Supplemental Material to "Methods for generating and evaluating synthetic longitudinal patient data: a systematic review"

Katariina Perkonoja[1], Kari Auranen[1], Joni Virta[1]

Contents

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

A Search algorithms

A.1 Web of Science (Core Collection)

Search date 2021-06-11, 3795 hits

```
#1 TS = ((synthetic OR artificial)
    NEAR/3 (*data* OR record*))
AND TS = ((generat* OR produc* OR simula*))
AND TS = ((longitudinal OR correl* OR panel
           OR repeat* OR follow-up
           OR multivariate OR lifespan*
           OR traject* OR health*
           OR medical OR patient))
NOT TS = (aperture OR insemination OR seism*)
AND LA = (English)
AND DT = (Article OR Abstract of Published Item
          OR Book OR Book Chapter OR Data Paper
          OR Early Access OR Proceedings Paper
          OR Review OR Software Review)
```

Search date 2022-11-22, 1734 hits

```
TS = ((synthetic OR artificial)
    NEAR/3 (*data* OR record*))
AND TS = ((generat* OR produc* OR simula*))
AND TS = ((longitudinal OR correl* OR panel
           OR repeat* OR follow-up
           OR multivariate OR lifespan*
           OR traject* OR health*
           OR medical OR patient))
NOT TS = (aperture OR insemination OR seism*)
AND LA = (English)
AND DT = (Article OR Abstract of Published Item
          OR Book OR Book Chapter OR Data Paper
          OR Early Access OR Proceedings Paper
          OR Review OR Software Review)
NOT #1
```

A.2 Embase (1947 onwards)

Search date 2021-06-11, 504 hits

```
#1 (((synthetic OR artificial)
      NEAR/3 (data OR record*)):ti,ab,kw)
      AND (generat* OR produc* OR simula*):ti,ab,kw
      AND (longitudinal OR correl* OR panel
            OR repeat* OR 'follow?up'
            OR multivariate OR lifespan*
            OR traject* OR health* OR medical
            OR patient):ti,ab,kw
      AND ([article]/lim OR [article in press]/lim
            OR [conference paper]/lim
            OR [conference review]/lim
            OR [data papers]/lim OR [letter]/lim
            OR [note]/lim OR [review]/lim
            OR [short survey]/lim)
     AND [english]/lim
     AND [embase]/lim
```

Search date 2022-11-22, 326 hits

```
(((synthetic OR artificial)
      NEAR/3 (data* OR record* OR microdata*)):ti,ab,kw)
      AND (generat* OR produc* OR simula*):ti,ab,kw
      AND (longitudinal OR correl* OR panel OR repeat*
                  OR 'follow?up' OR multivariate OR lifespan*
                  OR traject* OR health* OR medical OR patient):ti,ab,kw
      NOT (aperture OR insemination OR seism*):ti,ab,kw
      AND ([article]/lim OR [article in press]/lim
                  OR [conference paper]/lim OR [conference review]/lim
                  OR [data papers]/lim OR [letter]/lim OR [note]/lim
                  OR [review]/lim OR [short survey]/lim)
      AND [english]/lim NOT #1
```

A.3 MEDLINE (Ovid interface, 1946 onwards)

Search date 2021-06-12, 574 hits

```
#1 (((synthetic or artificial)
    adj3 (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)).ti,ab,kf.
 #2        limit #1 to ((english language or english)
    and (classical article or clinical conference
        or comparative study or congress
        or english abstract or evaluation study
        or festschrift or government publication
        or historical article
        or introductory journal article
        or journal article
        or letter or preprint or "review"
        or "systematic review" or technical report
        or validation study))
```

Search date 2022-11-22, 402 hits (contains duplicates with the previous search because the time range could not be specified more precisely)

```
#3 (((synthetic or artificial)
    adj3 (data* or record* or microdata*))
    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)
    not (aperture OR insemination OR seism*)).ti,ab,kf
#4 limit #3 to ((english language or english)
    and (classical article or clinical conference
    or comparative study or congress
    or english abstract or evaluation study
    or festschrift or government publication
    or historical article
    or introductory journal article
    or journal article or letter or preprint
    or "review" or "systematic review"
    or technical report or validation study))
#5 limit #3 not #2
```

A.4 Google Scholar (Publish or Perish, 1000 first hits)

Search date 2021-06-18, 980 hits

```
("synthetic data" OR "artificial data")
AND (generat* OR priduc* OR simula*)
AND (longitudinal OR correl* OR panel
    OR repeat* OR "follow up" OR "follow-up"
    OR "multivariate OR lifespan* OR traject*
    OR health* OR medical OR patient)
```

A.5 arXiv

Open-source metadata were downloaded from Kaggle [1] and R software (version 4.2.2) [2] was used to extract the relevant articles. The source code is presented below.

Search date 2022-11-22, 628 hits

```
# libraries
library(jsonlite)
library(data.table)
```

```
library(synthesisr)



# imoporting ArXiv results

arxiv <- stream_in(file(paste0(getwd(), "/articles/source_searches/arxiv-
metadata-oai-snapshot.json")))

arxiv <- as.data.table(arxiv)



# regex developed according to database search queries

# synthetic data

search_data <-
"\\b(synthetic|artificial)(?:\\W+\\w+){0,3}?\\W?(\\S*data|record\\S*)\\b"



# inclusion criteria

search_gener <- "(generat|produc|simula)"

search_type <- "(longitudinal|correl|panel|repeat|follow-
up|multivariate|lifespan|traject|health|medical|patient)"

# exclusion criteria

search_excl <- "(aperture|insemination|seism)"



# grepping abstracts according to criteria

arxiv_results_1 <- arxiv[grepl(search_data, abstract, ignore.case = T, perl =
T)]

arxiv_results_2 <- arxiv_results_1[grepl(search_gener, abstract, ignore.case =
T, perl = T)]

arxiv_results_3 <- arxiv_results_2[grepl(search_type, abstract, ignore.case = T,
perl = T)]

arxiv_results_4 <- arxiv_results_3[!grepl(search_excl, abstract, ignore.case =
T, perl = T)]



# modifying data for export

arxiv_results_4[, source_type := ifelse(is.na(`journal-ref`), "UNPB", "JOUR")]

arxiv_results_4[is.na(`journal-ref`), `journal-ref` := paste0("arXiv preprint
arXiv:", id)]

arxiv_results_4[, year := year(update_date)]
```

```
setnames(arxiv_results_4, "journal-ref" , "journal")

setnames(arxiv_results_4, "update_date" , "date_generated")

setnames(arxiv_results_4, "authors" , "author")


arxiv_results_4[, c("id", "submitter", "comments", "report-no",

                    "categories", "license", "versions", "authors_parsed") :=
NULL]


setcolorder(arxiv_results_4, c("date_generated", "source_type", "author",

                               "year", "title", "journal", "doi"))


arxiv_results_4[, author := gsub(",", " and", author)]

arxiv_results_4[, author := gsub("\n", "", author)]

arxiv_results_4[, author := gsub("\\\\", "", author)]

arxiv_results_4[, author := gsub("\\\"", "", author)]

arxiv_results_4[, author := gsub("[(]\\d[)](\\W?and)?", "and", author)]


# exporting as ris file

write_refs(as.data.frame(arxiv_results_4), format = "ris", file =
paste0(getwd(), "/articles/source_searches/arxiv_results.ris"))
```

B Selection process

B.1 Abstract screening chart

The flowchart presented in Figure 1 was used by KP and JV to independently screen the titles and abstracts yielded by the search.

