Journal of the American Medical Informatics Association, 28(1), 2021, 3–13
doi: 10.1093/jamia/ocaa249
Advance Access Publication Date: 13 November 2020
Research and Applications

Research and Applications

Optimizing the synthesis of clinical trial data using sequential trees

Khaled El Emam,[1,2,3] Lucy Mosquera,[3] and Chaoyi Zheng[3]

[1]School of Epidemiology and Public Health, Faculty of Medicine, University of Ottawa, Ottawa, Ontario, Canada, [2]Electronic Health Information Laboratory, Childrens Hospital of Eastern Ontario Research Institute, Ottawa, Ontario, Canada, [3]Replica Analytics Ltd, Ottawa, Ontario, Canada

Corresponding Author: Khaled El Emam, BEng, PhD, Children's Hospital of Eastern Ontario Research Institute, 401 Smyth Road, Ottawa, Ontario K1J 8L1, Canada; kelemam@ehealthinformation.ca

Received 2 June 2020; Editorial Decision 20 September 2020; Accepted 22 September 2020

ABSTRACT

Objective: With the growing demand for sharing clinical trial data, scalable methods to enable privacy protective access to high-utility data are needed. Data synthesis is one such method. Sequential trees are commonly used to synthesize health data. It is hypothesized that the utility of the generated data is dependent on the variable order. No assessments of the impact of variable order on synthesized clinical trial data have been performed thus far. Through simulation, we aim to evaluate the variability in the utility of synthetic clinical trial data as variable order is randomly shuffled and implement an optimization algorithm to find a good order if variability is too high.

Materials and Methods: Six oncology clinical trial datasets were evaluated in a simulation. Three utility metrics were computed comparing real and synthetic data: univariate similarity, similarity in multivariate prediction accuracy, and a distinguishability metric. Particle swarm was implemented to optimize variable order, and was compared with a curriculum learning approach to ordering variables.

Results: As the number of variables in a clinical trial dataset increases, there is a pattern of a marked increase in variability of data utility with order. Particle swarm with a distinguishability hinge loss ensured adequate utility across all 6 datasets. The hinge threshold was selected to avoid overfitting which can create a privacy problem. This was superior to curriculum learning in terms of utility.

Conclusions: The optimization approach presented in this study gives a reliable way to synthesize high-utility clinical trial datasets.

Key words: data synthesis, privacy enhancing technologies, data sharing, clinical trial transparency, secondary use

INTRODUCTION

It is important for analysts and researchers to get access to high-quality individual-level data for secondary purposes (such as for building statistical and machine learning models). In the case of clinical trials, the reanalysis of data from previous studies can provide new insights compared with the original publications.[1] Secondary analysis has produced informative research results including on drug safety, evaluating bias, and replication of studies, and for meta-analysis,[2] with the most common purposes being new analyses of the treatment effect and the disease state.[3] Also, there has been strong interest in making more clinical trial data available for secondary analysis by academia, the pharmaceutical industry, and regulators.[4–9]

However, data access remains a challenge. For example, an examination of trials registered on ClinicalTrials.gov found that only 15% of trials launched in 2019 plan to share data.[10] An analysis of the success rates of getting individual-level data for research projects

from authors found that the percentage of the time these efforts were successful varied significantly and was generally low at 58%,[11] 46%,[12] 14%,[13] and 0%.[14] Some researchers note that getting access to datasets from authors can take from 4 months to 4 years.[14]

One reason for this challenging environment is increasingly strict data protection regulations: a recent National Academy of Medicine and Government Accountability Office report highlights privacy as presenting a data access barrier for the application of AI and machine learning in healthcare.[15] While patient (re)consent is one legal basis for making data available for secondary purposes, it is often impractical to get retroactive consent under many circumstances and there is significant evidence of consent bias.[16]

Anonymization is another approach for addressing privacy concerns when making clinical trial data available for secondary analysis. However, there have been repeated claims of successful reidentification attacks on anonymized data,[17–23] eroding public and regulator trust in this approach.[23–32]

To solve this problem, there is growing interest in using and disclosing synthetic data instead. There are many use cases where synthetic data can provide a practical solution to the data access problem.[33,34] In fact, it was recognized some time ago that synthetic data are a key approach for data dissemination compared with more traditional disclosure control methods.[35] Furthermore, data synthesis has been highlighted as a key privacy enhancing technology to enable data access for the coming decade.[36]

Multiple researchers have noted that synthetic data does not have an elevated identity disclosure (privacy) risk because there is no unique or one-to-one mapping between the records in the synthetic data with the records in the original data.[35,37–43] Therefore, our focus in this article will be on ensuring that the synthesized data has sufficient utility, which is generally defined as the ability to replicate patterns and conclusions that were in the original data from the synthetic data.

Classification and regression trees[44] have been proposed for data synthesis when implemented in a sequential manner.[45] Using a scheme similar to sequential imputation,[46,47] trees are used quite extensively for the synthesis of health and social sciences data.[48–56] With these types of models, a variable is synthesized by using the values earlier in the sequence as predictors. Conceptually, sequential synthesis is similar to modeling multiple outcome variables using classifier chains[57] and regressor chains.[58] Compared with deep learning synthesis methods that require large datasets,[37,59–61] sequential decision trees work well for small datasets, such as clinical trial data, and work well with heterogenous variable types (such heterogeneity remains an area of research using other synthesis methods).[62]

It is known that the order of the variables can influence the accuracy of model chains.[63] The dependence of the synthetic data utility on the order that the variables are synthesized in is also a recognized issue[64] but has not been investigated in depth. For sequential data synthesis, variable order is important because each variable's generative model is fitted using only the variables before it in the order. When the preceding variables are weak predictors of subsequent variables, the synthesized values will have low utility, and synthesis errors will propagate, and potentially be amplified, through the chain. If the utility is dependent on variable order, then an arbitrary factor would effectively be driving nontrivial variation in the quality of synthesized data.

One approach to address this problem is to select the highest utility dataset among those generated through multiple random variable orders. However, this would not ensure that the utility meets acceptable threshold levels, and it is an inefficient way to search for a good utility variable order. One can instead model the dependence among the variables and select the variable order accordingly. However, dependence does not imply directionality, which is important for selecting an order.

The purpose of the current study is to assess the variation in the utility of synthetic clinical trial data generated using sequential decision trees, and if the variation is high, to optimize the order to meet data utility thresholds. In such a case, the optimal selection of variable order will ensure more consistent results and better data utility.

MATERIALS AND METHODS

Our methods had 2 main components. The first was to simulate and assess the variation in the utility of synthetic data due to variable order when using sequential techniques. This analysis would allow us to draw empirically supported conclusions about the extent to which variable order is indeed a problem. The second component was to evaluate an optimization algorithm that would target a threshold level of utility in the generated data. In that way we would have stronger assurances that variable order would not degrade utility in an arbitrary manner.

