Metalog Synthetic Data Generation for Healthcare

Raul Rios, Eric Haney, Randal Allen
Lone Star Analysis
Addison, TX
rrios@lone-star.com, ehaney@lone-star.com, rallen@lone-star.com

ABSTRACT

Deidentification is a common data anonymization technique for protecting sensitive data (e.g., personally identifying information). However, studies have shown that deidentification is often insufficient to prevent re-identification of individuals from their data; in some cases, zip code, gender, and age are sufficient characteristics to uniquely identify individuals. An alternative approach that guarantees anonymization is synthetic data generation (SDG), which algorithmically generates artificial data to mimic a desired real dataset. SDG irreversibly anonymizes the output data since synthetic data has no direct link to individual samples in the source, private data. Complex industries such as healthcare, finance, and national security often employ SDG when conducting analysis and simulation.

A traditional approach to SDG is to create a model by fitting a probability distribution to the data and then sampling that distribution. A plethora of distributions are available for smooth, continuous data (e.g., normal, Student's t, Cauchy, logistic, or beta) however, in many cases, it may be unclear which distribution would best capture the reality of the data (shape, skewness, tail behavior, etc.). In addition, the distribution fitting process for these approaches is often complex, subjective, and non-convergent. Metalog distributions (Keelin, 2016), address these issues by proposing a distribution that is more flexible and easier to use for fitting datasets. We demonstrate a workflow for using the metalog distribution for SDG and analysis of simulated health data and assess the model fidelity.

ABOUT THE AUTHORS

Raul Rios is a Senior R&D Engineer at Lone Star Analysis. He is responsible for designing and implementing analysis infrastructure and tools, analytics workflow processes, and complex data visualizations. His expertise spans algorithms for signal and image processing, developing, testing, and fielding custom software and hardware solutions for data collection, and actively supporting our Research and Development efforts. His technical skills include architecture design, modeling and simulation, testing and validation, and data analysis and interpretation. Raul received his B.S. in Aerospace Engineering from the Massachusetts Institute of Technology.

Eric Haney is the Chief Technology Officer at Lone Star Analysis. He is responsible for the development, deployment, and support of multiple analytics platforms, including TruNavigator™ and TruPredict™. He also leads the research and development division at Lone Star, Cipher Alchemy. As part of his work, he has been awarded two patents in edge analytics and digital twins. He holds a Ph.D. in Aerospace Engineering (University of Texas at Arlington) and a B.S. in Aerospace Engineering (Texas A&M University).

Randal Allen is the Chief Scientist of Lone Star Analysis. He is responsible for applied research and technology development across a wide range of M&S disciplines and manages intellectual property. He maintains a CMSP with NTSA. He has published and presented technical papers and is co-author of the textbook, "Simulation of Dynamic Systems with MATLAB and Simulink." He holds a Ph.D. in Mechanical Engineering (University of Central Florida), an Engineer's Degree in Aeronautical and Astronautical Engineering (Stanford University), an M.S. in Applied Mathematics and a B.S. in Engineering Physics (University of Illinois, Urbana-Champaign). He serves as an Adjunct Professor/Faculty Advisor in the MAE department at UCF where he has taught over 20 aerospace-related courses.

Synthetic Data Generation for Healthcare

Raul Rios, Eric Haney, Randal Allen
Lone Star Analysis
Addison, TX
rrios@lone-star.com, ehaney@lone-star.com, rallen@lone-star.com

INTRODUCTION

In the realm of data analytics, data silos are a real problem. Data silos arise when organizational structures, policies, workflows, and technology artificially separate data access from potential users. Historically, this was often a desired effect to protect sensitive data from unauthorized access. However, in the modern age and with the advent of big data, cloud computing, and zero-trust cybersecurity policies, the utility of data silos has diminished to the point of being a hindrance to organizations and can even be security risks (Tidwell, 2024). Data silos create losses in efficiency and value in many domains (e.g., healthcare, government and national security, business/finance) by limiting cross-collaboration, creating data inconsistencies, and wasting resources (Talend, 2020). Referring to the healthcare domain, Slabodkin (2021) writes that "a health data disconnect between clinicians and data scientists is wasting precious medical research and healthcare resources, hampering innovation and resulting in poorer outcomes than would otherwise be achievable." A Worldwide Business Research Insights (2022) survey of financial leaders indicates that data silos are the leading barrier to innovation. The Department of Defense (DoD) has a data management strategy that is focused on treating data as a product to break down data silos and improve data quality to make more effective decisions (Department of Defense, 2023).

One strategy for mitigating the negative effects of data silos while still protecting sensitive data contained within is by using synthetic data. Synthetic data generation (SDG) is a process by which an implicit or explicit model of the real dataset is constructed. This model can then synthesize new data samples with the same characteristics as the source data.

This paper proposes the use of an SDG framework for data anonymization that leverages the metalogistic distribution (also known as the metalog), an infinitely flexible distribution model (Keelin, 2016). As an exemplary use case, this framework is applied to a representative healthcare dataset. The output synthetic dataset is then tested for similarity with the source dataset.

Data Privacy

Personally identifiable information (PII) and protected health information (PHI) are found in every corner of the healthcare industry, including emergency departments, primary care practices, specialty clinics, and pharmaceutical companies. PHI is a subset of PII that is specifically "associated with or derived from a healthcare service event" according to the Northwestern University Institutional Review Board Office (2020). This includes any medical data linked to one or more of the identifiers defined in the Health Insurance Portability and Accountability Act of 1996 (HIPAA), such as names, birthdates, social security numbers, or geographic information, etc. Data silos naturally arise in healthcare both because of the nature of collecting PHI at the point of service and because of the legally and ethically required data protections for PHI. Because of these data silos, the medical research community has historically struggled in the process to access the types of data needed for studies and to access these large amounts of data at scale (Chevrier et al., 2019). For PHI, such as those stored in hospital and clinic electronic health records (EHR), to be used for medical research, direct patient consent is required unless the data are anonymized sufficiently according to ethical guidelines.

