Generation and Evaluation of Synthetic Endoscopy Free-Text Reports with Differential Privacy

Agathe Zecevic*[1,2], Xinyue Zhang*[3], Sebastian Zeki[1], Angus Roberts[3]
[1]Gastroenterology Department, Guy's and St Thomas' NHS Foundation Trust, United Kingdom
[2]Clinical Scientific Computing, Guy's and St Thomas' NHS Foundation Trust, United Kingdom
[3]Biostatistics and Health Informatics, Institute of Psychiatry, Psychology and Neuroscience
King's College London, United Kingdom
agathe.zecevic@gstt.nhs.uk, leo.xinyue.zhang@kcl.ac.uk, **Joint first authorship*

Abstract

The development of NLP models in the healthcare sector faces important challenges due to the limited availability of patient data, mainly driven by privacy concerns. This study proposes the generation of synthetic free-text medical reports, specifically focusing on the gastroenterology domain, to address the scarcity of specialised datasets, while preserving patient privacy. We fine-tune BioGPT on over 90 000 endoscopy reports and integrate Differential Privacy (DP) into the training process. 10 000 DP-private synthetic reports are generated by this model. The generated synthetic data is evaluated across multiple dimensions: similarity to real datasets, language quality, and utility in both supervised and semi-supervised NLP tasks. Results suggest that while DP integration impacts text quality, it offers a promising balance between data utility and privacy, improving the performance of a real-world downstream task. Our study underscores the potential of synthetic data to facilitate model development in the healthcare domain without compromising patient privacy.

1 Introduction

The development of computer-aided tools in medicine, including natural language processing (NLP), requires real patient data for model training. However, this development has been significantly limited due to the lack of availability of patient data due to privacy concerns, restricted access to hospital data, a scarcity of labeled data, barriers to sharing pretrained models, and a lack of capable computational resources in many healthcare settings (Wu et al., 2022). The lack of specialised datasets when developing NLP models can lead to biased or ungeneralizable models (Panch et al., 2019; Daneshjou et al., 2021). Recent literature highlights that open-source, synthetic datasets could mitigate data scarcity and lead to robust AI model training, particularly in NLP (Ive et al., 2020). However, very few studies tackle the generation of synthetic free text in the medical domain, with no known studies focusing on gastroenterology text reports.

While synthetic data presents a viable solution to dataset scarcity, ensuring the privacy of patient data in the original dataset used for training remains essential. Recent findings suggest that simply de-identifying the training set by removing names and unique identifiers is insufficient to prevent patient re-identification (Sarkar et al., 2024). Despite this, it is not common practice to include a robust data privacy framework when generating synthetic medical data (Begoli et al., 2018; Guan et al., 2021). To maintain stringent patient confidentiality, our approach incorporates Differential Privacy (DP), a framework that mathematically guarantees the level of inability to identify an individual's data within a dataset (Dwork et al., 2006). Our approach is motivated by the fact that only a limited number of academic papers investigate the application of differential privacy in the generation of synthetic data within healthcare (Klymenko et al., 2022).

The quality and utility of these generated reports must also be rigorously assessed to ensure their practical application in clinical settings. Utility, in our context, refers to the degree to which the synthetic data can be used to perform real-world tasks, such as text classification. It is crucial to compare how differential privacy impacts the quality and utility of synthetic data and whether it can be used to enhance performance on various tasks. These tasks can be supervised, such as text classification, unsupervised or semi-supervised, like Task Adaptive Pre-Training Tasks (TAPT).

2 Aims

- We create free text endoscopy reports generated with differential privacy by fine-tuning a medical domain GPT-based model.

Proceedings of the 23rd Workshop on Biomedical Language Processing, pages 14–24
August 16, 2024. ©2024 Association for Computational Linguistics

- We assess the similarity of DP-generated reports to the original patient data (training dataset), using a set of experiments that includes outliers re-generation.

- We aim to quantify the potential quality reduction induced by DP by assessing the text quality of synthetic text reports with and without DP.

- We assess and compare the utility of DP-generated synthetic reports across supervised and semi-supervised tasks.

3 Related Work

Generating Synthetic Medical Notes with Differential Privacy. The generation of synthetic text data with Differential Privacy (DP) is an emerging field with limited research. While (Yue et al., 2022) has provided a comprehensive framework for generating synthetic text data with DP, none of the investigated datasets include medical data. Sarkar et al. (2024) also propose a framework for synthetic data generation with DP.

