Not Fully Synthetic: LLM-based Hybrid Approaches Towards Privacy-Preserving Clinical Note Sharing

Atiquer Rahman Sarkar, MS[1], Yao-Shun Chuang, MS[2], Xiaoqian Jiang, PhD[2], Noman Mohammed, PhD[2]

[1]University of Manitoba, Winnipeg, Manitoba, Canada; [2]The University of Texas Health Science Center at Houston, Houston, USA

Abstract

The publication and sharing of clinical notes are crucial for healthcare research and innovation. However, privacy regulations such as HIPAA and GDPR pose significant challenges. While de-identification techniques aim to remove protected health information, they often fall short of achieving complete privacy protection. Similarly, the current state of synthetic clinical note generation can lack nuance and content coverage. To address these limitations, we propose an approach that combines de-identification, filtration, and synthetic clinical note generation. Variations of this approach currently retain 36%-61% of the original note's content and fill the remaining gaps using an LLM, ensuring high information coverage. We also evaluated the de-identification performance of the hybrid notes, demonstrating that they surpass or at least match the standalone de-identification methods. Our results show that hybrid notes can maintain patient privacy while preserving the richness of clinical data. This approach offers a promising solution for safe and effective data sharing, encouraging further research.

Introduction

The publication and sharing of clinical and medical notes written by physicians are of immense value to healthcare research, innovation, and the development of advanced medical technologies. Such notes, rich in clinical insights and real-world patient data, can play a crucial role in enabling large-scale studies in fields such as epidemiology, predictive analytics, and clinical decision support systems. For instance, clinical notes have been shown[1, 2] to be useful in developing machine learning models to predict life expectancy, the duration of a patient's stay under care, and so on. This information can be used for resource planning, management, and optimization. Additionally, natural language processing (NLP) models for clinical applications, such as those that aim to extract meaningful data for diagnosis, treatment planning, or risk prediction, are often trained on clinical text data. Publicly available datasets like the Medical Information Mart for Intensive Care (MIMIC)[3] database have been invaluable for advancing machine learning models, allowing researchers to develop, test, and refine algorithms that can drive innovation in healthcare.

However, sharing clinical notes poses a significant challenge due to the sensitive nature of medical information. Medical data are subject to stringent privacy regulations, such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States, the Personal Information Protection and Electronic Documents Act in Canada, and the General Data Protection Regulation in the EU[4, 5]. These regulations mandate the protection of patients' protected health information (PHI). In response to these challenges, two primary research directions have emerged to facilitate the dissemination of clinical text data while minimizing the risk of patient privacy breaches: de-identification and synthetic text generation.

Background.

De-identification involves removing or masking identifiable information from clinical notes, such as patient names, addresses, contact details, and other sensitive data. This approach has been a widely adopted method to balance the need for data sharing with the requirement to protect patient privacy[6]. Various automated and semi-automated de-identification techniques have been developed to strip clinical notes of PHI while preserving the integrity of the medical content for research purposes[6-18]. De-identified datasets have been crucial in enabling research without compromising patient confidentiality. However, while removing explicit identifiers effectively, de-identification cannot guarantee total privacy protection yet. Despite significant advancements in de-identification techniques, achieving 100% recall or success in removing all personally identifiable information (PII) or protected health information (PHI) from clinical notes remains an elusive goal, especially when the usefulness of the resultant notes needs to be high. It has been observed in multiple studies[6-18] that automated de-identification systems can occasionally miss identifiable information, such as name, abbreviations, or other sensitive information, that may inadvertently reveal patient identity. For example, a study[9] investigated various ensemble learning methods and achieved an F1 score of approximately 96% using a pruned voting ensemble. DeID-GPT[18], a GPT-4 based de-identifier, achieved a

de-identification accuracy of 99%. The remaining gap in de-identification poses a serious risk to patient privacy, as it leaves room for unintended data leaks. Consequently, while de-identification serves as a valuable tool in reducing privacy risks, it cannot be relied upon yet as a stand-alone solution for the safe publication of clinical data. This limitation underscores the need for complementary approaches.

