

6 Evaluating the Utility and Privacy of Synthetic Breast Cancer Clinical Trial Data Sets

Samer El Kababji, PhD, MEng, MSc¹ ; Nicholas Mitsakakis, PhD, MSc¹; Xi Fang, MSc² ; Ana-Alicia Beltran-Bless, MD^{3,4} ; Greg Pond, PhD⁵; Lisa Vandermeer, MSc³; Dhenuka Radhakrishnan, MD, MSc^{1,6}; Lucy Mosquera, MSc^{1,2} ; Alexander Paterson, MD⁷; Lois Shepherd, MD⁸ ; Bingshu Chen, PhD⁸ ; William E. Barlow, PhD⁹ ; Julie Gralow, MD¹⁰ ; Marie-France Savard, MD^{3,4} ; Mark Clemons, MD, MB^{3,4} ; and Khaled El Emam, PhD, BEng^{1,2,11} 

DOI <https://doi.org/10.1200/JCO.23.00116>

ABSTRACT

PURPOSE There is strong interest from patients, researchers, the pharmaceutical industry, medical journal editors, funders of research, and regulators in sharing clinical trial data for secondary analysis. However, data access remains a challenge because of concerns about patient privacy. It has been argued that synthetic data generation (SDG) is an effective way to address these privacy concerns. There is a dearth of evidence supporting this on oncology clinical trial data sets, and on the utility of privacy-preserving synthetic data. The objective of the proposed study is to validate the utility and privacy risks of synthetic clinical trial data sets across multiple SDG techniques.

METHODS We synthesized data sets from eight breast cancer clinical trial data sets using three types of generative models: sequential synthesis, conditional generative adversarial network, and variational autoencoder. Synthetic data utility was evaluated by replicating the published analyses on the synthetic data and assessing concordance of effect estimates and CIs between real and synthetic data. Privacy was evaluated by measuring attribution disclosure risk and membership disclosure risk.

RESULTS Utility was highest using the sequential synthesis method where all results were replicable and the CI overlap most similar or higher for seven of eight data sets. Both types of privacy risks were low across all three types of generative models.

DISCUSSION Synthetic data using sequential synthesis methods can act as a proxy for real clinical trial data sets, and simultaneously have low privacy risks. This type of generative model can be one way to enable broader sharing of clinical trial data.

ACCOMPANYING CONTENT

 [Data Supplement](#)

Accepted September 19, 2023

Published November 27, 2023

JCO Clin Cancer Inform

7:e2300116

© 2023 by American Society of
Clinical Oncology

Creative Commons Attribution
Non-Commercial No Derivatives
4.0 License

INTRODUCTION

The large amount of patient-centered information contained in clinical trial data sets is rarely fully utilized for secondary purposes such as testing new hypotheses through the use of statistical and machine learning (ML) models. The secondary analysis of data from previous clinical trials can provide new insights compared with the original publications,¹ and has produced informative research results on drug safety, evaluating bias, replication of studies, and meta-analysis.² Therefore, there has been strong interest in making more clinical trial data available for secondary analysis by journals, funders, the pharmaceutical industry, and regulators.³⁻¹¹

However, data access for secondary analysis remains a challenge,¹² sometimes taking many months to access data.¹³ Analyses of the success of getting individual-level data for meta-analysis projects from authors found that the

percentage of the time these efforts were successful ranged from 0% to 58%.¹⁴⁻¹⁹ Some researchers note that getting access to data sets from authors can take from 4 months to 4 years.¹⁸ Although there are many repositories that are potentially suitable for sharing individual-level data, only a handful do so for data from clinical studies.²⁰ Early experiences with some of these noted that the process is lengthy,²¹ and that it takes 6 months from proposal submission to data access.²²

A key reason why individual-level clinical trial data are not made readily available to data users is patient privacy. Privacy concerns by patients and regulators have historically acted as a barrier to the sharing of health data.^{23,24} Recent reports highlight the difficulties in accessing data for health research and ML analytics.²⁵⁻²⁷

Although patient (re)consent is one legal basis for making data available for secondary purposes, it is often impractical

CONTEXT

Key Objective

To compare different methods for generating synthetic clinical trial data sets as a means of enabling privacy-preserving sharing of data.

Knowledge Generated

Published analyses for the eight clinical trials were replicated on the synthetic data generated using sequential synthesis with decision trees, and the CI overlap for parameter estimates was high. The privacy risks in the form of attribution disclosure and membership disclosure were evaluated and deemed to be below the generally accepted thresholds.

Relevance

Generative artificial intelligence that is based on deep neural networks is receiving much attention. However, the authors demonstrate that other machine learning methods which are more readily understood, such as sequential decision trees, may yield superior results when applied to generating synthetic datasets.

to get retroactive consent under many circumstances, especially in studies of patients being treated for advanced cancer. In addition, there is significant evidence of consent bias.²⁸ Anonymization is another approach to making data available for secondary analysis. However, recently, there have been repeated claims of successful reidentification attacks on anonymized data,^{29–35} eroding the trust of the public and regulators in this approach.^{36–45}

In this paper, we evaluate multiple synthetic data generation (SDG) methods for addressing the privacy concerns in making clinical trial data sets available.⁴⁶ SDG has been applied quite often on health and social sciences data.^{47–55} To create synthetic data, a (typically) ML generative model is trained on the real individual-level data, capturing its patterns and statistical properties. Then new data are generated from that model. Because the generated data do not have a one-to-one mapping to the original data sets, synthetic data are believed to have low privacy risks.^{56–64} However, a recent systematic review noted that most studies proposing and applying SDG methods do not evaluate the privacy risks of synthetic data.⁶⁵

Furthermore, there is a known tradeoff between privacy and utility when privacy enhancing methodologies are applied.^{66,67} This means that the more a method protects privacy, there is also a decrease in the utility of the data. Thus, it is important to assess both the utility and privacy when evaluating data protection methods.

Thus far, there has been an evaluation of SDG on a simulated nononcology clinical trial data set without an evaluation of privacy risks,⁶⁸ an application of SDG to a real nononcology clinical trial data set,⁶⁹ one study that evaluated the privacy risks in synthetic oncology clinical trial data sets but did not simultaneously consider the impact on utility,⁷⁰ and one feasibility study demonstrating that a synthetic oncology clinical trial data set can have high utility in that it can

replicate the analysis conclusions using the real data but did not examine the privacy risks.⁷¹

We aim to fill this knowledge gap by performing a more comprehensive evaluation of both data utility and privacy of SDG on eight breast cancer clinical trial data sets across three common SDG techniques. This will allow us to determine the extent to which synthetic clinical trial data sets can serve as a proxy for real data sets, and enable the data to be made responsibly available more broadly to the research community for secondary analysis.

