Original Paper

Augmenting Insufficiently Accruing Oncology Clinical Trials Using Generative Models: Validation Study

Samer El Kababji[1,2], PhD; Nicholas Mitsakakis[2], MSc, PhD; Elizabeth Jonker[2], BA; Ana-Alicia Beltran-Bless[3], MD; Gregory Pond[4], PhD; Lisa Vandermeer[3], MSc; Dhenuka Radhakrishnan[2], MD; Lucy Mosquera[5], MSc; Alexander Paterson[6], MD; Lois Shepherd[7], MDCM, FRCP; Bingshu Chen[7], PhD; William Barlow[8], PhD; Julie Gralow[9], MD; Marie-France Savard[3], MD; Christian Fesl[10], PhD; Dominik Hlauschek[10], MSc; Marija Balic[11], MD, PhD; Gabriel Rinnerthaler[11], MD; Richard Greil[12], MD; Michael Gnant[13], MD; Mark Clemons[3], MD; Khaled El Emam[1,2], BEng, PhD

[1]School of Epidemiology and Public Health, Faculty of Medicine, University of Ottawa, Ottawa, ON, Canada

[2]CHEO Research Institute, Ottawa, ON, Canada

[3]Division of Medical Oncology, University of Ottawa, Ottawa, ON, Canada

[4]Department of Oncology, McMaster University, Hamilton, ON, Canada

[5]Generate Ops & Data Science, Aetion, Ottawa, ON, Canada

[6]Oncology, Alberta Health Services, Edmonton, AB, Canada

[7]Public Health Sciences, Queens University, Kingston, ON, Canada

[8]Biostatistics, University of Washington, Seattle, WA, United States

[9]Medical Oncology, University of Washington, Seattle, WA, United States

[10]Clinical Statistics, Austrian Breast & Colorectal Cancer Study Group (ABCSG), Vienna, Austria

[11]Division of Clinical Oncology, Medical University Graz, Graz, Austria

[12]Paracelsus Medical University Salzburg, Salzburg, Austria

[13]Comprehensive Cancer Center, Medical University of Vienna, Vienna, Austria

Corresponding Author:
Khaled El Emam, BEng, PhD
School of Epidemiology and Public Health
Faculty of Medicine
University of Ottawa
75 Laurier Ave E
Ottawa, ON, K1N 6N5
Canada
Phone: 1 6137975412
Email: kelemam@ehealthinformation.ca

Abstract

Background: Insufficient patient accrual is a major challenge in clinical trials and can result in underpowered studies, as well as exposing study participants to toxicity and additional costs, with limited scientific benefit. Real-world data can provide external controls, but insufficient accrual affects all arms of a study, not just controls. Studies that used generative models to simulate more patients were limited in the accrual scenarios considered, replicability criteria, number of generative models, and number of clinical trials evaluated.

Objective: This study aimed to perform a comprehensive evaluation on the extent generative models can be used to simulate additional patients to compensate for insufficient accrual in clinical trials.

Methods: We performed a retrospective analysis using 10 datasets from 9 fully accrued, completed, and published cancer trials. For each trial, we removed the latest recruited patients (from 10% to 50%), trained a generative model on the remaining patients, and simulated additional patients to replace the removed ones using the generative model to augment the available data. We then replicated the published analysis on this augmented dataset to determine if the findings remained the same. Four different generative models were evaluated: sequential synthesis with decision trees, Bayesian network, generative adversarial network, and a variational autoencoder. These generative models were compared to sampling with replacement (ie, bootstrap) as a simple alternative.

Replication of the published analyses used 4 metrics: decision agreement, estimate agreement, standardized difference, and CI overlap.

Results: Sequential synthesis performed well on the 4 replication metrics for the removal of up to 40% of the last recruited patients (decision agreement: 88% to 100% across datasets, estimate agreement: 100%, cannot reject standardized difference null hypothesis: 100%, and CI overlap: 0.8-0.92). Sampling with replacement was the next most effective approach, with decision agreement varying from 78% to 89% across all datasets. There was no evidence of a monotonic relationship in the estimated effect size with recruitment order across these studies. This suggests that patients recruited earlier in a trial were not systematically different than those recruited later, at least partially explaining why generative models trained on early data can effectively simulate patients recruited later in a trial. The fidelity of the generated data relative to the training data on the Hellinger distance was high in all cases.

Conclusions: For an oncology study with insufficient accrual with as few as 60% of target recruitment, sequential synthesis can enable the simulation of the full dataset had the study continued accruing patients and can be an alternative to drawing conclusions from an underpowered study. These results provide evidence demonstrating the potential for generative models to rescue poorly accruing clinical trials, but additional studies are needed to confirm these findings and to generalize them for other diseases.

Introduction

Background

Recruiting a sufficient number of patients for clinical trials is challenging [1], and the inability to recruit participants is the cause of failure for many clinical trials [2]. Approximately, 25% of clinical trials are discontinued before completion [3], with insufficient recruitment being the most frequent reason in 31% of the cases [4]. For adult cancer trials, between 20% and 50% fail to complete or were unable to reach recruitment goals [5-9]. This has been exacerbated by the recent pandemic where many trials experienced a considerable reduction in recruitment rates [10-13], which has continued after the pandemic [12]. While poor accrual is a problem in all trials, it is a greater problem in government (ie, academic) sponsored trials [14,15].

When a study is unable to recruit a sufficient number of patients, the study can be stopped, and the relevant analyses are performed on the available data. However, not reaching accrual targets results in underpowered analyses, and the smaller sample sizes increase the risk of unstable parameter estimates.

Patients have an expectation that their trial participation will lead to some advancement in knowledge that can be beneficial to the community [16], but many enroll in trials that do not answer the primary question adequately [15]. They are, therefore, enrolled in a study and exposed to toxicity and additional costs, with limited scientific benefit, which is considered unethical [17]. In addition to the wasted resources, it also means that those resources were not used for other studies that could have produced useful results.

Data augmentation is one approach to address insufficient accrual by either using real-world data (RWD) or by simulating additional observations.

RWD can be used for matched controls [18] where patient data from external sources are used instead of recruiting patients to the trial itself. In such a case, previous similar trials, registries, or eHealth record datasets on patients under the standard of care are matched to the treatment arm patients, and the matched patients' data are used as the control arm [19,20]. Such an approach with external controls has been used for running single-arm oncology trials [21,22]. However, external controls are challenging for a number of reasons [18]. First, the patients from RWD may have different observed and unobserved characteristics than the treatment arm patients, despite the use of matching. Second, unaccounted for environmental factors, such as seasonal effects, may lead to outcome differences. Third, changes in medical practice may have occurred over time and since the external control data were collected. Fourth, there may be measurement differences between the treatment and external controls resulting in the pooling of incompatible datasets, even for objective metrics. Fifth, if there are no adequate matches in the external data for some of the treatment arm patients, then these treatment arm patients may need to be dropped resulting in loss of valuable data. Sixth, the outcome variables need to be available in the external control dataset to allow a comparison, which can be challenging for surrogate end points or patient-reported outcomes. Finally, and specific to our context, insufficient accrual would occur in all arms in a study and not only in the controls; therefore, external controls would not address our problem.

Augmentation through simulation can be a potential solution when there are insufficient data and is a common practice for imaging data [23-25] and time series data [26,27]. In the case of cross-sectional RWD, augmentation methods, such as sampling with replacement (henceforth referred to as *bootstrap*), and generative models, such as sequential synthesis using decision trees, generative adversarial networks (GANs), and variational autoencoders (VAEs), have been evaluated with encouraging results [28-32]. Augmentation methods have also been applied to small clinical trial datasets as a first step in synthetic data generation [33,34]. One study used a VAE

generative model to simulate additional patients as a mechanism to design smaller clinical trials [35].

Objective

In this study, we therefore adopt augmentation through simulation and expand on this body of work by more comprehensively evaluating multiple types of generative models and on a larger number of clinical trials. We make 2 hypotheses that we evaluate in this work as follows: (1) patients recruited early in a clinical trial are similar on the treatment effect to patients recruited later in a clinical trial and (2) because of hypothesis 1, we can train a generative model on early patients to simulate the remaining patients in insufficiently accruing trials to reach target recruitment and replicate the results of the original study that reached target recruitment.

Specifically, generative models are used to augment breast cancer clinical trials that do not reach target recruitment. We start with datasets from 9 completed breast cancer clinical trials and simulate different levels of insufficient accrual, and in each case, use generative machine learning models to simulate patients to compensate for the insufficient accrual. We then replicate the analyses of the published studies using the augmented datasets to determine if they produce similar findings as if the target number of patients were actually recruited.

Methods

Overview

Data augmentation methods using generative models were applied on 9 breast cancer clinical trial datasets. Insufficient accrual was simulated and augmentation then applied to compensate for that. The question was whether this can produce similar findings to the published analyses with the full data.

Datasets

The clinical trials that were included are summarized in Table 1, with further details described in Table S1 in Multimedia Appendix 1 [36-38]. The Rethinking Clinical Trials (REaCT) were supported by the REaCT program at the Ottawa Hospital [39]. The remaining datasets included a larger number of patients and multiple sites.

Table 2 shows the countries that the patients were recruited from. The studies spanned multiple jurisdictions in North America, Europe, Australia, and South Africa.

Table 1. Key features of the clinical trial datasets used in this study.

Dataset	NCT identifier	Participants, N	Variables in dataset[a], n	Strata info?	Control arm	Patients in the control arm, n (%)	Treatment arms considered	Patients in the treatment arm, n (%)	Enrollment period
REaCT[b]-ILI-AD	NCT02861859	218	8	True	Placebo	105 (48)	Active olanzapine	113 (52)	December 2016 to June 2019
REaCT-BTA[c]	NCT02721433	230	8	True	4 weekly BTA	118 (51)	12 weekly BTA	112 (49)	August 3, 2016, to June 5, 2018
CCTG[d] MA27	NCT00066573	7576	25	True	Anastrozole	3787 (50)	Exemestane	3789 (50)	June 2, 2003, to July 31, 2008
NSABP[e] B34	NCT00009945	3310	49	True	Placebo	1656 (50)	Clodronate	1654 (50)	January 22, 2001, to March 31, 2004
REaCT-G/G2	NCT02428114 and NCT02816164	401	10	True	7 or 10 d of granulocyte colony stimulating factor	248 (62)	5 d of granulocyte colony stimulating factor	153 (38)	May 2015 to September 2018
REaCT-HER2+[f]	NCT02632435	50	47	True	Peripherally inserted central catheter	26 (52)	PORT; totally implanted vascular access device	24 (48)	March 2016 to March 2018
ABCSG[g]-12	NCT00295646	1803	35	False	Tamoxifen[h]	900 (50)	Anastrozole[h]	903 (50)	1999 to 2006
REaCT-ZOL[i]	NCT03664687	211	11	False	__j	—	—	—	November 1, 2018, to April 2, 2020
SWOG[k] 0307[l]	NCT00127205	6018	23	False	Clodronate	2268 (38)	Zoledronic acid	2262 (38)	January 2006 to February 2010

[a]This is the total number of variables that were included in the generative models or bootstrap.

[b]REaCT: Rethinking Clinical Trials.

[c]BTA: bone-targeted agents.

[d]CCTG: Canadian Cancer Trials Group.

[e]NSABP: National Surgical Adjuvant Breast and Bowel Project.

[f]HER2+: human epidermal growth factor receptor-2 positive.

[g]ABCSG: Austrian Breast and Colorectal Cancer Study Group.

[h]Tamoxifen represents the arms of Nolvadex/control and Nolvadex/zoledronate in the clinical trial while anastrozole represents the arms of Arimidex/control and Arimidex/zoledronate.

[i]ZOL: zoledronate.

[j]Not applicable.

[k]SWOG: Southwest Oncology Group.

[l]Initially, the trial included 3 arms; however, only the indicated 2 arms had patients assigned to them throughout the duration of the study. Moreover, the available data for our study did not include any randomization codes, and as such, the original primary analysis comparing the outcomes between the 2 arms could not be replicated. Instead, we compared the 5-year survival probabilities between those with negative and positive or equivocal HER2 status. The estimate of the difference of survival probabilities and SE were produced.

Table 2. Countries of recruitment for the clinical trials used in this study.

Dataset	Countries of recruitment
REaCT[a]-ILIAD	Canada
REaCT-BTA[b]	Canada
CCTG[c] MA27	Australia, Canada, Hungary, Italy, Puerto Rico, South Africa, Switzerland, and the United States
NSABP[d] B34	United States
REaCT-G/G2	Canada
REaCT-HER2+[e]	Canada
ABCSG[f]-12	Austria and Germany
REaCT-ZOL[g]	Canada
SWOG[h]0307	Canada and the United States

[a]REaCT: Rethinking Clinical Trials.

[b]BTA: bone-targeted agents.

[c]CCTG: Canadian Cancer Trials Group.

[d]NSABP: National Surgical Adjuvant Breast and Bowel Project.

[e]HER2+: human epidermal growth factor receptor-2 positive.

[f]ABCSG: Austrian Breast and Colorectal Cancer Study Group.

[g]ZOL: zoledronate.

[h]SWOG: Southwest Oncology Group.

Ethical Considerations

This study was a secondary analysis of datasets from already completed clinical trials. The secondary analysis was approved by the Children's Hospital of Eastern Ontario Research Ethics Board (protocol: 23/47X) and the Ontario Cancer Research Ethics Board (project ID: 3749).

Nonmonotonic Treatment Effect Size Hypothesis

We will use the terms "early participants" to indicate participants who were recruited in the earlier stages of a study and "late participants" to indicate those who were recruited in later stages of a study. For early participants to be good candidates for training a generative model that can be used to simulate late participants, there should not be a systematic difference in the estimated effect size between these 2 groups.

Estimated effect sizes tend to vary as patients are recruited and converge to the true value as more information is collected [40]. Instability of estimates at small sample sizes is a contributing factor. This means that training a generative model on earlier patients may enable the simulation of realistic patients that are representative of those that would be recruited later in the trial if the effect over time is not systematic.

However, existing sites gain experience with conducting a trial and this may result in process adjustments along that learning curve that may have an impact on the outcome. This can also happen, for example, when there is treatment effect heterogeneity whereby some patient characteristic (eg, disease severity or age) interacts with the intervention and more patients at one end of the severity or age scale are recruited earlier in the study [41]. One simulation demonstrated a monotonic change

in effect size as more patients are recruited [42]. In that example, earlier patients had a meaningfully different estimated effect size compared to later patients with a trend over time. In such a case, training a generative model on earlier patients may not produce simulated patients that are representative of the later patients.

Therefore, it is an empirical question whether such a monotonic effect can be observed in practice.

As can be seen in Table 1, our 9 studies were conducted over an extended period. If the studies were very short, then there is a higher likelihood that early and late participants would be similar. However, the extended enrollment periods suggest that there was ample opportunity for the characteristics of the participants to change over time, as well as adjustments to the trial processes to occur as study staff gain more experience with time.

To test the hypothesis that early and late participants are similar on the estimated effect size, the monotonic relationship of the effect size and order of participant recruitment was examined with a regression model of treatment and recruitment order main effects and their interaction. Point estimates and 95% CIs of the interaction term were obtained, indicating statistical significance at $P<.05$ if they do not include 0. This investigation was not conducted for the trials where the main analysis did not use a statistical model with treatment as predictor (ie, REaCT-zoledronate [ZOL], Southwest Oncology Group 0307, and REaCT–human epidermal growth factor receptor-2 positive [HER2+]). If the effect size was not monotonic, then the interaction term would not be statistically significant.