The DoD faces some of the same data silo issues that healthcare researchers face. As a federal agency, the DoD, its components, and its business partners must also adhere to data privacy standards. Office of the Under Secretary of Defense for Research and Engineering (2019) established that the DoD must follow HIPPA privacy rules in conjunction with the Privacy Act (1974). The Privacy Act prohibits disclosure of PII by federal agencies except under certain circumstances, one of which is for statistical research purposes. When PII is disclosed for research, it is

specified that the private records should not be "individually identifiable," meaning that the records must be anonymized before distribution. In addition to general protections for PHI and PII, there are specific federal regulations for human subjects in research (Protection of Human Subjects, 2018). Solutions that address data silos and ensure privacy protections are needed to aid research in both the healthcare industry and in the DoD.

Data Anonymization

In medical research literature, anonymization and deidentification are sometimes used interchangeably, perhaps in part because of how HIPAA refers to "de-identification." In this paper, anonymization refers to any process that transforms PHI data such that it would be virtually impossible to link the data back to a specific individual i.e., for a patient to be reidentified. Deidentification herein refers to an anonymization technique that explicitly removes, in whole or in part, the explicit patient identifiers, e.g., names and social security numbers. Deidentification is often the first step in any anonymization process, but this has been shown to be insufficient to anonymize data – 87% of Americans can be uniquely (or near-uniquely) identified by just their birthdate, gender, and 5-digit zip code (Sweeney, 2000).

One can broadly categorize data anonymization techniques into two groups: data manipulation and SDG. The categorization is displayed in Figure 1. Data manipulation techniques are ones in which the raw patient data is modified directly before being shared, and so contains some portion of the original data. Data manipulation includes a) simple operations such as perturbation and data masking, b) privacy models (e.g., k-anonymity), and c) collaborative privacy models such as those that use homomorphic encryption (Zuo et al., 2021). However, all data manipulation techniques have an inherent tradeoff between degree of data anonymization and the data utility. Thus, there is always some risk in reidentification of patients when using data manipulation for anonymization.

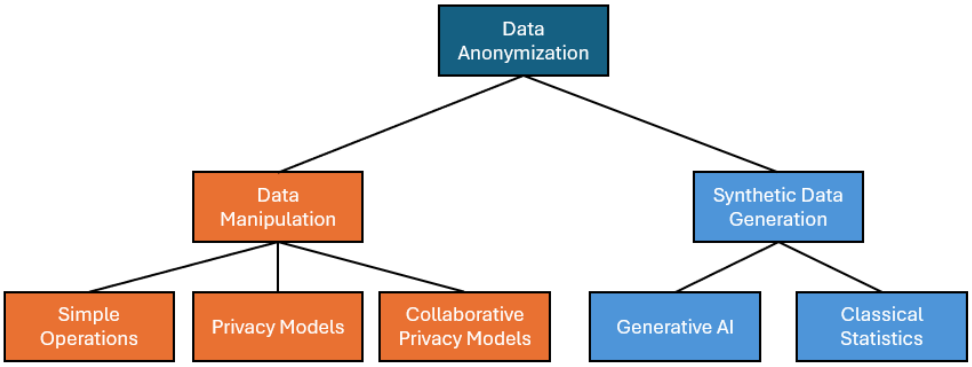

Figure 1. Categories of Data Anonymization Techniques

SDG offers a path towards removing the risk of patient reidentification by instead creating a model of the important attributes of the data and generating new, representative data. There are two primary toolsets for SDG: generative artificial intelligence (AI) such as generative adversarial networks (GANs), and more established classical statistical models, such as distribution-fitting. Generative AI has created a sensation via prompt-based generation as seen in ChatGPT and DALL-E. However, these AI models require an inordinate amount of data, compute resources (see Figure 2), and time to train. In addition, research has shown that these large models can leak data to the end-user and thus compromise the privacy of that data (Carlini et al., 2023).

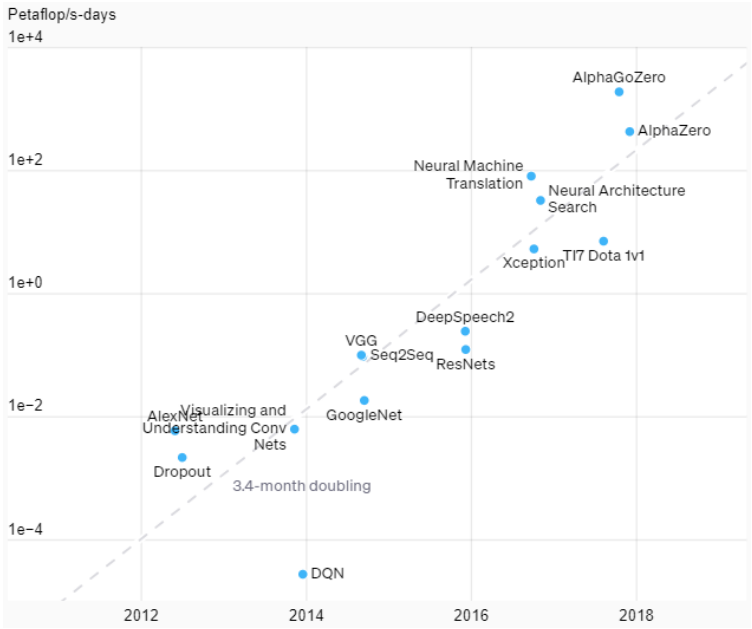

Figure 2. OpenAI Study on AI Training Compute Power (Amodei et al., 2018)

Fitting a probability distribution to data is not new, however, a practitioner must consider many potential distributions and examine many potential parameters before finding the most appropriate model for the data. This is a manual and time-intensive task. In addition, using inaccurate models can lead to poor correspondence between the generated synthetic data and the original data. However, a recent innovation, the metalog distributions (Keelin, 2016), addresses these issues by proposing a distribution that is more flexible and easier to use for fitting datasets. By using this distribution, it is possible to model any number of probabilistic distributions, which removes the need to explicitly choose a specific one by trial and error.