METHODS

Data Sets

Eight breast cancer clinical trials were included in the analysis as summarized in [Table 1](#). The first five clinical trials were supported by the Rethinking Clinical Trials Program at the Ottawa Hospital.⁸⁰ The remaining trials used data on recurrence risk in patients from three large international clinical studies based on both patient and cancer characteristics.^{77–79}

Methods for SDG

We used several ML-based SDG methods to synthesize the analytic data sets from the clinical trials. These may be classified under three broad synthesis techniques, namely sequential synthesis using decision trees, generative adversarial networks (GANs), and variational autoencoders (VAEs). For the latter two, we used the Synthcity comprehensive open-source generative model library.⁸¹

Sequential Tree-Based Synthesizers (SEQ)

The first type of generative model was a sequential decision tree.⁸² This has been used to synthesize health and social

TABLE 1. A Description of the Eight Clinical Trials Used in This Analysis

Trial (ClinicalTrials.gov identifier)	Description
REaCT-G/G2 (NCT02428114 and NCT02816164)	A total of 401 patients with early breast cancer were randomly assigned to receive filgrastim as primary FN prophylaxis. The trial evaluates whether 5 days of filgrastim was noninferior to the 7-10 days' dosing duration. The primary outcome was a composite of either FN or treatment-related hospitalization. ⁷²
REaCT-HER2 (NCT02632435)	A total of 48 patients with early-stage breast cancer were randomly assigned to receive trastuzumab-based chemotherapy using either PICC or totally implanted vascular access device (PORT). The trial feasibility was evaluated through a combination of end points that mainly included patient engagement and physician engagement, both as percentages. ⁷³
REaCT-BTA (NCT02721433)	A total of 263 patients with bone metastases from breast or castration-resistant prostate cancer were randomly assigned to treatment using BTAs once every 4 weeks or every 12 weeks. The data set comprised data for 230 patients. The primary end point was change in health-related quality of life. ⁷⁴
REaCT-ZOL (NCT03664687)	A total of 211 patients with EBC were randomly assigned to receive either one dose of zoledronate or seven doses with dosing once every 6 months for 3 years. The study was conducted to evaluate the feasibility of performing a large trial to study the effect of a single dose of zoledronate. As a primary outcome, the feasibility was assessed by a combination of metrics, including activation of sites and active participation in the trial. ⁷⁵
REaCT-ILIAD (NCT02861859)	A total of 218 patients with breast cancer with high risk of CINV were randomly assigned to triple therapy with added 5 mg olanzapine or placebo. The primary end point was frequency of self-reported significant nausea. ⁷⁶
CCTG MA27 (NCT00066573)	The CCTG MA27 study was a large phase III trial of 7,576 postmenopausal women with hormone receptor–positive early-stage breast cancer who were randomly assigned to receive exemestane v anastrozole. Median follow-up was 4.1 years. Follow-up care was semiannually during year 1 and annually thereafter, with yearly mammogram. Event-free survival was defined as the time from random assignment to the time of locoregional or distant disease recurrence, new primary breast cancer, or death from any cause. In addition, this study has competing risk results, making it even more applicable for global oncology practice. ⁷⁷
SWOG 0307 (NCT00127205)	The SWOG 0307 trial was a large phase III trial of 6,097 postmenopausal women with early-stage breast cancer who were randomly assigned to receive either intravenous zoledronate, oral clodronate, or oral ibandronate. Median follow-up was 4.1 years at the time of the initial publication, and follow-up recently closed at 10 years. Follow-up care was semiannually during the first 5 years and then annually until year 10 or death, with yearly mammogram. The primary outcome was DFS, defined as the time from registration to first disease recurrence (local, regional, distant), new breast primary, or death from any cause. A secondary outcome was OS, defined as the time from registration to death from any cause. Patients not experiencing DFS or OS events were censored at the date of last contact. ⁷⁸
NSABP B34 (NCT00009945)	The NSABP B34 trial was a multicenter, randomized, double-blind placebo-controlled trial that enrolled 3,323 patients with stage I-III breast cancer. After tumor removal, patients were stratified by age, auxiliary nodes, and estrogen and progesterone receptor status. They were assigned to either oral clodronate once daily for 3 years (n = 1,662) or placebo. The primary end point was DFS, defined as the time from random assignment to local, regional, or distant breast cancer recurrence, contralateral breast cancer, second primary malignant disease, or death from any cause before breast cancer recurrence. ⁷⁹

Abbreviations: BTAs, bone-targeted agents; CCTG, Canadian Cancer Trials Group; CINV, chemotherapy-induced nausea and vomiting; DFS, disease-free survival; EBC, early-stage breast cancer; FN, febrile neutropenia; HER2, human epidermal growth factor receptor 2; NSABP, National Surgical Adjuvant Breast and Bowel Project; OS, overall survival; PICC, peripherally inserted central catheter; REaCT, Rethinking Clinical Trials; SWOG, Southwest Oncology Group.

sciences data,^{47–85} and applied in research studies on synthetic data. Alternative implementations can use a gradient boosted decision tree.^{86,87}

GAN

A basic GAN consists of two artificial neural networks (ANNs), viz, a generator and a discriminator.⁸⁸ The generator *G* and the discriminator *D* play a min–max game. The input to the generator is noise, while its output is synthetic data. The discriminator has two inputs: the real training data and the synthetic data generated by the generator. The output of the discriminator indicates whether its input is real or synthetic. The generator is trained to trick the discriminator by generating samples that look real. However, the discriminator is trained to maximize its discriminatory capability. This type of generative model has been used extensively for the synthesis of health data.^{57,89–93}

There are many variations of the vanilla GAN that are widely used in tabular data synthesis.⁹⁴ The conditional tabular GAN (CTGAN) builds on conditional GANs⁹⁵ by addressing the multimodal distributions of continuous variables and the highly imbalanced categorical variables.⁹⁶ CTGAN solves the first problem by proposing a per-mode normalization technique. For the second problem, each category of a categorical variable serves as the condition passed to the GAN.

VAEs

Autoencoders use ANNs to compress (ie, encode) the input vector into lower dimensionality in the latent space and then reconstruct (ie, decode) it. The neural network is optimized by minimizing the reconstruction loss between the output and the input. In a VAE, the input data are mapped by the encoder to the data distribution represented by its statistical parameters rather than a lower-dimensionality set of

vectors. The VAE involves sampling from this distribution during the learning process.⁹⁷ A triplet-based VAE captures the interpretable latent representation by incorporating the additional triplet loss.⁹⁸

Evaluation of Utility

Utility was evaluated in terms of replicability by comparing the published analysis results using the real data sets for these clinical trials with the results of the same analysis performed on the synthetic data, as illustrated in Figure 1. Replicability is when a study is repeated using the same analytical methods but different data, which is the case here.⁹⁹ The published analyses are assumed to be representative of the types of analyses that would be performed on clinical trial data sets, and are summarized in the Data Supplement (Appendix SA).

Because of the stochasticity in the synthesis process, the additional variance introduced needs to be accounted for in the analysis using the synthetic data sets. The original proposal for SDG treated it as a form of multiple imputation.¹⁰⁰ Under the multiple imputation model, multiple data sets, say m , are synthesized for each clinical trial data set and combining rules (described in the Data Supplement, Appendix SB) are used to compute the parameter estimates and variances across the m synthetic data sets.^{101,102} This process is illustrated in Figure 2. We set $m = 10$, which is consistent with current practice for SDG.^{53,103,104,105}

A model is fitted to obtain a parameter estimate and its 95% CI on the real data as per the published study for each clinical trial. The same is applied to each of the m synthetic data sets. We combine the results from all the synthetic versions using

the combining rules discussed above. Subsequently, the estimates and CIs of the original and synthetic data are compared in terms of the *estimate agreement*, the *decision agreement*, and the *CI overlap*. These are defined below:

1. *Estimate agreement* is a Boolean indicator of whether the estimate produced by the synthetic data is within the 95% CI produced by the real data. This requires that a synthetic data effect estimate be within the range of plausible values for the true effect on the basis of evidence from the real data.
2. *Decision agreement* is a Boolean indicator of whether the same conclusion is drawn from the real and synthetic estimates. This means that the synthetic data estimates have the same direction and statistical significance as the real data. The decision agreement does not apply if the analysis is descriptive.
3. *CI overlap* is the overlapping proportion of the real and synthetic CIs,¹⁰⁶ which is a commonly used utility metric.