Privacy Assessment of Synthetic Medical Notes and Text Similarity. Numerous studies have shown that Large Language Models (LLMs) and Generative Models can efficiently produce synthetic text reports (Melamud and Shivade, 2019; Abdollahi et al., 2021; Li et al., 2021; Guan et al., 2018; Tang et al., 2023). However, the privacy aspect of the synthetic data is often overlooked or relies on simple downstream analyses. A common practice in studies involving synthetic patient text data, especially in those not using DP, is to employ metrics like the Hamming distance or the Levenshtein distance to assess the privacy level of generated reports. These methods measure how closely synthetic data can be linked to their original counterparts. A threshold is established, and synthetic reports are considered vulnerable if their distances fall below this threshold (Ghosheh et al., 2024; Yan et al., 2021; Zhang et al., 2020).

Text Quality Assessment of Synthetic Medical Notes. The text quality of synthetic reports is often evaluated on a per-report basis, using metrics (Zhou et al., 2023) such as BLEU, ROUGE, or BERTScore (Zhang et al., 2019). However, these measurements require a set of references with which to compare the synthetic text for report-level evaluation. For synthetic text that does not have references, research tends to measure the distribution similarity on a corpus level using metrics such as generation perplexity (Fan et al., 2018), self-BLEU (Zhu et al., 2018) or Mauve (Pillutla et al., 2021). However, these methods do not give a score for each single report. In our recent work (in process of publication), we trained a language quality model that scores any generated report without the need for a reference text. The model is trained on a dataset that is corrupted by shuffling and inflection of real text. The model learns the mapping from each corrupted text, which can be seen as a proxy for model output, to a quality score, which is calculated by comparing the corrupted text with its original, unaltered form. This approach has proven to align well with human judgment and is effective in distinguishing higher-quality real texts from synthetic counterparts of lower language quality based on the generated scores.

Synthetic Data Utility. Sarkar et al. (2024) assess the utility of DP-generated synthetic reports through downstream tasks. However, their downstream tasks focus on ICD-10 code classification models trained on synthetic data, which differs significantly from our study. We explore the utility of DP-generated synthetic data both in a supervised setting, using it for data augmentation, and in a semi-supervised setting, employing it for further pre-training of the classifiers, which has not previously been documented in the literature.

4 Methods

A summary of the overall pipeline is depicted in Figure 1.

4.1 Data Access

Inclusion criteria Endoscopy reports were extracted from the electronic patient records (EPR) of St Thomas' Hospital in London. Data acquisition was authorised through an institutional board review. The dataset includes the following unfiltered procedures: Colonoscopy, Gastroscopy, Endoscopic ultrasound (EUS), Sigmodoiscopy and Endoscopic retrograde cholangiopancreatography (ERCP). The records spanned from January 2017 to October 2023.

Exclusion criteria To ensure patient privacy and comply with UK health service national data

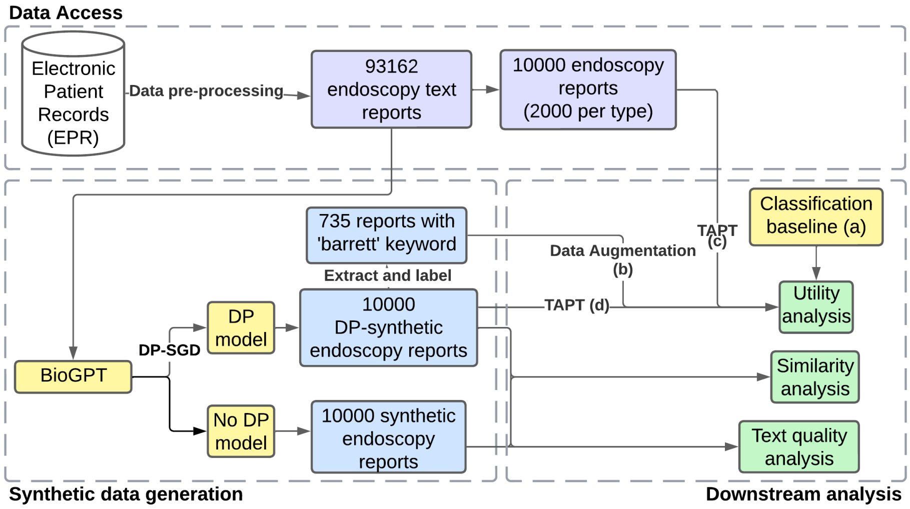

Data Access
Electronic Patient Records (EPR)
Data pre-processing
93162 endoscopy text reports
10000 endoscopy reports (2000 per type)
735 reports with 'barrett' keyword
Extract and label
Data Augmentation (b)
TAPT (c)
Classification baseline (a)
DP model
DP-SGD
10000 DP-synthetic endoscopy reports
TAPT (d)
Utility analysis
BioGPT
Similarity analysis
No DP model
10000 synthetic endoscopy reports
Text quality analysis
Synthetic data generation
Downstream analysis

Figure 1: Summary of the methodology of the study. (a), (b), (c), (d) correspond to the experiments described in Section 4.8.

opt-out policy (nhs, 2023), all individuals who had explicitly opted out of having their data used for research purposes were excluded from the study.