In the following sections we describe the clinical trial datasets and their sources, the variable order simulation process, the data utility metrics that we used, and the optimization method we used and its evaluation.

Sources of clinical trial data

To perform our simulations, we need access to clinical trials data. The following discussion presents the criteria and considerations for identifying usable datasets for our analysis. However, it also highlights the challenges in getting access to clinical trial datasets for this type of research project.

We defined the following criteria when selecting a data source for the clinical trial datasets:

- Access should be provided to individual-level patient data (instead of documents or summary statistics). The individual data is needed to perform the simulations.
- The datasets should be downloaded. We have access to significant high-performance computing CPU and GPU capacity to enable us to perform the simulations in a reasonable amount of time and with negligible incremental costs. Alternatively, any data source that only allows access to data in a virtual secure enclave would need to provide comparable cost-effective high-performance computing capacity for us to perform the study.
- The datasets should be readily available to other researchers to enable replication and extension of the current analysis.
- Relative time from request to data access, and incremental costs would also be relevant factors in selecting data sources.

Three data sources for clinical trial datasets were examined: (1) regulatory authorities (ie, the European Medicines Agency, Health Canada, and the Food and Drug Administration); (2) individual investigators; and (3) data sharing platforms, namely the Yale University Open Data Access project,[65] Project Data Sphere,[66] clinicalstudydatarequest.com,[67] and Vivli.[68] In Supplementary Appendix A, we provide a review of potential data sources and assess them with respect to our previous criteria.

We concluded that, relative to other options, the Project Data Sphere (PDS) data sharing platform met our criteria, while other options would not have met all of our criteria or would not have

Table 1. Summary of the 6 oncology trials used in the analysis with the National Clinical Trial number and the primary sponsor indicated, as well as the number of patients and variables used in the synthesis

Dataset	Individuals	Variables		
		Total	Binary/ Categorical	Discrete/ Continuous
Trial 1 (NCT00041197): National Cancer Institute				
Tests if postsurgery receipt of imatinib could reduce the recurrence of GISTs. Imatinib is an Food and Drug Administration approved protein-tyrosine kinase inhibitor for treating certain cancers of the blood cells. This drug is hypothesized to be effective against GIST as imatinib inhibits the kinase which experiences gain of function mutations in up to 90% of GIST patients.[69] At the time of this trial the efficacy of imatinib for GISTs as well as the optimal dosage for treatment of GISTs was unknown.	773	129	71 (55)	58 (45)
Trial 2 (NCT01124786): Clovis Oncology				
Most pancreatic cancer patients have advanced inoperable disease and potentially metastases. At the time of this trial the first line therapy for patients with inoperable disease was gemcitabine monotherapy. One transporter (hENT1: human equilibrative nucleoside transporter-1) has been identified as a potential predictor of successful treatment via gemcitabine. This trial compares standard gemcitabine therapy to a novel fatty acid derivative of gemcitabine. This is hypothesized to be superior to gemcitabine in metastatic pancreatic adenocarcinoma patients with low hENT1 activity as it exhibits anticancer activity independent of nucleoside transporters like hENT1, while gemcitabine seems to require nucleoside transporters for anticancer activity.	367	88	24 (27.2)	64 (72.3)
Trial 3 (NCT00688740): Sanofi				
This phase 3 trial compares adjuvant anthracycline chemotherapy (fluorouracil, doxorubicin, and cyclophosphamide) with anthracycline taxane chemotherapy (docetaxel, doxorubicin, and cyclophosphamide) in women with lymph node positive early breast cancer.	746	239	148 (61.9)	91 (38.1)
Trial 4 (NCT00113763): Amgen				
This was a randomized phase 3 trial examining whether panitumumab, when combined with best supportive care, improves progression-free survival among patients with metastatic colorectal cancer, compared with those receiving best supportive care alone.[70,71] Patients included in the study had failed other chemotherapy options available at the time of the study. Participants were enrolled between 2004 and 2005.	463 (sponsor only provided 370 in the dataset)	59	22 (37.2)	37 (62.8)
Trial 5 (NCT00460265): Amgen				
This was also a randomized phase 3 trial on panitumumab but among patients with metastatic and/or recurrent squamous cell carcinoma of the head and neck. The treatment group received panitumumab in addition to other chemotherapy (cisplatin and fluorouracil), while the control group received cisplatin and fluorouracil as first-line therapy.[72] Participants were enrolled between 2007 and 2009.	657 (sponsor only provided 520 in the dataset)	401	162 (40.3)	239 (59.6)
Trial 6 (NCT00119613): Amgen				
This was a randomized and blinded Phase 3 trial aimed at evaluating whether "increasing or maintaining hemoglobin concentrations with darbepoetin alfa" improves survival among patients with previously untreated extensive-stage small cell lung cancer. The treatment group received darbepoetin alfa with platinum-containing chemotherapy, whereas the control group received placebo instead of darbepoetin alfa.	600 (sponsor only provided 479 in the dataset)	382	82 (21.4)	300 (78.6)

Values are n (%). Variables are classified as either binary/categorical or ordered discrete/continuous. Dates are converted to relative days and therefore are considered continuous.

GIST: gastrointestinal stromal tumor.

been practical for our purposes: we were able to download individual-level patient datasets within days of making the request to run the simulations on our own high-performance computing infrastructure, other researchers can readily request the same data under the same conditions, and there would be no incremental costs. PDS has holdings of oncology clinical trial datasets sponsored by industry and public funders.

We selected 6 studies that could be downloaded from PDS, ensuring that we had variability in the number of patients, the number of variables within the range typical of clinical trials, disease areas, and therapeutic interventions. This would allow for a broader generalization of the results.

Datasets

We performed our simulations on the 6 oncology clinical trial datasets from PDS as summarized in Table 1. We considered the screening criteria, demographic variables, medical history, baseline characteristics, and the endpoints in this analysis.

Simulation and synthesis processes

For the simulations, we repeated the synthesis 1000 times for each dataset, and each time randomly shuffling the variable order that was used in the sequential tree generation process. The method we used to generate synthetic data is called conditional trees,[73] although

