Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240515

Preliminary Evaluation of Fine-Tuning the OpenDeLD Deidentification Pipeline Across Multi-Center Corpora

Shalini GUPTA [a], Jiaxing LIU [b], Zoie Shiu-Yee WONG [c,d] and Jitendra JONNAGADDALA [e,f,1]

[a] *CGD Health Pvt. Ltd. Hyderabad, Telangana, India*
[b] *School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan, China*
[c] *Graduate School of Public Health, St. Luke's International University, Tokyo, Japan*
[d] *The Kirby Institute, University of New South Wales, Sydney, Australia*
[e] *School of Population Health, UNSW Sydney, Kensington, Australia*
[f] *NMC Royal Hospital, Khalifa City, Abu Dhabi, United Arab Emirates*

ORCiD ID: Shalini Gupta https://orcid.org/0009-0008-2537-7791, Jiaxing Liu https://orcid.org/0000-0002-5184-4313, Zoie Shiu-Yee Wong https://orcid.org/0000-0003-4499-9779, Jitendra Jonnagaddala https://orcid.org/0000-0002-9912-2344

Abstract. Automatic deidentification of Electronic Health Records (EHR) is a crucial step in secondary usage for biomedical research. This study introduces evaluation of an intricate hybrid deidentification strategy to enhance patient privacy in secondary usage of EHR. Specifically, this study focuses on assessing automatic deidentification using OpenDeID pipeline across diverse corpora for safeguarding sensitive information within EHR datasets by incorporating diverse corpora. Three distinct corpora were utilized: the OpenDeID v2 corpus containing pathology reports from Australian hospitals, the 2014 i2b2/UTHealth deidentification corpus with clinical narratives from the USA, and the 2016 CEGS N-GRID identification corpus comprising psychiatric notes. The OpenDeID pipeline employs a hybrid approach based on deep learning and contextual rules. Pre-processing steps involved harmonizing and addressing encoding and format issues. Precision, Recall, F-measure metrics were used to assess the performance. The evaluation metrics demonstrated the superior performance of the Discharge Summary BioBERT model. Trained on three corpora with a total of 4,038 reports, the best performing model exhibited robust deidentification capabilities when applied to EHR. It achieved impressive micro-averaged F1-scores of 0.9248 and 0.9692 for strict and relaxed settings, respectively. These results offer valuable insights into the model's efficacy and its potential role in safeguarding patient privacy in secondary usage of EHR.

Keywords. Electronic Health Records, BioBERT, Deidentification, Multi-Center Corpora

[1] Corresponding Author: Dr Jitendra Jonnagaddala; E-mail: z3339253@unsw.edu.au.

1. Introduction

Electronic Health Records (EHR) signify a crucial progression in healthcare informatics, fundamentally transforming the acquisition, storage, and utilization of patient information within the healthcare system and create a dynamic and interconnected system that facilitates seamless information exchange among healthcare providers, enhancing patient care, safety, and overall healthcare outcomes [1, 2]. The implementation of EHR systems facilitates the standardization of data collection, enabling systematic analysis of extensive datasets by healthcare professionals [3]. The incorporation of machine learning, artificial intelligence [1] and Large Language Models (LLM) [4] within EHR systems further enhances analytical capabilities, providing predictive modeling and personalized healthcare interventions.

Deidentification of EHR text notes are a critical aspect of safeguarding patient privacy and promoting the responsible use of healthcare data across various corpora. This process involves the systematic removal or obfuscation of Sensitive Health Information (SHI) from health records to protect patient confidentiality. However, challenges such as unstructured data, biasness, preserving context, and dynamic updates persist. Notably, the AI Cup 2023 - Privacy Protection and Medical Data Standardization competition and the 2024 International Workshop on the Deidentification of Electronic Medical Record Notes (IW-DMRN) were organized to address these deidentification challenges [5-7]. Effective deidentification necessitates technological progress, ethical adherence, and a delicate balance between privacy and data utility in healthcare's evolving data landscape [8]. Ensuring privacy and security in today's data-driven world requires rigorous verification of deidentification techniques for SHI. The objective of this study is to assess the enhanced efficacy of OpenDeID pipeline [9] by finetuning the model with three diverse corpora. This investigation seeks to examine how variations in corpus selection influence the performance of deidentification approaches, with a focus on advancing our understanding of the optimal strategies for safeguarding sensitive information within diverse datasets. The study aims to contribute valuable insights into the refinement and optimization of a deidentification technique, thereby enhancing the overall efficacy of privacy-preserving measures in handling various types of data.

2. Methods

In this section, we introduce the deidentification corpus, detail the data processing steps, present our hybrid deidentification method, and explain the evaluation metrics utilized. For the experimental procedure, three distinct corpora were employed. Firstly, the OpenDeID v2 corpus of 3244 reports, comprising 2,100 pathology reports from OpenDeID v1 corpus [10, 11] and an additional 1,144 reports from four urban hospitals in Australia. Secondly, the 2014 i2b2/UTHealth deidentification corpus [12, 13], comprising 1,304 longitudinal clinical narratives from 296 patients in the USA. Lastly, the 2016 CEGS N-GRID [14] identification corpus encompasses 1,000 psychiatric notes from the USA. The 2014 i2b2/UTHealth deidentification and 2016 CEGS N-GRID deidentification datasets are accessible via the official i2b2 website, requiring interested users to complete and submit a Data Use Agreement (DUA) form, ensuring adherence to ethical and legal requirements.