A total of **93162** reports were included, representing a diverse range of gastrointestinal conditions and providing a comprehensive dataset for the generation of synthetic endoscopy text reports.

4.2 Data Pre-processing and De-identification

Our dataset was anonymized, as required by NHS England and the UK Information Commissioner's Office (ICO)'s anonymisation code of practice[1]. Direct identifiers (such as names, addresses, and contact numbers) and indirect identifiers (such as clinician names and dates) were systematically removed from the dataset without replacement using regular expressions.

The remaining pre-processing was performed using the EndoMineR package[2], a tool designed for the analysis of free-text in endoscopy reports (Zeki, 2018). The package enabled the extraction of relevant sections from endoscopy reports.

4.3 Differential Privacy

Differential Privacy ensures that the output of a randomized function applied to a dataset is statistically indistinguishable, up to a specified degree of error, regardless of whether any single individual's data is included in the dataset or not. The notion of (ϵ, δ)-differential privacy, as defined by (Dwork et al., 2006) and further elaborated in recent literature (Yue et al., 2022) is as follows:
A randomized function F provides (ϵ, δ)-differential privacy if for all datasets D_1 and D_2 differing on at most one element, and for all subsets S of the possible outputs of F:

$$\Pr[F(D_1) \in S] \leq e^{\epsilon} \times \Pr[F(D_2) \in S] + \delta \quad (1)$$

ϵ (epsilon), also called *privacy budget*, is a small non-negative parameter that quantifies the strength of the privacy guarantee. δ (delta), typically close to zero, represents the small probability that the ϵ-differential privacy guarantee may be exceeded. ϵ is a key feature of differential privacy: a lower value guarantees greater privacy but generally reduces the utility of the generated data. In this study, we set ϵ = 4, and δ to

$$\delta = \frac{1}{N \cdot \log N} \quad (2)$$

with N being the number of training samples. These values have proved to guarantee a robust level of privacy in previous studies (Yue et al., 2022).

4.4 Fine-tuning Bio-GPT with DP for Text Generation

To generate the synthetic reports, we fine-tuned BioGPT (generative pre-trained transformer for biomedical text generation) (Luo et al., 2022) on our dataset. BioGPTis a transformer-based

[1] https://transform.england.nhs.uk/information-governance/guidance/artificial-intelligence/

[2] https://docs.ropensci.org/EndoMineR/

sequence-to-sequence model that relies on the GPT-2 architecture and comprises 345 million parameters. BioGPT has been pre-trained on over 15 million PubMed abstracts and has demonstrated increased performance compared to its general domain counterparts for downstream tasks when fine-tuned on biomedical data (Turbitt et al., 2023). We conducted the fine-tuning using the Hugging-Face Transformers library (Wolf et al., 2020) along with the Meta AI Opacus library to implement Differential Privacy (Yousefpour et al., 2021). Opacus ensures privacy by applying Differentially Private Stochastic Gradient Descent (DP-SGD), which clips the gradients' L2 norm and adds Gaussian noise to maintain the privacy of the model parameters during the training process.

All the experiments were executed on an NVIDIA DGX server running GNU/Linux 5.4.0-125-generic x86_64, with MLflow integrated into the CSC MLOPs[3] environment to ensure experiment reproducibility, collaboration, and scalability. The specific hyperparameters considered included learning rate, batch size, number of epochs, maximum sequence length, and temperature for the generation process. Fine-tuning was performed with causal language modeling (CLM) objective. The hyperparameter values are described in Table 1.

Table 1: BioGPT fine-tuning hyperparameters

Hyperparameter	Value
Batch size per GPU	16
Learning rate	1e-5
Number of training epochs	25
Epsilon ϵ	4

4.5 Generation of DP synthetic endoscopy text reports

4.5.1 Generation process

Control codes were used to steer the generation of specific report types (e.g. OGD, colonoscopy, EUS, ERCP) by the fine-tuned BioGPT model. This technique facilitated the targeted generation of texts according to the different kinds of endoscopic procedures. The input format for this generation process can be conceptualised as: Input = Control Code + Separator + Initial Context.

We built a text generation pipeline, refined through iterative clinician feedback to optimise the authenticity and relevance of the generated reports. The key generation hyperparameters were set as follows:

- **Length Constraints**: The generated reports' maximum length was set to 400 words to reflect the typical lengths of endoscopy reports.

- **Temperature**: This parameter controls the randomness of the generated output by scaling the logits before applying softmax, defined by the equation:

$$P(token) = \frac{\exp(\frac{\log(o_i)}{T})}{\sum_j \exp(\frac{\log(o_j)}{T})} \tag{3}$$

Here, T represents the temperature, o_i the logits, and $P(token)$ the probability of selecting *token* as the next token. The temperature was set to $T = 0.9$ based on recommendations from domain experts to balance creativity with accuracy.