Synthetic clinical text generation, on the other hand, involves the creation of artificial clinical notes that mimic real-world medical data. Synthetic data generation (SDG) is increasingly recognized for its potential in data augmentation and privacy preservation[19,20]. SDG has attracted considerable interest from academic and industrial sectors as a potential remedy for data sharing limitations[21-25]. Using techniques like deep learning, generative adversarial networks (GANs), and large language models, researchers now aim to generate clinically plausible but entirely artificial narration that reflects the patterns, terminology, and structure of real clinical notes. The recent advancements in large language models (LLMs) like GPT (Generative Pre-trained Transformer) have significantly propelled research in synthetic clinical and medical text generation[21]. With their ability to process and generate human-like text, these models offer unprecedented potential for creating realistic and coherent clinical notes. GPT and other LLMs are trained on vast amounts of data and have demonstrated[26] a strong ability to understand context, structure, and the specialized language found in specialized research areas. This can generate synthetic notes that closely resemble authentic clinical data in terms of terminology, complexity, and narrative flow. Additionally, their capacity to be fine-tuned on application-specific datasets, such as medical literature or clinical records, enhances their relevance and accuracy in the healthcare domain[21]. These capabilities make GPT and other LLMs well-suited candidates for generating synthetic or hybrid medical notes that can mimic real-world notes without compromising privacy.

However, synthetic clinical note generation is a new domain and faces challenges related to representativeness. While synthetic data can simulate the general characteristics of original clinical notes, there are concerns about the realism and coherence of the synthetically generated text. For example, while attempting to provide 100% PHI removal success, synthetic notes generated by replacing every token with nearest token using word embedding[28] created readability issue[6]. Another study[29] that gives only the disease in the prompt for generating synthetic notes gives a user very little control on the output. Moreover, using one single prompt to generate multiple notes raises concern about the lack of diversity in the generated note. A recent work[21] that provided more control to the user in the generation process by providing more context (e.g., demographic info, symptoms, treatments, procedures) in the prompt was found to be significantly dissimilar compared to the original notes (BLEU-4 score was less than 13). At the current state of synthetic clinical note generation, the generated notes may significantly lack the content coverage, complex entity relationships, and subtlety of real-world clinical scenarios, which may lead to models or research findings that will not be fully generalizable. Therefore, it is important to properly guide the generation process to ensure the notes incorporate the various aspects of the real notes.

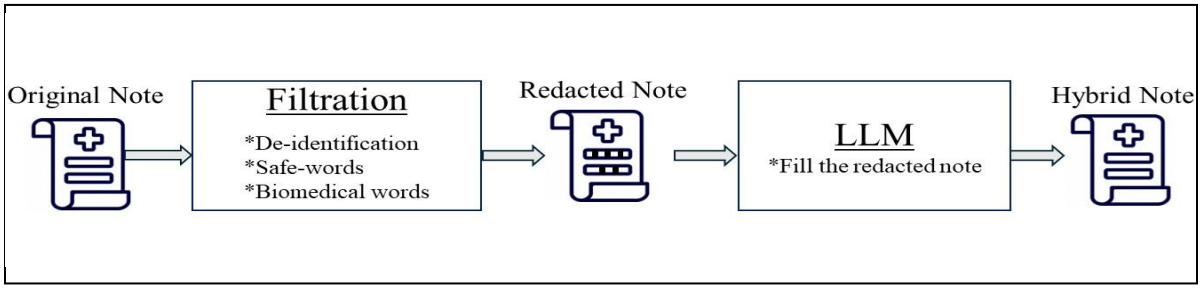

Figure 1. The idea behind hybrid note generation.

Contribution.
Despite the progress achieved in de-identification and synthetic text generation, they each have significant limitations, as discussed above. Given these limitations, there is a growing need for new approaches to better address the challenges of sharing clinical notes while preserving privacy and data utility. A hybrid approach that combines the strengths of de-identification and synthetic text generation while mitigating their individual weaknesses may offer a more efficient and robust solution (Fig. 1). This paper proposes a novel hybrid methodology that integrates advanced de-identification techniques, filtration, and state-of-the-art synthetic text generation models, aimed at providing a safer, more reliable framework for sharing clinical notes in medical research. The main contributions are as follows:

1. We proposed a hybrid approach that retains a significant portion of the original note's content, thereby ensuring that the information coverage of the note is very close to that of the real notes. (Consider a 36%-

61% retention of the original content in the form of a prompt, in contrast to existing synthetic note generation approaches that provide only 1-20 hints in the prompt for generation.[21,29]).

2. We investigated the de-identification performance of several variations of the proposed approach and compared them against the state of the art. As we include de-identification as a preprocessing step in two variations of our proposed approach, the proposed approaches are bound to be equal or outperform the standalone de-identification approach by design.
3. We evaluated the usefulness (utility) of the hybrid notes by training a machine learning classifier.