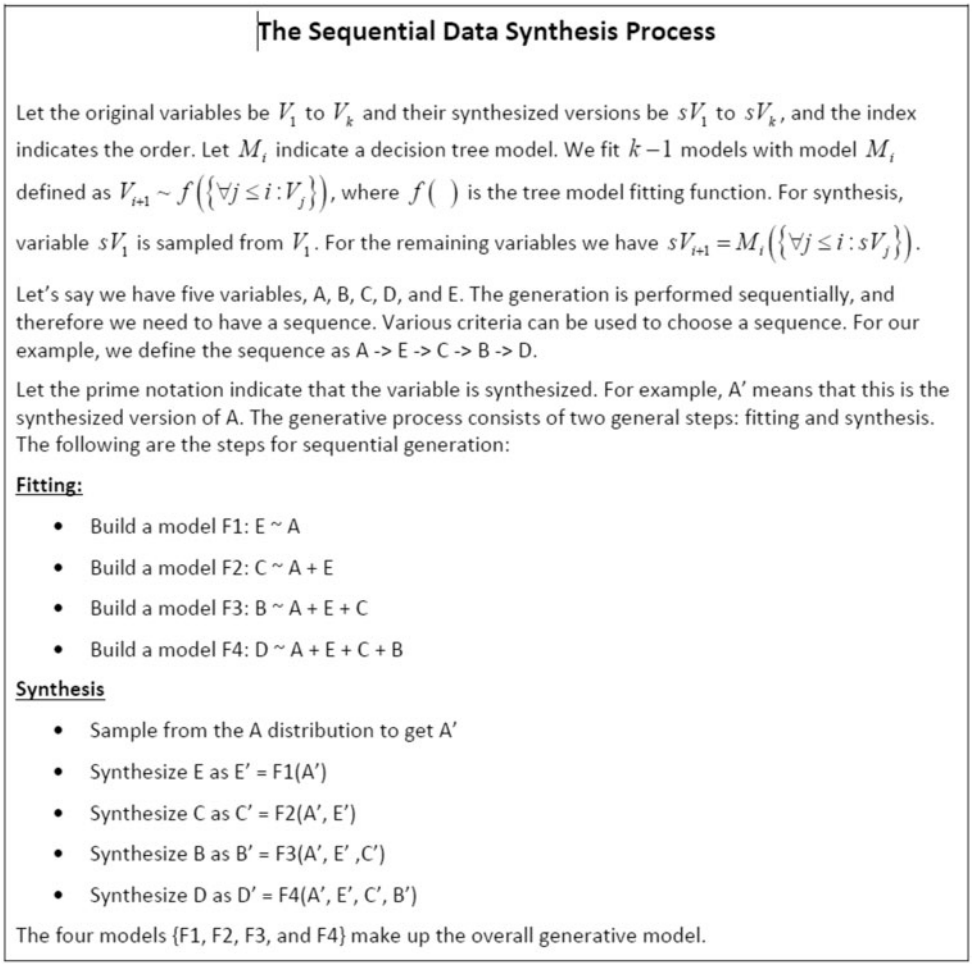

Figure 1. A description of the sequential data synthesis process using classification and regression trees. Although any set of classification and regression methods can be used in principle.

other tree algorithms could also be used. A summary of the algorithm is provided in Figure 1, and additional implementation and data preparation details are included in Supplementary Appendix B. When a fitted model is used to generate data, we sample from the predicted terminal node in the tree to get the synthetic values.

Evaluating synthetic data utility

Data utility is broadly defined as the ability to replicate patterns and conclusions that were in the original data from the synthetic data. The type of results that should be examined will depend on the data uses. For example, a statistical analysis of the data will require different utility evaluations and benchmarks than using the data to test statistical software.

A recent review identified 7 types of utility assessments for synthetic data.[74] Our focus here is on analytic uses of the data and therefore we examine general metrics comparing real and synthetic data analysis results. These metrics are not workload aware, but they reflect common data analysis tasks and are used often within the data synthesis and machine learning community.

For each simulation iteration, the synthetic data utility was estimated using 3 metrics, as summarized in El Emam et al[34]: comparisons of univariate distributions, "all-models" comparisons of multivariate prediction accuracy, and distinguishability. The former 2 reflect a broad spectrum of common statistical analysis on clinical datasets. The latter is an omnibus comparison of multivariate distributions using a binary classifier,[75,76] and is equivalent to the discriminator function used to evaluate (and train) performance in generative adversarial networks (an architecture of deep artificial neural networks).[77]

Univariate comparisons

We first compared the univariate distributions between the real and synthetic datasets on all variables. The comparison of univariate distributions as a utility metric is common in the synthesis literature.[37,78] For that purpose we utilized the Hellinger distance.[79] This has been shown to behave in a consistent manner as other distribution comparison metrics in the context of evaluating disclosure control methods when comparing original and transformed data.[80] It also has the advantage of being bounded between 0 and 1, which makes it easier to interpret. We computed the median Hellinger distance across all variables for each iteration during the simulation.

Comparing multivariate predictions

The second metric was a measure of multivariate prediction accuracy. It tells us the extent to which the prediction accuracy of synthetic data models is the same as the models from the real data.[81,82] The comparison of prediction model accuracy has been used, for ex-

ample, to compare the prediction of hospital readmissions between real and synthetic data[59] and to predict treatment arms from a large synthetic clinical study dataset.[83]

We built general boosted regression models,[84] taking each variable as an outcome to be predicted by all of the other variables. Hence, we built "all multivariate models" for the synthetic and real datasets. For each model 10-fold cross validation was used to compute the area under the receiver-operating characteristic curve (AUROC)[85] as a measure of model accuracy. We then compared the synthetic data and the real data accuracy by computing the relative absolute difference in the median AUROC measures for each dataset. Because AUROC requires a discrete variable, we discretized all continuous outcome variables using univariate k-means.[86]

Distinguishability

The third utility metric is based on propensity scores.[87,88] The real and synthetic datasets are pooled, and a binary indicator is assigned to each record depending on whether it is a real data record or a synthetic data record. A binary classification model is then constructed to distinguish between the real and synthetic records where the original variables are predictors and the binary indicator variable is the outcome. A 10-fold cross-validation is used to compute the propensity score (the predicted probability). The specific classification technique we use is generalized boosted models.[89]

The distinguishability score is computed as the mean square difference of the predicted probability from 0.5, which is the value where it is not possible to distinguish between the 2 datasets:[87]

$$d = {}^{1}\!/_{N} \sum_i (p_i - 0.5)^2 \quad (1)$$

where N is the size of the synthetic dataset and p_i is the propensity score for observation i.

If the 2 datasets are the same, then there will be no distinguishability between them. One reason for such a result would be if the synthetic data generator was overfit and effectively recreated the original data. In such a case the propensity score of every record will be close to or at 0.5, in that the classifier is not able to distinguish between real and synthetic data, and d approaches 0. If the 2 datasets are completely different, then the classifier will be able to distinguish between them. In such a case the propensity score will be either 0 or 1, with d approaching 0.25.

Across all 1000 simulation runs, we examined the median and 95% confidence interval on each dataset for the 3 utility metrics (the 2.5th percentile and the 97.5th percentile). This will indicate how stable the utility of the datasets are as the variable order is shuffled.

Because the generation of synthetic data is stochastic, there will be confounding variability in the utility metrics due to the synthesis process itself. Therefore, we average this out by generating 50 synthetic datasets for each of the 1000 variable orders, compute the utility metrics, and take the average of these 50 values to represent the value for that variable order. That way, we can factor out the impact of the stochastic synthesis process from the variability that we are interested in measuring. We did not observe meaningful fluctuations in that average when we used 100 or 150 generated datasets, and therefore the 50 iterations were deemed to be representative.