- **No Repeat Ngram Size**: This parameter was established at 4 to prevent the repetition of any four-word sequence within the generated text, enhancing the uniqueness and readability of the reports.

4.6 Assessment of the Similarity of DP-generated reports

While DP theoretically offers a high level of privacy, its practical effectiveness in safeguarding patient data still requires empirical verification. Recent work has indeed shown that misuses of DP in Deep Learning have often led to limited actual privacy (Blanco-Justicia et al., 2022). As discussed in Section 3, the Hamming and Levensthein distances are often used to assess the privacy of generated reports. However, considering the varying lengths and content complexity of medical reports, these methods may not fully capture the nuances of text similarity, and therefore, may not appropriately assess the privacy of generated reports.

ROUGE-L (Lin, 2004) is a metric which is particularly valuable for evaluating text similarity in generation tasks where structural coherence and order of information are crucial. Unlike BLEU, which focuses on precision by measuring how many words in the generated text appear in the reference texts, ROUGE-L relies on recall, assessing how much of the original report is captured in the generated text. Specifically, ROUGE-L

[3] https://github.com/GSTT-CSC/MLOps

relies on the longest common sub-sequence (LCS) shared between the generated and reference texts, providing a measure of the longest sequence of words appearing in both texts in the same order. ROUGE-L is a normalized metric, therefore making it robust to length variations between the original and generated reports. Our approach to assess the similarity of DP-generated synthetic reports is the following:

1. **Distribution Analysis of ROUGE-L scores**: We compute the ROUGE-L score between each synthetic report and each of the original patient reports. We then keep the highest ROUGE-L score for each of the synthetic reports and compute the resulting distribution. This process is done both for synthetic data generated with and without DP. We then compare DP and non-DP distributions to assess the impact of DP on text similarity and privacy enhancement.
2. **Inclusion of distinctive outliers**: 34 outliers with unique combinations of endoscopic findings are included in the training set. These outliers are text reports of typical length, containing phrasings or combinations of medical conditions not typically found in endoscopy reports. Developed in collaboration with a gastroenterologist, they are distinctive enough that reproducing them directly in synthetic reports could result in patient re-identification.

4.7 Evaluating the Text Quality of DP-generated reports

To evaluate the language quality of generated reports with and without DP, we use the language scoring model introduced in Section 3 (in process of publication), this model takes an individual report as input and assign a score to it based on its language quality. The score ranges from 0 to 1, with higher scores indicating better language quality of the text.

4.8 Evaluating the Utility of DP-generated reports in Downstream Tasks

4.8.1 Baseline Description and Evaluation Metrics

The utility of the generated synthetic reports was evaluated by trying to improve a clinically relevant 4-class classification problem. This involves categorising endoscopy free-text reports based on the length of an endoscopically detectable premalignant lesion: Barrett's Oesophagus (BO) (Fitzgerald et al. (2014); Hameeteman et al. (1989)). The categories are: Long, Short, No Barrett's, and Insufficient, relating to the detection of a long or short segment of BO, a definite lack of detection of BO, or an insufficient description, respectively. The baseline model, detailed in Table 3, is a BERT-based transformer with a linear layer for classification, currently used in clinical practice. It was trained with optimized hyperparameters described in the Appendix (4).

This baseline (Figure 1.a) will be compared against three distinct approaches (Figure 1.b,c,d), as detailed in the subsequent sections (4.8.2, 4.8.3). Given the slightly imbalanced nature of the original training set and the varying clinical relevance of the classes, per-class metrics such as AUC-ROC and F1-Score were recorded. The performance of the baseline and subsequent models was assessed on a test set, using an 80/20 stratified split for each random seed. Results were averaged across three random seeds to ensure robustness.

4.8.2 Synthetic Data Augmentation

The first approach to enhancing the baseline model involved augmenting the training set with 735 DP synthetic gastroscopy reports specifically related to Barrett's Oesophagus. Each report was manually annotated by a domain expert. An overview of the class distributions before synthetic data augmentation is presented in the Appendix (5). The BERT-based classifier was then retrained using the augmented dataset while maintaining the same hyperparameters to allow for a direct comparison of performance changes.

4.8.3 Task-Adaptive Pretraining with Synthetic Data

In the current NLP landscape, LLMs are typically pre-trained on general domain dataset using tasks such as Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) (Devlin et al., 2019; Liu et al., 2019). Although these models exhibit strong performance across various downstream tasks, research has shown that continued pre-training on domain-specific texts can further enhance their effectiveness (Gururangan et al., 2020; Li et al., 2023; Shi et al., 2023; Margatina et al., 2022). In this study, the target domain is gastroenterology text reports. Our second experiment involves task-adaptive pre-training (TAPT) of the baseline model using two separate datasets:

1. Synthetic Data: 10,000 synthetic endoscopy text reports generated with our Differential Privacy pipeline.

2. Real Patient Data: 10,000 endoscopy text reports extracted from the hospital's Electronic Patient Records (EPR) that were not part of the training or evaluation sets of the baseline model. These reports were selected to match the variety and number found in the synthetic dataset.

