Using Synthetic Data in Communication Sciences and Disorders to Promote Transparency and Reproducibility

James C. Borders1, Austin Thompson2, & Elaine Kearney3,4

1. Department of Speech, Language, and Hearing Sciences, Boston University
2. Department of Communication Sciences and Disorders, University of Houston
3. School of Health and Rehabilitation Sciences, University of Queensland, Brisbane, Australia
4. Department of Speech Pathology, Princess Alexandra Hospital, Brisbane, Australia

**Disclosures**:  
The authors have no financial or non-financial disclosures.

**Corresponding Author**:  
James C. Borders, PhD, CCC-SLP

**Authorship Contributions** (CRediT taxonomy - https://casrai.org/credit/)  
*Author Roles*: 1conceptualization, 2data curation, 3formal analysis, 4funding acquisition, 5investigation, 6methodology, 7project administration, 8resources, 9software, 10supervision, 11validation, 12visualization, 13writing – original draft, 14writing – reviewing & editing

JCB: 1, 2, 3, 6, 9, 12, 13  
AT: 1, 2, 3, 6, 9, 11, 12, 14  
EK: 1, 2, 3, 6, 9, 11, 12, 14

**Ethical Approval**: This study was deemed exempt by the Human Research Ethics Committee at the University of Queensland (#2024/HE001484).

**Keywords**: Open data; Reproducibility; Meta-science

# Abstract

**Purpose**: Reproducibility is a core principle of science and access to a study’s data is essential to reproduce its findings. However, data sharing is uncommon in the field of Communication Sciences and Disorders (CSD), often due to concerns related to privacy and disclosure risks. Synthetic data offers a potential solution to this barrier by generating artificial datasets that do not represent real individuals yet retain statistical properties and relationships from the original data. This study aimed to explore the feasibility and preliminary utility of synthetic data to promote transparency and reproducibility in the field of CSD.

**Method**: Ten open datasets were obtained from previously published research within the American Speech-Language-Hearing Association ‘Big Nine’ domains (articulation, cognition, communication, fluency, hearing, language, social communication, voice and resonance, and swallowing) across a range of study outcomes and designs. Synthetic datasets were generated with the *synthpop* R package. General utility was assessed visually and with the standardized ratio of the propensity mean squared error (*S\_pMSE*). Specific utility assessed whether inferential relationships from the original data were preserved in the synthetic dataset by comparing model fit indices, coefficients, and *p*-values.

**Results**: All synthetic datasets showed strong general utility, maintaining univariate and bivariate distributions. Six of nine synthetic datasets that used inferential statistics showed strong specific utility, maintaining inferential relationships from the original analyses. Specific utility was low in three datasets with hierarchical structures.

**Conclusion**: Findings suggest that synthetic data can effectively maintain statistical properties and relationships across a wide range of non-hierarchical data commonly seen in the field of CSD. Other approaches for hierarchical data need to be explored in future work. Researchers who use synthetic data should assess its utility in preserving their results for their own data and use-case.

# Introduction

Transparency and openness are fundamental tenets of science, with transparency referring to the ability of others to clearly understand how scientific conclusions were reached, including the data, methods, and analyses used (Aguinis et al., 2018). A key aspect of transparency is computational reproducibility, which refers to the ability to recreate a study’s results using the original data. Nowadays, the vast majority of scientific studies use some degree of computation in processing data, conducting descriptive or inferential statistics, and visualizing results. When these computations are reproducible, the transparency and confidence in findings are enhanced. Achieving computational reproducibility, however, requires authors to share their data. Both the National Institutes of Health and the National Science Foundation mandate data sharing and management plans to ensure that scientific data supporting a study is shared upon publication and aligns with FAIR (Findability, Accessibility, Interoperability, and Reuse) principles of digital assets (Watson et al., 2023; Wilkinson et al., 2016).

Providing open, publicly available data benefits scientists, funding bodies, and society at large by enabling researchers to verify results, generate new knowledge (e.g., meta-analyses, secondary analyses), develop hypotheses, and minimize redundant data collection (Chow et al., 2023). In this sense, sharing data promotes a cumulative and self-correcting science. Despite the clear benefits of open data and its growing adoption in other fields like psychology and the biobehavioral sciences (Quintana, 2020), only 26% of a sample of researchers in the field of Communication Sciences and Disorders (CSD) reported sharing their data publicly at least once (El Amin et al., 2023).

Understanding the nuances of data sharing requires a closer look at the different types of data generated throughout a research project’s life cycle. These include raw collected data, processed intermediate data, and final analysis data (Table 1). However, a common misconception is that open data refers solely to sharing raw data (e.g., audio recordings, videos, MRI data) (Pfeiffer et al., 2024). In reality, sharing intermediate or analysis data can also support reproducibility while reducing privacy and confidentiality concerns associated with sharing raw data. However, these different types of data offer varying levels of utility: sharing raw data enables maximum reproducibility and secondary research opportunities, while analysis data (although easier to share) primarily supports computational reproducibility.

##### Table 1 here.

Both individual and system-level barriers hinder data sharing, including a lack of time, knowledge, support from colleagues, and perceived incentives (Pfeiffer et al., 2024). Furthermore, each type of data comes with unique challenges regarding data sharing. For raw data, it is common that researchers often do not obtain consent to share data or cannot contact participants after data collection. Additionally, sharing de-identified raw or intermediate data may require additional approval from the institutional review board. Even when de-identification is possible, anonymized intermediate or analysis datasets can still carry re-identification risks, especially in small samples or vulnerable populations where indirect identifiers (e.g., gender, age, or race) may compromise participant confidentiality (Rocher et al., 2019). Therefore, although sharing de-identified analysis data is the minimum requirement for ensuring computational reproducibility and promoting cumulative science, concerns about privacy must be addressed when sharing sensitive data.

