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

Ippei Akiya, Takuma Ishihara, Keiichi Yamamoto

Submitted to: JMIR Medical Informatics
on: December 03, 2023

Table of Contents

A 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 JP
[2]Innovative and Clinical Research Promotion Center Gifu University Hospital Gifu JP
[3]Division of Data Science, Center for Industrial Research and Innovation Translational Research Institute for Medical Innovation Osaka Dental University Osaka JP

Corresponding Author:
Ippei Akiya MSc
Biometrics
ICON Clinical Research GK
Sumitomo-ShinTranomon Building
4-3-9 Toranomon Minato-ku
Tokyo
JP

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 CTGAN, have been employed 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 endpoints in oncology trials. To achieve this goal, we conducted a comparative simulation of 4 generation methods: CART, RF, BN, and the CTGAN, and the performance of each method was evaluated.

Methods: Using multiple clinical trial datasets, 1000 datasets were generated by using each method for each clinical trial dataset and evaluated as follows: 1) mean 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 MSTSs falling within the range of 95% confidence interval (CI) of the MSTA. 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 demonstrated that HRD values were concentrated at approximately 0.9. Conversely, for the other methods, no consistent trend was observed for either PFS or OS.
The reason why CART demonstrated better similarity than RF was that CART caused overfitting and RF, which is a kind of ensemble learning, 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 datasets are not suitable.

Conclusions: As a method for generating SPD for survival data from small datasets, 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 Preprints 03/12/2023:55118)
DOI: https://doi.org/10.2196/preprints.55118

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Ippei Akiya MSc [a], Takuma Ishihara PhD [b], Keiichi Yamamoto PhD [c]

[a] Biometrics, ICON Clinical Research GK
[b] Innovative and Clinical Research Promotion Center, Gifu University Hospital, Gifu, Japan.
[c] Division of Data Science, Center for Industrial Research and Innovation, Translational Research Institute for Medical Innovation, Osaka Dental University

Corresponding author:
Ippei Akiya MSc [a]
Sumitomo-ShinTranomon Building, 4-3-9, Toranomon, Minato-ku, Tokyo, 105-0001, Japan.
Email: ippei.akiya@gmail.com, Tel: +81-3-3830-1763

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

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 CTGAN, have been employed 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 endpoints in oncology trials. To achieve this goal, we conducted a comparative simulation of 4 generation methods: CART, RF, BN, and the CTGAN, and the performance of each method was evaluated.

Methods:
Using multiple clinical trial datasets, 1000 datasets were generated by using each method for each clinical trial dataset and evaluated as follows: 1) mean 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 MSTSs falling within the range of 95% confidence interval (CI) of the MSTA. 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 demonstrated that HRD values were concentrated at approximately

0.9. Conversely, for the other methods, no consistent trend was observed for either PFS or OS.
The reason why CART demonstrated better similarity than RF was that CART caused overfitting and RF, which is a kind of ensemble learning, 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 datasets are not suitable.

Conclusions:
As a method for generating SPD for survival data from small datasets, 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.

Keywords: oncology clinical trials, survival analysis, synthetic patient data

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 participants [10, 14, 25]. The challenge of subject 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 healthcare financing.
In recent years, the use of real-world data (RWD) has emerged as a potential solution for addressing these issues. The FDA has also released draft guidance [12], garnering attention on the application of RWD as an external control arm in clinical trials [4, 38]. 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 [13].
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 healthcare research field [27, 33, 34], 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 healthcare research while addressing data privacy concerns [19, 27, 29].
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 [7, 15, 16, 20] 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 endpoints such as overall survival (OS) and progression-free survival (PFS)-related survival

functions and median survival time (MST) [23]. In previous studies on SPD, there has been a notable emphasis on reporting patient background data and single time-point data [7, 15, 16, 20]. However, research focusing specifically on the relationship between SPD and survival data remains relatively insufficient.

As the first step in examining our hypothesis that the utilization of SPD can be beneficial in accelerating healthcare 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 endpoints in oncology trials. To achieve this goal, we conducted a comparative simulation of 4 generation methods: classification and regression trees (CART) [9], random forest (RF) [30], Bayesian network (BN) [6], and the CTGAN approach [26], and the performance of each method was evaluated.

Methods

To generate the SPD, subject-level clinical trial data were obtained from Project Data Sphere for the following 4 clinical trials and are summarized in Table 1. The following 4 clinical trials were selected: 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

1.

Clinical Trial ID	Titles	Phase	Cancer Type	Intervention for Control Arm	Number of Subjects of Control Arm
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 Alone	476

Clinical Trial ID	Titles	Phase	Cancer Type	Intervention for Control Arm	Number of Subjects of Control Arm
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 and/ or Metastatic Head and Neck Cancer	Cisplatin and 5-FU	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

Preparation of the Training Dataset

The patient data for the control arm contained within each trial dataset were extracted and used as the actual data for the training dataset. The selection of variables in the training dataset aimed to include as many variables related to the subject background as possible, excluding variables concerning tests and evaluations conducted during the trials. Furthermore, variables that had completely the same value were excluded, even if they were related to the subject 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 was utilized, 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 utilized.

In all these generation methods, to ensure the absence of conflicting data concerning 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 that did not meet these constraints were dropped, 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 datasets 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% confidence interval (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

An 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 hazard ratio distance (HRD) for PFS and OS from the SPD and the actual data were computed and evaluated:

HRD = 1 - abs(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 the worst-case for each SPD generation method in both PFS and OS were compared to actual survival by using 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 and Python version 3.10.

Results

Figure 1 shows the 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 three 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 1. Histogram of the MSTSs for PFS in the NCT00703326 trial. The dashed vertical line represents the MSTA, and the light blue background indicates its 95% CI.