Generative Modeling Methods

Overview

We used 4 common machine learning–based generative modeling methods for structured tabular data to synthesize the analysis datasets from the clinical trials under investigation. These methods include sequential synthesis using decision trees, Bayesian networks, GAN, and VAE. The last 3 methods were implemented as an adaptation of an open-sourced Python package *Synthcity* [43]. Our implementation, which is publicly available (*Data Availability* section), provides further preprocessing and postprocessing on top of *Synthcity*. No further hyperparameter tuning was performed beyond what was available in the *Synthcity* implementation. For each generative model, the number of variables indicated in Table 1 were synthesized.

In addition, we used the "bootstrap" technique for comparison as a baseline. Bootstrap is simply sampling with replacement from the training data to add the missing patients.

Sequential Decision Trees

Similar to using a chaining method for multilabel classification problems, sequential decision trees generate synthetic data using conditional trees in a sequential fashion [44-46]. It has been commonly used in the health care and social-science domains for data synthesis [28,47-54]. The details of the implementation procedures are described elsewhere [44].

Bayesian Networks

Bayesian networks are models based on directed acyclic graphs that consist of nodes representing the random variables and arcs representing the dependencies among these variables. To construct the Bayesian networks model, the first step is to find the optimal network topology and then to estimate the optimal parameters [55]. Starting with a random initial network structure, the Hill Climb heuristic search is used to find the optimal structure. Then, the conditional probability distributions are estimated using the maximum a posteriori estimator [56]. Once the network structure and the parameters are estimated, we can initialize the nodes with no incoming arcs by sampling from their marginal distributions and predict the rest of the connected variables using the estimated parameters.

Conditional GAN

A basic GAN consists of 2 artificial neural networks, a generator and a discriminator [57]. The generator and the discriminator play a min-max game. The input to the generator is noise, while its output is synthetic data. The discriminator has 2 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. In contrast, the discriminator is trained to maximize its discriminatory capability.

Among all the variations of GAN architectures, the conditional tabular GAN (CTGAN) is often used in tabular data synthesis [58]. CTGAN builds on conditional GANs by addressing the multimodal distributions of continuous variables and the highly imbalanced categorical variables [59]. 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.

Variational Autoencoder

VAEs use artificial neural networks and involve 2 steps (ie, encoding and decoding) to generate new samples [60]. First, an encoder is generated to compress input data into a lower-dimensional latent space, in which the data points are represented by distributions. The second step is a decoding process, in which new data samples are reconstructed as output from the latent space. The neural network is optimized by minimizing the reconstruction loss between the output and the input. VAEs are known to generate complex data of various types due to their ability to learn more complex distributions and relationships [61]. Many variants have been proposed as an extension of VAE, such as triplet-based VAE [62], conditional VAE [63], and Gaussian VAE [64]. In particular, the tabular VAE was proposed as an adaptation of standard VAE to model and generate mixed-type tabular data with a modified loss function [59].

Augmentation

The augmentation procedure ensured that the augmented data had the same number of patients as the original clinical trial. Therefore, there is no difference in size between the published study and the augmented datasets used in our analyses.

Figure 1 illustrates the main steps of generating the augmented datasets for the trials given in Table 1. A trial's *original* dataset, with N number of patients, was first reduced by *r*. The variable *r* signifies the fraction of the *last* patients who were deliberately removed from the input dataset. This results in a *reduced* dataset with "(1-r) N" patients. In practice, the reduced dataset represents a poorly accruing clinical trial that needs to be rescued. The shaded area outlines the typical steps taken by a practitioner during the implementation process.

In our study, the *r* value was varied incrementally from 0.1 to 0.5 in steps of 0.1. For instance, a 0.2 *r* value indicates that the last 20% of patients were deliberately removed from the original trial dataset. The reduced dataset was then used to train a generative model. In case of bootstrapping, the reduced dataset was used for sampling with replacement. After training, the necessary samples for augmentation were generated from the trained model and concatenated onto the reduced dataset. To account for stochasticity, and as depicted by the dotted lines in the Figure 1, ten versions of the synthetic data were generated leading to multiple versions of the augmented datasets. Subsequently, each of the augmented datasets were analyzed in the same way as the published analysis. The 10 analyses results were then combined to obtain a single augmented dataset result, and that was compared with the results obtained from the original dataset (as described in the subsequent section).

Figure 1. Augmentation of clinical trial datasets using the generative model.
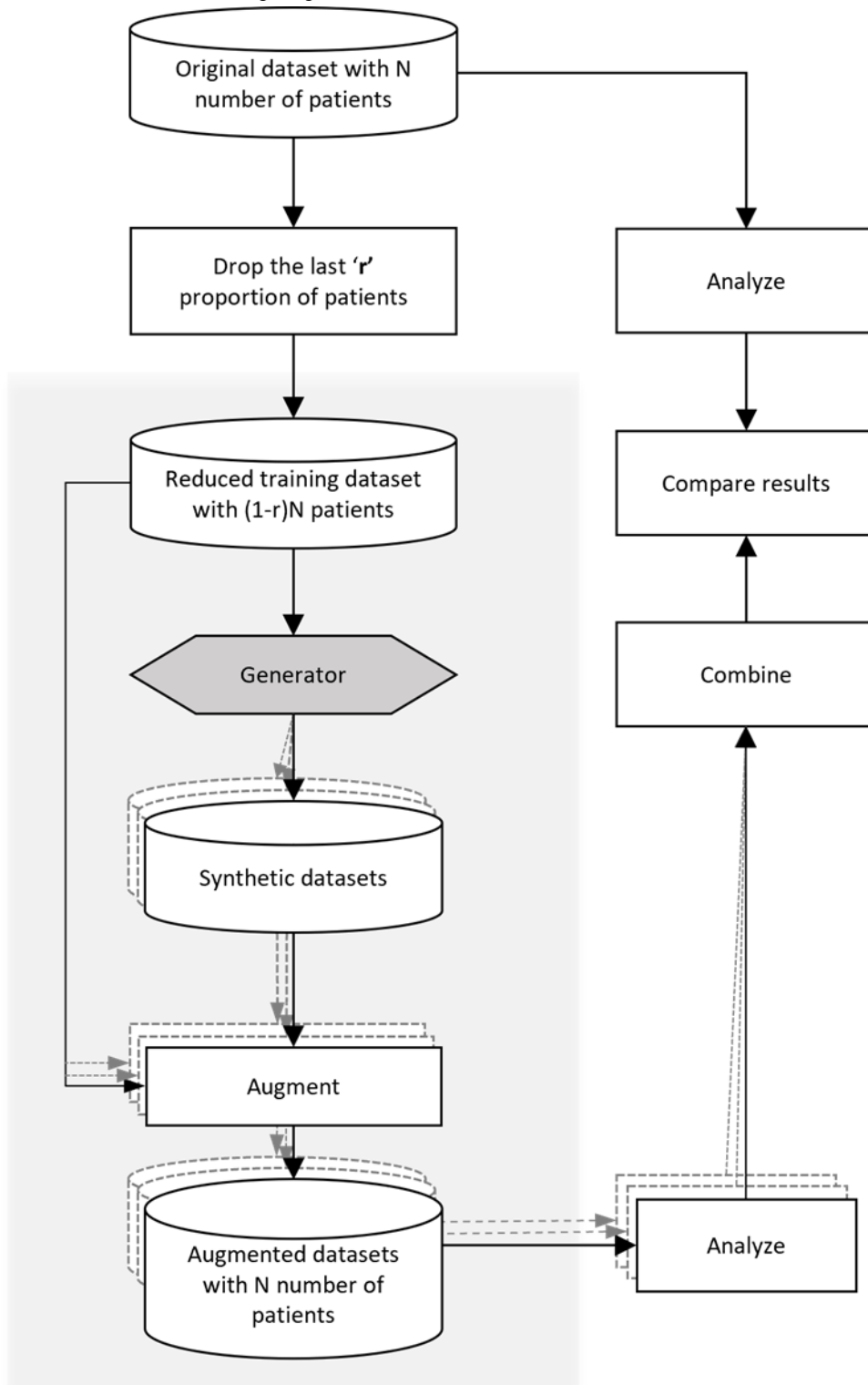
Some of the trials had implemented stratified randomization, and in those cases, we modified the basic process to accommodate stratification, for example, consider that as per the predefined protocol, the original dataset is stratified by 2 variables, the "Cancer Type" and the "Site Number." Rejection sampling was used to draw from the generated datasets to achieve the desired strata proportions along these 2 dimensions, in addition to the appropriate numbers of patients in each arm of the study.

Combining Rules

The original proposal for synthetic data generation treated it as a form of multiple imputation [65]. Under the multiple imputation model, multiple datasets, say m, are synthesized and analyzed. The combining rules are used to compute the parameter estimates and variances across the analysis results from the m synthetic datasets [66-68]. Such corrections for the parameter estimates and variances ensured that variability introduced by the generative process are accounted for when

estimating parameters and making population inferences from synthetic datasets.

Once the generative model was trained on the underlying distribution of the input data, it was used to create m synthetic datasets of size $r \times N$, and then each was added to that training dataset to augment it. This resulted in m versions of augmented datasets, each containing N records.

The m augmented datasets were analyzed using the same methodology applied to the original data in the relevant publications. This analysis yielded estimated parameters for each of the augmented versions. Subsequently, these estimated parameters were combined in accordance with the following partial synthesis rules.

For a particular model parameter q_i with variance v_i using synthetic dataset i where $i=1...m$. The adjustment for the model parameters and variances is as follows [52,68,69]. The combined model parameter $\bar{q}_m$. is the mean across the m model parameters from the synthetic datasets $\bar{q}_m = 1/m \sum_i q_i$, and $\bar{v}_m$ is the mean variance across the m model parameters from the synthetic datasets where $\bar{v}_m = 1/m \sum_i v_i$. The between imputation variance is given by $b_m = \frac{1}{m-1} \sum_{i=1}^{m} (q_i - \bar{q}_m)^2$, the adjusted variance is computed as $T_p = b_m/m + \bar{v}_m$, and the adjusted large sample 95% CI of the model parameter is computed as $\bar{q}_m \pm 1.96 \sqrt{T_p}$. For this study, we set $m=10$, which is consistent with current practice for the analysis of synthetic data [52,68-70] and has been recommended based on a recent simulation [71].

Augmentation Fidelity

For the generated datasets, we evaluated the fidelity relative to the training dataset. Fidelity indicates the extent to which the distributions of generated data deviate from the training data. For example, if $r=0.1$, then 90% of the dataset is used for training the generative model, and 10% is generated. We assessed the fidelity using the Hellinger distance [72], which has the advantage of being interpretable as it varies from 0 to 1. The Hellinger distance is averaged across the 10 generated datasets.

Evaluation of Study Replicability

We evaluated the replicability of the analysis results using the augmented datasets. Replicability is the reliability of findings when an existing study is repeated using the same analytical methods but different data [73]. We assessed it by comparing the published analysis results using the real datasets for these clinical trials with the results of the same analysis performed on the partially synthetic (ie, augmented) data. The details of the published analyses that were replicated are summarized in Multimedia Appendix 1.

For each clinical trial, a model was fitted to obtain a parameter estimate and its SE. For instance, if generalized estimating equations [36] were used for modeling the real data, an estimate would be the coefficient associated with the selected predictor as per the published study. We obtained a value for the estimate and its 95% CI. The same was applied to each of the m versions of the augmented data. We combined the results from all the augmented versions using the combining rules discussed above. Subsequently, the estimates and CIs of the original and augmented data were compared in terms of the *estimate agreement*, the *decision agreement*, *standardized difference*, and the *CI overlap*. These criteria have been used in the literature to assess the replicability of analyses using synthetic data [71,74]. The criteria are defined in Textbox 1.

Textbox 1. Criteria to assess the replicability of analyses using synthetic data.

- *Estimate agreement*: It is a Boolean indicator of whether the estimate produced by the augmented data is within the 95% CI produced by the real data. This requires that an augmented data effect estimate be within the range of plausible values for the true effect based on evidence from the real data. Under the assumption that the parameter variances are equal between the real and augmented datasets, estimate agreement is expected 83% of the time under no bias [75].

- *Decision agreement*: It is a Boolean indicator of whether the same conclusion is drawn from the real and augmented data estimates. This means that the augmented data estimates have the same direction and statistical significance as the real data. The decision agreement does not apply if the analysis is descriptive. We would expect decision agreement to occur at a rate equal to power, which would be at least 80% of the time (ie, assuming the 9 trials are powered by design for at least 80%) [75].

- *Standardized difference*: It is a Boolean indicator of whether the difference in the parameter estimate between real and augmented data is consistent with the null hypothesis of no difference [75]. The Z value is computed and compared with the standard normal ($|Z| \leq 1.96$).

- *CI overlap*: It is a proportion of the real and augmented data parameter CIs overlap [76], which is a commonly used synthetic data utility metric. We would want this to be as close to 100% as possible but set 80% as a minimal value.

The 2 agreement metrics are consistent with previous measures of replicability [77-79], have been used to compare RWD analysis results against a clinical trial reference [75,80-83], and have been used to assess the replicability of psychological studies [77].

Results

The first set of results are shown in Table 3 where the monotonic relationship of the effect size over time is investigated. In none of the trials used for the study was the interaction term (ie, recruitment order by treatment) found to be statistically significant, indicating lack of evidence of monotonically varying treatment effect with respect to the order of participant recruitment.

Table 3. Evaluation of monotonic relationships over recruitment order in the estimated effect size.

Trial	Main effect (95% CI)	Interaction effect (95% CI)
REaCT[a]-ILIAD	0.52 (0.023 to 1.01)	-0.00000371 (-0.0000116 to 4.16×10^{-6})
REaCT-BTA[b]	-1.85 (-4.75 to 1.04)	0.011 (-0.032 to 0.0542)
CCTG[c] MA27	0.033 (-0.117 to 0.18)	-0.0000135 (-0.0000854 to 5.84×10^{-5})
NSABP[d] B34	-0.036 (-0.18 to 0.11)	8.28×10^{-5} (-0.000072 to 0.00024)
REaCT-G/G2	0.015 (-0.0019 to 0.032)	6.6×10^{-6} (-0.000005 to 1.85×10^{-5})
ABCSG[e]-12 (tamoxifen vs anastrozole)	0.11 (-0.13 to 0.35)	7.36×10^{-5} (-0.0004 to 0.00054)
ABCSG-12 (zoledronic acid vs no zoledronic acid)	-0.24 (0.48 to 0.002)	6.69×10^{-5} (-0.00041 to 0.00054)

[a]REaCT: Rethinking Clinical Trials.

[b]BTA: bone-targeted agents.

[c]CCTG: Canadian Cancer Trials Group.

[d]NSABP: National Surgical Adjuvant Breast and Bowel Project.

[e]ABCSG: Austrian Breast and Colorectal Cancer Study Group.

The results on the data with insufficient accrual (ie, the "reduced" datasets with no augmentation) for 2 values of r at 0.2 and 0.5 are shown in Table 4.

We can make 4 observations from Table 4:

1. For the REaCT-HER2+ study, which had the smallest sample size, there is no decision agreement as r increases. This is not surprising as the low sample sizes would mean unstable parameter estimates and larger CIs. The small sample size also explains the lack of difference in the standardized difference comparison.
2. The REaCT-ILIAD study had a statistically significant result with the full data. Decision agreement is no longer attained as r increases due to the smaller sample sizes, lower power, and hence, wider CIs.
3. The Austrian Breast and Colorectal Cancer Study Group (ABCSG)-ZOL analysis had a marginal nonsignificant outcome in the full trial. At small r values, small changes can have an impact on the statistical significance of the results, but that becomes less of an issue for large r values as the results will no longer be marginal with wider CIs. Although, one would expect that if the original results were marginally statistically significant, there would not be decision agreement even for larger values of r.
4. For the remaining trial analyses, the results were not significant in the original data. Therefore, the decision agreement would not be affected with the lower power as r increases, and the parameter estimates retained the same direction as the full trial. The standardized difference comparison indicates that the parameters were not different even as r increases, which is consistent with the inability to detect a monotonic effect with recruitment order presented above, and detecting a difference becomes more difficult as the sample size decreases.

