The Review of Contemporary Scientific and Academic Studies
An International Multidisciplinary Online Journal
www.thercsas.com

ISSN: 2583-1380 Vol. 3 | Issue No. 9 | September 2023 Impact Factor: 4.736 (SJIF)

Exploring the Untapped Potential of Synthetic Data: A Comprehensive Review

Shashank Agarwal (shashank.agarwal@cvshealth.com), Senior Decision Scientist, CVS Health, Illinois, USA
Siddharth Sharma (siddharth.x3.sharma@jpmchase.com), Vice President, JP Morgan Chase & Co, New Jersey, USA
Sachin Parate (sparate@twilio.com), Principal Product Manager, Twilio, New Jersey, USA

Crossref/DOI: https://doi.org/10.55454/rcsas.3.09.2023.002

Abstract: *Synthetic data generation (SDG) is known as the method of training a model with machine learning techniques to recognize patterns in a real dataset. The trained model can then be used to produce fresh, or synthetic data. Synthetic data generation stands as a pivotal solution at the intersection of data privacy and medical research. This review paper covers in detail synthetic data generation approaches and their role in the field of healthcare. The article first describes the three major types of synthetic data generation approaches. The paper then continues to explore the concept of language modeling that is adapted to AI-generated tools, uses of synthetic data, and its applications in healthcare. Moreover, the paper finally highlights the potential challenges that may be faced by the healthcare industry associated with the adoption of synthetic data. It was concluded that challenges like complexity, bias, and regulation persist, but SDG remains promising for data-driven healthcare. It bridges technology, ethics, and innovation for transformative insights.*

Keywords: AI-Generated Data, Healthcare Industry, Medical Research, SDG, Word Embedding

Article History: **Received: 15 Sept- 2023;** **Accepted: 26 Sept- 2023;** **Published/Available Online: 30 Sept- 2023;**

Introduction

Synthetic data is described as data that has been generated artificially based on a use-case relevant context and that accurately captures the appropriate meaning for statistical evaluation in the intended context (which includes training and analysis by AI). Synthetic data generation (SDG) is known as the method of training a model with machine learning techniques to recognize patterns in a real dataset. The trained model can then be used to produce fresh or synthetic data. If appropriately generated, the synthetic data has the potential to offer privacy-preserving qualities because it does not map directly to the original data or to actual patients.

A vast collection of patient records, images, laboratory findings, measurements done on physiological environments along with a myriad of other artifacts are produced annually, contributing to the exponential growth in the amount of data generated in the healthcare industry. In the future, data-driven learning techniques will probably rank among the most important tools in scientific and medical research.

Algorithms based on artificial intelligence (AI) are being more widely used to facilitate the automation of data analysis and to enhance the effectiveness and efficiency of decision-making. Applications of artificial intelligence (AI) in healthcare have been studied in a number of areas, such as the detection of pathology involved in medical imaging (such as radiography, magnetic resonance imaging (MRI), computed tomography (CT), diagnosis of cardiovascular disorders (e.g., electrocardiograms, or ECGs), assessment and forecasting of health outcomes utilizing electronic healthcare records (EHRs), as well as data mining from the medical literature.

Data from electronic health records collected from entire populations can be used to test novel ideas, develop and evaluate various methodological and statistical strategies, and aid in producing real-world evidence. It is also beneficial to carry out secondary analyses of original research data, such as those from clinical trials, for conducting meta-analyses of participant-specific data. However, obtaining these data is difficult due to a number of intricate privacy restrictions.

Highly sensitive information might be found in digital health records or data from clinical trials, and accessing these datasets can be a costly and extensive process. The main hurdle in accessing data for healthcare and research purposes is data privacy regulations. Using generated or synthetic data that offers a realistic depiction of the source of the original data is one way around this problem. Synthetic data mimic the look of the actual data source but do not include any data of real individuals. Some statistical characteristics of the original (real) data source, such as the proportions of categorical data, continuous data distributions, the correlations among variables, as well as other model parameters, can be attempted to be preserved in synthetic data.