Curriculum learning

As a baseline method to defining an appropriate order for variables in sequential synthesis, we examined a curriculum learning approach.[90] For machine learning tasks, it is hypothesized that starting the training with easier or more general examples, and then continuing on to the more complex ones accelerates convergence and improves model accuracy. In the context of sequential data synthesis, it was argued that curriculum learning would better capture the dependence among the attributes.[61]

To operationalize this we discretized all continuous variables using univariate k-means.[86] Variables with fewer categories were put first in the order because they were deemed to be simpler. That curriculum learning order was evaluated as the baseline order, and all utility metrics were computed for that (note that the synthesis was performed on the original data and not the discretized values).

Particle swarm optimization

To optimize on the variable order we used a particle swarm algorithm.[91,92] This uses a search heuristic to find the global optimum without requiring the objective function to be continuous. For the objective function we computed the utility metrics and used a hinge loss function that was being minimized.[93] The hinge loss considers the utility metric to be 0 if it is below a threshold. For example, the distinguishability loss is 0 if distinguishability is below 0.05. The threshold ensures that we do not overfit the generated trees to the data. The overall loss for each of the utility metrics is therefore:

$$\begin{aligned} loss_1 &= \max(0, d - 0.05) \\ loss_2 &= \max(0, h - 0.1) \\ loss_3 &= \max(0, a - 0.1) \end{aligned} \quad (2)$$

where h is the Hellinger distance and a is the median relative absolute AUROC difference. A compound loss was also computed as the unweighted sum of all 3 losses as an optimization criterion:

$$loss_4 = \max(0, d - 0.05) + \max(0, h - 0.1) + \max(0, a - 0.1) \quad (3)$$

Therefore, in total the optimization was evaluated with 4 different loss functions.

RESULTS

We present 3 sets of graphs showing the different utility scores across all 6 trials.

Figure 2 shows the results across the 6 trials for the Hellinger distance. While there is a little bit of variation, in general the distance was relatively low and the variation within a narrow range.

Figure 3 has the results for the multivariate prediction models with the AUROC accuracy results. Although trial 4 has the most variation. This trial had the fewest variables and this may have created more instability in prediction performance relative to the other trials, hence the wider variation (although the amount of variation in this case is not that large).

In Figure 4, we see nontrivial variation in the distinguishability score. Specifically, trials 3, 5, and 6 show a large amount of variation due to variable order. The unoptimized (default) variable order does not necessarily capture the dependencies among the variables, and therefore in a sequential synthesis process, a particular variable may have been poorly modeled by the variables before it in the sequence. These poorly synthesized variables are then easier for a discriminator to use to distinguish between the real and synthetic datasets. This inability to capture dependencies is more pronounced the more variables there are in a dataset, as is the case with trials 3, 5, and 6.

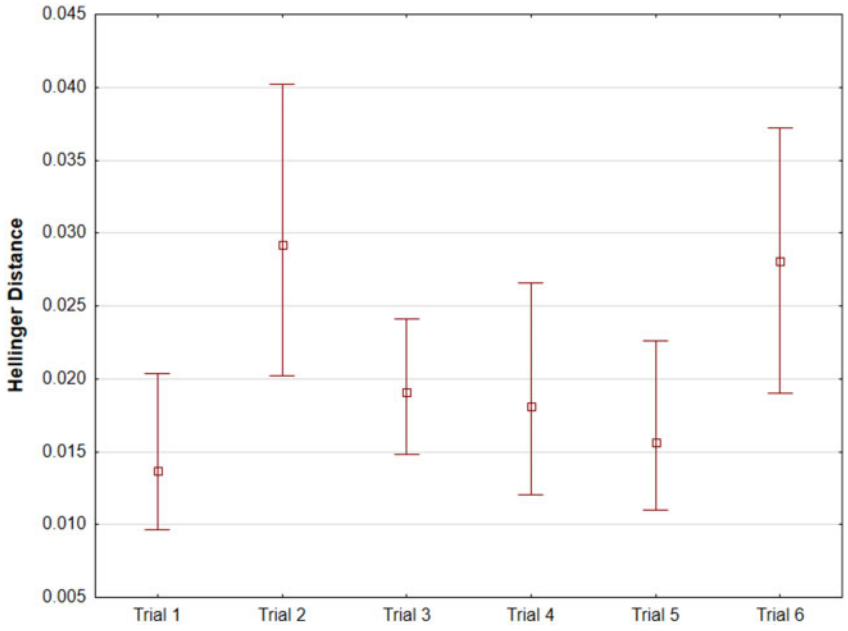

Figure 2. The Hellinger distance median value and 95% confidence intervals for the 6 clinical trial datasets. This is a value between 0 and 1, with lower values indicating that the univariate distributions of the real and synthetic variables are similar. In general, values in the lowest decile (≤0.1) would be indicative of reasonable similarity.

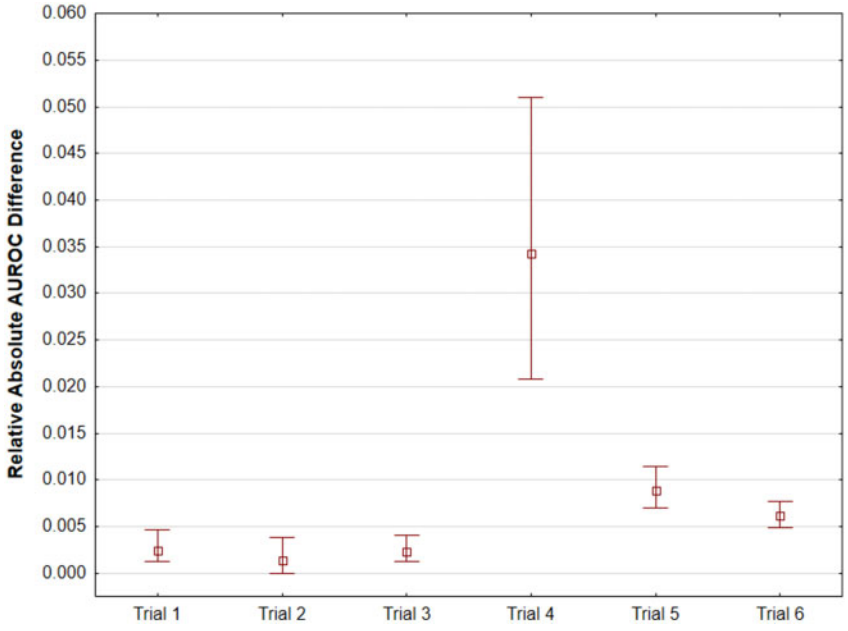

Figure 3. The relative absolute difference in area under the receiver-operating characteristic curve (AUROC) median value and 95% confidence intervals for the 6 clinical trial datasets. This is a value between 0 and 1, with lower values indicating that the multivariate models built using the real and synthetic datasets are similar. In general, values in the lowest decile (≤0.1) would be indicative of reasonable similarity.

