Handling Name Errors of a BERT-Based Text De-Identification System: Insights from Stratified Sampling and Markov-based Pseudonymization

Dalton Simancek
Department of Learning Health Sciences
University of Michigan
daltonsi@umich.edu

V.G.Vinod Vydiswaran
School of Information
University of Michigan
vgvinodv@umich.edu

Abstract

Missed recognition of named entities while de-identifying clinical narratives poses a critical challenge in protecting patient-sensitive health information. Mitigating name recognition errors is essential to minimize risk of patient re-identification. In this paper, we emphasize the need for stratified sampling and enhanced contextual considerations concerning Name tokens using a fine-tuned Longformer BERT model for clinical text de-identification. We introduce a Hidden in Plain Sight (HIPS) Markov-based replacement technique for names to mask name recognition misses, leading to a significant reduction in name leakage rates. Our experimental results underscore the impact on addressing name recognition challenges in BERT-based de-identification systems for heightened privacy protection in electronic health records.

1 Introduction

Clinical narratives and unstructured documentation within electronic health records (EHRs) are considered valuable assets in epidemiological research (Sheikhalishahi et al., 2019; Patra et al., 2021) and the creation of prognostic clinical prediction models (Seinen et al., 2022). The advancement of these applications is frequently impeded by the limited availability of de-identified clinical text corpora. Clinical notes must remove protected health information (PHI) to safeguard patient privacy and align with privacy regulations exemplified by the United States' HIPAA Safe Harbor privacy guidelines (OfC, 2022).

Bidirectional Encoder Representations from Transformers (BERT) models have shown promise in automatically identifying sensitive PHI in clinical texts (Johnson et al., 2020; Ahmed et al., 2020). Previous research has explored BERT variants with hyper-parameter tuning, comparing their efficacy in clinical text de-identification across PHI subcategories, including dates, phone numbers and names (Meaney et al., 2022). Less focus has been given to the shrinking but persistent margins of error that plague even the best-performing de-identification pipelines. For instance, our base de-identification model uses a Longformer BERT variant (Beltagy et al., 2020) fine-tuned using discharge summaries from a US-based tertiary healthcare institution. In preliminary work as a part of Alkiek et al. (2023), we discovered that the Longformer model performed exceptionally well during preliminary testing among the pretrained models we sampled. Our base model achieved an impressive but far-from-optimal overall F1 score of 0.90 for names (Table 1). Considering the inherent privacy risks of missed PHI, even small margins of error have the potential to result in substantial numbers of exposed patient records with identifiable health information when de-identification systems are deployed in the real world.

This study aims to contribute to characterizing Name errors in BERT-based transformer models on real-world discharge summaries. The goal is to provide insights into the development of comprehensive de-identification strategies capable of both correcting bias and tolerating mistakes related to names. We first compare the effect on recognizing names using a fine-tuned Longformer model with a stratified sample that uses demographic information against a model fine-tuned with a standard randomized sample. We suggest that enhancing the recognition of name tokens in our de-identification system is achieved by including sets of name tokens found in stratified training samples compared to those in random training samples. Next, we investigate a Hidden in Plain Sight (HIPS) Markov-based replacement technique for names. Using real-world discharge summaries from an academic healthcare institution, we compare the effectiveness of a Markov-based replacement strategy against a random replacement strategy in reducing name leakage.

Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization
CALD-pseudo 2024, pages 1–7

PHI	Mean True Token Count	Precision	Recall	F1-Score
PROVIDER NAMES	503	.984 (.007)	.984 (.007)	.984 (.007)
NON-PROVIDER NAMES	1836	.904 (.019)	.850 (.027)	.876 (.010)
NAMES IN WHITE RECORDS	1756	.923 (.022)	.889 (.019)	.905 (.006)
NAMES IN URM RECORDS	583	.925 (.018)	.860 (.024)	.892 (.014)
ALL NAMES	2339	.923 (.016)	.880 (.019)	.900 (.007)
ALL PHI	13103	.953 (.005)	.952 (.005)	.952 (.004)

Table 1: Baseline fine-tuned Longformer performance, averaged over 5 runs, on identifying name tokens over a test set of discharge summaries (n=80). For performance metrics, the numbers in parentheses show standard deviation. Name tokens from full name entities that matched with doctor, nurse or specialist names in a given EHR provider list were labeled as provider names. Name tokens from unmatched full name entities in the EHR provider list were labeled as non-provider. Underrepresented minority (URM) records are notes associated with patients reporting any non-white racial identity or a Hispanic/Latino ethnicity.

