Original Paper

Comparison of Synthetic Data Generation Techniques for Control Group Survival Data in Oncology Clinical Trials: Simulation Study

Ippei Akiya[1], MSc; Takuma Ishihara[2], PhD; Keiichi Yamamoto[3], PhD

[1]Biometrics, ICON Clinical Research GK, Tokyo, Japan

[2]Innovative and Clinical Research Promotion Center, Gifu University Hospital, Gifu, Japan

[3]Division of Data Science, Center for Industrial Research and Innovation, Translational Research Institute for Medical Innovation, Osaka Dental University, Osaka, Japan

Corresponding Author:
Ippei Akiya, MSc
Biometrics
ICON Clinical Research GK
Sumitomo-ShinTranomon Building, 4-3-9 Toranomon Minato-ku
Tokyo, 105-0001
Japan
Phone: 81 3-4510-4933
Email: ippei.akiya@gmail.com

Abstract

Background: Synthetic patient data (SPD) generation for survival analysis in oncology trials holds significant potential for accelerating clinical development. Various machine learning methods, including classification and regression trees (CART), random forest (RF), Bayesian network (BN), and conditional tabular generative adversarial network (CTGAN), have been used for this purpose, but their performance in reflecting actual patient survival data remains under investigation.

Objective: The aim of this study was to determine the most suitable SPD generation method for oncology trials, specifically focusing on both progression-free survival (PFS) and overall survival (OS), which are the primary evaluation end points in oncology trials. To achieve this goal, we conducted a comparative simulation of 4 generation methods, including CART, RF, BN, and the CTGAN, and the performance of each method was evaluated.

Methods: Using multiple clinical trial data sets, 1000 data sets were generated by using each method for each clinical trial data set and evaluated as follows: (1) median survival time (MST) of PFS and OS; (2) hazard ratio distance (HRD), which indicates the similarity between the actual survival function and a synthetic survival function; and (3) visual analysis of Kaplan-Meier (KM) plots. Each method's ability to mimic the statistical properties of real patient data was evaluated from these multiple angles.

Results: In most simulation cases, CART demonstrated the high percentages of MSTs for synthetic data falling within the 95% CI range of the MST of the actual data. These percentages ranged from 88.8% to 98.0% for PFS and from 60.8% to 96.1% for OS. In the evaluation of HRD, CART revealed that HRD values were concentrated at approximately 0.9. Conversely, for the other methods, no consistent trend was observed for either PFS or OS. CART demonstrated better similarity than RF, in that CART caused overfitting and RF (a kind of ensemble learning approach) prevented it. In SPD generation, the statistical properties close to the actual data should be the focus, not a well-generalized prediction model. Both the BN and CTGAN methods cannot accurately reflect the statistical properties of the actual data because small data sets are not suitable.

Conclusions: As a method for generating SPD for survival data from small data sets, such as clinical trial data, CART demonstrated to be the most effective method compared to RF, BN, and CTGAN. Additionally, it is possible to improve CART-based generation methods by incorporating feature engineering and other methods in future work.

JMIR Med Inform 2024;12:e55118; doi: 10.2196/55118

Keywords: oncology clinical trial; survival analysis; synthetic patient data; machine learning; SPD; simulation

Introduction

When submitting an application for the approval of a new pharmaceutical product to health authorities, it is imperative to demonstrate its efficacy and safety through multiple clinical trials. However, 86% of these trials encounter difficulties meeting the targeted number of subjects within the designated recruitment period, often leading to extensions of the trial duration or completion of the trial without reaching the target number of subjects [1-3]. The challenge of patient recruitment not only delays the submission of regulatory applications but also hinders the timely provision of effective treatment to patients, which consequently contributes to increased development costs and the escalation of drug prices and potentially exacerbates the strain on health care financing.

In recent years, the use of real-world data (RWD) has emerged as a potential solution for addressing these issues. The Food and Drug Administration has also released draft guidelines [4], garnering attention on the application of RWD as an external control arm in clinical trials [5,6]. Furthermore, it has been reported that it is possible to optimize eligibility using RWD and machine learning, thereby increasing the number of eligible subjects that can be included [7].

In addition to these approaches, we hypothesize that it is possible to generate synthetic patient data (SPD) from control arm data in past clinical trials and use it to establish a control arm for a new clinical trial. The use of SPD, an emerging research approach in the health care research field [8-17], involves the generation of fictitious individual patient-level data from real data, which possess statistical properties similar to those of actual data. This approach is anticipated to facilitate health care research while addressing data privacy concerns [14,18-21].

Regarding its application in clinical trials, concerns have been raised about the feasibility of generating SPDs with statistical properties similar to those of actual data due to the relatively smaller volume of clinical trial data compared to RWD, such as electronic health records or registry data. However, previous studies [22-25] have reported the successful generation of SPDs with statistical properties generally comparable to the actual data, although there are certain limitations. Additionally, with the expansion of clinical trial data-sharing platforms such as ClinicalStudyDataRequest.com, Project Data Sphere, and Vivli, acquiring subject-level clinical trial data has become more accessible. Consequently, advancements in research on the utility of SPD and the expansion of clinical trial data-sharing platforms are expected to have potential applications in clinical trials.

Our focus lies in the application of this technology in oncology clinical trials that evaluate popular efficacy end points such as overall survival (OS) and progression-free survival (PFS)–related survival functions and median survival time (MST) [26]. In previous studies on SPD, there has been a notable emphasis on reporting patient background data and single–time point data [22-25]. However, research focusing specifically on the relationship between SPD and survival data remains relatively insufficient [27].

As the first step in examining our hypothesis that the use of SPD can be beneficial in accelerating health care research, the aim of this study was to determine the most suitable SPD generation method for oncology trials, specifically focusing on both OS and PFS, which are set as the primary evaluation end points in oncology trials. To achieve this goal, we conducted a comparative simulation of 4 generation methods: classification and regression trees (CART) [28], random forest (RF) [29], Bayesian network (BN) [30], and the conditional tabular generative adversarial network (CTGAN) approach [31], and the performance of each method was evaluated.

Methods

Overview

To generate the SPD, subject-level clinical trial data were obtained from Project Data Sphere for the following 4 clinical trials (Table 1): (1) each had a different cancer type, (2) included control arm data, (3) contained both OS and PFS data, and (4) had a ready data format for analysis.

Table 1. List of selected oncology clinical trials in this study.

