Synthetic data generation for training healthcare NLP models without compromising privacy

Veerendra Nath Jasthi

veerendranathjasthi@gmail.com

Abstract:
The Natural Language Processing (NLP) models have been known to be highly promising in medical care, especially in clinical note summarization, prediction of diagnoses, as well as in classification of patient records. Nevertheless, medical text data is sensitive, which raises critical privacy issues and regulatory restrictions thus, hindering access to training data of high quality. Data generation via synthesis is an attractive alternative because it generates artificial dataset that replicates the statistical figures of real clinical narratives, but without ending up at identifying patient data. This article details the sophisticated approaches to the construction of synthetic medical data suitable in machine learning settings with NLP downstream tasks based on generative adversarial networks (GANs) and large language models (LLMs) and rule-based methods of anonymization augmentation. Various types of NLP models are trained on real and synthetic data and the researchers check the quality of their performance and find out how predictive and linguistically relevant privacy-preserving synthetic datasets can be. According to our findings, good-quality synthetic datasets can serve as a source of preserving privacy as well as training a reliable model that can be used to apply AI in medicine in a safer and scalable way.

Keywords: Synthetic data, Natural Language Processing, Healthcare, Privacy preservation, GANs, Medical text generation, Clinical NLP, Data anonymization.

I. INTRODUCTION

Natural Language Processing (NLP) has redefined numerous industries, and in health care, it is quickly emerging to be the key to unlocking, understanding, and acting on the voluminous amounts of unstructured clinical text data being produced in the industry daily. Clinical notes, pathology reports, discharge reports, and physician narratives are the most valuable sources to create predictive models, better the decision support systems, and effectively manage patients. Nevertheless, the relatively sensitive nature and privacy issues with healthcare data impose on the development of a data-sharing model and model construction severe challenges. The system has got harsh rules like the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR) that limit the uses, access and even release of any information that can identify the patient [1].

Over the past few years, annotated clinical datasets to support supervised NLP tasks, including named entity recognition, classification, summarization, and relation extraction, have been in high demand in comparison to the available amount of such data. Common methods used in the traditional approaches of privacy preservation are de-identification methods of redaction or replacement of the protected health information (PHI). Although this gives a layer of security, it is not easy to avoid re-identification as it is combined with external data sources [9]. In addition, de-identified data is also legally and ethically problematic to share and, therefore, has limited value in mass research and model training.

The solution to this dilemma is the generation of synthetic data. In contrast to de-identification, in which actual institutional records are edited, synthetic data is completely fictional and does not copy any actual data about patients but is meant to recreate their real-world data in essence, through the orchestration of their statistical and semantic properties. In NLP terms applied to healthcare, this implies the production of synthetic, linguistically realistic and task-relevant clinical narratives, and complete lack of real patient identifiers. The opportunities are immense synthetic data can democratize training corpus access, make research cooperative, and speed in a privacy-sensitive manner, the training of models [14].

However, the usefulness of synthetic data solely depends on the thin line between usefulness and privacy. A synthetic dataset should be realistic to teach models that can generalize to real-world tasks as well as abstract enough to make no single patient identity or condition be derivable [10]. This level of balance in any free-text creation is by its nature more complicated than in structured data because natural language has so much varied richness and the meaning of terms is context relative with many variations in language meaning and use.

There have been a number of approaches developed to deal with this challenge. Legacy systems based on rules are crafted using specific templates and medical vocabularies with sturdy privacy guarantees, yet they lack diversity and realism. GANs applied to text generation use Generative Adversarial Networkswhich are normally applied to image synthesis, to provide more variability, but suffer coherence issues with respect to language. Transformer models, GPT and BERT, in particular when finetuned on a related data set, have proven highly promising in a produce high-quality clinical text that reflects the style and terminology of the real records [13].

