Methods for generating and evaluating synthetic longitudinal patient data: a protocol for a methodological systematic review

Administrative information

Registration

In accordance with the PRISMA guidelines (Moher et al. 2015; Page et al. 2021), our systematic review protocol was registered with the International Prospective Register of Systematic Reviews (PROSPERO) on 4 June 2021.

Authors

Katariina Perkonoja*[1,2], Martin Closter Jespersen[3], Henning Langberg[4], Antti Airola[5], Arho Virkki[1,2], Kari Auranen[1,6], Joni Virta[1]

[1] Department of Mathematics and Statistics, University of Turku, Finland
[2] Auria Clinical Informatics, Turku University Hospital, Hospital District of Southwest Finland, Finland
[3] Analytics and Cognitive, Deloitte Consulting, Denmark
[4] Department of Public Health, University of Copenhagen, Denmark
[5] Department of Computing, University of Turku, Finland
[6] Department of Clinical Medicine, University of Turku, Finland

*Corresponding author: Katariina Perkonoja, Department of Mathematics and Statistics, 20014 University of Turku, Finland, kakype@utu.fi

Email: Katariina Perkonoja kakype@utu.fi, Martin Closter Jespersen majespersen@deloitte.dk, Henning Langberg henning.langberg@regionh.dk, Antti Airola ajairo@utu.fi, Arho Virkki arho.virkki@tyks.fi, Kari Auranen kari.auranen@utu.fi, Joni Virta joni.virta@utu.fi

Author contributions

KP is the guarantor. KP drafted the manuscript. All authors contributed to the development of the selection criteria, the risk of bias assessment strategy and data extraction criteria. KP developed the search strategy and data collection form. JV and KA provided statistical expertise. MJ and AA provided expertise on machine learning and HL on the healthcare domain. All authors read, provided feedback and approved the final manuscript.

Amendments

In case there will be a need to amend this protocol, we will give the date of each amendment, describe the changes and give the rationale in this section. Changes will not be incorporated into the protocol.

There are two amendments to this protocol (dated 8 March 2023 and 25 January 2022) and they are presented in the Appendix 3.

Support

This systematic review is part of the Synthetic Health and Research Data (SHARED) project and is funded by the Novo Nordisk Foundation (grant NNF19SA0059129). This funding will support the collection of individual participant data by the original investigators, data management and analyses. The Novo Nordisk Foundation is not involved in any other aspect of the project, such as the design of the project protocol or analysis plan, data collection and actual analyses. The funder will have no input on the interpretation or publication of the results of this review.

Introduction

Rationale

Patient data are generally considered as highly sensitive personal information and are thus regulated by international and national legislation, such as the General Data Protection Regulation (GDPR) or the Health Insurance Portability and Accountability Act (HIPAA). Different regulatory regimes typically imply extended permit processing times so that access to data can take a long time, if at all possible. However, if data can be considered anonymous in the sense that no personal information can be inferred, they are no longer subject to laws on personal data protection, thus facilitating data sharing.

Synthetic data generation seeks to create artificial data that resembles real-world (i.e., empirical) data as much as possible without being genuinely personal information. Availability of synthetic data has thus been suggested as a means to facilitate secondary use of existing data for various purposes, such as research or development and innovation activities. Nevertheless, it is not always guaranteed that synthetic data protect the privacy of the subjects in the original data. There also exist cases, such as augmenting or balancing data, where privacy is not needed if the data are not processed by a third party. Moreover, synthetic data may resemble the original data in a number of ways or may not be of sufficient quality to be used in practice.

Although there are many methods for synthetic data generation, not all methods are suitable for longitudinal data, where at least some of the unit-specific variables have been measured repeatedly over time creating a special dependency structure between the observations. Patient data are usually longitudinal because new information about each patient is stored at each health-care visit or treatment. Longitudinal data often contain unique combinations, especially for

repeated measurements, making it easier to single out individuals. As a result, it is difficult to generate synthetic data that would not only preserve data utility but also be private.

Further research is needed to identify and develop suitable methods to generate synthetic longitudinal data that are safe and of sufficient quality to be used in real-life settings. Such methods can then be provided directly to the data controllers and thus expedite utilization of patient data without compromising patient safety and privacy. The results of this review can be used to select an appropriate synthetic data generation method for a particular longitudinal data synthesis setup, as well as to select methods for evaluating the utility and privacy of synthetic data.

Objectives

The aim of this systematic review is to map the currently available methods suitable for generating synthetic longitudinal patient data in real-life settings and to evaluate their performance. To this end, our primary objective is to describe the current methods and their feasibility to enable data controllers and other researchers to choose appropriate methods for their own needs. The secondary objectives are to develop a generic evaluation framework that can be used to assess the utility and privacy of synthetic data, and to test what we consider to be the most promising methods with both simulated and real-world patient data using the evaluation framework.

Methods

Eligibility criteria

In this review, we consider synthetic data as data which have been generated on the basis of some existing data using a randomized algorithm. A randomized algorithm is one that exploits randomness as part of its operation and whose operation is not based on direct re-sampling. Thus, we exclude the cases of purely simulated data not based on approximation of some existing data (e.g., simulating observations from multivariate normal distribution that has not been estimated from empirical data), re-sampling methods (e.g., jackknife, bootstrap), and deterministic methods (e.g., rule-based algorithms).

Patient data are highly diverse. In this review we confine ourselves to longitudinal data that contain numerical and/or categorical variables (covariates) that describe different patient attributes and of which at least one variable (numerical or categorical) has been measured at least twice (responses). The measurement times or the number of measurements can be different for different subjects, i.e., we allow unbalanced data. Due to the above specifications, methods developed to generate, for example, text or image data or univariate time series are not within the scope of this review, although they are common types of patient data.

