**Comments from the Editor:**

The relevance of this paper to speech and language researchers is high, and JSLHR is clearly the best home for this work. However, the reviewers both point out an important dilemma when publishing a method-focused paper to the JSLHR audience. Some readers will be highly familiar with the topic and many others will be novices. I happen to be somewhere in between since I have been exposed to this sort of work and understand the premises, but have not developed the skills to follow the technical approaches that are recommended. The reviewers both suggest revisions that either narrow, expand, or split the scope of this work in future versions of this or other papers. Based on my opinion of the high importance of this work and the reviewers' comments, I would like to give the authors the opportunity to revise this work. I am partial to Reviewer 1's suggestion that a revision could focus on the methodological aspects, and thus be resubmitted as a Research Note, and a second (new) paper could serves as a tutorial.

**Response**: Thank you for these thoughtful reviews. Per the reviewers’ suggestions, we have decided to resubmit the manuscript as a Research Note that is focused on the application of synthetic data in the field of Communication Sciences and Disorders. We plan to submit a second tutorial paper focused on describing how to implement and evaluate synthetic data for various use-cases.

**Comments from Reviewers:**

The reviewers were asked to consider and address the items below in your evaluation of this manuscript:    
1. Overall Strengths    
2. Importance    
3. Justification/Rationale    
4. Methods/Approach    
5. Results/Findings    
6. Discussion/Conclusions    
    
Please use the comments below to guide your revisions.

**Reviewer #1:**

Reviewer 1: Overall Strengths / Importance / Justification/Rationale  
  
Thank you for the opportunity to review this manuscript. Improving reproducibility and transparency is important for research in communication sciences and disorders and related fields. Synthetic data is one potential tool for improving reproducibility and transparency, though it is not used often, as noted by the authors. The present manuscript evaluates the use of a well-established R package to produce synthetic data for 9 studies across ASHA practice domains and provides tutorial code and a web application to facilitate uptake within the field. The topic is in scope and tailored to the readership of JSLHR.

General Introduction:  
  
(1) I think it would be useful for the authors to clearly lay out the different use-cases for generating and sharing synthetic data when the source data cannot be shared. It seems to me that there is no single purpose for synthetic data – the use-cases are somewhat variable, and this directly impacts the level of scrutiny that needs to be applied to a synthetic data set.

For example, if sharing synthetic data to improve transparency around workflows or statistical models, it may not be completely necessary for synthetic data to reproduce the source data in every way. The same goes for educational contexts – when creating synthetic data for educational purposes.

However, when the synthetic data might be used to generate additional research findings, for example in the case of meta-analysis or methods development, the standards for synthetic data is higher.

**Response**: We have added a section in the introduction outlining the different use-cases for generating and evaluating synthetic data, as follows: “Synthetic data are, however, not without limitations; the extent to which statistical properties of the original data are retained varies based on the dataset and the model used to synthesize the data (Latner et al., 2024; Matthews & Harel, 2011). The intended use of synthetic data influences the level of rigor and scrutiny required. For example, synthetic data can serve as a pedagogical tool to teach data analysis skills or novel statistical methods (Shepherd et al., 2017). In such cases, preserving general statistical properties is sufficient, even if precise relationships between variables are not fully maintained. Similarly, synthetic data accompanying publications can facilitate reproducible workflows to illustrate data pre-processing steps or statistical models without reproducing exact study results. However, higher standards are required when synthetic data is used for hypothesis testing, meta-analyses, or methodological development (Raab et al., 2017). In these scenarios, synthetic datasets must accurately preserve multivariable relationships to ensure their validity and utility.” (lines 112 - 123)  
  
(2) More information is needed about how the synthpop R package generates synthetic data. In my opinion, it is not enough to demonstrate the implementation of an R package towards a purpose, but it is critical to explain the underlying statistical methods that are being implemented tailored to the background knowledge of the audience.

I glanced through the synthpop authors paper in JOSS, and acknowledge that this level of mathematical detail is likely less accessible to the JSLHR general readership. However, conceptual explanations of how synthetic data is generated, and perhaps for each method included in synthpop is necessary here so that readers understand (at least conceptually) the underlying methods. Whether this belongs in the introduction or methods – I leave that up to the authors. But I believe that clear and thorough explanations here stand to be a clear value-add over the existing papers from the synthpop authors, at least as it relates to the JSLHR readership.