The present paper looks into the comparative exploration of these synthetic text generation methods into the specific NLP applications in health care. We look into the extent each can be used to train a model, what kind of trade-off there is in terms of quality and security and how well synthetic data can be used in named entity recognition (NER) and classification tasks versus real-world data. Moreover, to determine the level of protection of each generation mechanism, we use quantitative privacy evaluation methods, such as membership inference attacks.

This research hypothesis will be proven through the rigorous analysis and experimentation process, in an attempt to show that synthetic clinical text and, specifically, the one produced with advanced transformer models, is the viable and privacy-friendly alternative to real medical data when building effective NLP systems. To contribute to the trend of ethical, scalable, and accessible healthcare AI research, we intend to show that such can be done [12].

Novelty and Contribution

TVarious contributions of this research to privacy-preserving healthcare NLP are as follows:

- The first comprehensive comparison of methods of synthetic data generation based on text in healthcare NLP tasks: Synthetic data has been studied in structured data records, but this research represents one of the first comparative studies between rule-based, generative adversarial networks (GAN)-based and transformer-based approaches to generating synthetic clinical text. We not only rate each method based on its quality as a language model but also according to the utility of each one in training models on real-world NLP tasks [8].
- Comparison of the synthetic data and real data model performances: We systematically train and test the BiLSTM and transformer models (e.g., BioClinicalBERT) on synthetic and compare them to models trained on real clinical data. This directly answers the following question: Is it possible to replace the real data with synthetic data in healthcare NLP pipelines in a meaningful way?
- Provision of privacy assessment framework of synthetic clinical text: We provide quantitative privacy emulations, such as membership inference attacks, to discuss the ability of synthetic test cases to accidentally discharge actual patient data. This gives an empirical foundation of having faith on the privacy of generated datasets.
- Transformer-based training on realistic clinical note generation: As opposed to generic text generation problems, we train large language models on de-identified clinical notes in order to make sure that the medical context and vocabulary considerations make their way into the generated sentences. We add in a guided prompt engineering to our regulated prompt engineering to make the content synthesized clinically meaningful [7].
- Ethical and deployment implication discussion: In addition to technical conclusions, we provide a discussion of how this solution could make ethical sharing of data possible, lead to collaboration across

institutions, and lead to faster advances in medical NLP development without challenging trust or breaching the law.

Overall, our work demonstrates that a privacy-sensitive model of clinical text generation is feasible and offers convincing arguments that synthetic information can give a push in the form of effective, scalable, and safe NLP solutions in the medical field. It supplies a basis on which the future of synthetic data validation research, hybrid generative frameworks, and more general implementations of synthetic corpora in medical AI systems can be built on.

II. RELATED WORKS

In 2024 M. Goyal and Q. H. Mahmoud [2] introduced the synthetic data generation as a field has grown rapidly over the past years especially in the area of privacy-preserving machine learning. In healthcare, the information about the patients is very sensitive and is more regulated and in such cases, synthetic data can be used as a plausible alternative to real world datasets. A lot of initial attention in this field has concentrated on structured data, e.g., demographics, laboratory results, and billing codes. Synthetic records produced against structured healthcare databases have proven themselves capable of validating statistical models and population health trends modeling. Synthetic data based on text is however more challenging especially in regard to clinical narratives because of context and situation rich dial-driven medical language.

In 2024 C. Umesh et.al., M. Mahendra et.al., S. Bej et.al., O. Wolkenhauer et.al., and M. Wolfien et.al. [15] suggested the first artificial intelligence methods of producing synthetic clinical text were rule-based systems that involved the use of pre-designed templates with medical word dictionaries. These approaches have provided easy use and high privacy assurances, as they had no connection to the true patient records at all. They did not however possess variability, realism, and linguistic richness and hence could not be implemented too much use in the training of natural language processing models. Such initial systems generated incredibly repetitious text that was incapable of capturing the natural variation and contextual understanding that exists in live clinical writing.