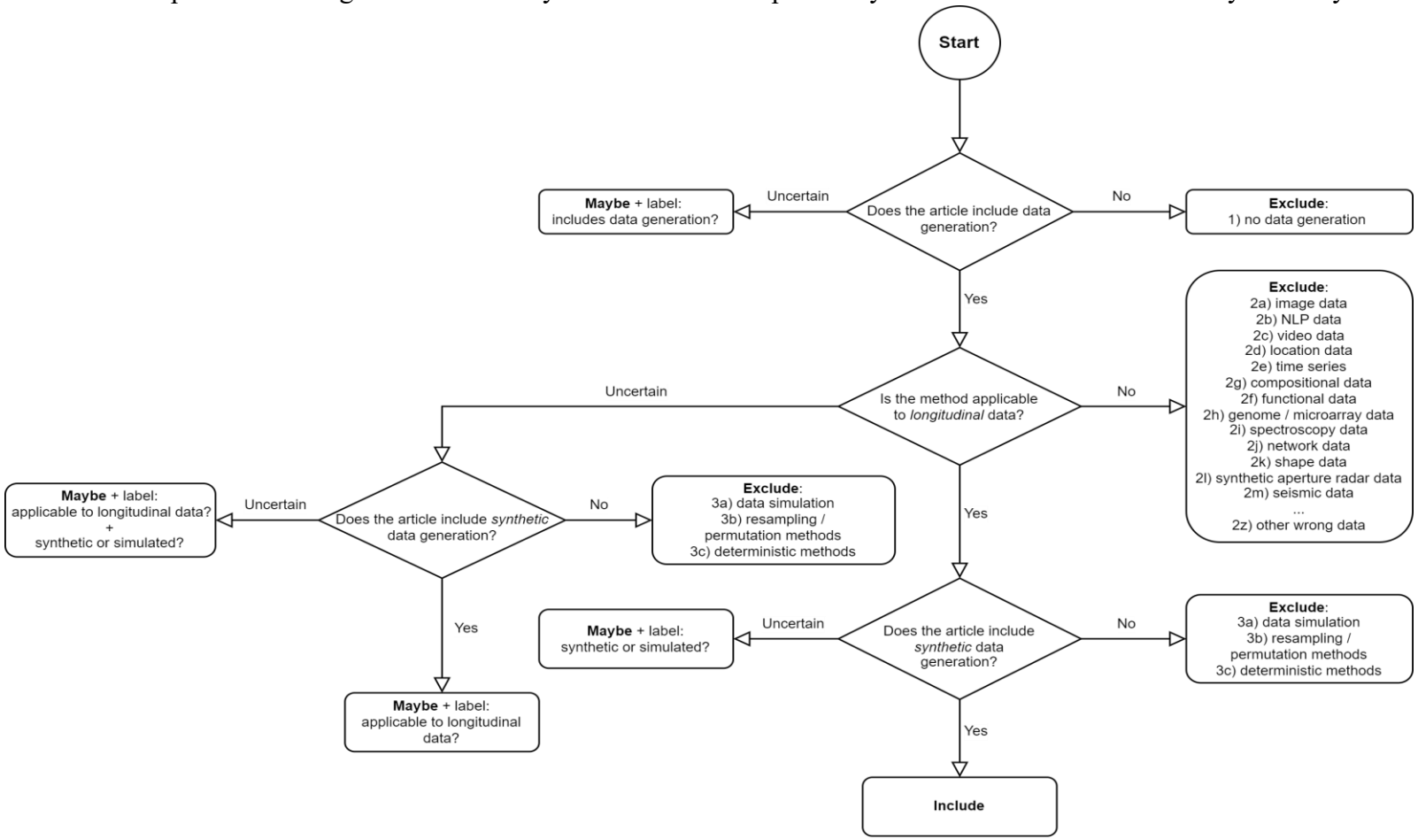

Figure 1. Title and abstract screening flowchart. Each included search result was screened by KP and JV independently using Rayyan [3] and this flowchart. The process started at the top of the chart (Start) and progressed in the directions indicated by the arrows, depending on the selection. The terminations of the process and the selection of the search result (include, maybe, exclude) are indicated in bold

B.2 Full-text screening chart

The flowchart presented in Figure 2 was used by KP and JV to independently screen the full texts of publications that had been deemed as potentially eligible (classified as 'Maybe' or 'Included') following the title and abstract screening.

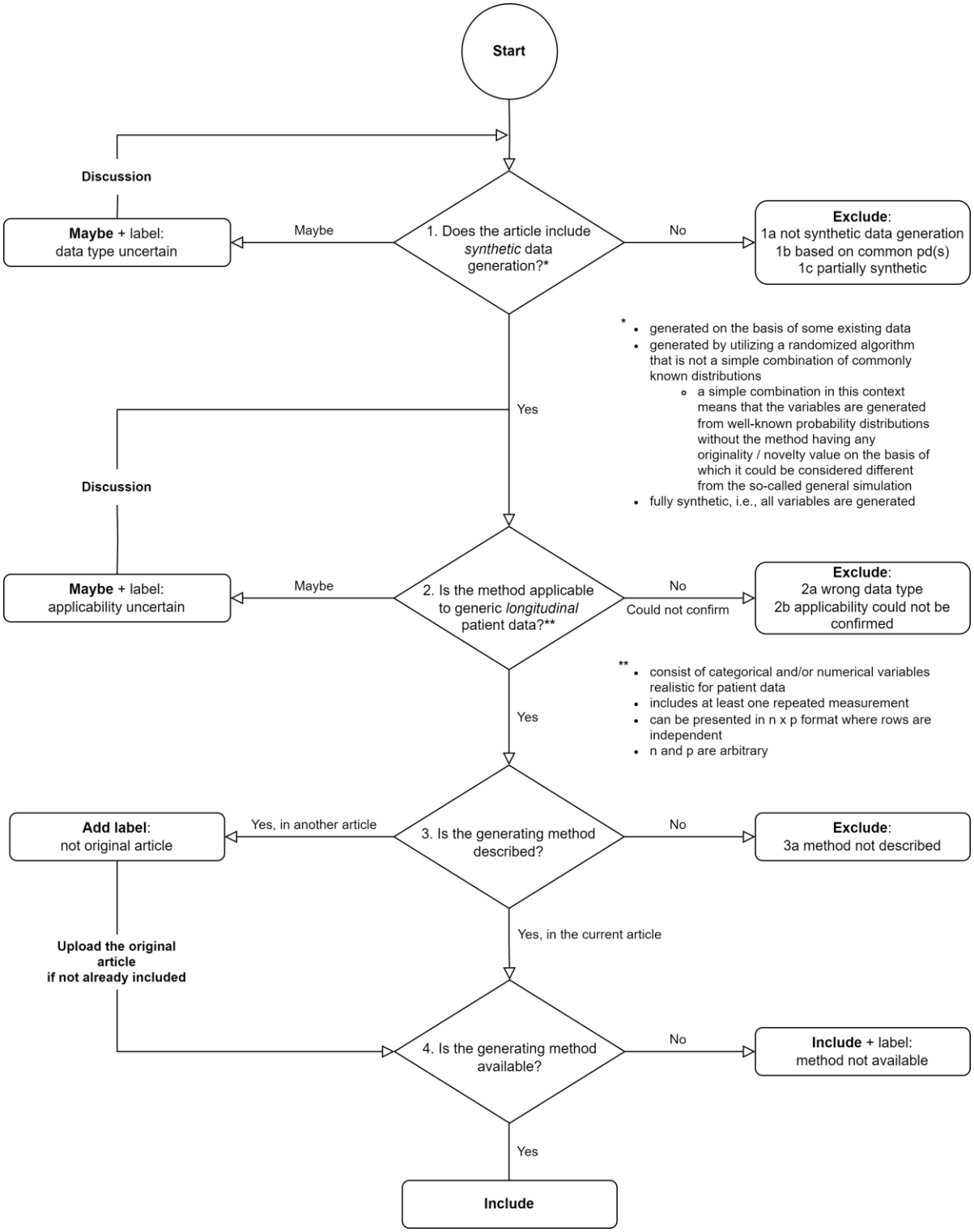

Figure 2. Full-text screening flowchart. Full texts of each publication that was included after screening the titles and abstracts was screened by KP and JV independently using Rayyan [3] and this flowchart. The process started at the top of the chart (Start) and progressed in the directions indicated by the arrows, depending on the selection. The actions and terminations of the process and the selection of the search result (include, maybe, exclude) are indicated in bold

C Data collection process

Data were collected and managed by the corresponding author using a structured form designed in REDCap electronic data capturing tools hosted at University of Turku [4,5]. The forms are presented below.

C.1 Literature information

Literature information

Please complete the survey below.

Type of publication	◯ Journal article ◯ Poster ◯ Conference paper ◯ Book chapter ◯ Dissertation or thesis ◯ Research report ◯ Review article ◯ Other
Name the other publication type	_____
Authors	_____
Year	_____
Title	_____
Publication platform (journal, conference, book...)	_____ (Give the name of the journal/conference etc.)
Volume	_____
Issue	_____
Page numbers	_____
Is the publication peer-reviewed?	◯ Yes ◯ No
What was the purpose of the study?	_____ (In this context, study refers to the article)

C.2 Method characteristics

Method characteristics

Please complete the survey below.

