Synthetic Data in Communication Sciences and Disorders:  
Promoting an Open, Reproducible, and Cumulative Science [preprint]

James C. Borders1, Austin Thompson2, & Elaine Kearney^3, 4^

1. Department of Biobehavioral Sciences, Teachers College Columbia 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  
jcb2271@tc.columbia.edu

**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

**Funding**: None.

**Ethical Approval**: This study was deemed exempt by the Institutional Review Board 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; however, data sharing is uncommon in the field of Communication Sciences and Disorders and exacerbated by 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. The present study evaluates the performance of synthetic data generation using open data from previously published studies across the ‘Big Nine’ domains defined by the American Speech-Language Hearing Association.

**Method**: Open datasets were obtained from previously published research across the domains of Articulation, Cognition, Communication, Fluency, Hearing, Language, Social Communication, Voice and Resonance, and Swallowing. Synthetic datasets were generated with the *synthpop* R package. Results from synthetic datasets were compared to those from the original published datasets.

**Results**: Synthetic datasets maintained the direction of the *p*-value in six studies and effect size categorization in five out of nine studies. **Add more here**.

**Conclusion**: Findings indicate that… We provide a general framework to promote sharing open data to facilitate computational reproducibility and a cumulative science.

# Introduction

Transparency and openness are fundamental tenets of science. One aspect of transparency and openness in science relates to computational reproducibility, or the ability to recreate a study’s results using the original data. Nowadays, the vast majority of scientific studies use some degree of computation, including processing data, conducting descriptive or inferential statistics, or visualizing results. When these computations are reproducible, the transparency and confidence in findings is enhanced. Achieving computational reproducibility, however, requires authors to share their data. Both the National Institutes of Health and 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. 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 researchers in the field of Communication Sciences and Disorders (CSD) reported sharing their data publicly at least once (El Amin et al., 2023). Both individual and system-level barriers hinder data sharing, including a lack of time, knowledge, support from colleagues, and perceived incentives. Privacy and confidentiality concerns, particularly in low-incidence populations, also pose significant challenges (Pfeiffer et al., 2024). Researchers have traditionally attempted to minimize disclosure risks by anonymizing datasets, aggregating results, or releasing a subset of the dataset; however, these practices do not fully eliminate the risk of identification in low-incidence populations. For example, re-identifying an individual in an incomplete dataset requires only a few demographic attributes (Rocher et al., 2019). A further challenge in sharing data can occur when researchers did not prospectively obtain consent to share data and may not be able to contact participants after data collection (Pfeiffer et al., 2024).

Synthetic data generation offers a potential solution to maintaining participants’ privacy and confidentiality in publicly available datasets (Drechsler & Haensch, 2024; Rubin, 1993). Synthetic data involves creating artificial datasets that do not represent real individuals, ensuring no risk of disclosure since participants in the synthetic dataset do not correspond to real individuals. Importantly, synthetic data retains the statistical properties and relationships of the original data, allowing researchers to reproduce study findings, explore the dataset, and develop new questions and 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). Though synthetic data methods were proposed more than 30 years ago (Rubin, 1993), recent analytic and practical developments have made it easier and more efficient to generate high-quality synthetic data (Nowok et al., 2016).

Despite the potential utility of synthetic data to promote open data in the field of CSD, this approach 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 (Borders et al., 2022; Gaeta & Brydges, 2020). Therefore, the present study aimed to examine the utility of synthetic data generation with open datasets from the ‘Big Nine’ American Speech-Language Hearing Association (ASHA) domains. We hypothesize that synthetic datasets will maintain the statistical properties and relationships of the original datasets, and that synthetic data will remain stable when generating multiple datasets. A secondary goal is to provide a framework for researchers in CSD to use data synthesis as a means to share fully de-identified data, thereby addressing concerns regarding researcher knowledge and participant confidentiality in sharing data.

# Method

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

Authors performed a manual search to obtain publicly available datasets from previously published research articles related to ‘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), cognitive aspects of communication (Clough et al., 2023), and social aspects of communication (Chanchaochai & Schwarz, 2023). Authors then reproduced an analysis from each study. Table 1 provides a description of the population, analysis, and open materials for each study.

##### Table 1 here.

## Generation of Synthetic Datasets and Comparison with Original Dataset

All data generation and analyses were conducted in R version 4.2.1 (**rcoreteam22?**). Synthetic data was generated with the *synthpop* R package version 1.8.0 (Nowok et al., 2016). Specifically, *synthpop* uses a non-parametric classification and regression tree (CART) approach that can handle any data type and generates data by sampling from a probability distribution. Our aims were twofold: (1) to determine whether a synthetic dataset maintained statistical properties and relationships of the original dataset and (2) to examine whether this remained stable when generating multiple synthetic datasets. In light of these aims, our approach involved generating 100 different synthetic datasets for each original dataset from an ASHA ‘Big Nine’ domain. A statistical model with the original dataset was fit and the *p*-value and effect size 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. 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 (e.g., a ‘medium’ effect size). Measures of effect size and their interpretation for each study are provided in Table 2. If variability between the 100 synthetic datasets was appreciated, we described the dispersion of this distribution. The analysis plan for this study was preregistered on the Open Science Framework (https://osf.io/vhgq2).