In the pre-processing phase, several key steps were taken. Firstly, the critical removal of the Byte Order Mark (BOM) character was executed from the tab-separated annotation file to facilitate smooth reading within the OpenDeID pipeline. Python, using 'utf-8-sig' encoding, ensured the exclusion of the BOM character. Additionally, conversion from TXT to XML format was essential for the OpenDeID pipeline. This step, supported by a Document Type Definition (DTD) file and a mask annotation file, standardized annotations for clarity. Manual intervention was necessary to address encoding and annotation errors during testing, ensuring data integrity for further processing. To conduct the experiments, we employ the OpenDeID pipeline [9], a multi-stage system designed for executing the deidentification process. Initially, the XML pathology reports underwent tokenization in which reports were broken down into individual tokens. Subsequently, these tokens were annotated with labels in the BIESO format, indicating the beginning, inside, end, single, or outside status of each token. Following this, the processed tokens were organized into sequences, creating a structured representation that captures the contextual relationships among tokens within the pathology reports. OpenDeID pipeline is also capable to generating surrogates [15].

In our model construction, we utilized Discharge Summary BioBERT, a specialized NLP framework built upon BERT [16], implemented in Python 3.9 using PyTorch (version 1.10.1). The hyperparameters used for fine-tuning the Discharge Summary BioBERT were aligned with those of the OpenDeID pipeline [9]. Table 1 described the experimental setting which involved four distinct runs, each with varying configurations of training, validation, and testing datasets. Some of the experiment was conducted as a part of AI- Cup 2023 challenge [5-7] and due to time constraints, the validation size is notably reduced to ensure prompt results.

The evaluation assesses the effectiveness of the methods using micro-averaged precision, recall, and F1-scores under both strict and relaxed matching criteria. Strict matching requires exact offsets, while relaxed matching allows for character offset tolerance. Predictions are then merged with the input and post-processed to generate tagged XML output, ensuring robust methodology validation.

Table 1. Experimental settings of different training data selection during fine-tuning Discharge Summary BioBERT.

	Training Data			**Validation Data**	**Test Data**
	OpenDeID v2 corpus (N=1734)	i2b2 2014 corpus (N=1304)	2016 CEGS N-GRID corpus (N=1000)	OpenDeID v2 corpus (N=10)	OpenDeID v2 corpus (N=560)
Run 1	Included	Not Included	Not Included	Included	Included
Run 2	Included	Included	Not Included	Included	Included
Run 3	Included	Not Included	Included	Included	Included
Run 4	Included	Included	Included	Included	Included

3. Results

Table 2 presents the performance metrics of OpenDeID pipeline using the Discharge Summary BioBERT model under various settings. Among the different experimental settings, the Discharge Summary BioBERT exhibited the most favorable overall performance, particularly in Run4, when all three corpora were included into training. This specific model was trained on a dataset comprising 4038 reports, validated on 10 reports, and tested on 560 OpenDeID v2 corpus reports. The achieved micro-averaged F1-scores for strict and relaxed matching were 0.9248 and 0.9692, respectively. These results underscore the efficacy of the Discharge Summary BioBERT model in accurately deidentifying EHR text notes, demonstrating robust performance in both stringent and more lenient evaluation criteria. This performance evaluation yields important insights into the model's capabilities and underscores its potential for protecting patient privacy in healthcare data.

Table 2. Performance of the all the BERT-base models with different settings

Runs	Strict Precision	Strict Recall	Strict F1	Relaxed Precision	Relaxed Recall	Relaxed F1
Run1	0.9525	0.8887	0.9195	0.953	0.8891	0.9199
Run2	0.9594	0.8884	0.9225	0.959	0.8887	0.9228
Run3	0.9571	0.886	0.9202	0.9575	0.8864	0.9206
Run4	0.9642	0.8884	0.9248	0.9647	0.8889	0.9692

4. Discussion

This research evaluates a hybrid approach to automatically deidentify sensitive healthcare data, utilizing three distinct corpora. By combining these corpora, it evaluates the OpenDeID pipeline's effectiveness across diverse clinical notes. In our pursuit of methodological rigor, we meticulously scrutinized the performance of the OpenDeID deidentification pipeline.[9] By incorporating the OpenDeID v2, the 2014 i2b2/UTHealth deidentification, and the 2016 CEGS N-GRID identification corpus, our methodology not only broadens its applicability but also ensures a thorough examination of its efficacy across varied healthcare contexts. [17] In addition to the experiments detailed in this study, we also conducted several experiments for the 2023 SREDH/AI Cup challenge, utilizing different configuration reports from the same corpora and their results are presented elsewhere. [5-7]