Based on these definitions, we will include literature in which the presented synthetic data generation method is capable of producing longitudinal patient data. We do not limit to the health data domain but do require that the generation method can be applied to longitudinal patient data.

We will also include literature that does not address the privacy of the method or the synthetic data generated, or where the generation method is not free of charge or open-source. The following forms of publication written in English language will be included: articles published in peer-reviewed journals and proceedings as well as pre-prints, books, book chapters and reviews.

Information sources

Literature search strategies will be developed using topic (title, abstract, keywords) and text words related to synthetic longitudinal patient data. We will search EMBASE (1947 onwards), MEDLINE (Ovid interface, 1946 onwards), Web of Science (1900 onwards) and Google Scholar (Publish or Perish software, current content, first 1000 hits).

To ensure literature saturation, we will scan the reference lists of the included literature as identified through the search. We will also search the authors' homepages and may contact them to make sure that all relevant material has been captured. Finally, we will circulate a bibliography of the included literature to the systematic review team (the authors of this protocol).

Search strategy

Studies in which the synthetic data generation method is suitable for producing synthetic longitudinal patient data will be sought (see Eligibility criteria). No date limits will be set for the search. The specific search strategies will be created by the corresponding author. The search algorithm will be developed with input from the other authors and by using the Web of Science advanced search. The strategy will then be reviewed by an author (AV) that was not involved in its development, using the PRESS standard (McGowan et al. 2016). A draft Web of Science search strategy is included in Appendix 1. The search algorithm may be updated toward the end of the review. After the Web of Science strategy is finalized, it will be adapted to the syntax of the other databases.

Study records

Data management

Literature search results will be uploaded to EndNote™ Online, an online-based software that allows efficient search and management of digital references. The software is used to assess eligibility and to remove duplicates. All literature considered eligible constitute records of this study. REDCap®, a secure web platform for building and managing online databases and surveys, is used to collect and store the data items from all study records.

Selection process

The review authors (KP and MJ) will independently screen the titles and abstracts yielded by the search against the eligibility criteria. Full text for all titles that appear to meet the eligibility criteria or where there is any uncertainty about meeting the eligibility criteria will be obtained. If the full

text is not available, the title cannot be included in the review and is defined as excluded. Any disagreements will be resolved through discussion and, if necessary, a third-party arbitration (JV or HL) will be used. Reasons for excluding reports will be recorded and the PRISMA flow diagram (Page et al. 2021) will be presented in the final review article.

Data collection process

Data will be collected by using a structured form designed in REDCap®. The current form, presented in Appendix 2, was developed on the basis of a preliminary version piloted by KP, MJ and JV using two publications previously unknown to them (Abay et al. 2019; Albuquerque et al. 2011). The publications were selected from the results of a preliminary search. Prior to the data collection phase, both reviewers will appraise the form together to ensure full consensus on its use. The two reviewers will extract data independently and in duplicate from each study record. Data abstracted will include basic information of the synthetic data generation method, procedures used to evaluate it and the evaluation results. Reviewers will resolve disagreements by discussion. One of two arbitrators (JV or HL) will adjudicate unresolved disagreements. If needed, we may contact authors of the method to resolve any uncertainties. The data collection form may be updated during the review if it turns out that something relevant is missing.

Data items

For each synthetic data generation method, we will extract general information (e.g., name, type, release year, public availability, software used in implementation), type and characteristics of data used (e.g., number of observations and variables, type of the variables, number of timepoints), procedures used in the evaluation (e.g., simulated or real data, visual or quantitative assessment, consideration of data anonymity) and evaluation results, as well as the advantages and disadvantages of the method, according to both the authors and our own views. Missing information will also be recorded. All data items to be collected are presented in the data collection form (see Appendix 2).

Outcomes and prioritization

The following primary and secondary outcomes will be collected from all study records.

Primary outcomes

The primary outcome will be the reported synthetic data generation method, including information on the properties and limitations of the method, e.g., types of variables in the data, number and nature of repeated measurements, handling of missing values, and level of privacy.

Secondary outcomes

Secondary outcomes are the different approaches and data sets used to assess the utility and privacy of synthetic data generated by the proposed method.

Risk of bias in individual studies

Since this review is methodological, the assessment of biases differs from the corresponding frameworks for clinical trials. However, since different sources of bias are possible also in methodological studies, Table 1 presents a framework adapted to accommodate our research design.

Table 1: The table describes different sources of biases that have been adapted to fit the context of methodological review concerning synthetic data generation. These biases will be assessed from the study records by reviewing whether bias is present in a record and by further defining how the conclusion was reached.

Bias	Examples	Rationale	Assessment plausibility
Selection bias	Using a data set that is known in advance to perform poorly with another method that is used as a reference for the developed method Post hoc alteration of data or model selection based on arbitrary or subjective reasons Using different training, validation, or test sets when evaluating the method performance.	When evaluating the performance of a method, it may be possible to select the data set so that it works poorly with other methods, or to select a subset of data or models in order to achieve better results. Thus, the data used and the choice of model should always be justified.	Selection bias can be difficult to identify from a record because usually only the data used have been reported or are not addressed in detail. The reviewers may also not know if the data used tend to perform poorly with some other method. The bias can be observed if, for example, some public data have been used in the study or the reviewers happen to know in advance that the data are not suitable for the setup in question, or learn this during the review.
Performance bias	If the performance of the method is compared against other methods, no fine-tuning is performed on the reference methods while the method in question is fine-tuned.	When comparing with other methods, the comparison should be fair in the sense that if it is possible to improve the reference methods, this should be done. Therefore, the comparison procedure should be carefully described.	Performance bias can be difficult to identify if the sections on model selection and / or training have not been clearly addressed. The bias can be detected if, for example, the selection of hyperparameters in the reference models is not addressed or the lack of fine-tuning of the models is discussed as a limitation of the study.
Reporting bias	The performance of the method has been found to be measured in some way, but the results are only partially or not at all presented.	All metrics used in the study to evaluate the performance of the method should be described in the study and the results for these should be available to the reader.	The bias should be relatively easy to detect on the assumption that the study report has been written truthfully by including all the metrics actually used, and that they can be found in the study report, its appendices, or supplementary materials.

