Generating Synthetic Free-text Medical Records with Low Re-identification Risk using Masked Language Modeling

Samuel Belkadi[1], **Libo Ren**[2], **Nicolo Micheletti**[3],
Lifeng Han[2,4*], **Goran Nenadic**[2]
[1] Department of Engineering, University of Cambridge, UK
[2] Department of Computer Science, The University of Manchester, UK
[3] Department of Computer Science and Technology, Tsinghua University, China
[3] LIACS & LUMC, Leiden University, Leiden, NL * *corresponding author*

Abstract

The abundance of medical records holds great promise for enhancing healthcare and advancing biomedical research. However, due to *privacy* constraints, access to such data is typically limited to internal use. Recent studies have attempted to overcome this challenge by generating synthetic data through Causal Language Modelling. Yet, this approach often fails to ensure patient anonymity and offers limited control over output diversity—unless additional computational cost is introduced. In response, we propose a method for generating synthetic free-text medical records based on *Masked Language Modelling*. Our approach retains key medical details while introducing variability in the generated texts and reducing the risk of patient re-identification. With a relatively lightweight architecture of approximately 120 million parameters, the system ensures low inference costs. Experimental results show that our method produces high-quality synthetic data, achieving a HIPAA-compliant PHI recall of 96% and a re-identification risk of only 3.5%. Furthermore, downstream evaluations reveal that models trained on the synthetic data perform comparably to those trained on real-world data. Our trained models are publicly available on Github as SYNDEIDMLM (at `https://github.com/SamySam0/SynDeidMLM`) (meaning **syn**thetic and **de-id**entified data generation using **MLM**).

1 Introduction

The widespread adoption of electronic medical record (EMR) systems has led to the accumulation of substantial volumes of patient data, offering considerable opportunities to improve healthcare delivery and biomedical research (Beam and Kohane, 2018; Shah et al., 2018). However, access to such data is heavily restricted due to privacy concerns, aiming to safeguard patients' personal information (Price and Cohen, 2019). One promising alternative is the use of synthetic data, which allows the generation of documents—such as discharge summaries—that retain medically relevant information while reducing privacy risks. This approach enables broader data sharing for purposes like healthcare system testing (Tucker et al., 2020), medical training (Li et al., 2024), and the development of artificial intelligence tools (Belkadi et al., 2023).

Much of the previous work on synthetic medical text generation has primarily relied on *Causal Language Modelling* (CLM), with comparatively limited attention paid to *Masked Language Modelling* (MLM). While CLM approaches have shown promise in replicating the statistical patterns of real-world clinical data, they present several challenges—specifically, ensuring privacy protection, managing diversity in generated texts, and mitigating the computational cost of generation.

Recent findings by Micheletti et al. (2024) demonstrate that Masked Language Modelling can perform on par with Causal Language Modelling across a wide range of synthetic generation tasks, while offering greater flexibility in controlling contextual information. Building on these insights, this paper introduces a system designed to generate synthetic English-language medical texts—such as discharge notes, admission records, and doctor-to-doctor communications—using Masked Language Modelling. The system integrates a cutting-edge de-identification tool capable of automatically detecting protected health information (Radhakrishnan et al., 2023), thereby removing the need for manual pre-processing. It also incorporates two named entity recognition (NER) models to help retain essential clinical information and strike a balance between diversity and fidelity in the generated output. Importantly, the system is based on an encoder-only, non-autoregressive architecture, significantly reducing both its size and inference cost. The code will be released for public access.

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), pages 200–206

2 Related Work

In their recent study, Yan et al. (2024) proposed a Generative Adversarial Network (GAN) to produce synthetic electronic health records. While effective in some respects, their method struggled to manage the similarity between synthetic and original data and failed to accurately capture temporal dependencies within medical histories. Building on similar techniques, Kasthurirathne et al. (2021) presented a system for generating synthetic medical records with a low risk of re-identification. Despite encouraging results, the authors noted that limited diversity in the generated samples reduced their usefulness for tasks like oversampling. They also assumed that synthetic generation alone sufficiently mitigates re-identification risk, signalling the continued need for explicit de-identification mechanisms. In one of the most recent contributions to synthetic medical data research, Falis et al. (2024) assessed GPT-3.5's ability to generate discharge summaries. Their findings revealed that GPT-3.5 often closely reproduced input concepts, thereby heightening the risk of re-identification. Additionally, the generated texts were often unnatural, omitting key medical details while introducing irrelevant or misleading information. Clinicians involved in the evaluation acknowledged the presence of correct content but highlighted weaknesses in narrative structure, variety, and supporting details. Another concern raised was the model's lack of data governance, as it is not maintained by the institution that owns the original data. Taken together, these studies highlight common challenges in synthetic medical text generation: persistent *privacy* concerns and limited *control* over output variability. In response, our work advocates for the use of Masked Language Modelling, which offers enhanced control over the content being generated while reducing privacy risks and maintaining lower computational costs.