We used these datasets to conduct domain-adaptive pre-training on the pre-trained BERT model before fine-tuning it for classification tasks and evaluation. Domain-adaptative pre-training was performed using a Masked Language Modeling (MLM) objective. The aim of this experiment is to quantify the utility of synthetic data in comparison to original patient data in terms of enhancing the model's performance on downstream tasks. In both cases, TAPT experiments were conducted using the hyperparameters listed in Table (2).

Table 2: Hyperparameters for TAPT Experiments

Hyperparameter	Value
Number of Pretraining Epochs	100
Number of Fine-tuning Epochs	8
Pretraining Learning Rate	1e-4
Fine-tuning Learning Rate	6.85e-5
Warming Up Steps	1000

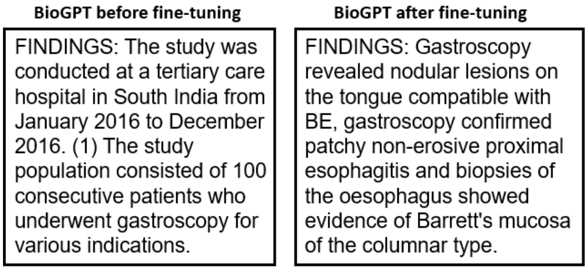
5 Results

5.1 Generation of DP Synthetic Reports

10,000 DP synthetic reports were generated using the input defined in 4.5. We compare the output of the model, before and after fine-tuning, as depicted in Figure 2.

We observe that before fine-tuning, the output of the model resembles a PubMed abstract that mentions the report type, gastroscopy in this case. However, it appears like a study rather than an individual patient's endoscopy report. After fine-tuning, the model's output presents as a well-formatted gastroscopy report, with findings related to Barrett's Oesophagus.

5.2 Text Similarity Analysis



BioGPT before fine-tuning

FINDINGS: The study was conducted at a tertiary care hospital in South India from January 2016 to December 2016. (1) The study population consisted of 100 consecutive patients who underwent gastroscopy for various indications.

BioGPT after fine-tuning

FINDINGS: Gastroscopy revealed nodular lesions on the tongue compatible with BE, gastroscopy confirmed patchy non-erosive proximal esophagitis and biopsies of the oesophagus showed evidence of Barrett's mucosa of the columnar type.

Figure 2: Comparison of BioGPT output for the generation of synthetic reports, before (left) and after (right) fine-tuning.

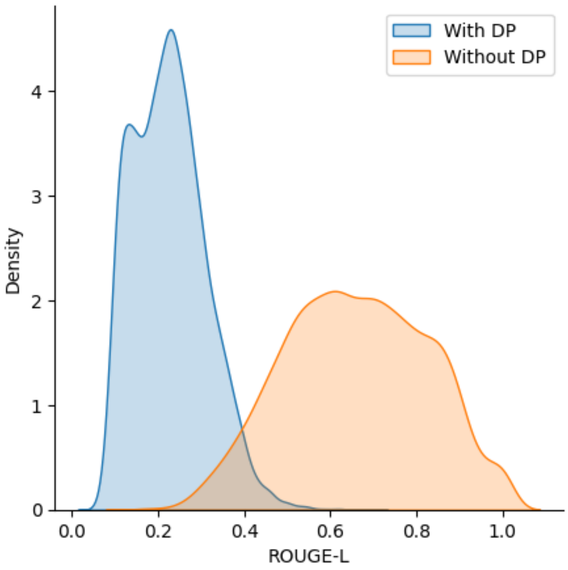


Figure 3: Distributions of maximum ROUGE-L scores between original and synthetic reports, with (left) and without (right) Differential Privacy.

The distributions of maximum ROUGE-L scores between original and synthetic reports (both DP and non DP) are depicted in Figure 3. We observe a significant shift in distribution between the synthetic reports generated with Differential Privacy compared to those generated without it. The ROUGE-L scores for the DP-generated reports span from 0.058 to 0.690, with an average of 0.226, indicating that the DP-generated reports significantly differ from the training set. In contrast, the ROUGE-L scores for the non-DP-generated reports span from 0.165 to 1.0, with an average of 0.660, indicating that some generated reports are highly similar to the training set.

After careful review of the synthetic reports generated with and without Differential Privacy (DP) with a domain expert, we confirmed that no outlier was directly regenerated in either the synthetic DP or non-DP reports.