Basic information	
Type of the method	☐ Generative adversarial network ☐ Recurrent neural network ☐ Auto-encoder (variational or other) ☐ Bayesian network ☐ Hidden Markov model ☐ Density estimation ☐ Imputation method ☐ Dimensionality reduction ☐ Data partitioning ☐ Decision tree (classification, regression) ☐ Posterior predictive sampling ☐ Clustering ☐ Other deep learning method ☐ Other
Type of the method (other/other deep learning)	_____
Is expert knowledge required/used in the method?	◯ Yes ◯ No
How is expert knowledge needed?	_____
Describe the method as concisely as possible	_____
Programming language	☐ R ☐ Python ☐ C++ ☐ Java ☐ Scala ☐ Julia ☐ Fortran(77/90/95/...) ☐ Matlab/Octave ☐ SAS ☐ Other ☐ Not specified
Programming language (other)	_____
Is the pseudocode of the method presented?	◯ Yes ◯ No
Is the source code/software provided?	◯ Yes ◯ No ◯ Upon request

Method's source code location	_____ (e.g. URL)
Is the software	◯ Library ◯ Standalone software ◯ Other ◯ Not specified
Define the other software type	_____
Is the software used to apply the method free?	◯ Yes ◯ No ◯ Not specified
Software licence	_____
Was the method	☐ Originally desinged for longitudinal data ☐ Altered/modified for longitudinal data ☐ Implemented to longitudinal data without any modifications
How the method model/approaches longitudinal data	_____

Used system and complexity (requirements)

Does the article mention anything about the used system or its requirements?	◯ Yes ◯ No
Operating system	_____
Other system requirements or used system information	_____
Method's running time in terms of the input size (Big O notation) if reported	_____

Input (original) and output (synthetic) data properties

Is the method capable of handling unbalanced longitudinal data?	◯ Yes (number of time points / timing / spacing for intervals is different for different subjects) ◯ Yes (some variables are collected less often than others, but still for everyone at the same time point) ◯ No ◯ Not specified

Is the method capable of generating unbalanced longitudinal data?	◯ Yes (number of time points / timing / spacing for intervals is different for different subjects) ◯ Yes (some variables are collected less often than others, but still for everyone at the same time point) ◯ No ◯ Not specified
The method is capable of	☐ Handling categorical original data ☐ Handling numerical original data ☐ Generating categorical synthetic data ☐ Generating numerical synthetic data (numerical = continuous / interval, categorical = binary / multiclass)
The numerical data values generated	☐ Will not necessarily fall within the corresponding range in the original data set ☐ Will fall within the corresponding range in the original data set ☐ Will be replicates of values in the original data set ☐ Not specified
Is the method capable of handling missing values in the original data?	◯ Yes ◯ No ◯ Not specified
Is the method capable of producing missing values for synthetic data?	◯ Yes ◯ No ◯ Not specified

C.3 Method evaluation

Method performance evaluation

Please complete the survey below.

Data used to generate synthetic data, i.e., original or input data	
Synthetic data was generated based on	☐ Real-world data ☐ Simulated data ☐ Synthetic data ☐ Other (i.e., what type/form was the original data?)
Give the name of the data set(s)	_____ (Separate the names with a comma)
What kind of data was used? Separate different data sets with a comma.	_____ (e.g., patient data, other data related to people, non-human data)
Is the used data set(s) available?	☐ Publicly ☐ Upon request ☐ No
Give the source(s) of the available data set(s)	_____ (If multiple, in same order as given the data sets above)
The number of independent observations (subjects) in the input data set(s). If multiple data sets, separate with a comma.	_____ (If not reported or recoverable, write 'Not specified')
The number of variables in the input data set, including variables with repeated measurements. If multiple data sets, separate with a comma.	_____ (If not reported or recoverable, write 'Not specified')
Number of variables with repeated measurements (subset of the total number of variables).	_____
Number of repeated measurements. For unbalanced data or varying ranges, give the range [min, max]. If multiple data sets are used, separate them with a comma.	_____
Number of categorical variables. If multiple data sets are used, separate them with a comma.	_____ (Options: 0,1,,2,..., not specified)
Number of numerical variables. If multiple data sets are used, separate them with a comma.	_____ (Options: 0,1,,2,..., not specified)

Is the pattern of missingness similar to the original data?	◯ Yes ◯ No ◯ Not specified
Does the method have other limitations or requirements for the original data that have not already been mentioned?	◯ Yes ◯ No (e.g. input data have to be scaled)
Describe the requirements / limitations regarding to original data	_____

Evaluation setup of the generated synthetic data

Was the variable(s) in the synthetic data with repeated measurement treated as	◯ Response ◯ Explanatory ◯ Both ◯ Not specified (Option "Not specified" should be used only if the article is otherwise relevant and the nature of the variable cannot be determined even through discussion.)
The evaluation of the generated synthetic data was based on	☐ Qualitative assessment ☐ Quantitative assessment ☐ Other (Select all suitable options)
Describe the other approach used to evaluate the synthetic data and/or the method	_____
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 (i.e., multiple data sets)(< 50) ☐ A large amount of repetitions (>= 50) (Select all suitable options)
Was any of the following used to describe or evaluate the generated synthetic data and/or the method	☐ Descriptive statistics ☐ Statistical inference ☐ Prediction/classification (synthetic vs. real) ☐ Prediction/classification (some other variable) ☐ Privacy ☐ Externally assessed realism (Select all suitable options)
Was the generated synthetic data evaluated	☐ Against (resamples) 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)

Name the other methods used in the comparison	_____
Describe the other approach to used to evaluate the generated synthetic data (in terms of data)	_____
Describe the simulation approach.	_____
Was the training process(es) described and/or available in the source code?	◯ Yes ◯ Partially ◯ No
What was lacking from the training process description if it was only partially described?	_____

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

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

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

Specify all descriptive statistics (e.g., measures, estimates) 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 differential privacy used to enhance/secure the privacy of the generated synthetic data?	◯ Yes ◯ No (e.g., as a part of the method / applied post-hoc)
What was the epsilon used? If multiple epsilons were used, separate them with a comma	_____
Specify delta if applicable. If value not specified, write not specified.	_____

Was any of the following used to test the privacy of the synthetic data?	☐ Membership attack / identity disclosure ☐ Attribute disclosure ☐ Inferential disclosure ☐ Other ☐ No other approaches were used
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 (other than DP asked previously) if reported or write a summary of the authors' discussion on the subject if no specific approach was used.	_____

Externally assessed realism

How was the realism assessed externally?	_____

Other limitations or requirements for the generated synthetic data

Does the method have other limitations or requirements for synthetic data that have not already been mentioned?	◯ Yes ◯ No
Describe the requirements / limitations regarding to synthetic data	_____

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	_____
Write down the disadvantages of the method according to you	_____

General remarks on the method and the synthetic data

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

C.4 Assessment of bias and reporting quality

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? 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	◯ Yes ◯ No ◯ Possibly (The option "Possibly" can be used in a situation where there is no clear evidence of bias, but there is something to point out about the subject.)
Describe the (possible) selection bias present	_____

Performance bias	
Does the study show evidence of performance bias? 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.	◯ Yes ◯ No ◯ Possibly (The option "Possibly" can be used in a situation where there is no clear evidence of bias, but there is something to point out about the subject.)
Describe the (possible) performance bias present	_____
In how many comparisons out of all reported comparisons did the method perform worse than another comparison method.	_____ (Give a fraction worse/total or write 0/1 if the method performed best/worst in every comparison)
List the situations in which the method performed worse than the other methods or write all and give the amount of comparisons reported, if the method performed worst in every comparison.	_____

Reporting bias

Does the study show evidence of reporting bias?

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.

- ◯ Yes
- ◯ No
- ◯ Possibly

(The option "Possibly" can be used in a situation where there is no clear evidence of bias, but there is something to point out about the subject.)

Describe the (possible) 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

_____

Competing interests

Were competing interests reported?

- ◯ Yes
- ◯ No
- ◯ Not available

D Risk of bias and reporting quality assessment

D.1 Risk of bias assessment framework

Table 1. The framework used to assess the risk of bias. This table outlines different biases that may influence the evaluation of the methods' performance. The risk of these biases were assessed from the included publications. The table presents the fundamental principles (Rationale) that guided the assessment as well as the challenges involved in recognizing each type of bias (Assessment plausibility), along with illustrative examples of each type of bias.

Bias	Rationale	Assessment plausibility	Examples
Selection bias	Assessing the method's performance requires fairness in data representation, use of suitable metrics, and equal potential across methods to perform specific tasks. This necessitates clear justifications for input data, metrics, and reference method selection.	Detecting selection bias is difficult because any assessment approaches taken prior to the final publication may not be fully disclosed, making it difficult to assess favoritism towards the primary method. The reviewers may also be unaware of instances where a particular dataset did not work well with a particular method.	- Adjusting data or models based on arbitrary factors. - Using different datasets to evaluate different methods - Selectively using data or methods to favor the primary method.
Performance bias	To ensure a fair performance evaluation across methods, it is essential that a transparent and detailed description of the comparison and training procedures has been provided.	Detecting performance bias is challenging when the model selection and training details are incomplete or not reported. It becomes possible when the authors provide these details and mention using reference methods without task optimization.	- Not giving the reference methods a fair opportunity to perform well, e.g., through intentionally inadequate model training compared to the primary method.
Reporting bias	To ensure research transparency, it is important that all research evaluation metrics are comprehensively documented and the results are shared.	Detecting the bias should be straightforward when a publication or its supplementary material lacks or incompletely presents results for the evaluation approaches mentioned in the study.	- Results are either incomplete or missing

D.2 Risk of bias in individual studies (detailed explanations)

Table 2. Detailed explanations of the identified risk of bias present within each study.