ClinicalTrials.gov ID	Titles	Phase	Cancer type	Intervention for the control arm	Subjects in the control arm, n
NCT00119613	A Randomized, Double-Blind, Placebo-Controlled Study of Subjects With Previously Untreated Extensive-Stage Small-Cell Lung Cancer (SCLC) Treated With Platinum Plus Etoposide Chemotherapy With or Without Darbepoetin Alfa.	III	Small cell lung cancer	Placebo	232
NCT00339183	A Randomized, Multicenter Phase 3 Study to Compare the Efficacy of Panitumumab in Combination With Chemotherapy to the Efficacy of Chemotherapy Alone in Patients With Previously Treated Metastatic Colorectal Cancer.	III	Metastatic colorectal cancer	FOLFIRI[a] Alone	476

ClinicalTrials.gov ID	Titles	Phase	Cancer type	Intervention for the control arm	Subjects in the control arm, n
NCT00339183	A Phase 3 Randomized Trial of Chemotherapy With or Without Panitumumab in Patients With Metastatic and/or Recurrent Squamous Cell Carcinoma of the Head and Neck (SCCHN).	III	Recurrent or metastatic (or both) head and neck cancer	Cisplatin and 5-fluorouracil	260
NCT00703326	A Multicenter, Multinational, Randomized, Double-Blind, Phase III Study of IMC-1121B Plus Docetaxel versus Placebo Plus Docetaxel in Previously Untreated Patients With HER2-Negative, Unresectable, Locally-Recurrent or Metastatic Breast Cancer.	III	Breast cancer	Placebo and docetaxel	382

[a]FOLFIRI: panitumumab plus fluorouracil, leucovorin, and irinotecan.

Preparation of the Training Data Set

The patient data for the control arm contained within each trial data set were extracted and used as the actual data for the training data set. The selection of variables in the training data set aimed to include as many variables related to the subjects' background as possible, excluding variables concerning tests and evaluations conducted during the trials. Furthermore, variables that had the same value were excluded, even if they were related to the subjects' background (Multimedia Appendices 1-4).

Generation of Synthetic Data

The SPDs in this study were generated using the following 4 methods:

1. CART: the synthpop package (version 1.8) in R (The R Foundation) was used, specifying the cart method for the syn function's method argument.
2. RF: the synthpop package (version 1.8) in R was used, specifying the Ranger method for the syn function's method argument.
3. BN: the bnlearn package (version 4.9) in R was used to conduct structural learning through the score-based algorithm hill-climbing, followed by parameter estimation using the bn.fit function. The default maximum likelihood estimator was used for parameter estimation.
4. CTGAN: the CTGANSynthesizer module included in the Python package sdv (version 1.3) was used.

In all these generation methods, to ensure the absence of conflicting data regarding the relationship between PFS and OS, constraints were set to ensure that the values of PFS and OS were greater than zero and that PFS was less than or equal to OS. Specific individual patient data in the generated SPD, which did not meet these constraints, were excluded, and new individual patient data were regenerated. The SPDs were generated in a manner that equaled the number of subject-level data to the record count in the actual data.

To ensure the reproducibility of SPD generation, 1000 random numbers were generated as seed values using the Mersenne Twister algorithm. The same seed value set was used for all generation methods.

Statistical Analysis

Histogram

Histograms were created to visually inspect the distributions of the MST of the synthetic data (MSTS) for PFS and OS for the 1000 SPD data sets generated by each method. The histograms also included the MST of the actual data (MSTA) as a vertical line and the range of its 95% CI as a rectangular background. For PFS and OS, a higher percentage of MSTS covered by the 95% CI of the MSTA was determined to indicate a greater level of reliability for the generation method.

Evaluation of Similarity

A hazard ratio (HR) of 1 signifies that the 2 survival functions are entirely identical. Thus, the closer the HR is to 1, the more similar the 2 survival functions are. Accordingly, based on the following calculation formula, the HR distance (HRD) for PFS and OS from the SPD and the actual data were computed and evaluated:

$$\mathrm{HRD} = 1 - \mathrm{abs}(\mathrm{HR} - 1)$$

Kaplan-Meier Plot

In the evaluation of similarity, the SPD that showed the highest HRD value was considered the best case, and the SPD with the lowest HRD value was considered the worst case. Three groups of Kaplan-Meier (KM) plots were created, including the actual data, the best case, and the worst case for each SPD generation method. The best case and worst case for each SPD generation method in both PFS and OS were compared to actual survival by using the log rank test. Multiple comparisons were not performed, nor were *P* values adjusted because controlling for the type I error rate does not affect the conclusions of this study.

Since the purpose of this study was to evaluate the method of generating SPD that closely resemble actual survival data, it might be unnecessary to calculate a *P* value that indicates

a significant difference from actual survival, but the *P* value was calculated in this study from the viewpoint that if a significant difference is also observed in the best-case, that method should not be adopted.

All analyses and data generation were performed using R (version 4.3.1; The R Foundation) and Python (version 3.10; Python Software Foundation).

Ethical Considerations

Ethical review was not needed for this simulation study for methodology comparison. All actual clinical trial data sets obtained from Project Data Sphere were used in accordance with relevant guidelines and regulations when the clinical trials were conducted.

Results

Figure 1 shows a histogram of the MSTS for PFS in the NCT00703326 trial. Using CART, RF, and BN, most of the generated MSTS values were within the 95% CI of the MSTA. In contrast, when CTGAN was used, SPD generation resulted in a widened variance in the distribution of MSTS. For the MSTS of PFS in the other 3 trials, RF exhibited a shift in the distribution of the MSTS, shortening the survival period, while BN displayed a shift in the distribution and prolonged the survival period. Similar trends to Figure 1 were observed for CART and CTGAN (Multimedia Appendices 5-7).

Figure 2 displays a histogram of the MSTS for OS in the NCT00460265 trial. The divergence from the PFS findings is that the MSTS of RF was more frequently included within the 95% CI of the MSTA, with similar results observed in other trials (Multimedia Appendices 8-10). In other aspects, similar findings were obtained as with the PFS.

Table 2 presents the number and proportion of the generated MSTS values included within the 95% CI of the MSTA for each trial and each method. In the case of CART for PFS, a high percentage ranging from 88.8% to 98.1% was exhibited for all trials. However, the OS ranged from 60.8% to 96.1%, with some trials displaying a lower percentage than the PFS.

Figure 1. Histogram of the median survival time of the synthetic data for progression-free survival in the NCT00703326 trial. The dashed vertical line represents the median survival time of the actual data, and the light blue background indicates its 95% CI. BN: Bayesian network; CART: classification and regression tree; CTGAN: conditional tabular generative adversarial network; MST: median survival time; RF: random forest.