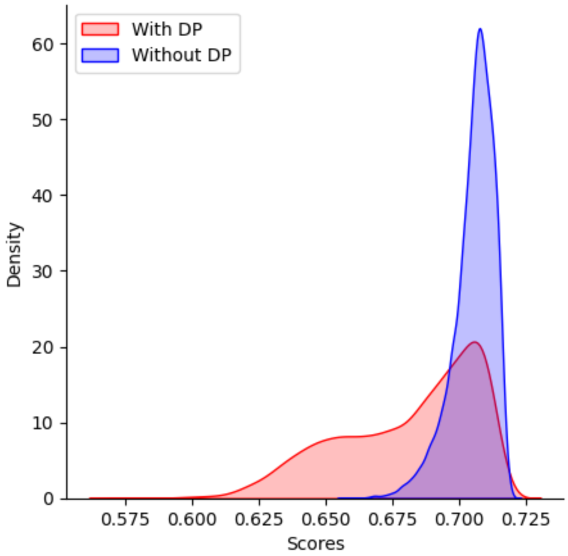


Figure 4: Distributions of synthetic data language quality scores with (left) and without (right) Differential Privacy.

5.3 Language Quality Evaluation

Figure 4 shows that synthetic reports generated without DP exhibit language scores centred around higher values. While synthetic reports generated with DP have a similar tendency, their scores are more distributed toward the lower end, resulting in a broader, shorter-tailed distribution. Despite this variation, the overall spread remains relatively constrained, indicating a slight reduction in the text quality of reports generated with DP.

5.4 Utility Evaluation

The results of the three utility evaluation experiments are summarized in Table 3. The optimized baseline model already achieves high performance across all classes, with the ‘Long’ class showing the highest average F1-score (0.958) and the ‘Insufficient’ class the lowest (0.822).

The most notable conclusion from these experiments is that task-adaptive pretraining considerably improves the baseline performance for all classes, especially for the ‘Insufficient’ class, which sees an F1-score increase of 0.089. The ‘Long’ class, which already performed well, also shows an improvement of 0.022.

The primary goal of this paper is to assess the utility of synthetic data generated with differential privacy. As expected, the baseline improvement using DP synthetic data is not as significant as with real patient data, likely due to the set privacy level (epsilon = 4). However, TAPT using synthetic data with DP still enhances the F1 scores across all classes, with the ‘Insufficient’ class showing the most significant improvement of 0.034. The Long class, despite its high performance, also showed an improvement while performing TAPT with DP synthetic data, with an F1 score improvement of 0.003.

Data augmentation with labelled synthetic DP text reports also improved performance across most classes, though results were more inconsistent. This variability may be due to the limited number of additional annotated reports, as the annotation process is time-consuming and constrained by a shortage of expert annotators.

6 Discussion and Conclusion

We have fine-tuned a pre-trained large language model with Differential Privacy to generate privacy-preserved synthetic endoscopy reports. We leveraged a highly specific in-house training set of over 90,000 endoscopy free-text reports. Using our pipeline, we generated a set of 10,000 diverse synthetic endoscopy reports, available for further research on a per-query basis. The utility of the synthetic reports was assessed by attempting to improve a clinically useful high performing classification baseline. The synthetic reports were used to augment the training set of the baseline and to pretrain the baseline classifier using a task-adaptive pretraining framework. A pre-training experiment with real patient data was also conducted for direct comparison.

Table 3 demonstrates that DP-generated synthetic data can significantly improve the performance of a real-world downstream task. Specifically, our study demonstrates the superiority of TAPT methods. It is important to highlight that, in comparison to supervised-learning approaches, TAPT does not require additional labeled data points, which significantly reduces the need for data annotation resources, a primary bottleneck in developing robust supervised learning models. However, the best-performing model remains the one pre-trained with real patient data.

The privacy preservation of DP-generated reports is quantified by assessing their similarity to the original data compared to that of synthetic data generated without DP. We observed an important difference in ROUGE-L scores between the DP and non-DP generated synthetic reports. Therefore, we can conclude that, in our study, DP effectively pre-

Table 3: Comparison of model performance across different approaches, including the baseline BERT-based model (a), synthetic data augmentation with 735 differentially private reports (b), task-adaptive pretraining with 10,000 real patient reports (c), and task-adaptive pretraining with 10,000 differentially private synthetic reports (d).