In addition, we will collect information on whether the source code of the method is publicly available. However, we do not validate the source code except for those methods that will be selected for further evaluation (see Objectives).

Data synthesis

All collected data will be combined by the corresponding author. Any discrepancies in the collected data between the reviewers will be resolved through discussion and, if necessary, a third-party arbitration (JV or HL) will be used.

The results of this review will be presented by describing the characteristics and findings about the eligible synthetic data generation methods. All identified methods together with their characteristics will be presented in a table that can be used, for example, to select a suitable data synthesis method. The estimated risk of biases in the conduct or reporting of the method will be included in the table, and no method will be excluded because we want to offer readers an opportunity to evaluate the identified methods themselves. Observed frequencies or proportions will be used to describe the distributions of the characteristics of the collected data items.

In addition, based on the results and other related literature, an evaluation framework will be developed to assess the quality of synthesized data. The exact instruments of the evaluation framework cannot be determined in advance but will most likely include both qualitative and quantitative criteria. The framework will be validated by generating synthetic data using a subset of reviewed methods which we consider to be the most promising. Both simulated and real-world patient data will be used in data generation.

In order for the method to be selected for further evaluation, it must at least have open-source code available freely or on request, as well as adequate privacy guarantees. In other respects, the selection will be based on the results of the review and the selection criteria will be described in the final review article.

Meta-bias(es)

In order to determine if any reporting bias is present, we will evaluate whether selective reporting appears to be present in the study records. Since the review is methodological, it is likely that there will be some publication or dissemination bias present, as it is more likely that written work on inoperative methods will not be sent for publication.

Confidence in collected evidence

To assess the level of confidence that can be placed on the collected evidence, we will gather information on the bias in individual studies (see Risk of bias in individual studies) as well as on inconsistency, imprecision, or indirectness of the reporting. Given these as well as the potential publication bias and the limitations of our search algorithm, we will discuss our confidence in the evidence gathered in the final review article.

Reference

Abay, N. C., Zhou, Y., Kantarcioglu, M., Thuraisingham, B., and Sweeney, L. (2019), “Privacy Preserving Synthetic Data Release Using Deep Learning,” in *Machine Learning and Knowledge Discovery in Databases*, Springer International Publishing, pp. 510–526. https://doi.org/10.1007/978-3-030-10925-7_31.

Albuquerque, G., Löwe, T., and Magnor, M. (2011), “Synthetic generation of high-dimensional datasets,” *IEEE Transactions on Visualization and Computer Graphics*, 17, 2317–2324. https://doi.org/10.1109/TVCG.2011.237.

Dahmen, J., and Cook, D. (2019), “SynSys: A Synthetic Data Generation System for Healthcare Applications,” *Sensors* , 19. https://doi.org/10.3390/s19051181.

McGowan, J., Sampson, M., Salzwedel, D. M., Cogo, E., Foerster, V., and Lefebvre, C. (2016), “PRESS Peer Review of Electronic Search Strategies: 2015 Guideline Statement,” *Journal of Clinical Epidemiology*, 75, 40–46. https://doi.org/10.1016/j.jclinepi.2016.01.021.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L. A., and PRISMA-P Group (2015), “Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement,” *Systematic Reviews*, 4, 1. https://doi.org/10.1186/2046-4053-4-1.

Nowok, B., Raab, G. M., Dibben, C., and Others (2016), “synthpop: Bespoke creation of synthetic data in R,” *Journal of Statistical Software*, Foundation for Open Access Statistics, 74, 1–26.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., and Moher, D. (2021), “The PRISMA 2020 statement: an updated guideline for reporting systematic reviews,” *BMJ* , British Medical Journal Publishing Group, 372. https://doi.org/10.1136/bmj.n71.

Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., Duffett, C., Dube, K., Gallagher, T., and McLachlan, S. (2018), “Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record,” *Journal of the American Medical Informatics Association: JAMIA*, 25, 230–238. https://doi.org/10.1093/jamia/ocx079.

Zhang, J., Cormode, G., Procopiuc, C. M., Srivastava, D., and Xiao, X. (2017), “PrivBayes: Private Data Release via Bayesian Networks,” *ACM Transactions on Database Systems*, New York, NY, USA: Association for Computing Machinery, 42, 1–41. https://doi.org/10.1145/3134428.

Appendix 1

The following search algorithm has been developed with the Web of Science advanced search. In the development of the algorithm, already known articles have been used as a benchmark (Dahmen and Cook 2019; Nowok et al. 2016; Walonoski et al. 2018; Zhang et al. 2017) and two articles from the preliminary search results were selected to pilot the data collection form (Abay et al. 2019; Albuquerque et al. 2011).

The following algorithm provides 3 690 hits (may include duplicates).

