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; Communication sciences and disorders

# 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 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 with 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 computational reproducibility playing a key role in maintaining these values. Computational reproducibility 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: Description of types of data.

|  | Raw Data | Intermediate Data | Analysis Data |
| --- | --- | --- | --- |
| Description | Original, unmodified data collected from studies.  Also known as primary data, microdata, individual-level data. | Cleaned, de-identified, and processed data, used for creating the analysis data.  Also known as cleaned data, transactional data, processed data. | Final dataset containing variables used in statistical analyses.  Also known as derived data, result data, aggregate data. |
| **Examples** |  |  |  |
| Acoustic Data (Thompson et al., 2023) | Raw audio recordings. | Extracted formant data. | Dataset containing average vowel space area per speaker. |
| Swallowing Data (Curtis et al., 2023) | Video files from flexible endoscopic evaluations of swallowing. | Not applicable. | Dataset containing ratings of swallowing safety and efficiency. |
| Eye Tracking Data (Baron et al., 2023) | Eye movement recordings (gaze, saccades, fixations). | Cleaned data with merged fixations and removed artifacts. | Dataset containing summary of fixation durations, reading times, and target proportions. |
| Focus Group/Interview Data (Pfeiffer et al., 2024) | Audio recordings and transcripts. | De-identified transcripts. | Dataset containing coded categories and themes from qualitative analysis. |
| Survey Data (Riccardi, 2024) | Raw survey responses. | Cleaned and coded responses. | Dataset containing summary scores and frequencies of survey responses. |
| Assessment Tool Data (Pfeiffer & Landa, 2024) | Raw, unscored assessment protocols. | Scored protocols with calculated totals and subscales. | Dataset containing standard scores, confidence intervals, percentile ranks. |

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., demographic information, outcome measures) and 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 (e.g., data pre-processing, statistical modeling), reproduce study findings, explore datasets, and develop new questions or hypotheses. Synthetic data generation is widely used across medical research, industry, and government agencies, most notably by the United States Census Bureau (Jarmin et al., 2014). 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., 2014; 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 synthetic data generation 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).

Table 2: Characteristics of included studies by ASHA domain.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Study** | **Domain** | **Study Design** | **Sample Size** | **Population(s)** | **Analysis of Interest** | **Outcome Type(s)** | **Statistics** |
| Curtis et al. (2023) | Swallowing | Descriptive, observational,  Cohort study | 39 | Community-dwelling adults | Distribution of laryngeal vestibule residue ratings | Continuous | Descriptive |
| Thompson et al. (2023) | Articulation | Analytical, observational,  Between-subject, cross-sectional study | 40 | Parkinson’s disease, amyotrophic lateral sclerosis, Huntington’s disease, cerebellar ataxia | Relationship between vowel space area and intelligibility | Continuous | Hierarchical linear regression |
| Elsherif et al. (2021) | Fluency | Analytical, observational,  Between-subject, cross-sectional study | 164 | Dyslexia, stuttering, neurotypical | Group difference in nonword repetition | Continuous | Independent t-test |
| Novotný et al. (2016) | Voice and resonance | Analytical, observational,  Between-subject, cross-sectional study | 111 | Parkinson’s disease, Huntington’s disease, neurotypical | Relationship between overall perceptual rating and variability of nasality | Continuous | Pearson correlation |
| Battal et al. (2019) | Hearing | Analytical, experimental,  Between- and within-subject, cross-sectional study | 34 | Congenitally blind, sighted | Group difference in auditory localization | Continuous | Linear mixed-effects model with 3-way interaction |
| King et al. (2022) | Communication modalities | Analytical, observational,  Between-subject, cross-sectional study | 160 | Speech-language pathologists | Timepoint difference in lack of/limited internet and technology barriers | Ordinal | Chi-square |
| Kearney et al. (2023) | Receptive and expressive language | Analytical, observational,  Between- and within-subject, cross-sectional study | 34 | Brain tumor | Relationship between years of education and reading score | Continuous | Spearman’s rank correlation coefficient |
| Robinaugh et al. (2024) | Receptive and expressive language | Analytical, experimental,  Within-subject, single-case study | 1 | Primary progressive aphasia with a history of traumatic brain injury |  | Binary | Bayesian generalized mixed effects model |
| Clough et al. (2023) | Cognitive aspects of communication | Analytical, experimental,  Between- and within-subject, cross-sectional study | 102 | Traumatic brain injury, neurotypical | Group x Condition interaction in emotion recognition accuracy | Binary | Generalized linear mixed-effects model with 3-way interaction |
| Chanchaochai & Schwarz (2023) | Social aspects of communication | Analytical, observational,  Between-subject, cross-sectional study | 96 | Autism spectrum disorder, neurotypical | Group difference in non-verbal IQ | Continuous | Analysis of Variance |

## 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.

## Generation of Synthetic Datasets with the Synthpop Package

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 2011). 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 three variables: participant ID, age, and weight. The process would begin by synthesizing participant ID through random sampling from its observed distribution. Age would then be synthesized conditionally based on the synthetic participant ID values, with synthetic values drawn from predictions informed by the original data. Finally, weight would be synthesized conditionally on both participant ID and age, with synthetic values similarly sampled from predictions.

*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 (van Buuren, 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 statistical models, only general utility was examined. The pre-registered analysis plan and corresponding deviations are publicly available on the Open Science Framework (<https://osf.io/vhgq2>).

# 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 1A). 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.

##### Figure 1 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 1). 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).

### 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 1C). 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).

### 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 1D). 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.

### 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 1E). 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 (r = 1.69). The random effect estimates were stable between the original (mean = 0, 90% CI: -0.027, 0.027) and synthetic (mean = 0, 90% CI: -0.036, -0.036) datasets.