**Response**: We have provided additional information on the underlying *synthpop* methods that are used to generate synthetic data, as follows:

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

For example, consider a dataset containing 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 (Audigier et al., 2018) but with a key distinction: instead of imputing only missing values, synthpop generates entirely synthetic data (Raghunathan, 2021), significantly reducing disclosure risk.

Nowok et al. (2016) provide an in-depth overview of the synthpop package’s features. Briefly, synthesis is largely automated using the syn() function. Users can customize various options, including the modeling approach, choice of predictors, order of synthesized variables, smoothing parameters for continuous variables to enhance privacy, and rules for maintaining logical relationships.” (lines 171 - 196)

Specific:  
105 – concerns about privacy may persist or perhaps even outweigh reproducibility in some cases?

**Response**: We agree with this sentiment that maintaining participant privacy should be a priority when sharing data, particularly in the case of confidential data. We have revised the manuscript, as follows: “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.” (lines 94 - 97)

110 – does synthetic guarantee that reidentification is not possible?  
111 – successful synthetic data does this, but it is not a guarantee of all synthetic data

**Response**: Synthetic data does not guarantee that reidentification is impossible, but it does significantly reduce the risk of disclosure. This must be confirmed by the user before public release of the synthetic dataset. We emphasize this point throughout the manuscript (e.g., lines 52 - 54; lines 112 - 147; lines 373 - 376).  
  
Methods/Approach  
  
In my opinion, the present manuscript would benefit from shifting either to a more comprehensive methods study examining the effectiveness of the methods within the synthpop R package to reproduce existing data or towards an explicit tutorial manuscript focusing on research designs and data structures that the synthpop package has been validated for (or perhaps if the authors are so inclined, both as separate manuscripts). As it stands the current study straddles between these two types of manuscripts, which sacrifices methodological detail and the extent to which the findings are reflective of the potential for synthetic data across the field as a methods study or the comprehensiveness needed as a tutorial paper to guide less informed researchers/readers.  
 **Response**: The current manuscript focuses on the application of synthetic data in the field of Communication Sciences and Disorders. We agree that a tutorial is outside the scope of this manuscript and plan to submit a separate paper focused on describing how to implement and evaluate synthetic data for various use-cases. Of note, the current study aims to examine the feasibility and preliminary utility of synthetic data in study designs and outcome types that are seen in the literature. It is not feasible to evaluate and provide recommendations for the utility of synthetic data for every possible scenario that a researcher may encounter. Instead, critical thinking on the part of the researcher, as well as appropriate approaches to evaluate utility, are necessary.

For the former (methods approach) to be successful, the authors need to take further steps to ensure that the studies included are representative of research designs and sample sizes in the field. To me, it is less important from a methodological standpoint that the topics of research are representative of the ASHA domains – but it is more important that the tools within synthpop are effective for the range of research designs and data structures typical of the field (and which ones it is less likely to be able to accommodate). As it stands, authors have not established that the studies included are representative of the research conducted by the field at large. Similarly, it may be useful in a methods study to vary certain parameters of the source data (e.g., the sample size, or synthetic data method) to understand how that might affect the effectiveness of the synthpop methods as they pertain to the research types and sample sizes that are typical in CSD. This is particularly true in the case of this manuscript, where 3/9-4/9 datasets were not reproduced successfully.