## Synthetic Data as an Approach to Promote Transparency and Reproducibility

Synthetic data generation offers a promising solution to safeguarding participants’ privacy and confidentiality in publicly available datasets (Drechsler & Haensch, 2024; Rubin, 1993). This approach can be applied to a wide variety of data types (e.g., raw, intermediate, or analysis data) and variables (e.g., qualitative demographic information or quantitative outcome measures). It involves creating artificial datasets that do not represent real individuals, thereby significantly reducing the risk of disclosure. Importantly, synthetic data retains the statistical properties and relationships of the original data, enabling readers to evaluate key aspects of the study’s analysis workflow, such as data pre-processing and statistical modeling, as well as develop and extend methodologies. Synthetic data generation has been used in government agencies, such as the United States Census Bureau (Jarmin et al., 2014) and the Government of Canada’s Directive on Open Government (Gauvin et al., 2021)to promote greater access to data and information. Although the concept of synthetic data methods was first proposed more than 30 years ago (Rubin, 1993), recent analytic and software developments have streamlined the process, making it easier and more efficient to generate high-quality synthetic data (Nowok et al., 2016).

Synthetic data are, however, not without limitations; the extent to which statistical properties of the original data are retained varies based on the dataset and the model used to synthesize the data (Latner et al., 2024; Matthews & Harel, 2011). The intended use of synthetic data influences the level of rigor and scrutiny required. For example, synthetic data can serve as a pedagogical tool to teach data analysis skills or novel statistical methods (Shepherd et al., 2017). In such cases, preserving general statistical properties is sufficient, even if precise relationships between variables are not fully maintained. Similarly, synthetic data accompanying publications can facilitate reproducible workflows to illustrate data pre-processing steps or statistical models without reproducing exact study results. However, higher standards are required when synthetic data is used for hypothesis testing, meta-analyses, or methodological development (Raab et al., 2017). In these scenarios, synthetic datasets must accurately preserve multivariable relationships to ensure their validity and utility.

Two main approaches are used to assess the utility of synthetic datasets: general and specific (Snoke et al., 2018). General utility evaluates whether the synthetic dataset maintains the overall statistical properties of the original dataset. This includes visual comparisons of univariate (e.g., bar charts, histograms) and bivariate joint distributions (e.g., scatterplots), as well as metrics to determine to what degree synthetic data is distinguishable from the original data (e.g., standardized propensity mean squared error; *S\_pMSE*). Specific utility assesses whether inferential relationships from the original dataset are preserved in the synthetic dataset by comparing model fit indices and coefficients.

## Application of Synthetic Data in Communication Sciences and Disorders

Despite its potential to enhance data sharing in the field of CSD, synthetic data is not widely known or adopted in the field. Data commonly collected in CSD research poses unique challenges, including smaller sample sizes than are typically recommended for synthetic data generation and a wide range of study designs, outcomes, and analyses (Borders et al., 2022; Gaeta & Brydges, 2020). Moreover, reproducible workflows that detail important steps for data wrangling or statistical modeling are rarely provided in publications, further hindering transparency and reproducibility.

To address this gap, the present study aimed to explore the feasibility and preliminary utility of generating synthetic analysis data in CSD. We applied synthetic data methods to open datasets from the ‘Big Nine’ American Speech-Language-Hearing Association (ASHA) domains and hypothesized that synthetic datasets would preserve both the statistical properties (general utility) and the inferential results (specific utility) of the original data. It’s important to recognize that synthetic data must be evaluated on a case-by-case basis and that the utility of the datasets included in this manuscript may not apply to one’s own dataset. To this end, the broad goal of the current investigation was to provide a proof-of-concept to the interested reader.

# Method

## Description of Original Datasets from ASHA ‘Big Nine’ Domains

A convenience sampling approach was used to identify publicly available datasets from previously published research articles related to the ‘Big Nine’ ASHA domains: swallowing (Curtis et al., 2023), articulation (Thompson et al., 2023), fluency (Elsherif et al., 2021), voice and resonance (Novotný et al., 2016), hearing (Battal et al., 2019), communication modalities (King et al., 2022), receptive and expressive language (Kearney et al., 2023; Robinaugh et al., 2024), cognitive aspects of communication (Clough et al., 2023), and social aspects of communication (Chanchaochai & Schwarz, 2023). The datasets were identified through searching keywords related to the ASHA domains on the Open Science Framework and other data aggregator sites (e.g., UK Data Service), as well as through the authors’ prior research. Given the prevalence of single subject experimental designs in the field of CSD, an additional study was included to ensure adequate representation (Robinaugh et al., 2024), resulting in ten studies. These studies were classified by their study design, population, and statistical analysis (Table 2).

It is important to note that not *all* research designs are represented due to the limited availability of public data in the field of CSD and the inherent challenge of including every possible study design. Instead, this approach was chosen to prioritize representation across all subfields to illustrate the application of synthetic data methods in CSD. To demonstrate the feasibility and preliminary utility of synthetic data, an analysis was chosen from each study and synthetic data was generated for those variables, as described below.

##### Table 2 here.

## Generation of Synthetic Datasets and Comparison with Original Dataset

Synthetic data generation and statistical analyses were conducted in R version 4.2.1 (R Core Team, 2022). The *synthpop* R package (version 1.8.0) (Nowok et al., 2016) was used to generate synthetic data via complete conditional specification (Drechsler & Haensch, 2024). This method synthesizes one variable at a time: the first variable is generated by random sampling from the original dataset, and subsequent variables are synthesized conditionally based on previously synthesized variables. This stepwise approach captures relationships between variables incrementally rather than attempting to synthesize all relationships simultaneously.