METHOD

To showcase the potential for this metalog modeling process in anonymizing data, we first explain how a multivariate metalog model is created and exercised. We then use this method with an exemplary Covid-19 dataset to generate a synthetic dataset. We compare synthetic and original datasets for similarity.

For the explanation that follows, the problem we are trying to solve is given a dataset X consisting of N samples of a multivariate distribution of D dimension, create a model that generates a synthetic dataset X' with N' samples that approximates the original dataset X. We use x_i to denote individual samples of a dataset X.

Metalog Framework

The metalog distribution is a parametric model on the quantiles of the cumulative distribution function (CDF), Y, of a univariate dataset X in $\mathbb{R}^1$. In constructing a metalog model, the main choice to be made is the "order" of the metalog fit; that is, how many terms are to be used in the model. For those familiar with Taylor series expansions, the metalog order is analogous to the number of terms in a Taylor series approximation to the true quantile function. The quantile function is simply the inverse CDF of a distribution. Equation 1 shows the general equation for the metalog quantile function of order m=4 with parameterized coefficients a_j, j in [1, m].

$$M(y; X, Y) = a_1 + a_2 \ln\frac{y}{1-y} + a_3(y-0.5)\ln\frac{y}{1-y} + a_4(y-0.5) \qquad (1)$$

The order is upper bounded by the number of unique CDF values in Y. To remove m as a hyperparameter of the process, we fit many metalog distributions with CDF Y', each of a different order, and algorithmically select the best

fit to Y from among them. We use the Kolmogorov-Smirnov (KS) statistic on the CDFs as a metric for the quality of the distribution fit, though other metrics could be used for the selection of the appropriate m. Since the metalog order, m, can be arbitrarily high (up to the number of unique CDF samples) care should be taken to avoid overfitting. Figure 3 shows an example comparing the distribution shapes of metalogs of different orders fit to the same Gumbel-distributed dataset.

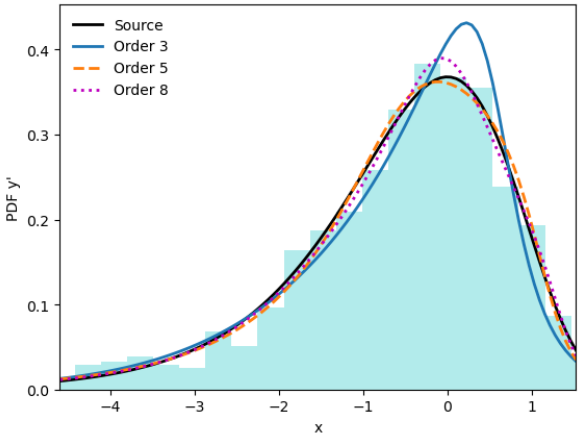

Figure 3. Metalog Distribution Fits of Different Order

The metalog distribution is only defined for univariate distributions. To extend the application to a multivariate dataset X in $\mathbb{R}^D$ we create D metalog distributions over the dataset's marginal distributions and use a copula, C, to capture the correlation information between the D random variables as Keelin (2023) suggests. Copulas are functions that can generate structured multidimensional correlated uniform random samples, which we use as CDF samples. Figure 4 shows 2D gaussian examples, with different correlations on each plot. The left plot shows copulas for marginal distributions with different variances, while the right plot shows copulas for marginal distributions with the same variance.

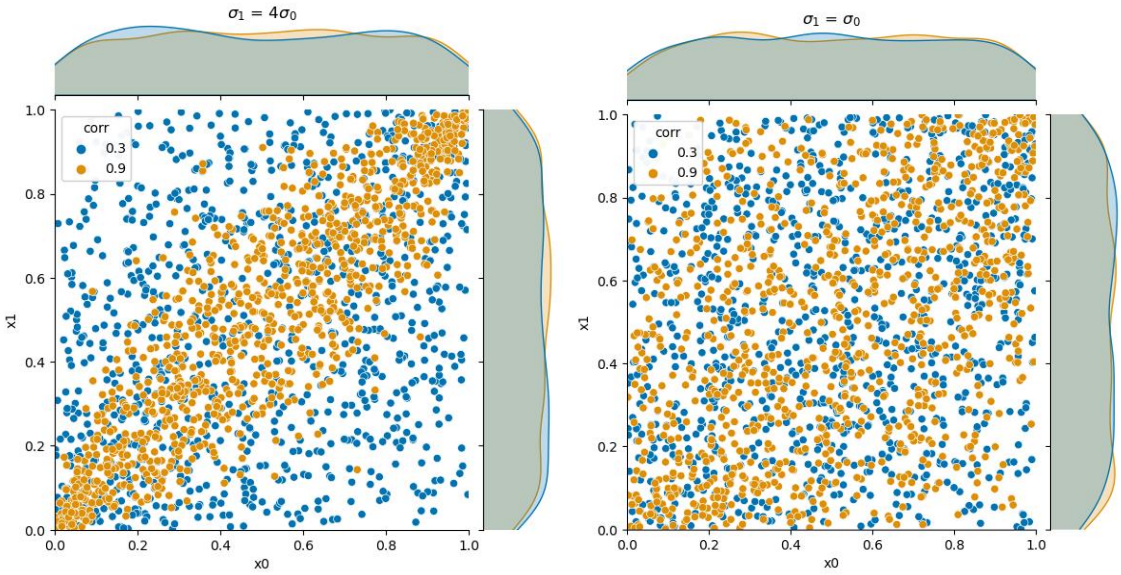

Figure 4. Gaussian Copula Samples

Once a multivariate metalog model consisting of D marginal metalogs and the gaussian copula, C, is constructed, the process of generating N' synthetic data samples occurs as follows:

1. Generate N' random samples from the D-variate copula. Each sample, y_i', represents a sample of the underlying joint CDF Y' for the metalog distributions.
2. For each y_i' sample, iterate over all $j \in D$ marginal metalog distributions, calculating the inverse CDF value, x_{ij}', for each y_{ij}' in the copula sample. The result will be a synthetic datapoint x_i' for each y_i' and in aggregate forms X'.