The two agreement metrics have been used in the past to compare the real-world data analysis results against a clinical trial reference¹⁰⁷ and to assess the replicability of psychological studies.¹⁰⁸

Evaluation of Privacy Risks

For the evaluation of privacy risks, we use two metrics: *attribution disclosure* and *membership disclosure*.

Attribution Disclosure

Previous work has considered similarity between real and synthetic data as a privacy metric on the premise that a replicated record means that it is a record about a real person.^{94,109,110} However, in practice, the high similarity between a synthetic record and a real record does not necessarily entail a high privacy risk.¹¹¹ For instance, if the real data were fully deidentified, the identity disclosure risk in the synthetic data would be extremely small even if synthetic records were similar to real records. Therefore, such record similarity metrics are not good indicators of privacy risks.

If a real record has a high risk of reidentification, and it is similar to a synthetic record, then an adversary can use the synthetic record to learn something new about the real person. This is the basic definition of the attribution disclosure metric.⁶⁴ It is expressed as a probability of learning something new conditional on identity disclosure. Identity disclosure was measured using a Gaussian copula-based estimator.¹¹² This metric has also been used in recent SDG benchmarking studies.^{113,114}

The European Medicines Agency has recently established a policy on the publication of clinical data for medicinal products.¹¹⁵ The guidelines accompanying the policy recommend a reidentification risk threshold of 0.09,¹¹⁶ which is the Health Canada threshold for the sharing of clinical trial data.¹¹⁷ This is consistent with the thresholds used by other

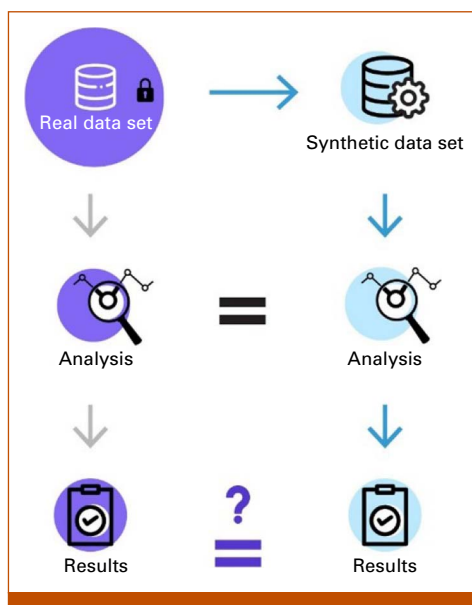


FIG 1. Utility (replicability) is evaluated by comparing the results from the real data to the synthetic data.

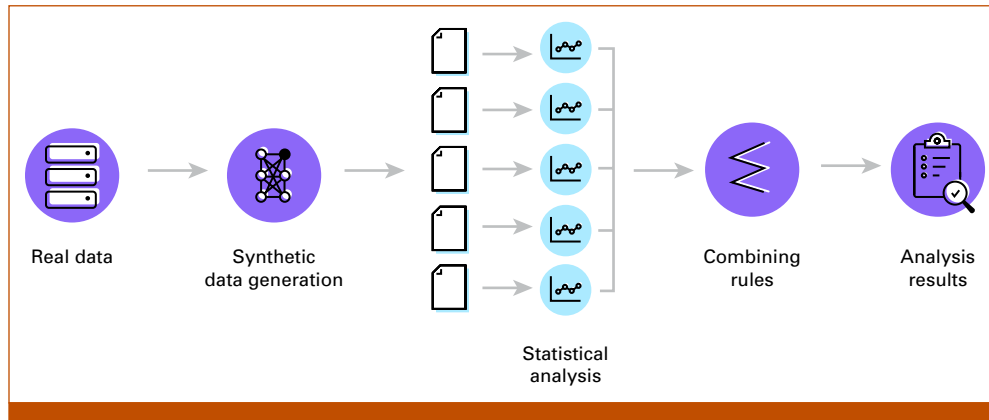


FIG 2. The process for computing valid parameters and making inferences from synthetic data using combining rules applied to multiple generated synthetic data sets.

agencies,^{118–122} and recommendations in a recent international standard on deidentification.¹²³ We therefore use that threshold for the interpretation of the attribution disclosure risk value.

Membership Disclosure

There has been a growing literature on assessing membership disclosure risks for synthetic data.^{70,92,93,113,124–130} Membership disclosure is when an adversary, using the information in synthetic data, determines that a target individual was included in the real data set used as input for SDG. Knowing that an individual was in the real data can reveal sensitive attributes about that individual if the data set pertains to a particular disease, condition, or process. The target individual is assumed to be from the same population as the real data set.

For example, if the real data set pertains to a clinical trial of patients with early breast cancer (EBC), membership disclosure would reveal that the target individual has EBC and satisfies the inclusion/exclusion criteria, or that they had participated in the study. Both would be deemed inappropriate disclosures of private information.

In the Data Supplement (Appendix SC), we provide an overview of membership disclosure measurement and extend current metrics to account for the number of patients in the clinical trial relative to the population. As an example, if a clinical trial represents 80% of the population, then an adversary would have a high success rate if they predict that every target individual sampled from the population is a member of the clinical trial data set that was synthesized. Therefore, it is important to also consider the sampling fraction.

The measure of membership disclosure is the F1 score, which reflects the accuracy of an adversary using the synthetic data to predict that a target record is a member of the training

data set. The threshold used in the literature for this measure is 0.2, and we show in the Data Supplement (Appendix SC) that the sampling fraction should also be <0.33 , otherwise the membership disclosure risk may still be high even if the F1 score is below the threshold.

Definition of Population

The size of the population that the clinical trial data set represents is an important parameter for the calculation of risk for both privacy risk metrics. The Data Supplement (Appendix SD) provides further details on the population sizes for each of the eight clinical trials.

Ethics

This study is approved by the Children's Hospital of Eastern Ontario Research Ethics Board—protocol no: 23/47X and the Ontario Cancer Research Ethics Board—project ID: 3749.

RESULTS

We replicated the published analyses for all eight clinical trials using both real and synthetic data. The utility results are shown in Table 2.

These results indicate that the utility of the sequential synthesis is generally better than the two other methods with estimate agreement and decision agreement that are always positive, where applicable. The CI overlap for the sequential synthesis is the highest among the methods for six of eight studies, and for one of these, the sequential synthesis overlap was very close to that of the next one (0.93 v 0.95). In all cases, the CI overlap for sequential synthesis was above 0.75.

In Table 3, we show the results for the attribution risk measurement. We can see here that all the three methods

TABLE 2. Utility Comparison Using Three Generative Models

Data Set	Sample Size	SEQ			GAN			VAE		
		Estimate Agreement	Decision Agreement	CI Overlap	Estimate Agreement	Decision Agreement	CI Overlap	Estimate Agreement	Decision Agreement	CI Overlap
REaCT-HER2+	48	1	1	0.77	1	1	0.88	1	1	0.94
REaCT-G/G2	401	1	1	0.91	^a	^a	^a	1	1	0.67
REaCT-ILIAD	218	1	1	0.99	1	1	0.85	1	0	0.74
REaCT-ZOL	211	1	^b	0.98	1	^b	0.88	0	^b	0.61
REaCT-BTA	230	1	1	0.85	1	0	0.68	1	0	0.72
CCTG MA27	7,576	1	1	0.90	1	1	0.62	1	1	0.82
SWOG 0307	6,097	1	1	0.93	1	0	0.50	1	1	0.95
NSABP B34	3,323	1	1	0.93	1	1	0.83	1	1	0.61

Abbreviations: BTAs, bone-targeted agents; CCTG, Canadian Cancer Trials Group; GAN, generative adversarial network; HER2, human epidermal growth factor receptor 2; NSABP, National Surgical Adjuvant Breast and Bowel Project; REaCT, Rethinking Clinical Trials; SEQ, sequential analysis; SWOG, Southwest Oncology Group; VAE, variational autoencoder.