The study's evaluation shows promising performance, with Run 4 demonstrating high micro-averaged F1-scores for strict and relaxed matching scenarios using Discharge Summary BioBERT. Integrating the 2016 CEGS N-GRID and 2014 i2b2 datasets enhances the pipeline's effectiveness by leveraging shared linguistic patterns across the USA and Australia. This study introduces a framework for methodological benchmarking through meticulous corpora selection and model construction. These findings represent a substantial stride forward in enhancing patient confidentiality, prompting further exploration and practical implementation in diverse healthcare settings. While this study provides valuable insights into the effectiveness of the hybrid deidentification approach, it's important to acknowledge certain limitations. One limitation is the reliance on specific corpora, which may not fully represent all possible variations in healthcare data. Additionally, the evaluation primarily focuses on Australian and USA EHR, potentially limiting the generalizability of the findings to other healthcare systems. Future studies could address these limitations by incorporating more diverse datasets from different countries and exploring the applicability of the method across these datasets. In future study, we plan to extend this preliminary work by delving into deidentification methodologies beyond diverse corpora, using advanced techniques like LLM, and conducting comprehensive evaluations to compare with the OpenDeID pipeline.

5. Conclusions

In conclusion, this study evaluates the OpenDeID hybrid deidentification pipeline. By integrating diverse corpora such as OpenDeID, i2b2/UTHealth, and CEGS N-GRID, the pipeline undergoes a thorough evaluation across various clinical narratives and geographical regions, showcasing its adaptability and broad applicability. The evaluation framework, focusing on accuracy and performance metrics, highlights the Discharge Summary BioBERT model is proficient in achieving high micro-averaged F1-scores for both strict and relaxed matching criteria. These findings contribute significantly to automatic deidentification of EHR, paving the way for future research and applications to ensure the confidentiality and security of SHI.

References

[1] Menachemi N, Collum TH. Benefits and drawbacks of electronic health record systems. Risk Manag Healthc Policy. 2011;4:47-55.

[2] Jonnagaddala J, et al. Mining Electronic Health Records to Guide and Support Clinical Decision Support Systems. In: Management Association IR, editor. Healthcare Ethics and Training: Concepts, Methodologies, Tools, and Applications. Hershey, PA, USA: IGI Global; 2017. p. 184-201.

[3] Pai MMM, Ganiga R, Pai RM, Sinha RK. Standard electronic health record (EHR) framework for Indian healthcare system. Health Services and Outcomes Research Methodology. 2021;21(3):339-62.

[4] Clusmann J, Kolbinger FR, Muti HS, Carrero ZI, Eckardt J-N, Laleh NG, et al. The future landscape of large language models in medicine. Communications Medicine. 2023;3(1):141.

[5] ISLAB. Privacy protection and medical data standardization competition: decoding clinical cases and letting data tell stories: CodaLab; 2023 [Available from: https://codalab.lisn.upsaclay.fr/competitions/15425.

[6] Gupta S, Alla NLV, Panchal O, Jonnagaddala J, editors. Evaluation of OpenDeID Pipeline in the AI-Cup 2023 Challenge for Deidentification of Sensitive Health Information. 2024 International workshop on deidentification of electronic medical record notes (IW-DMRN); 2024.

[7] Consortium S. 2024 International workshop on deidentification of electronic medical record notes (IW-DMRN) 2024 [Available from: https://www.sredhconsortium.org/sredh-competitions/sredhai-cup-2023/2024-iw-dmrn.

[8] Xiao Y, Lim S, Pollard TJ, Ghassemi M. In the Name of Fairness: Assessing the Bias in Clinical Record De-identification. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency; Chicago, IL, USA: Association for Computing Machinery; 2023. p. 123–37.

[9] Liu J, Gupta S, Chen A, Wang CK, Mishra P, Dai HJ, et al. OpenDeID Pipeline for Unstructured Electronic Health Record Text Notes Based on Rules and Transformers: Deidentification Algorithm Development and Validation Study. J Med Internet Res. 2023;25:e48145.

[10] Jonnagaddala J, Chen A, Batongbacal S, Nekkantti C. The OpenDeID corpus for patient de-identification. Scientific Reports. 2021;11(1):19973.

[11] Alla NLV, Chen A, Batongbacal S, Nekkantti C, Dai H-J, Jonnagaddala J. Cohort selection for construction of a clinical natural language processing corpus. Computer Methods and Programs in Biomedicine Update. 2021;1:100024.

[12] Stubbs A, Kotfila C, Uzuner Ö. Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. Journal of Biomedical Informatics. 2015;58:S11-S9.

[13] Stubbs A, Uzuner Ö. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. Journal of Biomedical Informatics. 2015;58:S20-S9.

[14] Stubbs A, Filannino M, Uzuner Ö. De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID shared tasks Track 1. Journal of Biomedical Informatics. 2017;75:S4-S18.

[15] Chen A, Jonnagaddala J, Nekkantti C, Liaw ST. Generation of Surrogates for De-Identification of Electronic Health Records. Stud Health Technol Inform. 2019;264:70-3.

[16] Devlin J, et al, editors. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. North American Chapter of the Association for Computational Linguistics; 2019.

[17] Ahmed T, Aziz MMA, Mohammed N. De-identification of electronic health record using neural network. Scientific Reports. 2020;10(1):18600.