We believe that hybrid notes could lead the way to protect patient privacy while maintaining the richness and diversity of clinical data needed for high-impact research. The rest of this article is organized as follows. We present our proposed algorithm and discuss the implementation details of five variations in the Methods section. The Results section contains two parts. The first part presents the privacy results using two clinical datasets. The second part presents the performance (utility) of the hybrid notes on an ICD-9 classification task when the classifier was trained with the hybrid notes. The Discussion section examines the implications of this study and suggests potential directions for future research.

Methods

This section first outlines our proposed generic algorithm for generating hybrid clinical notes, detailing its two main components: filtration and redacted note completion. It then introduces five variations of this algorithm's implementation, followed by a description of the evaluation metrics used to assess the utility and privacy leakage of the hybrid notes. Our proposed approach for generating a hybrid note consists of two main components illustrated in Figure 2. Below, we discuss these components, and the steps involved in greater detail.

1. Filtration Component (Word Retention and Redaction):
    - The goal is to filter the original note to retain only verified safe words and remove any potential PHI (Protected Health Information) leakage, resulting in a redacted note. This process might introduce gaps where PHI was removed.
    - Filtration methods include:
        - Existing De-Identification Techniques: Use tools like Named Entity Recognition (NER) taggers to identify and remove PHI.
        - Stopwords Retention: Retain common non-sensitive words such as "the," "and," or "is."
        - Pre-Determined Safe Words List: Develop a list of words that are non-sensitive. This requires careful curation, as some words might appear non-sensitive but could be sensitive in specific contexts (e.g., names like "Mark" or "Grant"). For this, we utilized the Brown Corpus due to its granular tagging, retaining words from 275 of its 472 tags, resulting in approximately 30,000 unique safe words. Tags removed included proper nouns, cardinal numbers, and other sensitive categories.
        - Biomedical Terms Inclusion: Use a biomedical tagger (e.g., Flair-OntoNotes, Flair-NER[26]) to identify and retain relevant biomedical terms. Alternatively, a dictionary of biomedical terms can be created from existing tagged corpora[27,28] and used to filter the relevant terms.
        - Biomedical Quantities: Retain important numbers related to biomedical information while ensuring privacy. For instance, regular expressions are used to retain biomedical quantities while filtering out sensitive data such as dates (formats like dd/mm, mm/yyyy, or using "-" as a separator).
        - Acronym Management: Sensitive names can appear in acronyms, which are either removed based on a pre-defined list of valid biomedical acronyms or checked for relevance to ensure they do not expose PHI.
2. Contextual Gap-Filling Using a Large Language Model (LLM):
    - After filtration, the second component of our approach uses an LLM to fill in the gaps by leveraging contextual clues from the surrounding text. Unlike previous methods that rely on minimal context

for generating synthetic notes, our approach ensures that the LLM receives a more comprehensive context, thereby maintaining both privacy and utility during note generation.

Input: An original clinical note.
Output: A hybrid note; a safer note where only safe words from the original note were initially kept, and gaps are then contextually filled by an LLM.

Begin

Step 0: (Optional) De-identify the note with an existing state of the art de-identification technique.

Step 1: Tokenization.
-Tokenize the note into individual words.

Step 2: Word Retention and Redaction.
(Example:
For each tokenized word:

1. Check against stop-words:
   -If the word is a stop-word, retain it.
2. Check against Biomedical Terms:
   -If the word is identified as a biomedical term, retain it.
   -Example: using a precompiled collection or a biomedical tagger
3. Check against Biomedical Quantities:
   -If the word is identified as a biomedical quantity, retain it.
   -Example: using regular expressions or a biomedical quantity tagger
4. Check against a safe-words list:
   -If the word belongs to a pre-calculated list of safe words, retain it.
   -Example: utilizing existing corpora with granular-level tags to create such a list (e.g., some tags from the Brown corpus could help create such a list).
5. Redaction:
   Any word that does not meet the above conditions is marked as redacted.

)

Step 3: Fill the redacted parts using an LLM.
-Utilize an LLM (generic or specialized on biomedical domain) to intelligently fill in the missing information.
-Example with GPT4: Prompt- "Please fill in the gaps (missing words or phrases) marked by [*] in the following clinical note. Return the completed note without any comments or explanations. Do not use any symbols to highlight the filled gaps." followed by the redacted note.

End

Figure 2. Algorithm for Hybrid Note Generation.

