Advancing Cancer Research with Synthetic Data Generation in Low-Data Scenarios

Patricia A. Apellániz, Borja Arroyo Galende, Ana Jiménez, Juan Parras, and Santiago Zazo
Information Processing and Telecommunications Center, ETS Ingenieros de Telecomunicación, Universidad Politécnica de Madrid

Acknowledgments

This work was supported by the GenoMed4All and SYNTHEMA projects from European Union's Horizon 2020 Research and Innovation Program under Grant 101017549 and Grant 101095530. However, views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

Additional experiments

This document presents additional experiments conducted to validate our research findings further. These experiments were performed with supplementary datasets, and the results are presented in the same format as the main results for consistency and comparison purposes.

Table 1: Additional medical datasets used in experiments.

Dataset	Number of samples	Number of features	Data types	Task
Diabetes_H	253,680	22	Binary and discrete	Classification
Diabetes_130	101,766	45	Binary and discrete	Classification
Whas	1,638	7	Binary, continuous and discrete	Survival Analysis
Pbc	418	19	Binary, continuous and discrete	Survival Analysis
Std	877	23	Binary and discrete	Survival Analysis

Initially, we used two additional classification datasets containing sufficient samples to facilitate initial testing. These datasets, Diabetes_H[1] and Diabetes_130 [1], are related to diabetes and enabled us to perform variations in the number of samples, allowing us to compare the results under conditions of sample scarcity to those obtained with an adequate number of samples. Subsequently, we extended our research to more realistic scenarios involving datasets with limited samples. We include this phase's results from three additional SA datasets. Unlike previous experiments, these datasets are not cancer-related. Instead, they were selected from publicly available sources and possess similar characteristics to the cancer SA data used earlier, specifically the scarcity of samples and the diversity of data types. The survival datasets include Whas

[1] `https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset`

[2], which pertains to heart attack cases, Pbc [3], which is related to autoimmune liver disease, and Std [4], which concerns sexually transmitted diseases. Table 1 summarizes the additional datasets used in these experiments, detailing their key characteristics. This comprehensive approach ensures a thorough evaluation of the methodology across various data types, reinforcing our findings' robustness and applicability.

Including these additional datasets allows us to comprehensively test our proposed synthetic data generation methodology's performance and generalizability under both optimal and constrained conditions, thereby providing a more nuanced understanding of its efficacy and limitations.

Classification datasets

Table 2 presents the results for the Diabetes_H dataset. Focusing on the JS divergence, the 'Big data' scenario ($N = 10,000$) achieves an upper bound JS divergence value of 0.278 ± 0.094. In contrast, the more realistic 'Low data' scenario ($N = 100$) exhibits a significantly higher JS divergence of 0.860 ± 0.001, indicating considerable scope for improvement. Consistent with the results from the Heart dataset, Model-Averaging and DRS techniques are the most effective, yielding the lowest JS divergences. The KL divergence results similarly show improvements across all techniques, with the DRS method particularly standing out. Regarding clinical utility validation, the accuracy obtained with few real data samples (0.500 ± 0.114) overlaps with the 'Big data' scenario (0.606 ± 0.016). This trend matches the Heart dataset results and persists across all cases, indicating that generating synthetic data does not significantly impact classification performance. This suggests that accurately generating a few critical variables is sufficient to maintain classification accuracy, even if other variables are less well-represented.

Table 2: Validation results for the Diabetes_H dataset across different scenarios. The 'Big data' scenario ($N = 10,000$ samples) provides optimal conditions for generating reliable synthetic data, while the 'Low data' scenario ($N = 100$ samples) poses significant challenges for STDG. Similarity validation includes divergences (JS and KL), with lower values preferred, and **bold** indicating improvement from applied techniques. Clinical utility validation compares accuracy metrics for models trained on real, synthetic, and fine-tuned synthetic data, with higher values denoting better performance. **Bold** in clinical utility metrics indicates significance ($p < 0.01$). All results are expressed as *mean (std)*.

