Response to Comments

This document contains the response to the comments to the authors. In what follows, the original comments to the authors are in bold font, and the authors' responses are in regular font.

*** Please move the output data from github and into a repository that aligns with this policy - https://www.nature.com/sdata/policies/repositories**

Done (figshare)

*** Please add a data citation for the dataset to the reference list using these instructions (https://www.nature.com/sdata/publish/submission-guidelines#data_citations - note that a DOI URL should be used). Please add the reference number to wherever the dataset is mentioned in the text - the main position should be the first part of the Data Record in a sentence describing where the data has been deposited.**

Done

*** There needs to be some means for users/readers to access the input data as needed. If these are sensitive and contain personal data then the requirement is to provide some secure method of access, e.g. via application. Looking at the source data paper (https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-021-10326-6#availability-of-data-and-materials), these are noted to be available "on reasonable request". Please add details to the paper on how users can obtain these, what information etc, needs to be provided, and a guarentee that requests will be honoured to allow access to the input data as needed.**

Done.

Editorial Board Member (Comments to the Author):

We appreciate the time and effort you have invested in this work. We are pleased to inform you that the reviewers found your case study on anonymization tools to be interesting and relevant. However, they have provided valuable feedback that, if addressed, will significantly strengthen your paper. We encourage you to carefully consider their comments and revise your manuscript accordingly.

Reviewer #1 (Remarks to the Author):

In this article, the authors present a case study in which 3 anonymization tools (ARX, SDV, and SynDiffix) are evaluated on a single dataset containing cardiorespiratory fitness and commuting patterns for Slovenian school children.
The paper presents an interesting case study, however the application is limited to a single dataset which, although challenging to anonymize due to its small size, is limited in terms of being able to generalize or extrapolate the results derived from this study. My main concern is that since only one case study is presented, extrapolating the results is problematic. However, there are other concerns that have emerged during the reading of the paper.

(This comment is addressed below.)

- First of all, following the thread of the work is complicated, the results are given at the beginning (following the journal format), but it would be interesting to comment here on how the authors have arrived at these results, explaining briefly the anonymization processes carried out with each tool. This will be detailed later, but in this first section it would be good for the reader to have an idea of what is done in each case. Otherwise, it is difficult to understand the tables 1,2,3 and 4 and analyze them, since it is not known what is being done in each case to anonymize the data. I think the document needs to be reordered with more coherence.

A short description of the three anonymization methods and why they were chosen is added to the introduction. We think it is best to keep the detailed description at the end for interested readers, and in any event we don't want to depart from the recommended journal format.

-Another of my main concerns is the anonymization methods used in each case. It is clear the anonymity guarantees obtained with ARX by using k-anonymity. However, are the methods used with the 3 tools comparable? It would be important to analyze the usefulness of anonymized data in all 3 cases using metric scores, as well as to analyze the risk/protection that each of them can provide against different types of attacks.

A quality analysis across general metric scores has been done in other work already cited (and which has been added to a new prior work section). Adding a general metric evaluation to this paper would add only marginally to the prior work, would substantially lengthen the paper, and would distract from the novel contribution of this paper, which is to do an evaluation based on actual scientific conclusions. That cited work also includes a risk assessment. We have additionally added a different risk assessment to this paper.

It may be that some methods perform worse in view of the first four tables but that this is because they are providing much better privacy guarantees (it is just an assumption, it does not have to be so, but the authors should analyze and discuss it). This has to be properly analyzed and discussed. Similarly, as for the application of k-anonymity, it is

evaluated only for a value of k, it would also be interesting to see how the performance degrades as this value increases and what guarantees are provided in each case.

Added risk evaluation section. Made the point that the anonymity of all of these methods is not really in question, but nevertheless provides a risk assessment with respect to a well-regarded measure (Anonymeter attribute inference).

-Table 5 shows how ARX achieves a higher significance in relation to the conclusions obtained with the original data. This comes back to my previous point, it is necessary to analyze this carefully in terms of the tradeoff between privacy and utility obtained in each case. It may simply be that anonymization with ARX is not sufficiently strict and is therefore less guaranteed to succceed in the face of an extraction or inference attack than the other methods (again, it is a question and the authors should discuss it.).

The new risk evaluation section demonstrates that all three methods are strongly anonymous. This is in agreement with prior work (now cited in the prior work section).