2 Related Work

2.1 Name-related biases in BERT Models

Naming a person often involves a deliberate or subconscious choice conveying racial, ethnic, class-based, gender-normative, or religious affiliations (Seguin et al., 2021; Lindsay and Dempsey, 2017). These choices collectively contribute to naming trends or groups of names characterized by gender, race, or association with a specific locality or decade (Lockhart et al., 2023). Learned contextual embeddings in BERT models have been shown to capture such signals in socio-demographic phenomena, which may then contribute to discriminative biases in recruitment and other systems informed by trained contextual embeddings (Ramezanzadehmoghadam et al., 2021). Name tokens like "Smith" are expected to be prominent in United States based, English-language datasets, making them more easily identifiable by systems trained on those datasets. In contrast, names that are rare, unique, or correspond to underrepresented minority (URM) groups in the training data may carry a higher risk of being overlooked by the model. Our work is inspired by Xiao et al. (2023), who studied learned name biases in pretrained BERT models for de-identification using synthetic patient data to evaluate fairness across demographic patient groups. Others, such as Yue and Zhou (2020), have proposed solutions involving training data augmentation.

2.2 Hidden in Plain Sight (HIPS) strategies

Multiple strategies have been explored to enhance privacy through de-identification. Notable among these are the HIPS approaches, which involves substituting personally-identifiable information (PIIs) with authentic pseudonyms to mitigate false negatives (Carrell et al., 2013). For example, instead of replacing a real name token "John" with a placeholder tag "[NAME]", HIPS uses a realistic pseudonym like "Tony". As an extension of this methodology, Osborne et al. (2022) proposed Bratsynthetic, which utilizes a Markov-based substitution component to introduce randomness in the selection of pseudonyms. Building on this foundation, our study applies a similar approach to a real-world dataset of discharge summaries, extending the scope of its evaluation beyond simulated environments.

3 Methods

3.1 Stratification of Training Data

Our methodology challenges the conventional practice of randomly selecting clinical notes for training data annotation, that tends to predominantly include the majority patient population of white patients. Instead, we advocate for a stratified sampling approach to diversify the training data, that would select disproportionately more patients from underrepresented minority communities. This method aims to improve the generalizability of the Longformer-based de-identification model previously proposed by Alkiek et al. (2023), particularly in identifying patient names during inference. We propose that the name token sets found in a stratified training samples containing a higher proportion of documents from URM patients, would result in superior name de-identification performance than the performance achieved with the name token set found in random training samples.

Sampling	Demographic	# Names	Precision	Recall	F1 score
Random	All	654 (17)	0.918 (0.021)	0.843 (0.019)	0.878 (0.008)
	White	483 (24)	0.916 (0.014)	0.849 (0.022)	0.881 (0.009)
	URMs	171 (29)	0.901 (0.029)	0.837 (0.042)	0.867 (0.020)
Stratified	All	654 (17)	0.897 (0.032)	0.842 (0.018)	0.868 (0.014)
	White	483 (24)	0.900 (0.026)	0.857 (0.032)	0.877 (0.015)
	URMs	171 (29)	0.892 (0.030)	0.837 (0.025)	0.863 (0.020)

Table 2: Test set performance of patient name identification models fine-tuned on (a) random sample (URM patients: 21%, s.d. 0.01) and (b) stratified sample (URM patients: 34%, s.d. 0.01). Averaged over five runs. The numbers in parentheses are standard deviations.