^aTraining the generative model failed.

^bThe analysis is descriptive and hence decision agreement does not apply.

ensure that this type of risk is below the predefined threshold of 0.09 and the values are overall quite small.

The membership disclosure results are shown in [Table 4](#). These indicate that all three generative models have low membership disclosure risks on seven of eight clinical trial data sets. For the Canadian Cancer Trials Group data set, given that it represents a large proportion of the population, an adversary would achieve a successful membership disclosure attack with a naïve approach that does not require the synthetic data, and therefore by definition will have an unacceptable membership disclosure risk.

DISCUSSION

In summary, although the benefits of data sharing are widely acknowledged, one reason why sharing clinical trial data has been a challenge is privacy concerns. In this

paper, we evaluated the extent to which SDG methods can address these challenges of sharing clinical trial data sets. We compared multiple SDG methods on how well they were able to protect privacy, and how well they can replicate the published analytic results from the real (original) data sets. Three classes of methods were considered and eight different breast cancer clinical trials were evaluated.

Our findings indicate that sequential synthesis produces the highest utility results across all clinical trial data sets. Utility was defined as the ability to replicate the findings from published analyses. Different utility metrics were used, including the extent to which the same decision would be drawn from the synthetic data as from the real data, the extent to which the parameter estimates from the synthetic data are within the range of plausible values for the true effect on the basis of real data estimates, and the CI overlap.

TABLE 3. Attribution Disclosure Risk Results Using Three Generative Models

Data Set	SEQ		GAN		VAE	
	Risk Value	Risk	Risk Value	Risk	Risk Value	Risk
REaCT-HER2+	2.56E-04	LO	2.35E-04	LO	2.35E-04	LO
REaCT-G/G2	1.10E-04	LO	1.10E-04	LO	1.10E-04	LO
REaCT-ILIAD	2.90E-05	LO	2.90E-05	LO	2.90E-05	LO
REaCT-ZOL	1.58E-03	LO	1.41E-03	LO	1.10E-03	LO
REaCT-BTA	6.48E-04	LO	6.43E-04	LO	6.43E-04	LO
CCTG MA27	1.37E-03	LO	1.37E-03	LO	1.38E-03	LO
SWOG 0307	2.09E-03	LO	2.17E-03	LO	2.02E-03	LO
NSABP B34	2.25E-02	LO	2.02E-02	LO	1.83E-02	LO

Abbreviations: BTAs, bone-targeted agents; CCTG, Canadian Cancer Trials Group; GAN, generative adversarial network; HER2, human epidermal growth factor receptor 2; LO, low risk; NSABP, National Surgical Adjuvant Breast and Bowel Project; REaCT, Rethinking Clinical Trials; SEQ, sequential analysis; SWOG, Southwest Oncology Group; VAE, variational autoencoder.

TABLE 4. Membership Disclosure Comparison Using the Three Types of Generative Models

Data Set	n/N (sampling fraction)	SEQ		GAN		VAE	
		F_rel	Risk	F_rel	Risk	F_rel	Risk
REaCT-HER2+	0.021	0.15	LO	0.07	LO	0.09	LO
REaCT-G/G2	0.062	0.06	LO	0.06	LO	0.06	LO
REaCT-ILIAD	0.004	0.02	LO	0.02	LO	0.02	LO
REaCT-ZOL	0.023	0.02	LO	0.02	LO	0.02	LO
REaCT-BTA	0.207	0.13	LO	0.18	LO	0.18	LO
CCTG MA27	0.573	0.31	HI	0.32	HI	0.34	HI
SWOG 0307	0.147	0.13	LO	0.13	LO	0.13	LO
NSABP B34	0.158	-0.02	LO	-0.15	LO	-0.19	LO

NOTE. The threshold for the sampling fraction is 0.33, and 0.2 for the relative F1 score (F_rel).

Abbreviations: BTAs, bone-targeted agents; CCTG, Canadian Cancer Trials Group; GAN, generative adversarial network; HER2, human epidermal growth factor receptor 2; HI, high risk; LO, low risk; NSABP, National Surgical Adjuvant Breast and Bowel Project; REaCT, Rethinking Clinical Trials; SEQ, sequential analysis; SWOG, Southwest Oncology Group; VAE, variational autoencoder.

These results are consistent with previous comparative evidence on oncology data from SEER whereby a sequential synthesis generative model using decision trees had better utility than a GAN.¹²⁸ More advanced sequential synthesis approaches use boosted decision trees. There is evidence that discriminative and regression models on heterogeneous data sets (ie, those with continuous and categorical values, and with missingness) using boosted trees perform better than ANN architectures on tabular data sets.¹³¹⁻¹³⁵ Therefore, we would expect even better utility performance with sequential boosted decision trees as a generative model.

In addition to a potential inherent advantage, there are implementation differences that may contribute to sequential synthesis performing better. Our sequential synthesis implementation had a more complete process for handling missing data compared with the open-source Synthcity library. We observed that Synthcity generative models were not able to reproduce the missingness patterns in the synthetic data as well. Furthermore, the Synthcity implementation had limited hyperparameter tuning. Another commonly used implementation of these types of generative models, the Synthetic Data Vault,^{97,136} does not address these limitations either.

All methods performed well on the privacy metrics. By removing privacy barriers and making data more shareable among researchers, SDG can help reduce cycles of discovery, possibly even without the need for institutional ethics review board approval (eg, see the study by Guo et al¹³⁷). This can vastly improve the efficiency of secondary research using clinical trial data and accelerate new discoveries that would improve care.

Our results indicate that sequential synthesis can be a reasonable SDG method to enable such data access and data sharing. Although all evaluated methods performed well on managing privacy risks, sequential synthesis with decision trees had the highest utility across all data sets considered.

This evaluation was performed on eight clinical trials of various sizes, however, they were all patients with breast cancer. We chose breast cancer as this was the disease site for which the authors could bring together several large data sets relatively quickly. The authors' plan is to now expand into other tumor sites including, but not limited to, prostate and gastrointestinal malignancies. An analysis of trials in a different disease area may produce different findings.

We limited our evaluations to three commonly used generative models. Other methods could have been considered such as Bayesian networks and transformers. However, the latter requires much larger data set sizes than would normally be available in clinical trials.

