PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Can Synthetic Data Be A Proxy for Real Clinical Trial Data: A Validation Study
AUTHORS	Azizi, Zahra; Zheng, Mina; Mosquera, Lucy; Pilote, Louise; El Emam, Khaled

VERSION 1 – REVIEW

REVIEWER	Kristin Sheffield Eli Lilly and Company
REVIEW RETURNED	29-Sep-2020

GENERAL COMMENTS	This was a well-written manuscript that nicely introduces synthetic data methods to a readership that may be unfamiliar with this concept, particularly as applied to clinical trial data. Comments: “There have thus far been no studies that examine the ability of synthetic to replicate secondary analyses of clinical trial data.” The authors may wish to confirm this assertion in the literature. A brief search revealed at least one study that generated synthetic data based on the SPRINT trial (https://www.ahajournals.org/doi/full/10.1161/CIRCOUTCOMES.118.005122) Page 7: Please add the censoring rules for DFS and OS for Trial N0147. Related: OS and DFS are listed as quasi-identifiers. Did the authors mean to say that date of death, date of disease recurrence, and event/censor flags are quasi-identifiers? Please share more details about how time to event outcomes are synthesized On page 8 it is mentioned that sequential decision trees have the advantage of not requiring large training datasets. Was a training set used for this study? If not, please discuss the rationale and the potential limitations/implications for the results. Limitations section: “additional evaluations are necessary to increase the weight of evidence in support of using synthetic data as a proxy for the real dataset”. Please provide more detail about the necessary additional evaluations. Figure 1 is a helpful summary of the sequential data synthesis process. However, it would beneficial to also describe the specifics of the sequential data synthesis process in this study, i.e. what was the sequence of variables? Can these models be used for time-to-event outcomes? Was any adaptation necessary for Cox regression modeling? Figure 4: please comment on potential reasons for 0% overlap for race.

	Figures 7 and 8: The point estimate of the hazard ratio is different between the real and synthetic data for T stage and histology. Please address this point in the results section or discussion. It is my understanding that synthetic models need to be verified against models from the original data. That is, analysts may ask questions and conduct novel secondary analyses using the synthetic data, but they would need to acquire the real data and confirm an observed relationship or ask the custodian of the original data to conduct the analysis and confirm the relationship. Please comment on the implications in the discussion section and if/how this may limit the utility of synthetic data.

REVIEWER	Sarah Nevitt University of Liverpool United Kingdom
REVIEW RETURNED	10-Nov-2020

GENERAL COMMENTS	I have conducted a statistical review of the manuscript “Replicating Secondary Analyses of Clinical Trial Data Using Data Synthesis" This is a very interesting study, which shows promising results for the field of data sharing and secondary analysis of clinical data. I enjoyed reading this work. The methods used for this work are suitable and well described. Results are interpreted appropriately and written very well. My comments and thoughts are as follows: Title: I suggest that ‘synthetic data’ rather than ‘data synthesis’ may be more appropriate for the title. ‘Data synthesis’ is also a term to describe methods for combining data, such as meta-analysis. Given that meta-analysis and other synthesis techniques are common reasons for the secondary analysis of clinical trial data, it may be confusing to use the term ‘data synthesis’ in the title. Abstract (and methods): “Participants: There were 2,686 patients recruited in the original trial.” But if I understand correctly, only the control arm data is used so this is 1337 patient’s data included in this analysis? Abstract (results): “…the hazard ratio confidence interval overlap between the real and synthetic data varying from 61% to 86%.” It isn’t clear what this refers to? Is this the overlap for the outcomes DFS and OS respectively? Please be more specific when referring to results Strengths and limitations: This is a summary of the aims rather than strengths and limitations of the work. Introduction (second paragraph): “An analysis of the success rates of getting individual-level data for research projects from authors found that the percentage of the time these efforts were successful varied significantly and was generally low at 58% [4], 46% [5], 14% [6], and 0% [7]. “