Authors	Performance bias	Explanation	Reporting bias	Explanation
Li et al. [6]	Possibly	The method was compared to other methods but the training processes were not described.	Yes	Certain outcomes were exclusively or incompletely reported across methods and/or datasets. For instance, not all outcomes of t-tests were fully given, and patient trajectories were displayed only for the primary method and using only the MIMIC-III data.
Bhanot et al. [7]	Possibly	The method was compared to other methods but the training processes were not described.		
Yu, He & Raghunathan [8]	Possibly	The method was compared to other methods but the training processes were not described.	Yes	Certain findings, such as those shown in Table 2, pertained only to the primary method. Furthermore, the outcomes pertaining to IVEWare were excluded from the tabulated results of Tables 3 and 4. These specific outcomes were also omitted from the supplemental materials.
Zhang, Yan & Malin [9]	Possibly	The method was compared to other methods but the training processes were not described.	Yes	The primary method “Baseline + CFR + RS” was omitted from Figure 5 illustrating the drift in time.
Zhang et al. [10]			Yes	The authors asserted in their work (page 602, top of the second column) that statistical insignificance of FPR and TPR was observed. However, we could not find information about the specific statistical test they used in this context.

Biswal et al. [11]	Possibly	The method was compared to other methods but the training processes were not described.	Yes	In Figure 2, the VAE-Deconv component is absent. Within Figure 3, the depiction of outcomes is partial across various methods, and the rationale for excluding specific subfigures has not been presented. The evaluation of privacy remains either unaddressed or, at minimum, the outcomes pertaining to the alternative comparative methods and EVA_c are absent from the presentation.
Gootjes-Dreesbach et al. [12]			Yes	Comparative analyses between the actual patients and virtual patients were only shown for the PPMI dataset. In Figure 6, the depiction of decoded real patients was missing from the subset pertaining to SP513.
Sood et al. [13]			Yes	Comparisons between synthetic and original variables were selectively delineated for a subset of the variables under consideration.
Fisher et al. [14]			Yes	The authors had decided to confine the outcome section to a subset of data characterized as partially synthetic. Notably, some of the evaluation techniques could have been suitably extended to encompass fully synthetic data. The rationale behind this decision remains unclear.
Raab, Nowok & Dibben[15]			Yes	Analyses concerning the marginal distributions and the preservation of temporal correlations of discrete variables were not presented.

D.3 Individual study reporting quality assessment (detailed explanations)

Table 3. Detailed explanations of the identified reporting quality deviations within each study. *Inconsistency of reporting* refers to utilization of 1) identical terminology or notations to signify distinct phenomena and lacking clarification (e.g., using "noise" for both original data variation and additional privacy mechanism-induced noise without clear differentiation) or 2) disparate notations to represent the same phenomenon, both 1 and 2 introduce a potential risk of misunderstanding. *Imprecision of reporting* refers to the lack of precision (e.g. p-values reported with varying accuracies) or clarity in the presentation of information, which may lead to ambiguity or difficulty in understanding the reported data. *Indirectness of reporting* involves conveying information in a manner that is not straightforward or explicit, albeit to a lesser extent than observed in reporting bias, potentially requiring the reader to infer or deduce certain details. This can introduce a level of uncertainty or make the interpretation less direct.

Authors	Inconsistency	Explanation	Imprecision	Explanation	Indirectness	Explanation
Li et al. [6]			Yes	Statistically significant p-values are not reported as precisely as values above 0.05.	Yes	The meaning of mean and standard deviation for a discrete-valued feature is unclear (page 13, section 4.3). Figure 5 states that the y-axis represents the probability distribution of Mechanical Ventilation and Vasopressor being applied ("On"). It's unclear how the y-axis can exceed the range of [0,1]. In reference to differential privacy, it is stated that delta ≤ 0.001 (p. 21), but it is unclear what the exact delta was in each situation, e.g. if the delta remains constant for all values of epsilon presented in Figure 7b.
Bhanot et al. [7]			Yes	The number of patients was not reported precisely (“The data set has over 30 K records”, page 2).		
Zhang et al. [10]					Yes	The data description in Table 1 shows the gender distribution, but the article lacks clarity on whether this variable was utilized in data synthesis or analyses.

Biswal et al. [11]			Yes	The number of clinicians used to evaluate the realism score was not reported. Information on the minimum and maximum number of visits per patient and the minimum number of codes per visit was not given.	Yes	Full details about the presence disclosure test were not provided. Abbreviations like ELBO were unspecified. The nature of preliminary evaluations mentioned in the appendix remained unclear.
Gootjes-Dreesbach et al. [12]	Yes	Utilized three distinct notations for the differential privacy budget parameter.	Yes	Subfigure 9.1 did not specify the epsilon used in that figure.		
Sood et al. [13]			Yes	The number and types of variables employed in the actual synthesis of data remained unclear.		
Wendland et al. [16]			Yes	The p-value on page 3, right column, first paragraph, is reported with different precision compared to the subsequent p-values (which have two significant figures).		

E Study selection: excluded publications

The primary reason for exclusion was wrong data type (n = 165), mostly cross-sectional [17–20], survival [21–24] or time-series data [25–27]. Publications compromising the temporal structure in longitudinal data were categorized as having wrong data type [28–30]. Publications lacking SDG (n = 81) were typically introductions of a specific synthetic data framework [31–34] or data simulations [35–37]. Exclusions due to partially synthetic data (n = 49) were largely related to data augmentation using techniques such as Synthetic Minority Over-Sampling Technique (SMOTE) [38] or its variants [39–43].

We excluded 29 publications as we could not determine their eligibility, stemming from incomplete data, incomplete method description, or restricted access to the cited references, data, or algorithms [44–47]. Additionally, 28 studies were excluded for relying solely on standard probability distributions to simulate data [48–51]. Furthermore, 14 studies were excluded for failing to acknowledge the longitudinal nature of data [52–57], although the original datasets included variables with repeated measurements. Lastly, we identified three duplicates and two publications of wrong literature type (thesis or an extended abstract).

F Primary methods

F.1 Generative adversarial networks

Generative Adversarial Networks (GANs) [58] are a class of deep learning (DL) models of two neural networks. The generator network is trained to create synthetic data while the discriminator network learns to distinguish between real and generated data. The two networks are trained in a competitive setting, where the generator aims to produce increasingly realistic samples and the discriminator strives to improve its ability to differentiate between real and fake data.

AC-GAN

AC-GAN [59] (auxiliary classifier GAN) generates continuous synthetic data that includes a stratifying variable, e.g., a treatment group. Notably, AC-GAN offers options for both differentially private and non-private training approaches. The method models temporal relationships through convolutional layers [60] and by assuming that variables in the input dataset are ordered by time. Given that the objective is to concurrently generate realistic synthetic data while maintaining the inherent data stratification, its applicability in producing more generic longitudinal patterns is difficult to determine.

EHR-M-GAN

EHR-M-GAN [6] first maps variables into a shared latent space of reduced dimension using a dual variational autoencoder [61]. The method then generates correlated patient trajectories of different variable types through a coupled recurrent network that specifically focuses on learning temporal dependencies in the data. As EHR-M-GAN requires filtering outliers from the input data, it is not clear how well the method performs under data with long-tailed distributions.

HealthGAN

HealthGAN [62], applied in Bhanot et al. [7], implements a Wasserstein GAN gradient penalty (WGAN-GP) [63] and data transformation to generate mixed-type data. The transformation

involves scaling all variables to a unit range and reversing them back to their original scales after synthesis. HealthGAN, not initially developed for longitudinal data, relies on its ability to learn the multivariate distribution underlying the input data to capture temporal correlations. It may face challenges in learning and generating subpopulations.

Health Gym GAN

Health Gym GAN [64] generates mixed-type data and utilizes WGAN-GP and a bi-directional long short-term memory (biLSTM) network [65–67] to model dependencies in both temporal directions. To model multiple correlated categorical variables, Health Gym GAN requires fine-tuning.

MTGAN

Multi-label time series GAN (MTGAN) [68] generates patient-level illness sequences (diagnosis code indicator vectors). MTGAN utilizes a gated recurrent unit (GRU) generator [69] to recursively generate diagnosis probabilities and applies a conditional transition matrix to better address rare diagnoses. GRU also models temporal correlations between visits and diagnoses via latent variables and probabilities from previous iterations. The current MTGAN version is restricted to categorical variables and cannot generate continuous variables.

F.2 Autoencoders

Autoencoders (AEs) [70] are a type of neural network architecture that consists of an encoder and a decoder network, collectively trained to learn an efficient data representation that captures the most salient features of the input data. The encoder maps input data to a lower-dimensional latent space, while the decoder reconstructs the original input from the latent space. The goal of an autoencoder is to minimize the reconstruction error.

Variational autoencoders (VAEs) [61] differ from AEs by employing probabilistic encodings that capture uncertainty through probability distributions over latent variables. This approach offers greater flexibility in handling mixed-type data and enables VAEs to generate new samples by sampling from the latent space and decoding to the data domain.