For example, consider a dataset containing four variables: participant ID, sex (categorical), age (continuous), and weight (continuous). The process would begin by synthesizing sex by estimating its distribution based on the original data and sampling synthetic values from that distribution. Age would then be synthesized conditionally based on the synthetic sex values, using a model that captures the relationship between sex and age in the original data. Finally, weight would be synthesized conditionally on both sex and age, again based on relationships estimated from the observed data. Participant IDs, which serve only as identifiers and do not contain meaningful information, would be randomly assigned after data synthesis is complete.

Synthpop inherently manages missing data and maintains relationships between missingness and other variables using a tree-based algorithm, specifically classification and regression trees (CART), for data synthesis (Nowok et al., 2016). Alternatively, users can select other tree-based methods, such as random forests, or parametric models like linear or logistic regression. This process resembles multiple imputation by chained equations (MICE) for handling missing data (Audigier et al., 2018) but with a key distinction: instead of imputing only missing values, synthpop generates entirely synthetic data (Raghunathan, 2021), significantly reducing disclosure risk.

Nowok et al. (2016) provide an in-depth overview of the synthpop package’s features. Briefly, synthesis is largely automated using the syn() function. Users can customize various options, including the modeling approach, choice of predictors, order of synthesized variables, smoothing parameters for continuous variables to enhance privacy, and rules for maintaining logical relationships.

## Evaluation of General and Specific Utility

In the present study, we aimed to explore the feasibility and preliminary utility of synthetic data to promote transparency and reproducibility in CSD. Utility was operationalized as general (does the synthetic data resemble the original data in its statistical properties and distribution?) and specific (is the inferential relationship between variables maintained?). To evaluate general utility, we visually compared univariate (e.g.., bar charts, histograms) and bivariate joint distributions (e.g., scatterplots) between the original and synthetic dataset, and evaluated the predicted probability that a record comes from the synthetic versus original data, known as the standardized propensity mean squared error (*S\_pMSE*). Standardized propensity scores closer to zero indicate greater general utility (typically with a standard deviation of one), where a value of zero indicates that the original and synthetic data are identical (Snoke et al., 2018). Notably, a value of zero is highly unlikely since synthetic data generation aims to achieve distributional similarity.

To assess specific utility, a statistical analysis was selected from each study and performed separately with the original and the synthetic data. Greater overlap in effect size or coefficient confidence intervals and similar p-value inferences (i.e., significant or non-significant) indicated greater specific utility. Since Curtis et al. (2023) examined median and interquartile ranges (IQR) instead of inferential statistical models, only general utility was examined.

We also examined the stability of results when generating multiple synthetic datasets. Specifically, our approach involved generating 100 different synthetic datasets for each original dataset. A statistical model with the original dataset was fit, and the *p*-value and effect size were recorded. We then evaluated whether 95% of *p*-values and effect sizes from the synthetic datasets demonstrated a similar result as the original study, Measures of effect size and their interpretation for each study are provided in Table 3. Additional information related to these analyses, as well as results and figures are provided in Appendix A. Since Curtis et al. (2023) did not perform inferential statistical models, we directly compared each synthetic dataset to the original data with a zero-inflated beta multilevel model with the *gamlss* package (version 5.4.3) (Stasinopoulos & Rigby, 2007), which included a fixed effect of dataset type and a random intercept of participant. The *p*-value from both zero-inflated and beta portions of the model were evaluated and *p* < .05 was interpreted as no statistically significant difference between the synthetic and original dataset. Since Robinaugh et al. (2024) used a Bayesian analysis, we compared synthetic versus original model estimates. The pre-registered analysis plan and corresponding deviations are publicly available on the Open Science Framework (https://osf.io/vhgq2).

##### Table 3 here.

# Results

### Swallowing

Curtis et al. (2023) provided normative reference values for swallowing outcomes during flexible endoscopic evaluations of swallowing in a sample of 39 community-dwelling adults without dysphagia. In this observational cohort study, each participant completed 15 swallowing trials that varied by bolus size, consistency, contrast agent, and swallowing instructions. Several swallowing outcomes were measured, including the amount of laryngeal vestibule residue, which was rated using the Visual Analysis of Swallowing Efficiency and Safety. No inferential statistics were conducted; instead, the distribution of laryngeal vestibule residue ratings was summarized using median and interquartile ranges.

In the synthetic dataset, the primary outcome (laryngeal vestibule residue ratings) closely mirrored the original data (Figure 2A). The frequency of zero values was nearly identical and the distribution of values greater than zero was also similar, with only minor deviations at higher residue ratings. The *S\_pMSE* value was 0.09, indicating strong overall similarity and general utility between the synthetic and original data.

When examined across 100 synthetic datasets, findings from the zero-inflated beta multilevel models indicate that 100% and 98% of synthetic datasets were not statistically significantly different than the original dataset for the zero-inflated and beta portions of the model, respectively. Additionally, effect size categorizations were maintained for 100% of both zero-inflated and beta portions of the model.

##### Figure 2 here.

### Articulation

Thompson et al. (2023) examined the relationship between vowel space area and speech intelligibility among 40 speakers with dysarthria of varying etiologies (e.g., Parkinson’s disease, amyotrophic lateral sclerosis, Huntington’s disease, and ataxia). A linear regression model revealed a statistically significant relationship between vowel space area and intelligibility (*p* < .001) with a Cohen’s *f* of 0.59, corresponding to a conventionally “large” effect size (Table 3).

Compared to original data, the synthetic data demonstrated a similar distribution (Figure 2B). General utility was high for both variables of vowel space area (*S\_pMSE* = 1.07) and intelligibility (*S\_pMSE* = 0.74). Specific utility was high as the statistical model with the synthetic data maintained the direction of statistical significance (*p* = .002) and effect size magnitude (*f* = 0.54).