	I hate to be the reviewer that advertises their own work, but our systematic review from 2017 examining IPD retrieval rates is probably relevant here: Nevitt SJ, Marson AG, Davie B, Reynolds S, Williams L, Smith CT. Exploring changes over time and characteristics associated with data retrieval across individual participant data meta-analyses: systematic review. bmj. 2017 Apr 5;357. Introduction (general comment): I think a little more information is needed about 'synthetic data' and the process in this context for the general readership of the journal. Such as, describe that it must be the original data controller / sponsor / owner of the data who performs the synthesis as access to the original data is required for the synthesis. So, it wouldn't be possible using these methods for a secondary researcher to synthesise themselves a dataset based on published information of a clinical trial to overcome the time and issues associated with data retrieval as described in the introduction. So, it will still take time and resources to provide synthetic data and it won't necessarily be quicker or 'easier' for a researcher to get access to synthetic data. Rather the benefits are that data which previously may not have been able to be shared due to concerns regarding data protection and anonymisation / re-identification attacks may now be able to be shared as 'synthetic data' (partially or completely synthesised?) as an accurate proxy for the 'real' data? Introduction (paragraph 8): "This will inform us whether the sharing of synthetic clinical trial data will still allow researchers performing new analyses on that data to draw similar conclusions as they would have had the original data been shared." I would argue that while it is important for numerical results (e.g. the hazard ratio and 95% CI) and the magnitude of any effect or relationship to be similar between the synthetic data and the original data, it is essential that the SAME conclusions can be drawn, rather than similar conclusions (i.e. whether a relationship between a factor and an outcome exists, whether one intervention is significantly better than another etc.). Methods (Data sources): I'm not sure I understand the relevance to this particular work of the first paragraph which discusses the length of the data requesting processes on CSDR and Project Data Sphere. I suggest just stating the source of the dataset used is Project Data Sphere I also note that as data from Project Data Sphere is anonymised data, arguably this is not the 'original' data from the trial. This is a minor wording point but given that sharing synthetic data is proposed here is an alternative to sharing an anonymised version of the 'original' data where there are data privacy concerns, it should be made clear that the 'real' data (compared to synthetic data) is an anonymised version of the data used to produce the published results. Methods (Data synthesis, third paragraph): "The quasi-identifiers selected for the N0147 trial were age, gender, race, BMI, OS, DFS (since death status would be known by an adversary)."

	I'm a little confused by this sentence – specifically that OS is identified as a quasi-identifier and that 'death status would be known by an adversary.' OS data would include death status (i.e. whether a patient has died) as well as event time or censoring time. What has been synthesised with respect to OS? Methods (Section 2.4): The first sentence of this section appears twice. Results (Figures 3, 4, 6 and 7): These results figures are interesting. I agree that mostly the confidence interval overlap is high. It looks to me that often the confidence intervals for the synthetic data look to be narrower than the confidence intervals of the real data, particularly for parameters with the (relatively) widest confidence intervals such as race and T-stage. Is it to be expected that synthetic data is a better match for the 'real data' where the 'real data' less variable and more precise? Results: "1.56; 95% CI: 1.11-2.2 for real data, and 2.03; 95% CI: 1.44-2.87 for synthetic data." Please add in what these results are here and for the DFS results (presumably hazard ratios?) Please also state that the overlap for the DFS results. I note that, although there is a high overlap, the HRs and 95% CIs are larger for both OS and DFS from synthetic data compared to the real data. Is this just a coincidence for this example or could the data synthesis process systematically lead to slightly larger / smaller results across a dataset for synthetic data compared to real data? Discussion: The manuscript ends quite abruptly with the strengths and limitations section. Consider adding in a brief conclusion statement (such as the conclusions of the abstract).