Other approaches to combining multiple synthetic data sets could be explored, in particular, applying masking techniques followed by the synthesis, and then an analysis of multiple data sets with the application of the combining rules, as we did in this paper.¹³⁸ Such an approach can mitigate against generated data that are misaligned with the intended analysis, and was found to improve the robustness of the synthetic data in a recent study.¹³⁹

AFFILIATIONS

¹CHEO Research Institute, Ottawa, ON, Canada

²Replica Analytics Ltd, Ottawa, ON, Canada

³Ottawa Hospital Research Institute, Ottawa, ON, Canada

⁴Division of Medical Oncology, Department of Medicine, University of Ottawa, ON, Canada

⁵McMaster University, Hamilton, ON, Canada

⁶Department of Paediatrics, University of Ottawa, Ottawa, ON, Canada

⁷Alberta Health Services, Edmonton, AB, Canada

⁸Queen's University, Kingston, ON, Canada⁹Cancer Research and Biostatistics, Seattle, WA¹⁰University of Washington, Seattle, WA¹¹School of Epidemiology and Public Health, University of Ottawa, Ottawa, ON, Canada

CORRESPONDING AUTHOR

Khaled El Emam, PhD, BEng, CHEO Research Institute, 401 Smyth Rd, Ottawa, ON K1H 8L1, Canada; e-mail: kelemam@uottawa.ca.

PRIOR PRESENTATION

Presented in part at the ASCO Annual Meeting, Chicago, IL, June 2-6, 2023.

SUPPORT

Supported by a Data Transformation Grant of the Canadian Cancer Society (CCS grant #707600), the Canada Research Chairs program through the Canadian Institutes of Health Research, a Discovery Grant RGPIN-2022-04811 from the Natural Sciences and Engineering Research Council of Canada, and by the Government of Ontario.

AUTHOR CONTRIBUTIONS

Conception and design: Samer El Kababji, Ana-Alicia Beltran-Bless, Lucy Mosquera, Lois Shepherd, Bingshu Chen, Julie Gralow, Mark Clemons, Khaled El Emam

Financial support: Dhenuka Radhakrishnan, Khaled El Emam

Administrative support: Lisa Vandermeer, Mark Clemons

Provision of study materials or patients: Lisa Vandermeer, Mark Clemons, Khaled El Emam

Collection and assembly of data: Samer El Kababji, Xi Fang, Lisa Vandermeer, Alexander Paterson, Lois Shepherd, Bingshu Chen, William E. Barlow, Mark Clemons

Data analysis and interpretation: Samer El Kababji, Nicholas Mitsakakis, Xi Fang, Greg Pond, Dhenuka Radhakrishnan, Lois Shepherd, Bingshu Chen, Julie Gralow, Marie-France Savard, Khaled El Emam

Manuscript writing: All authors

Final approval of manuscript: All authors

Accountable for all aspects of the work: All authors

REFERENCES

1. Ebrahim S, Sohani ZN, Montoya L, et al: Reanalyses of randomized clinical trial data. *JAMA* 312:1024-1032, 2014
2. Ferran J-M, Nevitt S: European Medicines Agency Policy 0070: An exploratory review of data utility in clinical study reports for academic research. *BMC Med Res Methodol* 19:204, 2019
3. Phrma and E.F.P.I.A.: Principles for responsible clinical trial data sharing, 2013. <http://www.phrma.org/sites/default/files/pdf/PhRMAPrinciplesForResponsibleClinicalTrialDataSharing.pdf>
4. E. M. Agency: European Medicines Agency Policy on publication of data for medicinal products for human use: Policy 0070, 2014. http://www.ema.europa.eu/docs/en_GB/document_library/Other/2014/10/WC500174796.pdf
5. Taichman DB, Backus J, Baethge C, et al: Sharing clinical trial data: A proposal from the International Committee of Medical Journal Editors. *Ann Intern Med* 164:505-506, 2016
6. Institute of Medicine: Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk. Washington, DC, National Academies Press, 2015
7. International Committee of Medical Journal Editors: Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals, 2019. <http://www.icmje.org/icmje-recommendations.pdf>
8. The Wellcome Trust: Policy on data, software and materials management and sharing. Wellcome, 2017. <https://wellcome.ac.uk/funding/managing-grant/policy-data-software-materials-management-and-sharing>
9. National Institutes of Health: Final NIH statement on sharing research data, 2003. <https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html>
10. Protection of personal data in clinical documents—A model approach. TransCelerate Biopharma, 2017. <http://www.transceleratebiopharmainc.com/wp-content/uploads/2017/02/Protection-of-Personal-Data-in-Clinical-Documents.pdf>
11. De-identification and anonymization of individual patient data in clinical studies: A model approach. TransCelerate Biopharma, 2017. <http://www.transceleratebiopharmainc.com/wp-content/uploads/2015/04/TransCelerate-De-identification-and-Anonymization-of-Individual-Patient-Data-in-Clinical-Studies-V2.0.pdf>
12. Doshi P: Data too important to share: Do those who control the data control the message? *BMJ* 352:i1027, 2016
13. Rankin D, Black M, Bond R, et al: Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for data sharing. *JMIR Med Inform* 8:e18910, 2020
14. Polanin JR: Efforts to retrieve individual participant data sets for use in a meta-analysis result in moderate data sharing but many data sets remain missing. *J Clin Epidemiol* 98:157-159, 2018
15. Naudet F, Sakarovich C, Janiaud P, et al: Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: Survey of studies published in *The BMJ* and *PLOS Medicine*. *BMJ* 360:k400, 2018
16. Nevitt SJ, Marson AG, Davie B, et al: Exploring changes over time and characteristics associated with data retrieval across individual participant data meta-analyses: Systematic review. *BMJ* 357:1390, 2017

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.

Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians ([Open Payments](http://OpenPayments.org)).

Greg Pond

Employment: Roche Canada

Stock and Other Ownership Interests: Roche Canada

Honoraria: AstraZeneca

Consulting or Advisory Role: Takeda, Profound Medical

Lucy Mosquera

Employment: Aetion

Stock and Other Ownership Interests: Aetion

Patents, Royalties, Other Intellectual Property: Pending patents through work at Replica Analytics, an Aetion company

Alexander Paterson

Stock and Other Ownership Interests: Roche, Pfizer

William E. Barlow

Research Funding: Merck (Inst), AstraZeneca (Inst)

Marie-France Savard

Honoraria: Novartis Canada Pharmaceuticals Inc, Seagen, Knight Therapeutics, Merck, Roche Canada, AstraZeneca, Pfizer, Gilead Sciences, Lilly

Consulting or Advisory Role: Pfizer, Knight Therapeutics, seagen

Mark Clemons

Travel, Accommodations, Expenses: Pfizer

Khaled El Emam

Employment: Aetion

Leadership: Canary Medical, DistillerSR

Stock and Other Ownership Interests: Canary Medical, Aetion, DistillerSR

No other potential conflicts of interest were reported.