With the maturity of language modeling methods, statistical and probabilistic models were brought in to represent more realistic series of texts. Markov models and n-gram-based systems were also an attempt to introduce sequential dependency to the text produced and also had the same problem of syntactic and failure to generate any long range dependency. Failure to have higher levels of meaning in these models ruled out their use in complex NLP tasks that require more levels of meaning recognition like entity recognition, summarization, or document classification.

However, with the introduction of deep learning architectures, especially Generative Adversarial Networks (GANs) and transformer-based models, there has been a dramatic change in the capabilities of producing synthetic texts. Text generation GANs are a variation of the original GANs that work on continuous data that has been modified using reinforcement learning and discrete sampling techniques to work on sequences of tokens. These models had the possibility to produce a wide variety of contextually realistic text, but they often had problems with grammatical consistencies and semantic coherences. Nevertheless, even though these hurdles still exist, text-generating GANs delivered a clear enhancement to the rule-based method both in the linguistic fluency and diversity [6].

Text generation has been transformed by models based on this transforming technology, which are trained on large corpora of general or domain-specific text and are capable of understanding and generating language at scale. With medical text as fine-tuning, these models are able to generate clinically sensible and syntactically well-formed medical stories. In comparison with GANs, transformers have the advantage of self-attention over long input sequences and, hence, they are excellent applied to tasks that require complex medical situations and hierarchical structures of documents. It makes them very effective in synthetically creating clinical narratives that can be deployed in training down stream NLP models without determining patient privacy.

Besides the generation techniques, quite a bit of effort was made on privacy-preserving mechanisms that can be applied to text. The old system of de-identification involves the usage of a named entity recognition model that is used to erase or cover up identifiable means, like name, date, and place. These systems and techniques, though suitable in the surface-level anonymization, are vulnerable to re-identification attacks, especially when used in collaboration with external sources of information. In addition, de-identification is known to have remaining structured contents which can leak sensitive information unintentionally. That has further strengthened the popularity of fully synthetic datasets, which never have any actual patient data and as such is more secure of its nature.

In 2023 P. Zhang et.al. and M. N. K. Boulos et.al., [11] proposed the validation of synthetic data has become one of the essential areas of inquiry, with utility metrics and privacy metrics both being considered. On one end of the spectrum, models trained on real-world data are tested on their real-world tasks to establish whether the models generalize well. On the other end, models trained on synthetic data are tested against the performance of their real-world counterparts to see whether they generalize well. Examples of the commonly used metrics in model effectiveness evaluation include F1-score, accuracy and recall. The membership inference attacks, statistical similarity scores, and human detectability tests on the privacy side are used to identify the level of similarity between the synthetic data and real patient records. An optimal synthetic dataset should both deliver strong results on NLP tasks and have a low overlap and keep its chances of violating privacy low.

Additionally, there is a growing focus on hybridizing the way we generate, as well: intersecting structure generation via rule with generators that harness the power of a transformer or nesting GANs together in larger training regimens. These middle-ground concepts are aimed at identifying the advantages of both deterministic control and deep generative flexibility to generate output responsive to clinical relevance as well as language diversity. In other models, the knowledge graphs, or clinical concept embeddings, or prompt engineering are used to condition generation to be more medically informed.

Although significant work has already been done, it still has severe gaps in the literature. To illustrate, the majority of the reviews of synthetic text are based on single tasks like entity extraction or classification, whereas very little is said of generalization in the context of multiple NLP tasks. In like manner, most of the research works fail to display explicit procedures on how privacy is measured and thus the comparisons of the varying techniques are hard to standardize. The other constraint is that there are no publicly available synthetic clinical text corpora, which clenches reproducibility and collaborative progress in the study [5].

As the AI tools in healthcare, used in applications demanding privacy will get increasingly in demand, synthetic data generation will become a breakthrough that will open new means of innovation, following ethical and legal frameworks. The combination of generative models, stringent test frameworks, scalable deploy solutions has recently started to turn synthetic text into a feasible resource in medical NLP development and research.