3 System Design

The system architecture, illustrated in Figure 1, is designed to generate synthetic medical records—including discharge summaries, admission notes, and clinician correspondence—through a two-stage pipeline: a *Masker* and a *Mask-Filling System*. The Masker identifies which parts of the text should be hidden or retained, outputting a partially masked version of the original text. The Mask-Filling System then replaces the masked sections with context-aware content, producing one or more synthetic variants of the original record.

3.1 The Masker

The Masker operates in three sequential stages: I) **De-identification.** The initial step involves detecting Protected Health Information (PHI) using Philter (Norgeot et al., 2020), a rule-based tool that relies on regular expressions to extract six PHI categories: DATE, ID, NAME, CONTACT, AGE, and LOCATION. According to the authors, Philter achieves high recall scores—99.46% on the UCSF dataset and 99.92% on the i2b2 2014 dataset. To our knowledge, it is the first certified de-identification system that enables the release of clinical notes for nonhuman-subject research, exempt from further IRB approval during the time period outlined by Radhakrishnan et al.. II) **Medical Entity Recognition.** In the second stage, a medical named entity recognition (NER) model identifies essential clinical terms that should remain unmasked in the synthetic output. For this, we fine-tuned a pre-trained Stanza model[1] on the i2b2-2010 dataset to extract three categories of entities: PROBLEM, TEST, and TREATMENT. This model achieved an F1 score of 88.13% on the test set. The system is also adaptable—users can substitute the model to target other entity types (e.g. medication names or dosages), and control the degree of masking applied to each entity class. III) **Part-of-Speech Tagging.** The final phase involves part-of-speech (POS) tagging using Stanza's POS tagger. Based on user-specified ratios, a subset of the tagged tokens is randomly masked to influence the diversity of the synthetic output. For instance, a setting like NOUN: 0.7, VERB: 0.5 would randomly mask 70% of nouns and 50% of verbs, while leaving other word types untouched.

3.2 The Mask-Filling System

Once the Masker has produced masked letters, the Mask-Filling System reconstructs them into synthetic texts using a masked language model (MLM) and a replacement algorithm. I) **MLM Model.** The MLM is an encoder-based model that predicts context-sensitive replacements for masked tokens by generating a probability distribution over possible vocabulary items. In our system, we employ Bio_ClinicalBERT—a version of BioBERT (Lee et al., 2020) fine-tuned on clinical texts from

[1] stanfordnlp.github.io/stanza/available_biomed_models.html

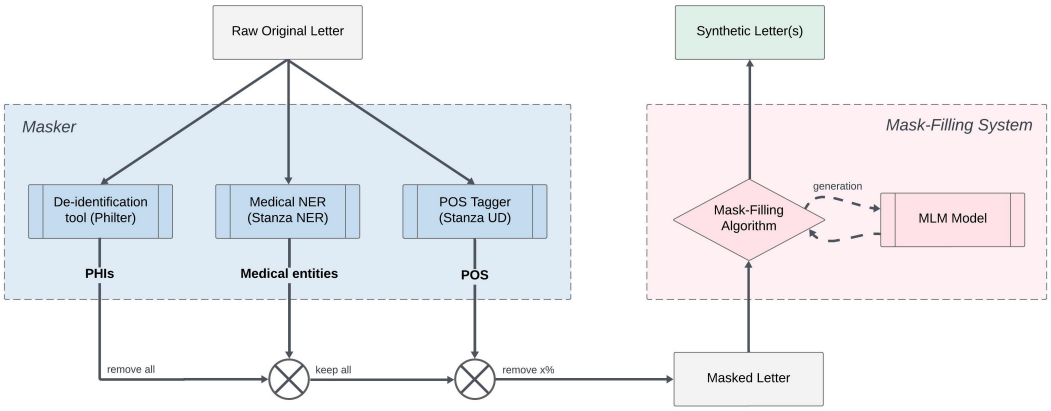

Figure 1: SYNDEIDMLM System Design: Masker and Mask-Filling two steps.