17. Villain B, Dechartres A, Boyer P, et al: Feasibility of individual patient data meta-analyses in orthopaedic surgery. *BMC Med* 13:131, 2015
18. Ventresca M, Schünemann HJ, Macbeth F, et al: Obtaining and managing data sets for individual participant data meta-analysis: Scoping review and practical guide. *BMC Med Res Methodol* 20:113, 2020
19. Iqbal SA, Wallach JD, Khoury MJ, et al: Reproducible research practices and transparency across the biomedical literature. *PLoS Biol* 14:e1002333, 2016
20. Banzi R, Canham S, Kuchin W, et al: Evaluation of repositories for sharing individual-participant data from clinical studies. *Trials* 20:169, 2019
21. Geifman N, Bollyky J, Bhattacharya S, et al: Opening clinical trial data: Are the voluntary data-sharing portals enough? *BMC Med* 13:280, 2015
22. National Academies of Sciences, Engineering, and Medicine: Reflections on Sharing Clinical Trial Data: Challenges and a Way Forward: Proceedings of a Workshop. Washington, DC, The National Academies Press, 2020
23. van Panhuis WG, Paul P, Emerson C, et al: A systematic review of barriers to data sharing in public health. *BMC Public Health* 14:1144, 2014
24. Kalkman S, Mostert M, Gerlinger C, et al: Responsible data sharing in international health research: A systematic review of principles and norms. *BMC Med Ethics* 20:21, 2019
25. Building Canada's health data foundation: Report 2. Expert Advisory Group, Pan-Canadian Health Data Strategy, Canada, 2021. <https://www.canada.ca/content/dam/phac-aspc/documents/corporate/mandate/about-agency/external-advisory-bodies/list/pan-canadian-health-data-strategy-reports-summaries/expert-advisory-group-report-02-building-canada-health-data-foundation/expert-advisory-group-report-02-building-canada-health-data-foundation.pdf>
26. Read KB, Ganshorn H, Rutley S, et al: Data-sharing practices in publications funded by the Canadian Institutes of Health Research: A descriptive analysis. *CMAJ Open* 9:E980-E987, 2021
27. Artificial Intelligence in Health Care. Washington, DC, National Academy of Medicine and the General Accountability Office, 2019
28. Emam KE, Jonker E, Moher E, et al: A review of evidence on consent bias in research. *Am J Bioeth* 13:42-44, 2013
29. de Montjoye YA, Hidalgo CA, Verleysen M, et al: Unique in the crowd: The privacy bounds of human mobility. *Sci Rep* 3:1376, 2013
30. Sweeney L, Yoo JS, Perovich L, et al: Re-identification risks in HIPAA safe harbor data: A study of data from one environmental health study. *J Technol Sci* 2017:2017082801, 2017
31. Yoo JS, Thaler A, Sweeney L, et al: Risks to patient privacy: A re-identification of patients in Maine and Vermont statewide hospital data. *J Technol Sci* 23:2018100901, 2018
32. Sweeney L: Matching Known Patients to Health Records in Washington State Data. Harvard University. Cambridge, MA, Data Privacy Lab, 2013
33. Sweeney L, Loewenfeldt M, Perry M: Saying it's anonymous doesn't make it so: Re-identifications of 'anonymized' law school data. *J Technol Sci* 23:2018111301, 2018
34. Zewe A: Imperiled information: Students find website data leaks pose greater risks than most people realize. Harvard John A. Paulson School of Engineering and Applied Sciences, 2020. <https://www.seas.harvard.edu/news/2020/01/imperiled-information>
35. de Montjoye Y-A, Radaelli L, Singh VK, et al: Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science* 347:536-539, 2015
36. Bode K: Researchers find 'anonymized' data is even less anonymous than we thought. https://www.vice.com/en_ca/article/dygy8k/researchers-find-anonymized-data-is-even-less-anonymous-than-we-thought
37. Clemons E: Online profiling and invasion of privacy: The myth of anonymization. *HuffPost*, 2013. https://www.huffpost.com/entry/internet-targeted-ads_b_2712586
38. Jee C: You're very easy to track down, even when your data has been anonymized. *MIT Technology Review*, 2019. <https://www.technologyreview.com/2019/07/23/134090/youre-very-easy-to-track-down-even-when-your-data-has-been-anonymized/>
39. Kolata G: Your data were 'anonymized'? These scientists can still identify you. *The New York Times*, 2019. <https://www.nytimes.com/2019/07/23/health/data-privacy-protection.html>
40. Lomas N: Researchers spotlight the lie of 'anonymous' data. *TechCrunch*, 2019. <https://techcrunch.com/2019/07/24/researchers-spotlight-the-lie-of-anonymous-data/>
41. Mitchell S: Study finds HIPAA protected data still at risks. *Harvard Gazette*, 2019. <https://news.harvard.edu/gazette/story/newsplus/study-finds-hipaa-protected-data-still-at-risks/>
42. Thompson SA, Warzel C: Twelve million phones, one dataset, zero privacy. *The New York Times*, 2019. <https://www.nytimes.com/interactive/2019/12/19/opinion/location-tracking-cell-phone.html>
43. 'Anonymised' data can never be totally anonymous, says study. *The Guardian*, 2019. <http://www.theguardian.com/technology/2019/jul/23/anonymised-data-never-be-anonymous-enough-study-finds>
44. Wolk A: The (im)possibilities of scientific research under the GDPR, 2020. <https://www.mofo.com/resources/insights/200617-scientific-research-gdpr.html>
45. Ghafur S, Dael JV, Leis M, et al: Public perceptions on data sharing: Key insights from the UK and the USA. *Lancet Digit Health* 2:e444-e446, 2020
46. El Emam K, Mosquera L, Hoptoff R: Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data. Sebastopol, CA, O'Reilly, 2020
47. Drechsler J, Reiter JP: An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Comput Stat Data Anal* 55:3232-3243, 2011