TS = ((synthetic OR artificial) NEAR/3 (data OR record*)
AND (generat* OR produc* OR simula*)
AND (longitudinal OR correl* OR panel OR repeat* OR follow-up OR multivariate OR lifespan* OR traject* OR health* OR medical OR patient))
AND LANGUAGE: (English)
AND DOCUMENT TYPES: (Article OR Abstract of Published Item OR Book OR Book Chapter OR Early Access OR Proceedings Paper OR Review OR Software Review)

Appendix 2

The following forms will be used to collect data from all eligible records. All possible questions are presented in the forms, but their occurrence may be conditional on previous choices. In addition, the layout of the form differs slightly from the electronic form that will be used to collect all the data.

Literature information

Please complete the survey below.

1) Reviewer
   ○ Katariina
   ○ Martin

2) Type of publication
   _____

3) Authors
   _____

4) Year
   _____

5) Title
   _____

6) Publication platform (journal, conference, book...)
   _____

7) Volume
   _____

8) Issue
   _____

9) Page numbers
   _____

Method characteristics

Please complete the survey below.

Basic information	
Type of the method	☐ Generative adversarial network ☐ Auto-encoder (variational or other) ☐ Other deep learning method ☐ Bayesian network ☐ Hidden Markov model ☐ Density estimation ☐ Imputation method ☐ Dimensionality reduction ☐ Data partitioning ☐ Decision tree (classification, regression) ☐ Posterior predictive sampling ☐ Clustering ☐ Other
Type of the method (other / other deep learning)	_____
Describe the method as concisely as possible	_____
Method's running time in terms of the input size (Big O notation) if reported	_____
Is the software used to apply the method available?	◯ Yes ◯ No
Is the software used to apply the method free?	◯ Yes ◯ No ◯ Not specified
Is the software a	◯ Library ◯ Standalone software ◯ Other ◯ Not specified
Define the other software type	_____
Software licence	_____

Programming language	◯ R ◯ Python ◯ C++ ◯ Java ◯ Scala ◯ Julia ◯ Fortran(77/90/95/...) ◯ Matlab/Octave ◯ SAS ◯ Other ◯ Not specified
Programming language (other)	_____
Method's source code location if provided	_____ (e.g. URL)

System requirements and complexity

Operating system	_____
Supports GPU acceleration?	◯ Yes ◯ No
Other system requirements	_____

Input (original) and output (synthetic) data properties

The method is capable of	☐ Handling categorical original data ☐ Handling numerical original data ☐ Generating categorical synthetic data ☐ Generating numerical synthetic data
The numerical data values generated	☐ Will not necessarily fall within the corresponding range in the input data set ☐ Will fall within the corresponding range in the input data set ☐ Will be replicates of values in the input data set
Is the method capable of handling unbalanced longitudinal data?	◯ Yes ◯ No ◯ Not specified
Is the method capable of generating unbalanced longitudinal data?	◯ Yes ◯ No ◯ Not specified
Is the method capable of handling missing values in original data?	◯ Yes ◯ No ◯ Not specified

Is the method capable of producing missing values with the same pattern for synthetic data as in the original data?	◯ Yes ◯ No ◯ Not specified
Give the number of repeated measurements (i.e., sequence length) the method is capable of generating in theory	_____ (Options: 1,2,..., unlimited, same as input, not specified)
Give the number of repeated measurements (i.e., sequence length) the method is capable of generating in practice (without losing accuracy) if reported	_____ (Options: 1,2,..., unlimited, same as input, not specified)
Does the method have other limitations or requirements for original data?	◯ Yes ◯ No
Describe the requirements / limitations regarding to original data	_____
Does the method have other limitations or requirements for synthetic data?	◯ Yes ◯ No
Describe the requirements / limitations regarding to synthetic data	_____

Method performance evaluation

Please complete the survey below.

Data used to generate synthetic data	
Synthetic data was generated based on	☐ Public data ☐ Simulated data ☐ Real-world data (not public) ☐ Synthetic data ☐ Other
Name and source of data if reported. Also specify the type of 'Other' if selected.	_____
Dimensions of the original data set (n x p). If multiple data sets are used, separate them with comma.	_____ (If not reported or recoverable, write 'Not specified')
Number of categorical variables. If multiple data sets are used, separate them by comma.	_____ (Options: 0,1,,2,..., not specified)
Number of numerical variables. If multiple data sets are used, separate them by comma.	_____ (Options: 0,1,,2,..., not specified)
Did the original data include repeated measurements?	◯ Yes ◯ No
Number of repeated measurements. For unbalanced data, give the range [min, max]. If multiple data sets are used, separate them by comma.	◯ Yes ◯ No

Evaluation setup of the generated synthetic data	
The evaluation of the generated synthetic data was based on	☐ Visual assessment ☐ Quantitative assessment ☐ Other (Select all suitable options)
The evaluation of the generated synthetic data was based on	☐ A single repetition (i.e., the assessment is based on a single generated synthetic data set) ☐ A small amount of repetitions (< 50) ☐ A large amount of repetitions (>= 50) (Select all suitable options)
Describe the other approach used to evaluate the synthetic data	_____

Was any of the following used to describe or evaluate the generated synthetic data and/or the method	☐ Descriptive statistics ☐ Statistical inference ☐ Prediction/classification ☐ Privacy ☐ Other (Select all suitable options)
Describe the other method(s) used to describe or evaluate the generated synthetic data	_____
Was the generated synthetic data evaluated	☐ Against original data ☐ Against other simulated data ☐ Against other real-world data (public / private) ☐ Against another synthetic data set(s) generated by the same method (e.g., using different parameters) ☐ Against another synthetic data set(s) generated by a different method or methods ☐ No comparisons to other data or methods were made (i.e., a single data set was generated) ☐ Other (Select all suitable options)
Describe the other approach to used to evaluate the generated synthetic data (in terms of data)	_____