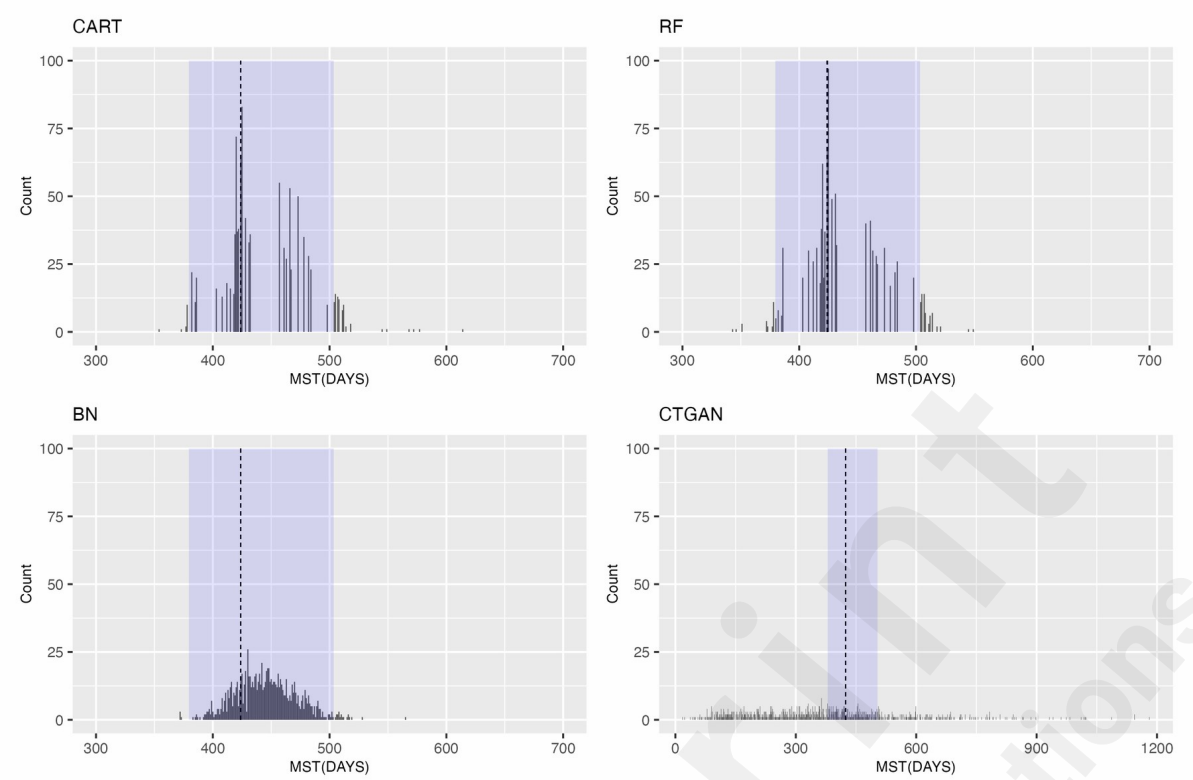


Figure 2 displays the histogram of the MSTS for OS in the NCT00460265 trial. The divergence from the PFS results is that the MSTS of RF was more frequently included within the 95% CI of the MSTA, with similar results observed in other trials (Multimedia Appendices 8-10). In other aspects of the results,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

Figure 2. Histogram of the MSTSs of OS in the NCT00460265 trial. The dashed vertical line represents the MSTA, and the light blue background indicates its 95% CI.

