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

James C. Borders1, Austin Thompson2, & Elaine Kearney3

1. Department of Biobehavioral Sciences, Teachers College Columbia University
2. Department of Communication Sciences and Disorders, University of Houston
3. Faculty of Health, School of Psychology and Counselling, Queensland University of Technology

**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).

# 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**: XXX

**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. 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 (Rubin, 1993; Drechsler et al., 2024). 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, articulation, fluency, voice and resonance, hearing, communication modalities, receptive and expressive language, cognitive aspects of communication, and social aspects of communication. Once a dataset in each domain was found, the authors reproduced the primary analysis from the 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

Synthetic data was generated with the *synthpop* R package (cite). 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 et al. (2023) 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 (cite). 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.

# 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.

To generate synthetic data, we first load in the original dataset and then create a synthetic dataset with the syn() function in the *synthpop* package.

# 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  
 )  
  
# 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). 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.

### 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 and generate a synthetic data set. *Synthpop* provides a warning message since our dataset has few observations than recommended.

# 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)  
 )  
  
# 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

# 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.