These results indicate that drawing conclusions from reduced datasets can be misleading and can produce incorrect findings relative to those that would be obtained if target recruitment was achieved. In addition, the nature of any error would not be known a priori.

The fidelity results for the augmented datasets are shown in Figure 2. At low values of r, the generated datasets were small, making fidelity comparisons unstable. For the smaller datasets, such as REaCT-HER2+, REaCT–bone-targeted agents (BTA), REaCT-ZOL, and REaCT-ILIAD, the Hellinger distance values were highest but still relatively low on an absolute scale. For the other datasets, the Hellinger values were quite small with variation in a very narrow range, demonstrating high fidelity.

Table 4. The baseline primary results for the original datasets, and for the reduced datasets at 2 different values of *r* (ie, 0.2 and 0.5)

Trial name	Trial short name	Full dataset			r=0.2						r=0.5					
		Effect size (SE)	Variables used in the analysis	Analysis method	Sample size, N	Effect size (SE)	Estimate agreement	Decision agreement	Standardized difference	CI overlap	Sample size, N	Effect size (SE)	Estimate agreement	Decision agreement	Standardized difference	CI overlap
REaCT[a]-ILIAD	ILIAD [84]	0.52 (0.25)	8	GEE[b]	174	0.64 (0.28)	1	1	1	0.88	109	0.39 (0.36)	1	0	1	0.85
REaCT-BTA[c]	BTA [85]	−1.85 (1.48)	8	LM[d]	184	−2.54 (1.63)	1	1	1	0.89	115	−3.46 (2.03)	1	1	1	0.79
CCTG[e] MA27	CCTG [86]	0.033 (0.076)	6	Cox[f]	6060	0.045 (0.084)	1	1	1	0.96	3788	−0.006 (0.102)	1	1	1	0.87
NSABP[g] B34	NSABP [87]	−0.036 (0.076)	6	Cox	2648	−0.034 (0.085)	1	1	1	0.95	1655	−0.124 (0.108)	1	1	1	0.78
REaCT-G/G2	G/G2 [88]	0.015 (0.009)	10	GEE	320	0.012 (0.010)	1	1	1	0.91	200	0.0005 (0.016)	1	1	1	0.76
REaCT-HER2+[h]	HER2+ [89]	0.0058 (0.0062)	3	GLM[i]	40	0.0075 (0.006)	1	1	1	0.93	25	0.017 (0.009)	1	0	1	0.62
ABC-SG[j]-12 (tamoxifen vs anastrozole)	ABC-SG [90]	0.11 (0.12)	3	Cox	1442	0.13 (0.14)	1	1	1	0.96	901	0.07 (0.17)	1	1	1	0.87
ABCSG-12 (ZOL[k] vs no ZOL)	ABC-SG-ZOL [90]	−0.24 (0.1237)	3	Cox	1442	−0.27 (0.14)	1	0	1	0.95	901	−0.26 (0.17)	1	1	1	0.87
REaCT-ZOL	ZOL [91]	58.98[l] (0.76)	2	*t* test	168	58.07 (0.89)	1	—	1	0.72	105	58.20 (1.10)	1	—	1	0.81
SWOG[m] 0307	SWOG [92]	−0.0087 (0.0145)	3	Surv[n]	4814	4.3×10^{-5} (0.016)	1	1	1	0.86	3009	−0.0096 (0.02)	1	1	1	0.86

[a]REaCT: Rethinking Clinical Trials.

[b]GEE: general estimating equations.

[c]BTA: bone-targeted agents.

[d]LM: linear model.

[e]CCTG: Canadian Cancer Trials Group.

[f]Cox: Cox regression

[g]NSABP: National Surgical Adjuvant Breast and Bowel Project.

[h]HER2+: human epidermal growth factor receptor-2 positive

[i]GLM; general linear model.

[j]ABCSG: Austrian Breast and Colorectal Cancer Study Group.

[k]ZOL: zoledronate.

[l]No effect was estimated for REaCT-ZOL—the figures shown correspond to descriptive analysis.

[m]SWOG: Southwest Oncology Group.

[n]Surv: difference in survival probabilities.

Figure 2. Hellinger distance results by comparing the training dataset with the generated dataset. The values were averaged across all generated 10 datasets. A value of 0 means that the 2 datasets are the same and of 1 indicates maximum difference. Note that the y-axis scales are not the same across all plots to provide better readability. ABCSG: Austrian Breast and Colorectal Cancer Study Group; BTA: bone-targeted agents; CCTG: Canadian Cancer Trials Group; CTGAN: conditional tabular generative adversarial network; HER2+: human epidermal growth factor receptor-2 positive; NSABP: National Surgical Adjuvant Breast and Bowel Project; REaCT: Rethinking Clinical Trials; SWOG: Southwest Oncology Group; TVAE: tabular variational auto encoder; ZOL: zoledronate.
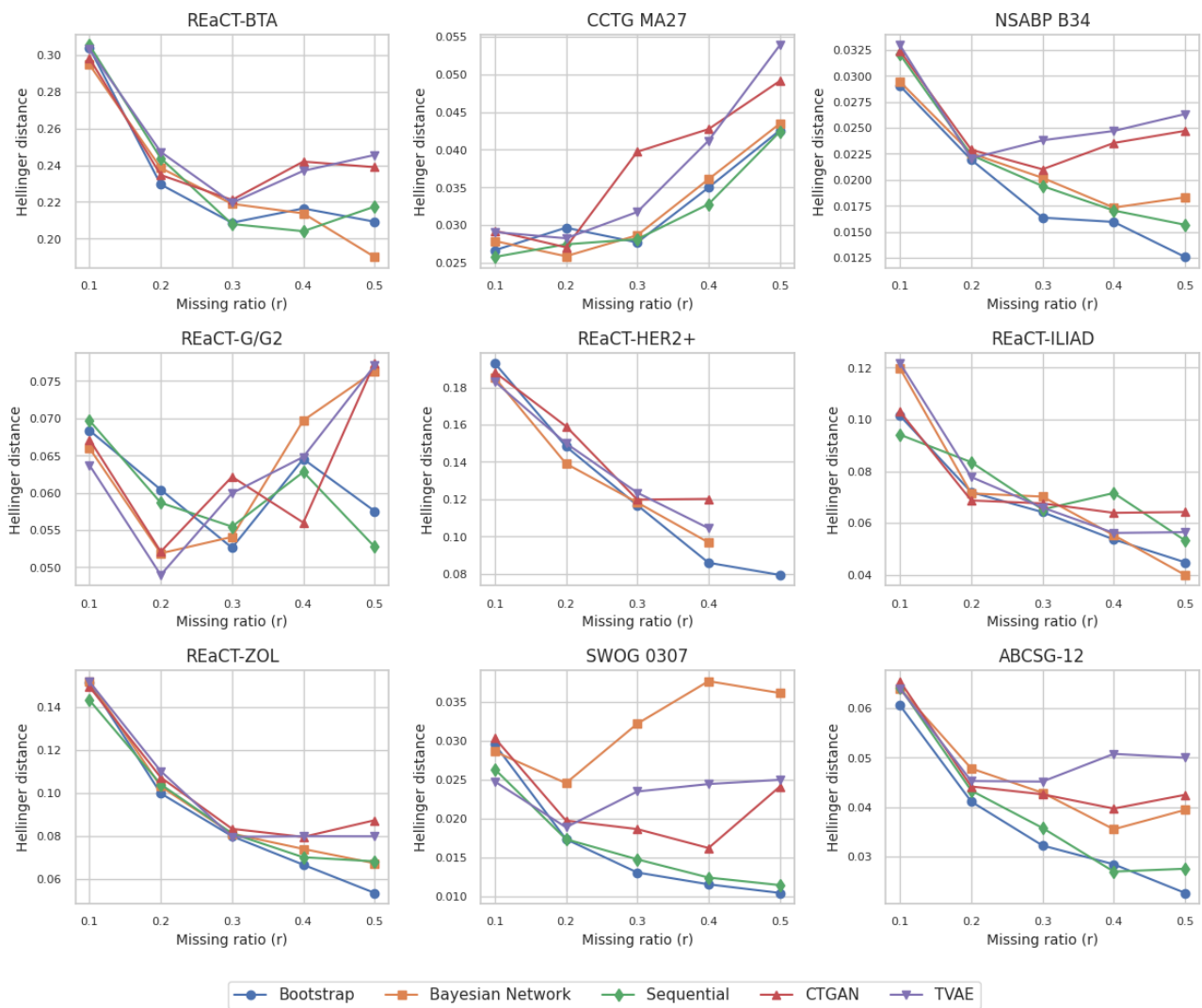
Another representation of fidelity is shown in Figure 3, where we compared the training dataset with a generated dataset of the same size. Here, we can see high fidelity values for all datasets as these results were less affected by small dataset sizes. As the training dataset size decreased with higher *r*, the fidelity decreased, but the changes in fidelity were modest.

All the detailed replicability results for all values of *r* are provided in Table S2 in Multimedia Appendix 1.

The first general observation is that the bootstrap and sequential synthesis tend to perform best among all the generative models. Therefore, in Figure 4 we only show their results across all datasets that were examined for the 4 measures of replicability. Both types of generative models achieved high estimate and decision agreement and maintained a CI overlap above 80% even as *r* values approached 0.5.

Figure 3. Hellinger distance results by comparing the training dataset with a generated dataset of the same size. The values were averaged across all generated 10 datasets. A value of 0 means that the 2 datasets are the same and of 1 indicates maximum difference. Note that the y-axis scales are not the same across all plots to provide better readability. ABCSG: Austrian Breast and Colorectal Cancer Study Group; BTA: bone-targeted agents; CCTG: Canadian Cancer Trials Group; CTGAN: conditional tabular generative adversarial network; HER2+: human epidermal growth factor receptor-2 positive; NSABP: National Surgical Adjuvant Breast and Bowel Project; REaCT: Rethinking Clinical Trials; SWOG: Southwest Oncology Group; TVAE: tabular variational auto encoder; ZOL: zoledronate.
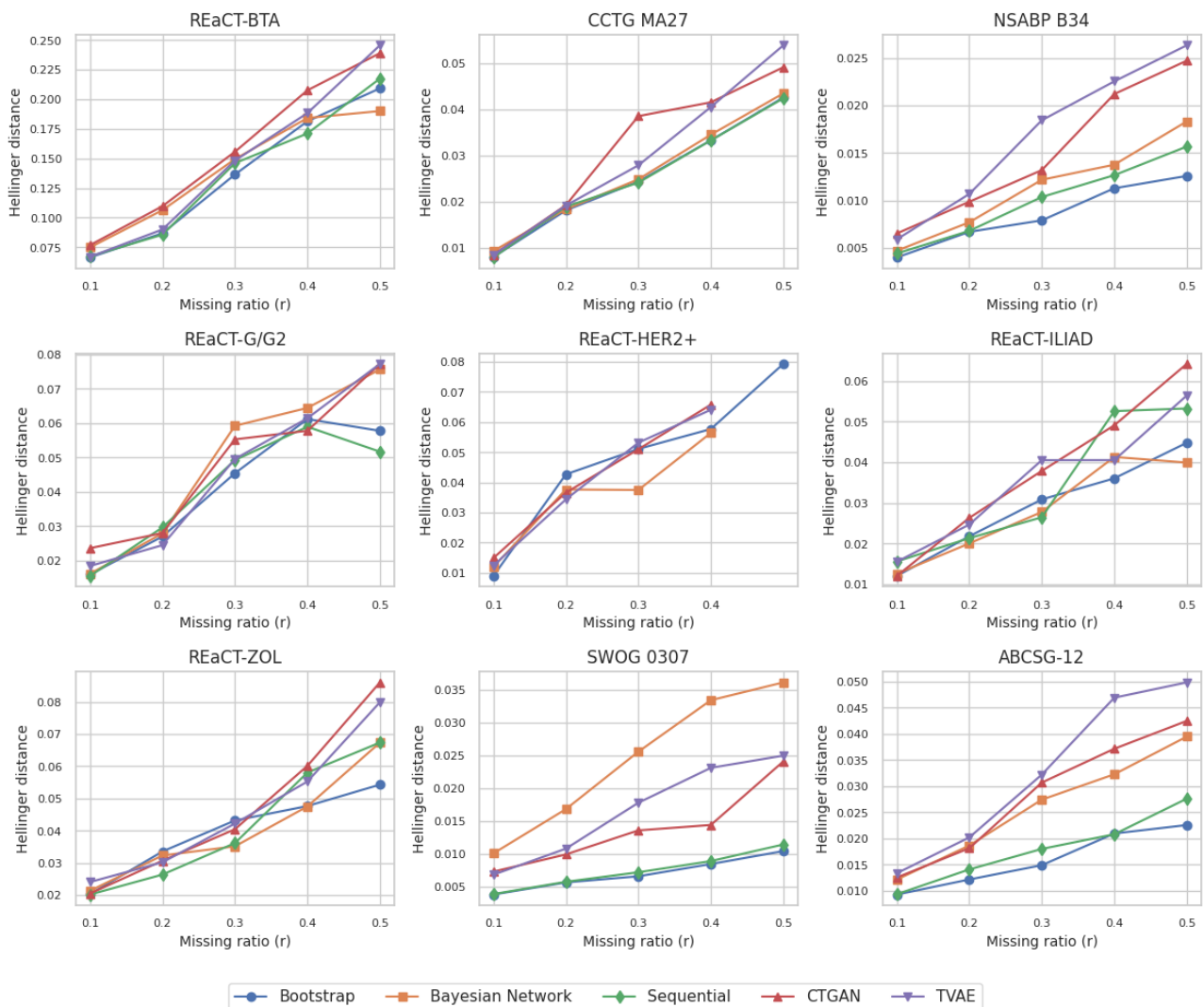
It should be noted that sequential synthesis fails for the smallest clinical trial and that was not included in the denominator for Figure 4. This failure was a design decision in the implementation that we used to not train a model with less than 50 observations. Also note that in the plots model failures are not counted in the denominator.

For the results by dataset, we focus on 4 trials that exemplify all the scenarios in our dataset across all 4 generative models, and are shown in Figure 5 for estimate agreement, Figure 6 for decision agreement, Figure 7 for the standardized difference, and Figure 8 for the CI overlap. The values were consistently high for estimate agreement, standardized difference, and CI overlap across all the scenarios, even at values of *r* approaching 0.5. The results for decision agreement in Figure 6 were more varied and can be characterized as follows:

1. For studies where there was a statistically significant result, for example, REaCT-ILIAD, the ability to maintain decision agreement deteriorates with higher values of *r*. The best performing generative model was sequential synthesis in that decision agreement was maintained for *r* as high as 0.4. Next was the bootstrap, which had a decision agreement for *r* up to 0.3.

2. For studies where the results were not statistically significant (eg, Canadian Cancer Trials Group MA27) or a marginal nonsignificant result (eg, ABCSG–ZOL), the value of *r* had no impact on the replicability of the results in that all the results were successfully replicated.

3. For smaller studies (eg, REaCT-ZOL), some generative models were able to maintain decision agreement for *r* as high as 0.3, although most models failed for *r* greater than that.