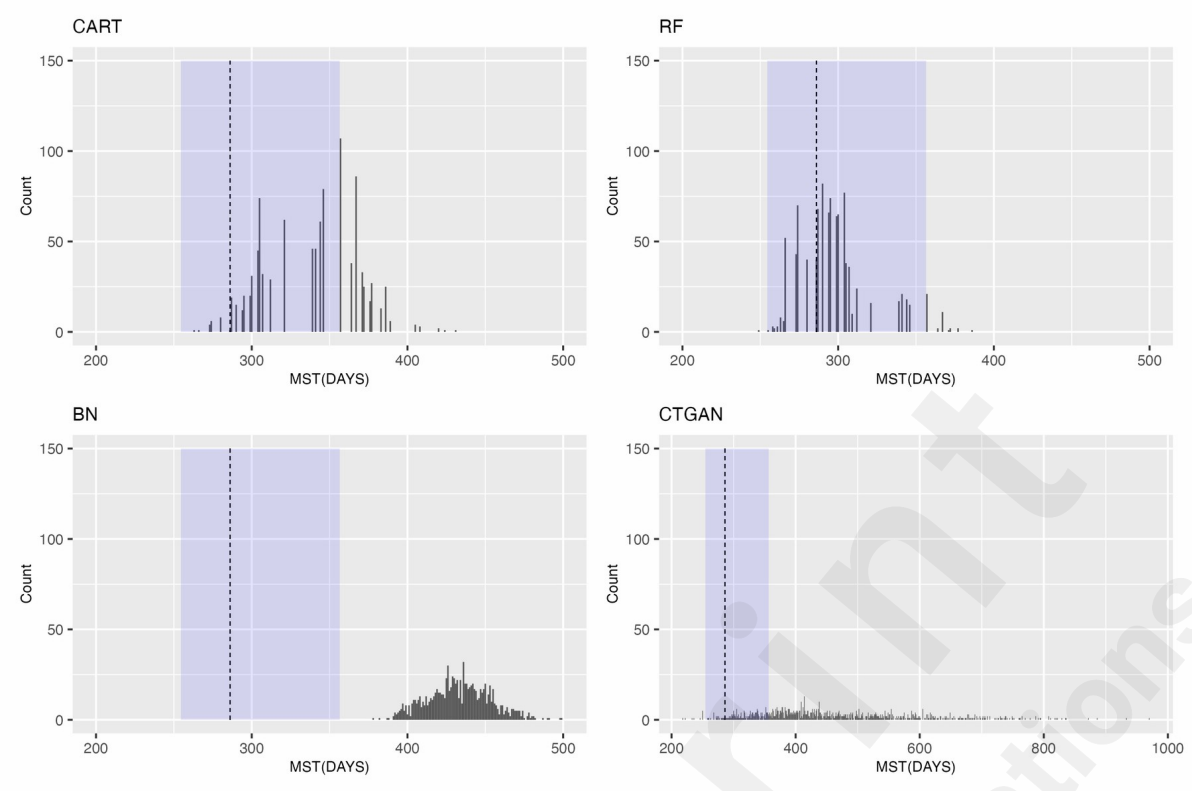


Table 2 presents the number and percentage of the generated MSTS values included within the 95% CI of the MSTA for each trial and each method. In the case of CART for PFS, a high percentage ranging from 88.8% to 98.1% was exhibited for all trials. However, the OS ranged from 60.8% to 96.1%, with some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

For RF, it was demonstrated as a high percentage that were 91.9% for PFS in the NCT00703326 trial and 98.0% for OS in the NCT00460265 trial, whereas in other cases, RF did not indicate as a high percentage as CART.

In the case of BN, 97.6% was demonstrated for PFS in the NCT00703326 trial, and 62.2% was demonstrated for OS, but in the other three trials, BN showed an extremely low percentage ranging from 0.0% to 3.7%.

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

Table 12. The number and percentage of MSTSs falling within the 95% CI of the MSTA.

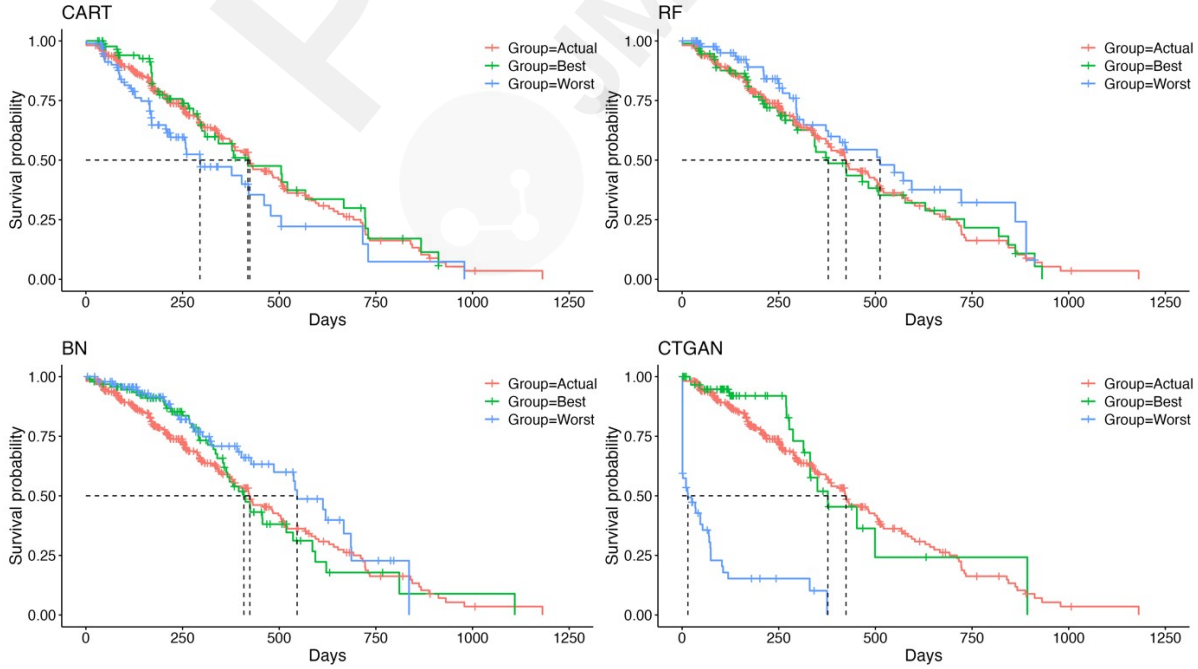
		Clinical Trial ID			
		NCT00119613	NCT00339183	NCT00460265	NCT00703326
	PFS				
MSTA (95% CI)		169 (163-183)	155 (121-168)	133 (121-167)	424 (380-504)
MSTSs					
	CART (N=1000)	981 (98.1%)	888 (88.8%)	955 (95.5%)	918 (91.8%)
	RF(N=1000)	693 (69.3%)	248 (24.8%)	426 (42.6%)	919 (91.9%)
	BN(N=1000)	10 (1.0%)	0 (0.0%)	37 (3.7%)	976 (97.6%)

		Clinical Trial ID			
		NCT00119613	NCT00339183	NCT00460265	NCT00703326
	CTGAN(N=1000)	65 (6.5%)	378 (37.8%)	322 (32.2%)	254 (25.5%)
	OS				
MSTA (95% CI)		276 (259-303)	361 (319-393)	286 (255-357)	1452 (1417-1507)
MSTSs					
	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%)

Figure 3 shows the KM plot for PFS in the NCT00703326 trial. For the best-case curves of CART and RF, the curves were similar to the actual data curve. On the other hand, for BN and CTGAN, even the best-case curves deviated from the actual data curve. In other trials, some SPD did not demonstrate a similar trend. However, at least for the best-case scenarios of CART and RF, the generated synthetic survival curves closely resembled those of 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 in BN and CTGAN, the best-case curves deviated from the actual data curve. These trends were also observed in other trials (Multimedia Appendices 14-16).

Figure 3. KM plots for PFS in the NCT00703326 trial.