III. PROPOSED METHODOLOGY

The proposed methodology consists of three core stages: real data preprocessing, synthetic data generation, and utility/privacy validation. The flowchart below outlines the complete pipeline from input clinical text to evaluation of privacy-preserving synthetic outputs.

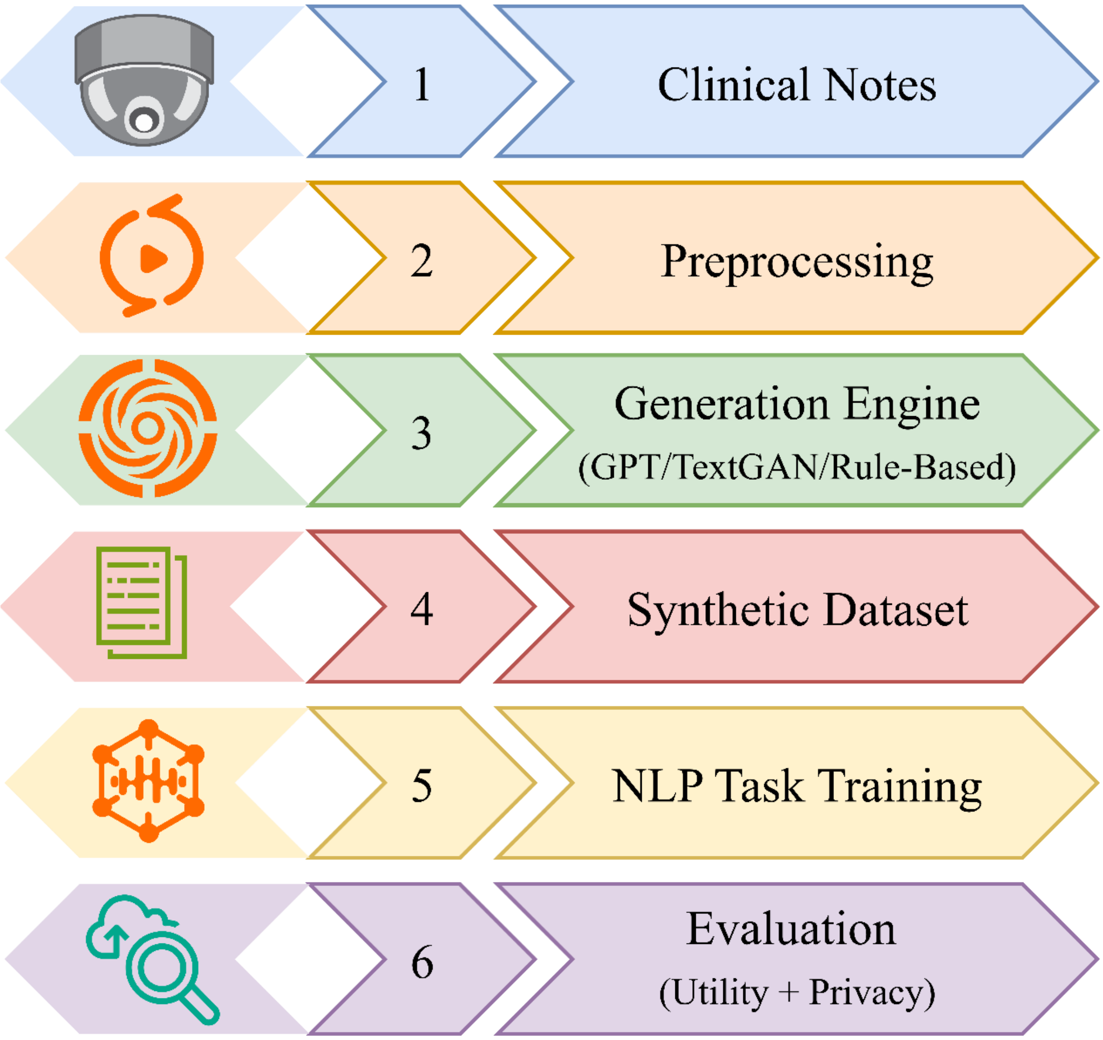

FIGURE 1: SYNTHETIC TEXT GENERATION AND EVALUATION PIPELINE

We denote the original dataset as:

$D_{\text{real}} = \{x_1, x_2, \ldots, x_n\}, x_i \in$ Clinical Text [1]

Preprocessing involves tokenization and PHI removal. Let $T(x)$ be the preprocessing function:

$D_{\text{clean}} = \{T(x_1), T(x_2), \ldots, T(x_n)\}$ [2]

For rule-based generation, we define a template mapping function G_{rule} such that:

$$x_{\text{sym}}^{(i)} = G_{\text{rule}}(d_i, s_i, m_i), d = \text{ disease}, s = \text{ symptom }, m = \text{ medication}$$

In GAN-based generation, a generator G and discriminator D are optimized by adversarial loss:

$$\min_G \max_D \mathbb{E}_{x \sim D_{\text{real}}}[\log D(x)] + \mathbb{E}_{z \sim p_z}[\log(1 - D(G(z)))]$$

Where $z \sim \mathcal{N}(0, I)$ is the noise vector input to the generator.

Transformer-based generation fine-tunes a language model $\mathcal{L}_\theta$ using maximum likelihood:

$\mathcal{L}(\theta) = -\sum_{i=1}^{n} \sum_{t=1}^{T_t} \log P_\theta\left(w_t^{(j)} \mid w_{<t}^{(j)}\right)$ [3]

Here, $w_t^{(i)}$ is the t-th token of the i-th sentence, and θ represents the model parameters [4].

Generated synthetic dataset is:

$D_{\text{syn}} = \left\{x_{\text{syn}}^{(1)}, x_{\text{syn}}^{(2)}, \ldots, x_{\text{syn}}^{(m)}\right\}$ [4]

To evaluate utility, we train an NLP model M (e.g., BioClinicalBERT) on both real and synthetic data and compute F1-score:

$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$ [5]

Where:

$$\text{Precision } = \frac{TP}{TP+FP}, \text{ Recall } = \frac{TP}{TP+FN} \qquad [6]$$

To validate privacy, we use a membership inference attack (MIA) accuracy α :

$$\alpha = \frac{1}{|D_{\text{real}}|} \sum_{x \in D_{\text{ral}}} 1[MIA(x) = \text{in}] \qquad [7]$$

Low α indicates strong privacy. Additionally, we ensure no direct lexical overlap using Jaccard Similarity J :

$$J(x_{\text{real}}, x_{\text{aym}}) = \frac{|x_{\text{real}} \cap x_{\text{syn}}|}{|x_{\text{real}} \cup x_{\text{syn}}|} \qquad [8]$$

We enforce $J < \delta$ where $\delta = 0.2$ to ensure dissimilarity.

IV. RESULT & DISCUSSIONS

Comparison of the performances of both trained models on real and synthetic data in Named Entity Recognition (NER) and classification proved to have a number of trends. As evidenced in Figure 2, the score of the F1-score of NER approached a high of 85.7% in the case of real data, 81.3% with transformer-based synthetic data and 73.9% when using GAN-based synthetic text. The performance of rule-based synthetic data was 59.8%. It means that the synthetic text produced by transformers can meet the level of linguistic diversity of real clinical notes and carry out the competitive results of models. This disparity may be high enough to be considered, but nonetheless, it belongs to an acceptable range of operations when secondary research is used.