REVIEWER	Jean-Marc Ferran Qualiance (Denmark) I lead the PHUSE Data Transparency Working Group and I am an appointed member of the EMA Technical Anonymisation Group and Health Canada Reference Group for the Public Release of Clinical Information. I also advise d-Wise Inc. on the development of their anonymization solutions.
REVIEW RETURNED	16-Nov-2020

GENERAL COMMENTS	Page 4, lines 13-17 There has not been successful re-identification attacks on the like of clinical trials report or IPD shared in the context of EMA Policy 0070/Health Canada PRCI nor the voluntary sharing initiatives from sponsors (e.g. CSDR, Vivli, YODA, etc.) and regulators are promoting anonymization methods in this field. The paragraph and its references should be put further in context as it is speculative at this stage. Page 4, lines 41-45 There are different types of secondary analyses and the appraisal ones related to checking the integrity of the study and any bias in its conduct or reporting may not be suitable for use of Synthetic Data. The paper should elaborate further and be more explicit on which types of analyses are legitimate to consider Synthetic Data for.

	Page 5, lines 39-46 The source dataset is coming from Project Data Sphere and is meant to be anonymized. Could you please elaborate on any risk (privacy in particular) to generate synthetic data from anonymized data rather than pseudonymized data? Are there any pitfalls to consider? Page 7, lines 49-54 While Hiding-in-Plain-Sight (HIPS) would require synthesizing certain quasi-identifiers within bands, the use of Synthetic Data for the like of Adverse Events (certain being considered quasi-identifiers) could be misleading in CSRs shared under EMA Policy 0070 or Health Canada PRCI. To which extent synthetic data would be acceptable under these policies? Or would it be in connection with anonymization of certain quasi-identifiers? Would partial synthesize of demographics be sufficient in this context?

VERSION 1 – AUTHOR RESPONSE

Reviewer #1

1 "There have thus far been no studies that examine the ability of synthetic to replicate secondary analyses of clinical trial data."

The authors may wish to confirm this assertion in the literature. A brief search revealed at least one study that generated synthetic data based on the SPRINT trial (https://www.ahajournals.org/doi/full/10.1161/CIRCOUTCOMES.118.005122)

Thank you for pointing this out. We have added the article on the synthesis of the SPRINT trial to the paper and adjusted the description of prior work in the introduction as follows:

"There have been limited replications of clinical studies using synthetic data, with only a handful of examples in the context of observational research [1], [2] and larger clinical trial data [3]. The current study adds to this body of work and contributes to the evidence base for enabling more access to clinical trial data through synthesis."

2 Page 7: Please add the censoring rules for DFS and OS for Trial N0147.
Related: OS and DFS are listed as quasi-identifiers. Did the authors mean to say that date of death, date of disease recurrence, and event/censor flags are quasi-identifiers? Please share more details about how time to event outcomes are synthesized

We have added the censoring rules from the original trial in the revised manuscript. Since this was an analysis dataset (ADaM) rather than a source dataset, the time-to-event was already calculated. We have clarified the variables that were synthesized in the revised manuscript.

The details on the quasi-identifiers (which are the variables that have been synthesized), their definitions, and their selection has been moved to the appendix.

3 On page 8 it is mentioned that sequential decision trees have the advantage of not requiring large training datasets. Was a training set used for this study? If not, please discuss the rationale and the potential limitations/implications for the results.

The process of creating a generative model is often referred to as training the model or fitting the model. Synthetic data is then produced from the generative model. We have clarified this terminology in Section 2.3. The input data that is used to create the generative model does not need to be large when sequential trees are used compared to other methods such as deep learning.

4 Limitations section: "additional evaluations are necessary to increase the weight of evidence in support of using synthetic data as a proxy for the real dataset". Please provide more detail about the necessary additional evaluations.