The differences among the 3 utility metrics are not surprising because they are measuring different things, and they are also influenced by outliers differently. However, it is clear that the larger the number of variables there are in the dataset, the greater the variability in the distinguishability score is. There was no discernable relationship between the proportion of categorical vs continuous variables and the utility outcomes.

The curriculum learning baseline order results are shown in Table 2. Compared with the median values that were observed during the simulation, curriculum learning was not consistently better than the random median result, and was not consistently lower than our utility thresholds (used in the loss metrics).

After optimization, we show the results in Table 3 for optimizing on distinguishability ($loss_1$). An example illustrating how the optimization has affected the ordering of variables is provided in Supplementary Appendix C, in which we show the original ordering and the ordering after optimization that produced the result in Table 3. The results for the Hellinger distance loss, relative absolute AUROC

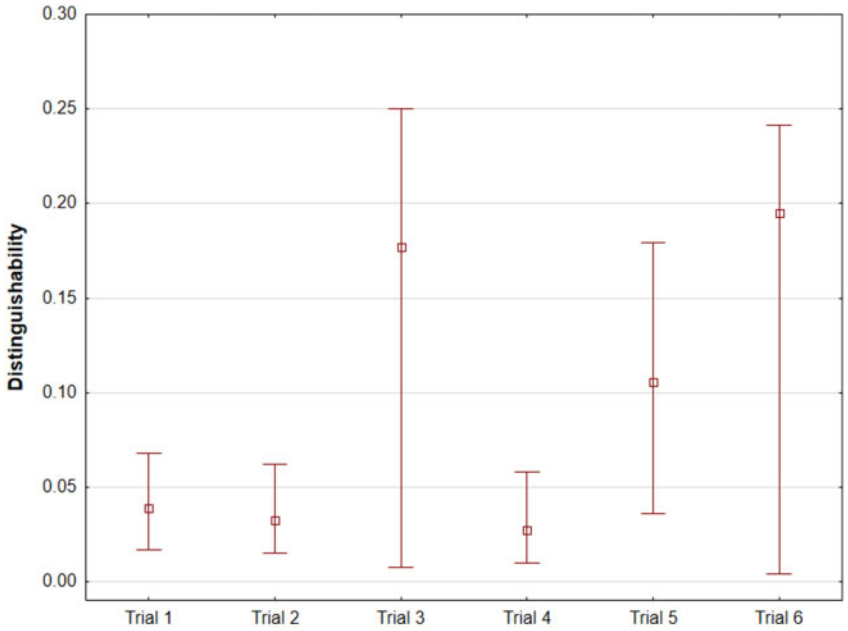

Figure 4. The distinguishability score median value and 95% confidence intervals for the 6 clinical trial datasets. This is a value between 0 and 0.25, with lower values indicating that the overall real and synthetic datasets are not distinguishable from each other by a discriminative model. In general, values in the lowest quintile ($\leq$0.05) would be indicative of reasonable nondistinguishability.

Table 2. Utility results for the curriculum learning variable order

Trial	Distinguishability	Hellinger	AUROC
1	0.114	0.0147	0.002
2	0.064	0.026	0.001
3	0.2	0.011	0.003
4	0.034	0.03	0.059
5	0.101	0.019	0.012
6	0.232	0.023	0.008

AUROC: area under the receiver-operating characteristic curve.

Table 3. Utility results after the optimal variable order was selected with optimization on distinguishability

Trial	Distinguishability	Hellinger	AUROC
1	0.0113	0.0118	0.0019
2	0.033	0.027	0.001
3	0.049	0.017	0.0026
4	0.02	0.0204	0.0584
5	0.044	0.0135	0.0118
6	0.0388	0.0277	0.009

AUROC: area under the receiver-operating characteristic curve.

difference loss, and the compound loss are included in Supplementary Appendix D.

We can make 3 observations. The first is that optimization on the Hellinger distance and the relative absolute AUROC difference do not ensure that the utility meets the threshold for the other utility metrics. The second is that with distinguishability as the main loss ($loss_1$) we meet the threshold values for the other utility metrics. The third observation is that the results for optimizing on the distinguishability loss ($loss_1$) are the same as for the compound loss ($loss_4$). When we checked the optimal orders, they were also the same when using distinguishability loss and compound loss. The optimization on the compound loss is more computationally intensive than just for distinguishability. Given that the results are the same across all 6 trials, then a strong case can be made for only optimizing on the distinguishability metric.

Additional details on the performance of the optimization approach compared with random search is included in Supplementary Appendix E.

DISCUSSION

Summary

Our results indicate that the variation in the data utility of synthesized clinical trial data using (unoptimized) sequential trees was impacted significantly by the variable order, after accounting for natural variation due to the stochastic nature of data synthesis. The variability in utility was more pronounced as the number of variables increased, meaning that some orders will result in quite poor utility results on some of the key utility metrics. The number of patients did not seem to play a prominent role in affecting the variation in data utility, although our datasets did err on the small side.

Optimization using the particle swarm algorithm combined with a distinguishability hinge loss reliably found the variable orders that ensure that the utility metrics are below an acceptable threshold level across multiple utility metrics. This optimization strategy also had better utility than a curriculum learning approach. Particle swarm optimization was faster than a random search for an acceptable variable order. Therefore, we recommend implementing this optimization strategy when using sequential synthesis techniques. Furthermore, using a curriculum learning strategy to set the order of the variables did not ensure that the utility was consistently acceptable across datasets and utility metrics, and therefore that approach is not recommended.

Practical implications

Sequential trees as a method for data synthesis have the advantage in that they can work with datasets with few patients as well as for very large datasets (this is a feature of decision trees in general), datasets that are heterogeneous in the variable types (eg, combining continuous and high-cardinality categorical variables), and for datasets with significant missingness. These are common characteristics of many real health datasets. While the impact on datasets with few variables will be smaller, there would be minimal downside to the optimization of variable order every time a clinical trial dataset is synthesized using sequential trees. And our results suggest that this optimization should be performed on a hinge loss function based on the distinguishability metric.

Order-optimized sequential trees can therefore be good methods for the synthesis of clinical trial data. Other methods, such as deep learning models that are also being applied to the synthesis of health data, do not scale down well to the small datasets typically encountered with clinical trials.

For data custodians sharing synthetic data, providing data utility metrics as part of the documentation accompanying the synthetic data would be desirable. This allows the data users to validate that the overall utility is acceptable. The key utility metric is distinguishability, but other univariate and multivariate utility metrics, as described in this article, would also be informative.

Research contributions

The contributions of this work are the following: (1) we empirically demonstrate that variable order has an impact on synthetic clinical trial data utility for a sequential synthesis method commonly used on health data, (2) we propose and evaluate a method for optimally selecting the variable order, and (3) this is the first study to examine optimal synthesis for clinical trial datasets. Given the growing demand for access to clinical trial data, this can be another technique to make such data broadly available to researchers. Further research can expand the loss functions to include privacy metrics, and optimize other hyperparameters (eg, tree depth).