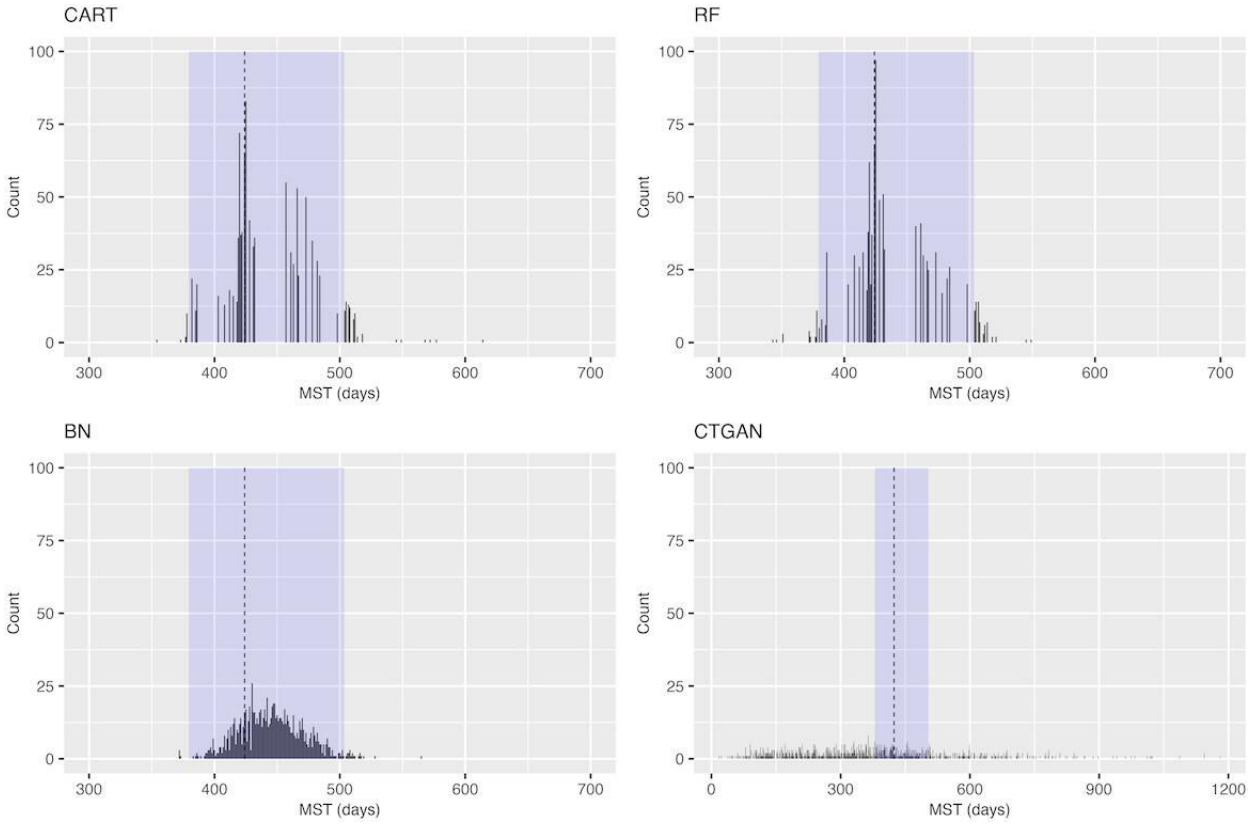


Figure 2. Histogram of the median survival time of the synthetic data of overall survival in the NCT00460265 trial. The dashed vertical line represents the median survival time of the actual data, and the light blue background indicates its 95% CI. BN: Bayesian network; CART: classification and regression tree; CTGAN: conditional tabular generative adversarial network; MST: median survival time; RF: random forest.

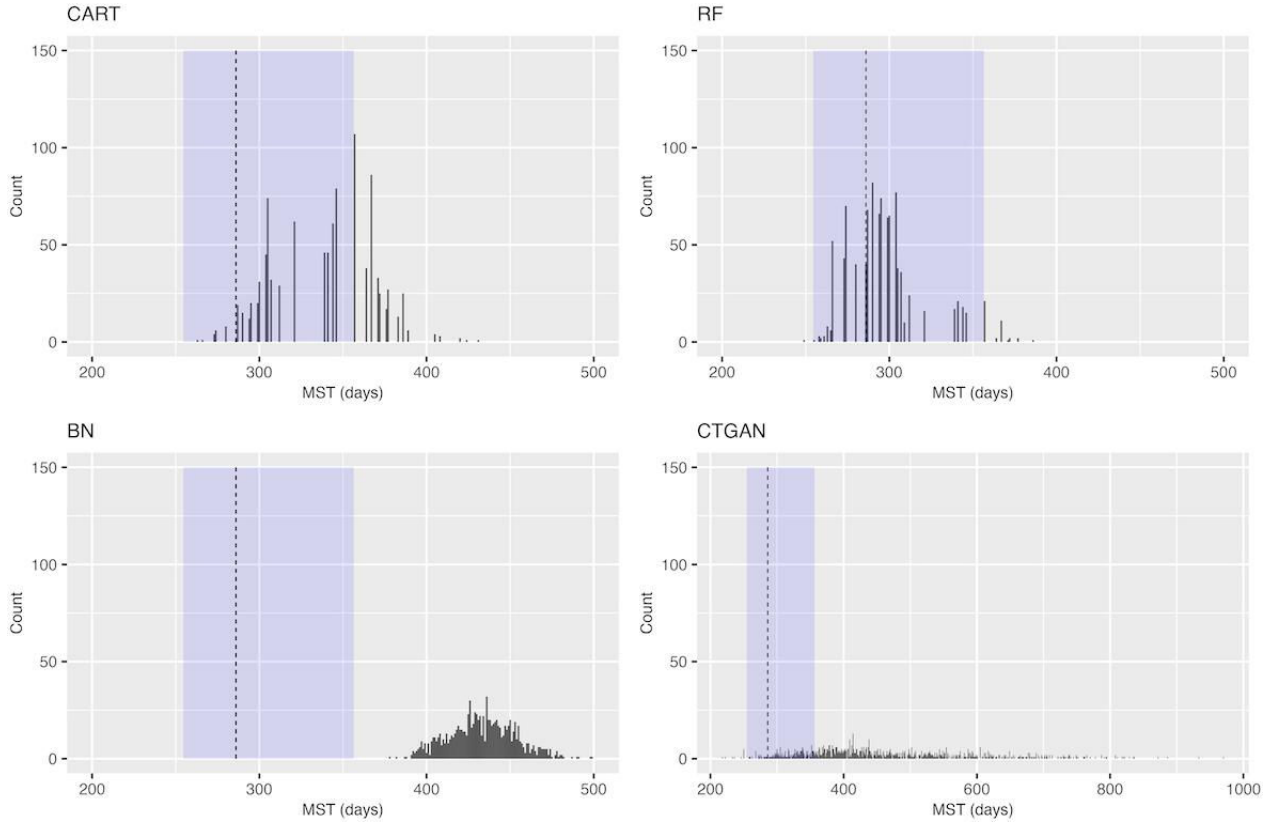


Table 2. The number and proportion of median survival times of the synthetic data (MSTSs) falling within the 95% CI of the median survival time of the actual data (MSTA).