As one specific example, it is not clear to me that the methods the authors use to evaluate synthetic data sufficiently test whether the synthetic data generated recreates multilevel/hierarchical data which is prevalent throughout CSD and related research. Examining p-values/effect sizes is an important step, but as part of this it would also be important to understand whether (for example) within- and between- subject variances are similar. Additionally, the web application provided by the authors (and the similar application from the synthpop author(s) only comparisons of single variables, not of any joint distributions, and so I worry that some researchers may take the fact that the distributions of any single variables being similar as an indication that the joint distributions of a dataset as a whole are also similar. A similar case could be made for single case experimental design studies that are highly prevalent in the field.  
 **Response**: It is not feasible to evaluate the utility of synthetic data for every possible research design, outcome type, or data parameter. This issue is further compounded by the lack of open data in our field. In fact, obtaining datasets that were representative across subfields was itself a difficult task. Additionally, it is not our goal to suggest that our investigation could provide sufficient evidence to suggest that synthetic data could be used in each of these scenarios. Instead, we aim to provide a proof-of-concept to the reader that synthetic data is a feasible approach in different scenarios that might be representative of different subfields in CSD (and thus, data types that the reader may encounter). We expand on this point in the manuscript, as follows: “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 all possible designs. Instead, this approach prioritizes representation across all CSD subfields to provide a proof of concept for the reader. To demonstrate the feasibility and potential utility of synthetic data to promote transparency and reproducibility, an analysis was chosen from each study to generate synthetic data for those variables.” (lines 163 - 168)

To ensure that we have included a range of study designs, outcome types, and analyses, we have also expanded our description of included studies to better characterize the different scenarios where synthetic data *may* be more or less effective for general and specific utility (see Table 2). In the limitations section, we also acknowledge that our studies are not representative of every research design or data parameter, and that it is imperative that the user evaluate its utility in the context of their own goals before publicly sharing synthetic data. This is described, as follows “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.” (lines 370 - 376)

Regarding the importance of hierarchical data, we have included datasets that have a hierarchical structure (Battal et al., 2019; Clough et al., 2023). We also provide results related to the stability of synthetic data to maintain between/within variances for these studies in the context of random effects from the hierarchical model, as follows:

“Battal et al., 2019… 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.” (lines 274 - 276)

“Clough et al., 2023… (mean = -0.011, 90% CI: -0.018, 0.005) and synthetic (mean = -0.016, 90% CI: -0.018, -0.005) datasets.” (lines 318 - 320)

It’s important to note that other approaches to more effectively synthesize hierarchical data have been proposed, though this is outside the scope of our manuscript. We have included this in the manuscript, as follows: “Alternative approaches specifically designed to handle hierarchical data (Gauvin et al., 2021) may offer a better solution and should be explored in future work.” (lines 359 - 360).

We also acknowledge that the studies we originally selected did not include a single subject experimental design, which is a common and important research design in CSD. Therefore, we have added a study that used this type of design (Robinaugh et al., 2024) and reported the results (lines 298 - 306).

Finally, we have more explicitly specified our evaluation of the utility of synthetic datasets by delineating general and specific utility. Within general utility, we have expanded our evaluation to include both univariate distributions (e.g.., bar charts, histograms) and bivariate joint distributions (e.g., scatterplots) [lines 126 - 129]. We have provided these evaluations for each included study in this manuscript and plan to provide a more thorough tutorial in our separate submission.

I note that the authors of the synthpop package warn against the packages ability to produce synthetic data for hierarchical data. (“Data sets with a complex data structure, e.g. hierarchical data, multiple events data, can not be easily synthesised in 'synthpop' at the moment. You can still attempt to synthesise such data but some pre-processing will be required.” <https://www.synthpop.org.uk/faq.html>). Unless this has changed, it is important that these nuances are clear in the present manuscript.

**Response**: We acknowledge the limitation of synthpop in synthesizing hierarchical data and now more explicitly highlight this point in the manuscript, as follows: “...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.” (lines 355 - 358).  
  
On the other hand, the present manuscript lays a strong foundation for an extremely useful and accessible tutorial for JSLHR that is focused on helping researchers create and evaluate synthetic data for their own studies when data sharing is not possible. The code examples and range across the ASHA domains are well-suited to this purpose. However, more attention is needed to help guide researchers on interpreting the results of the code and making decisions around whether or not they should share the synethetic data, how they can tweak the parameters of the data generation, and how they should write about the synthetic data generation when sharing in their own research studies.

Perhaps a successful tutorial might even include a sort of ‘faq’ section to help less informed researchers evaluate if synthetic data is a good option for their current study. Like ‘is my research design / data structure supported’. ‘How do I know if the synthetic data generated is close enough?’ ‘which variables should I generate synthetic data for’? Regardless, a successful revision in this respect should seek to ensure that readers are well prepared to generate and share rigorous synthetic data that serves their specific purpose, from the decision to use synthetic data, to evaluating its appropriateness for the use case, to being able to clearly write about the data’s generation methods.