This statement was intended to clarify that it is reasonable to expect that as replications are conducted, this will increase the acceptance of synthetic data as a proxy for real data. We are starting to see that happening with studies being conducted only on synthetic derivatives. We have restated the point as follows:

"It is a reasonable expectation that as more similar replications using synthetic data demonstrate equivalent results and conclusions as real data, there will be greater acceptance of synthetic derivatives as a reliable way to share clinical trial datasets. In fact, we are already starting to see published (observational) health research using synthetic derivatives only [40]."

5 Figure 1 is a helpful summary of the sequential data synthesis process. However, it would beneficial to also describe the specifics of the sequential data synthesis process in this study, i.e. what was the sequence of variables? Can these models be used for time-to-event outcomes? Was any adaptation necessary for Cox regression modeling?

We have added another figure to illustrate the process that was used and included in the body of the paper the sequence that was used for synthesis. Given that this was an analysis-ready dataset, there was no adaptation needed to the synthesis process – it just takes in a tabular dataset and creates a synthetic version of that data. We have added that point as well. Note that to remain within recommended word lengths the more detailed description of the synthesis method has been moved to the appendix.

6 Figure 4: please comment on potential reasons for 0% overlap for race.

Given the weak relationship, and that synthesis is stochastic, it is not surprising that there will be examples of non- overlapping confidence intervals. The differences are also very small in absolute terms. The following is the new text:

"Given the weak relationship between race and DFS, this lack of confidence interval overlap is likely due to the stochastic nature of synthesis. In addition, the relationship is quite weak in both datasets and of very similar magnitude, therefore the conclusions would still be the same in both cases."

7 Figures 7 and 8: The point estimate of the hazard ratio is different between the real and synthetic data for T stage and histology. Please address this point in the results section or discussion.

We have added a point about these specific results in the Results section of the paper. The new wording is as follows:

"The point estimates for the T stage covariates differ the most in Figure 6 for overall survival and Figure 7 for the disease-free survival model, with lower confidence interval overlap than the other covariates. The same is true for histology in Figure 6. While some variation in the numeric values is expected in the synthetic data, the parameters were directionally the same, and the inclusion of these covariates did allow us to control for their effect in the assessment of obstruction, which was the main objective of the analysis."

8 It is my understanding that synthetic models need to be verified against models from the original data. That is, analysts may ask questions and conduct novel secondary analyses using the synthetic data, but they would need to acquire the real data and confirm an observed relationship or ask the custodian of the original data to conduct the analysis and confirm the relationship. Please comment on the implications in the discussion section and if/how this may limit the utility of synthetic data.

The use of a verification server is one way to deploy synthetic data, and we have added a discussion of that setup in the "Relevance and Application of Results" section of the paper. The explanation of a verification server is as follows:

"While we are already starting to see published (observational) health research using synthetic data only [40], there will be situations where there is a requirement for additional verification that the model parameters produced from synthetic data are numerically similar to the real data, and that the conclusions are the same. This step can be achieved by implementing a verification server. With such a setup synthetic data is shared, and the analysts build their models on the synthetic data. Then their analysis code (say an R or SAS program) is sent to a verification server which is operated by the data controller / custodian. The analysis code is executed on the real data, and the results returned to the analysts. The returned results would either be the model parameters on the real data or the difference in parameter values between the real data model and the synthetic data model. That way the analysts can get feedback as to the accuracy of the synthetic data model parameters without having direct access to the real data themselves. The deployment of a verification server balances the need for rapid access to data with minimal

constraints with the need for ensuring model accuracy from the synthetic data. On the other hand, it does introduce an additional process step.

The need for a verification server can arise, for example, when results are going to be submitted to a regulator. Generally, in the early days of adoption of data synthesis there will likely be a greater need for verification, and one would expect that need would dissipate as successful applications of data synthesis increase over time."

Reviewer #2

1 Title: I suggest that 'synthetic data' rather than 'data synthesis' may be more appropriate for the title. 'Data synthesis' is also a term to describe methods for combining data, such as meta-analysis. Given that meta-analysis and other synthesis techniques are common reasons for the secondary analysis of clinical trial data, it may be confusing to use the term 'data synthesis' in the title.