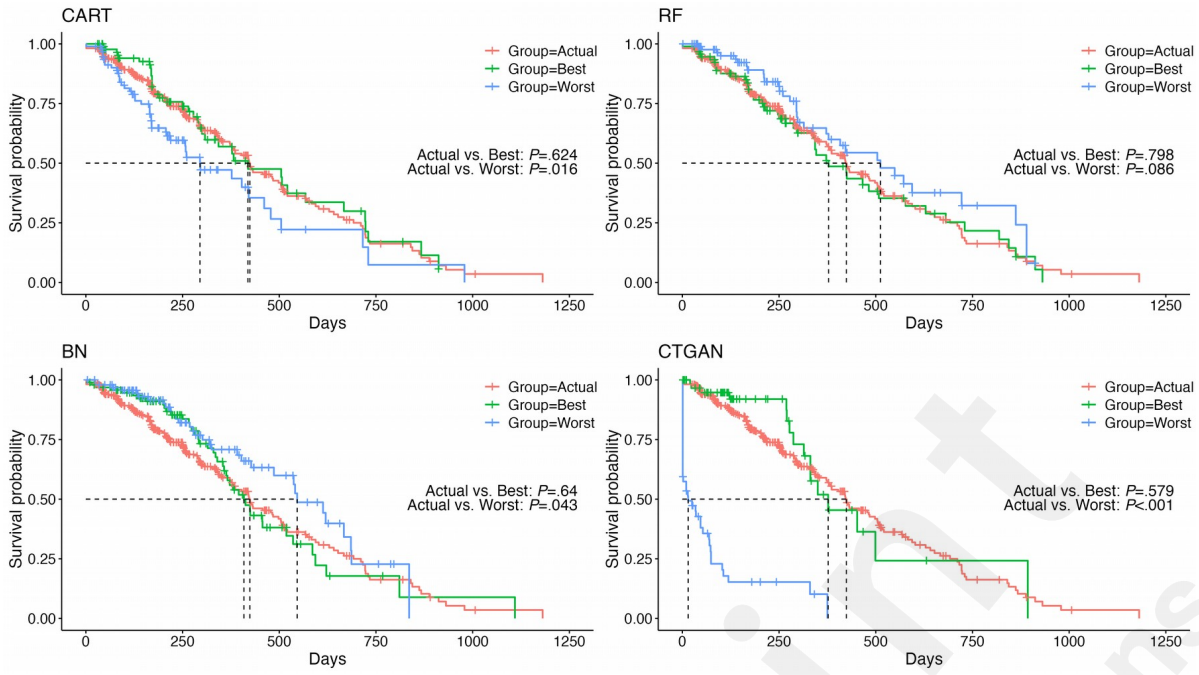
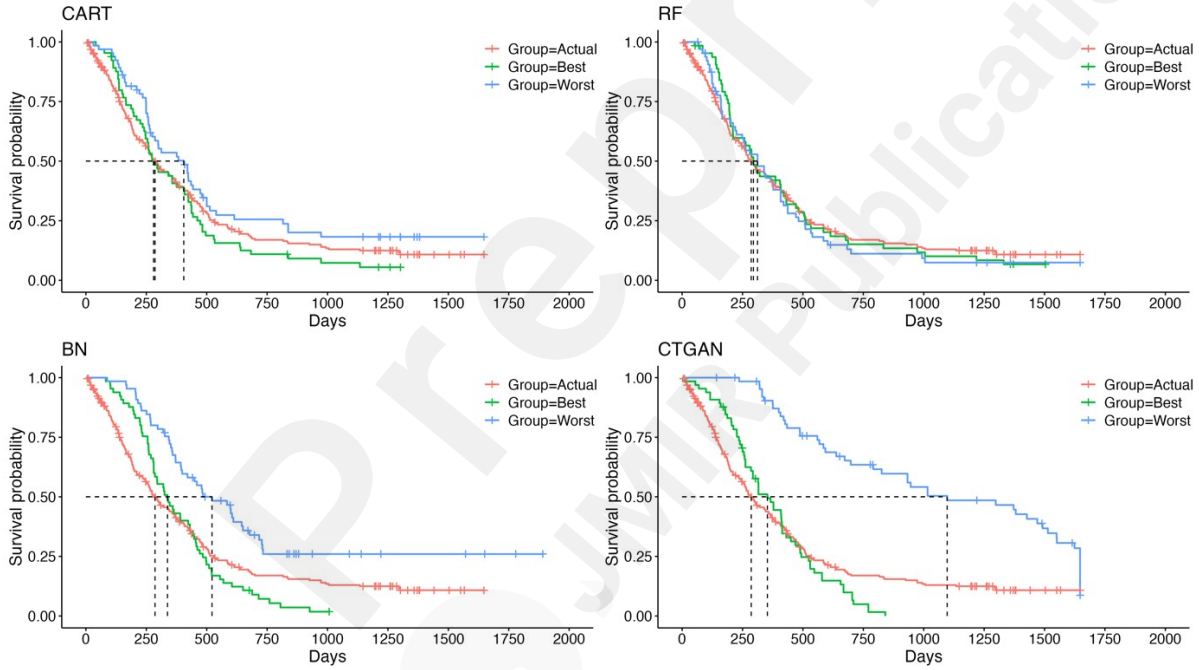
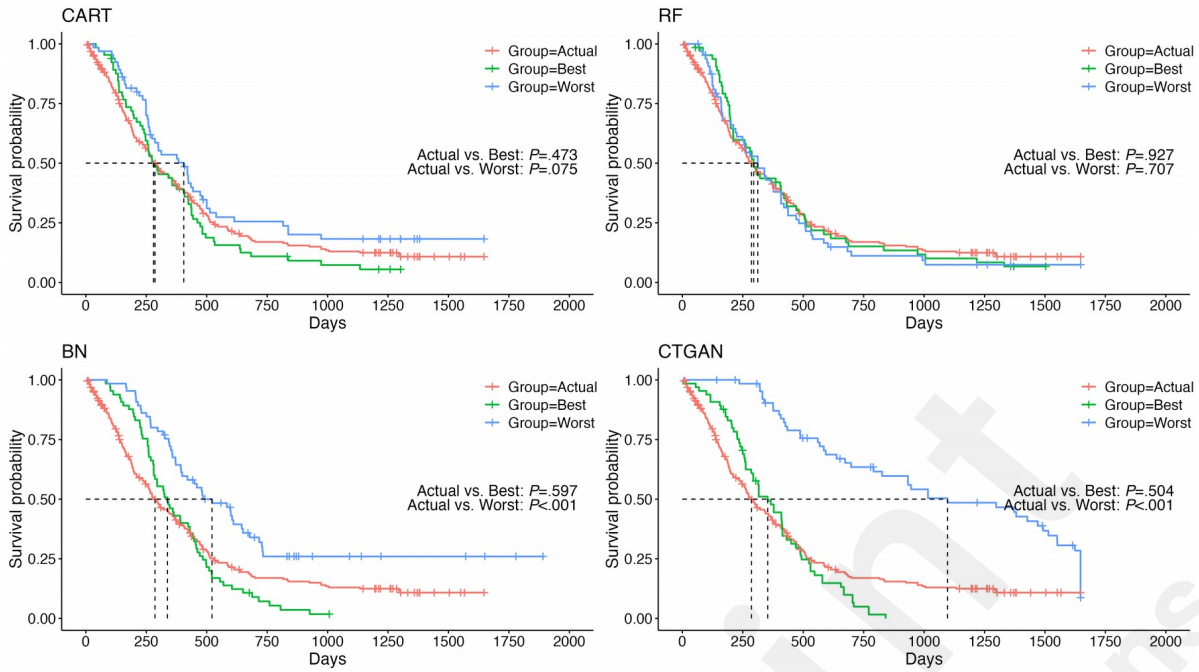