Figure 4. All 4 metrics calculated across all the datasets for the bootstrap and sequential generators. For estimate agreement, decision agreement, and standardized difference, the y-axis is the proportion across all datasets. For CI overlap, the y-axis is the average across all data sets. Modeling failures are not considered.
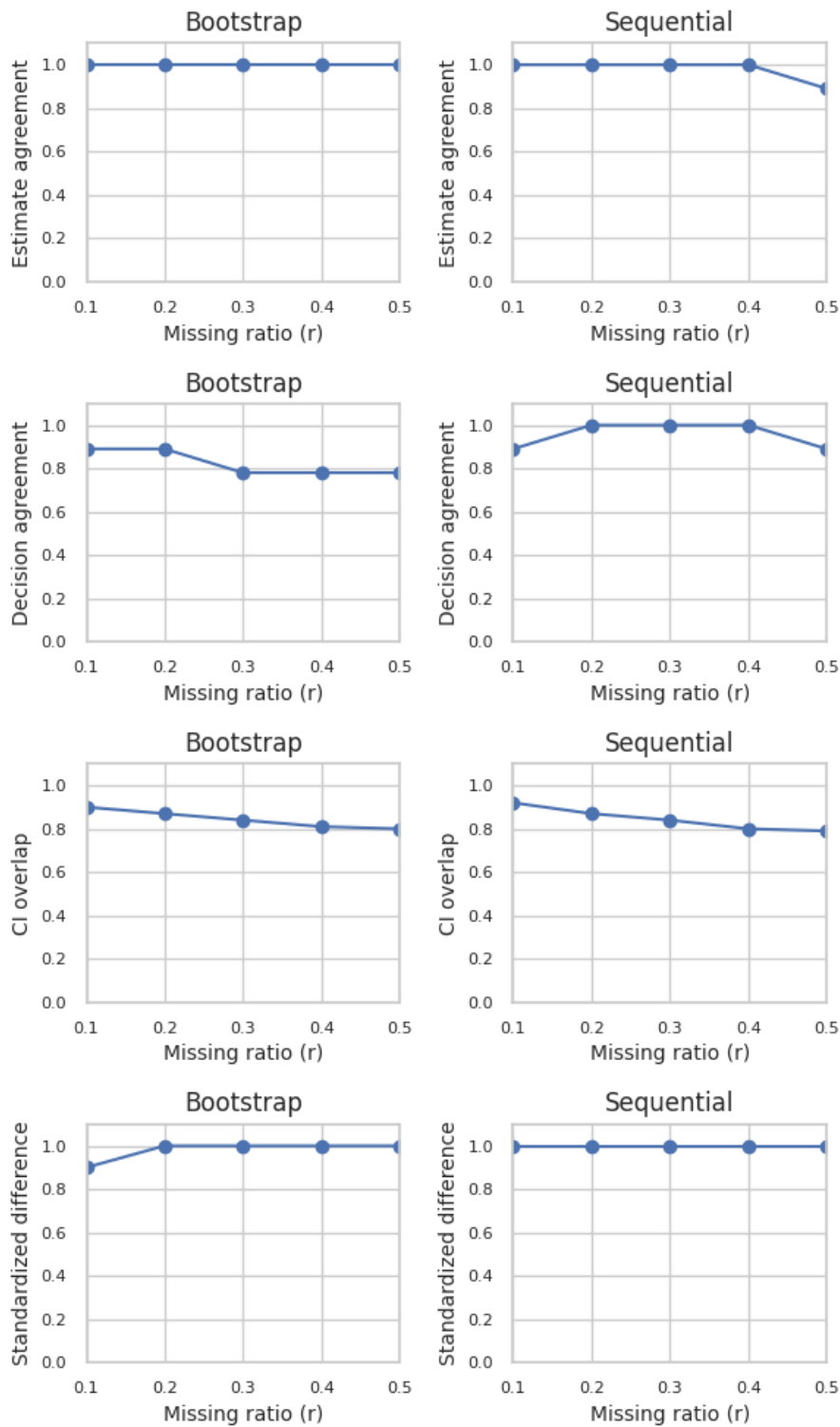
Figure 5. Estimate agreement for selected datasets—proportion across all generators. ABCSG: Austrian Breast and Colorectal Cancer Study Group; CCTG: Canadian Cancer Trials Group; HER2+: human epidermal growth factor receptor-2 positive; REaCT: Rethinking Clinical Trials; ZOL: zoledronate.
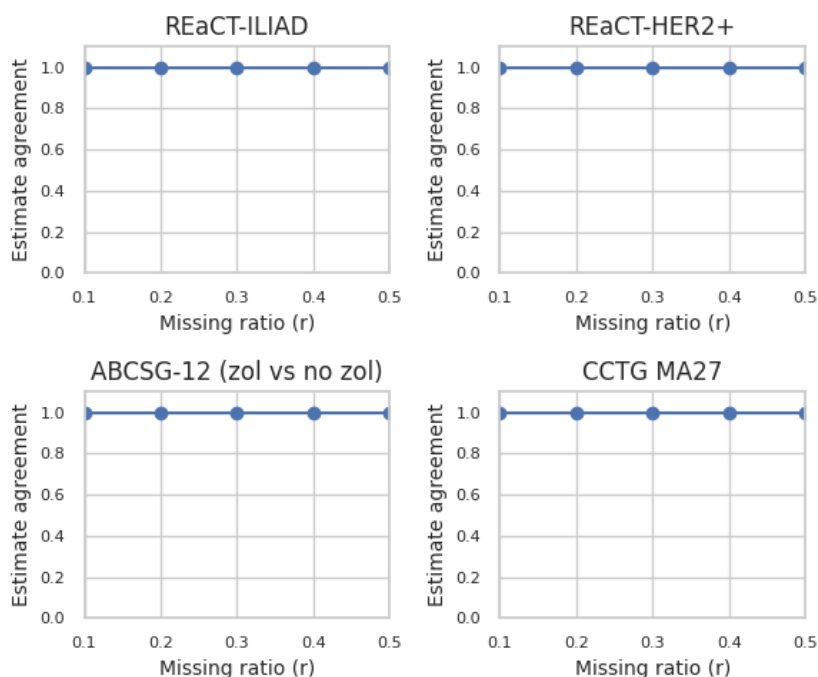
Figure 6. Decision agreement for selected datasets—proportion across all generators. ABCSG: Austrian Breast and Colorectal Cancer Study Group; CCTG: Canadian Cancer Trials Group; HER2+: human epidermal growth factor receptor-2 positive; REaCT: Rethinking Clinical Trials; ZOL: zoledronate.
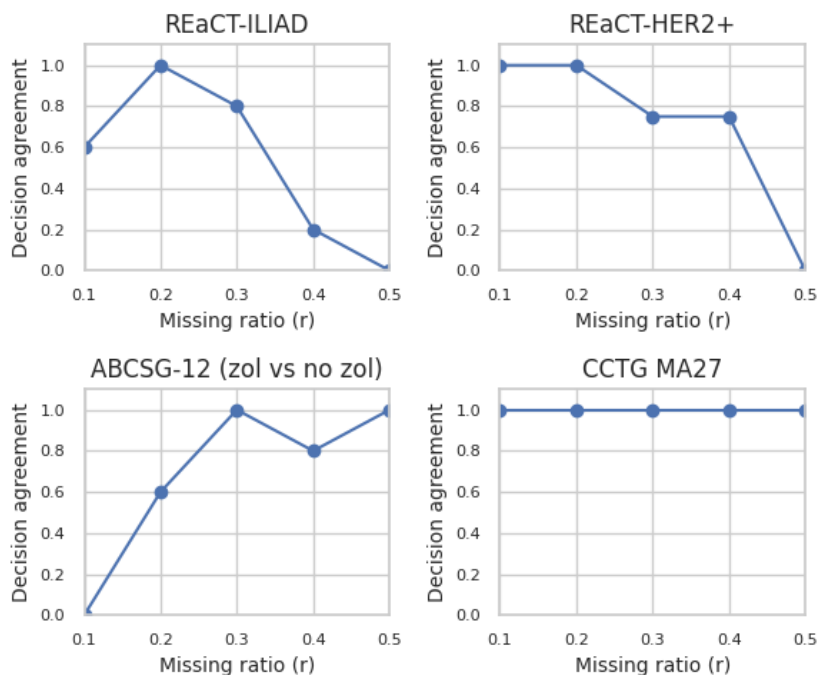
Figure 7. Standardized difference indicator for selected datasets—proportion across all generators. ABCSG: Austrian Breast and Colorectal Cancer Study Group; CCTG: Canadian Cancer Trials Group; HER2+: human epidermal growth factor receptor-2 positive; REaCT: Rethinking Clinical Trials; ZOL: zoledronate.
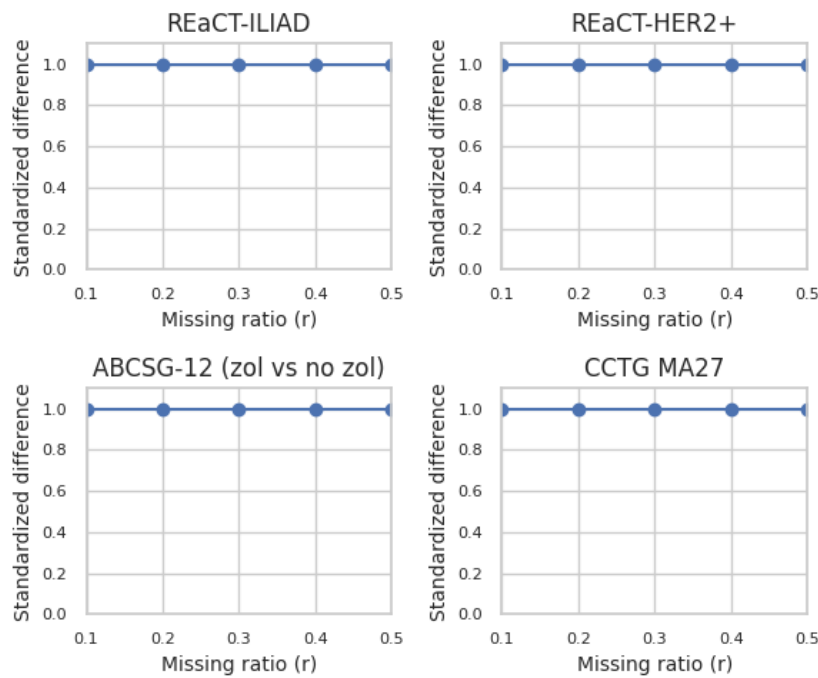
Figure 8. CI overlap for selected datasets—average across all generators. ABCSG: Austrian Breast and Colorectal Cancer Study Group; CCTG: Canadian Cancer Trials Group; HER2+: human epidermal growth factor receptor-2 positive; REaCT: Rethinking Clinical Trials; ZOL: zoledronate.
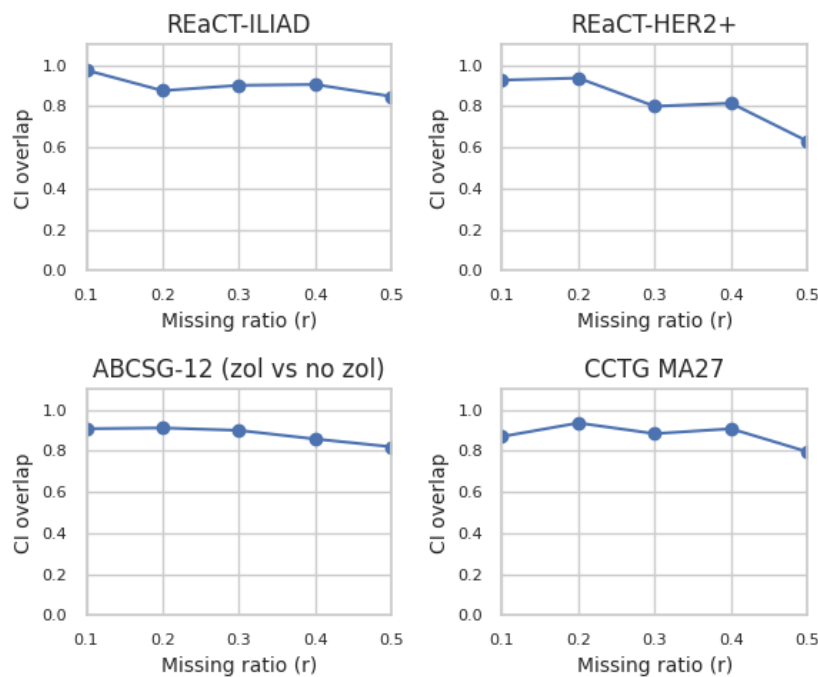
Discussion

Summary

Many clinical trials face accrual problems, resulting in an inability to reach target recruitment and resulting in analysts drawing conclusions from potentially underpowered studies. This exposes patients to toxicity and additional costs, with potentially no scientific benefit. Accrual problems may be due to genuine difficulty with recruitment of patients or with execution quality challenges during the trial itself.

When a study is unable to recruit more patients, the study can be stopped, and the relevant analyses is performed on the available data. For small trials, analyzing the data with insufficient accrual results in even smaller sample sizes, which

can produce unstable parameter estimates and direction. For larger trials, when the complete trial results are statistically significant, an analysis with insufficient accrual can be underpowered and result in nonsignificant findings. For marginal results with the full data, insufficient accrual can reverse their statistical significance. When the full study results are not significant, then insufficient accrual would have less of an impact. A priori, it would not be known which one of these situations pertains to a particular study, making it difficult to interpret the results when analysis is performed with unplanned accrual deficiencies.

The objective of this study was to determine whether generative models can be a useful tool to rescue clinical studies that have insufficient accrual, through augmentation. Generative models have been applied to simulate participants [93-96] and counterfactuals [95,96] in the context of in silico clinical trials. While there have been concerns about generative models overfitting for the small datasets typically encountered in clinical trials [96,97], recent studies have been able to generate synthetic variants of full clinical trial datasets with high utility [74,95,97-102].

To test the ability of generative models to augment clinical trial datasets, we evaluated 4 different types of commonly used generative models (ie, sequential synthesis using decision trees, Bayesian network, GAN, and VAE) on 9 different breast cancer clinical trials and 10 different analyses. The study simulated different degrees of insufficient accrual and the generative models simulated replacement patients to compensate. The last fraction of recruited patients were replaced with simulated ones ranging from $r=0.1$ to $r=0.5$. In addition to generative models, we evaluated a bootstrap approach. These augmented datasets were then used to replicate the published analyses (ie, using the complete datasets) on these 9 trials.

An important assumption for these augmentation methods to work was that participants recruited early were not systematically different than late participants in their estimated effect size. It has been argued that estimated effect sizes tend to vary as patients are recruited and converge to the true value with more information [40,42]. In contrast, as sites gain experience and adjust their processes, there could be a monotonic treatment effect over recruitment time. Also, consider, for example, if there was treatment effect heterogeneity on disease severity (ie, the impact of the treatment on the outcome depends on disease severity), with high severity patients having a bigger response to the intervention, and fewer high severity patients were recruited early in the study compared with late in the study. This would manifest itself as a monotonic relationship between the estimated effect size and patients recruited over time.

We tested that monotonic relationship hypothesis by examining the interaction effect between recruitment order and treatment on the outcome. We did not find evidence across the trials that this relationship was monotonic (ie, all interaction effects were small and not statistically significant), meaning that we could not detect an increasing or decreasing effect size as more patients were recruited. This is despite these trials having long enrollment periods, in some cases lasting years. Therefore, if

the early and late participants were, on average, similar in the estimated effect size, then generative models and a simple bootstrap would also be expected to work well.

These results are supportive of hypothesis 1, which stated that patients recruited early in a trial are similar to those recruited later in the trial.

Furthermore, the fidelity of the generated datasets was quite high relative to the training datasets from the different generative models. In all cases, the patterns in datasets that were synthesized had a high similarity on the Hellinger distance to the datasets of the participants already recruited.

Several observations can be made from our results:

1. A bootstrap would be attractive due to its simplicity and low computational burden. However, this method tended to lack consistent decision agreement when a trial result was statistically significant and when the trial was small, while in other scenarios, sampling with replacement performed well. However, the former are nontrivial failure modes.
2. All approaches struggled with marginal results (eg, marginally nonsignificant results) when the r value was low. This indicates a general sensitivity to that particular scenario.
3. For r values as high as 0.4, sequential synthesis performed well across all datasets. This means that both decision agreement and estimate agreement were achieved, CI overlap was at or above 0.8, and it was either equivalent to or better than the other methods evaluated.
4. Bayesian networks and the GAN had the next best performance after sequential synthesis; however, the Bayesian network had slightly better CI overlap overall up to an $r=0.4$.