Our annotated dataset consists of 400 discharge summary notes from a tertiary academic medical center, annotated by trained medical professionals to label eighteen PHI categories, including provider and patient names. The average discharge summary consists of 1643 tokens containing 43 PHI entity annotations, of which 12 are names. The average name entity has two tokens, typically a first name and a surname. For this study, full-name entities that did not match with doctor, nurse, or specialist names in a given EHR provider list were labeled as non-providers and assumed to patient or family names.

The discharge summary notes were randomly split 80:20 into train-test splits. To implement the stratified sampling approach and evaluate its effectiveness, we created two subsamples of 200 notes each from the train set. One subsample followed random selection, while the other used stratified sampling that leveraged structured EHR race and ethnicity fields to create a white stratum and an underrepresented minority (URM) stratum. A patient was categorized into the URM stratum if their record indicated any non-white racial identity or Hispanic/Latino ethnicity. In the stratified sample, we included the maximum number of available notes from URM (Underrepresented Minority) patients in the train split. The remaining notes were randomly sampled from white patients. This process was repeated five times, starting from a new 80:20 split.

On average, the number of instances from the underrepresented minority patients increased from 21% in the random sample to 34% in the stratified sample. Subsequently, two new Longformer models were trained and tested on the same held-out test set (n=80): one fine-tuned with the random subsample and the other with the stratified subsample.

3.2 Markov-based name pseudonymization

Inspired by the strategy used by Osborne et al. (2022), we designed the following Markov-based strategy to replace the names identified by the Longformer models with surrogates. For each note, when the Longformer identifies k name token predictions, our approach will select k surrogates. There's a $p = 0.5$ chance that the preceding surrogate name will be reused, and a $p = 0.5$ chance that a new surrogate name will be chosen. In addition, each surrogate name is constrained to be used at most $t = 4$ times, which follows the observation that the mode frequency of name tokens in the training documents is four. Once identified, the surrogate names are used to replace the original names at random. This strategy was contrasted with random pseudonymization strategy, where the k surrogate names are selected at random without replacement. All surrogate names were selected from a United States Census name list (Bureau, 2021) and the sampling strategies are document-independent, ensuring that names are not sampled repetitively for the same name tokens across multiple documents. The values for the probability and repetition threshold was set empirically based on the mode of the name frequency distribution.

3.3 Name leakage

We defined name leakage as instances where a name token appears more frequently than allowed by a perfect re-identification system and replacement strategy. For example, a name token appearing five times in a document would flag that document as leaking real name tokens with our Markov-based replacement strategy, as it contains a name token count exceeding the maximum threshold of $t = 4$ uses. The overall leakage rate was calculated across the entire test set by considering the proportion of notes where the mode of the name

frequency distribution was indeed a true patient name. This follows a heuristic that in a clinical note, the patient's name is the one most likely to repeat. Wilcoxon signed-rank tests (Wilcoxon, 1945) were conducted to test statistical significance of differences in mode of name frequency distribution in the original data and each replacement strategy. McNemar's test (Sundjaja et al., 2023) was conducted to test if the leakage rate of Markov-based strategy is significantly lower than the random replacement strategy.

4 Results

4.1 Stratification of training data

Across the five iterations, the performance of the Longformer models fine-tuned with stratified samples exhibited comparable results to the models fine-tuned with random samples (Table 2). A slight reduction in precision and F1-score was observed across all test notes and within each demographic subgroup. Recall remained consistent for all test notes and URM notes, with a modest and statistically insignificant increase in recall for names in notes from white patients. Consequently, our demographic stratification rule, aimed at enhancing the representation of URMs in the training set of notes from 21% to 34%, did not yield an improvement in name recognition.

4.2 Markov-based name pseudonymization

Pseudonymization strategies significantly influenced the frequency of name tokens in documents, impacting the risk of name leakage. For each note, the random replacement strategy assigns a unique surrogate to each name token prediction, ensuring no repetition. This results in an expected mode of 1 among all name token frequencies in a given document. In contrast, Markov-based replacement may reuse a name token surrogate for up to four predictions, adjusting the expected mode to 4 among all name token frequencies. Our analysis showed that instances of a name being missed twice, even with random replacement, could signal potential name leakage.