	ClinicalTrials.gov ID			
	NCT00119613	NCT00339183	NCT00460265	NCT00703326
Progression-free survival				
MSTA (95% CI)	169 (163-183)	155 (121-168)	133 (121-167)	424 (380-504)
MSTSs, n (%)				
CART[a] (n=1000)	981 (98.1)	888 (88.8)	955 (95.5)	918 (91.8)
RF[b] (n=1000)	693 (69.3)	248 (24.8)	426 (42.6)	919 (91.9)
BN[c] (n=1000)	10 (1.0)	0 (0.0)	37 (3.7)	976 (97.6)
CTGAN[d] (n=1000)	65 (6.5)	378 (37.8)	322 (32.2)	254 (25.5)
Overall survival				
MSTA (95% CI)	276 (259-303)	361 (319-393)	286 (255-357)	1452 (1417-1507)
MSTSs, n (%)				
CART (n=1000)	831 (83.1)	608 (60.8)	719 (71.9)	961 (96.1)
RF (n=1000)	757 (75.7)	697 (69.7)	980 (98.0)	599 (59.9)
BN (n=1000)	0 (0.0)	0 (0.0)	0 (0.0)	622 (62.2)
CTGAN (n=1000)	72 (7.2)	155 (15.5)	197 (19.7)	81 (8.5)

[a]CART: classification and regression tree.
[b]RF: random forest.
[c]BN: Bayesian network.
[d]CTGAN: conditional tabular generative adversarial network.

For RF, a high proportion of 91.9% was observed for PFS in the NCT00703326 trial and 98.0% for OS in the NCT00460265 trial, whereas in other cases, the proportion for RF was not as high as that for CART.

In the case of BN, proportions of 97.6% and 62.2% were observed for PFS and OS, respectively, in the NCT00703326 trial, but in the other 3 trials, BN showed an extremely low

percentage ranging from proportion ranging from 0.0% to 3.7%.

CTGAN showed a low proportion ranging from 6.5% to 37.8% for both PFS and OS in all trials.

Figure 3 shows the KM plot for PFS in the NCT00703326 trial. The best-case curves of CART and RF were similar to the actual data curve. In contrast, for BN and CTGAN, even the best-case curves deviated from the actual data curve. In other trials, some SPD did not show a similar trend. However, at least for the best-case scenarios of CART and RF, the generated synthetic survival curves closely resembled the actual survival curve (Multimedia Appendices 11-13).

Figure 4 displays the KM plot for OS in the NCT00460265 trial. Similar to the KM plots for PFS, the best-case curves of CART and RF resembled the actual data curve, whereas those of BN and CTGAN deviated from the actual data curve. These trends were also observed in other trials (Multimedia Appendices 14-16).

Figures 5 and 6 present box plots of the HRD. When using CART, the HRD values for both PFS and OS in all trials were concentrated at approximately 0.9. Conversely, for the other methods, no consistent trend was observed for either PFS or OS.

Figure 3. Kaplan-Meier plots for progression-free survival in the NCT00703326 trial. BN: Bayesian network; CART: classification and regression tree; CTGAN: conditional tabular generative adversarial network; RF: random forest.

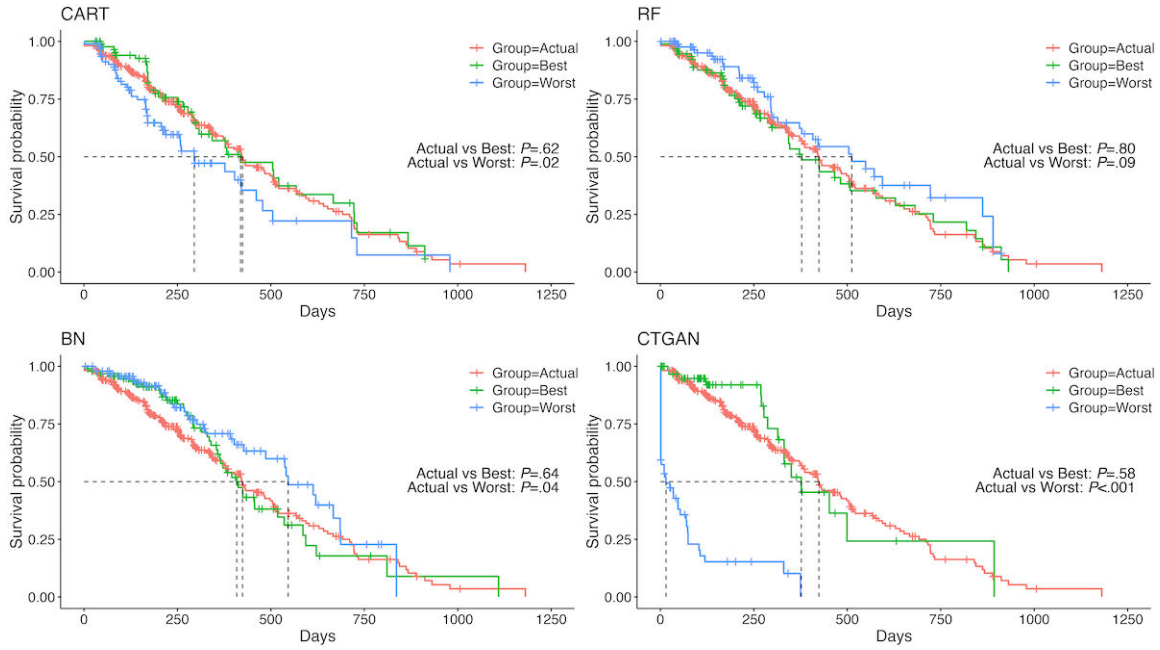


Figure 4. Kaplan-Meier plots for overall survival in the NCT00460265 trial. BN: Bayesian network; CART: classification and regression tree; CTGAN: conditional tabular generative adversarial network; RF: random forest.