MIMIC III (Johnson et al., 2016). We further trained this model on a set of 790 clinical letters described in Section 4.1. While we did not compare it with alternative models, we encourage future research to explore different baselines. II) **Mask-Filling Algorithm.** This module prepares the masked input for the MLM model and chooses suitable replacements for each masked segment based on the model's predictions. We compare two strategies:

- *Simultaneous Chunk Filling*: In this approach, masked text is processed in chunks. Each chunk is passed through the MLM to generate probabilities for the masked elements. Replacements can be selected deterministically (using the highest probability term) or stochastically (by sampling from the distribution). While stochastic replacement enhances variation, it may also reduce fidelity by adding noise.

- *Iterative Mask Filling* (Kesgin and Amasyali, 2023): This method processes one masked entity at a time within a defined context window. As each masked token is resolved, it is replaced in the text, whereas upcoming masked items remain untouched until processed. This allows the model to focus on a specific context, improving generation quality and encouraging output diversity. Replacements, like in the previous approach, can be selected either deterministically or stochastically.

4 Experimental Setup

This section presents the dataset used to train and evaluate the MLM model, alongside the four system variants assessed in our experiments.

4.1 Datasets

The experiments are conducted using the i2b2 2014 shared task dataset for PHI de-identification (Stubbs and Uzuner, 2015; Stubbs et al., 2015), which includes 1,304 English clinical documents from 296 diabetic patients. These records comprise various note types such as discharge summaries, admission notes, and inter-physician communications. The dataset is pre-split into 790 training samples and 514 test samples. This resource offers a wide range of clinical scenarios and treatment contexts, making it well-suited for generating diverse synthetic outputs. All entries are annotated with HIPAA-compliant PHI labels. Furthermore, the dataset includes additional PHI sub-categories to reinforce privacy protection. The categories of annotations are Name, Profession, Location, Age, Contact, and IDs. Among these categories, only the following align with the official HIPAA-PHI definitions: NAME-PATIENT, LOCATION-STREET, LOCATION-CITY, LOCATION-ZIP, LOCATION-ORGANIZATION, AGE, DATE, CONTACT-PHONE, CONTACT-FAX, CONTACT-EMAIL, along with all sub-categories under ID.

4.2 Hyperparameters Tuning

The main parameters for system optimisation are the learning rate of the MLM model, the training batch size, the PHI's masking proportion, and the overall masking probability. We evaluate each instance using perplexity as it reflects the MLM model's confidence. We divided the data set into 80% and 20% for training and validation and retrained the model using the full data when the best parameter set is selected.

4.3 System Instances

We evaluated four system variants differing in masking ratios and mask-filling strategies:

- **System_S_0.5** and **System_S_0.7**: Both use Simultaneous Chunk Filling with stochastic sampling, mask all PHI, and retain all medical entities. They differ in lexical diversity, masking 50% and 70% of NOUNS, VERBS, and ADJECTIVES, respectively.

- **System_I_0.7** and **System_I_0.9**: These use Iterative Mask Filling with stochastic replacement, also masking all PHI while keeping medical entities. They apply 70% and 90% masking, respectively, for increased diversity.

The selected masking ratios are based on insights from Micheletti et al. (2024) but can be customised depending on the intended use case.

4.4 Evaluation Metrics

Our evaluation focuses on three main criteria: similarity to real data, utility, and privacy. **Lexical similarity** measures how well synthetic data reflects the structure and meaning of real text using ROUGE, BERTScore, and readability metrics (FRE[2], FKG[3], SMOG). It captures information retention, meaning preservation, and diversity, as well as how easy the text is to read. **Data utility** evaluates the effectiveness of synthetic data in training machine learning models. We assess this via a downstream NER task, comparing performance against models trained on real data (Belkadi et al., 2025; Micheletti et al., 2024). **Data privacy** is assessed by computing the F1 score for PHI removal (based on annotated HIPAA-PHI labels) and estimating re-identification risk.

5 Experiments and Results

5.1 Lexical Similarity Evaluation against References

The ROUGE and BERTScore results for the four system configurations are presented in Table 1. As expected, higher masking ratios tend to lower both ROUGE and BERTScore metrics, due to the increased noise introduced during generation. This confirms the trade-off between diversity and content fidelity, as previously discussed in Section 3.