APPLICATION

To assess this method of generating synthetic data, we apply it to an exemplary healthcare dataset. For ease of access, we chose to rely on a dataset generated by an open-source patient population simulation (Walonoski et al., 2017) instead of real patient data. This realistic, simulated data does not require protections as real patient data would, which makes it easy to work with and share for this forum. For this example, we use the COVID-19 10K dataset.

There are several tables provided in the COVID-19 dataset, including patient information, logged conditions, observations, care plans, and encounters. For the following examples, we are interested in modeling populations partitioned on COVID-19 survival. We first combine information in care plans and observations to identify patient IDs that tested negative or completed isolation - these are the survivors. We cross-reference these with patient information to associate survivorship status to each patient. In this dataset, there are $N_S = 8473$ survivors and $N_F = 347$ fatalities.

Healthcare Costs

As an initial test we model healthcare expenses and healthcare coverage as the features of interest from the dataset. Figure 5 qualitatively compares the CDFs of the original data and the synthetic data for each feature and for each class of patient, i.e., fatalities and survivors. Each of the feature CDFs are virtually indistinguishable, indicating good agreement between the original and synthetic data.

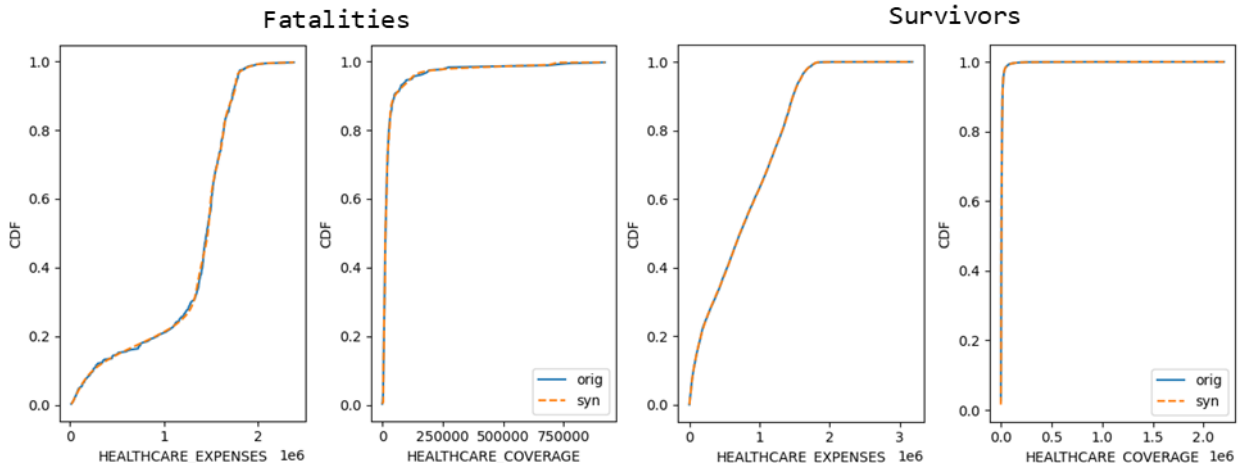

Figure 5. Patient Healthcare CDFs – (Orig)inal vs (Syn)thetic

In this work, we imposed a hard threshold of m_{max}=18, meaning that for each marginal metalog model $m \leq m_{max}$ which appeared to yield good results. Advanced methods for preventing overfitting are out of scope for this work. Table 1 contains the KS distances for each CDF plot, which is the maximum difference between the CDFs at any measured point. The largest deviation between the original and synthetic dataset CDFs corresponds to healthcare coverage of non-survivors at 0.061, which can be interpreted as a ~6% difference.

Table 1. Patient Healthcare KS Distances

	Patient Survivors	Patient Fatalities
Healthcare Expenses	0.005	0.030
Healthcare Coverage	0.020	0.061

Looking at the 2D data scatterplots will also help assess the goodness of fit. Figure 6 shows scatterplots for the features for each class of patient.