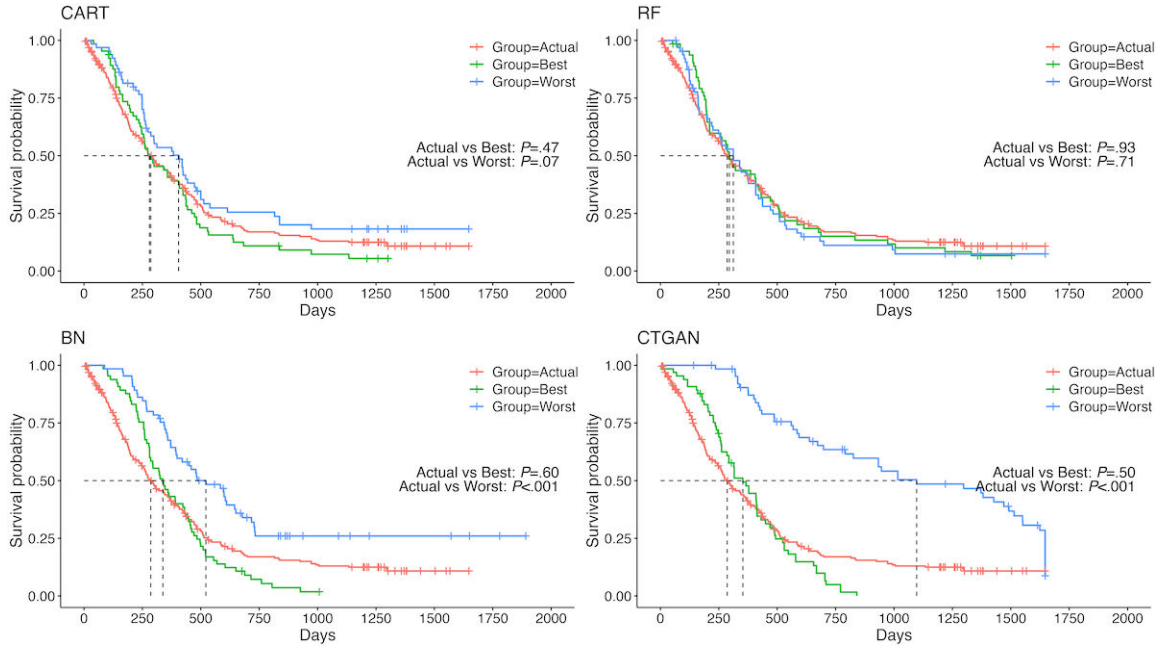


Figure 5. Box plot of progression-free survival hazard ratio distance (HRD) for each method and clinical trial. BN: Bayesian network; CART: classification and regression tree; CTGAN: conditional tabular generative adversarial network; RF: random forest.

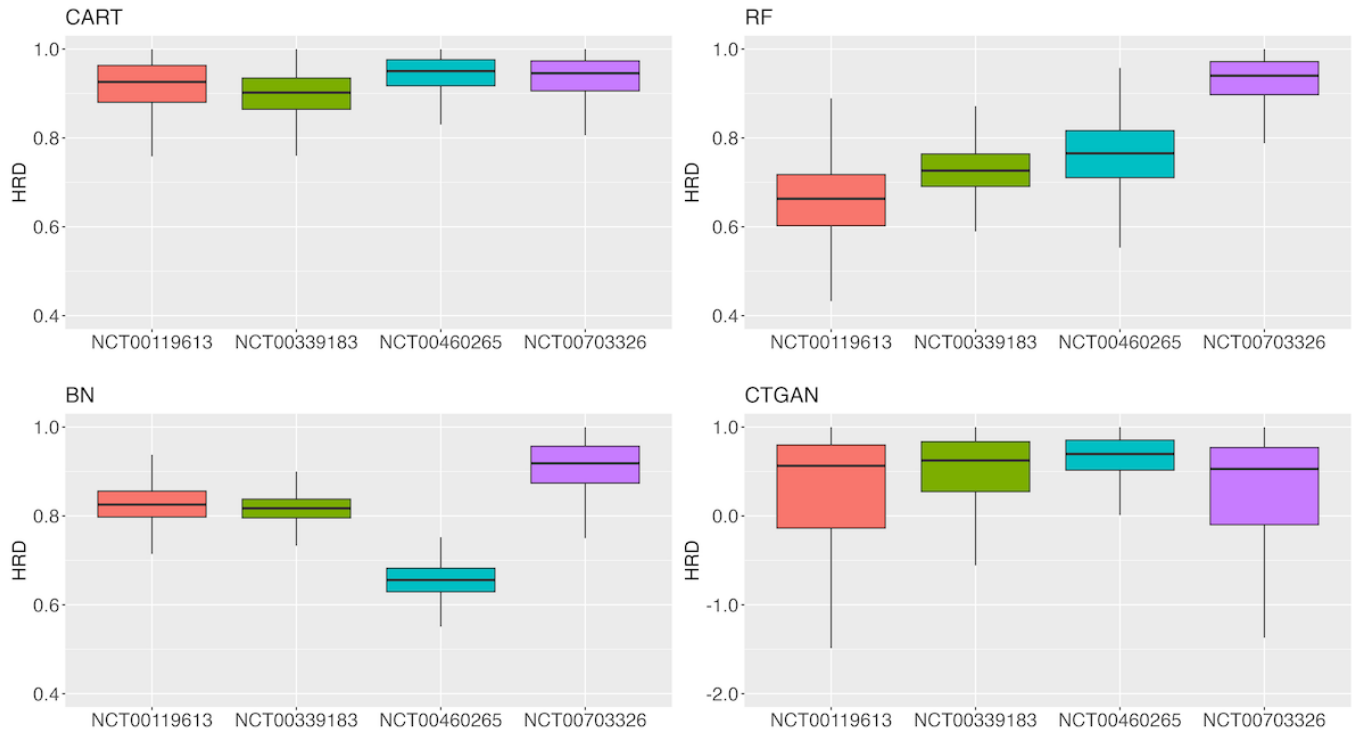
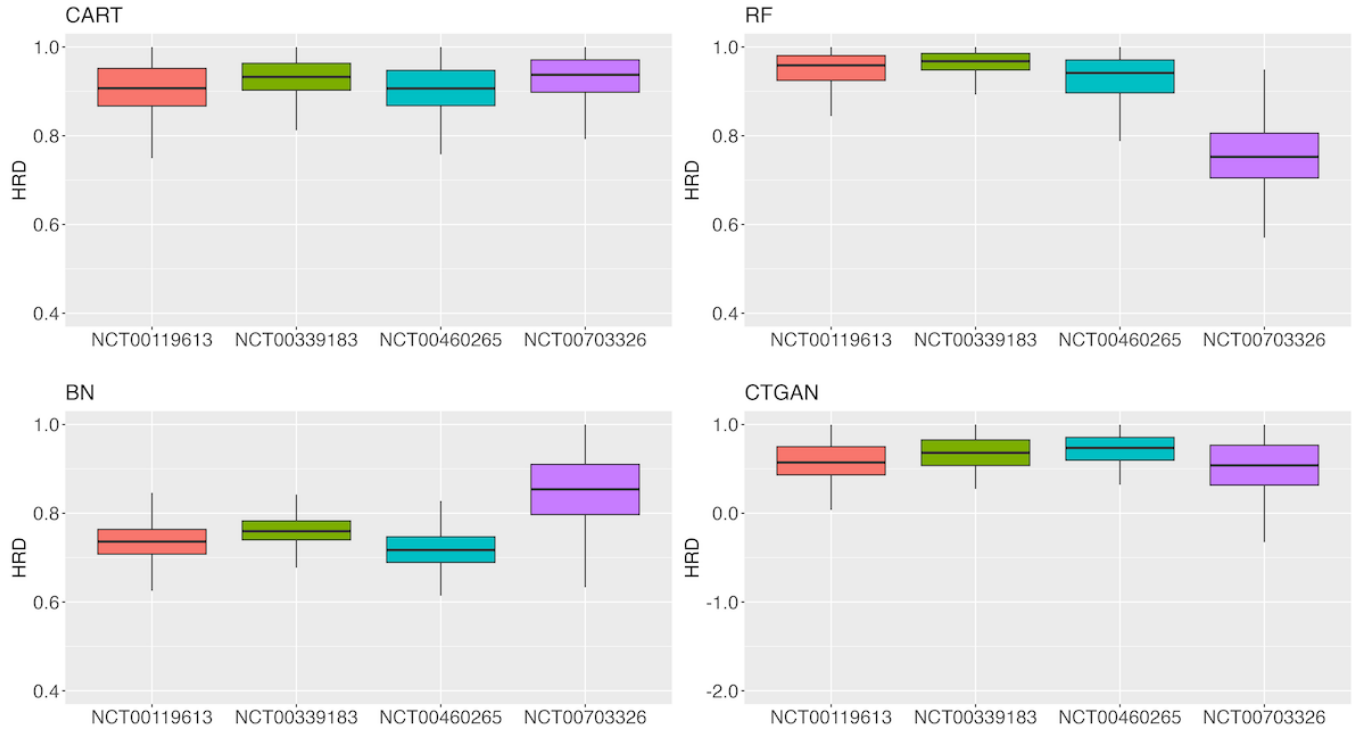