Methodology

A narrative review was conducted through a deep existing literature search using Google Scholar. The keywords used for the identification of the articles were “synthetic data”, “healthcare”, “AI-generated data”, “medical research”, and “AI models”. A total of 40 full-access articles were selected from a vast literature search. Our review specifically focused exclusively on the development of synthetic data, its use, potential role, and limitations in the field of healthcare and medicine, while the studies that focused on other fields were excluded from our review. Moreover, we kept our search generalized for synthetic data generation approaches and models as our study does not focus on any specific AI-generated approach/tool. Majority of the articles that were included in our work were the most recent ones (after 2010). However, 5 studies were from before 2010 (one from 1968, one from 2003,one from 2009 and two from 2011), considering the most pioneer and fundamental studies worth including in our review. The review was then broadly divided into sub-headings covering topics such as “Generation of synthetic data”, “Evaluation of synthetic data”, “Language models”, “Uses of synthetic data”, “Applications in healthcare”, and “Limitations”.

Synthetic Data Generation

Bayesian Networks (Probabilistic Model)

Bayesian networks (BN) are a form of probabilistic, graphical model in which each node of the graph reflects a random variable, and the edges present between the nodes reflect probabilistic dependencies among the respective random variables. Through the use of a Bayesian network, the structure of the graph along with the conditional probability distributions is calculated from the original data for the purpose of generating synthetic data. In Bayesian network, the whole joint distribution is simplified as follows

$$\rho(x) = 11\prod_{v\epsilon V}\rho\left(x_v \,|x_{pa(v)}\right) \qquad (1)$$

Where,

V= set of random variables which represents categorical variables

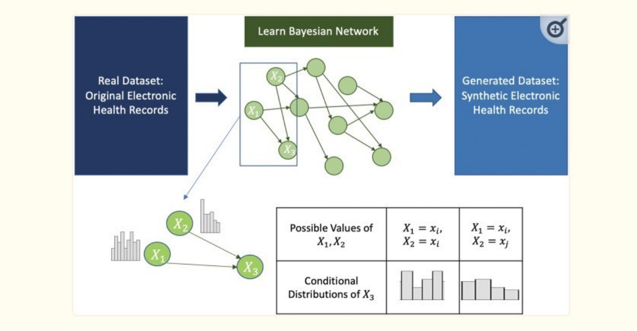
$x_{pa(v)}$= subset of parent variables of “*v*”

Two processes are involved in the learning process: (i) learning an acyclic graph with a directed structure from the data that expresses every pairwise conditional (in)dependence across the variables, and (ii) predicting conditional probability tables for all variables using maximum likelihood. In the first phase, we employ the Chow-Liu tree approach, which looks for a first-order approximation of the dependency tree for the actual whole joint probability distribution. The approximation provided by the Chow-Liu method fails to reflect higher-order dependencies. However, it has been demonstrated to work well for a variety of practical problems. The conditional dependence between the variables is encoded in the structure of the graph derived from the real data. Additionally, the derived graph also offers a visual depiction of the relationships among the variables. By selecting samples from the determined Bayesian network, synthetic data can be produced. The main pros and cons of this approach include:

- BN is technically effective and goes well with the dataset's dimensionality.
- The causal associations among the variables can also be investigated using the directed acyclic graph.
- Although, factorization of full joint distributions, as shown in Equation 1, is sufficiently general to encompass any possible dependence structure, the practical application shows that simplifying assumptions about the graphical structure are made to facilitate model inference. These presumptions might not adequately depict higher-order dependencies.

The BN approach is shown in Figure 1 where X1 and X2 are nodes that do not contain incoming edges, thus, their values must first be independently simulated. The value for variable X3 is then created from distributions that depend on the values of variables X1 and X2

Figure 1: Process of Bayesian Network Approach



Multiple Imputations (Classification-based Imputation Models)

In regard to the process of creating synthetic data, methods based on multiple imputations have become increasingly popular, particularly in applications where a portion of the data is deemed sensitive. Among all the available imputation techniques, the MICE or "Multivariate Imputation by Chained Equations" has recently emerged as a systematic approach for masking private information in datasets with safety restrictions. The main idea is to present sensitive information as missing data. The "missing" data is then imputed using values provided by randomly sampling models that are trained using the non-sensitive variables.

The first variable is randomly selected from the empirical distribution during the sampling phase, and then the subsequent variables are drawn at random from the deduced conditional distributions in accordance with the topological order. Although generalized linear frameworks are particularly common for modeling conditional distributions, additional non-linear techniques like neural nets and random forests can be simply included into this framework. The simplified complete joint probability distribution for the MICE variation is as follows:

$$\rho(x) - \prod_{v \in V} \rho\left(x_v \,|\mathrm{X}_{;(v)}\right) \qquad (2)$$