### 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 1F). 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.

### Receptive and Expressive Language

Two studies were included in the domain of Receptive and Expressive Language (Kearney et al., 2023; **robinaugh\_etal24a?**).

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 2A). 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.

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 2B). 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.

### 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 = 1.90, *p* = .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 X). 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 = .059), even though the effect size magnitude was still considered large (OR = 1.38). The random effect estimates were stable between the original (mean = -0.011, 90% CI: -0.018, 0.005) and synthetic (mean = -0.016, 90% CI: -0.018, -0.005) datasets.

### 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 2D). 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.

##### Figure 3 here.

##### Figure 4 here.

##### Table 4 here.

# Discussion

Although computational reproducibility is a core principle of science, data sharing remains uncommon in the field of CSD, often due to concerns regarding disclosure risk (Pfeiffer et al., 2024). This study demonstrates the feasibility of synthetic datasets to promote transparency and reproducibility while protecting participant confidentiality. Overall, synthetic datasets showed strong general utility, preserving univariate and bivariate distributions. However, inferential relationships (specific utility) were only maintained in datasets without hierarchical structures. Key findings and considerations for the implementation of synthetic data in CSD are outlined below.

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 specifically designed to handle hierarchical data (Gauvin et al., 2021) may offer a better solution and should be explored in future work.

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).

***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, synthetic data generated using *synthpop* 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.

# Acknowledgements

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

**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.

**Funding:** None.

# References

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>

Drachen, T. M., Ellegaard, O., Larsen, A. V., & Dorch, S. B. F. (2016). Sharing data increases citations. *LIBER Quarterly: The Journal of the Association of European Research Libraries*, *26*(2, 2), 67–82. <https://doi.org/10.18352/lq.10149>

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

Eisenhauer, J. G. (2021). Meta‐analysis and mega‐analysis: A simple introduction. *Teaching Statistics*, *43*(1), 21–27. <https://doi.org/10.1111/test.12242>

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>

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics*, *15*(3), 651–674. <https://doi.org/10.1198/106186006X133933>

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>

Lewis, C. (2024). *Data Management in Large-Scale Education Research*. CRC Press. <https://books.google.com/books?hl=en&lr=&id=z54LEQAAQBAJ&oi=fnd&pg=PT12&dq=Data+Management+in+Large-Scale+Education+Research&ots=6h09wWzI1A&sig=-Eq8b942JpPzrApSvO2M2y8MhCo>

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>

Ohmann, C., Banzi, R., Canham, S., Battaglia, S., Matei, M., Ariyo, C., Becnel, L., Bierer, B., Bowers, S., Clivio, L., Dias, M., Druml, C., Faure, H., Fenner, M., Galvez, J., Ghersi, D., Gluud, C., Groves, T., Houston, P., … Demotes-Mainard, J. (2017). Sharing and reuse of individual participant data from clinical trials: Principles and recommendations. *BMJ Open*, *7*(12), e018647. <https://doi.org/10.1136/bmjopen-2017-018647>

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., Thompson, A., Ciullo, B., Hirsch, M. E., Amin, M. E., Ford, A., Riccardi, J. S., & Kearney, E. (2024, May 31). *1-800-Help-Me-With-Open-Science-Stuff: A Qualitative Examination of Open Science Practices in Communication Sciences and Disorders*. <https://doi.org/10.31219/osf.io/9kxa7>

Piwowar, H. A., Day, R. S., & Fridsma, D. B. (2007). Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLOS ONE*, *2*(3), e308. <https://doi.org/10.1371/journal.pone.0000308>

Piwowar, H. A., & Vision, T. J. (2013). Data reuse and the open data citation advantage. *PeerJ*, *1*, e175. <https://doi.org/10.7717/peerj.175>

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/>

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.

Soule MD, M. C., Beale BA, E. E., Suarez MD, L., Beach MD, S. R., Mastromauro MSW, L., Carol A., Celano MD, C. M., Moore BA, S. V., & Huffman MD, J. C. (2016). Understanding motivations to participate in an observational research study: Why do patients enroll? *Social Work in Health Care*, *55*(3), 231–246. <https://doi.org/10.1080/00981389.2015.1114064>

Stasinopoulos, D. M., & Rigby, R. A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, *23*(7), 1–46.

Tedersoo, L., Küngas, R., Oras, E., Köster, K., Eenmaa, H., Leijen, Ä., Pedaste, M., Raju, M., Astapova, A., Lukner, H., Kogermann, K., & Sepp, T. (2021). Data sharing practices and data availability upon request differ across scientific disciplines. *Scientific Data*, *8*(1), 192. <https://doi.org/10.1038/s41597-021-00981-0>

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>

Vazire, S., & Holcombe, A. O. (2022). Where Are the Self-Correcting Mechanisms in Science? *Review of General Psychology*, *26*(2), 212–223. <https://doi.org/10.1177/10892680211033912>

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>

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. <https://doi.org/10.21105/joss.01686>

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>

Yu, Y., & Romero, D. M. (2024). Does the use of unusual combinations of datasets contribute to greater scientific impact? *Proceedings of the National Academy of Sciences*, *121*(41), e2402802121. <https://doi.org/10.1073/pnas.2402802121>

# 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.

Table 4: Stability of synthetic datasets across ASHA domains.

Figure 1. Visualization of data distributions from synthetic and original data for Study #1 (Curtis et al., 2023).

*Caption*: Panel A displays the overall distribution of laryngeal vestibule residue. Panel B displays the frequency of values by bolus consistency.

Figure 2. Visualization of data distributions from synthetic and original data for Study #2 (Thompson et al., 2023).

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

Figure 3. 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.

Figure 4. 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.

Figure 5. Decision tree for data sharing.