Limitations

Our empirical analysis was performed on oncology clinical trial datasets. There is no intrinsic reason to believe that the therapeutic area covered by a dataset would impact the results. The main driver of utility variation was the number of variables being synthesized. There will be nononcology clinical trial datasets with many variables, and nonclinical trial datasets which have many variables. Consequently, these findings should generalize outside oncology and to nonclinical trial data.

Three clinical trials (numbers 4, 5, and 6) had some records missing. They were all from the same sponsor, in which they consistently provided only 80% of the data. It is not clear whether this was a measure to protect patient privacy, or if there was another reason for these omissions (eg, these were screening failure patients). Given that we did not find the number of patients or observations to be a driver of the outcomes we were interested in, the impact on our results is expected to be limited.

Our utility metrics were not workload aware. Future studies could further evaluate the optimization strategy presented here on replications of published research analyses.

We did not examine the privacy risks from optimizing sequential synthesis. These risks can be controlled by adjusting the threshold in the loss function, and therefore privacy considerations would not take away from the importance of dealing with the arbitrary utility impacts of variable order.

CONCLUSIONS

There is growing demand to access clinical trial data. Data synthesis is one way to address this demand and simultaneously satisfy privacy concerns. The objective of this study was to evaluate and optimize the variable order for sequential synthesis of clinical trial data, with the loss being a measure of synthetic data utility. Sequential synthesis is commonly used for health and social science datasets, and it is suited to synthesizing small datasets such as clinical trial data. Furthermore, it can handle heterogeneous inputs and missingness well.

Through a simulation on oncology trial data we found that variable order affected utility, and more so with a higher number of variables in a dataset. Particle swarm optimization was able to identify variable permutations that ensure consistent data utility. Optimized sequential synthesis can provide a reliable way to synthesize clinical trial data.

FUNDING

This work was partially funded by a Discovery Grant RGPIN-2016-06781 (KEE) from the Natural Sciences and Engineering Research Council of Canada, and by Replica Analytics Ltd.

AUTHOR CONTRIBUTIONS

KEE contributed to designing the study, performed the analysis, and contributed to writing the article. LM contributed to designing the study, prepared the datasets for the study, implemented some of the code used to perform the study, and contributed to writing the article. CZ contributed to designing the study, prepared the datasets for the study, implemented some of the code used to perform the study, and contributed to writing the article.

ETHICS APPROVAL

This project was approved by the Children's Hospital of Eastern Ontario Research Institute Research Ethics Board, protocol number CHEOREB# 20/39X.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online

ACKNOWLEDGMENTS

This article is based on research using information obtained from the Project Data Sphere website (http://projectdatasphere.org), which is maintained by Project Data Sphere, LLC. Neither Project Data Sphere, LLC, nor the owner(s) of any information from the web site have contributed to, approved, or are in any way responsible for the contents of this article. This research was enabled in part by support provided by Compute Ontario (computeontario.ca) and Compute Canada (http://computecanada.ca). We also thank Fida Dankar for reviewing an earlier version of this manuscript.

CONFLICT OF INTEREST STATEMENT

This work was performed in collaboration with Replica Analytics Ltd. This company is a spin-off from the Children's Hospital of Eastern Ontario Research Institute. KEE is co-founder and has equity in this company. LM and CZ are data scientists employed by Replica Analytics Ltd.