Figure 6. Box plot of overall survival hazard ratio distance (HRD) for each method and clinical trial. BN: Bayesian network; CART: classification and regression tree; CTGAN: conditional tabular generative adversarial network; RF: random forest.



Discussion

Regarding the survival SPD, CART often yielded better results than the other methods in evaluations using MST, HRD, and visual analysis via KM plots. Given the crucial importance of the hazard ratio and MST as end points in oncology trials [26], demonstrating the utility of both of these evaluation metrics is essential. Therefore, using CART for generating survival SPD was suggested as a beneficial approach.

While both CART and RF generally yielded preferable results across all trials, they share the common characteristic of using tree models. RF, with its use of the bootstrap method for resampling and constructing tree models for ensemble learning, is known to prevent overfitting. In general, in terms of constructing machine learning models with high generalization performance, RF performs better than CART. However, CART is prone to overfitting as the layers of the tree become deeper [32]. Although RF is considered a superior method for constructing high–generalization-performance machine learning models, the results from Table 2 and the KM plots in this study suggest that CART is a better approach than RF. This discrepancy might be due to differing views on what is a higher performance between the machine learning prediction model and SPD. In the machine learning prediction model, it is important to prevent overfitting and reduce bias; however, SPD is expected to match its statistical properties with actual data. Thus, in the case of SPD, the overfitting suppression mechanism possessed by RF might have resulted in inferiority to that of CART from the perspective of improving similarity.

In the case of using BN, the percentage of MSTSs falling within the 95% CI of MSTAs was 0% for the PFS of the NCT00339183 trial, and for OS, this phenomenon also occurred in the NCT00119613, NCT00339183, and NCT0046265 trials. This implies that the SPD failed to accurately reflect the statistical properties of the actual data. Conversely, a high value of 97.6% was observed for the PFS in the NCT00703326 trial. The reason for this discrepancy could not be determined on the basis of the results of this study. Tucker et al [24] reported that they could generate data highly similar to actual data when using BN for the generation of SPD, which differs from our findings. One notable difference is that while Tucker et al [24] used a large-scale actual data set of 27.5 million patients for their study, this study used only a few hundred patients for training data. This difference likely had a significant impact on the accuracy of the SPD generation model, resulting in conflicting results. However, the SPD generated by BN were not distributed in the direction of shortening PFS or OS. Thus, this would not be harmful when the SPD generated by BN is used as a more conservative control arm in clinical trials.

Using CTGAN, the percentage of the MSTSs falling within the 95% CI of the actual data was low, indicating low performance associated with the generation of SPD that reflect the statistical properties of the actual data. However, Krenmayr et al [23] reported favorable performance results when using the same generative adversarial network (GAN)–based methods and RWD. The differences between their study and our study were as follows: their study did not include SPD on survival time or generate multiple SPD data sets from the same actual data, and there was a large amount of individual patient data in their study. In particular, focusing

on the amount of individual patient data, the number of patients in each trial included in this study was relatively small, with the NCT00119613 trial having 232 patients, the NCT00339183 trial having 476 patients, the NCT0046265 trial having 260 patients, and the NCT00703326 trial having 382 patients, while the trial conducted by Krenmayr et al [23] had 500 or more patients. GAN-based methods using deep neural networks are known to perform poorly with small amounts of data [25,33]. In this study, although the NCT00339183 trial had the largest number of individual patient data, the best case of CTGAN for NCT00339183 produced a KM plot similar to the actual data, suggesting that a larger data set yields better results. Thus, there is no contradiction. Another characteristic of using CTGAN in this study was the larger variance in the estimated MSTSs, as indicated in Figures 1 and 2. Goncalves et al [34] showed that using MC-MedGAN, a GAN-based method, to generate an SPD from small data resulted in a large SD of the data utility metrics, leading to results with larger variance, similar to those of this study. Therefore, it is extremely challenging to generate useful SPD by applying GAN-based methods to small data sets, such as clinical trial data.

When generating SPDs for survival data and using them as a certain arm in a clinical trial, it is important to verify that the statistical properties closely match those of the actual data with the MST and the hazard ratio with the actual data being close to 1. Based on our results, we conclude that CART, which can concentrate the MSTSs within the range of 95% CI of MSTAs and approximately 0.9 for HRD, is an efficient method for generating SPD that meets the above-mentioned conditions. However, even when using CART, slight variations were observed in the MSTSs, and some cases fell outside the 95% CI of the MSTAs, as revealed by our results. Therefore, for practical use, it is necessary to verify that the MSTSs are included in the 95% CI of the MSTAs and that both are close in value. It is also necessary to verify whether the HRD of the actual data and the SPD are close to 1 and then decide whether to adopt the generated SPD. Hence, the generation process must be repeated until an acceptable SPD is obtained. There may also be a need to use statistical methods to match characteristics between the SPD and the actual treatment arm in clinical trials.