Scenario	SIMILARITY VALIDATION		CLINICAL UTILITY VALIDATION		
	JS	KL	Real Acc	Synth Acc	Synth Fine-Tuned Acc
Big data	0.258 (0.122)	0.724 (0.267)	0.605 (0.018)	0.598 (0.023)	0.621 (0.016)
Low data	0.841 (0.005)	7.094 (0.936)	0.529 (0.143)	0.634 (0.113)	0.607 (0.067)
Pre-train	**0.723 (0.015)**	**3.765 (0.075)**	N/A	0.540 (0.274)	0.602 (0.049)
AVG	**0.697 (0.013)**	**3.833 (0.192)**	N/A	0.548 (0.046)	0.571 (0.050)
DRS	**0.709 (0.010)**	**3.611 (0.148)**	N/A	0.528 (0.060)	0.532 (0.055)

The results of the last sufficient sample dataset, Diabetes_130, are presented in Table 3. This table illustrates higher divergences across all scenarios, a phenomenon likely ascribed to the dataset's intricate nature and elevated dimensionality, encompassing 45 covariates. Despite this complexity, there is a

Table 3: Validation results for the Diabetes_130 dataset across different scenarios. The 'Big data' scenario ($N = 10,000$ samples) provides optimal conditions for reliable synthetic data generation, while the 'Low data' scenario ($N = 100$ samples) presents challenges for STDG. Similarity validation includes divergences (JS and KL), where lower values are preferred, and **bold** highlights improvements from applied techniques. Clinical utility validation compares accuracy metrics for models trained on real, synthetic, and fine-tuned synthetic data, with higher values indicating better performance. **Bold** in clinical utility metrics denotes significance ($p < 0.01$). All results are expressed as *mean (std)*.

Scenario	SIMILARITY VALIDATION		CLINICAL UTILITY VALIDATION		
	JS	KL	Real Acc	Synth Acc	Synth Fine-Tuned Acc
Big data	0.469 (0.054)	0.968 (0.181)	0.452 (0.019)	0.452 (0.014)	0.455 (0.019)
Low data	0.979 (0.002)	16.217 (2.972)	0.298 (0.160)	0.252 (0.160)	0.256 (0.156)
Pre-train	**0.953 (0.002)**	21.535 (1.580)	N/A	0.386 (0.051)	0.381 (0.051)
AVG	**0.943 (0.002)**	15.335 (2.071)	N/A	0.279 (0.153)	0.292 (0.155)
DRS	**0.944 (0.003)**	19.544 (1.908)	N/A	0.444 (0.103)	0.438 (0.098)

noticeable enhancement in JS divergence through applying various techniques outlined in the methodology, underscoring the approach's efficacy even when handling challenging data. Conversely, no improvement is observed in KL divergence, highlighting the inherent difficulty of generating high-quality synthetic data as the dataset's dimensionality increases, complicating the accurate capture of inter-variable dependencies. The results of clinical utility validation parallel those of the previous dataset, indicating no significant difference between utilizing real or synthetic data for classifier training. Accuracy metrics remain consistent across all scenarios, reflecting the dataset's complex nature. These findings, consistent with results from other datasets, affirm that JS divergence is a robust and reliable metric for similarity validation. Additionally, clinical utility validation alone is not a reliable measure of synthetic data quality.

Survival Analysis data

Our study's additional SA datasets maintain the heterogeneity characteristic of the previously examined cancer-related datasets. Specifically, the Gbsg dataset was sourced from the Pycox package in Python [5], while the Whas, Pbc, and Std datasets were obtained from the SAVAE repository, a state-of-the-art SA model [6]. This selection is consistent with the practical challenges of SA studies, which often involve smaller sample sizes and intricate data relations.

Tables 4, 5 and 6 present the C-index and IBS results for the three cases (Real, Synthetic, and Synthetic Fine-Tuned) in each scenario, paralleling the previous experiments. The findings indicate no significant difference in performance metrics when employing the methodologies compared to not using them. Furthermore, there is no significant difference between utilizing a relatively high number of samples versus a small sample size, suggesting that clinical utility validation alone is insufficient to determine the effectiveness of STDG. Given the limitations of this validation, we also present the Kaplan-Meier estimations in Fig. A.1 to compare survival functions between real data and synthetic data generated from high samples (upper bounds) and low data with and without the methodology. Although these cases present more similar curves, this is

Table 4: Validation results for the Whas dataset across different scenarios. The 'Big data' scenario ($N = 1,311$ samples, 80% of the data) provides optimal conditions for synthetic data generation, while the 'Low data' scenario ($N = 100$ samples) presents significant challenges for STDG. SA metrics include C-index (higher values indicate better performance) and IBS (lower values are preferable). **Bold** values indicate significant improvements using the methodology, while * highlights significant disadvantages. All results are expressed as *mean (std)*.