EVA

EHR Variational Encoder (EVA) [11] generates patient-level visit sequences (indicator vectors of diagnosis codes, medications, and procedures) as autoregressive time-ordered transitions, with latent variables accounting for between-patient heterogeneity across the sequences. EVA models the temporal structure by incrementally expanding the latent space's spatial dimensions (deconvolution). While EVA can generate unbalanced data, it does not model the actual time between the visits. In addition, EVA's performance may be suboptimal when dealing with less frequent sequences in input data.

F.3 Bayesian Networks

Bayesian Networks (BNs) [71] are probabilistic modeling techniques that capture relationships between variables using a directed acyclic graph (DAG). The graph's nodes represent random variables while the edges indicate between-node dependencies. Each node is associated with a conditional probability distribution that describes the probability of the variable given its parental nodes.

MBN

A Modular Bayesian Network (MBN) [13] generates Gaussian and categorical synthetic data by learning conditional probabilities between predefined modules of semantically similar variables. Learning the network structure is improved by enforcing edge constraints, such as the correct temporal order of the nodes, and by reducing the module dimensionality via sparse autoencoders. In the case of non-Gaussian variables, MBN performs better when these variables are discretized, but this process also reduces data resemblance. Moreover, defining the modules and constraints requires expert knowledge.

VAMBN

A Variational Autoencoder Modular Bayesian Network (VAMBN) [12] expands on MBN by introducing a variational autoencoder (HI-VAE) [72] that considers data heterogeneity and missingness within modules. Temporal ordering is maintained by preventing edges from pointing backward in time for variables with repeated measurements. Similarly to MBN, VAMBN requires expert knowledge. In addition, the current implementation does not allow Gaussian nodes to have discrete-node children and necessitates a modern parallel computing architecture.

GMB model

Wang et al. [73] used a Generative Markov-Bayesian-based (GMB) approach to generate disease progression sequences (diagnosis codes). The method is a hierarchical model, with three layers: disease progression is modelled as a continuous-time Markov jump process [74], possible complications as conditionally independent Markov processes [74], and the presence of comorbidities is inferred through a bipartite noisy-or Bayesian Network [75,76]. GMB transforms unbalanced discrete-time input data into continuous-time illness sequences. For improved computational efficiency, expert knowledge is needed to establish prior probabilities that link complications and observed comorbidities.

F.4 Ensembles

Ensemble methods are machine learning techniques that combine multiple individual models [77]. The underlying idea is that by aggregating predictions or decisions from multiple models, the overall performance is improved over a single model. Common ensemble methods include bagging, boosting, and stacking [77]. Bagging involves training multiple models independently on different subsets of the training data and averaging their predictions. Boosting focuses on sequential model training, where each subsequent model tries to correct mistakes made by the previous models. Stacking combines predictions from multiple models using another model, called a meta-learner.

LS-EHR

The Longitudinal Simulation framework for EHR (LS-EHR) [9] combines GAN and recurrent neural network (RNN) with condition fuzzing and regularization (CFR) [9] to generate patient-level visit sequences (indicator vectors of diagnosis and procedure codes). To further improve data quality, LS-EHR incorporates Gaussian noise to add variability to synthetic observations and uses rejection sampling to improve data resemblance. CFR enables learning from both previous and subsequent episodes, mitigating gradual synthetic sequence divergence (drift) from the real sequence. While the LS-EHR was developed to address drifting, the problem was not fully resolved. Additionally, the performance of LS-EHR on datasets with high sparsity or a mix of categorical and continuous variables remains uncertain.

MultiNODEs

The Multimodal Neural Ordinary Differential Equations (MultiNODEs) [16] uses latent NODEs [78] to generate continuous repeated measurements, HI-VAE [72] to generate static variables (both categorical and numerical) and an imputation layer to replace any missing values present in the input data. The method is currently limited to generating continuous repeated measurements and its optimal performance depends on tuning several sensitive hyperparameters.

SynTEG

The Synthetic Temporal EHR Generator (SynTEG) [10] utilizes a self-attention architecture of transformer encoders [79] and a recurrent model to generate patient-level visit sequences (diagnosis code indicator vectors) conditionally on the previous visits. Subsequently, GAN is used to capture the multivariate distribution and to generate the sequences. SynTEG is limited to generate only diagnosis codes and it is possible that the method generates sequences conflicting with medical knowledge.

F.5 Other

CRBM

Fisher et al. [14] used a Conditional Restricted Boltzmann Machine (CRBM) to generate mixed-type disease progression data. CRBM is a probabilistic graphical model that incorporates latent variables and conditional distributions. The temporal dependence structure was learned by training the model with all possible pairs of two consecutive observations. As such, CRBM can generate both static and time-varying variables. However, the method requires balanced, numerically formatted data.

SCM

Barrientos et al. [80] used Sequential Conditional Modeling (SCM) to generate synthetic career data. Specifically, they modelled each input variable based on its type, utilizing techniques like classification and regression trees (CARTs) [81] and parametric probability distributions. Data were generated sequentially, variable-by-variable, and the future values of any time-varying variables were assumed to depend on the past only through the variables' current values. This method resembles traditional simulation and relies on expert knowledge to determine the approach and sequence for modeling each variable.

SPMI

Yu, He and Raghunathan [8] used Semiparametric Multiple Imputation (SPMI) to generate synthetic mixed-type survey data. Missing observations were first imputed using a Sequential Regression Multiple Imputation (SRMI) [82] framework. Subsequently, a Bayesian bootstrap sample [83] was extracted from these data and Alternating Conditional Expectation (ACE) [84] and a Ridge-Penalized Logistic (RPL) [85] imputation models were used to generate synthetic observations of continuous and discrete variables, respectively. Temporal dependencies were assumed to be learned by the imputation models as part of the overall correlation structure. SPMI is designed for datasets with around a hundred variables and may not be suitable for significantly larger or smaller datasets. Additionally, the method's generalizability beyond specific types of survey data, such as EHR or census data, is uncertain.

Synthea

Synthea [86] generates synthetic EHR data using modules and state-transition machines to model patient data. The modules are built based on Publicly Available Data Approach to the Realistic Synthetic EHR (PADARSER) framework [33] utilizing publicly available data and predefined healthcare trajectory templates (care maps). Users can build their own disease models using a dedicated module builder, but this requires expert knowledge of the disease. Synthea's module-based approach may not fully capture real-world complexity, and it primarily generates snapshots of patients at specific times, lacking long-term health data representation.

Synthpop

Raab, Nowok and Dibben [15] generated mixed-type data with Synthpop [87]. This R-package enables the use of several different parametric and non-parametric methods for generating synthetic mixed-type data by drawing each variable sequentially from its conditional distribution given the already synthesized variables. The authors applied both non-parametric (CART) and parametric (polychotomous, logistic, and linear regression) models to estimate these conditional distributions. Temporal modeling is based on the models' abilities to learn the general correlation structure. Applying methods provided by Synthpop requires expert knowledge akin to SCM. In addition, the parametric methods may oversimplify the underlying distributions and structure in the input data and thus may not work with complex datasets.

G Reference methods

Table 4. Reference methods used to benchmark the primary method.

Study	Primary method	Reference methods
Li at al. [6]	EHR-M-GAN	C-RNN-GAN [88] R(C)GAN [89] TimeGAN [90] medGAN [91] seqGAN[92] SynTEG [10] (included) DualAEE [93] PrivBayes [20]
Lu et al. [68]	MTGAN	medGAN [91] CTGAN [94] EMR-WGAN [95] RDP-CGAN [96] WGAN-GP [63] TimeGAN [90] T-CGAN [97]
Biswal et al. [11]	EVA	EVA_c biLSTM [65] VAE-LSTM [98] VAE-Deconv [99]
Wendland et al. [16]	MultiNODEs	VAMBN [12] (included)
Yu, He & Raghunathan [8]	SPMI	IVEware Version 0.3 [100] Synthpop [87] (included)

H Datasets used in the included publications

Table 5. Details regarding the datasets utilized within the studies.