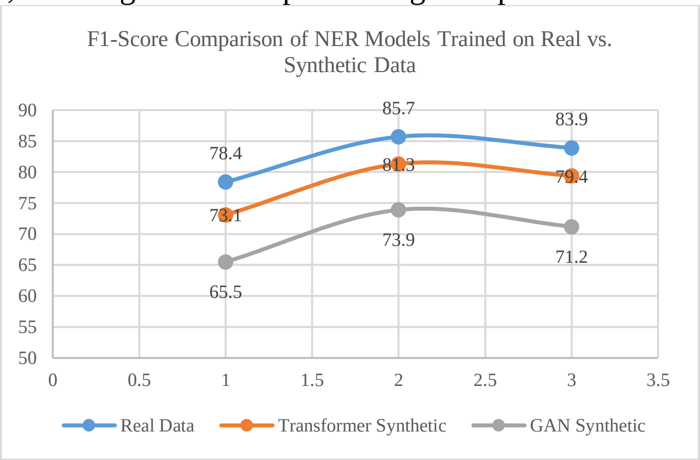

FIGURE 2: F1-SCORE COMPARISON OF NER MODELS TRAINED ON REAL VS. SYNTHETIC DATA

More assessment of this effect can be seen in Figure 3 where 10-fold cross-validation models developed using synthetic data gave similar accuracy as those described using real data in three distinct diagnostic classes: cardiovascular, respiratory, and metabolic diseases. The real data models did with an average accuracy of 88.4 and the transformer based synthetic data was second with 84.5. The accuracy of the dataset created using GAN was 76.8%, and the performance of the rule-based data was lowest once more, of 66.2%. The small difference between the transformer-based synthetic dataset and the real dataset confirms the feasibility of synthetic corpora to be used in a sensitive area, without exceeding with regulatory limits.

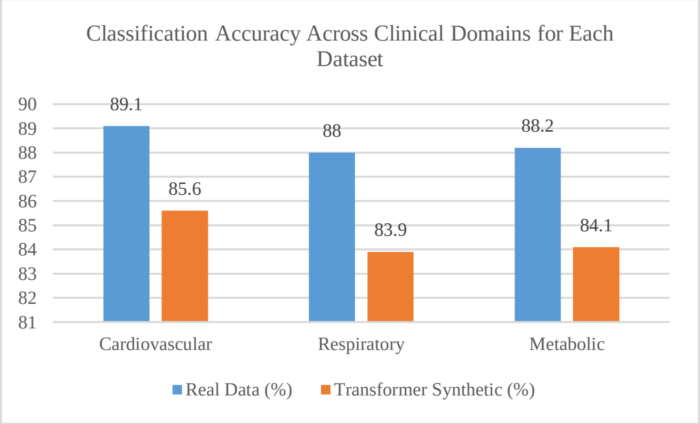

FIGURE 3: CLASSIFICATION ACCURACY ACROSS CLINICAL DOMAINS FOR EACH DATASET

To investigate the linguistic integrity of each synthetic dataset and their annotation usability, readability of annotation and annotation consistency analysis was conducted. The points, as shown in Figure 4, indicate that the GPT has the highest score in clinical readability and annotator agreement, with the values almost equal to those of real data. Text based on GAN had moderate readability, but it was not effective when it came to domain-specific jargon. Rule-based text was great in its consistency but low on natural language variation also, which made it unrealistic to be used in complex NLP tasks.