48. Arslan RC, Schilling KM, Gerlach TM, et al: Using 26,000 diary entries to show ovulatory changes in sexual desire and behavior. *J Pers Soc Psychol* 121:410-431, 2021
49. Bonnéry D, Feng Y, Henneberger AK, et al: The promise and limitations of synthetic data as a strategy to expand access to state-level multi-agency longitudinal data. *J Res Educ Eff* 12:616-647, 2019
50. Sabay A, Harris L, Bejugama V, et al: Overcoming small data limitations in heart disease prediction by using surrogate data. *SMU Data Sci Rev* 1:12, 2018
51. Freiman M, Lauger A, Reiter J: Data synthesis and perturbation for the American Community Survey at the U.S. Census Bureau. US Census Bureau, Working Paper, 2017. <https://www2.census.gov/adrm/CED/Papers/CY17/2017-09-FreimanLaugerReiter-ACSSynthesisPerturbation.pdf>
52. Nowok B: Utility of synthetic microdata generated using tree-based methods. UNECE Statistical Data Confidentiality Work Session, Helsinki, Finland, October 5-7, 2015
53. Raab GM, Nowok B, Dibben C: Practical data synthesis for large samples. *J Priv Confidentiality* 7:67-97, 2018
54. Nowok B, Raab GM, Dibben C: Providing bespoke synthetic data for the UK Longitudinal Studies and other sensitive data with the synthpop package for R 1. *Stat J IAOS* 33:785-796, 2017
55. Quintana DS: A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation. *eLife* 9:e53275, 2020
56. Reiter JP: New approaches to data dissemination: A glimpse into the future (?) *Chance* 17:11-15, 2004
57. Park N, Mohammadi M, Gorde K, et al: Data synthesis based on generative adversarial networks. *Proc VLDB Endow* 11:1071-1083, 2018
58. Hu J: Bayesian estimation of attribute and identification disclosure risks in synthetic data. *arXiv:1804.02784 stat*, 2018. <http://arxiv.org/abs/1804.02784>
59. Taub J, Elliot M, Pampaka M, et al: Differential correct attribution probability for synthetic data: An exploration, in Domingo-Ferrer J, Montes F (eds): *Privacy in Statistical Databases, Lecture Notes in Computer Science*. Cham, Switzerland: Springer International Publishing, 2018, pp 122-137
60. Hu J, Reiter JP, Wang Q: Disclosure risk evaluation for fully synthetic categorical data, in Domingo-Ferrer J (ed), *Privacy in Statistical Databases, Lecture Notes in Computer Science*. Cham, Switzerland: Springer International Publishing, 2014, pp 185-199
61. Wei L, Reiter JP: Releasing synthetic magnitude microdata constrained to fixed marginal totals. *Stat J IAOS* 32:93-108, 2016
62. Ruiz N, Muralidhar K, Domingo-Ferrer J: On the privacy guarantees of synthetic data: A reassessment from the maximum-knowledge attacker perspective, in Domingo-Ferrer J, Montes F (eds): *Privacy in Statistical Databases, in Lecture Notes in Computer Science*. Cham, Switzerland: Springer International Publishing, 2018, pp 59-74
63. Reiter JP: Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study. *J R Stat Soc Ser A: Stat Soc* 168:185-205, 2005
64. El Emam K, Mosquera L, Bass J: Evaluating identity disclosure risk in fully synthetic health data: Model development and validation. *J Med Internet Res* 22:e23139, 2020
65. Hernandez M, Epelde G, Alberdi A, et al: Synthetic data generation for tabular health records: A systematic review. *Neurocomputing* 493:28-45, 2022
66. El Emam K, Guide to the De-Identification of Personal Health Information. Boca Raton, FL, CRC Press (Auerbach), 2013
67. El Emam K, Mosquera L, Hoptoff R: Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data. Sebastopol, CA, O'Reilly Media, 2020
68. Krenmayr L, Frank R, Drobic C, et al: GANerAid: Realistic synthetic patient data for clinical trials. *Inform Med Unlocked* 35:101118, 2022
69. Beaulieu-Jones BK, Wu ZS, Williams C, et al: Privacy-preserving generative deep neural networks support clinical data sharing. *Circ Cardiovasc Qual Outcomes* 12:e005122, 2019
70. El Emam K, Mosquera L, Fang X: Validating a membership disclosure metric for synthetic health data. *JAMIA Open* 5:00ac083, 2022
71. Azizi Z, Zheng C, Mosquera L, et al: Can synthetic data be a proxy for real clinical trial data? A validation study. *BMJ Open* 11:e043497, 2021
72. Clemons M, Fergusson D, Simos D, et al: A multicentre, randomised trial comparing schedules of G-CSF (filgrastim) administration for primary prophylaxis of chemotherapy-induced febrile neutropenia in early stage breast cancer. *Ann Oncol* 31:951-957, 2020
73. Clemons M, Stober C, Kehoe A, et al: A randomized trial comparing vascular access strategies for patients receiving chemotherapy with trastuzumab for early-stage breast cancer. *Support Care Cancer* 28:4891-4899, 2020
74. Clemons M, Ong M, Stober C, et al: A randomised trial of 4- versus 12-weekly administration of bone-targeted agents in patients with bone metastases from breast or castration-resistant prostate cancer. *Eur J Cancer* 142:132-140, 2021
75. Awan A, Ng T, Conter H, et al: Feasibility outcomes of a randomised, multicentre, pilot trial comparing standard 6-monthly dosing of adjuvant zoledronate with a single one-time dose in patients with early stage breast cancer. *J Bone Oncol* 26:100343, 2021
76. Clemons M, Dranitsaris G, Sienkiewicz M, et al: A randomized trial of individualized versus standard of care antiemetic therapy for breast cancer patients at high risk for chemotherapy-induced nausea and vomiting. *Breast* 54:278-285, 2020
77. Goss PE, Ingle JN, Pritchard KI, et al: Exemestane versus anastrozole in postmenopausal women with early breast cancer: NCIC CTG MA.27—A randomized controlled phase III trial. *J Clin Oncol* 31:1398-1404, 2013