[2]Flesch Reading Ease
[3]Flesch-Kincaid Grade

Additionally, systems that utilise *iterative* mask filling consistently outperform those using simultaneous filling in terms of lexical *similarity* to the original text. For instance, with a masking ratio of 0.7, the iterative approach achieves ROUGE scores that are over 3 points higher and BERTScore improvements exceeding 0.3. This emphasises the benefit of iterative replacement, where each masked term is filled in using richer contextual information—either from unmasked or previously generated tokens—thereby reducing ambiguity. Moreover, even at a high masking ratio of 0.9, iterative systems exhibit a relatively small drop in BERTScore (0.04), whereas ROUGE scores decline more significantly (by 4 points). This indicates that although the surface wording may deviate more from the original, the core meaning is largely preserved.

These observations are further supported by findings in Table 2, where lexical variations between real and synthetic datasets are assessed through word overlap comparisons. Overall, all system configurations were capable of striking a balance between variation and content retention. The results illustrate a clear diversity–fidelity trade-off, which can be fine-tuned by adjusting the masking ratio and choice of mask-filling strategy, offering flexibility for different downstream tasks.

5.2 Readability Evaluation against References

As shown in Table 3, the synthetic medical letters generally exhibit higher readability scores compared to their original counterparts. This improvement is more pronounced at higher masking ratios, likely because the MLM model tends to substitute masked terms with simpler and more frequently used vocabulary. When comparing the different system configurations, there is no single system that consistently outperforms the others in terms of readability. This outcome is beneficial, as it suggests that users have the freedom to adjust the balance between fidelity and diversity without negatively impacting the readability of the generated

	RGE1	RGE2	RGE-L	BERTS
Sys_S_0.5	0.861	0.760	0.852	0.729
Sys_S_0.7	0.828	0.703	0.815	0.674
Sys_I_0.7	0.852	0.732	0.841	0.706
Sys_I_0.9	0.826	0.686	0.811	0.668

Table 1: Lexical similarities of the generated synthetic letters against references on the testing dataset.

	Top 5	Top 20	Top 50	Top 100
System_S_0.5	3.848	15.593	38.420	78.670
System_S_0.7	3.601	14.607	35.971	73.695
System_I_0.7	3.712	15.095	37.233	76.093
System_I_0.9	3.537	14.551	35.510	72.298

Table 2: Average number of overlap between the top 5, 20, 50 and 100 words identified across the real and synthetic datasets, without stopwords. Additional results on lexical similarities.

	FRE	FKG	SMOG
System_S_0.5	64.024	7.647	10.823
System_S_0.7	65.091	7.466	10.696
System_I_0.7	63.792	7.707	10.878
System_I_0.9	64.294	7.636	10.832
References	61.597	8.06	11.067

Table 3: Readability scores of the generated synthetic letters against references on the testing dataset.

text.

5.3 Data Utility Evaluation

We investigate how effectively the synthetic data replicates key characteristics of real clinical text. To do so, we compare the performance of a medical NER model trained on synthetic data to that of a model trained on real data.

5.3.1 Downstream NER Task

For this task, the original test set is divided into new training and testing subsets. The real clinical letters are first passed through our system to generate synthetic equivalents. Both the real and synthetic texts are then processed using SciSpacy[4] (*en_ner_bc5cdr_md*), a named entity recognition model trained on the BC5CDR dataset, which achieves an F1 score of 0.84. This model identifies entities related to DISEASE and CHEMICAL terms. The extracted entities from the original and synthetic datasets are then used to create two distinct training sets. One SpaCy[5] model is trained using entities derived from the real data, while the other is trained on those extracted from the synthetic data. Both SpaCy models are then evaluated on the same test split. To study the effect of data *augmentation*, the experiment is repeated with twice as many synthetic letters generated per real letter. It is worth noting that while SciSpacy may introduce some errors during entity extraction, we assume these errors affect both the real and synthetic data consistently, preserving the fairness of the comparison.

		Precision	Recall	F1
x1	System_S_0.5	0.842	0.792	0.816
	System_S_0.7	0.851	0.797	0.823
	System_I_0.7	0.831	0.812	0.821
	System_I_0.9	0.846	0.810	0.827
x2	System_S_0.5	0.844	0.800	0.821
	System_S_0.7	0.850	0.805	0.828
	System_I_0.7	0.838	0.819	0.829
	System_I_0.9	0.855	0.819	**0.836**
	References	0.86	0.824	0.842

Table 4: Average Precision, Recall and F1 score for two labels (DISEASE and CHEMICAL) using Synthetic data $\times 1$, $\times 2$ and Real data, on the testing dataset.