The title has been changed to reflect this suggestion. The new title is: Can Synthetic Data Be A

Proxy for Real Clinical Trial Data: A Validation Study

2 Abstract (and methods): "Participants: There were 2,686 patients recruited in the original trial." But if I understand correctly, only the control arm data is used so this is 1337 patient's data included in this analysis?

We have clarified this point in the abstract and also in the body of the revised manuscript.

3 Abstract (results): "…the hazard ratio confidence interval overlap between the real and synthetic data varying from 61% to 86%." It isn't clear what this refers to? Is this the overlap for the outcomes DFS and OS respectively? Please be more specific when referring to results

The detailed results have been added to the abstract..

4 Strengths and limitations: This is a summary of the aims rather than strengths and limitations of the work.

The strengths and limitations have been rewritten. See the response to comment #2 from the Editors.

5 Introduction (second paragraph): “An analysis of the success rates of getting individual-level data for research projects from authors found that the percentage of the time these efforts were successful varied significantly and was generally low at 58% [4], 46% [5], 14% [6], and 0% [7]. “

I hate to be the reviewer that advertises their own work, but our systematic review from 2017 examining IPD retrieval rates is probably relevant here:

Nevitt SJ, Marson AG, Davie B, Reynolds S, Williams L, Smith CT. Exploring changes over time and characteristics associated with data retrieval across individual participant data meta-analyses: systematic review. bmj. 2017 Apr 5;357.

Thank you for bringing this to our attention. Yes, this is relevant and we have added that reference in the introduction.

6 Introduction (general comment): I think a little more information is needed about ‘synthetic data’ and the process in this context for the general readership of the journal.

Such as, describe that it must be the original data controller / sponsor / owner of the data who performs the synthesis as access to the original data is required for the synthesis. So, it wouldn’t be possible using these methods for a secondary researcher to synthesise themselves a dataset based on published information of a clinical trial to overcome the time and issues associated with data retrieval as described in the introduction.

So, it will still take time and resources to provide synthetic data and it won’t necessarily be quicker or ‘easier’ for a researcher to get access to synthetic data. Rather the benefits are that data which previously may not have been able to be shared due to concerns regarding data protection and anonymisation / re-identification attacks may now be able to be shared as ‘synthetic data’ (partially or completely synthesised?) as an accurate proxy for the ‘real’ data?

We have added the point in the introduction that the data controller would need to create the synthetic datasets from real individual-level datasets. That is the specific context for the current study. However, there are techniques for creating synthetic individual-level datasets from aggregate summaries as well, although these are not our focus and out-of-scope for this paper. Therefore, we also clarified by saying that the real data is also at the individual level. The new wording is as follows:

“To create synthetic data, a machine learning generative model is constructed from the real individual- level data, capturing its patterns and statistical properties. Then new data is generated from that model. This step is performed by the data controller / custodian who has access to that real data. The synthetic version of the data would then be provided to analysts to conduct their studies.”

There are also recent examples of academic medical centers allowing research studies on synthetic data without the need for ethics board reviews [4]. Therefore, the benefits are potentially beyond addressing the concerns around anonymization methods. We have added these points in the Conclusions section of the revised manuscript.

7 Introduction (paragraph 8): “This will inform us whether the sharing of synthetic clinical trial data will still allow researchers performing new analyses on that data to draw similar conclusions as they would have had the original data been shared.”

I would argue that while it is important for numerical results (e.g. the hazard ratio and 95% CI) and the magnitude of any effect or relationship to be similar between the synthetic data and the original data, it is essential that the SAME conclusions can be drawn, rather than similar conclusions (i.e. whether a relationship between a factor and an outcome exists, whether one intervention is significantly better than another etc.).