Where,

V= set of randomized variables depicting the variables that need to be created

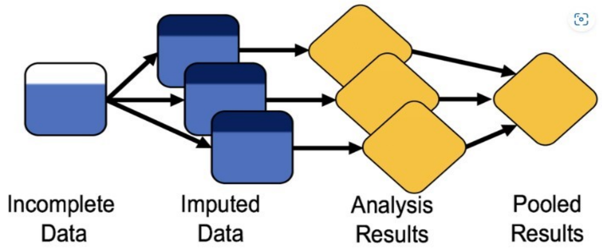
$p(x_v|\mathbf{x}_{:v})$= conditional probability distribution of v-th randomized variable

It is apparent that the definition of the topological order is very important for the construction of the model. Sorting the variables by numbering the levels, either in order of ascending or descending, is a typical common approach. Figure 2 shows the process framework of a multiple imputation approach. In the first stage, for constructing M complete data sets (M ¼ 3 shown in figure 2), missing data (represented in white color) is imputed (represented in dark blue color) in the first stage. Then, each whole set of imputed data is examined making use of the conventional techniques like linear regression. Finally, Rubin's rules are used to pool the results

Next, we list the primary pros and cons of the MICE approach:

- MICE scales to very big datasets, both in terms of the quantity of variables and samples and is computationally fast.
- By appropriately selecting a Softmax or Gaussian modeling for conditional probability distribution on a particular variable, it is capable of dealing with continuous and categorical data easily.
- Although this approach is probabilistic, there is no assurance that the generative model it produces is an accurate representation of the data's underlying joint distribution.
- MICE heavily depend on the topological order of the produced acyclic graph as well as the fact that how much the model is flexible for the conditional probability distributions.

Figure 2: Diagrammatic Representation Showing the Process of Multiple Imputation Approach



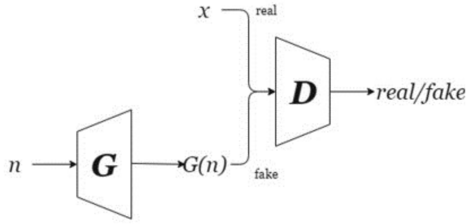
Generative Adversarial Networks or (GANs)

Recent research has demonstrated that generative adversarial networks or (GANs) are very effective at generating complicated synthetic data, like text and images. This approach involves training two neural networks in a competitive way, as shown in Figure 3. The first network attempts to produce realistic synthetic data and the second network tries to distinguish between real and synthetic data that is produced by the first network. A normal convolutional classifier is responsible for down sampling an example to generate a probability, while the generator is responsible for taking a vector of the random noise and up sampling it. The first one throws away data via down sampling methods such as max pooling, while the second one creates new data. During the process of training, each network stimulates the other network to perform better. A well-known shortcoming of GANs is that they cannot be used directly for producing categorical synthetic datasets since gradients, based on implicit categorical variables needed for backpropagation training, cannot be computed. Recent attempts, such as medGAN have used autoencoders for converting categorical data into a continuous domain because clinical patient data are frequently mostly categorical. After this transformation, GANs can be used to produce synthetic health records or EHR i.e., electronic health records. MedGAN, however, only works with data that is binary and countable; it does not work with data that is multi-categorical.

Next, we list the main pros and cons of this approach:

- MC-MedGAN being a generative technique does not necessitate rigid probabilistic model assumptions, in contrast to other techniques such as CLGP, POM, and BN. As a result, it is more adaptable than CLGP, POM and BN.
- It is simple to adapt GANs-based models to handle mixed data types, such as categorical as well as continuous variables.
- The MC-MedGAN model is deep and has a lot of parameters. It takes time and effort to choose various tuning parameters (or hyper-parameters) correctly.
- Due to the instability of the underlying problem regarding min-max optimization, GANs are renowned for being challenging to train. But newly proposed GAN variants, like Wasserstein GANs, and their modifications, have greatly reduced the stability issue of training GANs.

Figure 3: Architectural Representation of GAN Network



Language Models

Basically, AI language models create language by building sentences out of words they have learned from a learned lexicon. Their word choices are influenced by the words' probability distributions discovered through the examination of massive amounts of text, often known as training data. For instance, to determine their

https://doi.org/10.55454/rcsas.3.09.2023.002

relative probability distributions, we can count the number of times the words "book" and "bag" appear after the sentence "He dropped the..." (for example, 45% for "bag" and 10% for "book")

A concept known as "word embedding" [e.g., BERT] is employed to capture this combination of probability distributions of words and textual contexts. In order to forecast word probabilities, advanced neural language models (like GPT-3) learn a multitude of parameters using vast amounts of unmoderated data from the Internet . Generic generators for these models are freely accessible online through libraries like Huggingface

Uses of Synthetic Data

Creating code or performing the generation of preliminary hypotheses and testing prior to deployment in real datasets are two important uses for synthetic data. Before accessing original data, researchers might create and verify procedures for a certain purpose. Since data access applications are able to be processed simultaneously or while waiting for access to data to be granted, this technique saves time. Synthetic data has a prominent contribution to preserving the privacy and confidentiality of patient data as it takes less time for researchers to access sensitive patient data.