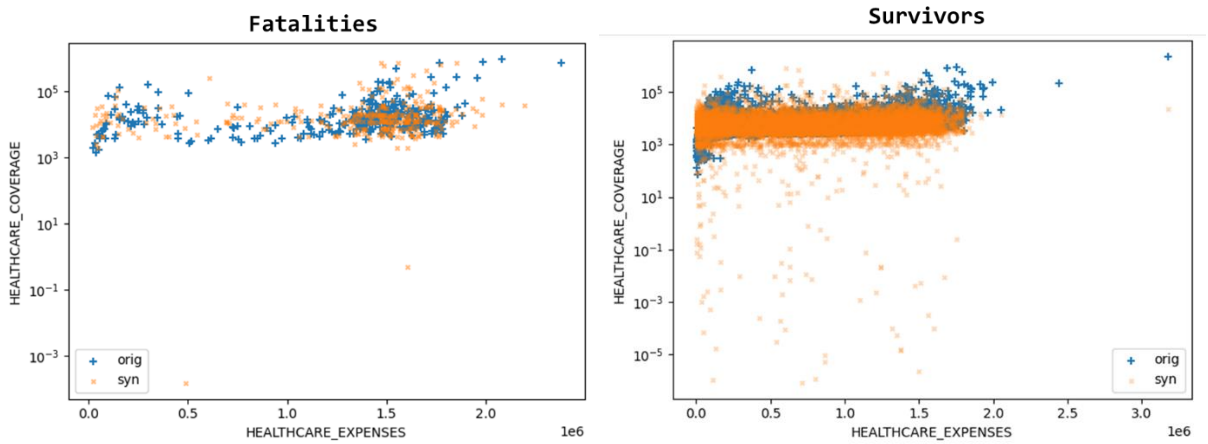

Figure 6. Patient Healthcare Scatterplots – (Orig)inal vs (Syn)thetic

For each plot, an equal number of samples were drawn for each class of patient (N_S for the survivor class and N_F for the fatality class). Healthcare coverage is plotted on a log scale for clarity. We note that while the main distribution matches well, there are a number of synthetic artifacts at low values of healthcare coverage. This is in part because in the original data, there are entries with 0 healthcare coverage which do not show up on the log scale. As a test, patients with 0 healthcare coverage were excluded from the dataset and the metalog modeling and SDG were repeated. The resulting scatterplots are shown in Figure 7.

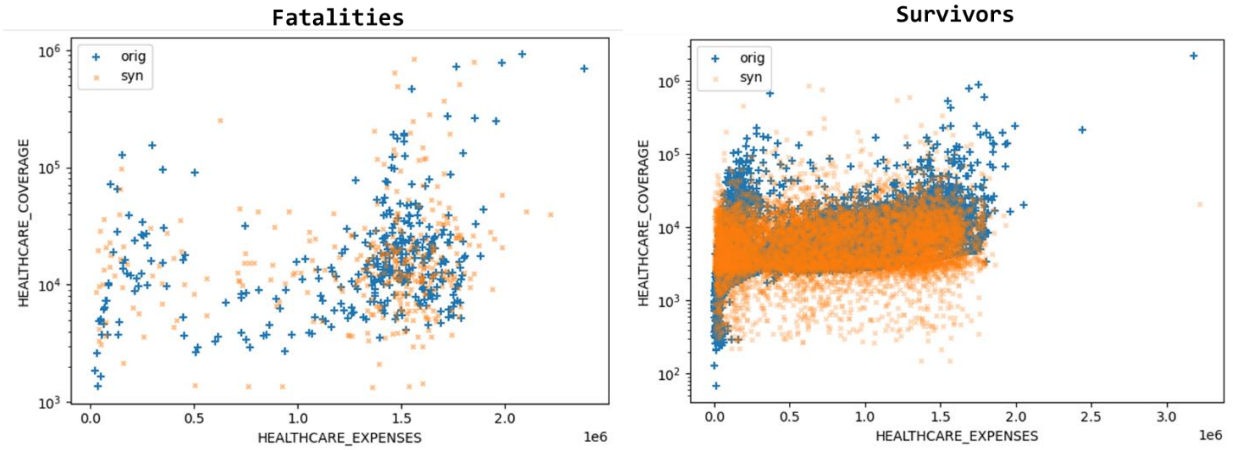

Figure 7. Nonzero Patient Healthcare Scatterplots – (Orig)inal vs (Syn)thetic

The synthetic samples for the patient data now match more closely to the original dataset, with the KS distance for healthcare coverage shrinking by a factor of three from 0.020 to 0.006. The distribution shape of the survivors' synthetic samples at low healthcare coverage could still be improved. It's possible the metalog model may be overfitting in this case. Another way to potentially improve the distribution shape is by changing the copula, for

example using a Clayton copula instead of a gaussian copula. Different copulas have different correlation structures and so can influence the joint distribution shape.

To statistically test the similarity between the original data distribution and the synthetic data distribution, we utilize the maximum mean discrepancy (MMD) test (Gretton et al., 2012). This is a distribution-free multivariate probability density function (PDF) comparison. It is agnostic to the distribution type, i.e. does not assume a normal distribution. We follow the procedure laid out in Heine et al. (2023); specifically, we use the bound on the unbiased statistic, MMD_u^2, and use the median heuristic to determine the kernel parameter for the test sample. For these tests, the datasets that include patients with 0 healthcare coverage are used. We conduct the test 1000 times, drawing random synthetic samples for each test. We use a hypothesis test level alpha of 0.05. For both patient populations (fatalities and survivors), the test fails to distinguish between the original and synthetic samples across all 1000 iterations.

Lab Values

As a more complex example, we look at modeling certain lab results from simulated patients. As before we partition the COVID-19 patients on survivorship status. The six features in this case are lab values for: d-dimer, serum ferritin (Fe), high sensitivity cardiac troponin (cTn) I, lymphocytes, and lactate dehydrogenase (LDH). Table 2 contains the KS distances for the metalog model fits, with the largest deviation being 0.018 (1.8% difference in the CDF) for Fe in the fatality group. Notably, lymphocytes and d-dimer for survivors have exquisite distribution fits.

Table 2. Patient Lab Values KS Distances

	Patient Survivors	Patient Fatalities
D-dimer	**3e-13**	0.008
Fe	0.007	0.018
cTn I	0.007	0.012
Lymph.	**2e-13**	**1e-13**
LDH	0.007	0.016

The corresponding CDFs are plotted in Figure 8 below and demonstrate a qualitatively good fit for all the features.