Findings from the 100 generated synthetic datasets indicate that 71% of datasets demonstrated the same inferential result (i.e., a statistically significant *p*-value) and 57% maintained their effect size category.

### Fluency

Elsherif et al. (2021) compared non-word repetitions between 80 neurotypical adults and 34 adults who stutter. An independent samples t-test demonstrated a statistically significant difference in non-word repetitions between these groups (*p* < .001) with a large effect size (Δ = 1.26).

Compared to original data, the synthetic data similar distributions (Figure 2C). General utility was high for the outcome of non-word repetitions (*S\_pMSE* = 1.27) and the statistical model with the synthetic data maintained the direction of statistical significance (*p* = .004) and effect size magnitude (*f* = 1.32).

Findings from the 100 generated synthetic datasets indicate that 100% of datasets demonstrated the same inferential result (i.e., a statistically significant *p*-value) and maintained a ‘large’ effect size categorization.

### Voice and Resonance

Novotny et al. (2016) examined the relationship between acoustic measures of nasality variability and overall perceptual ratings in a heterogenous cohort of individuals with Parkinson’s disease, Huntington’s disease, and neurotypical adults. Results indicated a statistically significant relationship (*r* = 0.51, *p* < .001) between perceptual ratings and acoustic nasality variability. Compared to original data, the synthetic data demonstrated similar values for nasality variability within perceptual ratings of ‘Normal’ and ‘Mild’, though values for ‘Moderate’ perceptual ratings showed larger variance likely due to its low occurrence in the dataset (Figure 2D). General utility was high for both nasality variability (*S\_pMSE* = 0.68) and perceptual rating (*S\_pMSE* = 0.44). The statistical model with the synthetic data maintained the direction of statistical significance (*p* < .001) and effect size magnitude (*r* = 0.44), indicating adequate specific utility.

Findings from the 100 generated synthetic datasets indicate that 100% of datasets demonstrated the same inferential result (i.e., a statistically significant *p*-value) and 46% maintained their effect size category.

### Hearing

Battal et al. (2019) compared auditory localization abilities between 17 congenitally blind and 17 sighted individuals. A linear mixed effects model indicated that congenitally blind individuals had enhanced spatial hearing abilities (*OR* = 1.56, *p* = .016). Compared to original data, the synthetic data showed similar distributions for auditory localization in both sighted and congenitally blind individuals, as well as similar auditory localization at the subject-level (Figure 2E). General utility was high for subject (*S\_pMSE* = 0.53) and group (*S\_pMSE* = 0.17) variables. Specific utility was high as the statistical model with the synthetic data maintained the direction of statistical significance (*p* .018) and effect size magnitude (*OR* = 1.73). The random effect estimates were stable between the original (mean = 0, 95% CI: -0.027, 0.027) and synthetic (mean = 0, 95% CI: -0.036, 0.036) datasets.

Findings from the 100 generated synthetic datasets indicate that 78% of datasets demonstrated the same inferential result (i.e., a statistically significant *p*-value) and 100% maintained their effect size categorization.

### Communication Modalities

King et al. (2022) collected survey responses from speech-language pathologists to assess the impact of the COVID-19 pandemic on service provision for emergent bilinguals who use augmentative and alternative communication. Results indicated that speech-language pathologists reporting that a lack of or limited access to internet increased during the initial phase of the pandemic (Cohen’s 𝜔 = 34.60, *p* < .001). Compared to original data, the synthetic data showed similar frequencies of responses for the barrier of ‘lack of/limited internet’ (Figure 2F). General utility was high for assessment type (*S\_pMSE* = 0.03) and time point (*S\_pMSE* = 0.17) variables. The statistical model with the synthetic data maintained the direction of statistical significance (*p* < .001) and effect size magnitude (*r* = 35.38), indicating high specific utility.

Findings from the 100 generated synthetic datasets indicate that 100% of datasets demonstrated the same inferential result (i.e., a statistically significant *p*-value) and maintained their effect size categorization.

### Receptive and Expressive Language

Two studies were included in the domain of Receptive and Expressive Language (Kearney et al., 2023; Robinaugh et al., 2024).

Kearney et al. (2023) examined the relationship between years of education and reading performance among 36 individuals following left-hemisphere tumor resection. Results indicated a large relationship between these variables (*r* = 0.59, *p* < .001). Compared to original data, the synthetic data showed maintained a similar visual relationship between years of education and reading scores (Figure 3A). General utility was high for both years of education (*S\_pMSE* = 0.6) and reading scores (*S\_pMSE* = 0.6) variables. The statistical model with the synthetic data maintained the direction of statistical significance (*p* = .004) and effect size magnitude (*r* = 0.47), indicating high specific utility. Findings from the 100 generated synthetic datasets indicate that 98% of datasets demonstrated the same inferential result (i.e., a statistically significant *p*-value) and 70% of synthetic datasets maintained their effect size category.

Robinaugh et al., (2024) examined the effectiveness of a naming treatment in a single-case experimental design for an individual presenting with semantic variant primary progressive aphasia and a history of traumatic brain injury. An item-level Bayesian generalized mixed-effects model revealed that the treatment resulted in a gain of 35 out of 60 trained words (β = 35.3; 90% CI: 30.6, 39.5). Compared to original data, the synthetic data showed similar frequencies of responses, but not sessions (Figure 3B). General utility was high for id (*S\_pMSE* = 0.22), set (*S\_pMSE* = 0.09), session (*S\_pMSE* = 0.22), and phase (*S\_pMSE* = 0.03) variables. The statistical model with the synthetic data overestimated the effect size (β = 60.11; 90% CI: 54.41, 65.38), indicating that specific utility was low. Findings from the 100 generated synthetic datasets indicate that the average model estimate was 69 words, which was a very large overestimation compared to the original model estimate of 35 words.