Approach	**Long**		**No Barretts**		**Short**		**Insufficient**	
	AUROC	**F1**	**AUROC**	**F1**	**AUROC**	**F1**	**AUROC**	**F1**
(a) Baseline	$0.988_{0.008}$	$0.958_{0.005}$	$0.992_{0.005}$	$0.924_{0.020}$	$0.984_{0.012}$	$0.925_{0.013}$	$0.979_{0.013}$	$0.822_{0.034}$
(b) Data augmentation	$0.997_{0.001}$	$0.961_{0.016}$	$0.982_{0.005}$	$0.914_{0.017}$	$0.987_{0.004}$	$0.944_{0.010}$	$0.967_{0.031}$	$0.798_{0.029}$
(c) TAPT with real data	$0.997_{0.002}$	$0.980_{0.003}$	$0.997_{0.002}$	$0.961_{0.014}$	$0.994_{0.002}$	$0.967_{0.010}$	$0.988_{0.004}$	$0.911_{0.022}$
(d) TAPT with DP synthetic	$0.991_{0.005}$	$0.961_{0.025}$	$0.989_{0.006}$	$0.935_{0.025}$	$0.987_{0.007}$	$0.945_{0.024}$	$0.980_{0.005}$	$0.856_{0.053}$

vents the replication of sensitive training samples. A set of outliers was also introduced in the training set, and we concluded that unique clinical findings could not be regenerated by the models.

6.1 Limitations and Future Work

While this study aims to explore various methods to assess the utility and privacy of DP-generated endoscopy text reports, we acknowledge several limitations. First, we used a fixed value of the privacy parameter epsilon throughout this study. Future work should assess the impact of epsilon on the privacy-utility trade-off.

We have compared the performance of a classification baseline to several approaches leveraging generated synthetic reports (Table 3), but we have not compared the baseline to a classifier solely trained on synthetic data due to the limited availability of high-quality annotation resources. Moreover, future work should also consider comparing our approach with other existing methods of synthetic data generation and privacy protection to provide a more comprehensive evaluation.

It is also important to highlight that the generative models were fine-tuned on endoscopy data from a London hospital, which might introduce population bias or an over-representation of specific endoscopic conditions due to the local context.

In this study, we assessed the text quality of the generated text; however, no evaluation of clinical relevance was conducted. Clinical accuracy is essential for specific downstream tasks as it ensures the medical reliability of the generated reports and prevents confusion. Users should remain mindful of this aspect when using our synthetic reports.

7 Acknowledgements

We would like to express our gratitude to the Clinical Scientific Department at Guy's and St' Thomas' NHS Foundation Trust for providing us with compute resources.

8 Availability of Generated Reports and Code

The generated synthetic endoscopy text reports, along with the corresponding code used for their generation, are made available on a per-request basis.

References

2023. Nhs national data opt-out.

Mahdi Abdollahi, Xiaoying Gao, Yi Mei, Shameek Ghosh, Jinyan Li, and Michael Narag. 2021. Substituting clinical features using synthetic medical phrases: Medical text data augmentation techniques. *Artificial Intelligence in Medicine*, 120:102167.

Edmon Begoli, Kris Brown, Sudarshan Srinivas, and Suzanne Tamang. 2018. Synthnotes: A generator framework for high-volume, high-fidelity synthetic mental health notes. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 951–958.

Alberto Blanco-Justicia, David Sánchez, Josep Domingo-Ferrer, and Krishnamurty Muralidhar. 2022. A critical review on the use (and misuse) of differential privacy in machine learning. *ACM Comput. Surv.*, 55(8).

Roxana Daneshjou, Mary P. Smith, Mary D. Sun, Veronica Rotemberg, and James Zou. 2021. Lack of transparency and potential bias in artificial intelligence data sets and algorithms. *JAMA Dermatology*, 157(11):1362.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. *Calibrating Noise to Sensitivity in Private Data Analysis*, page 265–284. Springer Berlin Heidelberg.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*.

Rebecca C Fitzgerald, Massimiliano di Pietro, Krish Ragunath, Yeng Ang, Jin-Yong Kang, Peter Watson, Nigel Trudgill, Praful Patel, Philip V Kaye, Scott Sanders, Maria O'Donovan, Elizabeth Bird-Lieberman, Pradeep Bhandari, Janusz A Jankowski, Stephen Attwood, Simon L Parsons, Duncan Loft, Jesper Lagergren, Paul Moayyedi, Georgios Lyratzopoulos, John de Caestecker, and British Society of Gastroenterology. 2014. British society of gastroenterology guidelines on the diagnosis and management of barrett's oesophagus. *Gut*, 63(1):7–42.

Ghadeer O. Ghosheh, Jin Li, and Tingting Zhu. 2024. A survey of generative adversarial networks for synthesizing structured electronic health records. *ACM Computing Surveys*, 56(6):1–34.

Jiaqi Guan, Runzhe Li, Sheng Yu, and Xuegong Zhang. 2018. Generation of synthetic electronic medical record text. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 374–380.

Jiaqi Guan, Runzhe Li, Sheng Yu, and Xuegong Zhang. 2021. A method for generating synthetic electronic medical record text. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(1):173–182.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *Preprint*, arXiv:2004.10964.

W Hameeteman, GNJ Tytgat, HJ Houthoff, and JG Van Den Tweel. 1989. Barrett's esophagus; development of dysplasia and adenocarcinoma. *Gastroenterology*, 96(5):1249–1256.