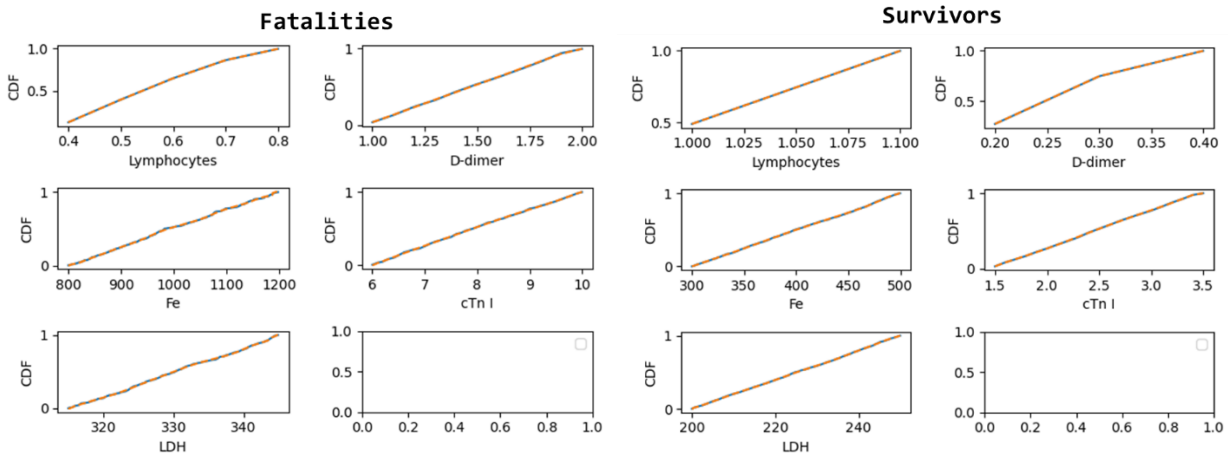

Figure 8. Patient Lab Values CDFs – (Orig)inal vs (Syn)thetic

We also examine scatterplots of pairwise features for the fatality patient class in Figure 9. Scatterplots are symmetric across the diagonal. Kernel density estimates of the underlying marginal distributions for singular features are plotted along the diagonal.

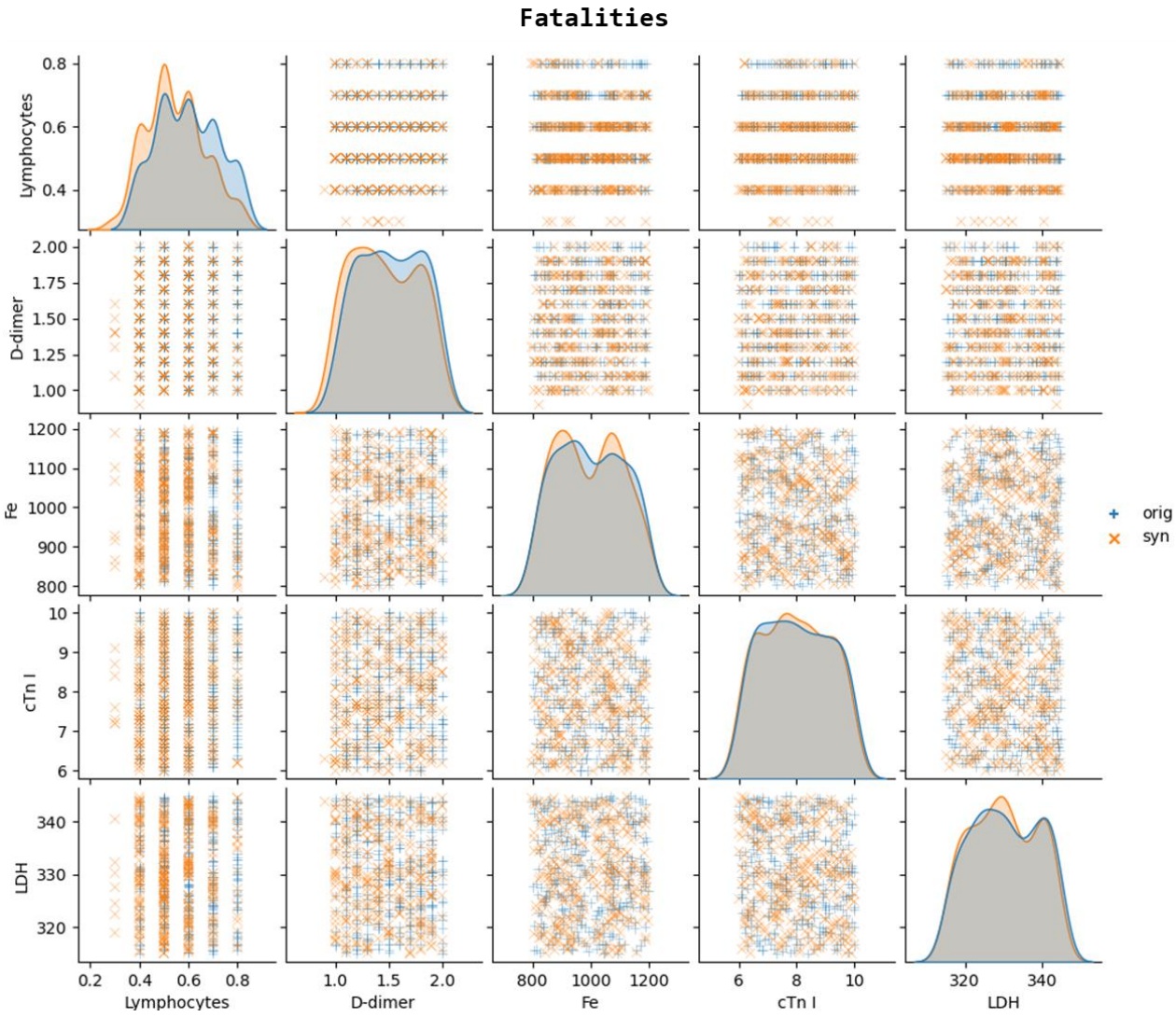

Figure 9. Patient Fatality Lab Values Scatterplots – (Orig)inal vs (Syn)thetic

The marginal distributions demonstrate that the original and synthetic dataset match very closely in most features. There is a slight shift in the lymphocytes distribution – perhaps due to a discretization artifact as described later. The cross correlations between the features in this case are quite low (largest absolute value is 0.1), so there is very little structure in the scatterplots. We note that some of the features (e.g., lymphocytes and d-dimer) had values with limited precision (~0.1) in the original data, which show up as discrete lines in some of the plots. We construct the same plots for the patient survivor class in Figure 10 below.