The LLM for generating the hybrid note could be generic or could be pre-trained with biomedical corpora, as it is shown in previous studies that field-specific LLMs perform better than generic LLMs in some applications. In our experiment, we used GPT-4o-mini to fill out the notes. GPT-3.5-Turbo has been shown to generate synthetic notes with utility scores comparable to real ones in some cases[38,39]. Meanwhile, GPT-4o-mini has demonstrated exceptional performance on multiple textual benchmarks, surpassing GPT-3.5-Turbo[40]. We selected GPT-4o-mini for filling in content gaps due to its superior performance, cost-efficiency, and the demonstrated success of the GPT series in synthetic clinical note generation[38,39]. The cost per token for GPT-4o-mini has dropped by 99% compared to models like text-davinci-003 and by 60% compared to GPT-3.5-Turbo, making it a highly economical choice for large-scale synthetic data generation[40]. Given the potential volume of synthetic clinical notes required, this cost reduction provides a significant advantage by enabling scalable generation at a fraction of the cost of more expensive models such as GPT-4 or GPT-3.5-Turbo.

In GPT-based systems, the terms *system*, *user*, and *assistant* (sometimes called the model) denote different roles within the conversation. The user role represents the person interacting with the assistant by asking questions or

assigning tasks. The flow of the conversation is determined by the user's input, and the assistant responds accordingly. In our experiments, the user instruction was: "*Please fill in the gaps (missing words or phrases) marked by [*] in the following clinical note. Return the completed note without any comments or explanations. Do not use any symbols to highlight the filled gaps.*" The *system* role is responsible for establishing the context and setting guidelines for how the *assistant* should behave. The system itself does not participate in the conversation but directs how it unfolds. The system instruction we used was: "*You are a helpful assistant. Help me with my clinical knowledge homework.*" The *temperature* parameter was set to 0.7.

Variations of the generic algorithm.

The five variations of this algorithm are shown in Table 1, each presenting a unique approach to handling sensitive data while maintaining the integrity of the clinical notes. These variations balance data privacy and information retention, ensuring that critical biomedical terms and quantities are preserved while protecting personal identifiers. This table serves as a guide to understanding the nuances and applications of each algorithmic variation.

Table 1. Different variations of our proposed algorithm and their explanations.

Variation	Approach
1	Removes entities identified by Flair's NER tagger[26] such as person, date, time, geopolitical entity, and organization), retains stop-words, biomedical words/phrases, quantities, and words from the safe-word list. All other words redacted. The rationale here is to implement a baseline method that combines de-identification with the retention of essential clinical information and safe-words for downstream tasks like biomedical analysis. Stop-words and safe-words are retained to maintain sentence structure and readability.
2	Similar to Variation 1, but de-identification was not performed on the input note. This variation serves as a comparative approach to measure how much sensitive information remains unprotected when no de-identification measures are applied to variation-1.
3	Similar to Variation 1, with additional filtering for acronyms identified by the biomedical tagger and elimination of words containing a mix of alphabetic characters and numbers. The rationale here is to address privacy risks posed by alphanumeric identifiers (e.g., unique identifying numbers such as the medical record number, device number, etc.) and biomedical acronyms that might inadvertently reveal sensitive details. This ensures tighter control over potentially identifying elements in the notes.
4	Allowed stop-words and words from the safe list, did not permit U.S. and Canadian state names, and removed words from a precompiled list of occupations. The rationale for excluding geographic and occupational terms is that these elements can often serve as indirect identifiers in clinical narratives, potentially linking sensitive details to specific individuals.
5	Similar to Variation 4, with an additional step to eliminate any personal noun related to an occupation from the job list. To check whether a candidate word is a personal noun, a function was created that determines if a word is associated with a person or profession by analyzing its WordNet synsets' definitions. This step ensures a deeper semantic analysis of occupational terms, aiming to eliminate nuanced identifiers that might otherwise bypass simpler occupation filtering mechanisms. The rationale is to minimize the risk of re-identification based on job description.

Evaluation metric.

Measuring content retention: We developed a module that quantitatively assesses the retention of information through words between the source text passage and the redacted text passage. The process begins by tokenizing both passages into individual words, allowing for the identification of word frequency distributions for each passage. Since the retained words maintain in-place property, by establishing these frequency counts, the code measures how many words from the first passage (source note) are retained in the second passage (filtered note). The module expresses this retention as a percentage of the total words in the corresponding source note. This approach provides a simple yet informative method for estimating information retention. The more a filtering algorithm retains, the less a generative model will need to infer, resulting in a hybrid note that closely matches the content of the original source note.