##### Table 2 here.

In addition to these inferential comparisons, we provide a tutorial to walk through the required steps to generate synthetic data for the reader. This is accomplished in the context of two datasets (Curtis et al., 2023; Thompson et al., 2023) with additional data visualization and detailed R code. Since curtis\_etal23a did not perform inferential tests, 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 (**stasinopoulosGeneralizedAdditiveModels2007?**). This model included fixed effects of dataset type (synthetic/original) and bolus consistency (thin liquid/extremely thick/regular) and a random intercept of participant. Due to issues with model convergence, the fixed effect structure was simplified to only include dataset type. The *p*-value from both zero-inflated and beta portions of the model were evaluated and *p* < .05 was interpreted as evidence of no statistically significant difference between the synthetic and original dataset.

The tutorial data and accompanying code can be accessed on the Open Science Framework (OSF: **add link**). To get started, download… Open the *open-and-synthetic-data.Rproj* file in RStudio and then X file.

# Results

### Study 1: Normative Reference Values for Swallowing Outcomes

Curtis et al. (2023) examined normative reference values for swallowing outcomes during flexible endoscopic evaluations of swallowing among 39 non-dysphagic, community-dwelling adults. In this observational cohort study, participants were administered 15 swallowing trials that varied by bolus size, consistency, contrast agent, and swallowing instructions. A variety of swallowing outcomes were measured, including the amount of laryngeal vestibule residue rated with the Visual Analysis of Swallowing Efficiency and Safety. Median and interquartile ranges (IQR) were used to describe the distribution of laryngeal vestibule residue ratings.

To generate synthetic data, we first load in the original dataset, wrangle the dataset using the *tidyverse* packages Wickham et al. (2019) and then create a synthetic dataset with the syn() function in the *synthpop* package. The data wrangling steps include loading required R packages, importing the original dataset csv file, ‘cleaning’ variable names to ensure they are all lowercase, selecting only relevant variables in the dataset, converting relevant categorical variables to factors, and expressing the outcome variable as a percentage out of 100.

# load required packages  
library(tidyverse) # data wrangling  
library(synthpop) # R package to generate synthetic data  
  
# load original data  
swallowing\_original\_data <-  
 # read csv file from appropriate path  
 read.csv(here::here("Data/01\_Swallowing/norms\_ratings.csv")) |>  
 # clean variable names  
 janitor::clean\_names() |>  
 # select only relevant variables from dataset  
 dplyr::select(c(study\_id, bolus\_consistency,   
 laryngeal\_vestibule\_severity\_rating)) |>   
 mutate(  
 # convert study\_id and bolus\_consistency to factors  
 study\_id = as.factor(study\_id),  
 bolus\_consistency = as.factor(bolus\_consistency),  
 # express laryngeal\_vestibule\_severity\_rating as a %  
 laryngeal\_vestibule\_severity\_rating = laryngeal\_vestibule\_severity\_rating/100  
 )

Next, we create a synthetic dataset with the syn() function from the *synthpop* package. Within the function, ‘method’ specifies the synthesising method for the data. The default in synthpop is “cart” (Classification and Regression Tree). If a synthetic dataset fails to generate with this method, Nowok et al. (2018) recommend an alternative implementation of the CART technique from package *party* (Hothorn et al., 2006). This dataset, for example, required the ‘ctree’ CART specification. Next, specify the number of synthetic datasets to generate within ‘m’. Once the synthetic dataset is generated, extract and convert it to a dataframe for additional wrangling and visualization.

# Create a synthetic dataset  
synthetic\_data <-  
 syn(swallowing\_original\_data, # name of the original data  
 method = "ctree", # CART model to generate synthetic data  
 m = 1 # number of synthetic datasets to generate  
 )

Synthesis  
-----------  
 study\_id bolus\_consistency laryngeal\_vestibule\_severity\_rating

# Extract the synthetic dataset and convert into a data frame  
synthetic\_dataset <- as.data.frame(synthetic\_data$syn)

An important step in the process is to assess the general utility of the synthetic dataset by visualizing any obvious differences compared to the original dataset. This can be easily accomplished with the compare() function in the *synthpop* package or manually with data wrangling and the ggplot package. Figure 1 suggests that the synthetic dataset demonstrated similar distributions for the variables of bolus consistency and laryngeal vestibule residue rating.