Yes - thank you for catching that. We have made the change accordingly. The revised wording in the introduction is as follows:

“An important question with the analysis of synthetic data is whether similar results and the same conclusions would be obtained as with the real data. To answer this question, we compared the analysis results and conclusions using real and synthetic data for a published oncology trial. ”

8 Methods (Data sources): I’m not sure I understand the relevance to this particular work of the first paragraph which discusses the length of the data requesting processes on CSDR and Project Data Sphere. I suggest just stating the source of the dataset used is Project Data Sphere

I also note that as data from Project Data Sphere is anonymised data, arguably this is not the ‘original’ data from the trial. This is a minor wording point but given that sharing synthetic data is proposed here is an alternative to sharing an anonymised version of the ‘original’ data where there are data privacy concerns, it should be made clear that the ‘real’ data (compared to synthetic data) is an anonymised version of the data used to produce the published results.

The section mentioning PDS has been edited to remove the requesting process.

We have clarified the terminology. We do not use the term ‘original’ data in the revised manuscript any more. We refer to the PDS dataset as the ‘real’ data. This is a simpler construct than ‘real but anonymized’, and is consistent with the definitions that we have added. The new wording is as follows:

"In the current paper, we will refer to this PDS dataset as the "real" data since that is our source dataset for synthesis. PDS data is already perturbed to anonymize it. The level of perturbation is dependent on the sponsor. Therefore, the use of term "real" should be interpreted to mean "real and anonymized" data."

9 Methods (Data synthesis, third paragraph): "The quasi-identifiers selected for the N0147 trial were age, gender, race, BMI, OS, DFS (since death status would be known by an adversary)."

I'm a little confused by this sentence – specifically that OS is identified as a quasi-identifier and that 'death status would be known by an adversary.' OS data would include death status (i.e. whether a patient has died) as well as event time or censoring time. What has been synthesised with respect to OS?

We have clarified the exact quasi- identifiers and how these are related to death. The definitions of the quasi-identifiers and details on how they have been selected has been moved to the appendix.

10 Methods (Section 2.4): The first sentence of this section appears twice.

The extra sentence has been deleted.

11 Results (Figures 3, 4, 6 and 7): These results figures are interesting. I agree that mostly the confidence interval overlap is high. It looks to me that often the confidence intervals for the synthetic data look to be narrower than the confidence intervals of the real data, particularly for parameters with the (relatively) widest confidence intervals such as race and T-stage. Is it to be expected that synthetic data is a better match for the 'real data' where the 'real data' less variable and more precise?

That is a very interesting hypothesis. A generative model captures the patterns in the data (except that there is no specific outcome variable defined). It is possible that the machine learning methods used during synthesis capture the signal in the data well and these are produced more clearly (or with less noise) in the synthetic data. We have added that point in the results section. The added wording is as follows:

"One other observation from the overall survival model in Figure 6 and the disease-free survival model in Figure 7 is that the confidence intervals from the synthetic data are narrower than the real data. A generative model captures the patterns in the data. A plausible explanation is that the machine learning methods used during synthesis capture the signal or patterns in the data well and these are produced more clearly (or with less noise) in the synthetic data. "

12 Results: “1.56; 95% CI: 1.11-2.2 for real data, and 2.03; 95% CI: 1.44-2.87 for synthetic data.” Please add in what these results are here and for the DFS results (presumably hazard ratios?) Please also state that the overlap for the DFS results.

Done. The revised wording is as follows:

“The main hypothesis being tested in the published secondary analysis pertains to obstruction. For the OS model the hazard ratio for obstruction overlap was high at 61% (HR of 1.56; 95% CI: 1.11-2.2 for real data, and HR of 2.03; 95% CI: 1.44-2.87 for synthetic data) with both models showing a strong effect of obstruction on OS (No obstruction related to higher OS). Similarly, for the DFS model, the overlap was 86% (real data HR of 1.51; 95% CI: 1.18-

1.95, and the synthetic data HR of 1.63; 95% CI: 1.26-2.1), indicating that the model shows an association between obstruction and DFS. Therefore, one would draw the same conclusion about the impact of obstruction on event-free survival.”

