Synthetic Clinical Trial Data while Preserving Subject-Level Privacy

Mandis Beigi, PhD[1], Afrah Shafquat, PhD[1], Jason Mezey, PhD[2], Jacob Aptekar, MD, PhD[1]

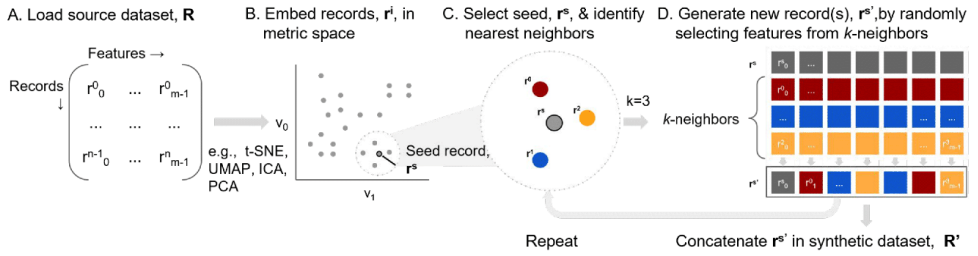
[1]Medidata, New York, NY, USA, [2]Cornell University, Ithaca, NY, USA

Motivation:

Around 23,000+ clinical trials, 6.3M+ subjects are exclusively available to Medidata. This data is valuable for research and drug development but is siloed data due to privacy concerns and HIPAA regulations. Generating synthetic data will provide the means for data sharing. It also allows us to up-sample under-represented populations as well as augmenting the training data when datasizes are limited.

Synthesis Method:

- Co-segregation of attributes
- Truncated gaussian noise
- Multivariate outlier detection and removal
- K-anonymity check
- Multi-sponsor patient generation



Fidelity Evaluation:

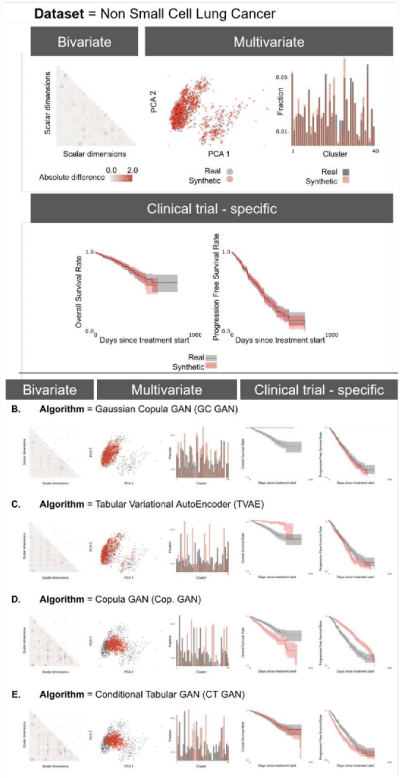
Compared with four State-of-the-art synthesizers from the SDGym benchmark: GaussianCopulaGAN, TVAE, CopulaGAN and CTGAN.

Tested on four clinical trial datasets ranging from 698 to 4369 patients as well as other canonical datasets from the benchmark.

To compare the synthesized data with the original, we performed univariate tests such as the chi-square and Kolmogorov-Smirnov tests as well as bivariate tests and multivariate tests such as bag-of-words and silhouette coefficients.

Benchmark

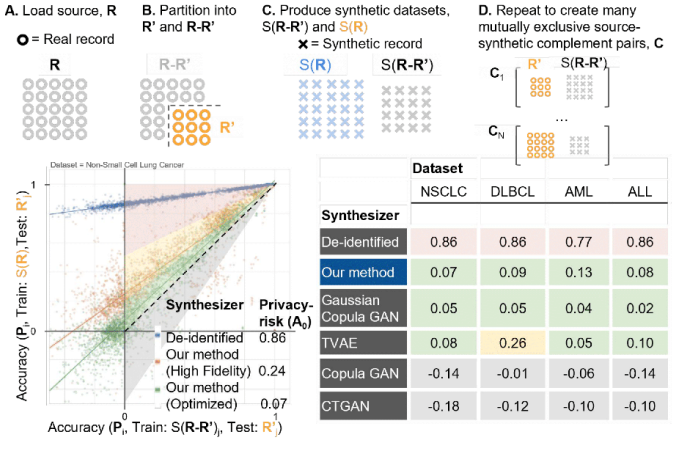
SDGym SDV



F. Fidelity quantification summary

Category		Univariate	Bivariate	Multivariate		Clinical trial - specific	
Metric		Dimensions where K-S, Chi-Square p<0.05	Correlation, mean absolute difference	Silhouette score	"Bag of Words" dist	log-rank Overall Survival K-M	log-rank Progression Free Survival K-M
Best		0/N	0	0	0	1	1
Worst		N/N	1	1		0	0
Dataset	Algorithm						
Non small cell lung cancer	**Our method**	**0/171**	0.11	**0.00**	**0.04**	0.47	**0.89**
	GC GAN	56/171	**0.08**	0.00	0.11	**8.75E-26**	0.46
	TVAE	**98/171**	0.10	0.01	0.22	5.62E-16	0.39
	Cop. GAN	64/171	**0.22**	**0.04**	0.24	9.63E-08	**9.03E-09**
	CTGAN	58/171	0.20	0.03	**0.24**	**0.56**	0.6
Diffuse Large B Cell Lymphoma	**Our method**	**2/174**	**0.07**	**0.00**	**0.02**	**0.99**	NA
	GC GAN	88/174	0.09	0.01	0.21	**3.31E-35**	NA
	TVAE	**94/174**	0.12	0.03	**0.16**	3.82E-09	NA
	Cop. GAN	82/174	**0.22**	0.01	0.33	2.49E-04	NA
	CTGAN	83/174	0.22	**0.02**	0.32	3.02E-11	NA
Acute Lymphoblastic Leukemia	**Our method**	**11/142**	**0.08**	**0.00**	**0.01**	NA	NA
	GC GAN	72/142	0.20	**0.03**	**0.26**	NA	NA
	TVAE	73/142	0.48	0.00	0.08	NA	NA
	Cop. GAN	101/142	**1.14**	0.02	0.17	NA	NA
	CTGAN	**115/142**	1.10	0.03	0.15	NA	NA
Acute Myeloid Leukemia	**Our method**	**0/108**	**0.19**	**0.00**	**0.08**	**0.06**	NA
	GC GAN	36/108	0.29	0.01	0.19	0.68	NA
	TVAE	**58/108**	0.30	**0.07**	**0.39**	**5.86E-35**	NA
	Cop. GAN	44/108	**0.42**	0.03	0.30	0.11	NA
	CTGAN	36/108	**0.42**	0.02	0.31	0.20	NA

Privacy Evaluation:



Synthesizer	Dataset: NSCLC	DLBCL	AML	ALL
De-identified	0.86	0.86	0.77	0.86
Our method	0.07	0.09	0.13	0.08
Gaussian Copula GAN	0.05	0.05	0.04	0.02
TVAE	0.08	0.26	0.05	0.10
Copula GAN	-0.14	-0.01	-0.06	-0.14
CTGAN	-0.18	-0.12	-0.10	-0.10

Conclusion:

Configurable – Tunable for desired privacy and fidelity levels; up-sampling desired cohorts

Efficient – Runs on a single CPU

High fidelity – Outperform all leading synthesizers

Private – on par privacy level with the state-of-the-art

Light weight – Open and extensible; designed for the community

Scalable – Outperforms for all data sizes especially small trial sizes