To ensure reasonable performance across multiple scenarios, the results suggest that sequential synthesis can be used to address insufficient accrual up to $r=0.4$ (ie, only 60% of the target is recruited).

These results are supportive of hypothesis 2, which stated that generative models can simulate the remaining patients in a clinical trial with insufficient recruitment, and the augmented dataset would replicate the results if the trial did reach target recruitment.

It should be noted that using generative models to simulate additional patients as described in this study would not be preplanned, as opposed to a planned interim analysis. Insufficient accrual becomes a problem when there are no budgeted resources available to continue recruiting patients, for example, by adding sites, extending recruitment time, or changing the inclusion and exclusion criteria. Although, if the results from the augmentation show positive findings, the case may be made to allocate more resources to continue recruitment. Additionally, if the results from augmentation show negative findings, then that would provide a stronger case for terminating the study.

The countries of recruitment for the clinical trials used in this study cover multiple regions around the world as shown in Table

2. Our results were consistent across the different jurisdictions. Therefore, it would be reasonable to have confidence that the findings are generalizable across multiple jurisdictions and not specific to a particular region.

Comparison With Prior Work

Previous studies have demonstrated that sequential synthesis performs well (ie, in terms of replicability of published studies) on oncology clinical trial datasets [74,98], and therefore, our findings are consistent with that evidence. More generally, sequential synthesis has been found to have superior utility across different types of datasets relative to other types of generative models [103-105]. Also, it should be noted that most of the published analyses that were replicated across all clinical trials used datasets that were low dimensional, which imposes lower sample size requirements for the generative models.

An earlier study found that a VAE generative model trained on early patients could augment a clinical trial with simulated patients [35]. In that study, the authors argued that generative models can also enable the design of smaller studies to start off with. This means that studies would be designed to be smaller, with augmentation used to reach target recruitment for the final analysis. However, that analysis only considered a single simulated clinical trial (ie, not real data).

The argument for using augmentation to prospectively design smaller studies is appealing. The largest factor driving up the cost of trials is the number of participants required to achieve sufficient statistical power [106,107]. The median cost per participant in drug trials (ie, in general) was estimated to be US $41,413 [107], the median cost per participant specifically in oncology drug trials was US $100,271 [107], and an earlier study found the 1-year cost to be US $17,003 per patient in the treatment arm and US $15,516 per control participant [108]. Designing studies that require fewer patients to be recruited can improve the cost-effectiveness of studies.

Our current analysis generalizes this work to other types of generative models, including a VAE, but we did find that a VAE did not perform well on our criteria. Plus, we performed the evaluation on 9 real clinical trial datasets of different sizes and durations and compared the generative models to a simple bootstrap. Furthermore, an important difference is that rescuing a study due to insufficient accrual using augmentation is not planned, whereas designing a small study is planned.

Nevertheless, using augmentation to design smaller studies deserves further investigation, and it remains a necessity that researchers aim to recruit the target sample size wherever possible.

Clinical trials are known to underrepresent certain groups, and hence, there is the potential for introducing bias in the results. For instance, a recent study in Canada found that the underrepresentation of Black patients in cancer research remains a significant concern, with 15 out of the 20 most common types of cancer not being studied in Black communities [109,110]; studies on underrepresented populations in clinical trials show racial and ethnic disparities worldwide [111-118]; and there is a consistent underrepresentation of various other groups, such as older adults [111,112,119-122], women

[111,113,119,120,123], and individuals of lower socioeconomic status and educational level [111,112,124]. Furthermore, synthetic data generation has been shown to introduce bias in the generated data relative to the training data [125], and these biases are propagated across multiple generations of generative models, where the output of one is used as training for the next one [126].

Our analysis did not explicitly evaluate representation bias as we were replicating the published analyses rather than identifying and correcting any weaknesses and did not explicitly attempt to mitigate such underrepresentation to the extent that it existed in the original datasets. Nevertheless, analysts can also apply augmentation methods to compensate for any biases that may exist in the training datasets [127-129].

Another risk of bias is if there is a relationship between a particular characteristic and the order of recruitment. For example, if the first 60% of patients recruited were aged mostly >70 years and the last 40% were aged mostly <70 years, then the trained generative model and their simulated patients would not include sufficient younger patients. However, such an age bias would only impact the randomized study outcomes if there was an interaction between recruitment order, or factors correlated to it, such as age in this example, and the treatment. We explicitly tested for such an interaction effect, and as seen in the results, there were none detected.

Generative models can simulate a larger number of patients than what is needed to reach target recruitment (ie, data amplification). On the surface, this may seem to be a mechanism to amplify the statistical power of the study and solve the problem of drawing conclusions from small studies. However, with the necessary adjustments using the combining rules described in our methodology, it has been shown that amplification does not increase statistical power for fully synthetic data as the adjusted SEs of parameter estimates are also increased [71]. The same study shows that population inferences and replicability diminish markedly without using the combining rules during the analysis of synthetic data. Further examination is needed to determine whether the same conclusions would hold for hybrid data that is only part synthetic.

The incorporation of generative models and simulated patients in industry-sponsored clinical trials would necessitate collaboration with sponsors to apply and evaluate these methods in their studies across multiple therapeutic areas to determine consistency in performance. However, recent surveys indicate that sponsors see uncertainty about regulators' expectations and requirements for evidence as a critical barrier for the adoption of computer modeling and simulation methods in clinical trials [130]. Some efforts have identified high level principles that can be applied for the quality assurance evaluation of in silico trials [131], as well as a general good practice guidance for the application of simulations [132]. Regulators have noted the potential of simulated patients in clinical trials [133] and have further suggested adopting the Government Accountability Office accountability framework [134] for the application of machine learning models to in silico trials, which includes addressing challenges related to governance, accountability,

and transparency; data considerations; and model development, performance, and validation [135]. Furthermore, synthetic data, because of their privacy-protective properties, can serve as a more readily available proxy for RWD when used in regulatory submissions [136], which can solve the data access challenge and accelerate the generation of real-world evidence. Some experts at regulatory agencies have expressed cautious optimism that synthetic data can be used by manufacturers [137], with replicability of results on real data and the need for further experimental exemplars being emphasized specifically for synthetic data generation [94]. Our study contributes to that evidence base.

Recruitment is a challenging issue not only in clinical trials but in other clinical studies as well [138,139]. These challenges are exacerbated with studies on rare diseases [140] and pediatric studies. Pediatric datasets are typically small due to a scarcity of potential study participants: there are fewer children in the population with severe disease [141]. Additional challenges include the complex ethical issues surrounding research involving children and an extra layer of consent required for pediatric participants (ie, parental consent is required in addition to patients' assent) [142]. Many trials recruited a very small number of children [143], and studies with placebo arms are also at a disadvantage as patients are less likely to participate in case they do not get randomized to the treatment arm [144]. Therefore, the results from our study would have broader applicability beyond oncology trials in adults.

On a methodological point, this study was possible because of the ability to obtain access to the original datasets across 9 different clinical trials. While there has been strong interest in making more clinical trial data available for secondary analysis by journals, funders, the pharmaceutical industry, and regulators [145-153], data access for secondary analysis remains a challenge [154], sometimes taking many months to get data [155,156]; for example, an analysis 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% [156-161]. Specifically, recent reports highlight the difficulties in accessing data for health research and machine learning analytics [162-164]. In our case, the process of getting access to all the datasets used in this study took approximately 2 years, including executing the necessary data sharing agreements and establishing collaborations with the original investigators.

One of the reasons that access to individual-level clinical trial data faces friction is concern over patient privacy by the patients and regulators [165,166]. However, the general assumption has been that synthetic data produced through generative models have low identity disclosure vulnerability because there is no unique or one-to-one mapping between the records in the synthetic data with the records in the original (ie, real) data [167-174]. While there are other types of disclosure vulnerabilities that are also relevant [175,176], some authors have argued that synthetic data can be considered nonpersonal information under statuary definitions in North America and Europe [177-179]. This would arguably be the case if disclosure vulnerability measurements demonstrate vulnerability values that are below acceptable thresholds. To that end, it is

encouraging that recently study authors have been making synthetic variants of data used in their research papers publicly available to enable open science [180-183]. However, given that generating fully synthetic variants of the clinical trial datasets used in this study were not part of the original protocol, readers interested in data access can make a request to the individual data custodians, with the necessary contacts in the *Data Availability* section.

Limitations

Although our datasets covered single site as well as multisite studies, and trials performed across multiple regions of the world with a wide range of sizes and durations, our results were obtained only from oncology trials (mostly breast cancer). There is no a priori reason for these methods not to work for other diseases and conditions, and different populations. To generalize the findings, it is necessary to replicate the findings in this work for other diseases, particularly those that impose high societal costs and where the acceleration of clinical trial evidence can be most impactful.

It is not known whether there would be evidence of a monotonic effect size over the enrollment period in other types of clinical trials and with different populations. For example, one can argue that surgery trials may exhibit a monotonic effect because surgeons become more experienced with new procedures over time. None of our studies were surgery trials. Therefore, this monotonicity relationship would need to be investigated further in these other contexts before drawing broader conclusions.

This study did not consider safety data, which tend to have fewer observations in a clinical trial. Because of the relatively smaller number of adverse events, it is more challenging to train a generative model on that kind of information. Therefore, studies comparing existing interventions or new indications would be more suited for the application of augmentation methods.

Furthermore, our retrospective analysis was limited to 9 clinical trials that were actually completed. Studies that do not reach accrual targets may have different characteristics, which could lead to different results. While our results are encouraging based on a retrospective analysis, future work should evaluate this approach through a prospective design.

When the r value increases, the available dataset size to train a generative model decreases. For the smallest clinical trials, in some rare instances, this resulted in generative model failure. The generative models that did not fail under those conditions had a high risk of overfitting to the training data, although they still performed better than a simple bootstrap on decision agreement. For most of our small studies (REaCT-ILIAD, REaCT–BTA, and REaCT-ZOL), the number of variables that were used for training was also quite small (8, 8, and 11, respectively; Table 1). To the extent that low dimensionality has a diluting impact on the rate of overfitting for a fixed sample size, the small number of variables would reduce that risk. Furthermore, most of the variables in some of the small trials were categorical with very few categories (ie, mostly binary), suggesting quite simple datasets were being modeled (eg, REaCT-HER2+).

We did not find that the deep learning generative models performed as well as the other approaches that were considered. One possible explanation is that additional hyperparameter tuning for these models was not performed and the default implementation model size and characteristics were used. The tuning could have improved the performance of these models. Although the default hyperparameters were suitable for low dimensional data, which was the case for the analysis datasets in many of these clinical trials.

In drawing our conclusions, we used a 0.8 value as a threshold of acceptability on CI overlap. Should one apply a more stringent threshold, then the values of r where the results would be acceptable would decrease further.

Future Work

Further research is needed to provide evidence-based parameters for the number of variables that can be simulated with different types of generative models and the appropriate number of patients in a trial for training these models. This will further ensure that augmentation is not applied in inappropriate contexts.

Studies on rare diseases would have few observations, making it more challenging to train a generative model using the methods described here. Evaluations using pretrained generative models to simulate clinical trial patients should be investigated as these may be more applicable to rare disease trials.

Given that our focus was to address the insufficient accrual problem, the augmentation that was performed covered all arms of the study. However, the methods described here can be applied to individual arms as well. For example, it is possible to augment only the control arm of a trial if there were challenges accruing patients in the control arm or it could be by design to only recruit a subset of control patients and simulate the rest. Future work can evaluate such alternative augmentation strategies.

An alternative approach for augmentation to address insufficient accrual is to use pretrained generative models. For example, these models can be trained on historical clinical trial datasets, and then the pretrained models are used to simulate additional patients. This is different from the approach presented in this paper, whereby already collected data from the trial are used to train the generative models. It would be informative to compare both approaches to determine which would work better in practice, and to understand the types of historical datasets that give the best results for training these generative models.

Acknowledgments

Data Availability

The datasets used in this study are not publicly available. Access to the datasets can be requested from the contacts listed in Multimedia Appendix 1 for each of the datasets used in this study. The synthetic data generation code is available in the *pysdg* package [184].

Authors' Contributions

SEK, NM, GP, MC, and KEE designed the study. All authors participated in collecting and curating the datasets and writing and reviewing the paper. SEK, NM, GP, MC, CF, DH, and KEE performed the data analysis.

Conflicts of Interest

KEE owns shares in Aetion. AABB received a travel grant from Janssen. GP has received consulting fees from Traferox Technologies and Merck, honorariums from Astra-Zeneca and Takeda, and has a close family member who is employed by Roche Canada and who owns stock in Roche Ltd. MFS has received honoraria from GILEAD, AstraZeneca, Merck, Seagen, Novartis, Pfizer, Roche, and Lilly. RG has advisory roles and provided expert testimony to: Celgene, Novartis, Roche, BMS, Takeda, Abbvie, Astra Zeneca, Janssen, MSD, Merck, Gilead, Daiichi Sankyo; Honoria: Celgene, Roche, Merck, Takeda, AstraZeneca, Novartis, Agmen, BMS, MSD, Sandoz, Abbvie, Gilead, Daiichi Sankyo; financing of scientific research: Celgene, Merck, Takeda, AstraZeneca, Novartis, Agmen, BMS, MSD, Sandoz, Gilead, Roche; travel support: Roche, Agmen, Janssen, Astra Zeneca, Novatris, MSD Gilead, Abbvie, Daiichi Sankyo. M. G. reports personal fees / travel support from Amgen, AstraZeneca, Bayer, DaiichiSankyo, EliLilly, EPG Health (IQVIA), Menarini-Stemline, MSD, Novartis, PierreFabre, Veracyte.
KEE is the Editor-in-Chief of JMIR AI at the time of this publication but played no role in the editorial handling or peer review process of this manuscript.

Multimedia Appendix 1

Methodology details and analysis results.
[PDF File (Adobe PDF File), 300 KB-Multimedia Appendix 1]

References

1. Prescott RJ, Counsell CE, Gillespie AJ, Grant AM, Russell IT, Kiauka S, et al. Factors that limit the quality, number and progress of randomised controlled trials: a review. Health Technol Assess. 1999;3(20). [doi: 10.3310/hta3200]

2. Gul RB, Ali PA. Clinical trials: the challenge of recruitment and retention of participants. J Clin Nurs. Jan 17, 2010;19(1-2):227-233. [doi: 10.1111/j.1365-2702.2009.03041.x] [Medline: 20500260]

3. Kasenda B, von Elm E, You J, Blümle A, Tomonaga Y, Saccilotto R, et al. Prevalence, characteristics, and publication of discontinued randomized trials. JAMA. Mar 12, 2014;311(10):1045-1051. [FREE Full text] [doi: 10.1001/jama.2014.1361] [Medline: 24618966]

4. Kitterman DR, Cheng SK, Dilts DM, Orwoll ES. The prevalence and economic impact of low-enrolling clinical studies at an academic medical center. Acad Med. Nov 2011;86(11):1360-1366. [FREE Full text] [doi: 10.1097/ACM.0b013e3182306440] [Medline: 21952064]

5. Stensland KD, McBride RB, Latif A, Wisnivesky J, Hendricks R, Roper N, et al. Adult cancer clinical trials that fail to complete: an epidemic? J Natl Cancer Inst. Sep 2014;106(9):dju229. [doi: 10.1093/jnci/dju229] [Medline: 25190726]

6. Feller S. One in four cancer trials fails to enroll enough participants. United Press International. Dec 30, 2015. URL: https://www.upi.com/Health_News/2015/12/30/One-in-four-cancer-trials-fails-to-enroll-enough-participants/2611451485504/ [accessed 2021-03-25]

7. McDonald AM, Knight RC, Campbell MK, Entwistle VA, Grant AM, Cook JA, et al. What influences recruitment to randomised controlled trials? A review of trials funded by two UK funding agencies. Trials. Apr 07, 2006;7:9. [FREE Full text] [doi: 10.1186/1745-6215-7-9] [Medline: 16603070]

8. Institute of Medicine. Transforming Clinical Research in the United States: Challenges and Opportunities: Workshop Summary. Washington, DC. The National Academies Press; 2010.