# Comparison of original and synthetic datasets with synthpop package  
swallowing\_comparison <- compare(  
 synthetic\_dataset, # synthetic dataset  
 swallowing\_original\_data, # original dataset  
 vars = c("bolus\_consistency",  
 "laryngeal\_vestibule\_severity\_rating"), # variables for comparison  
 stat = "counts", # Present the raw counts for each variable  
 cols = c("#62B6CB", "#1B4965") # Setting the colours in the plot  
)

Descriptively, the synthetic dataset classified 64% of laryngeal vestibule ratings on thin liquid boluses as ‘absent’ (i.e., 0% residue) compared to 68% in the original dataset. In the synthetic dataset, the median value on thin liquids was 0.03 (IQR: 0.02 - 0.045) compared to 0.03 (IQR: 0.02 - 0.04) in the original dataset. 98.61% of extremely thick liquids were classified as having no laryngeal vestibule residue compared to 100% in the original dataset. A similar pattern was appreciated for regular solids (96.43% in synthetic vs. 100% in original dataset). Number of trials was lower for extremely thick (77 trials) and regular solid (78 trials) boluses compared to thin liquid (429 trials). When examined across 100 synthetic datasets, findings from the zero-inflated beta multilevel models indicate that 98% of synthetic datasets were not statistically significantly different than the original dataset for both the zero-inflated and beta portions of the model (Table 3). Additionally, effect size categorizations were maintained for 100% of both zero-inflated and beta portions of the model.

### Study 2: Vowel Acoustics as Predictors of Speech Intelligibility in Dysarthria

Thompson et al. (2023) examined the relationship between vowel space area and speech intelligibility among 40 speakers with dysarthria of varying etiologies, including Parkinson’s disease, amyotrophic lateral sclerosis, Huntington’s disease, and cerebellar 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 2).

Below we import the original dataset, wrangle the data, and generate a synthetic data set. Data wrangling stpes include importing the original dataset, removing a string from the SpeakerID variable, and selecting only relevant variables.

# import original data  
articulation\_original\_data <- rio::import(  
 file = here::here("Data", "02\_Articulation", "data\_Acoustic Measures.csv")  
) |>  
 # Remove the reliability trials that have "\_rel" in the SpeakerID variable  
 dplyr::filter(  
 !grepl(  
 pattern = "\_rel",  
 x = SpeakerID  
 )) |>  
 # Selecting just the variables we need  
 dplyr::select(  
 SpeakerID, # ID  
 VSA\_b, # VSA in Bark  
 Int = Int\_OT # intelligibility (orthographic transcriptions)  
 )

Next, we generate a synthetic dataset with the syn() functionm extract the dataset, and convert it to a dataframe. *Synthpop* provides a warning message since this dataset has fewer observations than recommended (> 130).

# generate synthetic dataset  
articulation\_synthetic\_dataset <- syn(articulation\_original\_data,  
 m = 1,  
 seed = 2024)

CAUTION: Your data set has fewer observations (40) than we advise.  
We suggest that there should be at least 130 observations  
(100 + 10 \* no. of variables used in modelling the data).  
Please check your synthetic data carefully with functions  
compare(), utility.tab(), and utility.gen().  
  
  
Variable(s): SpeakerID have been changed for synthesis from character to factor.  
  
Synthesis  
-----------  
 SpeakerID VSA\_b Int

# Extract the synthetic dataset and convert into a data frame  
articulation\_synthetic\_dataset <- as.data.frame(articulation\_synthetic\_dataset$syn)

Next, we compare the distributions for vowel space area and speech intelligibility between the synthetic and original dataset. Figure 2 suggests that while the synthetic data largely approximates the original dataset, there are several values that are oversampled in the synthetic dataset. This might affect the quality of inferences with this dataset.

# Comparison of original and synthetic datasets with synthpop package  
articulation\_comparison <- compare(  
 articulation\_synthetic\_dataset, # synthetic dataset  
 articulation\_original\_data, # original dataset  
 vars = c("VSA\_b",  
 "Int"), # variables for comparison  
 stat = "counts", # Present the raw counts for each variable  
 cols = c("#62B6CB", "#1B4965") # Setting the colours in the plot  
)

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

### Results for Studies 3 - 9

##### Table 3 here.

# Discussion

XXX

# Conclusions

XXX

# Preregistration and Data Availability

Preregistration, data, and analysis scripts are publicly available on the Open Science Framework (https://osf.io/vhgq2 - **note: update this link when preregistration goes public**).

# Acknowledgements

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

# 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>

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>

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>

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>

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>

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>

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.

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>

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>

# Table and Figure Captions

Table 1. Table 1: Characteristics of included studies by ASHA domain.

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.