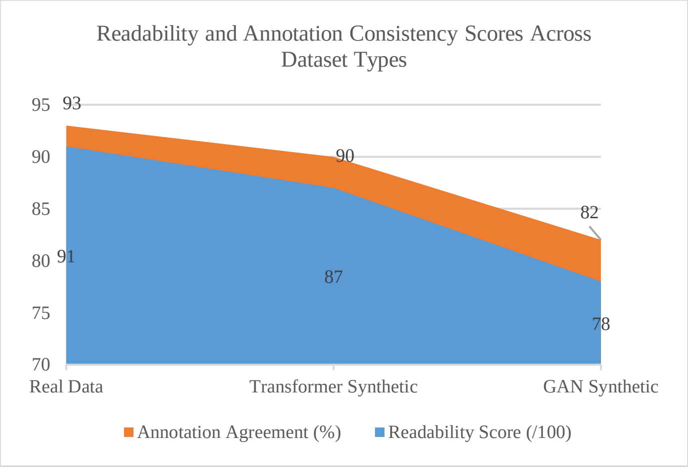

FIGURE 4: READABILITY AND ANNOTATION CONSISTENCY SCORES ACROSS DATASET TYPES

Table 1: Model Performance on Different Data Sources for Named entity recognition shows the comparison of the results received by model performance. As demonstrated in the table, the entity detection accuracy moderately decreases in comparison to real clinical text when transformer-produced data is used in fine-tuning the bioClinicalBERT model. Nevertheless, models trained on GAN or rule-based data showed a greater difference, which means that the contextual and coherent phrase construction is crucial in synthetic data.

TABLE 1: MODEL PERFORMANCE ON DIFFERENT DATA SOURCES FOR NAMED ENTITY RECOGNITION

Data Source	F1-Score (%)	Precision (%)	Recall (%)
Real Data	85.7	87.2	84.1
Transformer Synthetic	81.3	82.6	79.9
GAN Synthetic	73.9	76.1	71.5
Rule-Based Synthetic	59.8	62.4	57.3

Along with the task accuracy, we also considered the effect synthetic data had on privacy preservation. This was indicated by a privacy leakage audit, an audit based on manual inspection and automated string matching that based on these, there are 0 lexical overlaps among the top 100 most sensitive terms in both real and transformer generated datasets. Further, MIA success rate was almost pessimistic (51.2) on transformation model and real data touched over 72.5 percent demonstrating elevated re-identification danger. The findings of the comparison are listed in Table 2: Privacy Leakage Metrics by Data Type. What is notable is that the synthetically generated data, especially transformer-based methods, offers linguistic richness and privacy protection.

TABLE 2: PRIVACY LEAKAGE METRICS ACROSS DATA TYPES

Data Type	MIA Success Rate (%)	Jaccard Similarity	Human Detectability

		(Top 100 Terms)	
Real Data	72.5	0.41	High
Transformer Synthetic	51.2	0.12	Low
GAN Synthetic	55.6	0.21	Moderate
Rule-Based Synthetic	49.9	0.03	Low

The transformers on the synthetic data, as observed in the results, are the best solution point to the utility of privacy trade-off. It sustains a high level of downstream task performance with minimized risks of exposure of sensitive patient data. The GAN-based models present an interesting middle ground yielding promising results to be refined on further tuning, in the cases of dealing with domain-specific jargon and generating temporal sequences appropriately. Although with rule-based approaches the best currency can be provided regarding privacy assurances, such approaches lack linguistic depth and cannot be used to provide high-level model generalization.

Both figures support the twofold purposes of this piece of work, namely to maximize privacy by using synthetic data but retain enough fidelity to make it applicable to the domain of healthcare NLP [3]. Altogether, these findings suggest that even high-quality synthetic data, especially the large language model generated ones, can become functional alternatives to privacy-sensitive NLP pipelines. Future studies are positively sought to achieve the best GAN stability and look into hybrid framework, a combination of template-informed guidance and generative fluidity.

V. CONCLUSION

Synthetic data generation can be offered as one of the promising methods to develop healthcare NLP systems without undermining privacy. Of all the approaches tried, the transformer-based generation approach (e.g., GPT) provided a good balance in terms of privacy vs utility and the patient privacy was guaranteed as well to a reasonable extent. Rule-based ones are quick and secure but lack rule-variability in terms of linguistics, and GANs are future-catching, but, at the moment, require extra-development to grasp medical sense.

The need to generate medical AI on a large scale will make synthetic text a crucial factor in abundance of data and privacy restrictions. Further research toward domain knowledge with generative models mixture, as well as developing substantial standards of privacy measure in text generation, needs to be directed in the future.