Figure 4. KM plots for OS in the NCT00460265 trial.





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 5. Box plot of PFS HRD for each method and clinical trial.

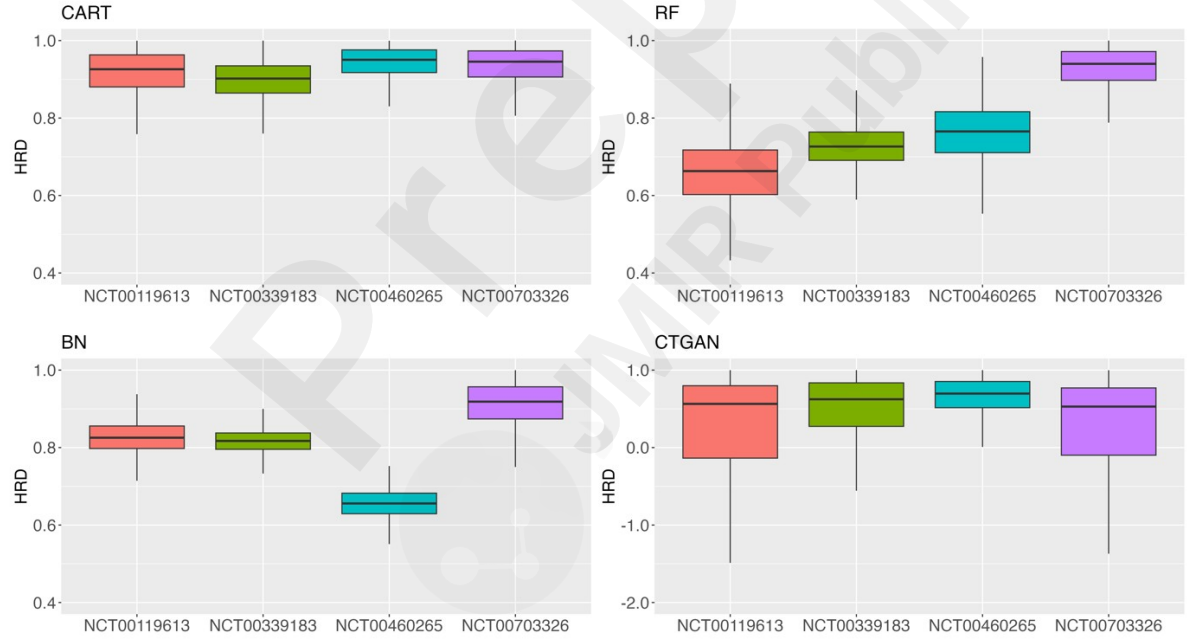
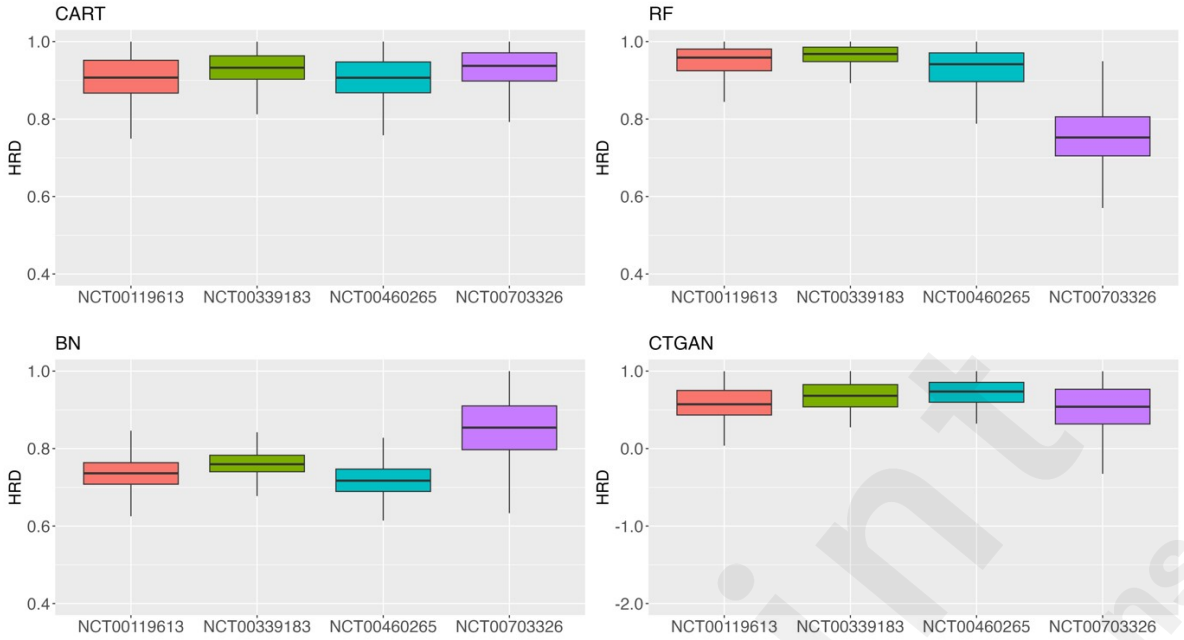