Descriptive methods used to characterize and/or evaluate the generated synthetic data set(s)

Specify all descriptive statistics (e.g., estimates, figures) used to describe and/or evaluate the generated synthetic data set(s)	_____

Inferential statistics used to evaluate the generated synthetic data set(s)

Specify all inferential statistics (e.g., tests, models) used to evaluate the generated synthetic data set(s)	_____

Predictive and classification approaches used to evaluate the generated synthetic data set(s)

Specify all the predictive and classification approaches (e.g., models, accuracy measures) used to evaluate the generated synthetic data set(s) in terms of synthetic data performance	_____

Privacy of the method and the generated synthetic data set(s)

Was the privacy of the generated synthetic data and/or the method addressed?	◯ Yes ◯ No

Specify how the privacy of the method and/or the generated synthetic data set(s) was addressed: specify the approach (e.g., distinguishing records with a model) and parameters used (e.g., epsilon in differential privacy) if reported or write a summary of the authors' discussion on the subject if no specific approach was used.	_____

Advantages and disadvantages of the method

Did the authors discuss the advantages / disadvantages of the method?	◯ Yes ◯ No
Write down the advantages of the method according to the authors	_____
Write down the disadvantages of the method according to the authors	_____
Write down the advantages of the method according to you (if not mentioned earlier)	_____
Write down the disadvantages of the method according to you (if not mentioned earlier)	_____

General remarks on the evaluation of the method and the synthetic data

General remarks on the evaluation of the method and the synthetic data that were not addressed here	_____

Assessment of bias and reporting quality

Please complete the survey below.

For more information, see "Risk of bias in individual studies" in the review protocol.

Selection bias

Does the study show evidence of selection bias?

◯ Yes
◯ No

Assumption: The data used and the choice of model(s) should always be justified.

Examples:

Using a data set that is known in advance to perform poorly with another method that is used as a reference for the developed method Post hoc alteration of data or model inclusion based on arbitrary or subjective reasons Using different training, validation, or test sets when evaluating the method performance

Describe the selection bias present

_____

Performance bias

Does the study show evidence of performance bias?

◯ Yes
◯ No

Assumption: Method comparison procedures should be fair and carefully described.

Examples:

No fine-tuning is performed on the reference methods while the method in question is fine-tuned.

Describe the performance bias present

_____

Reporting bias

Does the study show evidence of reporting bias?

◯ Yes
◯ No

Assumption: All metrics used in the study to evaluate the performance of the method should be described in the study and the results for these should be available to the reader.

Examples:

The performance of the method has been found to be measured in some way, but the results are only partially or not at all presented.

Describe the reporting bias present

_____

Inconsistency, imprecision and indirectness of reporting	
Did the study show evidence of	☐ Inconsistency of reporting ☐ Imprecision of reporting ☐ Indirectness of reporting ☐ None of the above
Describe the type of inconsistency present	_____
Describe the type of imprecision present	_____
Describe the type of indirectness present	_____

Appendix 3

Protocol amendment: Methods for generating and evaluating synthetic longitudinal patient data

Administrative information

Registration

In accordance with the PRISMA guidelines (Moher et al. 2015; Page et al. 2021), our systematic review protocol was registered with the International Prospective Register of Systematic Reviews (PROSPERO) on 4 June 2021 with the identification number of CRD42021259232.

Authors

Katariina Perkonoja*[1,2], Martin Closter Jespersen[3], Henning Langberg[4], Antti Airola[5], Arho Virkki[1,2], Kari Auranen[1,6], Joni Virta[1]

[1] Department of Mathematics and Statistics, University of Turku, Finland
[2] Auria Clinical Informatics, Turku University Hospital, Hospital District of Southwest Finland, Finland
[3] Analytics and Cognitive, Deloitte Consulting, Denmark
[4] Department of Public Health, University of Copenhagen, Denmark
[5] Department of Computing, University of Turku, Finland
[6] Department of Clinical Medicine, University of Turku, Finland

*Corresponding author: Katariina Perkonoja, Department of Mathematics and Statistics, 20014 University of Turku, Finland, kakype@utu.fi

Email: Katariina Perkonoja kakype@utu.fi, Martin Closter Jespersen majespersen@deloitte.dk, Henning Langberg henning.langberg@regionh.dk, Antti Airola ajairo@utu.fi, Arho Virkki arho.virkki@tyks.fi, Kari Auranen kari.auranen@utu.fi, Joni Virta joni.virta@utu.fi

Amendments

This is a protocol amendment to “Methods for generating and evaluating synthetic longitudinal patient data: a protocol for a methodological systematic review”. All amendments together with rationale are provided in Table 1 below. The table contains the amendments added in this version as well as in previous versions in chronological and section wise order.

Table 1: The table describes the amendment version, effective date of the change in protocol, section where the change would be found in the protocol, the original and revised parts (in bold) and justification for the change.