**Response**: Thank you for this suggestion. We plan on adding this in a future tutorial in greater detail.  
  
Specific:  
  
130 – similar to as noted above: how were these articles selected. How did the authors ensure they are representative of their subfields or the field as a whole?

**Response**: Given the lack of open data in our field, we used a convenience sampling approach. We have added this to the manuscript, as follows: “A convenience sampling approach was used to identify publicly available datasets from previously published research articles related to the ‘Big Nine’ ASHA domains… 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.” (lines 156 - 158)

5. Results/Findings  
  
One of the main concerns about sharing data are when data are combined with demographic variables in a study. How do the authors think about combining synthetic data from a study with participant demographics. Should synthetic data have participant identifiers that connect back to “real” participants? Or should they be randomly assigned? What about for individual difference studies where demographics or testing scores are included in statistical analyses?

**Response**: This is an important consideration. Ideally, if researchers choose to share demographic data without participant consent they should synthesize the demographic data to ensure it is not identifiable and does not represent ‘real’ individuals.  
  
256/257: So based on this information, do the authors conclude that the synthetic data set sufficiently recaptures the source data? How do they choose between the 100 data sets?

**Response**: Although our initial goal was to assess the stability of synthetic data across multiple (100) datasets, we recognize that this approach may be excessive and impractical for implementation. Therefore, we have revised the manuscript to focus on a single synthesized dataset compared to the original. For transparency, we have included the results across all 100 datasets in an appendix for the interested reader (see Supplemental Table 1 and Supplemental Figures 1 and 2).  
  
322: similarly, in this case are these results sufficient to think that the synthetic data is usable? How do the authors choose which set? Does the % of synthetic data sets that demonstrate the same result impact whether we choose ultimately to share synthetic data?

**Response**: We have revised the manuscript to focus on a single synthesized dataset compared to the original. The utility of synthetic data depends on the researcher’s specific goals and intended use case, as discussed in the manuscript (lines 361 – 368). To support this, we provide information on both general and specific utility (lines 124 - 131 and lines 198 - 209). Ultimately, the decision to use and share synthetic data should be evaluated on a case-by-case basis, considering whether the synthetic dataset meets the needs of the intended analyses.

337: For all studies included, what other checks are needed to examine the synthetic data. Synthpop methods appear to incorporate joint distributions from the source data in the creation of synthetic data – do we need to evaluate the extent to which these distributions hold?

**Response**: One approach to evaluate whether these distributions are similar between synthetic and original datasets is with the standardized standardized ratio of the propensity mean squared error (*S\_pMSE*). We describe its use in the manuscript, as follows: “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.” (lines 198 - 209)  
  
6. Discussion/Conclusions  
  
345: The utility of synthetic data is further strengthened by the range of datasets included in the current study, which varied by domain (across nine ASHA domains), sample size (from 40 to >8,000 data points) statistical models (from simple correlations to multilevel model with 3-way interactions), and effect sizes (from conventionally “small” to “large”).”

Moreover, while there is a range of studies included in this study, it remains unclear the extent to which they represent research in the field. Moreover, single only 6/9 studies maintained the direction of p-values and 5/9 effect sizes, its not clear that the present study justifies the following conclusion: “These results suggest that synthetic datasets can be effectively used across a wide range of studies in the field of CSD to preserve participant confidentiality when sharing data.” I think an appropriate hedge is warranted here.

**Response**: We revised the wording of the overarching finding in the discussion/conclusion sections to reflect the updated results and to emphasize the need for researchers to critically evaluate the utility of data synthesis for their purposes:

“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.” (lines 396 - 401).  
  
383: I’m not entirely clear on what the framework is that is laid out by the authors. Important points are discussed in this area of the discussion, though arguably many of them would be useful in the introduction. The decision tree figure is useful, but does not constitute a framework. Overall, how this information can act as an overarching, board guide for how researchers should think about data sharing and synthetic data is not clear. Perhaps framework is not the best word choice here. Otherwise, I think more attention is warranted to strengthen this section to reflect what is expected of a framework - whether supported by conceptual diagrams or other methods. I’m not sure whether this should be the focus of a revision or the focus of separate work.