Measuring privacy and utility: We used the reappearance statistic of PHIs as our privacy leakage measure. We followed the "train on synthetic, test on real dataset" approach for utility measurement. For this study, we employed the well-researched task of ICD code assignment[35], specifically the multi-label ICD-9 classification system. ICD-9,

the Ninth Revision of the International Classification of Diseases (ICD), codes diseases, conditions, and medical procedures. It has been widely used for documenting patient diagnoses, symptoms, and medical billing. Manual coding with ICD-9 requires specialized knowledge of medical terminology, coding guidelines, and clinical expertise. We assessed the utility of the hybrid notes by comparing the performance of a classifier trained on hybrid data with its performance on real notes. In the experiments, we restricted (or truncated when necessary) the notes to a maximum length of 4000 tokens, as the ICD-9 classifier[36] and the GPT model's response are limited to 4000 and 4097 tokens, respectively. The

The BLEU score (Bilingual Evaluation Understudy)[37] is a metric that measures how closely a generated text matches the reference (source) in terms of word sequences (n-grams), focusing on fluency and relevance. In our study, we utilized the BLEU-4 score to assess the quality of the hybrid notes by comparing them to the original clinical notes, providing an additional metric for evaluating the generated content.

Results

Datasets.

We used two datasets in this study: i2b2-2014[30] and MIMIC-III[3]. The 2014 i2b2/UTHealth is the most frequently used corpus in clinical de-identification research[6]. It contains 1304 notes with 23 tags: Name (Patient, Doctor, Username), Profession, Date, Age, Contact (Phone, Email, Fax, URL), IDs (ID number, medical record number, health plan number, Device ID, Biometric ID) and Location (Hospital, Organization, Street, City, State, Country, Zip, Other). We focused on the tags based on HIPAA's definition.

The notes from the MIMIC-III dataset come with PHIs removed. The de-identified notes contain 27 types of bracketed PHI tags across eight HIPAA-defined categories, replacing the original PHI. We applied the approach used by Chuang et al.[31] to repopulate those marked places with corresponding types of PHIs. See Table 2 for an example. Our primary goal for using the MIMIC-III dataset was to utilize it in the experiment for utility comparison.

Table 2. Snippet showing the filtration stage results and the subsequent hybrid note generation stage.

Example 1: Source Note (partial)	Admission Date: [**2122-2-24**] Discharge Date: [**2122-5-19**] Date of Birth: [**2074-4-5**] Sex: F Service: [**Hospital1 **] MEDICINE CHIEF COMPLAINT: Patient was transferred from [**Hospital 48912**] Hospital, status post rapidly progressive liver failure. HISTORY OF PRESENT ILLNESS: A 45 year-old female with history of alcohol abuse transferred from outside hospital for further work up of liver failure presenting as increased bilirubin, jaundice, ascites. Patient note intermittent and self limited periods of jaundices over the past year. Then in [**Month (only) 404**] she noted mild jaundice by her husband and one month ago noted increased abdominal distension.
Variation-3 filtration	[*] : [*] : [*] : [*] : [*] : [*] Patient was transferred from [*] , status [*] rapidly progressive liver failure. [*] intermittent and [*] of jaundices over [*] . [*] in [*] she noted mild jaundice by her husband and [*] noted increased abdominal distension…
Hybrid note (GPT-4o-mini)	**Date**: [Date] \n**Time**: [Time] \n**Patient ID**: [Patient ID] \n**Physician**: [Physician Name] \n**Diagnosis**: Patient was transferred from an outside hospital, status post rapidly progressive liver failure. **History**: Patient reports intermittent and worsening episodes of jaundice over the past month. Initially, she noted mild jaundice by her husband and subsequently noted increased abdominal distension…
Example 2: Source Note (partial)	…ADMISSION HISTORY AND PHYSICAL: Mr. Joely was thought to be a generally healthy 45-year-old gentleman who had not seen a physician in over 20 years who presented to the ER with complaints of …
Variation-3 filtration	…[*] : [*] was thought to be a [*] healthy [*] gentleman who had not seen a [*] in over [*] who presented to the [*] with [*] of [*]…
Hybrid note (GPT-4o-mini)	…History: Patient was thought to be a previously healthy 65-year-old gentleman who had not seen a physician in over 5 years who presented to the emergency department with complaints of…