Over the five iterations, random strategy displayed significant differences in mode values, whereas Markov distributions exhibited smaller differences, compared to the original name frequency distribution. In Iteration 5 (Figure 1), the Markov name distribution closely resembled the original distribution and was statistically similar per Wilcoxon's signed rank test.

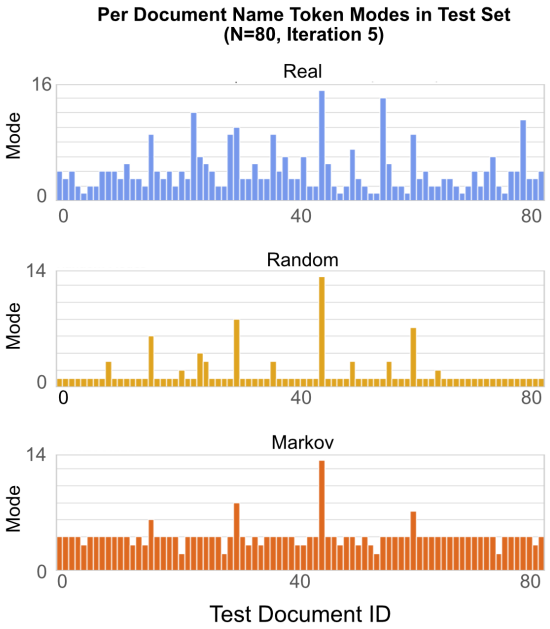

Figure 1: Distributions of unique name modes in the test set by replacement strategy on a sample iteration.

Further, the average leakage rate decreased from 13.1% with random replacement to 3.8% with Markov replacement across all iterations (Table 3). Markov replacement significantly reduced the name leakage rate compared to random replacement, as evidenced by McNemar's test.

5 Discussion

While previous research has emphasized the effectiveness of strategies such as data augmentation for improving name diversity and contextual patterns (Yue and Zhou, 2020), we focused on enhancing name performance through stratification. We aimed to increase the proportion of URM patients in the train dataset by using a stratified sampling approach. Surprisingly, this approach did not significantly improve the name performance of our de-identification system. This lack of improvement could be attributed to a relatively small sample size, where the change in URM representation was not substantial enough to impact name performance in a predominately White patient population. Additionally, higher URM representation in the training sample through stratification may not perfectly correlate with a more robust set of name tokens. URM patients may have names more commonly associated with white demographics and white patients may also have rare or unique names.

Iter.	Original data		Random replacement		Markov-based replacement	
	Avg. Mode	Leakage (%)	Avg. Mode	Leakage (%)	Avg. Mode	Leakage (%)
1	3.5	100	1.2	7.5	3.8	1.3*
2	3.8	100	1.5	17.7	4.1	3.8***
3	3.7	100	1.4	11.3	4.0	3.8*
4	3.9	100	1.6	14.1	4.0	5.1**
5	4.0	100	1.6	15.0	4.0	5.0**
AVG	3.8	100	1.5	13.1	4.0	3.8

Table 3: Average mode and rate of name leakage on the test set across five runs. Statistically significant drop relative to leakage in random approach using McNemar's test:*(p<0.05), **(p<0.01),***(p<0.001)

In our discharge summaries, there is a single patient with two notes. To prevent name token leakage, we deliberately included both notes from this patient exclusively in the training sets. This ensures that the names associated with this particular patient do not leak between the training and test sets through their notes.

However, this method does not address the potential leakage of provider names between training and test sets, as a provider's name may appear in notes from multiple patients. Additionally, our stratification rule for creating the white and URM strata does not consider the demographics of the providers. Consequently, not only may provider names leak between the training and test sets, but the set of provider name entities could also remain similar between the stratified and random training samples, posing a challenge to the effectiveness of our methodology." The higher performance for provider names in Table 1 is therefore unsurprising considering the possibility of name leakage for this subgroup. This observation reinforces the importance of name token distributions represented in the training data as a vector for improving de-identification performance on names.