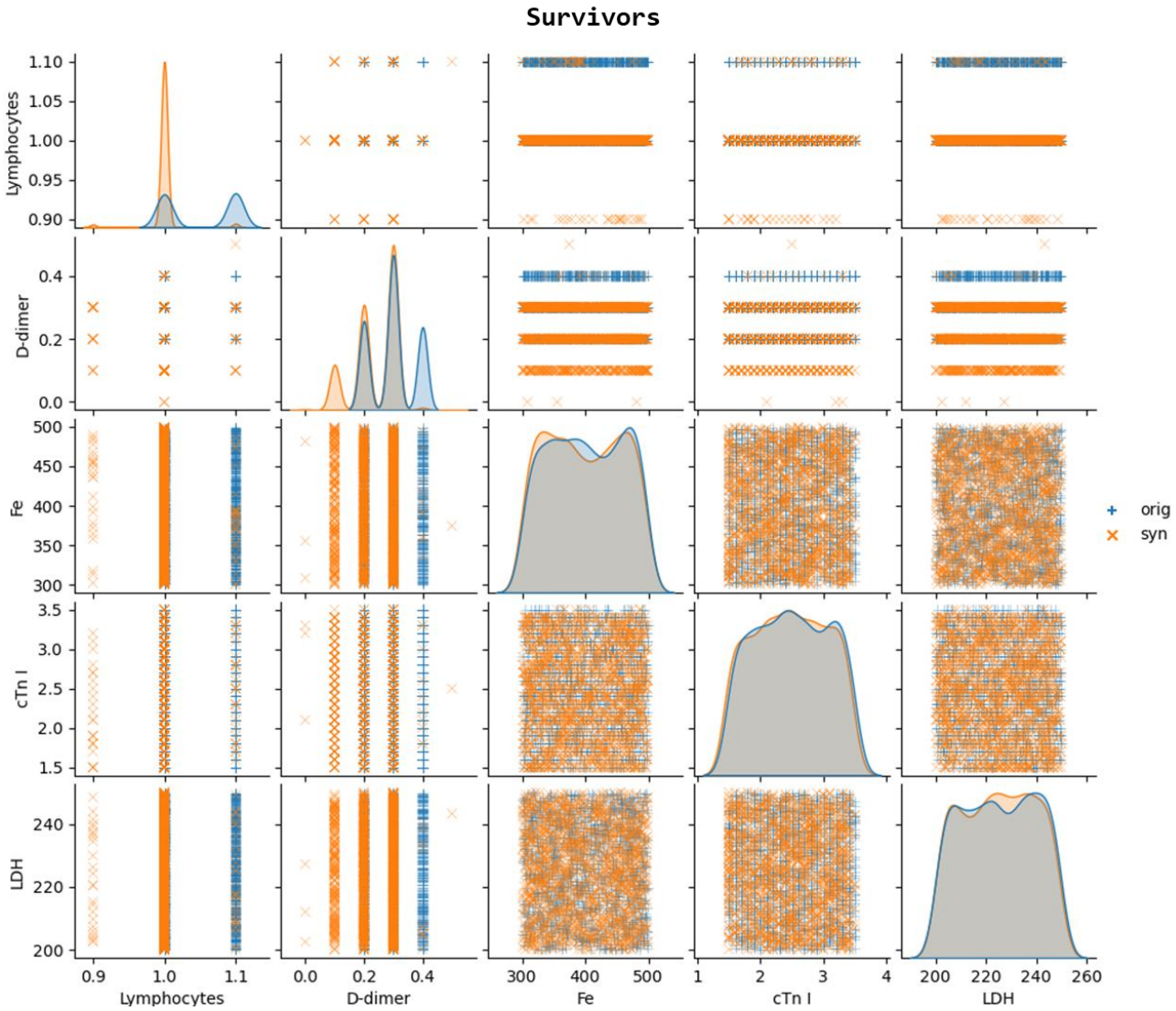

Figure 10. Patient Survivor Lab Values Scatterplots – (Orig)inal vs (Syn)thetic

As in the patient fatality class, the cross-correlations of features for patient survivors are very low, so little structure is seen in the scatterplots. In this case, while the Fe, cTn I, and LDH features show good agreement in the marginal distribution estimates, we observe some discrepancy for the lymphocytes and d-dimer features. These discrepancies can be attributed to the automated discretization process we implemented for sampling the metalog model. The underlying metalog model in all cases is a continuous distribution. If the distribution is skewed relative to the discrete levels, then discretization of the continuous samples can lead to the rounding artifacts observed. Further work is needed to refine the discretization method for the metalog models.

As in the healthcare expenses example, we conduct MMD tests for distribution similarity for both patient classes. We find again in these cases that the test is not able to distinguish between the original and synthetic datasets over 1000 trials.