9. Sully BG, Julious SA, Nicholl J. A reinvestigation of recruitment to randomised, controlled, multicenter trials: a review of trials funded by two UK funding agencies. Trials. Jun 09, 2013;14:166. [FREE Full text] [doi: 10.1186/1745-6215-14-166] [Medline: 23758961]

10. Mirza M, Siebert S, Pratt A, Insch E, McIntosh F, Paton J, et al. Impact of the COVID-19 pandemic on recruitment to clinical research studies in rheumatology. Musculoskeletal Care. Mar 2022;20(1):209-213. [FREE Full text] [doi: 10.1002/msc.1561] [Medline: 33938621]

11. Mitchell EJ, Ahmed K, Breeman S, Cotton S, Constable L, Ferry G, et al. It is unprecedented: trial management during the COVID-19 pandemic and beyond. Trials. Sep 11, 2020;21(1):784. [FREE Full text] [doi: 10.1186/s13063-020-04711-6] [Medline: 32917258]

12. Slow recruitment due to Covid-19 disruptions continues to climb in 2022. Clinical Trials Arena. Jan 31, 2022. URL: https://www.clinicaltrialsarena.com/analyst-comment/slow-recruitment-covid-19-disruptions/ [accessed 2023-10-05]

13. McDonald K, Seltzer E, Lu M, Gaisenband SD, Fletcher C, McLeroth P, et al. Quantifying the impact of the COVID-19 pandemic on clinical trial screening rates over time in 37 countries. Trials. Apr 04, 2023;24(1):254. [FREE Full text] [doi: 10.1186/s13063-023-07277-1] [Medline: 37013558]

14. Hauck CL, Kelechi TJ, Cartmell KB, Mueller M. Trial-level factors affecting accrual and completion of oncology clinical trials: a systematic review. Contemp Clin Trials Commun. Dec 2021;24:100843. [FREE Full text] [doi: 10.1016/j.conctc.2021.100843] [Medline: 34765799]

15. Carlisle B, Kimmelman J, Ramsay T, MacKinnon N. Unsuccessful trial accrual and human subjects protections: an empirical analysis of recently closed trials. Clin Trials. Feb 2015;12(1):77-83. [FREE Full text] [doi: 10.1177/1740774514558307] [Medline: 25475878]

16. Lièvre M, Ménard J, Bruckert E, Cogneau J, Delahaye F, Giral P, et al. Premature discontinuation of clinical trial for reasons not related to efficacy, safety, or feasibility. BMJ. Mar 10, 2001;322(7286):603-605. [FREE Full text] [doi: 10.1136/bmj.322.7286.603] [Medline: 11238162]

17. Halpern SD, Karlawish JH, Berlin JA. The continuing unethical conduct of underpowered clinical trials. JAMA. Jul 17, 2002;288(3):358-362. [doi: 10.1001/jama.288.3.358] [Medline: 12117401]

18. Schmidli H, Häring DA, Thomas M, Cassidy A, Weber S, Bretz F. Beyond randomized clinical trials: use of external controls. Clin Pharmacol Ther. Apr 17, 2020;107(4):806-816. [doi: 10.1002/cpt.1723] [Medline: 31725899]

19. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behav Res. May 2011;46(3):399-424. [FREE Full text] [doi: 10.1080/00273171.2011.568786] [Medline: 21818162]

20. Baumfeld Andre E, Reynolds R, Caubel P, Azoulay L, Dreyer NA. Trial designs using real-world data: The changing landscape of the regulatory approval process. Pharmacoepidemiol Drug Saf. Oct 01, 2020;29(10):1201-1212. [FREE Full text] [doi: 10.1002/pds.4932] [Medline: 31823482]

21. Gökbuget N, Kelsh M, Chia V, Advani A, Bassan R, Dombret H, et al. Blinatumomab vs historical standard therapy of adult relapsed/refractory acute lymphoblastic leukemia. Blood Cancer J. Sep 23, 2016;6(9):e473. [FREE Full text] [doi: 10.1038/bcj.2016.84] [Medline: 27662202]

22. Davi R, Mahendraratnam N, Chatterjee A, Dawson CJ, Sherman R. Informing single-arm clinical trials with external controls. Nat Rev Drug Discov. Dec 18, 2020;19(12):821-822. [doi: 10.1038/d41573-020-00146-5] [Medline: 32811986]

23. Mumuni A, Mumuni F. Data augmentation: a comprehensive survey of modern approaches. Array. Dec 2022;16:100258. [doi: 10.1016/j.array.2022.100258]

24. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. J Big Data. Jul 6, 2019;6:60. [doi: 10.1186/s40537-019-0197-0]

25. Goceri E. Medical image data augmentation: techniques, comparisons and interpretations. Artif Intell Rev. Mar 20, 2023:1-45. [FREE Full text] [doi: 10.1007/s10462-023-10453-z] [Medline: 37362888]

26. Wen Q, Sun L, Yang F, Song X, Gao J, Wang X, et al. Time series data augmentation for deep learning: a survey. arXiv. Preprint posted online on Feb 27, 2020. [FREE Full text] [doi: 10.24963/ijcai.2021/631]

27. Iwana BK, Uchida S. An empirical survey of data augmentation for time series classification with neural networks. PLoS One. Jul 15, 2021;16(7):e0254841. [FREE Full text] [doi: 10.1371/journal.pone.0254841] [Medline: 34264999]

28. 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). [FREE Full text]

29. Nakhwan M, Duangsoithong R. Comparison analysis of data augmentation using bootstrap, GANs and autoencoder. In: Proceedings of the 14th International Conference on Knowledge and Smart Technology. 2022. Presented at: KST 2022; January 26-29, 2022; Chon Buri, Thailand. [doi: 10.1109/kst53302.2022.9729065]

30. Zhao Y, Duangsoithong R. Empirical analysis using feature selection and bootstrap data for small sample size problems. In: Proceedings of the 16th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology. 2019. Presented at: ECTI-CON 2019; July 10-13, 2019; Pattaya, Thailand. [doi: 10.1109/ecti-con47248.2019.8955366]

31. Pedersen PM. Exploring the value of GANs for synthetic tabular data generation in healthcare with a focus on data quality, augmentation, and privacy. Oslo Metropolitan University. 2023. URL: https://oda.oslomet.no/oda-xmlui/handle/11250/3101091#:~:text=While%20various%20methods%20exist%20for,is%20organized%20in%20tabular%20format [accessed 2025-01-01]

32. Ahmadian M, Bodalal Z, van der Hulst HJ, Vens C, Karssemakers LH, Bogveradze N, et al. Overcoming data scarcity in radiomics/radiogenomics using synthetic radiomic features. Comput Biol Med. May 2024;174:108389. [doi: 10.1016/j.compbiomed.2024.108389] [Medline: 38593640]

33. Wang W, Pai TW. Enhancing small tabular clinical trial dataset through hybrid data augmentation: combining SMOTE and WCGAN-GP. Data. Aug 23, 2023;8(9):135. [doi: 10.3390/data8090135]

34. Shafquat A, Beigi M, Gao C, Mezey J, Sun J, Aptekar J. An interpretable data augmentation framework for improving generative modeling of synthetic clinical trial data. In: Proceedings of the 3rd Workshop on Interpretable Machine Learning in Healthcare. 2023. Presented at: IMLH 2023; July 28, 2023; Virtual event. URL: https://icml.cc/virtual/2023/27750

35. Papadopoulos D, Karalis VD. Variational autoencoders for data augmentation in clinical studies. Appl Sci. Jul 30, 2023;13(15):8793. [doi: 10.3390/app13158793]

36. Carey VJ, Lumley TS, Moler C, Ripley B. gee: generalized estimation equation solver. The Comprehensive R Archive Network. Dec 11, 2024. URL: https://cran.r-project.org/web/packages/gee/index.html [accessed 2024-12-12]

37. Oehlert GW. A note on the delta method. Am Stat. Feb 1992;46(1):27. [doi: 10.2307/2684406]

38. Ver Hoef JM. Who invented the delta method? Am Stat. May 2012;66(2):124-127. [doi: 10.1080/00031305.2012.687494]

39. Beltran-Bless AA, Clemons M, Vandermeer L, El Emam K, Ng TL, McGee S, et al. The Rethinking Clinical Trials Program Retreat 2023: creating partnerships to optimize quality cancer care. Curr Oncol. Mar 06, 2024;31(3):1376-1388. [FREE Full text] [doi: 10.3390/curroncol31030104] [Medline: 38534937]

40. The Coronary Drug Project Research Group. Practical aspects of decision making in clinical trials: the coronary drug project as a case study. Control Clin Trials. May 1981;1(4):363-376. [doi: 10.1016/0197-2456(81)90041-6] [Medline: 7261627]

41. Degtiar I, Rose S. A review of generalizability and transportability. Annu Rev Stat Appl. Mar 10, 2023;10(1):501-524. [doi: 10.1146/annurev-statistics-042522-103837]

42. Ciolino JD, Kaizer AM, Bonner LB. Guidance on interim analysis methods in clinical trials. J Clin Transl Sci. May 15, 2023;7(1):e124. [FREE Full text] [doi: 10.1017/cts.2023.552] [Medline: 37313374]

43. Qian Z, Cebere BC, van der Schaar M. Synthcity: facilitating innovative use cases of synthetic data in different data modalities. arXiv. Preprint posted online on January 18, 2023. [FREE Full text]

44. El Emam K, Mosquera L, Zheng C. Optimizing the synthesis of clinical trial data using sequential trees. J Am Med Inform Assoc. Jan 15, 2021;28(1):3-13. [FREE Full text] [doi: 10.1093/jamia/ocaa249] [Medline: 33186440]

45. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: a conditional inference framework. J Comput Graph Stat. Sep 2006;15(3):651-674. [doi: 10.1198/106186006X133933]

46. Read J, Pfahringer B, Holmes G, Frank E. Classifier chains for multi-label classification. In: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases. 2009. Presented at: ECML PKDD 2009; September 7-11, 2009; Bled, Slovenia. [doi: 10.1007/978-3-642-04174-7_17]

47. Drechsler J, Reiter JP. An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. Comput Stat Data Anal. Dec 2011;55(12):3232-3243. [doi: 10.1016/j.csda.2011.06.006]

48. 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. Aug 2021;121(2):410-431. [doi: 10.1037/pspp0000208] [Medline: 30148371]

49. Bonnéry D, Feng Y, Henneberger AK, Johnson TL, Lachowicz M, Rose BA, 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. Aug 02, 2019;12(4):616-647. [doi: 10.1080/19345747.2019.1631421]

50. Lauger A, Freiman M, Reiter J. Data synthesis and perturbation for the American community survey at the U.S. Census Bureau. United States Census Bureau. 2016. URL: https://www.census.gov/content/dam/Census/newsroom/press-kits/2016/20160803_lauger_data_synthesis_and_perturbation_for_acs.pdf [accessed 2025-02-20]

51. Nowok B. Utility of synthetic microdata generated using tree-based methods. synthpop. 2016. URL: https://synthpop.org.uk/assets/35_nowok_psd2016.pdf [accessed 2025-01-07]

52. Raab GM, Nowok B, Dibben C. Practical data synthesis for large samples. J Priv Confidentiality. Feb 02, 2018;7(3):67-97. [doi: 10.29012/jpc.v7i3.407]

53. 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. Stat J IAOS. Aug 21, 2017;33(3):785-796. [doi: 10.3233/sji-150153]

54. Quintana DS. A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation. Elife. Mar 11, 2020;9:9. [FREE Full text] [doi: 10.7554/eLife.53275] [Medline: 32159513]

55. Kaur D, Sobiesk M, Patil S, Liu J, Bhagat P, Gupta A, et al. Application of Bayesian networks to generate synthetic health data. J Am Med Inform Assoc. Mar 18, 2021;28(4):801-811. [FREE Full text] [doi: 10.1093/jamia/ocaa303] [Medline: 33367620]

56. Murphy KP. Machine Learning: A Probabilistic Perspective. Cambridge, MA. MIT Press; 2012.

57. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Proceedings of the 28th International Conference on Neural Information Processing Systems. 2014. Presented at: NIPS'14; December 8-13, 2014; Montreal, QC. URL: https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

58. Bourou S, El Saer A, Velivassaki TH, Voulkidis A, Zahariadis T. A review of tabular data synthesis using GANs on an IDS dataset. Information. Sep 14, 2021;12(9):375. [doi: 10.3390/info12090375]

59. Xu L, Skoularidou M, Cuesta-Infante A. Modeling tabular data using conditional GAN. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019. Presented at: NIPS'19; December 8-14, 2019; Vancouver, BC.

60. Kingma DP, Welling M. Auto-encoding variational bayes. arXiv. Preprint posted online on December 20, 2013. [FREE Full text]

61. Wan Z, Zhang Y, He H. Variational autoencoder based synthetic data generation for imbalanced learning. In: Proceedings of the IEEE Symposium Series on Computational Intelligence. 2017. Presented at: SSCI; November 27-December 1, 2017; Honolulu, HI. [doi: 10.1109/ssci.2017.8285168]

62. Ishfaq H, Hoogi A, Rubin D. TVAE: triplet-based variational autoencoder using metric learning. arXiv. Preprint posted online on February 13, 2018. [FREE Full text]

63. Sohn K, Yan X, Lee H. Learning structured output representation using deep conditional generative models. In: Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 2. 2015. Presented at: NIPS'15; December 7-12, 2015; Montreal, QC.

64. Salim AJ. Synthetic patient generation: a deep learning approach using variational autoencoders. arXiv. Preprint posted online on August 20, 2018. [FREE Full text] [doi: 10.1017/9781108891530.014]

65. Rubin DB. Discussion: statistical disclosure limitation. J Off Stat. 1993;9:461-468.

66. Raghunathan TE, Reiter JP, Rubin DB. Multiple imputation for statistical disclosure limitation. J Off Stat. 2003;19(1):1-16. [FREE Full text]

67. Reiter JP. Satisfying disclosure restrictions with synthetic data sets. J Off Stat. Jan 2002;18(4).

68. Reiter JP. Inference for partially synthetic, public use microdata sets. Survey Methodol. 2003;29(2):181-188.

69. Loong B, Zaslavsky AM, He Y, Harrington DP. Disclosure control using partially synthetic data for large-scale health surveys, with applications to CanCORS. Stat Med. Oct 30, 2013;32(24):4139-4161. [FREE Full text] [doi: 10.1002/sim.5841] [Medline: 23670983]

70. Taub J, Elliot MJ, Sakshaug JW. The impact of synthetic data generation on data utility with application to the 1991 UK samples of anonymised records. Transact Data Privacy. Jan 2020;13(1):1-23. [FREE Full text] [doi: 10.1201/9781315370910-16]

71. El Emam K, Mosquera L, Fang X, El-Hussuna A. An evaluation of the replicability of analyses using synthetic health data. Sci Rep. Mar 24, 2024;14(1):6978. [FREE Full text] [doi: 10.1038/s41598-024-57207-7] [Medline: 38521806]

72. El Emam K, Mosquera L, Fang X, El-Hussuna A. Utility metrics for evaluating synthetic health data generation methods: validation study. JMIR Med Inform. Apr 07, 2022;10(4):e35734. [FREE Full text] [doi: 10.2196/35734] [Medline: 35389366]

73. National Academies of Sciences, Engineering, and Medicine. Reproducibility and Replicability in Science. Washington, DC. The National Academies Press; 2019.