In this study, even the most useful CART method produced SPDs that did not meet the requirements of MST and HRD. We expect that this issue will be addressed by incorporating feature engineering, such as dimension reduction, imputing missing values, derived variable creation, and other processing. Additionally, in clinical research, as subgroup analyses are frequently conducted, it is necessary to improve the generation method to reflect the statistical properties of the actual data even when the data are divided into subgroups under certain conditions. Moreover, from the perspective of data privacy, it is essential to incorporate approaches to prevent data reidentification into the generation method [35].

In conclusion, as a method for generating SPD for survival data from small data sets, such as clinical trial data, CART is the most effective method for generating SPD that meet the 2 conditions of having an MSTSs close to the MSTAs and an HRD close to 1. However, as SPD might be generated, which do not meet these 2 conditions, it is necessary to incorporate mechanisms to improve a CART-based generation method in future studies. Overcoming these challenges would make it possible to reduce the recruitment period and costs of clinical trial participants to ≥50% in comparative trials of new drug development against existing therapeutic drugs. This approach could accelerate clinical development, similar to the use of RWD.

Acknowledgments

We would like to express our gratitude to Project Data Sphere, the platform that provided the necessary data for this study, and to the clinical trial data providers Amgen and Eli Lilly.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Variables used to generate synthetic patient data from the NCT00119613 trial.
[DOCX File (Microsoft Word File), 48 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Variables used to generate synthetic patient data from the NCT00339183 trial.
[DOCX File (Microsoft Word File), 48 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Variables used to generate synthetic patient data from the NCT00460265 trial.
[DOCX File (Microsoft Word File), 48 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Variables used for generating synthetic patient data from the NCT00703326 trial.
[DOCX File (Microsoft Word File), 48 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Histogram of the median survival time of the synthetic data for progression-free survival in the NCT00119613 trial. The dashed vertical line represents the median survival time for the actual data, and the light blue background indicates its 95% CI.
[DOCX File (Microsoft Word File), 203 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Histogram of the median survival times for the synthetic data for progression-free survival in the NCT00339183 trial. The dashed vertical line represents the median survival time of the actual data, and the light blue background indicates its 95% CI.
[DOCX File (Microsoft Word File), 193 KB-Multimedia Appendix 6]

Multimedia Appendix 7

Histogram of the median survival times of the synthetic data for progression-free survival in the NCT00460265 trial. The dashed vertical line represents the median survival time of the actual data, and the light blue background indicates its 95% CI.
[DOCX File (Microsoft Word File), 188 KB-Multimedia Appendix 7]

Multimedia Appendix 8

Histogram of the median survival times of the synthetic data for overall survival in the NCT00119613 trial. The dashed vertical line represents the median survival time of the actual data, and the light blue background indicates its 95% CI.
[DOCX File (Microsoft Word File), 187 KB-Multimedia Appendix 8]

Multimedia Appendix 9

Histogram of the median survival times of the synthetic data for overall survival in the NCT00339183 trial. The dashed vertical line represents the median survival time of the actual data, and the light blue background indicates its 95% CI.
[DOCX File (Microsoft Word File), 196 KB-Multimedia Appendix 9]

Multimedia Appendix 10

Histogram of the median survival times of the synthetic data for overall survival in the NCT00703326 trial. The dashed vertical line represents the median survival time of the actual data, and the light blue background indicates its 95% CI.
[DOCX File (Microsoft Word File), 188 KB-Multimedia Appendix 10]

Multimedia Appendix 11

Kaplan-Meier plots for progression-free survival in the NCT00119613 trial.
[DOCX File (Microsoft Word File), 216 KB-Multimedia Appendix 11]

Multimedia Appendix 12

Kaplan-Meier plots for progression-free survival in the NCT00339183 trial.
[DOCX File (Microsoft Word File), 230 KB-Multimedia Appendix 12]

Multimedia Appendix 13

Kaplan-Meier plots for progression-free survival in the NCT00460265 trial.
[DOCX File (Microsoft Word File), 218 KB-Multimedia Appendix 13]

Multimedia Appendix 14

Kaplan-Meier plots for overall survival in the NCT00119613 trial.
[DOCX File (Microsoft Word File), 230 KB-Multimedia Appendix 14]

Multimedia Appendix 15

Kaplan-Meier plots for overall survival in the NCT00339183 trial.
[DOCX File (Microsoft Word File), 253 KB-Multimedia Appendix 15]

Multimedia Appendix 16

Kaplan-Meier plots for overall survival in the NCT00703326 trial.
[DOCX File (Microsoft Word File), 265 KB-Multimedia Appendix 16]

References

1. Huang GD, Bull J, Johnston McKee K, et al. Clinical trials recruitment planning: a proposed framework from the clinical trials transformation initiative. Contemp Clin Trials. Mar 2018;66:74-79. [doi: 10.1016/j.cct.2018.01.003] [Medline: 29330082]

2. Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. Contemp Clin Trials Commun. Sep 2018;11:156-164. [doi: 10.1016/j.conctc.2018.08.001] [Medline: 30112460]
3. Treweek S, Lockhart P, Pitkethly M, et al. Methods to improve recruitment to randomised controlled trials: Cochrane systematic review and meta-analysis. BMJ Open. 2013;3(2):e002360. [doi: 10.1136/bmjopen-2012-002360] [Medline: 23396504]
4. Considerations for the design and conduct of externally controlled trials for drug and biological products. Guidance for industry. US Food and Drug Administration. 2023. URL: https://www.fda.gov/media/164960/download [Accessed 2024-06-04]
5. Yap TA, Jacobs I, Baumfeld Andre E, Lee LJ, Beaupre D, Azoulay L. Application of real-world data to external control groups in oncology clinical trial drug development. Front Oncol. 2021;11:695936. [doi: 10.3389/fonc.2021.695936] [Medline: 35070951]
6. Dagenais S, Russo L, Madsen A, Webster J, Becnel L. Use of real-world evidence to drive drug development strategy and inform clinical trial design. Clin Pharmacol Ther. Jan 2022;111(1):77-89. [doi: 10.1002/cpt.2480] [Medline: 34839524]
7. Liu R, Rizzo S, Whipple S, et al. Evaluating eligibility criteria of oncology trials using real-world data and AI. Nature. Apr 2021;592(7855):629-633. [doi: 10.1038/s41586-021-03430-5] [Medline: 33828294]
8. Azizi Z, Lindner S, Shiba Y, et al. A comparison of synthetic data generation and federated analysis for enabling international evaluations of cardiovascular health. Sci Rep. Jul 17, 2023;13(1):11540. [doi: 10.1038/s41598-023-38457-3] [Medline: 37460705]
9. El Emam K, Jonker E, Arbuckle L, Malin B. A systematic review of re-identification attacks on health data. PLoS One. 2011;6(12):e28071. [doi: 10.1371/journal.pone.0028071] [Medline: 22164229]
10. Kaur D, Sobiesk M, Patil S, et al. Application of Bayesian networks to generate synthetic health data. J Am Med Inform Assoc. Mar 18, 2021;28(4):801-811. [doi: 10.1093/jamia/ocaa303] [Medline: 33367620]
11. Mavrogenis AF, Scarlat MM. Artificial intelligence publications: synthetic data, patients, and papers. Int Orthop. Jun 2023;47(6):1395-1396. [doi: 10.1007/s00264-023-05830-w] [Medline: 37162553]
12. Meeker D, Kallem C, Heras Y, Garcia S, Thompson C. Case report: evaluation of an open-source synthetic data platform for simulation studies. JAMIA Open. Oct 2022;5(3):ac067. [doi: 10.1093/jamiaopen/ooac067] [Medline: 35958672]
13. Brownstein JS, Chu S, Marathe A, et al. Combining participatory influenza surveillance with modeling and forecasting: three alternative approaches. JMIR Public Health Surveill. Nov 1, 2017;3(4):e83. [doi: 10.2196/publichealth.7344] [Medline: 29092812]
14. Guillaudeux M, Rousseau O, Petot J, 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. [doi: 10.1038/s41746-023-00771-5] [Medline: 36899082]
15. El Emam K. Status of synthetic data generation for structured health data. JCO Clin Cancer Inform. Jun 2023;7:e2300071. [doi: 10.1200/CCI.23.00071] [Medline: 37390378]
16. D'Amico S, Dall'Olio D, Sala C, et al. Synthetic data generation by artificial intelligence to accelerate research and precision medicine in hematology. JCO Clin Cancer Inform. Jun 2023;7:e2300021. [doi: 10.1200/CCI.23.00021] [Medline: 37390377]
17. Gonzales A, Guruswamy G, Smith SR. Synthetic data in health care: a narrative review. PLOS Digit Health. Jan 2023;2(1):e0000082. [doi: 10.1371/journal.pdig.0000082] [Medline: 36812604]
18. Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. NPJ Digit Med. Oct 9, 2023;6(1):186. [doi: 10.1038/s41746-023-00927-3] [Medline: 37813960]
19. Ursin G, Sen S, Mottu JM, Nygård M. Protecting privacy in large datasets—first we assess the risk; then we fuzzy the data. Cancer Epidemiol Biomarkers Prev. Aug 1, 2017;26(8):1219-1224. [doi: 10.1158/1055-9965.EPI-17-0172] [Medline: 28754793]
20. 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. [doi: 10.2196/18910] [Medline: 32501278]
21. Summers C, Griffiths F, Cave J, Panesar A. Understanding the security and privacy concerns about the use of identifiable health data in the context of the COVID-19 pandemic: survey study of public attitudes toward COVID-19 and data-sharing. JMIR Form Res. Jul 7, 2022;6(7):e29337. [doi: 10.2196/29337] [Medline: 35609306]
22. 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. [doi: 10.1136/bmjopen-2020-043497] [Medline: 33863713]

23. Krenmayr L, Frank R, Drobig C, et al. GANerAid: realistic synthetic patient data for clinical trials. Inform Med Unlocked. 2022;35:101118. [doi: 10.1016/j.imu.2022.101118]
24. Tucker A, Wang Z, Rotalinti Y, Myles P. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. NPJ Digit Med. Nov 9, 2020;3(1):147. [doi: 10.1038/s41746-020-00353-9] [Medline: 33299100]
25. Santos M. How to generate real-world synthetic data with CTGAN. Medium. 2023. URL: https://medium.com/towards-data-science/how-to-generate-real-world-synthetic-data-with-ctgan-af41b4d60fde [Accessed 2024-06-04]
26. Ben-Aharon O, Magnezi R, Leshno M, Goldstein DA. Median survival or mean survival: which measure is the most appropriate for patients, physicians, and policymakers?. Oncologist. Nov 2019;24(11):1469-1478. [doi: 10.1634/theoncologist.2019-0175] [Medline: 31320502]
27. Smith A, Lambert PC, Rutherford MJ. Generating high-fidelity synthetic time-to-event datasets to improve data transparency and accessibility. BMC Med Res Methodol. Jun 23, 2022;22(1):176. [doi: 10.1186/s12874-022-01654-1] [Medline: 35739465]
28. Breiman L, editor. Classification and Regression Trees. Chapman and Hall; 1998.
29. Breiman L. Random forests. Mach Learn. 2001;45(1):5-32. [doi: 10.1023/A:1010933404324]
30. Pearl J. Bayesian networks: a model of self-activated memory for evidential reasoning. Presented at: Proceedings of the 7th Conference of the Cognitive Science Society; Aug 15 to 17, 1985; Irvine, CA. URL: https://ftp.cs.ucla.edu/tech-report/198_-reports/850017.pdf [Accessed 2024-06-04]
31. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling tabular data using conditional GAN. arXiv. Preprint posted online on Jul 1, 2019. [doi: 10.48550/arXiv.1907.00503]
32. Hayes T, Usami S, Jacobucci R, McArdle JJ. Using classification and regression trees (CART) and random forests to analyze attrition: results from two simulations. Psychol Aging. Dec 2015;30(4):911-929. [doi: 10.1037/pag0000046] [Medline: 26389526]
33. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved techniques for training GANs. arXiv. Preprint posted online on Jun 10, 2016. URL: http://arxiv.org/abs/1606.03498 [Accessed 2024-06-04] [doi: 10.48550/arXiv.1606.03498]
34. Goncalves A, Ray P, Soper B, Stevens J, Coyle L, Sales AP. Generation and evaluation of synthetic patient data. BMC Med Res Methodol. May 7, 2020;20(1):108. [doi: 10.1186/s12874-020-00977-1] [Medline: 32381039]
35. El Emam K, Mosquera L, Hoptroff R. Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data. O'Reilly Media; 2020.

Abbreviations

BN: Bayesian network
CART: classification and regression trees
CTGAN: conditional tabular generative adversarial network
GAN: generative adversarial network
HR: hazard ratio
HRD: hazard ratio distance
KM: Kaplan-Meier
MST: median survival time
MSTA: median survival time of actual data
MSTS: median survival time of synthetic data
OS: overall survival
PFD: progression-free survival
RF: random forest
RWD: real-world data
SPD: synthetic patient data

Edited by Christian Lovis; peer-reviewed by Danqing Hu, Jiangdian Song; submitted 03.12.2023; final revised version received 06.04.2024; accepted 08.05.2024; published 18.06.2024

Please cite as:
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 2024;12:e55118
URL: https://medinform.jmir.org/2024/1/e55118

doi: 10.2196/55118