CONCLUSION

This work demonstrates a novel modeling framework for synthetic data generation (SDG) using the metalog distribution as the basis. SDG using metalog distributions can be theoretically applied to diverse datasets such as pilot

training data, service personnel performance measures, and sensitive PII and PHI, though further study is needed to vet the suitability of this methodology. Data anonymization is important in the commercial healthcare sector for protecting sensitive information in medical research. Anonymization also has applications within the DoD for all human research initiatives, such as those undertaken by the Army Medical Research and Development Command (AMRDC). Data anonymization can address the pervasive data silo problem by enabling the sharing of data without fear of revealing sensitive attributes for individual data entries.

Compared to popular generative AI approaches to SDG, the metalog framework takes much less compute power and works on much smaller datasets than those required by traditional AI methods. This metalog approach is a type of distribution modeling, but unlike other distribution models, the metalog is highly flexible and computes its parameters directly from the CDF of the data, making it easy to use. This is advantageous because there is not a need to test many different distribution families (e.g., Gaussian, logistic, gamma, etc.). However, further work is needed to refine this multivariate metalog modeling framework. To better model discrete variables, a robust extension to this framework is needed. Integrating different copulas and algorithmic techniques to avoid overfitting into the modeling framework should lead to further improvements in the models. Adding more statistical tests for validation will also make it easier for a user to identify how well the synthetic data matches the original dataset.

This metalog modeling framework, tied together with a statistical copula for multivariate data, was applied to a representative, simulated medical dataset of COVID-19 patients. The objective was to create a model for the patient data and generate a synthetic dataset that captures the characteristics of the original data and guarantees patient privacy. We applied this approach to patient healthcare expenses and to clinical lab values successfully. The generated synthetic data is in good agreement with the original patient dataset as assessed in scatterplots, distance metrics in the data CDFs, and with statistical hypothesis tests.

REFERENCES

Amodei, D., Hernandez, D., Sastry, G., Clark, J., Brockman, G., & Sutskever, I. (2018). *AlexNet to AlphaGo Zero: 300,000x increase in compute*. AI and compute. OpenAI. Retrieved 2024, from https://openai.com/index/ai-and-compute/.

Carlini, N., Hayes, J., & Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., & Wallace, E. (2023). Extracting Training Data from Diffusion Models. In *Proceedings of the 32nd USENIX Security Symposium* (pp. 5253–5270). Anaheim; USENIX. Retrieved 2024, from https://www.usenix.org/conference/usenixsecurity23/presentation/carlini.

Chevrier, R., Foufi, V., Gaudet-Blavignac, C., Robert, A., & Lovis, C. (2019). Use and Understanding of Anonymization and De-Identification in the Biomedical Literature: Scoping Review. *Journal of Medical Internet Research*, *21*(5), e13484. https://doi.org/10.2196/13484

Department of Defense, DoD Data, Analytics, and Artificial Intelligence Adoption Strategy: Accelerating Decision Advantage (2023). Washington, D.C.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research*, *13*, 723–773. https://dl.acm.org/doi/10.5555/2188385.2188410

Health Insurance Portability and Accountability Act, (1996).

Heine, J., Fowler, E. E., Berglund, A., Schell, M. J., & Eschrich, S. (2023). Techniques to Produce and Evaluate Realistic Multivariate Synthetic Data. *Scientific Reports*, *13*(1). https://doi.org/10.1038/s41598-023-38832-0

Northwestern University Institutional Review Board Office. (2020). *HIPAA, PHI, & PII*. https://irb.northwestern.edu/resources-guidance/consent-templates-hipaa-requirements/consent-hipaa/hipaa-phi-pii.html

Keelin, T. W. (2016). The Metalog Distributions. *Decision Analysis*, *13*(4), 243–277. https://doi.org/https://doi.org/10.1287/deca.2016.0338

Keelin, T. W. (2023). *The Multivariate Metalog Distributions with Application to Strategic Decision-Making in Golf*. https://doi.org/10.31219/osf.io/cdmrv

Office of the Under Secretary of Defense for Research and Engineering. (2020, April). Protection of Human Subjects and Adherence to Ethical Standards in DoD-Conducted and -Supported Research. (DoDI 3216.02). Department of Defense.

Privacy Act, 5 U.S.C. § 552a. (1974).

Protection of Human Subjects, 32 C.F.R. pt.219 (2018).

Slabodkin, G. (2021, September 28). *Data silos holding back healthcare breakthroughs, outcomes*. Health Data Management. https://www.healthdatamanagement.com/articles/data-silos-holding-back-healthcare-breakthroughs-outcomes

Sweeney, L. (2000). *Simple Demographics Often Identify People Uniquely*. Carnegie Mellon University. https://dataprivacylab.org/projects/identifiability/paper1.pdf

Talend. (2020). *What are Data Silos?*. Data silos, why they're a problem, & how to fix it. https://www.talend.com/resources/what-are-data-silos/

Tidwell, N. (2024, February 16). *How advanced cybersecurity can solve data silos*. Sertainty. https://www.sertainty.com/blog/how-advanced-cybersecurity-can-solve-data-silos/

Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., Duffett, C., Dube, K., Gallagher, T., & McLachlan, S. (2017). Synthea: An approach, method, and software mechanism for generating synthetic patients and the Synthetic Electronic Health Care Record. *Journal of the American Medical Informatics Association*, *25*(3), 230–238. https://doi.org/10.1093/jamia/ocx079

World Business Research Insights, Financial Information Management US, & InterSystems. (2022). *Empowering Line of Business Users Through Data Democratization: How financial firms can use enterprise data to drive actionable insights across their business teams*. https://assets.intersystems.com/75/89/5c5130e54147b61aaf2d0c980195/empowering-line-of-business-data-democratization.pdf

Zuo, Z., Watson, M., Budgen, D., Hall, R., Kennelly, C., & Al Moubayed, N. (2021). Data Anonymization for Pervasive Health Care: Systematic Literature Mapping Study. *JMIR Medical Informatics*, *9*(10), e29871. https://doi.org/10.2196/29871