74. El Kababji S, Mitsakakis N, Fang X, Beltran-Bless AA, Pond G, Vandermeer L, et al. Evaluating the utility and privacy of synthetic breast cancer clinical trial data sets. JCO Clin Cancer Inform. Nov 27, 2023;7. [doi: 10.1200/cci.23.00116]

75. Franklin JM, Pawar A, Martin D, Glynn RJ, Levenson M, Temple R, et al. Nonrandomized real-world evidence to support regulatory decision making: process for a randomized trial replication project. Clin Pharmacol Ther. Apr 2020;107(4):817-826. [FREE Full text] [doi: 10.1002/cpt.1633] [Medline: 31541454]

76. Karr AF, Kohnen CN, Oganian A, Reiter JP, Sanil AP. A framework for evaluating the utility of data altered to protect confidentiality. Am Stat. Aug 2006;60(3):224-232. [doi: 10.1198/000313006x124640]

77. Open Science Collaboration. PSYCHOLOGY. Estimating the reproducibility of psychological science. Science. Aug 28, 2015;349(6251):aac4716. [doi: 10.1126/science.aac4716] [Medline: 26315443]

78. Patil P, Peng RD, Leek JT. What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. Perspect Psychol Sci. Jul 29, 2016;11(4):539-544. [FREE Full text] [doi: 10.1177/1745691616646366] [Medline: 27474140]

79. Camerer CF, Dreber A, Holzmeister F, Ho TH, Huber J, Johannesson M, et al. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nat Hum Behav. Sep 27, 2018;2(9):637-644. [doi: 10.1038/s41562-018-0399-z] [Medline: 31346273]

80. Crown W, Dahabreh IJ, Li X, Toh S, Bierer B. Can observational analyses of routinely collected data emulate randomized trials? Design and feasibility of the observational patient evidence for regulatory approval science and understanding disease project. Value Health. Feb 2023;26(2):176-184. [FREE Full text] [doi: 10.1016/j.jval.2022.07.003] [Medline: 35970705]

81. Yoon D, Jeong HE, Park S, You SC, Bang SM, Shin JY. Real-world data emulating randomized controlled trials of non-vitamin K antagonist oral anticoagulants in patients with venous thromboembolism. BMC Med. Sep 29, 2023;21(1):375. [FREE Full text] [doi: 10.1186/s12916-023-03069-1] [Medline: 37775786]

82. Wang SV, Schneeweiss S, RCT-DUPLICATE Initiative, Franklin JM, Desai RJ, Feldman W, et al. Emulation of randomized clinical trials with nonrandomized database analyses: results of 32 clinical trials. JAMA. Apr 25, 2023;329(16):1376-1385. [FREE Full text] [doi: 10.1001/jama.2023.4221] [Medline: 37097356]

83. Franklin JM, Patorno E, Desai RJ, Glynn RJ, Martin D, Quinto K, et al. Emulating randomized clinical trials with nonrandomized real-world evidence studies. Circulation. Mar 09, 2021;143(10):1002-1013. [doi: 10.1161/circulationaha.120.051718]

84. Clemons M, Dranitsaris G, Sienkiewicz M, Sehdev S, Ng T, Robinson A, 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. Dec 2020;54:278-285. [FREE Full text] [doi: 10.1016/j.breast.2020.11.002] [Medline: 33242754]

85. Clemons M, Ong M, Stober C, Ernst S, Booth C, Canil 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. Jan 2021;142:132-140. [FREE Full text] [doi: 10.1016/j.ejca.2020.08.019] [Medline: 33023785]

86. Goss PE, Ingle JN, Pritchard KI, Ellis MJ, Sledge GW, Budd GT, 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. Apr 10, 2013;31(11):1398-1404. [FREE Full text] [doi: 10.1200/JCO.2012.44.7805] [Medline: 23358971]

87. Paterson AH, Anderson SJ, Lembersky BC, Fehrenbacher L, Falkson CI, King KM, 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. Jul 2012;13(7):734-742. [doi: 10.1016/s1470-2045(12)70226-7]

88. Clemons M, Fergusson D, Simos D, Mates M, Robinson A, Califaretti N, 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. Jul 2020;31(7):951-957. [FREE Full text] [doi: 10.1016/j.annonc.2020.04.005] [Medline: 32325257]

89. Clemons M, Stober C, Kehoe A, Bedard D, MacDonald F, Brunet MC, et al. A randomized trial comparing vascular access strategies for patients receiving chemotherapy with trastuzumab for early-stage breast cancer. Support Care Cancer. Oct 30, 2020;28(10):4891-4899. [doi: 10.1007/s00520-020-05326-y] [Medline: 32002617]

90. Gnant M, Mlineritsch B, Schippinger W, Luschin-Ebengreuth G, Pöstlberger S, Menzel C, et al. Endocrine therapy plus zoledronic acid in premenopausal breast cancer. N Engl J Med. Feb 12, 2009;360(7):679-691. [doi: 10.1056/nejmoa0806285]

91. Awan A, Ng T, Conter H, Raskin W, Stober C, Simos D, 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. Feb 2021;26:100343. [FREE Full text] [doi: 10.1016/j.jbo.2020.100343] [Medline: 33425673]

92. Gralow JR, Barlow WE, Paterson AH, M'iao JL, Lew DL, Stopeck AT, et al. Phase III randomized trial of bisphosphonates as adjuvant therapy in breast cancer: s0307. J Natl Cancer Inst. Jul 01, 2020;112(7):698-707. [FREE Full text] [doi: 10.1093/jnci/djz215] [Medline: 31693129]

93. Wang Z, Draghi B, Rotalinti Y, Lunn D, Myles P. High-fidelity synthetic data applications for data augmentation. In: Domínguez-Morales MJ, Civit-Masot J, Damaševičius R, Muñoz-Saavedra L, Damaševičius R, Engelbrecht A, editors. Deep Learning - Recent Findings and Research. London, UK. IntechOpen; 2024.

94. Myles P, Ordish J, Tucker A. The potential synergies between synthetic data and in silico trials in relation to generating representative virtual population cohorts. Prog Biomed Eng. Jan 24, 2023;5:013001. [doi: 10.1088/2516-1091/acafbf]

95.   Das T, Wang Z, Sun J. TWIN: personalized clinical trial digital twin generation. In: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2023. Presented at: KDD '23; August 6-10, 2023; Long Beach, CA. [doi: 10.1145/3580305.3599534]

96.   Gootjes-Dreesbach L, Sood M, Sahay A, Hofmann-Apitius M, Fröhlich H. Variational autoencoder modular Bayesian networks for simulation of heterogeneous clinical study data. Front Big Data. May 28, 2020;3:16. [FREE Full text] [doi: 10.3389/fdata.2020.00016] [Medline: 33693390]

97.   Akiya I, Ishihara T, Yamamoto K. Comparison of synthetic data generation techniques for control group survival data in oncology clinical trials: simulation study. JMIR Med Inform. Jun 18, 2024;12:e55118. [FREE Full text] [doi: 10.2196/55118] [Medline: 38889082]

98.   Azizi Z, Zheng C, Mosquera L, Pilote L, El Emam K, GOING-FWD Collaborators. Can synthetic data be a proxy for real clinical trial data? A validation study. BMJ Open. Apr 16, 2021;11(4):e043497. [FREE Full text] [doi: 10.1136/bmjopen-2020-043497] [Medline: 33863713]

99.   Krenmayr L, Frank R, Drobig C, Braungart M, Seidel J, Schaudt D, et al. GANerAid: realistic synthetic patient data for clinical trials. Inform Med Unlock. 2022;35:101118. [doi: 10.1016/j.imu.2022.101118]

100.   Beaulieu-Jones BK, Wu ZS, Williams C, Lee R, Bhavnani SP, Byrd JB, et al. Privacy-preserving generative deep neural networks support clinical data sharing. Circ Cardiovasc Qual Outcomes. Jul 2019;12(7):e005122. [FREE Full text] [doi: 10.1161/CIRCOUTCOMES.118.005122] [Medline: 31284738]

101.   Beigi M, Shafquat A, Mezey J, Aptekar J. Synthetic Clinical Trial Data while Preserving Subject-Level Privacy. In: Proceedings of the Synthetic Data for Empowering ML Research. 2022. Presented at: NeurIPS 2022 SyntheticData4ML; December 2, 2022; New Orleans, LA. URL: https://neurips.cc/virtual/2022/58683

102.   Eckardt JN, Hahn W, Röllig C, Stasik S, Platzbecker U, Müller-Tidow C, et al. Mimicking clinical trials with synthetic acute myeloid leukemia patients using generative artificial intelligence. NPJ Digit Med. Mar 20, 2024;7(1):76. [FREE Full text] [doi: 10.1038/s41746-024-01076-x] [Medline: 38509224]

103.   Latner J, Neunhoeffer M, Drechsler J. Generating synthetic data is complicated: know your data and know your generator. In: Proceedings of the International Conference on Privacy in Statistical Databases. 2024. Presented at: PSD 2024; September 25-27, 2024; Antibes Juan-les-Pins, France. [doi: 10.1007/978-3-031-69651-0_8]

104.   Fössing E, Drechsler J. An evaluation of synthetic data generators implemented in the python library synthcity. In: Proceedings of the Privacy in Statistical Databases: International Conference. 2024. Presented at: PSD 2024; September 25-27, 2024; Antibes Juan-les-Pins, France. [doi: 10.1007/978-3-031-69651-0_12]

105.   Slokom M, Agrawal S, Krol NC, de Wolf PP. Relational or single: a comparative analysis of data synthesis approaches for privacy and utility on a use case from statistical office. In: Proceedings of the International Conference on Privacy in Statistical Databases. 2024. Presented at: PSD 2024; September 25-27, 2024; Antibes Juan-les-Pins, France. [doi: 10.1007/978-3-031-69651-0_27]

106.   Huynh L, Johns B, Liu SH, Vedula SS, Li T, Puhan MA. Cost-effectiveness of health research study participant recruitment strategies: a systematic review. Clin Trials. Oct 30, 2014;11(5):576-583. [FREE Full text] [doi: 10.1177/1740774514540371] [Medline: 24980278]

107.   Moore TJ, Heyward J, Anderson G, Alexander GC. Variation in the estimated costs of pivotal clinical benefit trials supporting the US approval of new therapeutic agents, 2015-2017: a cross-sectional study. BMJ Open. Jun 11, 2020;10(6):e038863. [FREE Full text] [doi: 10.1136/bmjopen-2020-038863] [Medline: 32532786]

108.   Fireman BH, Fehrenbacher L, Gruskin EP, Ray GT. Cost of care for patients in cancer clinical trials. J Natl Cancer Inst. Jan 19, 2000;92(2):136-142. [doi: 10.1093/jnci/92.2.136] [Medline: 10639515]

109.   Anand A. Study reveals serious cancer research gaps for Black Canadians. CBC News. Feb 7, 2023. URL: https://www.cbc.ca/news/canada/ottawa/black-canadian-cancer-research-gap-ottawa-problem-1.6737797 [accessed 2024-02-07]

110.   Cénat JM, Dromer É, Darius WP, Dalexis RD, Furyk SE, Poisson H, et al. Incidence, factors, and disparities related to cancer among Black individuals in Canada: a scoping review. Cancer. Feb 01, 2023;129(3):335-355. [FREE Full text] [doi: 10.1002/cncr.34551] [Medline: 36436148]

111.   Gross AS, Harry AC, Clifton CS, Della Pasqua O. Clinical trial diversity: an opportunity for improved insight into the determinants of variability in drug response. Br J Clin Pharmacol. Jun 2022;88(6):2700-2717. [FREE Full text] [doi: 10.1111/bcp.15242] [Medline: 35088432]

112.   Ford JG, Howerton MW, Lai GY, Gary TL, Bolen S, Gibbons MC, et al. Barriers to recruiting underrepresented populations to cancer clinical trials: a systematic review. Cancer. Jan 15, 2008;112(2):228-242. [FREE Full text] [doi: 10.1002/cncr.23157] [Medline: 18008363]

113.   Khan MS, Shahid I, Siddiqi TJ, Khan SU, Warraich HJ, Greene SJ, et al. Ten-year trends in enrollment of women and minorities in pivotal trials supporting recent US Food and Drug Administration approval of novel cardiometabolic drugs. J Am Heart Assoc. Jun 02, 2020;9(11):e015594. [FREE Full text] [doi: 10.1161/JAHA.119.015594] [Medline: 32427023]

114.   Hussain-Gambles M, Atkin K, Leese B. Why ethnic minority groups are under-represented in clinical trials: a review of the literature. Health Soc Care Community. Sep 2004;12(5):382-388. [doi: 10.1111/j.1365-2524.2004.00507.x] [Medline: 15373816]

115. Evelyn B, Toigo T, Banks D, Pohl D, Gray K, Robins B, et al. Participation of racial/ethnic groups in clinical trials and race-related labeling: a review of new molecular entities approved 1995-1999. J Natl Med Assoc. Dec 2001;93(12 Suppl):18S-24S. [Medline: 11798060]

116. George S, Duran N, Norris K. A systematic review of barriers and facilitators to minority research participation among African Americans, Latinos, Asian Americans, and Pacific Islanders. Am J Public Health. Feb 2014;104(2):e16-e31. [doi: 10.2105/ajph.2013.301706]

117. UyBico SJ, Pavel S, Gross CP. Recruiting vulnerable populations into research: a systematic review of recruitment interventions. J Gen Intern Med. Jun 21, 2007;22(6):852-863. [FREE Full text] [doi: 10.1007/s11606-007-0126-3] [Medline: 17375358]

118. Ramamoorthy A, Pacanowski MA, Bull J, Zhang L. Racial/ethnic differences in drug disposition and response: review of recently approved drugs. Clin Pharmacol Ther. Mar 20, 2015;97(3):263-273. [doi: 10.1002/cpt.61] [Medline: 25669658]

119. Tannenbaum C, Day D, Matera Alliance. Age and sex in drug development and testing for adults. Pharmacol Res. Jul 2017;121:83-93. [FREE Full text] [doi: 10.1016/j.phrs.2017.04.027] [Medline: 28455265]

120. Vitale C, Fini M, Spoletini I, Lainscak M, Seferovic P, Rosano GM. Under-representation of elderly and women in clinical trials. Int J Cardiol. Apr 01, 2017;232:216-221. [doi: 10.1016/j.ijcard.2017.01.018] [Medline: 28111054]

121. Ruiter R, Burggraaf J, Rissmann R. Under-representation of elderly in clinical trials: an analysis of the initial approval documents in the Food and Drug Administration database. Br J Clin Pharmacol. Apr 23, 2019;85(4):838-844. [FREE Full text] [doi: 10.1111/bcp.13876] [Medline: 30681181]

122. Crome P, Cherubini A, Oristrell J. The PREDICT (increasing the participation of the elderly in clinical trials) study: the charter and beyond. Expert Rev Clin Pharmacol. Jul 2014;7(4):457-468. [doi: 10.1586/17512433.2014.922864] [Medline: 24939469]

123. Heidari S, Babor TF, de Castro P, Tort S, Curno M. Sex and Gender Equity in Research: rationale for the SAGER guidelines and recommended use. Res Integr Peer Rev. 2016;1:2. [FREE Full text] [doi: 10.1186/s41073-016-0007-6] [Medline: 29451543]

124. Acharya KP, Pathak S. Applied research in low-income countries: why and how? Front Res Metr Anal. Nov 14, 2019;4:3. [FREE Full text] [doi: 10.3389/frma.2019.00003] [Medline: 33870035]

125. Bhanot K, Qi M, Erickson JS, Guyon I, Bennett KP. The problem of fairness in synthetic healthcare data. Entropy (Basel). Sep 04, 2021;23(9):1165. [FREE Full text] [doi: 10.3390/e23091165] [Medline: 34573790]

126. Wyllie S, Shumailov I, Papernot N. Fairness feedback loops: training on synthetic data amplifies bias. arXiv. Preprint posted online on March 12, 2024. [FREE Full text] [doi: 10.1145/3630106.3659029]

127. Juwara L, El-Hussuna A, El Emam K. An evaluation of synthetic data augmentation for mitigating covariate bias in health data. Patterns (N Y). Apr 12, 2024;5(4):100946. [FREE Full text] [doi: 10.1016/j.patter.2024.100946] [Medline: 38645766]

128. Chen F, Wang L, Hong J, Jiang J, Zhou L. Unmasking bias in artificial intelligence: a systematic review of bias detection and mitigation strategies in electronic health record-based models. J Am Med Inform Assoc. Apr 19, 2024;31(5):1172-1183. [doi: 10.1093/jamia/ocae060] [Medline: 38520723]

129. Theodorou B, Danek B, Tummala V, Kumar SP, Malin B, Sun J. FairPlay: improving medical machine learning models with generative balancing for equity and excellence. Research Square. Preprint posted online on December 13, 2024. [FREE Full text] [doi: 10.21203/rs.3.rs-5252769/v1]

130. Landscape report and industry survey on the use of computational modeling and simulation in medical device development. Medical Device Innovation Consortium. Feb 16, 2024. URL: https://mdic.org/resources/landscape-report-industry-survey-on-the-use-of-computational-modeling-simulation-in-medical-device-development/ [accessed 2025-01-07]

131. Redrup Hill E, Mitchell C, Myles P, Branson R, Frangi AF. Cross-regulator workshop: journeys, experiences and best practices on computer modelled and simulated regulatory evidence— workshop report. InSilicoUK Pro-Innovation Regulations Network. 2023. URL: https://www.pankhurst.manchester.ac.uk/wp-content/uploads/sites/278/2024/01/Cross_Regulator_Workshop_on_CMS_Report_December_2023_final.pdf [accessed 2025-01-07]

132. Viceconti M, Emili L. Toward Good Simulation Practice: Best Practices for the Use of Computational Modelling and Simulation in the Regulatory Process of Biomedical Products. Cham, Switzerland. Springer; 2024.