REFERENCES

1. Ebrahim S, Sohani ZN, Montoya L, and, *et al*. Reanalyses of randomized clinical trial data. *JAMA* 2014; 312 (10): 1024–32.
2. Ferran J-M, Nevitt S. European Medicines Agency Policy 0070: an exploratory review of data utility in Clinical Study Reports for research. *BMC Med Res Methodol* 2019; 19 (1): 204.
3. Navar AM, Pencina MJ, Rymer JA, Louzao DM, Peterson ED. Use of open access platforms for clinical trial data. *JAMA* 2016; 315 (12): 1283.
4. PhRMA and EFPIA. Principles for Responsible Clinical Trial Data Sharing. 2013. https://www.phrma.org/en/Codes-and-guidelines/PhRMA-Principles-for-Clinical-Trial-Data-Sharing Accessed July 9, 2019.
5. TransCelerate BioPharma. De-identification and anonymization of individual patient data in clinical studies: a model approach. 2017. https://iapp.org/resources/article/data-de-identification-and-anonymization-of-individual-patient-data-in-clinical-studies-a-model-approach/ Accessed July 11, 2019.
6. TransCelerate Biopharma. Protection of personal data in clinical documents – a model approach. 2017. https://www.phusewiki.org/docs/WorkingGroups/TransCelerate/PhUSE%20Deliverables%20Protection%20of%20Personal%20Data%20in%20Clinical%20Documents-%20A%20Model%20Approach.pdf Accessed July 9, 2019.
7. 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 Accessed July 9, 2019.
8. 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* 2016; 164 (7): 505–6.
9. Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk. Washington, DC: Institute of Medicine; 2015.
10. National Academies of Sciences, Engineering, and Medicine, *Reflections on Sharing Clinical Trial Data: Challenges and a Way Forward: Proceedings of a Workshop*. Washington, DC: National Academies Press; 2020.
11. 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* 2018; 98: 157–9.
12. Naudet F, Sakarovitch 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* 2018; 360: k40. doi: 10.1136/bmj.k400.
13. Villain B, Dechartres A, Boyer P, Ravaud P. Feasibility of individual patient data meta-analyses in orthopaedic surgery. *BMC Med* 2015; 13 (1): 131.
14. 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* 2020; 20 (1): 113.
15. National Academy of Medicine. *Artificial Intelligence in Health Care*. Washington, DC: U.S. General Accountability Office. 2019.
16. El Emam KE, Jonker E, Moher E, Arbuckle L. A review of evidence on consent bias in research. *Am J Bioeth* 2013; 13 (4): 42–4.
17. de Montjoye Y-A, Hidalgo CA, Verleysen M, Blondel VD. Unique in the crowd: the privacy bounds of human mobility. *Sci Rep* 2013; 3 (1). doi: 10.1038/srep01376.
18. de Montjoye Y-A, Radaelli L, Singh VK, Pentland AS. Unique in the shopping mall: on the reidentifiability of credit card metadata. *Science* 2015; 347 (6221): 536–9.
19. Sweeney L, Yoo JS, Perovich L, Boronow KE, Brown P, Green Brody J. Re-identification risks in HIPAA Safe Harbor Data: a study of data from one environmental health study. *J Technol Sci* 2017; 2017: 2017082801.
20. Yoo JS, Thaler A, Sweeney L, Zang J. Risks to patient privacy: a re-identification of patients in Maine and Vermont Statewide Hospital Data. *J Technol Sci* 2018; 2018; 2018100901.
21. Sweeney L. Matching known patients to health records in Washington State Data. Harvard University. Data Privacy Lab. 2013. https://privacytools.seas.harvard.edu/publications/matching-known-patients-health-records-washington-state-data Accessed August 12, 2019.
22. Sweeney L, von Loewenfeldt M, Perry M. Saying it's anonymous doesn't make it so: re-identifications of 'anonymized' law school data. *J Technol Sci* 2018; 2018; 2018111301.
23. 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 Accessed March 23, 2020.
24. Bode K. Researchers find 'anonymized' data is even less anonymous than we thought. *Motherboard: Tech by Vice*. https://www.vice.com/en_ca/article/dygy8k/researchers-find-anonymized-data-is-even-less-anonymous-than-we-thought Accessed May 11, 2020.
25. Clemons E. Online profiling and invasion of privacy: the myth of anonymization. *Huffington Post*. https://www.huffpost.com/entry/internet-targeted-ads_b_2712586 Accessed May 11, 2020.
26. Jee C. You're very easy to track down, even when your data has been anonymized. *MIT Technology Review*. https://www.technologyreview.com/2019/07/23/134090/youre-very-easy-to-track-down-even-when-your-data-has-been-anonymized/ Accessed May 11, 2020.
27. 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 Accessed May 11, 2020.
28. Lomas N. Researchers spotlight the lie of 'anonymous' data. *TechCrunch*. https://techcrunch.com/2019/07/24/researchers-spotlight-the-lie-of-anonymous-data/ Accessed May 11, 2020.
29. Mitchell S. Study finds HIPAA protected data still at risks. *Harvard Gazette*. https://news.harvard.edu/gazette/story/newsplus/study-finds-hipaa-protected-data-still-at-risks/Accessed May 11, 2020.
30. Thompson SA, Warzel C. Twelve million phones, one dataset, zero privacy. *The New York Times*. https://www.nytimes.com/interactive/2019/12/19/opinion/location-tracking-cell-phone.html Accessed 05/11/2020.
31. Hern A. 'Anonymised' data can never be totally anonymous, says study. *The Guardian*. https://www.theguardian.com/technology/2019/jul/23/anonymised-data-never-be-anonymous-enough-study-finds Accessed May 11, 2020.
32. van der Wolk A. The (im)possibilities of scientific research under the GDPR. *Cybersecurity Law Report*. https://www.mofo.com/resources/insights/200617-scientific-research-gdpr.html Accessed July 3, 2020.
33. El Emam KE, Hoptroff R. The synthetic data paradigm for using and sharing data. *Cutter Executive Update* 2019; 19 (6).
34. El Emam K, Mosquera L, Hoptroff R. *Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data*. Sebastopol, MA: O'Reilly; 2020.
35. Reiter JP. New approaches to data dissemination: a glimpse into the future (?). *Chance* 2004; 17 (3): 11–5.
36. Polonetsky J, Renieris E. Privacy 2020: 10 privacy risks and 10 privacy technologies to watch in the next decade. Future of Privacy Forum. https://fpf.org/2020/01/28/privacy-2020-10-privacy-risks-and-10-privacy-enhancing-technologies-to-watch-in-the-next-decade/ Accessed July 27, 2020.
37. Park N, Mohammadi M, Gorde K, Jajodia S, Park H, Kim Y. Data synthesis based on generative adversarial networks. *Proc Vldb Endow* 2018; 11 (10): 1071–83.
38. Hu J. Bayesian estimation of attribute and identification disclosure risks in synthetic data. *arXiv*: 1804.02784; 2018.
39. Taub J, Elliot M, Pampaka M, Smith D. Differential correct attribution probability for synthetic data: an exploration. In: Domingo-Ferrer J, Montes F, eds. *Privacy in Statistical Databases*, PSD 2018. Lecture Notes in Computer Science, vol 11126. Cham: Springer; 2018: 122–37.
40. Hu J, Reiter JP, Wang Q. Disclosure risk evaluation for fully synthetic categorical data. In: Domingo-Ferrer J, ed. *Privacy in Statistical Databases: PSD 2014*.Lecture Notes in Computer Science, vol 8744. Cham: Springer; 2014: 185–99.
41. Wei L, Reiter JP. Releasing synthetic magnitude microdata constrained to fixed marginal totals. *Stat J IAOS* 2016; 32 (1): 93–108.
42. 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, PSD 2018. Lecture Notes in Computer Science, vol 11126. Cham: Springer; 2018: 59–74.
43. Reiter JP. Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study. *J R Stat Soc A* 2005; 168 (1): 185–205.
44. Breiman L, Friedman J, Stone C, Olshen R. *Classification and Regression Trees*. New York, NY: Taylor & Francis; 1984.
45. Reiter J. Using CART to generate partially synthetic, public use microdata. *J Offic Stat* 2005; 21 (3): 441–62.
46. Conversano C, Siciliano R. Incremental tree-based missing data imputation with lexicographic ordering. *J Classif* 2009; 26 (3): 361–79.