Future work could expand on efforts to explore how shifts in training data effect embedding or attention mechanisms (Clark et al., 2019) to better understand causal mechanisms for targeting name distributions as a parameter to improving de-identification system performance on names.

Nevertheless, achieving complete elimination of name errors through incremental model improvements is unlikely, especially as the model is applied to different note types or health system contexts. In such scenarios, our Markov-based HIPS strategy holds theoretical efficacy in diminishing document-level leakage associated with false negatives related to name tokens. However, missed names that appear more than the allowable repeats in the Markov method remain at risk for PHI leakage. Our results propose that implementing such a strategy would be beneficial for BERT-based de-identification systems, effectively masking name errors and providing enhanced privacy protection for patients.

6 Conclusion

Our study aims to address errors in name recognition of a Longformer-based de-identification model fine-tuned on discharge summaries. Stratified sampling for fine-tuning did not significantly improve name recognition. However, the introduction of a HIPS Markov-based pseudonymization strategy showed promising results, significantly reducing name leakage rates compared to random replacement. This research contributes to the ongoing efforts to address the persistent challenges associated with de-identification of clinical texts, offering valuable guidance for the development of robust and privacy-conscious clinical text de-identification.

Limitations

While we recognize the importance of acknowledging and addressing disparities in healthcare, the term "Underrepresented Minorities (URM)" utilized in this study is acknowledged as not being an ideal or precise grouping, potentially oversimplifying the diverse range of minority patient populations in a tertiary, academic healthcare institution. The rationale behind the chosen stratification rule was to establish a straightforward method that could yield sufficient samples in each stratum to ensure a stable signal for analysis.

Moving forward, we advocate for future research endeavors in stratified sampling to prioritize annotation efforts aimed at including more substantial samples from demographic subgroups. It is impor-

tant to acknowledge and accept this limitation as a conscious choice made to address the inherent dominance and potential bias arising from a larger representation of White patient records in this specific experiment. Our commitment remains steadfast in contributing to a nuanced understanding of healthcare disparities, and we encourage ongoing efforts to refine and expand upon our methodology in future investigations.

Additionally, using name lists from US Census data for pseudonymization may not be ideal for all contexts due to potential biases and limitations. Careful consideration of the specific context and potential biases is crucial when selecting and utilizing name lists for pseudonymization purposes.

Finally, we acknowledge that fine-tuning of our BERT-based de-identification system used discharge summaries written exclusively in American English clinical language and not other languages.

Ethics Statement

The experiments in this study use 400 discharge summaries sampled from patient populations in a tertiary, academic healthcare institution. The handling of sensitive medical data prioritized participant privacy, with robust data handling protocols in place. The research was conducted under the oversight of an Institutional Review Board, ensuring continual adherence to ethical guidelines.

Acknowledgements

We acknowledge the authors of Alkiek et al. (2023), who developed the DeiDoc de-identification model and shared code, annotated data, and model insights used in conceiving the experiments conducted in this paper. These contributions significantly enhanced the quality of this research.

References

Tanbir Ahmed, Md Momin Al Aziz, and Noman Mohammed. 2020. De-identification of electronic health record using neural network. *Scientific Reports*, 10(1):18600.

Kenan Alkiek, Dalton Simancek, Jiaye Tan, Noah Weissman, Jane Ferraro, and Vydiswaran V.G. Vinod. 2023. DeiDoc: Automatic De-identification of Notes Using Long-Document Transformers. In *Annual Symposium of the American Medical Informatics Association*.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. ArXiv:2004.05150 [cs].

United States Census Bureau. 2021. Frequently Occurring Surnames from the 2010 Census.

David Carrell, Bradley Malin, John Aberdeen, Samuel Bayer, Cheryl Clark, Ben Wellner, and Lynette Hirschman. 2013. Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. *Journal of the American Medical Informatics Association*, 20(2):342–348.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What Does BERT Look At? An Analysis of BERT's Attention. ArXiv:1906.04341 [cs].

Alistair E. W. Johnson, Lucas Bulgarelli, and Tom J. Pollard. 2020. Deidentification of free-text medical records using pre-trained bidirectional transformers. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 214–221, Toronto Ontario Canada. ACM.