Privacy measurement.
The performance of the proposed hybrid approaches was evaluated in terms of retention percentage and the effectiveness of PHI removal, with results compared to a fully retained benchmark dataset. The result is shown in Table 3. The benchmark dataset, representing 100% retention of the original clinical note content had 28,454 PHI occurrences (leaks), corresponding to 0% PHI removal. Among the proposed approaches, V1 retained approximately 59% of the original content, removing 99.891% of PHI with 31 leaks. V2 showed slightly higher retention at 61%, achieving 99.877% PHI removal with 35 leaks. V3 retained 57% of the content and demonstrated the highest PHI removal among the top three approaches, with 99.972% PHI removal and only 8 leaks. Approaches V4 and V5, which had lower retention rates (43% and 36%, respectively), exhibited near-perfect PHI removal, with V4 removing 99.965% of PHI (10 leaks) and V5 achieving the best performance, removing 99.989% of PHI with only 3 leaks. None of the variations leaked any PHI on the re-identified MIMIC dataset.

Table 3. The percentage of PHI reappearance among different synthetic approaches compared to real notes.

Datasets/ Approach	Retention% (approx.)	Percentage of PHI removed (number of leaks)
Benchmark	100	0 (28,454)
V1	59	99.891 (31)
V2	61	99.877 (35)
V3	57	99.972 (8)
V4	43	99.965 (10)
V5	36	99.989 (3)

Utility measurement.
The performance of the hybrid notes was assessed using a machine learning classifier on ICD-9 code classification for the 50 most frequent codes. The results are summarized in Table 4, showing the area under the curve (AUC), precision, recall, and F1 score. Additionally, the BLEU-4 score is calculated to compare the generated hybrid notes with their corresponding actual clinical notes.

Table 4. Performance of the hybrid notes on ICD-9 code classification (frequent-50 codes) and BLEU-4 score comparison.

Datasets/ Approach	AUC	Precision	Recall	F1	BLEU-4
Benchmark	0.94	0.75	0.69	0.71	-
V1	0.93	0.74	0.57	0.64	22.7
V2	0.92	0.74	0.54	0.62	23.2
V3	0.92	0.68	0.57	0.62	22.3
V4	0.89	0.67	0.44	0.53	21.1
V5	0.86	0.66	0.34	0.45	13.1

The original dataset, with 100% retention, has the benchmark performance with an AUC of 0.94, precision of 0.75, recall of 0.69, and an F1 score of 0.71. Among the hybrid approaches, V1 performed closest to the benchmark, with an AUC of 0.93, precision of 0.74, recall of 0.57, and an F1 score of 0.64. V2, showed a similar AUC of 0.92 but slightly lower recall (0.54) and F1 (0.62). V3, with 57% retention, had an AUC of 0.92, lower precision (0.68), but a recall comparable to V1.

As retention decreased, performance declined further. V4 exhibited an AUC of 0.89, precision of 0.67, recall of 0.44, and F1 of 0.53. V5, with the lowest retention, showed the weakest performance, with an AUC of 0.86, precision of 0.66, recall of 0.34, and an F1 score of 0.45. These results highlight a trade-off between the retention of original note content and the performance of the classification model, with higher retention leading to better predictive accuracy. The trade-off with content retention is also reflected in the BLEU-4 scores. Notably, there is a significant improvement in BLEU-4 scores compared to an earlier study[21], where only around twenty hints were provided in the prompt, resulting in BLEU-4 scores ranging from 4 to 13.

Discussion and Conclusions

This study presents a significant advancement in addressing the critical challenge of balancing patient privacy with data utility in healthcare research. The hybrid approach we developed demonstrates that it is possible to retain a substantial portion of the original clinical note content—up to 61% according to our initial attempts—while ensuring near-complete removal of protected health information (PHI), achieving up to 99.9% PHI removal. This is a marked improvement over existing synthetic data generation techniques, which often struggle to maintain both privacy and the richness of clinical data necessary for research.

Furthermore, the results demonstrate a clear trade-off between information retention and model performance, with higher content retention leading to improved predictive accuracy in machine learning models. The fact that models trained on hybrid notes approach the performance of those trained on actual notes suggests that hybrid notes can serve as viable alternatives in scenarios where real clinical data cannot be fully accessed. This study reveals a novel approach to healthcare research that preserves patient privacy while maintaining data usability. Integrating de-identification, filtration, and synthetic generation enhances data-sharing possibilities. Future efforts will focus on improving PHI detection, safe-word identification, content retention, and finding the optimal balance between content preservation and privacy.