-The paper is missing several very important parts in a scientific paper, namely a detailed study of the state of the art, analyzing both studies that analyze the impact of anonymization, as well as anonymization techniques and tools available. I do not want to recommend any reference to the authors in order not to bias their perspective, but a study of state of the art of these aspects is essential to understand the field, where we come from and where we are going.

A prior work section has been added. We are not sure what reviewer 1 means by “a detailed study of the state of the art”. Such a study is a major effort (10s of pages easily) and outside the scope of this paper.

-In the same line as the previous item, the authors should develop a section of conclusions and future work. As for future work, it is understood that more datasets can be analyzed, but it may also be interesting to study the performance of the anonymized data with the 3 methods as input in ML/DL models (which can also be reviewed in the state of the art), use more anonymization methods, various parameters and techniques employed, etc.

Future work is now mentioned in the discussion section.

Minor comments:
- The footnotes to the tables are too long (especially tables 1, 2, 3 and 4), this could be more detailed in the discussion of the results, with more specifics on what each value means and an analysis of what leads to SDV being the worst performing model overall (is the comparison fair? to be analyzed).

Our preference is to adequately explain each table in the caption so that the reader does not have to search through the text to understand the tables. There is already a discussion of the target use cases of SDV (i.e. data enhancement rather than data replication), which suggests why SDV is the worst performing. A more detailed analysis is out of scope for this paper.

- There are some typos throughout the text, for example, on page 10 a citation is included without putting it in the \cite{} environment (toolscitepatki2016synthetic). The expression "research burden" is repeated several times, especially in the discussion.

The citation is fixed. We use the word burden often because one of the distinguishing features of this paper is that it takes into account the burden on both the data controller as well as the analyst.

To conclude, I think that although the topic of the paper is interesting and has potential, it needs a very extensive review to provide robust scientific conclusions and to contribute a real insight and novelty. Otherwise, the contribution is not clear. Although the main concern is the limited application to a single dataset (since there is no major contribution other than the analysis, as the 3 software are used as they are), the authors also need to analyze the results in more detail, include conclusions and a more elaborated discussion, analyze the usefulness of the data in each case and especially discuss whether the results of the anonymization of the 3 cases are comparable.

We've added some text in the introduction to clarify the contributions of this paper (a key contribution being that this is the first paper that uses actual scientific conclusions as the basis for evaluating data quality). We understand that this paper is limited in scope, and look forward to additional papers using our quality methodology for other methods and scientific analyses. The new risk section demonstrates that the three cases are comparable (which we had assumed to be the case in the earlier version of the paper based on prior work).

Reviewer #2 (Remarks to the Author):

In this paper the authors addressed the challenge of anonymizing personal data for open science. It presents a case study involving a dataset of cardio-respiratory fitness and commuting patterns of Slovenian school children. The study evaluates three anonymization tools: ARX, SDV, and SynDiffix. The goals are to determine if the scientific conclusions of the original study are supported by the anonymized datasets, assess the difficulty of using these tools for data generation and analysis, and evaluate the burden on researchers. The authors provide insights into the effectiveness and usability of these anonymization tools in preserving data utility while ensuring privacy, ultimately contributing to the advancement of open science practices.

Agree. (no response needed)

The quality of the data provided in the paper appears to be quite robust, but it varies depending on the anonymization tool used. The data quality in the paper could be improved by addressing the limitations of the anonymization tools used. ARX and SynDiffix, for example, performed reasonably well but required substantial configuration and expertise, which could be streamlined to make them more accessible. SDV, on the other hand, produced lower quality data and led to incorrect scientific conclusions; enhancing its algorithms and default settings could improve its performance.

Agree. (no response needed)

Additionally, the study could benefit from a more realistic open science scenario where researchers unfamiliar with the raw data analyze the anonymized data independently. This would provide a better assessment of the tools' effectiveness in real-world applications. Expansion of the research for providing multiple datasets and analyses would enhance the generalizability of the findings, providing a more comprehensive evaluation of the anonymization tools' capabilities.

We agree with this, as was mentioned in the paper. Given that the results of this paper are cautiously positive, we hope to undertake a more realistic approach in future work.

Overall, the integrity of the data files, repository record, and code is high. The authors have made significant efforts to ensure transparency, reproducibility, and accessibility of their work. However, there is room for improvement in the performance of the SDV tool and in simplifying the configuration process for ARX and SynDiffix to make them more user-friendly.

Agree. (no response needed)