5.3.2 Results of Downstream Task

The outcomes of the downstream evaluation are presented in Table 4. All system configurations performed *on par* with models trained using real data. Notably, systems with higher masking ratios achieved better F1 scores, likely due to the increased variability in the synthetic data, which may have provided SpaCy with a richer training set. In addition, when the volume of synthetic data was doubled, the F1 score rose to 0.836—just 0.006 below the performance of the model trained on authentic data.

5.4 Data Privacy Evaluation

To assess privacy preservation, we first measure the system's *de-identification* performance—specifically, how accurately the Masker detects all PHI instances in the test dataset. The Masker achieves a recall of 0.92 when considering all PHI categories, including additional sub-categories, and a recall of 0.96 when focusing solely on standard HIPAA-defined PHI types. Next, we assess the risk of *re-identification*, which refers to the likelihood that the MLM model inadvertently restores masked PHI entities. This step is crucial for safeguarding the privacy of individuals whose data contributed to model training. The results show that the model reintroduced PHI terms spanning more than two tokens at a very low rate of 0.035. In addition, we conducted a longest common substring analysis between original and synthetic texts for PHI segments. The overlap rates were minimal: 0.098 for substrings of 3 or more tokens, 0.020 for 5 or more, and just 0.009 for 7 or more. These findings demonstrate the system's strong performance in reliably removing HIPAA-sensitive information, while also

[4] https://allenai.github.io/scispacy/
[5] https://spacy.io/

maintaining a very low risk of re-identification.

6 Conclusion

In this work, we proposed a system using masked language models to generate synthetic clinical text, addressing challenges of data *scarcity* and *privacy*. The system includes a Masker (with de-identification, medical NER, and POS tagging) and a Mask-Filling module (supporting both simultaneous and iterative strategies). Key findings show that: **(1)** The system produces diverse yet clinically meaningful text, **(2)** Offers control over *diversity* and *fidelity* without reducing readability, **(3)** Performs well in downstream NER tasks—comparable to real data, **(4)** Ensures strong privacy protection (HIPAA-PHI recall of 0.96; re-ID risk of 0.035). The full system, SYNDEIDMLM, is available at https://github.com/SamySam0/SynDeidMLM.

Limitations and Future Work

Through close examination of the generated synthetic samples, we identified certain limitations—particularly with consistently reproducing temporal details and ensuring alignment with the original context. In some cases, maintaining *logical coherence* between related elements (e.g., referencing two names in the same scenario) proves difficult when the necessary context is outside the model's generation window. To address these issues, future work could incorporate a logic-based module for handling *temporal* data, which would enhance temporal consistency and further reduce re-identification risks. Another promising direction would be to supply the MLM model with the *type of entity* being replaced, which could increase the accuracy of PHI substitution and improve overall generation quality.

Regarding the masked language model itself, future research might explore the use of larger language models guided by prompt-based instructions to handle mask-filling. This strategy would specifically focus on the generation task, enabling a more in-depth comparison between Causal Language Models (CLMs) and Masked Language Models (MLMs) in terms of their ability to control fidelity and diversity in synthetic data. In such a setup, the Masker would remain unchanged, while the MLM and its mask-filling mechanism would be replaced by a CLM and a prompt-driven approach.

It is also important to acknowledge that our findings may have limited *generalisability*, as the experiments were conducted on a single dataset due to computational constraints. Future studies could expand the evaluation by testing the system on a wider variety of downstream tasks and datasets. For example, applying the system to specialised medical domains like radiology or oncology would be valuable. This would require replacing the current NER model with a more domain-specific one (e.g., *Stanza Radiology* or *Stanza Bionlp13cg*) to accurately extract relevant information. Such adaptation would likely necessitate re-evaluating masking strategies for both the NER component and the POS tagger to optimise performance.

We recognise that, at this stage, no alternative biomedical language models were assessed beyond Bio_ClinicalBERT. Nonetheless, future work should provide a more comprehensive rationale for selecting this model, including a comparative discussion of its strengths relative to other state-of-the-art options. A similar consideration applies to the use of Stanza for both NER and POS tagging tasks. In our readability evaluation, we reported that the synthetic letters appear easier to read than the originals, based solely on quantitative *evaluation metrics*. However, it is important to acknowledge the limitations of these metrics. Human evaluation will be necessary to more thoroughly assess the contextual appropriateness, narrative flow, and clinical usefulness of the generated content.