Jo Lindsay and Deborah Dempsey. 2017. First names and social distinction: Middle-class naming practices in Australia. *Journal of Sociology*, 53(3):577–591.

Jeffrey W Lockhart, Molly M King, and Christin Munsch. 2023. What's in a Name?: Name-Based Demographic Inference and the Unequal Distribution of Misrecognition. *Nat Hum Behav*, 7:1084–1095.

Christopher Meaney, Wali Hakimpour, Sumeet Kalia, and Rahim Moineddin. 2022. A Comparative Evaluation Of Transformer Models For De-Identification Of Clinical Text Data. ArXiv:2204.07056 [cs, stat].

Rights (OCR) OfC. 2022. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule.

John D Osborne, Tobias O'Leary, and Richard E Kennedy. 2022. BRATsynthetic: Text De-identification using a Markov Chain Replacement Strategy for Surrogate Personal Identifying Information. ArXiv:2210.16125 [cs.CR].

Braja G Patra, Mohit M Sharma, Veer Vekaria, Prakash Adekkanattu, Olga V Patterson, Benjamin Glicksberg, Lauren A Lepow, Euijung Ryu, Joanna M Biernacka, Al'ona Furmanchuk, Thomas J George, William Hogan, Yonghui Wu, Xi Yang, Jiang Bian, Myrna Weissman, Priya Wickramaratne, J John Mann, Mark Olfson, Thomas R Campion, Mark Weiner, and Jyotishman Pathak. 2021. Extracting social determinants of health from electronic health records using natural language processing: a systematic review. *Journal of the American Medical Informatics Association*, 28(12):2716–2727.

Maryam Ramezanzadehmoghadam, Hongmei Chi, and Edward L Jones. 2021. Inherent Discriminability of BERT towards Racial Minority Associated Data. volume 12951. Springer, Cham.

Charles Seguin, Chris Julien, and Yongjun Zhang. 2021. The stability of androgynous names: Dynamics of gendered naming practices in the United States 1880–2016. *Poetics*, 85:101501.

Tom M Seinen, Egill A Fridgeirsson, Solomon Ioannou, Daniel Jeannetot, Luis H John, Jan A Kors, Aniek F Markus, Victor Pera, Alexandros Rekkas, Ross D Williams, Cynthia Yang, Erik M Van Mulligen, and Peter R Rijnbeek. 2022. Use of unstructured text in prognostic clinical prediction models: a systematic review. *Journal of the American Medical Informatics Association*, 29(7):1292–1302.

Seyedmostafa Sheikhalishahi, Riccardo Miotto, Joel T Dudley, Alberto Lavelli, Fabio Rinaldi, and Venet Osmani. 2019. Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review. *JMIR Medical Informatics*, 7(2):e12239.

Joshua Henrina Sundjaja, Rijen Shrestha, and Kewal Krishan. 2023. McNemar And Mann-Whitney U Tests. *StatPearls Publishing*.

Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83.

Yuxin Xiao, Shulammite Lim, Tom Joseph Pollard, and Marzyeh Ghassemi. 2023. In the Name of Fairness: Assessing the Bias in Clinical Record De-identification. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 123–137, Chicago IL USA. ACM.

Xiang Yue and Shuang Zhou. 2020. PHICON: Improving Generalization of Clinical Text De-identification Models via Data Augmentation. ArXiv:2010.05143 [cs].

A Base De-identification Model Details

In our experiments, we employed a pre-trained Longformer model as the baseline for fine-tuning the de-identification approach over discharge summaries, based on work conducted by Alkiek et al. (2023). Each experiment consisted of five iterations of model training with different train-test splits, with each iteration consisting of 5 epochs. Training hyperparameters include a learning rate of 2e-5 and a batch size of 4. These parameters were chosen to balance the model's learning capacity and computational efficiency during the fine-tuning process. The utilization of a pre-trained Longformer model served as a foundation for our experiments, allowing us to leverage its inherent understanding of contextual information in medical text.