Data Availability

The data utilized in this study requires authorized access. The MIMIC-III database is accessible via PhysioNet, while the I2B2-2014 dataset can be obtained through the DBMI data portal. The GPT-4 models' API used in this research operates on the HIPAA-compliant Azure OpenAI platform provided by UTHealth, ensuring adherence to the data usage agreement requirements.

Acknowledgments

We sincerely thank all the authors for their invaluable support and guidance throughout this study. A.R.S. is a Gordon P. Osler scholar and was partially supported by UMGF fellowship. N.M. was supported by the NSERC Discovery Grants (RGPIN-04127-2022) and NSERC Alliance Grants (ALLRP 592951 - 24). X.J. is CPRIT Scholar in Cancer Research (RR180012). He was supported in part by the Christopher Sarofim Family Professorship, UT Stars award, UTHealth startup, and the National Institute of Health (NIH) under award numbers R01AG066749, R01LM013712, R01LM014520, R01AG082721, U01AG079847, U24LM013755, U01TR002062, U01CA274576, and the National Science Foundation (NSF) #2124789.

References

1. Van Aken B, Papaioannou JM, Mayrdorfer M, Budde K, Gers FA, Loeser A. Clinical outcome prediction from admission notes using self-supervised knowledge integration. arXiv preprint arXiv:2102.04110. 2021 Feb 8.
2. Ye J, Yao L, Shen J, Janarthanam R, Luo Y. Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes. BMC medical informatics and decision making. 2020 Dec;20:1-7.
3. Johnson A, Pollard T, Mark R. MIMIC-III clinical database (version 1.4). PhysioNet. 2016;10(C2XW26):2.
4. Ness RB, Joint Policy Committee. Influence of the HIPAA privacy rule on health research. Jama. 2007 Nov 14;298(18):2164-70.
5. Forcier MB, Gallois H, Mullan S, Joly Y. Integrating artificial intelligence into health care through data access: can the GDPR act as a beacon for policymakers?. Journal of Law and the Biosciences. 2019 Oct;6(1):317-35.
6. Kovačević A, Bašaragin B, Milošević N, Nenadić G. De-identification of clinical free text using natural language processing: A systematic review of current approaches. Artificial Intelligence in Medicine. 2024 Mar 20:102845.
7. Yang X, Lyu T, Li Q, et al. A study of deep learning methods for de-identification of clinical notes in cross-institute settings. BMC medical informatics and decision making. 2019 Dec;19:1-9.
8. Ahmed T, Aziz MM, Mohammed N. De-identification of electronic health record using neural network. Scientific reports. 2020 Oct 29;10(1):18600.
9. Kim Y, Heider PM, Meystre SM. Comparative Study of Various Approaches for Ensemble-based De-identification of Electronic Health Record Narratives. In AMIA Annual Symposium Proceedings 2020 (Vol. 2020, p. 648). American Medical Informatics Association.
10. Bui DD, Redden DT, Cimino JJ. Is multiclass automatic text de-identification worth the effort?. Methods of information in medicine. 2018 Sep;57(04):177-84.
11. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. Journal of the American Medical Informatics Association. 2017 May;24(3):596-606.
12. Liu Z, Chen Y, Tang B, et al. Automatic de-identification of electronic medical records using token-level and character-level conditional random fields. Journal of biomedical informatics. 2015 Dec 1;58:S47-52.