### Cognitive Aspects of Communication

Clough et al. (2023) examined the interaction between group (traumatic brain injury [TBI] or neurotypical) and condition (basic emotion or social emotion emojis) on the accuracy of emotion recognition. A generalized linear mixed effects model indicated that participants with TBI were more likely to correctly identify basic emotions than social emotions when presented as emoji (*OR* = 1.9), *p* = 0.013), whereas neurotypical participants did not differ in their ability to identify these emotions. Compared to original data, the synthetic data showed a similar distribution of responses for both basic and social emotions for the TBI group (Figure 3C). General utility was high for subject (*S\_pMSE* = 0.23) and condition (*S\_pMSE* = 0.02) variables. Specific utility was low as the statistical model with the synthetic data did not maintain the direction of statistical significance (*p* = 0.059)), even though the effect size magnitude was still considered large (*OR* = 1.51). The random effect estimates were stable between the original (mean = -0.011, 95% CI: -0.018, -0.005) and synthetic (mean = -0.012, 95% CI: -0.018, -0.005) datasets.

Findings from the 100 generated synthetic datasets indicate that 35% of datasets demonstrated the same inferential result (i.e., a statistically significant *p*-value) and 100% of synthetic datasets maintained their effect size category.

### Social Aspects of Communication

Chanchaochai & Schwarz et al. (2023) compared non-verbal IQ between individuals with autism spectrum disorder and neurotypical peers. An analysis of variance indicated that neurotypical individuals demonstrated higher non-verbal IQ (*d* = -0.85, *p* < .001). Compared to original data, the synthetic data showed similar distributions of non-verbal IQ for both groups (Figure 3D). General utility was high for both group (*S\_pMSE* = 0.01) and non-verbal IQ (*S\_pMSE* = 0.18) variables. The statistical model with the synthetic data maintained the direction of statistical significance (*p* = .018); however, the effect size magnitude (*r* = -0.54) was lower, indicating a low level of specific utility.

Findings from the 100 generated synthetic datasets indicate that 87% of datasets demonstrated the same inferential result (i.e., a statistically significant *p*-value) and 47% of synthetic datasets maintained their effect size category.

# Discussion

Although computational reproducibility is a core principle of science, data sharing is uncommon in CSD, partly due to concerns regarding disclosure risk (Pfeiffer et al., 2024). This study demonstrates the utility of synthetic datasets to protect participant confidentiality while preserving the statistical properties and relationships of the original analysis data. The utility of synthetic data is further strengthened by the range of datasets included in the current study, which varied by domain (across nine ASHA domains), sample size (from 40 to >8,000 data points), statistical models (from simple correlations to multilevel model with 3-way interactions), and effect sizes (from conventionally “small” to “large”). These results suggest that synthetic datasets can be effectively used across a wide range of studies in the field of CSD to preserve participant confidentiality when sharing data.

The current findings illustrate the feasibility of generating synthetic datasets across a range of studies in the field of CSD. Studies were conveniently selected based on domain and data availability, but varied in design, sample size, population, and statistics employed. While it was possible to synthesize a dataset for all included studies, it is important to consider the accuracy of the synthesis with regards to the purpose of data sharing. All synthetic datasets in the current study showed strong general utility, which means they would be suitable for sharing for educational purposes or to demonstrate computational reproducibility for published analyses, while mitigating confidentiality concerns. Six of the nine studies with inferential statistics also demonstrated strong specific utility. These datasets could form the basis for further hypothesis-testing or inclusion in meta-analyses, while those with low specific utility should be excluded from such analyses.

One key finding is that low specific utility of synthetic datasets was not necessarily attributed to sample size, despite the *synthpop* package’s recommendation of a minimum of 130 observations for generating synthetic datasets (Nowok et al., 2016). For example, specific utility was low for a synthetic dataset from the cognition domain with over 8,000 observations. Instead, low specific utility was primarily associated with datasets containing a hierarchical structure, such as repeated measure or nested designs, which are common in CSD. This suggests that current synthesis methods in synthpop may not adequately capture multilevel dependencies. Alternative approaches designed specifically to handle hierarchical data (Gauvin, 2021) might offer improved solutions and should be explored in future work. Additionally, another potential method for handling more complex, hierarchical data involves creating synthetic datasets by sampling posterior predictive distributions through a fully Bayesian analytical approach; however, this method does not guarantee anonymization and should be used with caution when sharing confidential and highly sensitive data.

These findings highlight the importance of evaluating the accuracy of synthetic datasets. To ensure synthetic data quality, researchers should clearly define their intended purpose (e.g., educational, exploratory, inferential) and assess general and/or specific utility accordingly. If synthetic datasets fail to retain key relationships of the original data, they should not be used or shared. Comparisons between synthetic and original analyses should be made available, ideally in supplementary materials, to promote transparency. This information can also give the end-user researcher confidence in using a synthetic dataset for their own purposes; a limitation cited in the re-use of synthetic data (Matthews & Harel, 2011).

It is important to highlight the many benefits of sharing either raw or intermediate data. Sharing these types of data enhances research transparency by enabling readers to re-analyze raw data for different purposes or reproduce the calculations behind analysis data. Different operational definitions or analysis steps are often a barrier to inclusion in a meta-analysis; therefore, sharing this type of data ensures that secondary analyses can be performed with alternate methodologies or operational definitions as the field progresses. In this sense, sharing raw or intermediate data facilitates the generation of new knowledge and accelerates scientific discovery. Despite its many benefits, there are instances where sharing raw or intermediate data may not be feasible. For example, researchers may not have obtained consent from participants for data sharing, or the institutional review board may impose project-specific guidelines that restrict sharing this type of data. In these instances, synthetic data fills an important gap by enhancing the transparency of analysis and methods workflows.