13 I note that, although there is a high overlap, the HRs and 95% CIs are larger for both OS and DFS from synthetic data compared to the real data. Is this just a coincidence for this example or could the data synthesis process systematically lead to slightly larger / smaller results across a dataset for synthetic data compared to real data?

It is difficult to draw strong conclusions about these differences. Because he synthesis process is stochastic, there will be slight variations when datasets are synthesized.

14 Discussion: The manuscript ends quite abruptly with the strengths and limitations section. Consider adding in a brief conclusion statement (such as the conclusions of the abstract).

We have added a conclusions section at the end of the paper with the key take-aways. The added conclusions section is as follows:

“As interest in the potential of synthetic data has been growing, an important question that remains is the extent to which similar results and the same conclusions would be obtained from the synthetic datasets compared to the real datasets. In this study we have provided one answer to that question. Our re-analysis of a published oncology clinical trial analysis demonstrated that the same conclusions can be drawn from the synthetic data. These results suggest that synthetic data can serve as a proxy for real data and would therefore make useful clinical trial data more broadly available for researchers.”

Reviewer #3

1 Page 4, lines 13-17

There has not been successful re-identification attacks on the like of clinical trials report or IPD shared in the context of EMA Policy 0070/Health Canada PRCI nor the voluntary sharing initiatives from sponsors (e.g. CSDR, Vivli, YODA, etc.) and regulators are promoting anonymization methods in this field. The paragraph and its references should be put further in context as it is speculative at this stage.

Thank you for pointing this out. We have clarified the text with the following qualification:

"Although, it should be noted that there are no known successful re-identification attacks on anonymized clinical trial data at the time of writing."

2 Page 4, lines 41-45

There are different types of secondary analyses and the appraisal ones related to checking the integrity of the study and any bias in its conduct or reporting may not be suitable for use of Synthetic Data. The paper should elaborate further and be more explicit on which types of analyses are legitimate to consider Synthetic Data for.

We have added a section "Relevance of Results" where this issue is discussed further. The added wording is as follows:

"If the objective of a secondary analysis of a clinical trial dataset is the replication / validation of a published study, then working with a synthetic variant of the dataset will not give the exact numeric results but would be expected to produce the same conclusions as was demonstrated in our study. Another type of secondary analysis is to assess bias in trial design, misreporting or selective outcome reporting where "keeping the same conclusions and comparable numerical results of all primary, secondary and safety endpoints […] is of utmost importance." [2]. The data synthesis approach we presented here achieves these objectives by including the primary and secondary endpoints in the generative model to ensure that relationships with other covariates are maintained, and it does not synthesize adverse event data to maintain the accuracy of safety data. More generally, a review of protocols found that most secondary analysis of clinical trial datasets focused on novel analyses rather than replication or validation of results [65]. In such cases, the conclusions from using synthetic data would be expected to be the same as using the real data. However, it is more difficult to make the case for using synthetic data for the primary analysis of a clinical trial dataset since the investigators and sponsors would have ready access to the real data."

3 Page 5, lines 39-46

The source dataset is coming from Project Data Sphere and is meant to be anonymized. Could you please elaborate on any risk (privacy in particular) to generate synthetic data from anonymized data rather than pseudonymized data? Are there any pitfalls to consider?

This point is now addressed in the limitations section of the revised manuscript. The added text is:

"The data we used in our analysis came from Project Data Sphere, which shares datasets that have already gone through a perturbation to anonymize the data. This would not affect our results or conclusions because the published study that we replicated used the same (perturbed) dataset from Project Data Sphere. More generally, synthetic data can be generated from pseudonymous data rather than from fully anonymized data. Multiple researchers have noted that synthetic data does not have an elevated identity disclosure (privacy) risk [60], [66]–[73]."