133. Warraich HJ, Tazbaz T, Califf RM. FDA perspective on the regulation of artificial intelligence in health care and biomedicine. JAMA. Jan 21, 2025;333(3):241-247. [doi: 10.1001/jama.2024.21451] [Medline: 39405330]

134. Artificial intelligence: an accountability framework for federal agencies and other entities. General Accountability Office. Jun 30, 2021. URL: https://www.gao.gov/products/gao-21-519sp [accessed 2024-02-07]

135. Using artificial intelligence and machine learning in the development of drug and biological products. U.S. Food & Drug Administration. URL: https://www.fda.gov/media/167973/download [accessed 2024-02-07]

136. Alloza C, Knox B, Raad H, Aguilà M, Coakley C, Mohrova Z, et al. A case for synthetic data in regulatory decision-making in Europe. Clin Pharmacol Ther. Oct 24, 2023;114(4):795-801. [doi: 10.1002/cpt.3001] [Medline: 37441734]

137. Myles P, Ordish J, Branson R. Synthetic data and the innovation, assessment, and regulation of AI medical devices. Clinical Practice Research Datalink. 2022. URL: https://www.cprd.com/sites/default/files/2022-02/Synthetic%20data_RA%20proof.pdf [accessed 2024-02-07]

XSL•FO

RenderX

138. Newington L, Metcalfe A. Factors influencing recruitment to research: qualitative study of the experiences and perceptions of research teams. BMC Med Res Methodol. Jan 23, 2014;14(1):10. [FREE Full text] [doi: 10.1186/1471-2288-14-10] [Medline: 24456229]

139. Price D, Edwards M, Carson-Stevens A, Cooper A, Davies F, Evans B, et al. Challenges of recruiting emergency department patients to a qualitative study: a thematic analysis of researchers' experiences. BMC Med Res Methodol. Jun 11, 2020;20(1):151. [FREE Full text] [doi: 10.1186/s12874-020-01039-2] [Medline: 32522265]

140. Friede T, Posch M, Zohar S, Alberti C, Benda N, Comets E, et al. Recent advances in methodology for clinical trials in small populations: the InSPiRe project. Orphanet J Rare Dis. Oct 25, 2018;13(1):186. [FREE Full text] [doi: 10.1186/s13023-018-0919-y] [Medline: 30359266]

141. Reilly K. Sharing data for the benefit of children with cancer. BioMed Central. Sep 9, 2019. URL: https://blogs.biomedcentral.com/on-medicine/2019/09/09/sharing-data-for-the-benefit-of-children-with-cancer/ [accessed 2020-10-02]

142. Bavdekar SB. Pediatric clinical trials. Perspect Clin Res. Jan 2013;4(1):89-99. [FREE Full text] [doi: 10.4103/2229-3485.106403] [Medline: 23533990]

143. Campbell H, Surry SA, Royle EM. A review of randomised controlled trials published in Archives of Disease in Childhood from 1982-96. Arch Dis Child. Aug 01, 1998;79(2):192-197. [FREE Full text] [doi: 10.1136/adc.79.2.192] [Medline: 9797608]

144. Silverman WA, Altman DG. Patients' preferences and randomised trials. Lancet. Jan 20, 1996;347(8995):171-174. [FREE Full text] [doi: 10.1016/s0140-6736(96)90347-5] [Medline: 8544555]

145. Principles for responsible clinical trial data sharing: our commitment to patients and researchers. European Federation of Pharmaceutical Industries and Associations. 2023. URL: https://www.efpia.eu/media/qndlfduy/phrmaefpiaprinciplesforresponsibledatasharing2023.pdf [accessed 2024-02-07]

146. European Medicines Agency policy on publication of clinical data for medicinal products for human use. European Medicines Agency. Mar 21, 2019. URL: https://www.ema.europa.eu/en/documents/other/policy-70-european-medicines-agency-policy-publication-clinical-data-medicinal-products-human-use_en.pdf [accessed 2024-01-11]

147. Taichman DB, Backus J, Baethge C, Bauchner H, de Leeuw PW, Drazen JM, et al. Sharing clinical trial data: a proposal from the International Committee of Medical Journal Editors. Ann Intern Med. Jan 26, 2016;164(7):505-506. [doi: 10.7326/m15-2928]

148. Institute of Medicine. Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk. Washington, DC. The National Academies Press; 2015.

149. Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals. International Committee of Medical Journal Editors. Jan 2015. URL: https://www.icmje.org/icmje-recommendations.pdf [accessed 2024-02-07]

150. Data, software and materials management and sharing policy. Wellcome. URL: https://wellcome.org/grant-funding/guidance/policies-grant-conditions/data-software-materials-management-and-sharing-policy [accessed 2024-02-07]

151. Final NIH statement on sharing research data. National Institutes of Health. Feb 26, 2003. URL: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-03-032.html [accessed 2020-06-29]

152. Protection of personal data in clinical documents – a model approach. TransCelerate Biopharma Inc. 2016. URL: http://www.transceleratebiopharmainc.com/wp-content/uploads/2017/02/Protection-of-Personal-Data-in-Clinical-Documents.pdf [accessed 2024-02-07]

153. Data de-identification and anonymization of individual patient data in clinical studies — a model approach. TransCelerate Biopharma Inc. 2016. URL: http://www.transceleratebiopharmainc.com/wp-content/uploads/2015/04/TransCelerate-De-identification-and-Anonymization-of-Individual-Patient-Data-in-Clinical-Studies-V2.0.pdf [accessed 2024-02-07]

154. Doshi P. Data too important to share: do those who control the data control the message? BMJ. Mar 02, 2016;352:i1027. [doi: 10.1136/bmj.i1027] [Medline: 26936297]

155. Rankin D, Black M, Bond R, Wallace J, Mulvenna M, Epelde G. Reliability of supervised machine learning using synthetic data in health care: model to preserve privacy for data sharing. JMIR Med Inform. Jul 20, 2020;8(7):e18910. [FREE Full text] [doi: 10.2196/18910] [Medline: 32501278]

156. Ventresca M, Schünemann HJ, Macbeth F, Clarke M, Thabane L, Griffiths G, et al. Obtaining and managing data sets for individual participant data meta-analysis: scoping review and practical guide. BMC Med Res Methodol. May 12, 2020;20(1):113. [FREE Full text] [doi: 10.1186/s12874-020-00964-6] [Medline: 32398016]

157. 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. Jun 2018;98:157-159. [doi: 10.1016/j.jclinepi.2017.12.014] [Medline: 29288135]

158. Naudet F, Sakarovitch C, Janiaud P, Cristea I, Fanelli D, Moher D, 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. Feb 13, 2018;360:k400. [FREE Full text] [doi: 10.1136/bmj.k400] [Medline: 29440066]

159. Nevitt SJ, Marson AG, Davie B, Reynolds S, Williams L, Smith CT. Exploring changes over time and characteristics associated with data retrieval across individual participant data meta-analyses: systematic review. BMJ. Apr 05, 2017;357:j1390. [FREE Full text] [doi: 10.1136/bmj.j1390] [Medline: 28381561]

160. Villain B, Dechartres A, Boyer P, Ravaud P. Feasibility of individual patient data meta-analyses in orthopaedic surgery. BMC Med. Jun 03, 2015;13(1):131. [FREE Full text] [doi: 10.1186/s12916-015-0376-6] [Medline: 26040278]

161. Iqbal SA, Wallach JD, Khoury MJ, Schully SD, Ioannidis JP. Reproducible research practices and transparency across the biomedical literature. PLoS Biol. Jan 4, 2016;14(1):e1002333. [FREE Full text] [doi: 10.1371/journal.pbio.1002333] [Medline: 26726926]

162. Expert advisory group report 2: building Canada's health data foundation. Government of Canada. Nov 2021. URL: https://tinyurl.com/zcsakfv9 [accessed 2024-02-07]

163. Read KB, Ganshorn H, Rutley S, Scott DR. Data-sharing practices in publications funded by the Canadian Institutes of Health Research: a descriptive analysis. CMAJ Open. Nov 09, 2021;9(4):E980-E987. [FREE Full text] [doi: 10.9778/cmajo.20200303] [Medline: 34753787]

164. National Academy of Medicine. Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril. Washington, DC. The National Academies Press; 2022.

165. van Panhuis WG, Paul P, Emerson C, Grefenstette J, Wilder R, Herbst AJ, et al. A systematic review of barriers to data sharing in public health. BMC Public Health. Nov 05, 2014;14:1144. [FREE Full text] [doi: 10.1186/1471-2458-14-1144] [Medline: 25377061]

166. Kalkman S, Mostert M, Gerlinger C, van Delden JJ, van Thiel GJ. Responsible data sharing in international health research: a systematic review of principles and norms. BMC Med Ethics. Mar 28, 2019;20(1):21. [FREE Full text] [doi: 10.1186/s12910-019-0359-9] [Medline: 30922290]

167. Reiter JP. New approaches to data dissemination: a glimpse into the future (?). CHANCE. Sep 20, 2012;17(3):11-15. [doi: 10.1080/09332480.2004.10554907]

168. Park N, Mohammadi M, Gorde K, Jajodia S, Park H, Kim Y. Data synthesis based on generative adversarial networks. Proc VLDB Endow. Jun 01, 2018;11(10):1071-1083. [doi: 10.14778/3231751.3231757]

169. Hu J. Bayesian estimation of attribute and identification disclosure risks in synthetic data. arXiv. Preprint posted online on April 9, 2018. [FREE Full text]

170. Taub J, Elliot M, Pampaka M, Smith D. Differential correct attribution probability for synthetic data: an exploration. In: Proceedings of the Privacy in Statistical Databases. 2018. Presented at: PSD 2018; September 26-28, 2018; Valencia, Spain. [doi: 10.1007/978-3-319-99771-1_9]

171. Hu J, Reiter JP, Wang Q. Disclosure risk evaluation for fully synthetic categorical data. In: Proceedings of the Privacy in Statistical Databases. 2014. Presented at: PSD 2014; September 17-19, 2014; Ibiza, Spain. [doi: 10.1007/978-3-319-11257-2_15]

172. Wei L, Reiter JP. Releasing synthetic magnitude microdata constrained to fixed marginal totals. Stat J IAOS. Feb 27, 2016;32(1):93-108. [doi: 10.3233/sji-160959]

173. Ruiz N, Muralidhar K, Domingo-Ferrer J. On the privacy guarantees of synthetic data: a reassessment from the maximum-knowledge attacker perspective. In: Proceedings of the Privacy in Statistical Databases. 2018. Presented at: PSD 2018; September 26-28, 2018; Valencia, Spain. [doi: 10.1007/978-3-319-99771-1_5]

174. Reiter JP. Releasing multiply imputed, synthetic public use microdata: an illustration and empirical study. J R Stat Soc Ser A Stat Soc. Jan 2005;168(1):185-205. [doi: 10.1111/j.1467-985X.2004.00343.x]

175. El Emam K, Mosquera L, Fang X. Validating a membership disclosure metric for synthetic health data. JAMIA Open. Dec 2022;5(4):ooac083. [FREE Full text] [doi: 10.1093/jamiaopen/ooac083] [Medline: 36238080]

176. El Emam K, Mosquera L, Bass J. Evaluating identity disclosure risk in fully synthetic health data: model development and validation. J Med Internet Res. Nov 16, 2020;22(11):e23139. [FREE Full text] [doi: 10.2196/23139] [Medline: 33196453]

177. Arora A, Wagner SK, Carpenter R, Jena R, Keane PA. The urgent need to accelerate synthetic data privacy frameworks for medical research. Lancet Digit Health. Feb 2025;7(2):e157-e160. [doi: 10.1016/s2589-7500(24)00196-1]

178. El Emam K, Mosquera L, Hoptroff R. Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data. Sebastopol, CA. O'Reilly Media; 2020.

179. Beduschi A. Synthetic data protection: towards a paradigm change in data regulation? Big Data Soc. Feb 14, 2024;11(1). [doi: 10.1177/20539517241231277]

180. Lun R, Siegal D, Ramsay T, Stotts G, Dowlatshahi D. Synthetic data in cancer and cerebrovascular disease research: a novel approach to big data. PLoS One. Feb 7, 2024;19(2):e0295921. [FREE Full text] [doi: 10.1371/journal.pone.0295921] [Medline: 38324588]

181. Rousseau O, Karakachoff M, Gaignard A, Bellanger L, Bijlenga P, Constant Dit Beaufils P, et al. Location of intracranial aneurysms is the main factor associated with rupture in the ICAN population. J Neurol Neurosurg Psychiatry. Feb 23, 2021;92(2):122-128. [doi: 10.1136/jnnp-2020-324371] [Medline: 33097563]

182. Guillaudeux M, Rousseau O, Petot J, Bennis Z, Dein CA, Goronflot T, et al. Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. NPJ Digit Med. Mar 10, 2023;6(1):37. [FREE Full text] [doi: 10.1038/s41746-023-00771-5] [Medline: 36899082]

183. Thomas A, Jaffré S, Guardiolle V, Perennec T, Gagnadoux F, Goupil F, et al. Does PaCO correction have an impact on survival of patients with chronic respiratory failure and long-term non-invasive ventilation? Heliyon. Feb 29, 2024;10(4):e26437. [FREE Full text] [doi: 10.1016/j.heliyon.2024.e26437] [Medline: 38420381]

184. pysdg. Open Science Framework Home. 2024. URL: https://osf.io/xj9pr/ [accessed 2025-02-20]

Abbreviations

ABCSG: Austrian Breast and Colorectal Cancer Study Group
BTA: bone-targeted agents
CTGAN: conditional tabular generative adversarial network
GAN: generative adversarial network
HER2+: human epidermal growth factor receptor-2 positive
REaCT: Rethinking Clinical Trials
RWD: real-world data
VAE: variational autoencoder
ZOL: zoledronate

XSL•FO
RenderX