## Limitations and future directions

This study is not without limitations. First, studies were selected in the present study because they were openly available and represented different subfields within CSD. Therefore, selection bias is likely present, and these studies are certainly not representative of every research design or data parameter that a researcher may encounter. It is imperative that the user evaluate the utility of synthetic data in the context of their own goals (e.g., educational, workflow transparency, or meta-analysis/hypothesis generation purposes) before publicly sharing the dataset.

Second, it is important to recognize that synthetic data are inherently a proxy and cannot entirely preserve all statistical properties of the original dataset. Whenever ethically permissible, researchers should prioritize sharing de-identified or identifiable data. Moreover, open data alone does not ensure computational reproducibility. Instead, open data must be accompanied by reproducible code and analysis scripts. In fact, recent research showed that a high percentage of findings from registered reports that provided open data alone were unable to be reproduced (Obels et al., 2020). Reproducible workflows in languages like R have been proposed and warrant consideration (Peikert et al., 2021).

Moving forward, broader systemic changes will be necessary to normalize and encourage responsible data sharing. Doctoral programs should offer formal training on open science, data sharing, and analysis practices that promote reproducibility. Fortunately, a wealth of resources is available to support researchers in learning these practices (Lewis, 2024). Academic institutions must also recognize open science activities as meaningful scholarly contributions. While ASHA’s implementation of open science badges is a positive step, more systemic efforts will be required to shift the culture away from individualism and toward a more collaborative, pro-social scientific community.

## Conclusions

This study evaluated the feasibility and use of the synthpop package in R for generating synthetic data in the field of CSD, particularly when sharing original data presents confidentiality risks. Findings suggest that synthetic data can effectively reproduce distributional and inferential properties in datasets without hierarchical structures. However, for hierarchical datasets, standard methods for synthesizing data generated using the synthpop package may not maintain key inferential relationships, limiting its suitability for some research applications. Therefore, researchers should rigorously assess the utility of synthetic datasets before sharing and ensure their intended purpose aligns with the capabilities of the synthesis method used.

Appendix A: Description, full results, and visualization of stability analysis

An additional stability analysis was performed to examine whether synthetic datasets maintained the statistical properties and relationships of the original dataset across multiple synthetic datasets. This approach involved generating 100 different synthetic datasets for each study’s original dataset. A statistical model was fit with the original dataset, and the *p*-value and effect size magnitude were recorded. If 95% of *p*-values and effect sizes from the synthetic datasets demonstrated a similar result as the original study, then this indicated that synthetic data maintained the statistical relationship across multiple generations. Specifically, we further defined this as a similar inferential result for *p*-values (i.e., a ‘significant’ or ‘non-significant’ p-value based on the original study’s alpha level) and effect sizes that maintained their categorization based on conventional thresholds (e.g., a ‘medium’ effect size). If variability between the 100 synthetic datasets was appreciated, we visualized and described the dispersion of this distribution. Full results are provided below in appendix table 1, and figures 1 and 2.

Of note, two studies required additional considerations. Since Curtis et al. (2023) did not perform inferential statistical models, we directly compared each synthetic dataset to the original data with a zero-inflated beta multilevel model with the gamlss package (version 5.4.3) (Stasinopoulos & Rigby, 2007), which included a fixed effect of dataset type and a random intercept of participant. The p-value from both zero-inflated and beta portions of the model were evaluated and p < .05 was interpreted as no statistically significant difference between the synthetic and original dataset.

Since Robinaugh et al. (2024) used a Bayesian analysis, we focused our analysis on the comparison of unadjusted model estimates between the synthetic and original data. No *p*-values are provided as this is not a component of a Bayesian statistical framework.

Appendix Table 1: Stability of synthetic datasets across ASHA domains.

| Domain | Study | Sample Size | *p*-value | Effect Size Measure | Effect Size | *p*-value Agreement | Effect Size Categorization Agreement |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Articulation | Thompson et al. (2023) | 40 | < .001 | Cohen’s f | .59 | 94% | 81% |
| Fluency | Elsherif et al. (2021) | 114 | < .001 | Glass' Δ | -2.18 | 100% | 100% |
| Voice and resonance | Novotný et al. (2016) | 111 | < .001 | Correlation coefficient | .51 | 100% | 46% |
| Hearing | Battal et al. (2019) | 372 | .016 | √(3/𝜋) x odds ratio | 1.56 | 78% | 100% |
| Communication modalities | King et al. (2022) | 463 | < .001 | Cohen's 𝜔 | 34.6 | 100% | 100% |
| Receptive and expressive language | Kearney et al. (2023) | 36 | < .001 | Correlation coefficient | .59 | 98% | 70% |
| Receptive and expressive language | Robinaugh et al. (2024) | 1 | N/A | Unadjusted model estimate | 35 | N/A | N/A |
| Cognitive aspects of communication | Clough et al. (2023) | 8,568 | .013 | √(3/𝜋) x odds ratio | 1.85 | 35% | 100% |
| Social aspects of communication | Chanchaochai & Schwarz (2023) | 96 | < .001 | Cohen’s d | -.85 | 87% | 47% |

N/A: Not applicable.

Appendix Figure 1: Distribution of log-transformed *p*-values in synthetic datasets across ASHA domains.

A group of graphs showing different types of data

AI-generated content may be incorrect.

*Each panel displays the distribution of log-transformed \*p\*-values across 100 synthetic datasets for a given ASHA domain. The dashed line indicates the threshold for statistical significance from the original study. Shaded green areas indicate synthetic \*p\*-values that maintained the statistical inferential result of the original study. The mean difference and standard deviation of raw \*p\*-values compared to the \*p\*-value reported in the original study is shown below each panel's title. Note that the domain of Receptive and Expressive Language refers to Kearney et al. (2023).*

Appendix Figure 2: Distribution of effect sizes in synthetic datasets across ASHA domains.