Scenario	Real CI	Synth CI	Synth Fine-Tuned CI	Real IBS	Synth IBS	Synth Fine-Tuned IBS
Big data	0.731 (0.035)	0.715 (0.037)	0.714 (0.036)	0.178 (0.030)	0.179 (0.030)	0.185 (0.030)
Low data	0.703 (0.036)	0.698 (0.036)	0.696 (0.038)	0.177 (0.030)	0.180 (0.030)	0.182 (0.030)
Pre-train	N/A	0.689 (0.037)	0.675 (0.037)	N/A	0.183 (0.030)	0.185 (0.030)
AVG	N/A	0.725 (0.036)	0.720 (0.037)	N/A	0.175 (0.030)	0.174 (0.030)
DRS	N/A	0.710 (0.037)	0.708 (0.036)	N/A	0.178 (0.030)	0.179 (0.030)

Table 5: Validation results for the Pbc dataset across different scenarios. The 'Big data' scenario ($N = 335$ samples, 80% of the data) provides optimal conditions for synthetic data generation, while the 'Low data' scenario ($N = 100$ samples) presents significant challenges for STDG. SA metrics include C-index (higher values indicate better performance) and IBS (lower values are preferable). **Bold** values indicate significant improvements using the methodology, while * highlights significant disadvantages. All results are expressed as *mean (std)*.

Scenario	Real CI	Synth CI	Synth Fine-Tuned CI	Real IBS	Synth IBS	Synth Fine-Tuned IBS
Big data	0.815 (0.069)	0.815 (0.065)	0.826 (0.063)	0.174 (0.062)	0.156 (0.058)	0.159 (0.058)
Low data	0.541 (0.063)	0.521 (0.055)	0.530 (0.054)	0.242 (0.051)	0.246 (0.046)	0.249 (0.046)
Pre-train	N/A	0.829 (0.063)	0.834 (0.059)	N/A	0.153 (0.060)	0.136 (0.054)
AVG	N/A	0.817 (0.062)	0.836 (0.059)	N/A	0.153 (0.058)	0.147 (0.056)
DRS	N/A	0.826 (0.063)	0.840 (0.061)	N/A	0.163 (0.058)	0.157 (0.060)

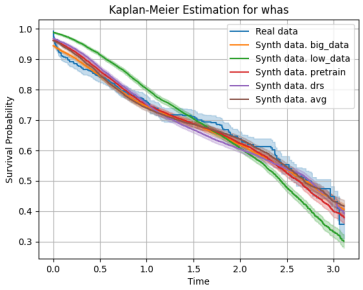
Table 6: Validation results for the Std dataset across different scenarios. The 'Big data' scenario ($N = 702$ samples, 80% of the data) provides optimal conditions for synthetic data generation, while the 'Low data' scenario ($N = 100$ samples) presents significant challenges for STDG. SA metrics include C-index (higher values indicate better performance) and IBS (lower values are preferable). **Bold** values indicate significant improvements using the methodology, while * highlights significant disadvantages. All results are expressed as *mean (std)*.

Scenario	Real CI	Synth CI	Synth Fine-Tuned CI	Real IBS	Synth IBS	Synth Fine-Tuned IBS
Big data	0.544 (0.055)	0.577 (0.057)	0.608 (0.052)	0.224 (0.046)	0.215 (0.044)	0.214 (0.044)
Low data	0.733 (0.079)	0.789 (0.072)	0.804 (0.066)	0.176 (0.061)	0.159 (0.060)	0.191 (0.069)
Pre-train	N/A	0.540 (0.057)	0.536 (0.058)	N/A	0.229 (0.045)	0.234 (0.046)
AVG	N/A	0.553 (0.060)	0.553 (0.057)	N/A	0.225 (0.045)	0.230 (0.045)
DRS	N/A	0.530 (0.057)	0.523 (0.054)	N/A	0.227 (0.045)	0.232 (0.045)

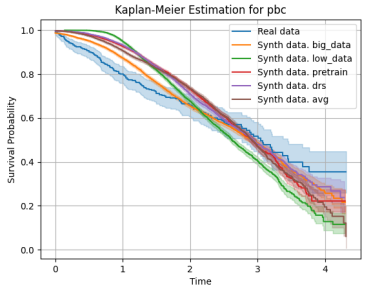
primarily due to the very small sample sizes of the Pbc and Std datasets. However, the curves representing the survival functions of synthetic data using the methodology more closely approximate the upper bounds, whereas the green curve diverges significantly more. Therefore, we continue to confirm that the methodology aids in more accurately generating time-to-event information in these cases.