**Response**: We plan to incorporate these suggestions into a new submission focused on a tutorial to implement and assess the utility of synthetic data.  
  
Could the manuscript benefit from the addition of supplemental material?  
Reviewer 1: No:  
  
Is additional information regarding the research methodology needed to replicate the study?  
Reviewer 1: No:

**Reviewer #2:**

Reviewer 2: This concise paper presents a compelling case for the creation of synthetic data as a means to address privacy and ethical concerns surrounding data sharing. I am quite supportive of the concept of synthetic data, as it has the potential to enhance reproducible research pipelines, and I believe that both critical discussion and demonstration in this area are timely contributions to the field. I fully encourage the authors' efforts in this regard.  
  
That said, I would appreciate a more thorough discussion of the limitations associated with synthetic data, which I believe could be more prominently addressed in the paper.

**Response**: We have added a more nuanced discussion of the limitations of synthetic data, emphasizing the need to evaluate the quality of synthesized data against the purpose of using the synthetic data:

“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.” (lines 112 - 123).

We have also noted the current limitations of synthesizing datasets with a hierarchical structure with the *synthpop* package.

When considering hierarchical data, such as repeated measurement designs commonly used in experimental phonetics, it seems that synthetic data can diverge significantly from the original data. Given this context, I find it challenging to see how synthetic datasets facilitate meaningful exploration of data or the development of novel questions and hypotheses. While they may offer some utility, particularly in terms of generating a rough statistical quality, the reliability of synthesized data to detect significant differences appears to be an area that warrants caution.  
I appreciate the authors’ suggestion to generate multiple synthetic datasets and select the best-fitting one. It would be beneficial if they could expand on the applicability of synthetic datasets in more complex data structures, especially hierarchical datasets with multiple observations per condition per subject.

**Response**: We have revised the discussion to emphasize that synthetic data may not be appropriate for hierarchical datasets and that the researcher must evaluate its utility on a case-by-case basis. The manuscript has been revised as follows,

“...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.” (lines 356 - 360)

“...for hierarchical datasets, synthetic data 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.” (lines 397 - 401)  
  
The authors mention (lines 111f.) that “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.” While I find that for simpler non-hierarchical datasets some statistical properties may be preserved, this retention does not seem reliable, consequently limiting reproducibility. It appears that what can be replicated most effectively is the analytical pipeline itself, which I view as a significant advantage of synthetic data, as it facilitates code execution and produces qualitatively similar results. However, I have found that reproducing specific numbers or statistical indices often lacks precision.

**Response**: As we described in a response above, we have revised the manuscript to emphasize the different use cases of synthetic data (e.g., lines 112 - 123), as well as its performance with different data types (e.g., worse performance with hierarchical data; lines 356 - 360, 397 - 401).  
  
Additionally, I believe that the demonstration would greatly benefit from providing clearer guidance for readers who may be less familiar with the subject matter. For instance, in lines 213ff., the mention of “cart” (Classification and Regression Tree) as the default method in synthpop could potentially confuse readers. A more detailed explanation of the CART approach would likely enhance understanding.

**Response**: As described in our response above, we have included more information about the underlying methods of synthetic data generation (lines 171 - 196).  
  
Overall, I commend the authors for advocating for transparency and a critical examination of practices within the field.  
  
Could the manuscript benefit from the addition of supplemental material?  
Reviewer 2: No:  
  
Is additional information regarding the research methodology needed to replicate the study?  
Reviewer 2: No:

**IMPORTANT COMMENTS FROM THE EDITORIAL OFFICE.**

In addition to completing the content revisions requested above please ensure you also attend to the following issues:

\* Please place the Data Availability statement and the Funding statement before the References listing with the Acknowledgments statement rather than at the beginning of the text file.   
 **Response**: The data availability and funding statements have been moved before the References.

\* Please remove all shading from your tables. Only bold, italic, or underlined text is allowed.

**Response**: Shading has been removed from the tables.