A collage of green and white graphs

Description automatically generated

*Each panel displays the distribution of effect sizes across 100 synthetic datasets for a given ASHA domain. The dashed line indicates the effect size reported in the original study and the light blue shaded area indicates the range of the effect size categorization. The mean difference and standard deviation of the effect size compared to the result reported in the original study is shown below each panel's title. Note that the x axis on panel F (Robinaugh et al., 2024) displays the difference in the unstandardized model estimate. Since there is no effect size categorization range, an overlay is not shown.*

# Acknowledgements

We would like to thank the authors of the studies included in this manuscript for making their data publicly available.

**Funding**: None.

**Study Preregistration and Data Availability**: The study preregistration (https://osf.io/vhgq2) and associated data and analysis scripts (https://osf.io/yhkqf/) are publicly available on the Open Science Framework.

# References

Audigier, V., White, I. R., Jolani, S., Debray, T. P., Quartagno, M., Carpenter, J., Van Buuren, S., & Resche-Rigon, M. (2018). *Multiple imputation for multilevel data with continuous and binary variables*. <https://projecteuclid.org/journals/statistical-science/volume-33/issue-2/Multiple-Imputation-for-Multilevel-Data-with-Continuous-and-Binary-Variables/10.1214/18-STS646.short>

Battal, C., Occelli, V., Bertonati, G., Falagiarda, F., & Collignon, O. (2019, May 16). *Ubiquitous enhancement of spatial hearing in congenitally blind people*. <https://doi.org/10.31234/osf.io/veh7t>

Borders, J. C., Grande, A. A., & Troche, M. S. (2022). Statistical Power and Swallowing Rehabilitation Research: Current Landscape and Next Steps. *Dysphagia*. <https://doi.org/10.1007/s00455-022-10428-2>

Chanchaochai, N., & Schwarz, F. (2023). Difficulties with pronouns in autism: Experimental results from Thai children with autism. *Language Acquisition*, *0*(0), 1–38. <https://doi.org/10.1080/10489223.2023.2262457>

Chow, J. C., Sandbank, M., & Hampton, L. H. (2023). Guidance for Increasing Primary Study Inclusion and the Usability of Data in Meta-Analysis: A Reporting Tutorial. *Journal of Speech, Language, and Hearing Research*. <https://doi.org/10.1044/2023_JSLHR-22-00318>

Clough, S., Morrow, E., Mutlu, B., Turkstra, L., & Duff, M. C. (2023). Emotion recognition of faces and emoji in individuals with moderate-severe traumatic brain injury. *Brain Injury*, *37*(7), 596–610. <https://doi.org/10.1080/02699052.2023.2181401>

Curtis, J. A., Borders, J. C., Dakin, A. E., & Troche, M. S. (2023). Normative Reference Values for FEES and VASES: Preliminary Data From 39 Nondysphagic, Community-Dwelling Adults. *Journal of Speech, Language, and Hearing Research*. <https://doi.org/10.1044/2023_JSLHR-23-00132>

Drechsler, J., & Haensch, A.-C. (2024). 30 Years of Synthetic Data. *Statistical Science*, *39*(2), 221–242. <https://doi.org/10.1214/24-STS927>

El Amin, M., Borders, J. C., Long, H. L., Keller, M. A., & Kearney, E. (2023). Open Science Practices in Communication Sciences and Disorders: A Survey. *Journal of Speech, Language, and Hearing Research*, *66*(6). <https://doi.org/10.1044/2022_JSLHR-22-00062>

Elsherif, M. M., Wheeldon, L. R., & Frisson, S. (2021). Do dyslexia and stuttering share a processing deficit? *Journal of Fluency Disorders*, *67*, 105827. <https://doi.org/10.1016/j.jfludis.2020.105827>

Gaeta, L., & Brydges, C. R. (2020). An Examination of Effect Sizes and Statistical Power in Speech, Language, and Hearing Research. *Journal of Speech, Language, and Hearing Research*, *63*(5), 1572–1580. <https://doi.org/10.1044/2020_JSLHR-19-00299>

Gauvin, H. (2021). Generating smart deep files: The example of synthesizing hierarchical data. *Proceedings of Statistics Canada Symposium*.

Jarmin, R. S., Louis, T., & Miranda, J. (2014, February 1). *Expanding the Role of Synthetic Data at the U.S. Census Bureau* (SSRN Scholarly Paper No. 2408030). <https://doi.org/10.2139/ssrn.2408030>

Kearney, E., Brownsett, S. L., Copland, D. A., Drummond, K. J., Jeffree, R. L., Olson, S., Murton, E., Ong, B., Robinson, G. A., & Tolkacheva, V. (2023). Relationships between reading performance and regional spontaneous brain activity following surgical removal of primary left-hemisphere tumors: A resting-state fMRI study. *Neuropsychologia*, *188*, 108631. <https://www.sciencedirect.com/science/article/pii/S0028393223001653>

King, M., Ward, H., Soto, G., & Barrett, T. S. (2022). Supporting Emergent Bilinguals Who Use Augmentative and Alternative Communication and Their Families: Lessons in Telepractice From the COVID-19 Pandemic. *American Journal of Speech-Language Pathology*, *31*(5), 2004–2021. <https://doi.org/10.1044/2022_AJSLP-22-00003>

Latner, J., Neunhoeffer, M., & Drechsler, J. (2024). Generating Synthetic Data is Complicated: Know Your Data and Know Your Generator. In J. Domingo-Ferrer & M. Önen (Eds.), *Privacy in Statistical Databases* (Vol. 14915, pp. 115–128). Springer Nature Switzerland. <https://doi.org/10.1007/978-3-031-69651-0_8>