Figure 6. Box plot of OS HRD for each method and clinical trial.



Discussion

Regarding the survival SPD, CART often indicated better results than other methods in evaluations using MST, HRD, and visual analysis via KM plots. Given the crucial importance of the hazard ratio and MST as endpoints in oncology trials [23], demonstrating the utility of both of these evaluation metrics is essential. Therefore, using CART for the generation of 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 employing tree models. RF, with its use of the bootstrap method for resampling and constructing tree models for ensemble learning, is known to reduce biases. In general, in terms of constructing machine learning models with high generalization performance, RF performs better than CART. On the other hand, CART is prone to overfitting as the layers of the tree become deeper [39]. 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 bias 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 MSTS falling within the 95% CI of MSTA was 0% for PFS of the NCT00339183 trial, and for OS, this phenomenon also occurred in the NCT00119613, NCT00339183, and NCT0046265. 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 of NCT00703326. The reason for this discrepancy could not be determined based on the results of this study. Tucker et al. [16] reported that they were able to generate data highly similar to actual data when using BN for the generation of SPD. This result differs from the findings of this study. One notable difference is that while Tucker et al. [16] used a large-scale actual data 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 MSTS falling within the 95% CI of the actual data was low, indicating low performance associated with creating the SPD that reflects the statistical properties of the actual data. On the other hand, Krenmayr et al. [15] reported good performance results when using the same GAN-based methods and RWD. The differences between their study and ours were as follows: their study did not include SPD on survival time or generate multiple SPD datasets 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 NCT00119613 having 232 patients, NCT00339183 having 476 patients, NCT0046265 having 260 patients, and NCT00703326 having 382 patients, while Krenmayr et al. [15] had 500 or more patients. GAN-based methods using deep neural networks are known to perform poorly with small amounts of data [20]. In this study, although NCT00339183 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 dataset yields better results. Thus, there is no contradiction. Another characteristic of using CTGAN in this study was the larger variance in the estimated MSTS, as indicated in Figure 1 and Figure 2. Goncalves et al. [18] showed that using MC-MedGAN, which is a GAN-based method, to generate an SPD from small data resulted in a large standard deviation of the data utility metrics, leading to results with larger variance, similar to those in this study. Therefore, it is extremely challenging to generate useful SPDs by applying GAN-based methods to small datasets, such as clinical trial data.

When generating SPDs for survival data and employing 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 the results of this study, it is concluded that CART, which can concentrate the MSTS within the range of 95% CI of MSTA and approximately 0.9 for HRD, is an efficient method for generating SPD that meets the above conditions. However, even when using CART, slight variations were observed in the MSTS, and some cases fell outside the 95% CI of the MSTA, as revealed by the results of this study. Therefore, it is necessary to verify for practical use that the MSTS is included in the 95% CI of the MSTA and that both are close in value. It is also necessary to verify whether the HRD of the actual data and the SPD is 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, it was observed instances that 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.

In conclusion, as a method for generating SPDs for survival data from small datasets, such as clinical trial data, CART was considered the most effective method for generating SPDs that meet the 2 conditions of having an MSTS close to the MSTA and an HRD close to 1. However, as SPD might be generated that does not meet these two conditions, it is necessary to incorporate mechanisms to improve a CART-based generation method in future work. Overcoming these challenges would make it possible to reduce the recruitment period and costs of clinical trial participants to 50% or more in comparative trials of new drug development against existing therapeutic drugs. This approach would be capable of accelerating 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.

Ethical Review

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

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Abbreviations

BN: Bayesian network
CART: classification and regression trees
CI: confidence interval
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
RF: random forest
RWD: real-world data
SPD: synthetic patient data

Multimedia Appendix 1

Variables used to generate the SPD from NCT00119613

Multimedia Appendix 2

Variables used to generate the SPD from NCT00339183

Multimedia Appendix 3

Variables used to generate the SPD from NCT00460265

Multimedia Appendix 4

Variables used for generating the SPD from NCT00703326

Multimedia Appendix 5

Histogram of the MSTSs for PFS in the NCT00119613 trial. The dashed vertical line represents the MSTA, and the light blue background indicates its 95% CI.

Multimedia Appendix 6

Histogram of the MSTSs for PFS in the NCT00339183 trial. The dashed vertical line represents the MSTA, and the light blue background indicates its 95% CI.

Multimedia Appendix 7

Histogram of the MSTSs for PFS in the NCT00460265 trial. The dashed vertical line represents the MSTA, and the light blue background indicates its 95% CI.

Multimedia Appendix 8

Histogram of the MSTSs for OS in the NCT00119613 trial. The dashed vertical line represents the MSTA, and the light blue background indicates its 95% CI.