Datasets	Data type	Availability	Study	Subjects	Numerical variables	Categorical variables	Variables with repeated measurements	Repeated measurements
MIMIC-III	Clinical database	Public	[59]	8 260	9	1	9	5
			[6]	28 344	78	20	98	24
			[68]	7 493	0	4 880	4 880	avg. 2.6
			[64]	3 910	9	13	20	48
			[64]	2 164	35	11	42	2–20
PPMI	Patient data	Public	[13]	362	NA	NA	38	2–12
			[16]	354	53	15	25	5–12
			[12]	557	NA	NA	38*	5
VUMC	EHR data	No	[9]	59 617	0	1 276	1 276	25–200
	Synthetic derivate	No	[10]	2 187 629	0	1 799	1 799	avg.12.1
ADNI	Patient data	Public	[13]	689	NA	NA	18	4
All of Us	EHR data	Public	[9]	59 617	0	526	526	10–200
ASD	Health data	No	[7]	> 280 000	7	2	7	10
ATUS	Behavioral data	Public	[7]	> 30 000	1	4	1	30
CDC	EHR data	No	[73]	9 298	1	100	88	2–11
CODR-AD	Clinical database	No	[14]	1 909	38	6	36	7
eICU	Clinical database	Public	[6]	99 015	55	19	74	24
HiRID	Clinical database	Public	[6]	14 129	50	39	89	24
HIV	EuResist integrated database	Public	[64]	8 916	3	12	13	10–100
HRS	Longitudinal survey	No	[8]	12 652	7	41	11	2–3
MIMIC-IV	Clinical database	Public	[68]	10 000	0	6 102	6 102	avg. 3.6
Multi-census	Census data	No	[86]	NA	NA	NA	NA	NA

NACC	Patient data	Public	[16]	2 284	4	3	3	4
PAMF EHR	EHR data	No	[11]	258 555	0	10 437	10 437	avg. 53.8
SP513	Clinical trial data	No	[12]	560	NA	NA	35*	2–11*
SPRINT	Clinical trial data	No	[59]	6 502	3	1	3	12
Status File	Employment data	No	[80]	3 511 824	5	24	22	24
UK LS	Admin-census data	No	[15]	> 186 000	1	4	5	2

NA: not available; avg.: average; *: calculated from presented materials by the corresponding author

References

[1] Cornell University, arXiv Dataset, (2022). https://www.kaggle.com/datasets/Cornell-University/arxiv (accessed November 22, 2022).

[2] R Core Team, R: A Language and Environment for Statistical Computing, (2021). https://doi.org/https://doi.org/10.59350/t79xt-tf203.

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

[4] P.A. Harris, R. Taylor, B.L. Minor, V. Elliott, M. Fernandez, L. O'Neal, L. McLeod, G. Delacqua, F. Delacqua, J. Kirby, S.N. Duda, The REDCap consortium: Building an international community of software platform partners, J Biomed Inform 95 (2019) 103208. https://doi.org/10.1016/j.jbi.2019.103208.

[5] P.A. Harris, R. Taylor, R. Thielke, J. Payne, N. Gonzalez, J.G. Conde, Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support, J Biomed Inform 42 (2009) 377–381. https://doi.org/10.1016/j.jbi.2008.08.010.

[6] J. Li, B.J. Cairns, J. Li, T. Zhu, Generating Synthetic Mixed-type Longitudinal Electronic Health Records for Artificial Intelligent Applications, ArXiv Preprint (2023). https://arxiv.org/abs/2112.12047v2.

[7] K. Bhanot, J. Pedersen, I. Guyon, K.P. Bennett, Investigating synthetic medical time-series resemblance, Neurocomputing 494 (2022) 368–378. https://doi.org/10.1016/j.neucom.2022.04.097.

[8] M. Yu, Y. He, T.E. Raghunathan, A Semiparametric Multiple Imputation Approach to Fully Synthetic Data for Complex Surveys, J Surv Stat Methodol 10 (2022) 618–641. https://doi.org/10.1093/jssam/smac016.

[9] Z. Zhang, C. Yan, B.A. Malin, Keeping synthetic patients on track: feedback mechanisms to mitigate performance drift in longitudinal health data simulation, J Am Med Inform Assoc 29 (2022) 1890–1898. https://doi.org/10.1093/jamia/ocac131.

[10] Z. Zhang, C. Yan, T.A. Lasko, J. Sun, B.A. Malin, SynTEG: a framework for temporal structured electronic health data simulation, Journal of the American Medical Informatics Association 28 (2021) 596–604. https://doi.org/10.1093/jamia/ocaa262.

[11] S. Biswal, S. Ghosh, J. Duke, B. Malin, W. Stewart, J. Sun, EVA: Generating Longitudinal Electronic Health Records Using Conditional Variational Autoencoders, in: K. Jung, S. Yeung, M. Sendak, M. Sjoding, R. Ranganath (Eds.), Proceedings of the 6th Machine Learning for Healthcare Conference, PMLR, 2021: pp. 260–282. https://proceedings.mlr.press/v149/biswal21a.html.

[12] L. Gootjes-Dreesbach, M. Sood, A. Sahay, M. Hofmann-Apitius, H. Fröhlich, Variational Autoencoder Modular Bayesian Networks for Simulation of Heterogeneous Clinical Study Data, Front Big Data 3 (2020).

[13] M. Sood, A. Sahay, R. Karki, M.A. Emon, H. Vrooman, M. Hofmann-Apitius, H. Fröhlich, Realistic simulation of virtual multi-scale, multi-modal patient trajectories using Bayesian networks and sparse auto-encoders, Sci Rep 10 (2020) 10971. https://doi.org/10.1038/s41598-020-67398-4.

[14] C.K. Fisher, A.M. Smith, J.R. Walsh, A.J. Simon, C. Edgar, C.R. Jack, D. Holtzman, D. Russell, D. Hill, D. Grosset, F. Wood, H. Vanderstichele, J. Morris, K. Blennow, K. Marek, L.M. Shaw, M. Albert, M. Weiner, N. Fox, P. Aisen, P.E. Cole, R. Petersen, T. Sherer, W. Kubick, Machine learning for comprehensive forecasting of Alzheimer's Disease progression, Sci Rep 9 (2019) 13622. https://doi.org/10.1038/s41598-019-49656-2.

[15] G.M. Raab, B. Nowok, C. Dibben, Practical Data Synthesis for Large Samples, Journal of Privacy and Confidentiality 7 (2018) 67–97. https://doi.org/10.29012/jpc.v7i3.407.

[16] P. Wendland, C. Birkenbihl, M. Gomez-Freixa, M. Sood, M. Kschischo, H. Fröhlich, Generation of realistic synthetic data using Multimodal Neural Ordinary Differential Equations, NPJ Digit Med 5 (2022) 122. https://doi.org/10.1038/s41746-022-00666-x.

[17] N.C. Abay, Y. Zhou, M. Kantarcioglu, B. Thuraisingham, L. Sweeney, Privacy Preserving Synthetic Data Release Using Deep Learning, in: Lecture Notes in Computer Science, Springer Verlag, 2019: pp. 510–526. https://doi.org/10.1007/978-3-030-10925-7_31.

[18] Y. Park, J. Ghosh, M. Shankar, Perturbed Gibbs Samplers for Generating Large-Scale Privacy-Safe Synthetic Health Data, in: 2013 IEEE International Conference on Healthcare Informatics, IEEE, 2013: pp. 493–498. https://doi.org/10.1109/ICHI.2013.76.

[19] M. Walia, B. Tierney, S. Mckeever, Synthesising Tabular Data using Wasserstein Conditional GANs with Gradient Penalty (WCGAN-GP), in: L. Longo, L. Rizzo, E. Hunter, A. Pakrashi (Eds.), Proceedings of The 28th Irish Conference on Artificial Intelligence and Cognitive Science, Technological University Dublin, Dublin, 2020: pp. 325–336. https://doi.org/https://doi.org/10.21427/E6WA-SZ92.

[20] J. Zhang, G. Cormode, C.M. Procopiuc, D. Srivastava, X. Xiao, PrivBayes, ACM Transactions on Database Systems 42 (2017) 1–41. https://doi.org/10.1145/3134428.

[21] J. Yoon, L.N. Drumright, M. van der Schaar, Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN), IEEE J Biomed Health Inform 24 (2020) 2378–2388. https://doi.org/10.1109/JBHI.2020.2980262.

[22] F. Bonofiglio, M. Schumacher, H. Binder, Recovery of original individual person data (IPD) inferences from empirical IPD summaries only: Applications to distributed computing under disclosure constraints, Stat Med 39 (2020) 1183–1198. https://doi.org/10.1002/sim.8470.

[23] K. El Emam, L. Mosquera, C. Zheng, Optimizing the synthesis of clinical trial data using sequential trees, Journal of the American Medical Informatics Association 28 (2021) 3–13. https://doi.org/10.1093/jamia/ocaa249.

[24] T. Khorchani, Y. Gadiya, G. Witt, D. Lanzillotta, C. Claussen, A. Zaliani, SASC: A simple approach to synthetic cohorts for generating longitudinal observational patient cohorts from COVID-19 clinical data, Patterns 3 (2022) 100453. https://doi.org/10.1016/j.patter.2022.100453.

[25] A. Torfi, E.A. Fox, CorGAN: Correlation-Capturing Convolutional Generative Adversarial Networks for Generating Synthetic Healthcare Records, The International FLAIRS Conference Proceedings 33 (2020).

[26] M. Hernandez, G. Epelde, A. Beristain, R. Álvarez, C. Molina, X. Larrea, A. Alberdi, M. Timoleon, P. Bamidis, E. Konstantinidis, Incorporation of Synthetic Data Generation Techniques within a Controlled Data Processing Workflow in the Health and Wellbeing Domain, Electronics (Basel) 11 (2022) 812. https://doi.org/10.3390/electronics11050812.