4 Page 7, lines 49-54

While Hiding-in-Plain-Sight (HIPS) would require synthesizing certain quasi-identifiers within bands, the use of Synthetic Data for the like of Adverse Events (certain being considered quasi-identifiers) could be misleading in CSRs shared under EMA Policy 0070 or Health Canada PRCI. To which extent synthetic data would be acceptable under these policies? Or would it be in connection with anonymization of certain quasi-identifiers? Would partial synthesize of demographics be sufficient in this context?

We have added a more detailed analysis and discussion of the selection of quasi-identifiers for sharing clinical trial datasets in the appendix. This is relevant for partial synthesis methods as well as the general approach described by the EMA and Health Canada.

The analysis shows that to manage re-identification risk, the focus needs to be on public quasi-identifiers. Not all SAEs would be public quasi-identifiers. For example, a hospitalization SAE is generally considered an acquaintance quasi-identifier rather than a public quasi-identifier and would not be affected by our quasi-identifier selection. However, if death is an SAE then it would be considered a public quasi-identifier.

A data protection method should not perturb the SAEs as this would affect the interpretation of the study. Under current practices following the EMA and Health Canada guidelines, these SAEs are considered as quasi-identifiers for measurement purposes but are typically not perturbed. We have added further clarifications on SAEs in the revised article.

In the current clinical trial, death in OS and DFS was the outcome rather than an SAE, and therefore it does not fall under the set of considerations above. The synthesis maintains the relationships between the outcomes and other variables, and hence the same conclusions were drawn in our study compared to the previously published analysis. We have also added that clarification to the appendix.

References

[1] A. R. Benaim *et al.*, “Analyzing Medical Research Results Based on Synthetic Data and Their Relation to Real Data Results: Systematic Comparison From Five Observational Studies,” *JMIR Medical Informatics*, vol. 8, no. 2, p. e16492, 2020, doi: 10.2196/16492.

[2] R. E. Foraker *et al.*, “Spot the difference: comparing results of analyses from real patient data and synthetic derivatives,” *JAMIA Open*, no. ooaa060, Dec. 2020, doi: 10.1093/jamiaopen/ooaa060.

[3] B. K. Beaulieu-Jones, Z. S. Wu, C. Williams, and C. S. Greene, “Privacy-preserving generative deep neural networks support clinical data sharing,” *bioRxiv*, p. 159756, Jul. 2017, doi: 10.1101/159756.

[4] A. Guo, R. E. Foraker, R. M. MacGregor, F. M. Masood, B. P. Cupps, and M. K. Pasque, “The Use of Synthetic Electronic Health Record Data and Deep Learning to Improve Timing of High-Risk Heart Failure Surgical Intervention by Predicting Proximity to Catastrophic Decompensation,” *Front. Digit. Health*, vol. 2, 2020, doi: 10.3389/fdgth.2020.576945.

VERSION 2 – REVIEW

REVIEWER	Kristin M Sheffield Eli Lilly and Company United States
REVIEW RETURNED	05-Feb-2021

GENERAL COMMENTS	The authors have adequately addressed my and other reviewer comments.

REVIEWER	Sarah Nevitt University of Liverpool United Kingdom
REVIEW RETURNED	18-Jan-2021

GENERAL COMMENTS	Thank you to the authors for their responses to the comments and for the edits made to their manuscript. All of my comments have been adequately addressed. This is excellent and very interesting work, I look forward to seeing it published.

REVIEWER	Jean-Marc Ferran Qualiance / PHUSE I'm a member of the EMA Technical Anonymisation Group and Health Canada Stakeholders Group for Public Release of Confidential Information. I also advise d-Wise Inc. (North Carolina) in the development of their anonymization solutions.
REVIEW RETURNED	04-Feb-2021

GENERAL COMMENTS	The updates in the first revision addresses my comments