Multimedia Appendix 9

Histogram of the MSTSs for OS in the NCT00339183 trial. The dashed vertical line represents the MSTA, and the light blue background indicates its 95% CI.

Multimedia Appendix 10

Histogram of the MSTSs for OS in the NCT00703326 trial. The dashed vertical line represents the MSTA, and the light blue background indicates its 95% CI.

Multimedia Appendix 11

KM plots for PFS in the NCT00119613 trial.

Multimedia Appendix 12

KM plots for PFS in the NCT00339183 trial.

Multimedia Appendix 13

KM plots for PFS in the NCT00460265 trial.

Multimedia Appendix 14

KM plots for OS in the NCT00119613 trial.

Multimedia Appendix 15

KM plots for OS in the NCT00339183 trial.

Multimedia Appendix 16

KM plots for OS in the NCT00703326 trial.

References

1. 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. 2023;13(1):11540. Doi:10.1038/s41598-023-38457-3
2. El Emam K, Jonker E, Arbuckle L, Malin B. A Systematic Review of Re-Identification Attacks on Health Data. Scherer RW, ed. PloS ONE. 2011;6(12):e28071. Doi:10.1371/journal.pone.0028071
3. Kaur D, Sobiesk M, Patil S, et al. Application of Bayesian networks to generate synthetic health data. Journal of the American Medical Informatics Association. 2021;28(4):801-811. Doi:10.1093/jamia/ocaa303
4. 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. 2022;11:695936. Doi:10.3389/fonc.2021.695936
5. Mavrogenis AF, Scarlat MM. Artificial intelligence publications: synthetic data, patients, and papers. International Orthopaedics (SICOT). 2023;47(6):1395-1396. Doi:10.1007/s00264-023-05830-w
6. Pearl J. Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning. In: Proceedings of the 7th Conference of the Cognitive Science Society. University of California; 1985.
7. Azizi Z, Zheng C, Mosquera L, Pilote L, El Emam K. Can synthetic data be a proxy for real clinical trial data? A validation study. BMJ Open. 2021;11(4):e043497. Doi:10.1136/bmjopen-2020-043497
8. 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. 2022;5(3):ooac067. Doi:10.1093/jamiaopen/ooac067
9. Breiman L, ed. Classification and Regression Trees. 1. CRC Press repr. Chapman & Hall/CRC; 1998.
10. Huang GD, Bull J, Johnston McKee K, Mahon E, Harper B, Roberts JN. Clinical trials recruitment planning: A proposed framework from the Clinical Trials Transformation Initiative. Contemporary Clinical Trials. 2018;66:74-79. Doi:10.1016/j.cct.2018.01.003
11. Brownstein JS, Chu S, Marathe A, et al. Combining Participatory Influenza Surveillance with Modeling and Forecasting: Three Alternative Approaches. JMIR Public Health Surveill. 2017;3(4):e83. Doi:10.2196/publichealth.7344
12. U.S. Food and Drug Administration. Considerations for the Design and Conduct of Externally Controlled Trials for Drug and Biological Products Guidance for Industry. Published online February 2023. https://www.fda.gov/media/164960/download
13. Liu R, Rizzo S, Whipple S, et al. Evaluating eligibility criteria of oncology trials using real-world data and AI. Nature. 2021;592(7855):629-633. doi:10.1038/s41586-021-03430-5

14. Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review. Contemporary Clinical Trials Communications. 2018;11:156-164. doi:10.1016/j.conctc.2018.08.001
15. Krenmayr L, Frank R, Drobig C, et al. GANerAid: Realistic synthetic patient data for clinical trials. Informatics in Medicine Unlocked. 2022;35:101118. doi:10.1016/j.imu.2022.101118
16. Tucker A, Wang Z, Rotalinti Y, Myles P. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. npj Digit Med. 2020;3(1):147. doi:10.1038/s41746-020-00353-9
17. Smith A, Lambert PC, Rutherford MJ. Generating high-fidelity synthetic time-to-event datasets to improve data transparency and accessibility. BMC Med Res Methodol. 2022;22(1):176. doi:10.1186/s12874-022-01654-1
18. Goncalves A, Ray P, Soper B, Stevens J, Coyle L, Sales AP. Generation and evaluation of synthetic patient data. BMC Med Res Methodol. 2020;20(1):108. doi:10.1186/s12874-020-00977-1
19. Giuffrè M, Shung DL. Harnessing the power of synthetic data in healthcare: innovation, application, and privacy. npj Digit Med. 2023;6(1):186. doi:10.1038/s41746-023-00927-3
20. Santos M. How to Generate Real-World Synthetic Data with CTGAN. Published April 14, 2023. https://medium.com/towards-data-science/how-to-generate-real-world-synthetic-data-with-ctgan-af41b4d60fde
21. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved Techniques for Training GANs. Published online June 10, 2016. Accessed September 18, 2023. http://arxiv.org/abs/1606.03498
22. Coulter B, Moritz L. Leveraging synthetic data to optimize clinical trials. Globant. Published November 15, 2022. Accessed June 11, 2023. https://stayrelevant.globant.com/en/technology/healthcare-life-sciences/leveraging-synthetic-data-to-optimize-clinical-trials/#:~:text=The%20use%20of%20synthetic%20data,made%20up%20of%20synthetic%20data.
23. 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? The Oncologist. 2019;24(11):1469-1478. doi:10.1634/theoncologist.2019-0175
24. Karmen C, Gietzelt M, Knaup-Gregori P, Ganzinger M. Methods for a similarity measure for clinical attributes based on survival data analysis. BMC Med Inform Decis Mak. 2019;19(1):195. doi:10.1186/s12911-019-0917-6
25. 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
26. Xu L, Skoularidou M, Cuesta-Infante A, Veeramachaneni K. Modeling Tabular data using Conditional GAN. Published online October 27, 2019. Accessed September 10, 2023. http://arxiv.org/abs/1907.00503
27. 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. 2023;6(1):37. doi:10.1038/s41746-023-00771-5
28. Emam K el, Mosquera L, Hoptroff R. Practical Synthetic Data Generation: Balancing Privacy and the Broad Availability of Data. First edition. O'Reilly; 2020.
29. Ursin G, Sen S, Mottu JM, Nygård M. Protecting Privacy in Large Datasets—First We Assess the Risk; Then We Fuzzy the Data. Cancer Epidemiology, Biomarkers & Prevention. 2017;26(8):1219-1224. doi:10.1158/1055-9965.EPI-17-0172
30. Breiman L. Random Forests. Machine Learning. 2001;45(1):5-32. doi:10.1023/A:1010933404324