[27] L. Wang, W. Zhang, X. He, Continuous Patient-Centric Sequence Generation via Sequentially Coupled Adversarial Learning, in: Lecture Notes in Computer Science, 2019: pp. 36–52. https://doi.org/10.1007/978-3-030-18579-4_3.

[28] M.K. Baowaly, C.C. Lin, C.L. Liu, K.T. Chen, Synthesizing electronic health records using improved generative adversarial networks, Journal of the American Medical Informatics Association 26 (2019) 228–241. https://doi.org/10.1093/jamia/ocy142.

[29] Y. Liu, J. Peng, J.J.Q. Yu, Y. Wu, PPGAN: Privacy-Preserving Generative Adversarial Network, in: 2019 IEEE 25th International Conference on Parallel and Distributed Systems (ICPADS), IEEE, 2019: pp. 985–989. https://doi.org/10.1109/ICPADS47876.2019.00150.

[30] S. Dash, R. Dutta, I. Guyon, A. Pavao, A. Yale, K.P. Bennett, Synthetic Event Time Series Health Data Generation, ArXiv Preprint (2019). http://arxiv.org/abs/1911.06411.

[31] M. Boedihardjo, T. Strohmer, R. Vershynin, Privacy of Synthetic Data: A Statistical Framework, IEEE Trans Inf Theory 69 (2023) 520–527. https://doi.org/10.1109/TIT.2022.3216793.

[32] J.S. Lombardo, L.J. Moniz, A method for generation and distribution of synthetic medical record data for evaluation of disease-monitoring systems, Johns Hopkins APL Technical Digest (Applied Physics Laboratory) 27 (2008).

[33] K. Dube, T. Gallagher, Approach and Method for Generating Realistic Synthetic Electronic Healthcare Records for Secondary Use, in: Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2014: pp. 69–86. https://doi.org/10.1007/978-3-642-53956-5_6.

[34] S. McLachlan, K. Dube, T. Gallagher, J.A. Simmonds, N. Fenton, Realistic Synthetic Data Generation: The ATEN Framework, in: Communications in Computer and Information Science, 2019: pp. 497–523. https://doi.org/10.1007/978-3-030-29196-9_25.

[35] S.D.P. Mendonca, Y.P.D.S. Brito, C.G.R. Dos Santos, R.D.A.D. Lima, T.D.O. De Araujo, B.S. Meiguins, Synthetic Datasets Generator for Testing Information Visualization and Machine Learning Techniques and Tools, IEEE Access 8 (2020) 82917–82928. https://doi.org/10.1109/ACCESS.2020.2991949.

[36] L.A. Garrow, T.D. Bodea, M. Lee, Generation of synthetic datasets for discrete choice analysis, Transportation (Amst) 37 (2010) 183–202. https://doi.org/10.1007/s11116-009-9228-6.

[37] J. Lobo, R. Henriques, S.C. Madeira, G-Tric: generating three-way synthetic datasets with triclustering solutions, BMC Bioinformatics 22 (2021) 16. https://doi.org/10.1186/s12859-020-03925-4.

[38] N. V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research 16 (2002) 321–357. https://doi.org/10.1613/jair.953.

[39] B. Tang, H. He, KernelADASYN: Kernel based adaptive synthetic data generation for imbalanced learning, in: 2015 IEEE Congress on Evolutionary Computation (CEC), IEEE, 2015: pp. 664–671. https://doi.org/10.1109/CEC.2015.7256954.

[40] S. Sharma, C. Bellinger, B. Krawczyk, O. Zaiane, N. Japkowicz, Synthetic Oversampling with the Majority Class: A New Perspective on Handling Extreme Imbalance, in: 2018 IEEE International Conference on Data Mining (ICDM), IEEE, 2018: pp. 447–456. https://doi.org/10.1109/ICDM.2018.00060.

[41] J.M. Martínez-García, C.P. Suárez-Araujo, P.G. Báez, SNEOM: A Sanger Network Based Extended Over-Sampling Method. Application to Imbalanced Biomedical Datasets, in: Lecture Notes in Computer Science, 2012: pp. 584–592. https://doi.org/10.1007/978-3-642-34478-7_71.

[42] Z. Wan, Y. Zhang, H. He, Variational autoencoder based synthetic data generation for imbalanced learning, in: 2017 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 2017: pp. 1–7. https://doi.org/10.1109/SSCI.2017.8285168.

[43] M. Peréz-Ortiz, P. Tiňo, R. Mantiuk, C. Hervás-Martínez, Exploiting Synthetically Generated Data with Semi-Supervised Learning for Small and Imbalanced Datasets, Proceedings of the AAAI Conference on Artificial Intelligence 33 (2019) 4715–4722. https://doi.org/10.1609/aaai.v33i01.33014715.

[44] M. Zare, J. Wojtusiak, Weighted Itemsets Error (WIE) Approach for Evaluating Generated Synthetic Patient Data, in: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2018: pp. 1017–1022. https://doi.org/10.1109/ICMLA.2018.00166.

[45] P. Stolfi, I. Valentini, M.C. Palumbo, P. Tieri, A. Grignolio, F. Castiglione, Potential predictors of type-2 diabetes risk: machine learning, synthetic data and wearable health devices, BMC Bioinformatics 21 (2020) 508. https://doi.org/10.1186/s12859-020-03763-4.

[46] A. Goncalves, P. Ray, B. Soper, J. Stevens, L. Coyle, A.P. Sales, Generation and evaluation of synthetic patient data, BMC Med Res Methodol 20 (2020) 108. https://doi.org/10.1186/s12874-020-00977-1.

[47] R. Indhumathi, S.S. Devi, Healthcare Cramér Generative Adversarial Network (HCGAN), Distrib Parallel Databases 40 (2022) 657–673. https://doi.org/10.1007/s10619-021-07346-x.

[48] S. Helfer, M. Kümmel, F. Bathelt, M. Sedlmayr, Generating Enriched Synthetic German Hospital Claims Data – A Use Case Driven Approach, in: German Medical Data Sciences: Bringing Data to Life, 2021: pp. 58–65. https://doi.org/10.3233/SHTI210051.

[49] A. Oganian, J. Domingo-Ferrer, Local synthesis for disclosure limitation that satisfies probabilistic k-anonymity criterion, Trans Data Priv 10 (2017) 61–81. https://doi.org/10.5555/3121409.3121412.

[50] M. Klein, R. Moura, B. Sinha, Multivariate Normal Inference based on Singly Imputed Synthetic Data under Plug-in Sampling, Sankhya B 83 (2021) 273–287. https://doi.org/10.1007/s13571-019-00215-9.

[51] H. Demirtas, Y. Yavuz, Concurrent Generation of Ordinal and Normal Data, J Biopharm Stat 25 (2015) 635–650. https://doi.org/10.1080/10543406.2014.920868.

[52] F.K. Dankar, M. Ibrahim, Fake It Till You Make It: Guidelines for Effective Synthetic Data Generation, Applied Sciences 11 (2021) 2158. https://doi.org/10.3390/app11052158.

[53] C.J. Andrews, M.S. Allacci, J. Senick, H.C. Putra, I. Tsoulou, Using synthetic population data for prospective modeling of occupant behavior during design, Energy Build 126 (2016) 415–423. https://doi.org/10.1016/j.enbuild.2016.05.049.

[54] J. Feldman, D.R. Kowal, Bayesian data synthesis and the utility-risk trade-off for mixed epidemiological data, Ann Appl Stat 16 (2022) 2577–2602.

[55] B. Li, S. Luo, X. Qin, L. Pan, Improving GAN with inverse cumulative distribution function for tabular data synthesis, Neurocomputing 456 (2021) 373–383. https://doi.org/10.1016/j.neucom.2021.05.098.

[56] M. Baak, S. Brugman, L. D'almeida, I.F. Rojas, J.-B. Oger, R.U. Ing, B. Ing, B. Ing Bank, Synthsonic: Fast, Probabilistic modeling and Synthesis of Tabular Data, 2022.

[57] F. Harder, K. Adamczewski, M. Park, DP-MERF: Differentially Private Mean Embeddings with Random Features for Practical Privacy-Preserving Data Generation, 2021.

[58] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Commun ACM 63 (2020) 139–144. https://doi.org/10.1145/3422622.

[59] B.K. Beaulieu-Jones, Z.S. Wu, C. Williams, R. Lee, S.P. Bhavnani, J.B. Byrd, C.S. Greene, Privacy-Preserving Generative Deep Neural Networks Support Clinical Data Sharing, Circ Cardiovasc Qual Outcomes 12 (2019) e005122. https://doi.org/10.1161/CIRCOUTCOMES.118.005122.

[60] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (1998) 2278–2324. https://doi.org/10.1109/5.726791.

[61] D.P. Kingma, M. Welling, Auto-encoding variational bayes, in: 2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings, 2014.