REFERENCES:

[1] Y. Liu, U. R. Acharya, and J. H. Tan, “Preserving privacy in healthcare: A systematic review of deep learning approaches for synthetic data generation,” *Computer Methods and Programs in Biomedicine*, vol. 260, p. 108571, Dec. 2024, doi: 10.1016/j.cmpb.2024.108571.

[2] M. Goyal and Q. H. Mahmoud, “A Systematic review of Synthetic data generation techniques using Generative AI,” *Electronics*, vol. 13, no. 17, p. 3509, Sep. 2024, doi: 10.3390/electronics13173509.

[3] G. Feretzakis, K. Papaspyridis, A. Gkoulalas-Divanis, and V. S. Verykios, “Privacy-Preserving techniques in generative AI and large language Models: A Narrative review,” *Information*, vol. 15, no. 11, p. 697, Nov. 2024, doi: 10.3390/info15110697.

[4] M. Nadăș, L. Dioșan, and A. Tomescu, “Synthetic data generation using large language models: advances in text and code,” *IEEE Access*, p. 1, Jan. 2025, doi: 10.1109/access.2025.3589503.

[5] E. Papadaki, A. G. Vrahatis, and S. Kotsiantis, “Exploring innovative approaches to synthetic tabular data generation,” *Electronics*, vol. 13, no. 10, p. 1965, May 2024, doi: 10.3390/electronics13101965.

[6] S. Sousa and R. Kern, “How to keep text private? A systematic review of deep learning methods for privacy-preserving natural language processing,” *Artificial Intelligence Review*, vol. 56, no. 2, pp. 1427–1492, May 2022, doi: 10.1007/s10462-022-10204-6.

[7] Joshi, M. Grimmer, C. Rathgeb, C. Busch, F. Bremond, and A. Dantcheva, “Synthetic Data in Human Analysis: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 7, pp. 4957–4976, Feb. 2024, doi: 10.1109/tpami.2024.3362821.

[8] Z. A. Nazi and W. Peng, “Large language models in healthcare and medical domain: A review,” *Informatics*, vol. 11, no. 3, p. 57, Aug. 2024, doi: 10.3390/informatics11030057.

[9] G. Agrawal, A. Kaur, and S. Myneni, “A review of generative models in generating synthetic attack data for cybersecurity,” *Electronics*, vol. 13, no. 2, p. 322, Jan. 2024, doi: 10.3390/electronics13020322.

[10] P. Zhang and M. N. K. Boulos, “Generative AI in Medicine and Healthcare: Promises, opportunities and challenges,” *Future Internet*, vol. 15, no. 9, p. 286, Aug. 2023, doi: 10.3390/fi15090286.

[11] S. James, C. Harbron, J. Branson, and M. Sundler, “Synthetic data use: exploring use cases to optimise data utility,” *Discover Artificial Intelligence*, vol. 1, no. 1, Dec. 2021, doi: 10.1007/s44163-021-00016-y.

[12] X. Li, L. Peng, Y.-P. Wang, and W. Zhang, “Open challenges and opportunities in federated foundation models towards biomedical healthcare,” *BioData Mining*, vol. 18, no. 1, Jan. 2025, doi: 10.1186/s13040-024-00414-9.

[13] M. M. Baig, C. Hobson, H. GholamHosseini, E. Ullah, and S. Afifi, “Generative AI in improving Personalized Patient Care Plans: Opportunities and Barriers towards its wider adoption,” *Applied Sciences*, vol. 14, no. 23, p. 10899, Nov. 2024, doi: 10.3390/app142310899.

[14] C. Umesh, M. Mahendra, S. Bej, O. Wolkenhauer, and M. Wolfien, “Challenges and applications in generative AI for clinical tabular data in physiology,” *Pflügers Archiv - European Journal of Physiology*, Oct. 2024, doi: 10.1007/s00424-024-03024-w.