31. 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. 2020;8(7):e18910. doi:10.2196/18910
32. El Emam K. Status of Synthetic Data Generation for Structured Health Data. JCO Clinical Cancer Informatics. 2023;(7):e2300071. doi:10.1200/CCI.23.00071
33. 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 Clinical Cancer Informatics. 2023;(7):e2300021. doi:10.1200/CCI.23.00021
34. Gonzales A, Guruswamy G, Smith SR. Synthetic data in health care: A narrative review. Johnson A, ed. PLOS Digit Health. 2023;2(1):e0000082. doi:10.1371/journal.pdig.0000082
35. Arora A, Arora A. Synthetic patient data in health care: a widening legal loophole. The Lancet. 2022;399(10335):1601-1602. doi:10.1016/S0140-6736(22)00232-X
36. Wolfien M, Ahmadi N, Fitzer K, et al. Ten Topics to Get Started in Medical Informatics Research. J Med Internet Res. 2023;25:e45948. doi:10.2196/45948
37. 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. 2022;6(7):e29337. doi:10.2196/29337
38. 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 Pharma and Therapeutics. 2022;111(1):77-89. doi:10.1002/cpt.2480
39. Hayes T, Usami S, Jacobucci R, McArdle JJ. Using Classification and Regression Trees (CART) and random forests to analyze attrition: Results from two simulations. Psychology and Aging. 2015;30(4):911-929. doi:10.1037/pag0000046
40. El Emam K, Mosquera L, Fang X, El-Hussuna A. Utility Metrics for Evaluating Synthetic Health Data Generation Methods: Validation Study. JMIR Med Inform. 2022;10(4):e35734. doi:10.2196/35734

Supplementary Files

Untitled.
URL: http://asset.jmir.pub/assets/a1c888a47e7565aa739d9ac19efbb33e.docx

Multimedia Appendixes

Variables used to generate SPD from NCT00119613.
URL: http://asset.jmir.pub/assets/43890a6f77523b973034a56b79bd3f8e.docx

Variables used to generate SPD from NCT00339183.
URL: http://asset.jmir.pub/assets/21b6bdba7ea8f966d838658a152e4460.docx

Variables used to generate SPD from NCT00460265.
URL: http://asset.jmir.pub/assets/50ebe3b4a99d9ee6ef506f128e58769d.docx

Variables used for generating SPD from NCT00703326.
URL: http://asset.jmir.pub/assets/97079e632630301d55c3a498e99f7d41.docx

Histogram of the MSTSs for PFS in the NCT00119613 trial. The dashed vertical line represents the MSTA, and the light blue background indicates its 95% CI.
URL: http://asset.jmir.pub/assets/8605cb3a3e244a9f32c838d6c606740f.docx

Histogram of the MSTSs for PFS in the NCT00339183 trial. The dashed vertical line represents the MSTA, and the light blue background indicates its 95% CI.
URL: http://asset.jmir.pub/assets/c7a50e7cd01ca73e5684e164b3406c94.docx

Histogram of the MSTSs for PFS in the NCT00460265 trial. The dashed vertical line represents the MSTA, and the light blue background indicates its 95% CI.
URL: http://asset.jmir.pub/assets/45fdef464e7def54d06f5ec2e18009dd.docx

Histogram of the MSTSs for OS in the NCT00119613 trial. The dashed vertical line represents the MSTA, and the light blue background indicates its 95% CI.
URL: http://asset.jmir.pub/assets/0407e8333e6f72ce8e61394a572871fb.docx

Histogram of the MSTSs for OS of the NCT00339183 trial. The dashed vertical line represents the MSTA, and the light blue background indicates its 95% CI.
URL: http://asset.jmir.pub/assets/13dc8d65240e7d433c553c342b9a3c95.docx

Histogram of the MSTSs for OS of the NCT00703326 trial. The dashed vertical line represents the MSTA, and the light blue background indicates its 95% CI.
URL: http://asset.jmir.pub/assets/1d2bb01e6017b3a88d0b359bd7328280.docx

KM plots for PFS in the NCT00119613 trial.

URL: http://asset.jmir.pub/assets/07ba8bc7bf16e69ea730d253375b8988.docx

KM plots for PFS in the NCT00339183 trial.
URL: http://asset.jmir.pub/assets/27e2fd2bdc3d5d7ab0c69ab69cae93e7.docx

KM plots for PFS in the NCT00460265 trial.
URL: http://asset.jmir.pub/assets/2f73949265e488ff2bf2b71b3b468ce1.docx

KM plots for OS in the NCT00119613 trial.
URL: http://asset.jmir.pub/assets/f59f6a4fa691eda170e44dcc55a90693.docx

KM plots for OS in the NCT00339183 trial.
URL: http://asset.jmir.pub/assets/5df9f5354a85176fb014cdfc3fd97b01.docx

KM plots for OS in the NCT00703326 trial.
URL: http://asset.jmir.pub/assets/431a498ef23d4a6114fecc68eadaf65c.docx