Acknowledgements

We thank the anonymous reviewers for valuable comments, which helped to make the paper better. LH and GN are grateful for the support from the grant "Assembling the Data Jigsaw: Powering Robust Research on the Causes, Determinants and Outcomes of MSK Disease", and the grant "Integrating hospital outpatient letters into the healthcare data space" (EP/V047949/1; funder: UKRI/EPSRC). LH is grateful for the 4D Picture EU project (https://4dpicture.eu/) on cancer patient journey support.

References

Andrew L Beam and Isaac S Kohane. 2018. Big data and machine learning in health care. *Jama*, 319(13):1317–1318.

Samuel Belkadi, Lifeng Han, Yuping Wu, and Goran Nenadic. 2023. Exploring the value of pre-trained language models for clinical named entity recogni-

tion. In *2023 IEEE International Conference on Big Data (BigData)*, pages 3660–3669. IEEE.

Samuel Belkadi, Nicolo Micheletti, Lifeng Han, Warren Del-Pinto, and Goran Nenadic. 2025. LT3: Generating medication prescriptions with conditional transformer. In *CL4Health 2025 Workshop at NAACL*.

Matúš Falis, Aryo Pradipta Gema, Hang Dong, Luke Daines, Siddharth Basetti, Michael Holder, Rose S Penfold, Alexandra Birch, and Beatrice Alex. 2024. Can gpt-3.5 generate and code discharge summaries? *arXiv preprint arXiv:2401.13512*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Suranga N Kasthurirathne, Gregory Dexter, and Shaun J Grannis. 2021. Generative adversarial networks for creating synthetic free-text medical data: a proposal for collaborative research and re-use of machine learning models. *AMIA Summits on Translational Science Proceedings*, 2021:335.

Himmet Toprak Kesgin and Mehmet Fatih Amasyali. 2023. Iterative mask filling: An effective text augmentation method using masked language modeling. In *International Conference on Advanced Engineering, Technology and Applications*, pages 450–463. Springer.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Zihao Li, Samuel Belkadi, Nicolo Micheletti, Lifeng Han, Matthew Shardlow, and Goran Nenadic. 2024. Investigating Large Language Models and Control Mechanisms to Improve Text Readability of Biomedical Abstracts . In *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, pages 265–274, Los Alamitos, CA, USA. IEEE Computer Society.

Nicolo Micheletti, Samuel Belkadi, Lifeng Han, and Goran Nenadic. 2024. Exploration of masked and causal language modelling for text generation. *arXiv preprint arXiv:2405.12630*.

Beau Norgeot, Kathleen Muenzen, Thomas A Peterson, Xuancheng Fan, Benjamin S Glicksberg, Gundolf Schenk, Eugenia Rutenberg, Boris Oskotsky, Marina Sirota, Jinoos Yazdany, et al. 2020. Protected health information filter (philter): accurately and securely de-identifying free-text clinical notes. *NPJ digital medicine*, 3(1):57.

W Nicholson Price and I Glenn Cohen. 2019. Privacy in the age of medical big data. *Nature medicine*, 25(1):37–43.

Lakshmi Radhakrishnan, Gundolf Schenk, Kathleen Muenzen, Boris Oskotsky, Habibeh Ashouri Choshali, Thomas Plunkett, Sharat Israni, and Atul J Butte. 2023. A certified de-identification system for all clinical text documents for information extraction at scale. *JAMIA open*, 6(3):ooad045.

Nilay D Shah, Ewout W Steyerberg, and David M Kent. 2018. Big data and predictive analytics: recalibrating expectations. *Jama*, 320(1):27–28.

Amber Stubbs, Christopher Kotfila, and Özlem Uzuner. 2015. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1. *Journal of biomedical informatics*, 58:S11–S19.

Amber Stubbs and Özlem Uzuner. 2015. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus. *Journal of biomedical informatics*, 58:S20–S29.

Allan Tucker, Zhenchen Wang, Ylenia Rotalinti, and Puja Myles. 2020. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *NPJ digital medicine*, 3(1):1–13.

Chao Yan, Ziqi Zhang, Steve Nyemba, and Zhuohang Li. 2024. Generating synthetic electronic health record data using generative adversarial networks: Tutorial. *JMIR AI*, 3:e52615.