Different data utility results

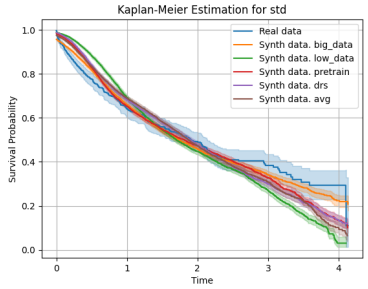
Finally, Tables 7 and 8 present the latest validation analyses for the additional datasets, encompassing both classification and SA. In this validation, we



(a) Whas dataset



(b) Pbc dataset



(c) Std dataset

Figure A.1: Kaplan-Meier estimations with confidence intervals for real and synthetic data across different scenarios for additional datasets. Blue and orange lines represent survival probabilities for real and synthetic data generated with many samples (upper bounds). Red, purple, and brown lines correspond to synthetic data generated using the proposed methodology, showing convergence towards the upper bounds. Green lines depict lower-bound synthetic data, illustrating significant deviations.

modified the final task of the analyzed datasets. We selected different target variables for the classification datasets, while for the SA datasets, we used variables other than time for prediction purposes. These results are consistent with previous findings, indicating that although the DRS technique does not make a significant difference in most cases, it improves feature generation and their relationships in specific instances. Overall, there are more improvements than declines, suggesting that the use of this methodology in cases where datasets are small is beneficial to the outcome.

References

[1] J. Clore, K. Cios, J. DeShazo, B. Strack, Diabetes 130-US Hospitals for Years 1999-2008, UCI Machine Learning Repository, DOI: https://doi.org/10.24432/C5230J (2014).

Table 7: Clinical utility validation results for additional classification datasets. Accuracy comparison between the $N = 100$ samples without methodology case ('Low data') and the DRS technique applied to the lower bound case ('DRS') for each feature used as a classification label. Higher accuracy values indicate better performance. **Bold** values denote a significant advantage, while * values indicate a significant disadvantage in using the methodology. Results are expressed as *mean (std)*.

Diabetes_H			**Diabetes_130**		
Feature	**Low data**	**DRS**	*Feature*	**Low data**	**DRS**
HighBP	0.710 (0.011)	0.708 (0.006)	*race*	0.179 (0.206)	0.343 (0.279)
HighChol	0.646 (0.005)	0.598 (0.011)*	*gender*	0.444 (0.132)	0.476 (0.017)
CholCheck	0.697 (0.209)	0.445 (0.070)	*num_procedures*	0.258 (0.049)	0.153 (0.066)
Smoker	0.530 (0.035)	0.540 (0.015)	*metformin*	0.056 (0.082)	0.064 (0.069)
Stroke	0.514 (0.063)	**0.761 (0.032)**	*repaglinide*	0.632 (0.379)	0.016 (0.019)
HeartDiseaseorAttack	0.635 (0.026)	0.638 (0.030)	*nateglinide*	0.552 (0.060)	0.417 (0.180)
PhysActivity	0.679 (0.010)	0.625 (0.049)	*chlorpropamide*	0.392 (0.234)	0.335 (0.107)
Fruits	0.384 (0.008)	0.388 (0.014)	*glimepiride*	0.166 (0.194)	**0.692 (0.197)**
Veggies	0.706 (0.090)	0.635 (0.030)	*glipizide*	0.216 (0.221)	0.157 (0.242)
HvyAlcoholConsump	0.757 (0.166)	0.534 (0.194)	*glyburide*	0.342 (0.237)	0.107 (0.063)
AnyHealthcare	0.471 (0.093)	**0.668 (0.022)**	*tolbutamide*	0.787 (0.112)	0.867 (0.025)
NoDocbcCost	0.741 (0.210)	0.630 (0.062)	*pioglitazone*	0.446 (0.393)	0.340 (0.304)
DiffWalk	0.783 (0.015)	0.785 (0.004)	*rosiglitazone*	0.218 (0.144)	0.505 (0.391)
Sex	0.531 (0.045)	0.529 (0.043)	*acarbose*	0.444 (0.263)	0.199 (0.335)
Diabetes	0.545 (0.029)	0.494 (0.087)	*miglitol*	0.123 (0.212)	0.240 (0.365)
			tolazamide	0.938 (0.027)	0.941 (0.014)
			insulin	0.284 (0.023)	**0.345 (0.020)**
			glyburide-metformin	0.235 (0.250)	0.480 (0.321)
			change	0.575 (0.007)	**0.615 (0.007)**
			diabetesMed	0.560 (0.045)	0.555 (0.047)
			readmitted	0.379 (0.149)	0.308 (0.165)