As it takes less time for researchers to access confidential patient data, synthetic data also contribute to privacy preservation. Due to the fact synthetic datasets may be easily shared with other investigators or outside parties to validate models and analytical approaches, they can also be employed to increase the reproducibility of research.

Synthetic data may also be employed to speed up methodological advancements in the field of healthcare. It helps in training people and improving their capacity building in techniques for managing high-dimensional and complex medical data. Moreover, synthetic data can also be a solution for scientists and researchers who are busy synthesizing evidence for clinical studies. For instance, researchers conducting a meta-analysis of individual patient data utilizing adequate statistics from entire data and who may want to merge data from clinical trials that yield individual patient data in addition to from those trials that do not. Additionally, synthetic data can also be used to simulate studies for calculating sample sizes for meta-analysis of individual patient data to signify previous knowledge in the available information.

Applications in Healthcare Industry

Synthetic Structured Data in Clinical studies

Demographics of the patient, previous medical history, drugs, disease diagnosis, any existing allergies, measurements of body vitals, the balance of fluids in the body, lab findings, microbiological findings and data, pathological information and certain information linked to, various procedures, therapies and photographic imaging are all forms examples of structured, also known as tabular data, in an electronic health record. Through various AI softwares, synthetic structured data can be generated from the original data of EHR which can further be used for observational clinical investigations.

Synthetic Natural Language Data in Disease Diagnosis

An electronic health record consists of large amounts of natural language text/data in the form of notes that provide a simplified narrative, highlighting the most important aspects of the patient's course and further treatment protocol. This is specifically essential for records of mental health that heavily depend on unstructured natural language data/text more than the structured data. Synthetic natural language text/data as training data might provide classification results which are comparable to the real data results. An example of this is a model that is trained with discharge summaries of the patients, generated synthetically, and can predict accurate diagnoses and phenotypes of patients with mental disorders

Processing of Physiological Signals

Any unstructured, continuous stream of data, including waveforms, a 12-lead ECG, along with ECG telemetry, linked to hemodynamics, is referred to as a physiological measurement. One of the areas in which synthetic data of patients have been applied extensively is in the physiological signals processing. Particularly, researchers have analyzed the generation of synthetic cardiac signals, also known as ECG's, photoplethysmograms and phonocardiograms. This in turn allows for the diagnosis of health conditions such as bradycardia, ventricular flutter, tachycardia, atrial fibrillation, etc.

Synthetic Image Generation

Improvement in categorizing hepatic lesions through CT scans, classifying brain cancers in MRI data, identifying COVID-19 via chest radiographs, classifying cancerous cells in histopathological slides, and categorizing dermatological lesions in photographs, has benefited from the use of synthetic image generation. More specifically, physics-based data augmentations are utilized to boost synthetic data using hybrid approaches.

Conclusion

The realm of synthetic data generation (SDG) holds immense promise in reshaping the landscape of healthcare and medical research. With the exponential growth of data in the healthcare industry and the increasing use of artificial intelligence (AI) techniques, the need for innovative data solutions has become more pressing. Synthetic data offers a path forward by enabling researchers and practitioners to harness the power of data-driven approaches without compromising patient privacy or facing complex data access challenges. Through physical and statistical models, synthetic data can emulate the characteristics of real-world data sources, serving as a bridge between privacy concerns and research needs. The ability to create synthetic data that closely mimics the original data's distribution holds tremendous potential in various healthcare applications, such as medical imaging, diagnostics, health outcome prediction, and literature mining. However, challenges persist, ranging from the intricacies of simulating real-world complexity to addressing unknown variables, biases, and lack of regulatory frameworks. Despite these obstacles, the progress in SDG underscores the importance of finding innovative ways to navigate data privacy, accessibility, and utilization concerns in the healthcare domain.

References

Azizi, Z., Zheng, C., Mosquera, L., Pilote, L., & El Emam, K. (2021). Can synthetic data be a proxy for real clinical trial data? A validation study. *BMJ open*, *11*(4), e043497.