Matthews, G. J., & Harel, O. (2011). *Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy*. <https://projecteuclid.org/journals/statistics-surveys/volume-5/issue-none/Data-confidentiality--A-review-of-methods-for-statistical-disclosure/10.1214/11-SS074.short>

Novotný, M., Rusz, J., Čmejla, R., Růžičková, H., Klempíř, J., & Růžička, E. (2016). Hypernasality associated with basal ganglia dysfunction: Evidence from Parkinson’s disease and Huntington’s disease. *PeerJ*, *4*, e2530. <https://doi.org/10.7717/peerj.2530>

Nowok, B., Raab, G. M., & Dibben, C. (2016). Synthpop: Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software*, *74*(11). <https://doi.org/10.18637/jss.v074.i11>

Obels, P., Lakens, D., Coles, N. A., Gottfried, J., & Green, S. A. (2020). Analysis of Open Data and Computational Reproducibility in Registered Reports in Psychology. *Advances in Methods and Practices in Psychological Science*, *3*(2), 229–237. <https://doi.org/10.1177/2515245920918872>

Peikert, A., van Lissa, C. J., & Brandmaier, A. M. (2021). Reproducible Research in R: A Tutorial on How to Do the Same Thing More Than Once. *Psych*, *3*(4, 4), 836–867. <https://doi.org/10.3390/psych3040053>

Pfeiffer, D. L., Thompson, A., Ciullo, B., Hirsch, M. E., El Amin, M., Ford, A., Riccardi, J., & Kearney, E. (2024). “1-800-Help-Me-With-Open-Science-Stuff”: A Qualitative Examination of Open Science Practices in Communication Sciences and Disorders. *Journal of Speech, Language, and Hearing Research*. <https://doi.org/10.1044/2024_JSLHR-24-00378>

Quintana, D. S. (2020). A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation. *eLife*, *9*, 1–12. <https://doi.org/10.7554/eLife.53275>

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>

Raab, G. M., Nowok, B., & Dibben, C. (2017, December 12). *Guidelines for Producing Useful Synthetic Data*. <https://doi.org/10.48550/arXiv.1712.04078>

Raghunathan, T. E. (2021). Synthetic Data. *Annual Review of Statistics and Its Application*, *8*(1), 129–140. <https://doi.org/10.1146/annurev-statistics-040720-031848>

Robinaugh, G., Henry, M. L., Cavanaugh, R., & Grasso, S. M. (2024). Computer-Based Naming Treatment for Semantic Variant Primary Progressive Aphasia With History of Traumatic Brain Injury: A Single-Case Experimental Design. *Journal of Speech, Language, and Hearing Research*, *67*(2), 524–544. <https://doi.org/10.1044/2023_JSLHR-23-00289>

Rocher, L., Hendrickx, J. M., & de Montjoye, Y.-A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications*, *10*(1), 3069. <https://doi.org/10.1038/s41467-019-10933-3>

Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of Official Statistics*, *9*, 462–468.

Shepherd, B. E., Blevins Peratikos, M., Rebeiro, P. F., Duda, S. N., & McGowan, C. C. (2017). A Pragmatic Approach for Reproducible Research With Sensitive Data. *American Journal of Epidemiology*, *186*(4), 387–392. <https://doi.org/10.1093/aje/kwx066>

Snoke, J., Raab, G. M., Nowok, B., Dibben, C., & Slavkovic, A. (2018). General and Specific Utility Measures for Synthetic Data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *181*(3), 663–688. <https://doi.org/10.1111/rssa.12358>

Thompson, A., Hirsch, M. E., Lansford, K. L., & Kim, Y. (2023). Vowel Acoustics as Predictors of Speech Intelligibility in Dysarthria. *Journal of Speech, Language, and Hearing Research*, *66*, 3100–3114. <https://doi.org/10.1044/2022_JSLHR-22-00287>

Watson, H., Gallifant, J., Lai, Y., Radunsky, A. P., Villanueva, C., Martinez, N., Gichoya, J., Huynh, U. K., & Celi, L. A. (2023). Delivering on NIH data sharing requirements: Avoiding Open Data in Appearance Only. *BMJ Health & Care Informatics*, *30*(1), e100771. <https://doi.org/10.1136/bmjhci-2023-100771>

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. <https://doi.org/10.1038/sdata.2016.18>

# Table and Figure Captions

Table 1: Description of types of data.

Table 2: Characteristics of included studies by ASHA domain.

Table 3: Effect size measures and interpretation by statistical test.

Figure 1. Illustration of a data sharing strategy.

Figure 2. Visualization of original and synthetic data for swallowing, articulation, fluency, voice, hearing, and communication modality domains.

Figure 3. Visualization of original and synthetic data for language, cognition, and social communication domains.

*Caption*: Panel A displays the distribution of vowel space area and panel B displays the distribution of speech intelligibility.

Supplemental Table 1: Stability of synthetic datasets across ASHA domains.

Supplemental Figure A. Distribution of log-transformed *p*-values in synthetic datasets across ASHA domains.

*Caption*: Each panel displays the distribution of log-transformed *p*-values across 100 synthetic datasets for a given ASHA domain. The dashed line indicates the threshold for statistical significance from the original study. Shaded green areas indicate synthetic *p*-values that maintained the statistical inferential result of the original study. The mean difference and standard deviation of raw *p*-values compared to the *p*-value reported in the original study is shown below each panel’s title.

Supplemental Figure B. Distribution of effect sizes in synthetic datasets across ASHA domains.

*Caption*: Each panel displays the distribution of effect sizes across 100 synthetic datasets for a given ASHA domain. The dashed line indicates the effect size reported in the original study and the light blue shaded area indicates the range of the effect size categorization. The mean difference and standard deviation of the effect size compared to the result reported in the original study is shown below each panel’s title.