[62] A. Yale, S. Dash, R. Dutta, I. Guyon, A. Pavao, K.P. Bennett, Generation and evaluation of privacy preserving synthetic health data, Neurocomputing 416 (2020) 244–255. https://doi.org/10.1016/j.neucom.2019.12.136.

[63] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. Courville, Improved training of wasserstein GANs, in: Adv Neural Inf Process Syst, 2017.

[64] N.I.-H. Kuo, M.N. Polizzotto, S. Finfer, F. Garcia, A. Sönnerborg, M. Zazzi, M. Böhm, R. Kaiser, L. Jorm, S. Barbieri, The Health Gym: synthetic health-related datasets for the development of reinforcement learning algorithms, Sci Data 9 (2022) 693. https://doi.org/10.1038/s41597-022-01784-7.

[65] M. Schuster, K.K. Paliwal, Bidirectional recurrent neural networks, IEEE Transactions on Signal Processing 45 (1997) 2673–2681. https://doi.org/10.1109/78.650093.

[66] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Networks 18 (2005) 602–610. https://doi.org/10.1016/j.neunet.2005.06.042.

[67] S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, Neural Comput 9 (1997) 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

[68] C. Lu, C.K. Reddy, P. Wang, D. Nie, Y. Ning, Multi-Label Clinical Time-Series Generation via Conditional GAN, ArXiv Preprint (2022). http://arxiv.org/abs/2204.04797.

[69] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, ArXin Preprint (2014). http://arxiv.org/abs/1412.3555.

[70] G.E. Hinton, R.R. Salakhutdinov, Reducing the Dimensionality of Data with Neural Networks, Science (1979) 313 (2006) 504–507. https://doi.org/10.1126/science.1127647.

[71] D. Koller, N. Friedman, Probabilistic Graphical Models: Principles and Techniques, The MIT Press, 2009.

[72] A. Nazábal, P.M. Olmos, Z. Ghahramani, I. Valera, Handling incomplete heterogeneous data using VAEs, Pattern Recognit 107 (2020) 107501. https://doi.org/10.1016/j.patcog.2020.107501.

[73] X. Wang, Y. Lin, Y. Xiong, S. Zhang, Y. He, Y. He, Z. Zhang, J.M. Plasek, L. Zhou, D.W. Bates, C. Tang, Using an optimized generative model to infer the progression of complications in type 2 diabetes patients, BMC Med Inform Decis Mak 22 (2022) 174. https://doi.org/10.1186/s12911-022-01915-5.

[74] D.W. Stroock, An Introduction to Markov Processes, Springer Berlin Heidelberg, Berlin, Heidelberg, 2014.

[75] M.A. Shwe, B. Middleton, D.E. Heckerman, M. Henrion, E.J. Horvitz, H.P. Lehmann, G.F. Cooper, Probabilistic Diagnosis Using a Reformulation of the INTERNIST-1/QMR Knowledge Base, Methods Inf Med 30 (1991) 241–255. https://doi.org/10.1055/s-0038-1634846.

[76] Y. Halpern, D. Sontag, Unsupervised learning of noisy-or Bayesian networks, in: Uncertainty in Artificial Intelligence - Proceedings of the 29th Conference, UAI 2013, 2013.

[77] T.G. Dietterich, Ensemble Methods in Machine Learning, in: Lecture Notes in Computer Science, 2000: pp. 1–15. https://doi.org/10.1007/3-540-45014-9_1.

[78] R.T.Q. Chen, Y. Rubanova, J. Bettencourt, D.K. Duvenaud, Neural Ordinary Differential Equations, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Adv Neural Inf Process Syst, Curran Associates, Inc., 2018.

[79] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Adv Neural Inf Process Syst, 2017.

[80] A.F. Barrientos, A. Bolton, T. Balmat, J.P. Reiter, J.M. de Figueiredo, A. Machanavajjhala, Y. Chen, C. Kneifel, M. Delong, Providing access to confidential research data through synthesis and verification: An application to data on employees of the U.S. federal government, Annals of Applied Statistics 12 (2018) 1124–1156. https://doi.org/10.1214/18-AOAS1194.

[81] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification And Regression Trees, Routledge, 2017.

[82] T.E. Raghunathan, J.M. Lepkowski, J. Van Hoewyk, P. Solenberger, A multivariate technique for multiply imputing missing values using a sequence of regression models, Surv Methodol 27 (2001).

[83] D.B. Rubin, The Bayesian Bootstrap, The Annals of Statistics 9 (1981) 130–134. https://projecteuclid.org/journals/annals-of-statistics/volume-9/issue-1/The-Bayesian-Bootstrap/10.1214/aos/1176345338.full.

[84] B. Liu, M. Yu, B.I. Graubard, R.P. Troiano, N. Schenker, Multiple imputation of completely missing repeated measures data within person from a complex sample: application to accelerometer data in the National Health and Nutrition Examination Survey, Stat Med 35 (2016) 5170–5188. https://doi.org/10.1002/sim.7049.

[85] M. Yu, E.J. Feuer, K.A. Cronin, N.E. Caporaso, Use of Multiple Imputation to Correct for Bias in Lung Cancer Incidence Trends by Histologic Subtype, Cancer Epidemiology, Biomarkers & Prevention 23 (2014) 1546–1558. https://doi.org/10.1158/1055-9965.EPI-14-0130.

[86] J. Walonoski, M. Kramer, J. Nichols, A. Quina, C. Moesel, D. Hall, C. Duffett, K. Dube, T. Gallagher, S. McLachlan, 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 25 (2018) 230–238. https://doi.org/10.1093/jamia/ocx079.

[87] B. Nowok, G.M. Raab, C. Dibben, synthpop: Bespoke Creation of Synthetic Data in *R*, J Stat Softw 74 (2016) 1–26.

[88] O. Mogren, C-RNN-GAN: Continuous recurrent neural networks with adversarial training, (2016). http://arxiv.org/abs/1611.09904.

[89] C. Esteban, S.L. Hyland, G. Rätsch, Real-valued (Medical) Time Series Generation with Recurrent Conditional GANs, (2017). http://arxiv.org/abs/1706.02633.

[90] J. Yoon, D. Jarrett, M. Van Der Schaar, Time-series Generative Adversarial Networks, in: 33rd Conference on Neural Information Processing Systems, 2019.

[91] E. Choi, S. Biswal, B. Malin, J. Duke, W.F. Stewart, S. Org, J. Sun, Generating Multi-label Discrete Patient Records using Generative Adversarial Networks, in: Proceedings of the 2nd Machine Learning for Healthcare Conference, PMLR, Boston, Massachusetts, 2017: pp. 286–305.

[92] L. Yu, W. Zhang, J. Wang, Y. Yu, SeqGAN: Sequence generative adversarial nets with policy gradient, in: 31st AAAI Conference on Artificial Intelligence, AAAI 2017, 2017. https://doi.org/10.1609/aaai.v31i1.10804.

[93] D. Lee, H. Yu, X. Jiang, D. Rogith, M. Gudala, M. Tejani, Q. Zhang, L. Xiong, Generating sequential electronic health records using dual adversarial autoencoder, Journal of the American Medical Informatics Association 27 (2020). https://doi.org/10.1093/jamia/ocaa119.

[94] L. Xu, M. Skoularidou, A. Cuesta-Infante, K. Veeramachaneni, Modeling tabular data using conditional GAN, in: Adv Neural Inf Process Syst, 2019.

[95] Z. Zhang, C. Yan, D.A. Mesa, J. Sun, B.A. Malin, Ensuring electronic medical record simulation through better training, modeling, and evaluation, Journal of the American Medical Informatics Association 27 (2020) 99–108. https://doi.org/10.1093/jamia/ocz161.

[96] A. Torfi, E.A. Fox, C.K. Reddy, Differentially private synthetic medical data generation using convolutional GANs, Inf Sci (N Y) 586 (2022) 485–500. https://doi.org/10.1016/j.ins.2021.12.018.

[97] G. Ramponi, P. Protopapas, M. Brambilla, R. Janssen, T-CGAN: Conditional Generative Adversarial Network for Data Augmentation in Noisy Time Series with Irregular Sampling, ArXiv Preprint (2018). http://arxiv.org/abs/1811.08295.

[98] S.R. Bowman, L. Vilnis, O. Vinyals, A.M. Dai, R. Jozefowicz, S. Bengio, Generating sentences from a continuous space, in: CoNLL 2016 - 20th SIGNLL Conference on Computational Natural Language Learning, Proceedings, 2016. https://doi.org/10.18653/v1/k16-1002.

[99] S. Semeniuta, A. Severyn, E. Barth, A hybrid convolutional variational autoencoder for text generation, in: EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings, 2017. https://doi.org/10.18653/v1/d17-1066.

[100] T.E. Raghunathan, P.W. Solenberger, J. Van Hoewyk, IVEware: Imputation and Variance Estimation Software User Guide, (2002).