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work?. *International journal of methods in psychiatric research*, *20*(1), 40-49.

Benaim, A. R., Almog, R., Gorelik, Y., Hochberg, I., Nassar, L., Mashiach, T., ... & Beyar, R. (2020). Analyzing medical research results based on synthetic data and their relation to real data results: systematic comparison from five observational studies. *JMIR medical informatics*, *8*(2), e16492.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).

Branciamore, S., Gogoshin, G., Di Giulio, M., & Rodin, A. S. (2018). Intrinsic properties of tRNA molecules as deciphered via Bayesian network and distribution divergence analysis. *Life*, *8*(1), 5

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Camino, R., Hammerschmidt, C., & State, R. (2018). Generating multi-categorical samples with generative adversarial networks. *arXiv preprint arXiv:1807.01202*.

Chen, D., Yu, N., Zhang, Y., & Fritz, M. (2020, October). Gan-leaks: A taxonomy of membership inference attacks against generative models. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security* (pp. 343-362).

Chen, J., Chun, D., Patel, M., Chiang, E., & James, J. (2019). The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures. *BMC medical informatics and decision making*, *19*(1), 1-9.

Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., & Sun, J. (2017, November). Generating multi-label discrete patient records using generative adversarial networks. In *Machine learning for healthcare conference* (pp. 286-305). PMLR.

Chow, C. K. C. N., & Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, *14*(3), 462-467.

Dattani, N., Hardelid, P., Davey, J., & Gilbert, R. (2013). Accessing electronic administrative health data for research takes time. *Archives of disease in childhood*, *98*(5), 391-392.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Drechsler, J. (2011). *Synthetic datasets for statistical disclosure control: theory and implementation* (Vol. 201). Springer Science & Business Media.

Ensor, J., Burke, D. L., Snell, K. I., Hemming, K., & Riley, R. D. (2018). Simulation-based power calculations for planning a two-stage individual participant data meta-analysis. *BMC medical research methodology*, *18*, 1-16.

Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., & Ranganath, R. (2020). A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings*, *2020*, 191.

Gogoshin, G., Branciamore, S., & Rodin, A. S. (2021). Synthetic data generation with probabilistic Bayesian Networks. *Mathematical biosciences and engineering: MBE*, *18*(6), 8603.

Grover, A., Song, J., Kapoor, A., Tran, K., Agarwal, A., Horvitz, E. J., & Ermon, S. (2019). Bias correction of learned generative models using likelihood-free importance weighting. *Advances in neural information processing systems*, *32*.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). Improved training of wasserstein gans. *Advances in neural information processing systems*, *30*.

Harutyunyan, H., Khachatrian, H., Kale, D. C., Ver Steeg, G., & Galstyan, A. (2019). Multitask learning and benchmarking with clinical time series data. *Scientific data*, *6*(1), 96.

Hinton, G. (2018). Deep learning—a technology with the potential to transform health care. *Jama*, *320*(11), 1101-1102.

Authors' Bio-Notes

Shashank Agarwal is a data science expert who has channeled his expertise into the healthcare space over the years. He has worked with several Fortune 500 healthcare companies, such as CVS Health, AbbVie, and IQVIA. His experience cuts across various areas in market access, brand analytics, predictive modeling, launch strategy, and multi-channel marketing. He has played a pivotal role in identifying new opportunities, fostering business growth, automating technical workflows, optimizing business processes, and leading multiple end-to-end data science implementations across his employers to generate substantial revenue. He is a highly sought-after industry judge, influencer, and technical writer and has multiple publications to his name. Additionally, he holds a Master of Science in Engineering Management from the prestigious Johns Hopkins University.

Siddharth Sharma is a Vice President for HR Digital Solutions at JP Morgan Chase & Co. With over a decade of experience in enterprise applications, he has led multiple end-to-end implementations for achieving significant cost reductions and business process optimization. He has played a key role in digital transformations for companies such as JP Morgan Chase, Securitas Security Services USA & Concentrix, with over half a million employees in over 60 countries worldwide and transformation projects budgeted up to $70 million.

Sachin Parate is a Principal Product Manager at Twilio where he leads Twilio's international tech expansion. Sachin has 13 years of progressive experience across Tech Product Management and Data Science. Sachin has built amazing products and great teams that he is extremely proud of, which include building and launching a multi-billion-dollar credit card product, building a successful start-up, being an executive sponsor for multiple start-ups, and leading direct and cross-functional teams. He also has an undergrad degree in technology from one of the top colleges in India known as the Indian Institute of Technology (IIT), Bombay.