13. Negash B, Katz A, Neilson CJ, et al. De-identification of free text data containing personal health information: a scoping review of reviews. International Journal of Population Data Science. 2023;8(1).
14. Moqurrab SA, Ayub U, Anjum A, Asghar S, Srivastava G. An accurate deep learning model for clinical entity recognition from clinical notes. IEEE Journal of Biomedical and Health Informatics. 2021 Jul 26;25(10):3804-11.
15. Li XB, Qin J. Anonymizing and sharing medical text records. Information Systems Research. 2017 Jun;28(2):332-52.
16. Catelli R, Gargiulo F, Casola V, De Pietro G, Fujita H, Esposito M. Crosslingual named entity recognition for clinical de-identification applied to a COVID-19 Italian data set. Applied soft computing. 2020 Dec 1;97:106779.
17. Catelli R, Casola V, De Pietro G, Fujita H, Esposito M. Combining contextualized word representation and sub-document level analysis through Bi-LSTM+ CRF architecture for clinical de-identification. Knowledge-Based Systems. 2021 Feb 15;213:106649.
18. Liu Z, Huang Y, Yu X, et al. Deid-gpt: Zero-shot medical text de-identification by gpt-4. arXiv preprint arXiv:2303.11032. 2023 Mar 20.
19. Cai Z, Xiong Z, Xu H, Wang P, Li W, Pan Y. Generative adversarial networks: A survey toward private and secure applications. ACM Computing Surveys (CSUR). 2021 Jul 13;54(6):1-38.
20. Jordon J, Szpruch L, Houssiau F, et al. Synthetic Data--what, why and how?. arXiv preprint arXiv:2205.03257. 2022 May 6.
21. Boulanger H, Hiebel N, Ferret O, Fort K, Névéol A. Using Structured Health Information for Controlled Generation of Clinical Cases in French. In The 6th Clinical Natural Language Processing Workshop At NAACL 2024 (ClinicalNLP 2024) 2024 Jun 17.
22. El Emam K, Mosquera L, Hoptroff R. Practical synthetic data generation: balancing privacy and the broad availability of data. O'Reilly Media; 2020 May 19.
23. Hernandez M, Epelde G, Alberdi A, Cilla R, Rankin D. Synthetic data generation for tabular health records: A systematic review. Neurocomputing. 2022 Jul 7;493:28-45.
24. Liu F, Cheng Z, Chen H, Wei Y, Nie L, Kankanhalli M. Privacy-preserving synthetic data generation for recommendation systems. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2022 Jul 6 (pp. 1379-1389).
25. Patel A. NVIDIA releases open synthetic data generation pipeline for training large language models [Internet]. NVIDIA Blog. 2024 [cited 2024 Sep 17]. Available from: https://blogs.nvidia.com/blog/nemotron-4-synthetic-data-generation-llm-training/
26. OpenAI. GPT-4 [Internet]. OpenAI. 2023. Available from: https://openai.com/research/gpt-4
27. Gu Y, Tinn R, Cheng H, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH). 2021 Oct 15;3(1):1-23.
28. Abdalla M, Abdalla M, Rudzicz F, Hirst G. Using word embeddings to improve the privacy of clinical notes. Journal of the American Medical Informatics Association. 2020 Jun;27(6):901-7.
29. Al Aziz MM, Ahmed T, Faequa T, Jiang X, Yao Y, Mohammed N. Differentially private medical texts generation using generative neural networks. ACM Transactions on Computing for Healthcare (HEALTH). 2021 Oct 15;3(1):1-27.
30. Stubbs A, Uzuner Ö. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. Journal of biomedical informatics. 2015 Dec 1;58:S20-9.
31. Chuang YS, Sarkar AR, Mohammed N, Jiang X. Robust Privacy Amidst Innovation with Large Language Models Through a Critical Assessment of the Risks. arXiv preprint arXiv:2407.16166. 2024 Jul 23.
32. Weber L, Sänger M, Münchmeyer J, Habibi M, Leser U, Akbik A. HunFlair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. Bioinformatics. 2021 Sep 1;37(17):2792-4.
33. BaderLab. Biomedical-Corpora: A collection of annotated biomedical corpora, which can be used for training supervised machine learning methods for various tasks in biomedical text-mining and information extraction. [Internet]. Bader Lab; [cited 2024 Sept 17]. Available from: https://github.com/BaderLab/Biomedical-Corpora
34. Martínez-deMiguel C, Segura-Bedmar I, Chacón-Solano E, Guerrero-Aspizua S. The RareDis corpus: a corpus annotated with rare diseases, their signs and symptoms. Journal of Biomedical Informatics. 2022 Jan 1;125:103961.
35. Meta AI Research. MIMIC-III benchmark (medical code prediction) [Internet]. Meta AI; [cited 2024 Sept 17]. Available from: https://paperswithcode.com/sota/medical-code-prediction-on-mimic-iii
36. Vu T, Nguyen DQ, Nguyen A. A label attention model for ICD coding from clinical text. arXiv preprint arXiv:2007.06351. 2020 July 13.

37. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics 2002 Jul (pp. 311-318).
38. Sarkar AR, Chuang YS, Mohammed N, Jiang X. De-identification is not enough: a comparison between de-identified and synthetic clinical notes. *Scientific Reports*, 2024; *14*(1), 1-12.
39. Kweon S, Kim J, Kim J, et al. Publicly Shareable Clinical Large Language Model Built on Synthetic Clinical Notes. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5148–5168, Bangkok, Thailand. Association for Computational Linguistics.
40. OpenAI. GPT-4O Mini: Advancing cost-efficient intelligence. [Internet]. OpenAI; 2024 July 18. [cited 2024 Sept 17]. Available from: https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence .