Journal of the American Medical Informatics Association, 2024, **31(7)**, 1578–1582
https://doi.org/10.1093/jamia/ocae092
Advance access publication 3 May 2024
Brief Communication

Brief Communication

Machine learning to predict notes for chart review in the oncology setting: a proof of concept strategy for improving clinician note-writing

Sharon Jiang, MEng*[,1,2], Barbara D. Lam, MD[3,4], Monica Agrawal, PhD[1,2], Shannon Shen, MS[1,2], Nicholas Kurtzman, MD[5], Steven Horng, MD[4,5], David R. Karger, PhD[1,2], David Sontag, PhD[1,2]

[1]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139, United States, [2]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, United States, [3]Division of Hematology and Oncology, Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA 02215, United States, [4]Division of Clinical Informatics, Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA 02215, United States, [5]Department of Emergency Medicine, Beth Israel Deaconess Medical Center, Boston, MA 02215, United States

*Corresponding author: Sharon Jiang, MEng, Department of Health Sciences and Technology, Harvard Medical School and Massachusetts Institute of Technology, 45 Carleton Street, E25-545, Cambridge, MA 02142, United States (jiangs@mit.edu)

Abstract

Objective: Leverage electronic health record (EHR) audit logs to develop a machine learning (ML) model that predicts which notes a clinician wants to review when seeing oncology patients.

Materials and Methods: We trained logistic regression models using note metadata and a Term Frequency Inverse Document Frequency (TF-IDF) text representation. We evaluated performance with precision, recall, F1, AUC, and a clinical qualitative assessment.

Results: The metadata only model achieved an AUC 0.930 and the metadata and TF-IDF model an AUC 0.937. Qualitative assessment revealed a need for better text representation and to further customize predictions for the user.

Discussion: Our model effectively surfaces the top 10 notes a clinician wants to review when seeing an oncology patient. Further studies can characterize different types of clinician users and better tailor the task for different care settings.

Conclusion: EHR audit logs can provide important relevance data for training ML models that assist with note-writing in the oncology setting.

Key words: machine learning; natural language processing; electronic health record; note writing.

Introduction

The amount of electronic health record (EHR) data for patients is rapidly increasing, making it harder than ever for clinicians to comprehensively review and synthesize information for optimal patient care.[1] While EHRs contain various types of data, much of the valuable information about a patient's story is found in free-text notes, which are often bloated due to their need to satisfy compliance and billing while also serving as clinical communication.[2] Clinicians are spending more time reading and writing in EHRs, which is leading to burnout and medical errors.[3–5] Recognizing this crisis, the American Medical Informatics Association announced its 25×5 initiative in 2022, a call to action to find innovative ways to reduce documentation burden by 25% over the next 5 years.[6]

Machine learning (ML) methods to automatically extract, analyze, and summarize information from the EHR offer a promising solution in this field.[7,8] Previous work has used natural language processing (NLP) to assist with structured data retrieval, speech recognition, and note generation.[9–13] However, these methods do not support clinicians in the laborious process of reviewing free-text notes, a part of chart review that is essential to constructing the patient story.[14–16] Clinician feedback on note-writing tools has emphasized the need to quickly and dynamically find relevant information in notes and the importance of reading those notes for clinical context.[11,17] No prior tools have attempted to predict which notes clinicians want to read. New work evaluating EHR audit logs, which capture granular information about clinician activity, show that logging data may offer an important signal for relevance when using ML to build a note retrieval tool.[18,19]

In this work, we leverage EHR audit logs to train a ML model to predict and display relevant notes for review when a clinician sees a patient.[20] We focus on the oncology setting in our proof of concept since these medically complex patients have regular follow up with many different providers and are likely to have a large quantity of notes for review.

Methods

Our dataset is derived from a large urban academic medical center with an audit logging system that captures information

Received: December 10, 2023; **Revised:** April 5, 2024; **Editorial Decision:** April 11, 2024; **Accepted:** April 25, 2024

on which users read and write notes, and when. The Beth Israel Deaconess Medical Center Institutional Review Board reviewed and approved this study.

The proactive information retrieval task

We considered proactive information retrieval as a binary classification task of whether a note should be retrieved or not, given the current writing context (Figure 1). In the rest of this work, "source documents" represent all available notes for review in the EHR and "written note" represents the most recently written oncology note. A note writing "session" is defined as the time between successively written oncology notes for 1 patient.

Formally, given the most recently written note $w^{(i)}$, the model predicts $y_j^{(i+1)} \in \{0,1\}$ to denote whether to retrieve a source document $d_j^{(i+1)}$ from a corpus of documents $D^{(i+1)}$ available in session S_{i+1}. We considered documents created before the start of $w^{(i+1)}$ to be available for session S_{i+1}. The candidate document set $D^{(i)}$ is updated across sessions to include newly created notes. Once the current oncology note is completed, it becomes an available source document for the next writing session. See Supplement A for inclusion criteria.

Modeling choices and feature construction

We used logistic regression (LR) models to predict whether to retrieve source documents because LR is interpretable and can provide valuable insights for understanding the data. The models used a combination of 3 types of features: text from the source documents and the previously written oncology note, metadata from source documents such as creation time and clinical service, and the document type of the written note (Table 1). All metadata is visible to the clinician during chart review, except the feature describing how many times the source document was read in previous sessions. User studies showed that high-yield, relevant notes were read repeatedly across writing sessions. We trained one model using only metadata from source documents and the written note, and another that added a text representation of the source documents and written note.

We represented all text using a bag-of-words (BoW) method known as Term Frequency Inverse Document

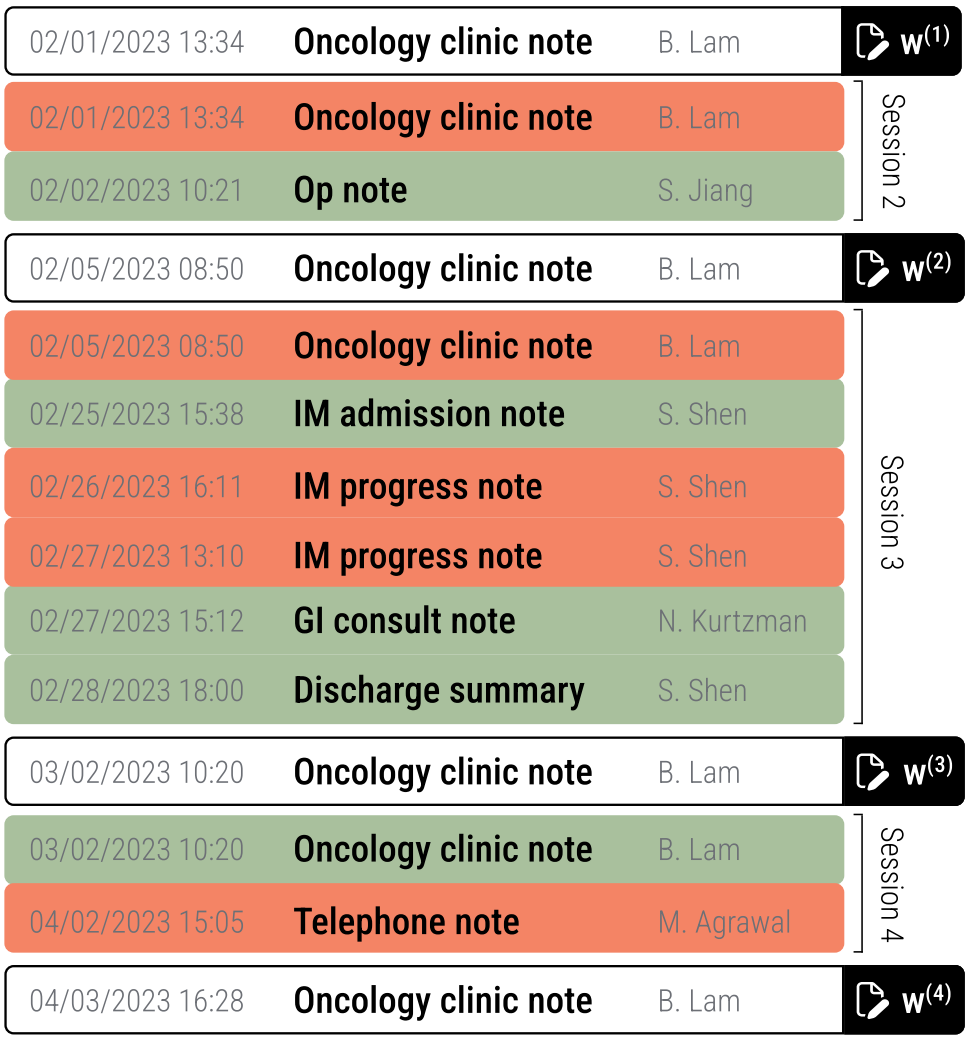

Figure 1. Visualization of the proactive information retrieval task. Each row represents a unique note defined by the written time, name, and author. The oncology clinic note $w^{(i)}$ serves as the writing context for the prediction of relevant documents for oncology clinic note $w^{(i+1)}$. The time between oncology clinic notes is defined as a session where source documents are created. In each session, the green documents are read in the time since the previous oncology clinic note was written, and the red documents were available but not read. After an oncology clinic note is completed, it becomes an available document in the next session.

Table 1. List of all available features and their dimensions in the model.

Feature description	Dimensions
Written note information	
Written note type (eg, progress note, initial note, telephone note, etc.)	2
Source document metadata	
Current session time t_i - source document creation time	6
How recent the source document is out of all available documents for a session	5
How many times the source document was read in previous sessions	5
Clinical service of the source document (eg, oncology, cardiology, etc.)	78
Source document type (eg, progress note, initial note, telephone note, etc.)	15
Source document and written note text	
Bag of words features for the source document text	21 000
Bag of words features for the written note text	21 000

Frequency (TF-IDF). The document is encoded as a vector that captures the presence of certain n-grams in the text. We include the top 20 000 unigrams that occur at least 10 times across the corpus and add 500 bigrams and 500 trigrams selected based on their Pointwise Mutual Information (PMI) score.

We converted the dataset into a training corpus $\{(x, y)\}$, where x is a concatenation of the features described above. For each session in a patient's visit, we created $|D^{(i)}|$ samples as the model generates a prediction $y_j^{(i)}$ per available source document $d_j^{(i)}$. We partitioned the data into training (80%) and testing (20%) sets based on patient IDs to avoid data leakage. For training, we used random undersampling of the majority class for each fold in the 5-fold cross validation to account for data imbalance. We grouped all non-categorical features into intervals and converted them to one-hot vectors.

Evaluation approach

We used traditional classification performance metrics and information retrieval metrics. Precision@k indicates how many items in the top k results were relevant, recall@k indicates how many relevant results were shown in the top k out of all relevant results, and F1@k is the harmonic mean of precision@k and recall@k. Based on clinician input, we chose $k = 10$ to present a suitable range of relevant documents. These top 10 notes are displayed chronologically, reflecting how clinicians typically gather information. We explain our reasoning for these choices in Supplement B. We reviewed 5 patients from the held-out set with 2 clinicians for qualitative evaluation. Each clinician was provided with the top 10 predicted notes and asked to "think aloud" as they judged if the note was relevant to their current note-writing session. The clinicians were asked to act as if they were writing a note for an oncology follow up visit and were allowed to conduct further chart review to assess if they needed additional documentation before writing their note.

Results

Our dataset included 66 762 positively labeled examples (notes that were read) and 1 648 068 negatively labeled (notes that were not read) for 1800 patients over 2 months. An analysis of source document metadata showed that positively labeled examples tend to be notes that were created since the last writing session. However, a significant portion of positively labeled notes are older than the most recent 10 documents (Table S1).

Table 2. Classification evaluation of ML models on held-out data and information retrieval evaluation metrics of predicted relevant documents.

	Metadata only	TF-IDF only	Metadata+TF-IDF
Precision	0.277	0.196	0.268
Recall	0.816	0.862	0.836
F1	0.414	0.320	0.406
AUC	0.930	0.910	0.937
Precision@10	0.522	0.361	0.524
Recall@10	0.610	0.434	0.614
F1@10	0.484	0.359	0.486

The metadata+TF-IDF model achieved an Area under the ROC curve (AUC) of 0.937, recall of 0.836 and precision@10 of 0.524 (Table 2). Thus, of the top 10 documents suggested, more than half will be relevant to the user. The recall@10 of 0.614 indicates that most relevant documents for a particular writing session are surfaced.

The most predictive features for the models are shown in Tables S2 and S3. Metadata, particularly written note time and note type, were strong predictors of whether the note will be read. Notes written by a surgical service were more likely to be read, while notes written by primary management services (emergency medicine, hospital medicine) and ancillary services (nursing, respiratory care, case management) were less likely to be read. The most predictive TF-IDF features suggested that phrases reflecting change ("we discussed," "follow up") indicate notes that were more likely to be read.

In the qualitative assessment of 5 cases, 38 of 50 predicted notes had been read by users. Clinician 1 judged 39 and Clinician 2 judged 34 of the 50 predicted notes relevant for current note-writing (Tables S4 and S5). In all cases, no other notes were identified as relevant on further chart review. Some notes were deemed irrelevant because they contained information already covered in a separately predicted note. Telephone notes in one case were irrelevant because they contained logistical information, while in another case were identified as relevant because they portrayed a patient's report of worsening health prior to a hospital admission. Qualitative evaluation differed most on cases 4 and 5. For case 4, Clinician 1 determined multiple cardiology progress notes to be relevant because of incrementally new information, while Clinician 2 determined these notes to be repetitive. For case 5, Clinician 1 determined oncology nursing notes to be repetitive, while Clinician 2 determined them relevant.

Discussion

Using EHR audit logs, we develop a ML model that evaluates the text and metadata of all available notes to predict the top 10 notes a clinician would want to review when seeing a patient. Our framework enables dynamic workflow support in the setting of complex patient care. The best performing model using metadata with a TF-IDF text representation achieved an AUC 0.937. In a qualitative assessment, the clinician judged more notes relevant for note-writing than had been read by users, suggesting the model is not only

predicting current state chart review, but potentially an improved state as well.

The qualitative assessment revealed nuances of chart review that can inform future iterations of modeling. Both clinicians identified "milestone" notes such as the most recent oncology progress note, a telephone call prior to a hospital admission, and a discharge summary as critical to information gathering. These note types could be encoded as prediction rules in a future model. The clinicians also noted different information needs when writing an inpatient versus an outpatient note. During one case's hospitalization, daily nursing progress notes were read by a primary provider because there are fewer notes to review between note-writing sessions. In the outpatient setting, a primary provider may be faced with many more notes between note-writing sessions, and what was important day-to-day may no longer be critical to patient care. Differences in qualitative evaluation between the 2 clinicians highlight inter-user variability. Future work should explore user-specific predictions that consider the user's role, specialty, and care setting, all of which demand different information needs.

Pre-trained transformers may better represent complex relationships between the clinical context and the note text.[21,22] Other NLP models trained for different tasks have shown that preprocessing a clinical note and only using dynamically changing sections can improve performance.[23,24] As new policy encourages clinicians to write shorter notes,[25] NLP model performance may improve.

Our study has several limitations, but the findings prompt exciting avenues for future work. We focused on the oncology setting because it includes medically complex patients who have frequent contact with the medical system. Clinicians in other specialties may have different approaches to chart review and note-writing. This work focused on the primary clinician workflow, but efforts need to be made to alleviate documentation burden for other healthcare staff.[26] Future models should incorporate other types of data, such as laboratory values and reports.

Positive labels were defined using prior audit logs, and while this strategy is novel, logs represent the current state—which may include bias for more recent notes written by physicians—rather than the ideal state of information gathering. During our initial user studies observing clinician workflows, we found that clinicians placed high value on metadata. Using metadata as features enables the learned models to mimic these existing clinician information gathering behaviors. Using NLP features derived from the notes themselves provides the models an opportunity to surface relevant information that clinicians may not currently be finding. Future work should consider augmenting audit log data with explicit labels of relevance for training data, which may enable the learned models to surface relevant information that clinicians are not currently able to find.

Finally, our model supports information gathering between visits but not during visits; for example, if a clinician learns new information from a patient and subsequently updates their note, an improved model could use this information to surface new data for real-time review. More granular audit logs are needed to support this iteration of our dynamic information retrieval system.

Conclusion

Our work is an effective proof of concept demonstrating that ML can be used to retrieve relevant notes in the EHR for review when clinicians see patients. In an era of increasing documentation burden and evolving NLP tools, our work demonstrates potential for provider-oriented use of novel technology.

Acknowledgments

The authors would like to thank Luke Murray for facilitating the research collaboration, Jerry Liu for contributing to the qualitative evaluation, and members of the MIT Clinical Machine Learning Group for their valuable feedback.

Author contributions

Sharon Jiang, Barbara D. Lam, Monica Agrawal, and Shannon Shen conceived and developed the computational framework and data analysis. Barbara D. Lam, Nicholas Kurtzman, and Steven Horng provided clinical interpretation of the results. Steven Horng, David R. Karger, and David Sontag guided the overall direction and planning. All authors provided critical feedback and helped shape the research, analysis, and manuscript.

Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

Funding

This work was supported by an award from the MIT Abdul Latif Jameel Clinic for Machine Learning in Health (J-Clinic), the National Science Foundation (NSF award no. IIS-2205320), the Machine Learning Core at Beth Israel Deaconess Medical Center, the MIT Deshpande Center, and the MachineLearningApplications@CSAIL initiative.

Conflicts of interest

The authors declare no competing interests.

Data availability

We are not planning to publicly release this dataset, which contains audit logs of user activities on real patients that cannot be reliably de-identified.

References

1. Varpio L, Rashotte J, Day K, King J, Kuziemsky C, Parush A. The EHR and building the patient's story: a qualitative investigation of how EHR use obstructs a vital clinical activity. *Int J Med Inform*. 2015;84(12):1019-1028. https://doi.org/10.1016/j.ijmedinf.2015.09.004
2. Steinkamp J, Kantrowitz JJ, Airan-Javia S. Prevalence and sources of duplicate information in the electronic medical record. *JAMA Netw Open*. 2022;5(9):e2233348. https://doi.org/10.1001/jamanetworkopen.2022.33348
3. Yan Q, Jiang Z, Harbin Z, Tolbert PH, Davies MG. Exploring the relationship between electronic health records and provider burnout: a systematic review. *J Am Med Inform Assoc*. 2021;28(5):1009-1021. https://doi.org/10.1093/jamia/ocab009
4. Nijor S, Rallis G, Lad N, Gokcen E. Patient safety issues from information overload in electronic medical records. *J Patient Saf*.

2022;18(6):e999-e1003. https://doi.org/10.1097/PTS.0000000000001002
5. Moy AJ, Schwartz JM, Chen R, et al. Measurement of clinical documentation burden among physicians and nurses using electronic health records: a scoping review. *J Am Med Inform Assoc*. 2021;28(5):998-1008.
6. Association AMI. AMIA and pacesetters comprehensively tackle burden reduction in healthcare. December 7, 2022. https://amia.org/news-publications/amia-and-pacesetters-comprehensively-tackle-burden-reduction-healthcare
7. Alsentzer E, Kim A. Extractive summarization of EHR discharge notes. arXiv. October 26, 2018. Accessed September 3, 2023. https://arxiv.org/pdf/1810.12085.pdf
8. Hossain E, Rana R, Higgins N, et al. Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review. *Comput Biol Med*. 2023;155:106649. https://doi.org/10.1016/j.compbiomed.2023.106649
9. Klann JG, Szolovits P. An intelligent listening framework for capturing encounter notes from a doctor-patient dialog. *BMC Med Inform Decis Mak*. 2009;9(Suppl 1):S3. https://doi.org/10.1186/1472-6947-9-S1-S3
10. Nuance. Explore Nuance DAX for clinicians. Accessed September 3, 2023. https://www.nuance.com/healthcare/dragon-ai-clinical-solutions/dax-copilot/explore-dax-for-clinicians.html
11. Wang J, Yang J, Zhang H, et al. PhenoPad: building AI enabled note-taking interfaces for patient encounters. *NPJ Digit Med*. 2022;5(1):12. https://doi.org/10.1038/s41746-021-00555-9
12. Liu PJ. Learning to write notes in electronic health records. arXiv. August 8, 2018. Accessed September 3, 2023. https://arxiv.org/pdf/1808.02622.pdf
13. King AJ, Cooper GF, Clermont G, et al. Using machine learning to selectively highlight patient information. *J Biomed Inform*. 2019;100:103327. https://doi.org/10.1016/j.jbi.2019.103327
14. Farri O, Pieckiewicz DS, Rahman AS, Adam TJ, Pakhomov SV, Melton GB. A qualitative analysis of EHR clinical document synthesis by clinicians. *AMIA Annu Symp Proc*. 2012;2012:1211-1220.
15. Hultman GM, Marquard JL, Lindemann E, Arsoniadis E, Pakhomov S, Melton GB. Challenges and opportunities to improve the clinician experience reviewing electronic progress notes. *Appl Clin Inform*. 2019;10(3):446-453. https://doi.org/10.1055/s-0039-1692164
16. Hripcsak G, Vawdrey DK, Fred MR, Bostwick SB. Use of electronic clinical documentation: time spent and team interactions. *J Am Med Inform Assoc*. 2011;18(2):112-117. https://doi.org/10.1136/jamia.2010.008441
17. Sultanum N, Naeem F, Brudno M, Chevalier F. ChartWalk: navigating large collections of text notes in electronic health records for clinical chart review. *IEEE Trans Vis Comput Graph*. 2023;29(1):1244-1254. https://doi.org/10.1109/TVCG.2022.3209444
18. Huilgol YS, Adler-Milstein J, Ivey SL, Hong JC. Opportunities to use electronic health record audit logs to improve cancer care. *Cancer Med*. 2022;11(17):3296-3303. https://doi.org/10.1002/cam4.4690
19. Rule A, Chiang MF, Hribar MR. Using electronic health record audit logs to study clinical activity: a systematic review of aims, measures, and methods. *J Am Med Inform Assoc*. 2020;27(3):480-490. https://doi.org/10.1093/jamia/ocz196
20. Jiang S, Shen S, Agrawal M, et al. Conceptualizing machine learning for dynamic information retrieval of electronic health record notes. In: *Proceedings of the 8th Machine Learning for Healthcare Conference 2023*. PMLR; 2023;219:343-359.
21. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234-1240. https://doi.org/10.1093/bioinformatics/btz682
22. Devlin J, Chang M-W, Lee K, Toutanova K. arXiv. May 24, 2019. BERT: pre-training on deep bidirectional transformers for language understanding. Accessed September 3, 2023. https://arxiv.org/pdf/1810.04805.pdf
23. Liu J, Capurro D, Nguyen A, Verspoor K. "Note Bloat" impacts deep learning-based NLP models for clinical prediction tasks. *J Biomed Inform*. 2022;133:104149. https://doi.org/10.1016/j.jbi.2022.104149
24. Fiszman M, Haug PJ, Frederick PR. Automatic extraction of PIOPED interpretations from ventilation/perfusion lung scan reports. *Proc AMIA Symp*. 1998:860-864.
25. Association AM. CPT evaluation and management (E/M). Accessed September 3, 2023. https://www.ama-assn.org/practice-management/cpt/cpt-evaluation-and-management
26. De Groot K, De Veer AJE, Munster AM, Francke AL, Paans W. Nursing documentation and its relationship with perceived nursing workload: a mixed-methods study among community nurses. *BMC Nurs*. 2022;21(1):34. https://doi.org/10.1186/s12912-022-00811-7

Journal of the American Medical Informatics Association, 2024, 31, 1578–1582
https://doi.org/10.1093/jamia/ocae092
Brief Communication