Version	Date	Section	Original protocol	Revised protocol	Rationale
1	25-Jan-22	Information sources	We will search EMBASE (1947 onwards), MEDLINE (Ovid interface, 1946 onwards), Web of Science (1900 onwards) and Google Scholar (Publish or Perish software, current content, first 1000 hits).	We will search EMBASE (1947 onwards), MEDLINE (Ovid interface, 1946 onwards), Web of Science (1900 onwards) and Google Scholar (Publish or Perish software, current content, first 1000 hits). **The use of arXiv® will be considered if suitable tools are available for exporting literature. If it is used, it will be reported in the final review article.**	The sentence was added because the platform in question often contains the latest methods that have not yet been published.
1	25-Jan-22	Search strategy	After the Web of Science strategy is finalized, it will be adapted to the syntax of the other databases.	After the Web of Science strategy is finalized, it will be adapted to the syntax of the other databases. **The search can be updated after the first title and abstract screening phase is completed to include the latest research in the review.**	The sentence was added because screening titles and abstracts is a time consuming process and we want to keep the option to include the latest methods.
1	25-Jan-22	Study records / Data management	Literature search results will be uploaded to EndNote™ Online, an online-based software that allows efficient search and management of digital references	Literature search results will be uploaded to **Rayyan — a web and mobile app for systematic reviews (Ouzzani et al. 2016).**	When we started screening abstracts, we found that EndNote™ Online was not well suited for screening abstracts, and collaborating on the platform was challenging. We ended up switching to Rayyan, as this tool has been specifically developed for systematic literature reviews and allows for collaboration between evaluators, including blinding

Version	Date	Section	Original protocol	Revised protocol	Rationale
					when assessing eligibility. This software was also used to remove some duplicates that EndNote™ Online did not recognize (n = 282).
1	25-Jan-22	Study records / Selection process	The review authors (KP and MJ)... a third-party arbitration (JV or HL) will be used.	The review authors (KP and **JV**)... a third-party arbitration (**MJ** or HL) will be used.	JV replaced MJ as the second review author and MJ replaced JV as a third-party arbitrator. The change was due to the time requirements of this study and the change was approved by the review team.
1	25-Jan-22	Study records / Selection process	Full text for all titles that appear to meet the eligibility criteria or where there is any uncertainty about meeting the eligibility criteria will be obtained.	Full text for all titles that appear to meet the eligibility criteria or where there is any uncertainty about meeting the eligibility criteria will be obtained. **A flowchart of the abstract screening process is presented in Figure 1 below.**	After piloting the abstract screening phase with approximately 200 literature search results, KP and JV developed a flowchart on how to screen the included literature. The flowchart is presented in Figure 1 of this protocol amendment.
1	25-Jan-22	Study records / Data collection process	One of two arbitrators (JV or HL)	One of two arbitrators (**MJ** or HL)	MJ replaced JV as a third-party arbitrator.
1	25-Jan-22	Data synthesis	a third-party arbitration (JV or HL) will be used	a third-party arbitration (**MJ** or HL) will be used	MJ replaced JV as a third-party arbitrator.
1	25-Jan-22	Reference		**Ouzzani, M., Hammady, H., Fedorowicz, Z., Elmagarmid, A. (2016), "Rayyan — a web and mobile app for systematic reviews", Systematic Reviews, 5, 210, https://doi.org/10.1**	Reference for Rayyan added.

Version	Date	Section	Original protocol	Revised protocol	Rationale
				186/s13643-016-0384-4	
2	9-March-23	Support	This systematic review is part of the Synthetic Health and Research Data (SHARED) project and is funded by the Novo Nordisk Foundation (grant NNF19SA0059129). This funding will support the collection of individual participant data by the original investigators, data management and analyses. The Novo Nordisk Foundation is not involved in any other aspect of the project, such as the design of the project protocol or analysis plan, data collection and actual analyses. The funder will have no input on the interpretation or publication of the results of this review.	This systematic review is part of the Synthetic Health and Research Data (SHARED) project and is funded by the Novo Nordisk Foundation (grant NNF19SA0059129) **and the Finnish Cultural Foundation (grant 00220801)**. This funding will support the collection of individual participant data by the original investigators, data management and analyses. The Novo Nordisk Foundation **and the Finnish Cultural Foundation are** not involved in any other aspect of the project, such as the design of the project protocol or analysis plan, data collection and actual analyses. The funder**s** will have no input on the interpretation or publication of the results of this review.	The literature review has been funded by the corresponding author's personal grant from January 1 2023 onward.
2	9-March-23	Eligibility criteria	We do not limit to the health data domain but do require that the generation method can be applied to longitudinal patient data. We will also include literature…	We do not limit to the health data domain but do require that the generation method can be applied to longitudinal patient data. **In order to make sure that the method is capable of producing the aforementioned**	The sentences were added to clarify the eligibility criteria used in the review.

Version	Date	Section	Original protocol	Revised protocol	Rationale
				data, we require that longitudinal data have been used as original data, i.e., input data, for the generated synthetic data. Literature in which data that were originally longitudinal are somehow manipulated so that they lose their temporal structure, e.g., through aggregation, is excluded from the review. Furthermore, we require that the generated data are fully synthetic. We will also include literature…	
2	9-March-23	Study records / Selection process	From amendment 1 (see above) The review authors (KP and JV)... a third-party arbitration (MJ or HL) will be used.	The review authors (KP and JV)... a third-party arbitration (**KA** or HL) will be used.	Update to amendment 1: KA replaced MJ as a third-party arbitrator. MJ is no longer part of the review team.
2	9-March-23	Study records / Selection process	From amendment 1 (see above) Full text for all titles that appear to meet the eligibility criteria or where there is any uncertainty about meeting the eligibility criteria will be obtained. A flowchart of the abstract screening process is presented in Figure 1 below.	Full text for all titles that appear to meet the eligibility criteria or where there is any uncertainty about meeting the eligibility criteria will be obtained. **Flowcharts** of the abstract screening **and full paper review** process**es are** presented in Figure**s 1 and 2, respectively.**	Update to amendment 1: After piloting the full paper screening phase with approximately 20 literature search results included from the abstract screening phase, KP and JV developed a flowchart on how to screen the included literature. The flowchart is presented in Figure 2 of this protocol amendment.

Version	Date	Section	Original protocol	Revised protocol	Rationale
2	9-March-23	Study records / Data collection process	The data collection form may be updated during the review if it turns out that something relevant is missing.		The updated form is available on https://users.utu.fi/kakype/research/synthetic-longitudinal-patient-data/
2	9-March-23	Study records / Data collection process	The publications were selected from the results of a preliminary search. Prior to the data collection phase, both reviewers will appraise the form together to ensure full consensus on its use. The two reviewers will extract data independently and in duplicate from each study record. Data abstracted will include basic information of the synthetic data generation method, procedures used to evaluate it and the evaluation results. Reviewers will resolve disagreements by discussion. One of two arbitrators (JV or HL) will adjudicate unresolved disagreements. If needed, we may contact authors of the method to resolve any uncertainties.	The publications were selected from the results of a preliminary search. **The corresponding author (KP) will extract data from each study record.** Data abstracted will include basic information of the synthetic data generation method, procedures used to evaluate it and the evaluation results. **In unclear situations and in the assessment of bias and reporting quality, KP will consult the two arbitrators JV and KA.** If needed, we may contact authors of the method to resolve any uncertainties.	The corresponding author will extract all data alone instead of the original plan of using two reviewers. However, KP consults JV and KA in unclear situations and in the assessment of bias and reporting quality. We ended up with this change because otherwise the review would have taken longer due to JV's limited time resources. However, we are aware that using only one reviewer for data collection increases the risk of error, and therefore the corresponding author must be especially careful while extracting information. However, with regard to bias and quality of reporting, we felt that the opinion of one reviewer was not sufficient. Therefore, if KP detects a bias or issues in the quality of reporting, she must confirm the conclusions with JV and KA.
2	9-March-23	Study records / Data collection process	From amendment 1 (see above) One of two arbitrators (MJ or HL)	One of two arbitrators (**KA** or HL)	Update to amendment 1: KA replaced MJ as a third-party arbitrator. MJ is no longer part of the review team.

Version	Date	Section	Original protocol	Revised protocol	Rationale
2	9-March-23	Data synthesis	All collected data will be combined by the corresponding author. Any discrepancies in the collected data between the reviewers will be resolved through discussion and, if necessary, a third-party arbitration (JV or HL) will be used.	All collected data will be combined by the corresponding author.	The last sentence was removed because only the corresponding author collects data.
2	9-March-23	Data synthesis	From amendment 1 (see above) a third-party arbitration (MJ or HL) will be used	a third-party arbitration (**KA** or HL) will be used	Update to amendment 1: KA replaced MJ as a third-party arbitrator. MJ is no longer part of the review team.