78. Gralow JR, Barlow WE, Paterson AHG, et al: Phase III randomized trial of bisphosphonates as adjuvant therapy in breast cancer: S0307. *J Natl Cancer Inst* 112:698-707, 2020
79. Paterson AHG, Anderson SJ, Lembersky BC, et al: Oral clodronate for adjuvant treatment of operable breast cancer (National Surgical Adjuvant Breast and Bowel Project protocol B-34): A multicentre, placebo-controlled, randomised trial. *Lancet Oncol* 13:734-742, 2012
80. REaCT: Rethinking Clinical Trials. <http://www.ohri.ca/auditfeedback>
81. Qian Z, Cebere B-C, van der Schaar M, Synthcity: Facilitating innovative use cases of synthetic data in different data modalities. <https://arxiv.org/abs/2301.07573v1>
82. Emam KE, Mosquera L, Zheng C: Optimizing the synthesis of clinical trial data using sequential trees. *J Am Med Inform Assoc* 28:3-13, 2020
83. Sabay A, Harris L, Bejugama V, et al: Overcoming small data limitations in heart disease prediction by using surrogate data. *SMU Data Sci Rev* 1:25, 2018
84. Freiman M, Lauger A, Reiter J: Data synthesis and perturbation for the American Community Survey at the U.S. Census Bureau. US Census Bureau, Working Paper, 2017. <https://www.census.gov/library/working-papers/2018/adrm/formal-privacy-synthetic-data-aacs.html>
85. Quintana DS: A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation. *eLife* 9:e53275, 2020
86. Bühlmann P, Hothorn T: Boosting algorithms: Regularization, prediction and model fitting. *Statist Sci* 22:477-505, 2007
87. Ke G, Meng Q, Finley T, et al: LightGBM: A highly efficient gradient boosting decision tree, in Guyon I, Luxburg UV, Bengio S, et al (eds): *Advances in Neural Information Processing Systems* 30. Long Beach, CA, Curran Associates, 2017, pp 3146-3154
88. Goodfellow I, Pouget-Abadie J, Mirza M, et al: Generative adversarial nets, in Ghahramani Z, Welling M, Cortes C, et al (eds): *Advances in Neural Information Processing Systems*. Montreal, Canada, Neural Information Processing Systems Foundation Inc, 2014, pp 2672-2680
89. Chin-Cheong K, Sutter T, Vogt JE: Generation of heterogeneous synthetic electronic health records using GANs. Presented at the Workshop on Machine Learning for Health (ML4H) at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Vancouver, BC, 2019
90. Choi E, Biswal S, Malin B, et al: Generating multi-label discrete patient records using generative adversarial networks. *arXiv:1703.06490 cs*, 2017. <http://arxiv.org/abs/1703.06490>
91. Goncalves A, Ray P, Soper B, et al: Generation and evaluation of synthetic patient data. *BMC Med Res Methodol* 20:108, 2020
92. Zhang Z, Yan C, Mesa DA, et al: Ensuring electronic medical record simulation through better training, modeling, and evaluation. *J Am Med Inform Assoc* 27:99-108, 2020
93. Yan C, Zhang Z, Nyemba S, et al: Generating electronic health records with multiple data types and constraints. *arXiv:2003.07904 cs stat*, 2020. <http://arxiv.org/abs/2003.07904>
94. Bourou S, El Saer A, Velivassaki TH, et al: A review of tabular data synthesis using GANs on an IDS dataset. *Information* 12:375, 2021
95. Mirza M, Osindero S: Conditional generative adversarial nets. *arXiv* 10.48550/arXiv.1411.1784
96. Xu L, Skoularidou M, Cuesta-Infante A, et al: Modeling tabular data using conditional GAN, in Wallach H, Larochelle H, Beygelzimer A, et al (eds): *Advances in Neural Information Processing Systems*. Vancouver, Canada, Neural Information Processing Systems Foundation Inc, 2019
97. Kingma DP, Welling M: Auto-encoding variational bayes 10.48550/arXiv.1312.6114
98. Ishfaq H, Hoogi A, Rubin D: TVAE: Triplet-based variational autoencoder using metric learning. *arXiv* 10.48550/arXiv.1802.04403
99. National Academies of Sciences, Engineering, and Medicine; Policy and Global Affairs; Committee on Science, Engineering, Medicine, and Public Policy; Board on Research Data and Information, et al: *Reproducibility and Replicability in Science*. Washington, DC, National Academies Press (US), 2019
100. Rubin DB: Statistical disclosure limitation. *J Off Stat* 9:461-468, 1993
101. Raghunathan TE, Reiter JP, Rubin DB: Multiple imputation for statistical disclosure limitation. *J Off Stat* 19:1, 2003
102. Reiter J: Satisfying disclosure restrictions with synthetic data sets. *J Off Stat* 18:531-543, 2002
103. Taub J, Elliot M, Sakshaug W: The impact of synthetic data generation on data utility with application to the 1991 UK samples of anonymised records. *Trans Data Priv* 13:1-23, 2020
104. Loong B, Zaslavsky AM, He Y, et al: Disclosure control using partially synthetic data for large-scale health surveys, with applications to CanCORS. *Stat Med* 32:4139-4161, 2013
105. Reiter J: Inference for partially synthetic, public use microdata sets. *Surv Methodol* 29:181-188, 2003
106. Karr A, Kohonen CN, Oganian A, et al: A framework for evaluating the utility of data altered to protect confidentiality. *Am Stat* 60:224-232, 2006
107. Franklin JM, Pawar A, Martin D, et al: Nonrandomized real-world evidence to support regulatory decision making: Process for a randomized trial replication project. *Clin Pharmacol Ther* 107: 817-826, 2020
108. Open Science Collaboration: PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science* 349:aac4716, 2015
109. Zhang Z, Yan C, Mesa DA, et al: Ensuring electronic medical record simulation through better training, modeling, and evaluation. *J Am Med Inform Assoc* 27:99-108, 2020
110. Zhao Z, Kunar A, Birke R, et al: CTAB-GAN: Effective table data synthesizing. *Proceedings of the 13th Asian Conference on Machine Learning, PMLR* 157:97-112, 2021
111. El Emam K: Status of synthetic data generation for structured health data. *JCO Clin Cancer Inform* 7:e2300071, 2023
112. Jiang Y, Mosquera L, Jiang B, et al: Measuring re-identification risk using a synthetic estimator to enable data sharing. *PLoS One* 17:e0269097, 2022
113. Mendelevitch O, Lesh MD: Fidelity and privacy of synthetic medical data, *arXiv:2101.08658 cs*, 2021. <http://arxiv.org/abs/2101.08658>
114. Yan C, Yan Y, Wan Z, et al: A Multifaceted benchmarking of synthetic electronic health record generation models. *Nat Commun* 13:7609, 2022
115. European Medicines Agency: European Medicines Agency Policy on publication of data for medicinal products for human use: Policy 0070, 2014. http://www.ema.europa.eu/docs/en_GB/document_library/Other/2014/10/WC500174796.pdf
116. European Medicines Agency: External guidance on the implementation of the European Medicines Agency Policy on the publication of clinical data for medicinal products for human use (v1.4), 2018. European Medicines Agency, London, UK
117. Health Canada: Guidance document on public release of clinical information, 2019. <https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/profile-public-release-clinical-information-guidance.html>
118. CMS: 2008 basic stand alone medicare claims public use files. <http://go.cms.gov/2itDh2o>
119. Erdem E, Prada S: Creation of public use files: Lessons learned from the comparative effectiveness research public use files data pilot project, 2011. <http://bit.ly/2xZKfyb>
120. Instructions for completing the limited data set ATA use agreement (DUA) (CMS-R-0235L). Department of Health & Human Services. <http://go.cms.gov/2yJ1KX4>
121. California Department of Health Care Services: Public reporting guidelines. <https://www.dhcs.ca.gov/dataandstats/Pages/PublicReportingGuidelines.aspx>
122. State of Vermont Agency of Education: Data governance. <https://education.vermont.gov/data-and-reporting/data-governance>
123. ISO/IEC 27559:2022: Information security, cybersecurity and privacy protection—Privacy enhancing data de-identification framework, 2023. <https://www.iso.org/standard/71677.html>
124. Choi E, Biswal S, Malin B, et al: Generating multi-label discrete patient records using generative adversarial networks, in Doshi-Velez F, Fackler J, Kale D, et al (eds): *Proceedings of Machine Learning for Healthcare 2017*, Boston, MA, MLResearchPress, 2017, pp 286-305
125. Stadler T, Oprisanu B, Troncoso C: Synthetic data—A privacy mirage, *arXiv:2011.07018 cs*, 2021. <http://arxiv.org/abs/2011.07018>
126. Torfi A, Fox EA: CorGAN: Correlation-capturing convolutional generative adversarial networks for generating synthetic healthcare records, *arXiv:2001.09346 cs*, 2020. <http://arxiv.org/abs/2001.09346>
127. Zhang Z, Yan C, Lasko TA, et al: SynTEG: A framework for temporal structured electronic health data simulation. *J Am Med Inform Assoc* 28:596-604, 2020
128. Goncalves A, Ray P, Soper B, et al: Generation and evaluation of synthetic patient data. *BMC Med Res Methodol* 20:108, 2020
129. Chen D, Yu N, Zhang Y, et al: GAN-leaks: A taxonomy of membership inference attacks against generative models. *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*. USA Virtual, Association for Computing Machinery, 2020
130. Hilprecht B, Härterich M, Bernau D: Monte Carlo and reconstruction membership inference attacks against generative models. *Proc Priv Enh Technol* 2019:232-249, 2019
131. Borisov V, Leemann T, Sessler K, et al: Deep neural networks and tabular data: A survey. *IEEE Trans Neural Netw Learn Syst* 10.1109/TNNLS.2022.3229161
132. Bojer CS, Meldgaard JP: Kaggle forecasting competitions: An overlooked learning opportunity. *Int J Forecast* 37:587-603, 2021
133. Shwartz-Ziv R, Armon A: Tabular data: Deep learning is not all you need. *Inf Fusion* 81:84-90, 2022
134. Grinsztajn L, Oyallon E, Varoquaux G: Why do tree-based models still outperform deep learning on typical tabular data? *Adv Neural Inf Process Syst* 35:507-520, 2022
135. Pathare A, Mangrulkar R, Suvana K, et al: Comparison of tabular synthetic data generation techniques using propensity and cluster log metric. *Int J Inf Manage Data Insights* 3:100177, 2023
136. Patki N, Wedge R, Veeramachaneni K: The Synthetic Data Vault. 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA). Montreal, QC, Canada, IEEE, 2016, pp 399-410
137. Guo A, Foraker RE, MacGregor RM, et al: The use of synthetic electronic health record data and deep learning to improve timing of high-risk heart failure surgical intervention by predicting proximity to catastrophic decompensation. *Front Digit Health* 2:576945, 2020
138. Jiang B, Raftery AE, Steele RJ, et al: Balancing inferential integrity and disclosure risk via model targeted masking and multiple imputation. *J Am Stat Assoc* 117:52-66, 2022
139. Grund S, Lütke O, Robitzsch A: Using synthetic data to improve the reproducibility of statistical results in psychological research. *Psychol Methods* 10.1037/met0000526 [epub ahead of print on August 4, 2022]