Julia Ive, Natalia Viani, Joyce Kam, Lucia Yin, Somain Verma, Stephen Puntis, Rudolf N. Cardinal, Angus Roberts, Robert Stewart, and Sumithra Velupillai. 2020. Generation and evaluation of artificial mental health records for natural language processing. *npj Digital Medicine*, 3(1).

Oleksandra Klymenko, Stephen Meisenbacher, and Florian Matthes. 2022. Differential privacy in natural language processing the story so far. In *Proceedings of the Fourth Workshop on Privacy in Natural Language Processing*, pages 1–11, Seattle, United States. Association for Computational Linguistics.

Jianfu Li, Yujia Zhou, Xiaoqian Jiang, Karthik Natarajan, Serguei Vs Pakhomov, Hongfang Liu, and Hua Xu. 2021. Are synthetic clinical notes useful for real natural language processing tasks: A case study on clinical entity recognition. *Journal of the American Medical Informatics Association*, 28(10):2193–2201.

Shiyang Li, Semih Yavuz, Wenhu Chen, and Xifeng Yan. 2023. Task-adaptive pre-training and self-training are complementary for natural language understanding. *Preprint*, arXiv:2109.06466.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

Katerina Margatina, Loic Barrault, and Nikolaos Aletras. 2022. On the importance of effectively adapting pretrained language models for active learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 825–836, Dublin, Ireland. Association for Computational Linguistics.

Oren Melamud and Chaitanya Shivade. 2019. Towards automatic generation of shareable synthetic clinical notes using neural language models. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 35–45, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Trishan Panch, Heather Mattie, and Leo Anthony Celi. 2019. The "inconvenient truth" about AI in healthcare. *npj Digital Medicine*, 2(1).

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828.

Atiquer Rahman Sarkar, Yao-Shun Chuang, Noman Mohammed, and Xiaoqian Jiang. 2024. De-identification is not always enough. *Preprint*, arXiv:2402.00179.

Zhengxiang Shi, Francesco Tonolini, Nikolaos Aletras, Emine Yilmaz, Gabriella Kazai, and Yunlong Jiao. 2023. Rethinking semi-supervised learning with language models. *Preprint*, arXiv:2305.13002.

Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. 2023. Does synthetic data generation of llms help clinical text mining? *Preprint*, arXiv:2303.04360.

Oisn Turbitt, Robert Bevan, and Mouhamad Aboshokor. 2023. Mdc at biolaysumm task 1: Evaluating gpt models for biomedical lay summarization. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Honghan Wu, Minhong Wang, Jinge Wu, Farah Francis, Yun-Hsuan Chang, Alex Shavick, Hang Dong, Michael T. C. Poon, Natalie Fitzpatrick, Adam P. Levine, Luke T. Slater, Alex Handy, Andreas Karwath, Georgios V. Gkoutos, Claude Chelala, Anoop Dinesh Shah, Robert Stewart, Nigel Collier, Beatrice Alex, William Whiteley, Cathie Sudlow, Angus Roberts, and Richard J. B. Dobson. 2022. A survey on clinical natural language processing in the united kingdom from 2007 to 2022. *npj Digital Medicine*, 5(1).

Chao Yan, Ziqi Zhang, Steve Nyemba, and Bradley A Malin. 2021. Generating electronic health records with multiple data types and constraints.

Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. 2021. Opacus: User-friendly differential privacy library in PyTorch. *arXiv preprint arXiv:2109.12298*.

Xiang Yue, Huseyin A. Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. 2022. Synthetic text generation with differential privacy: A simple and practical recipe. *arXiv preprint*.

Sebastian S. Zeki. 2018. Endominer for the extraction of endoscopic and associated pathology data from medical reports. *Journal of Open Source Software*, 3(24):701.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

Ziqi Zhang, Chao Yan, Thomas A Lasko, Jimeng Sun, and Bradley A Malin. 2020. Synteg: a framework for temporal structured electronic health data simulation. *Journal of the American Medical Informatics Association*, 28(3):596–604.

Yongxin Zhou, Fabien Ringeval, and François Portet. 2023. A survey of evaluation methods of generated medical textual reports. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 447–459, Toronto, Canada. Association for Computational Linguistics.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.

A Classification baseline details

Table 4: Classification Baseline Hyperparameters

Hyperparameter	Search Space	Baseline model value
Batch size	{16, 32, 64}	32
Learning rate	[1e-6, 1e-3]	6.85e-5
Number of training epochs	[1, 10]	8
Warming steps fraction	[0.1, 0.5]	0.4

Table 5: Class Distributions Before and After Data Augmentation

Class	Before Augmentation	After Augmentation
Insufficient	279	475
Long	1,649	1,688
Short	1,901	1,901
No Barrett's	288	788
Total Reports	4,117	4,852