Table 8: Clinical utility validation results for additional SA datasets. Accuracy comparison between the $N = 100$ samples without methodology case ('Low data') and the DRS technique applied to the lower bound case ('DRS') for each feature used as a classification label. Higher values indicate better performance. **Bold** values denote a significant advantage in using the methodology. Results are expressed as *mean (std)*.

Whas			**Pbc**			**Std**		
Feature	**Low data**	**DRS**	*Feature*	**Low data**	**DRS**	*Feature*	**Low data**	**DRS**
sex	0.532 (0.060)	0.642 (0.006)	treatment	0.626 (0.027)	0.579 (0.072)	race	0.439 (0.101)	0.600 (0.051)
chf	0.645 (0.007)	**0.680 (0.009)**	sex	0.448 (0.173)	0.617 (0.106)	marital	0.641 (0.027)	0.622 (0.054)
miord	0.552 (0.031)	0.574 (0.067)	ascites	0.414 (0.363)	0.533 (0.363)	iinfct	0.370 (0.112)	0.282 (0.088)
event	0.736 (0.015)	0.754 (0.001)	hepatom	0.588 (0.012)	0.621 (0.028)	os12m	0.501 (0.059)	0.581 (0.031)
			spiders	0.681 (0.056)	0.714 (0.023)	os30d	0.537 (0.209)	0.613 (0.015)
			edema	0.457 (0.072)	0.517 (0.034)	rs12m	0.674 (0.075)	0.651 (0.000)
			stage	0.362 (0.012)	0.367 (0.030)	rs30d	0.754 (0.162)	0.712 (0.347)
			event	0.836 (0.009)	0.833 (0.011)	abdpain	0.790 (0.086)	0.622 (0.247)
						discharge	0.512 (0.011)	0.507 (0.015)
						dysuria	0.539 (0.033)	0.651 (0.195)
						condom	0.339 (0.056)	0.259 (0.149)
						itch	0.690 (0.251)	0.567 (0.299)
						lesion	0.600 (0.045)	0.634 (0.033)
						rash	0.552 (0.207)	0.639 (0.234)
						lymph	0.659 (0.203)	0.486 (0.271)
						vagina	0.658 (0.146)	0.647 (0.136)
						dchexam	0.493 (0.233)	0.518 (0.292)
						event	0.561 (0.035)	0.565 (0.025)

[2] D. W. Hosmer Jr, S. Lemeshow, S. May, Applied survival analysis: regression modeling of time-to-event data, John Wiley & Sons, 2011.

[3] T. Therneau, P. Grambsch, Modeling Survival Data: Extending The Cox Model, Vol. 48, 2000. doi:10.1007/978-1-4757-3294-8.

[4] M. B. Rao, Survival analysis, techniques for censored and truncated data, Technometrics 40 (2) (1998) 159–160. arXiv:https://doi.org/10.1080/00401706.1998.10485206, doi:10.1080/00401706.1998.10485206.
URL https://doi.org/10.1080/00401706.1998.10485206

[5] H. Kvamme, Ørnulf Borgan, I. Scheel, Time-to-event prediction with neural networks and cox regression (2019). arXiv:1907.00825.

[6] P. A. Apellániz, J. Parras, S. Zazo, Leveraging the variational bayes autoencoder for survival analysis, Scientific Reports 14 (1) (2024) 24567.