47. Conversano C, Siciliano R. Tree-based classifiers for conditional incremental missing data imputation. In: Zani S, Cerioli A, Riani M, Vichi M, eds. Data Analysis, Classification and the Forward Search. Studies in Classification, Data Analysis, and Knowledge Organization. Berlin, Heidelberg: Springer; 2002: 271–8.
48. Drechsler J, Reiter JP. An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Comput Stat Data Anal* 2011; 55 (12): 3232–43.
49. Arslan RC, Schilling KM, Gerlach TM, Penke L. Using 26,000 diary entries to show ovulatory changes in sexual desire and behavior. *J Pers Soc Psychol* 2018 Aug 27 [E-pub ahead of print]; doi: 10.1037/pspp0000208.
50. 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 Effect* 2019; 12 (4): 616–47.
51. Sabay A, Harris L, Bejugama V, Jaceldo-Siegl K. Overcoming small data limitations in heart disease prediction by using surrogate data. *SMU Data Sci Rev* 2018; 1 (3): 12.
52. Freiman M, Lauger A, Reiter J. Data synthesis and perturbation for the American Community Survey at the U.S. Census Bureau. Working paper. Washington, DC: U.S. Census Bureau; 2017.
53. Nowok B. Utility of synthetic microdata generated using tree-based methods. 2015. https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/20150/Paper_33_Session_2_-_Univ._Edinburgh__Nowok_.pdf Accessed July 9, 2019.
54. Raab GM, Nowok B, Dibben C. Practical data synthesis for large samples. *J Priv Confid* 2016; 7 (3): 67–97. doi: 10.29012/jpc.v7i3.407.
55. 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* 2017; 33 (3): 785–96.
56. Quintana DS. A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation. *eLife* 2020; 9: e53275. doi: 10.7554/eLife.53275.
57. Read J, Pfahringer B, Holmes G, Frank E. Classifier chains for multi-label classification. In: Buntine W, Grobelnik M, Mladenic D, Shawe-Taylor J, eds. *Machine Learning and Knowledge Discovery in Databases*. Berlin, Germany: Springer; 2009: 254–69.
58. Spyromitros-Xioufis E, Tsoumakas G, Groves W, Vlahavas I. Multi-target regression via input space expansion: treating targets as inputs. *Mach Learn* 2016; 104 (1): 55–98.
59. Chin-Cheong K, Sutter T, Vogt JE. Generation of heterogeneous synthetic electronic health records using GANs. In: proceedings of the Workshop on Machine Learning for Health (ML4H) at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019); 2019, doi: 10.3929/ethz-b-000392473.
60. Choi E, Biswal S, Malin B, Duke J, Stewart WF, Sun J. Generating multi-label discrete patient records using generative adversarial networks. *arXiv*: 1703.06490; 2017.
61. Goncalves A, Ray P, Soper B, Stevens J, Coyle L, Sales AP. Generation and evaluation of synthetic patient data. *BMC Med Res Methodol* 2020; 20 (1): 108.
62. Yan C, Zhang Z, Nyemba S, Malin BA. Generating electronic health records with multiple data types and constraints. *arXiv*: 2003.07904; 2020.
63. Read J, Pfahringer B, Holmes G, Frank E. Classifier chains: a review and perspectives. arXiv: 1912.13405; 2019.
64. Raab GM, Nowok B, Dibben C. Guidelines for producing useful synthetic data. *arXiv*: 1712.04078; 2017.
65. Center for Outcomes Research & Evaluation (CORE), Yale School of Medicine. The YODA Project. http://medicine.yale.edu/core/projects/yodap/rhbmp/index.aspx Accessed July 11, 2019.
66. CEO Life Sciences Consortium. Share, integrate & analyze cancer research data. Project Data Sphere. https://projectdatasphere.org/projectdatasphere/html/home Accessed July 9, 2019.
67. CSDR: Clinical Study Data Request. 2015. https://www.clinicalstudydatarequest.com/ Accessed November 26, 2015.
68. Vivli - Center for Global Clinical Research Data. https://vivli.org/ Accessed July 15, 2020.
69. Sarlomo-Rikala M, Kovatich AJ, Barusevicius A, Miettinen M. CD117: a sensitive marker for gastrointestinal stromal tumors that is more specific than CD34. *Mod Pathol* 1998; 11 (8): 728–34.
70. Amado RG, Wolf M, Peeters M, *et al*. Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. *J Clin Oncol* 2008; 26 (10): 1626–34.
71. Van Cutsem E, Peeters M, Siena S, *et al*. Open-label phase III trial of panitumumab plus best supportive care compared with best supportive care alone in patients with chemotherapy-refractory metastatic colorectal cancer. *J Clin Oncol* 2007; 25 (13): 1658–64.
72. Vermorken JB, Stöhlmacher-Williams J, Davidenko I, *et al*. Cisplatin and fluorouracil with or without panitumumab in patients with recurrent or metastatic squamous-cell carcinoma of the head and neck (SPECTRUM): an open-label phase 3 randomised trial. *Lancet Oncol* 2013; 14 (8): 697–710.
73. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. *J Comput GraphStat* 2006; 15 (3): 651–74.
74. El Emam K. Seven ways to evaluate the utility of synthetic data. *IEEE Secur Priv* 2020; 18 (4): 56–9.
75. Friedman J. On multivariate goodness-of-fit and two-sample testing. Stanford University. 2003. http://statweb.stanford.edu/~jhf/ftp/gof Accessed May 6, 2020.
76. Hediger S, Michel L, Näf J. On the use of random forest for two-sample testing. *arXiv*: 1903.06287; 2019.
77. Ian G, Pouget-Abadie J, Mriza M, *et al*. Generative adversarial nets. In: NIPS'14: proceedings of 27th the International Conference on Neural Imaging Processing Systems; 2014: 2672–80.
78. Wang Z, Myles P, Tucker A. Generating and evaluating synthetic UK primary care data: preserving data utility patient privacy. In: proceedings of the 32nd International Symposium on Computer-Based Medical Systems (CBMS); 2019, 126–131. doi: 10.1109/CBMS.2019.00036.
79. Le Cam L, Yang GL. *Asymptotics in Statistics: Some Basic Concepts*. New York, NY: Springer; 2000.
80. Gomatam S, Karr A, Sanil A. Data swapping as a decision problem. *J Offic Stat* 2005; 21 (4): 635–55.
81. Howe B, Stoyanovich J, Ping H, Herman B, Gee M. Synthetic data for social good. *arXiv*: 1710.08874; 2017.
82. Kaloskampis I. Synthetic data for public good. Office of National Statistics. 2019. https://datasciencecampus.ons.gov.uk/projects/synthetic-data-for-public-good/ Accessed March 2, 2020.
83. Beaulieu-Jones BK, Wu ZS, Williams C, *et al*. Privacy-preserving generative deep neural networks support clinical data sharing. *Circ Cardiovasc Qual Outcomes* 2019; 12 (7): e005122.
84. Bühlmann P, Hothorn T. Boosting algorithms: regularization, prediction and model fitting. *Stat Sci* 2007; 22 (4): 477–505.
85. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction, 1 Edition*. Oxford, United Kingdom: Oxford University Press; 2004.
86. Charrad M, Ghazzali N, Boiteau V, Niknafs A. NbClust: an R package for determining the relevant number of clusters in a data set. *J Stat Soft* 2014; 61 (6): 1–36.
87. Snoke J, Raab GM, Nowok B, Dibben C, Slavkovic A. General and specific utility measures for synthetic data. *J R Stat Soc A* 2018; 181 (3): 663–88.
88. Woo M-J, Reiter JP, Oganian A, Karr AF. Global measures of data utility for microdata masked for disclosure limitation. *J Priv Confid* 2009; 1 (1): 111–24. doi: 10.29012/jpc.v1i1.568.

89. McCaffrey DF, Griffin BA, Almirall D, Slaughter ME, Ramchand R, Burgette LF. A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Stat Med* 2013; 32 (19): 3388–414.
90. Bengio Y, Louradour J, Collobert R, Weston J. Curriculum learning. In: proceedings of the 26th Annual International Conference on Machine Learning; 2009: 41–8. doi: 10.1145/1553374.1553380.
91. Bonyadi MR, Michalewicz Z. Particle swarm optimization for single objective continuous space problems: a review. *Evol Comput* 2016; 25 (1): 1–54.
92. Poli R. Analysis of the publications on the applications of particle swarm optimisation. *J Artif Evol Appl* 2008; 2008: 685175.
93. Rosasco L, Vito ED, Caponnetto A, Piana M, Verri A. Are loss functions all the same? *Neural Comput* 2004; 16 (5): 1063–76.