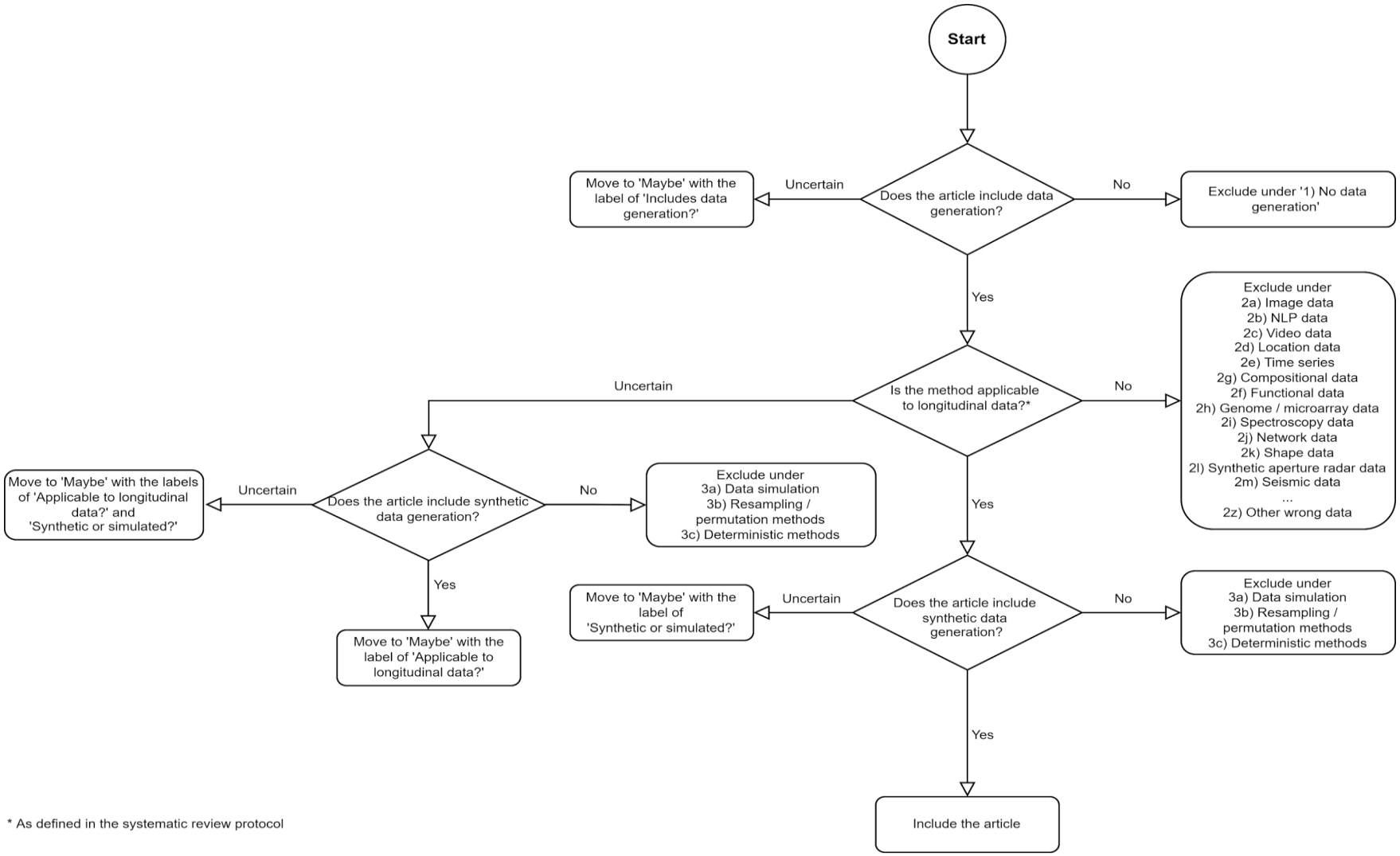

Figure 1: The flowchart describes the screening process for titles and abstracts. The first step is to assess whether or not data is being generated, regardless of the type of data generation method used. It is then assessed whether this generation method is suitable for longitudinal data (as defined in the review protocol). In order to facilitate the work of other researchers in the future, we have sought to classify different types of data and will present these classified bibliographies in the supplementary material of the final article. However, this classification may change, for example, some may be combined and more categories may be added. Finally, it is evaluated whether the article includes synthetic data generation according to our protocol or not.

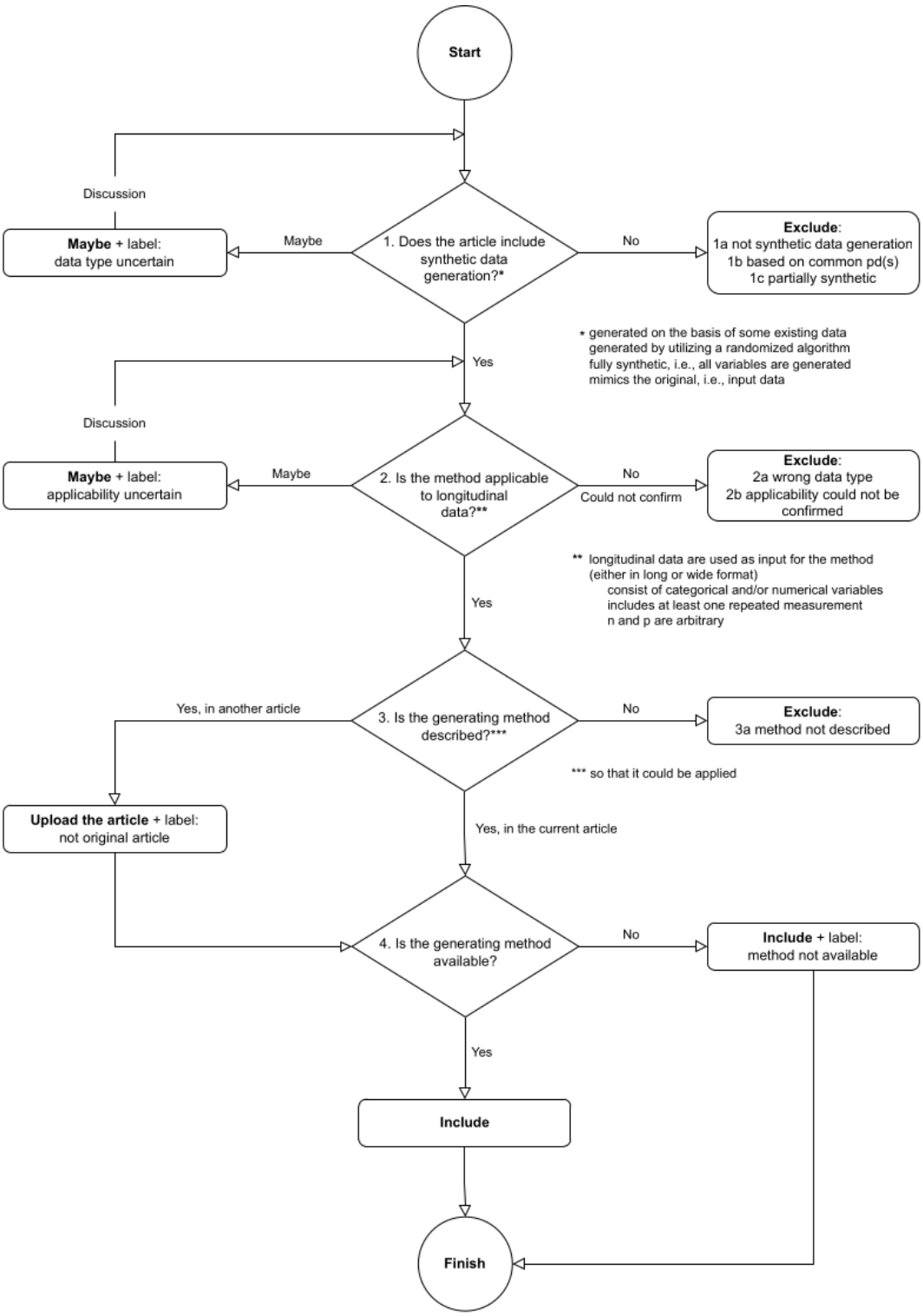

Figure 2: Flowchart used to assess the eligibility of the literature included from the abstract screening phase. Full text, including supplements and other related material, are considered while making the decisions. Both review authors, KP and JV, will independently screen the included literature based on this flow chart.

Reference

Ouzzani, M., Hammady, H., Fedorowicz, Z., Elmagarmid, A. (2016), “Rayyan — a web and mobile app for systematic reviews”, Systematic Reviews, 5, 210, https://doi.org/10.1186/s13643-016-0384-4