Efficient Prediction of COPD Outcomes from Clinical Notes Using Filtered Input and Language Model Embeddings

Mohammed Al-Garadi[1], Sharon E. Davis[1], Michael E. Matheny[1], Dax Westerman[1], Bradley W. Richmond[1], Adrienne K. Conger[1], Thomas A. Lasko[1], Iben M. Ricket [2], Laura M. Paulin [3], Jeremiah R. Brown [2], Ruth M. Reeves[1]

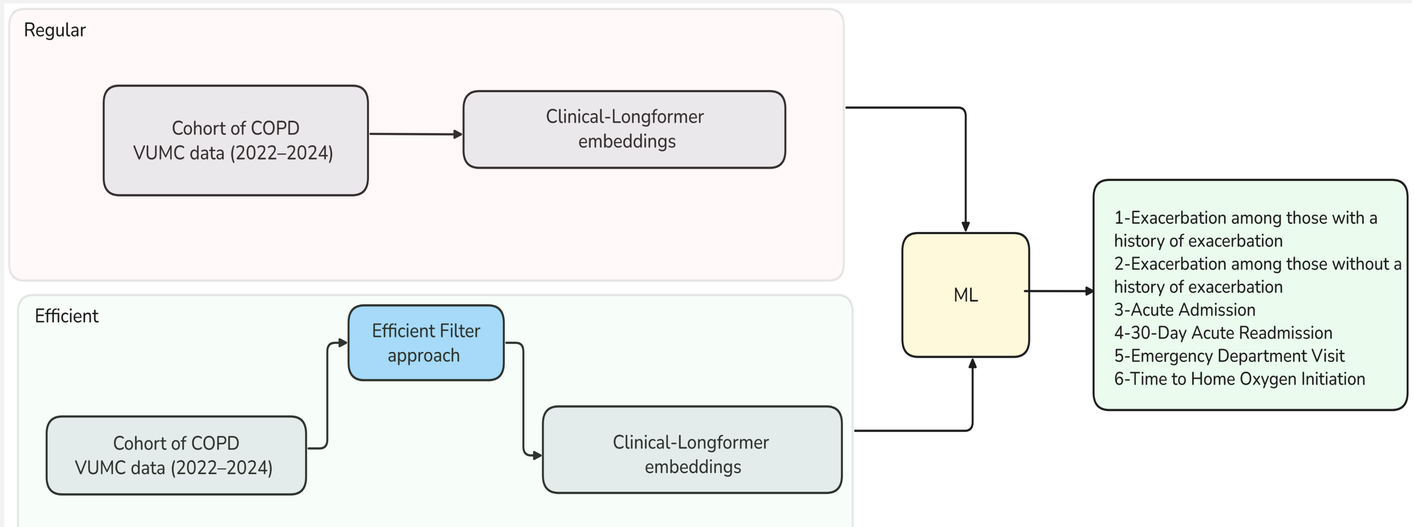
[1]Vanderbilt University Medical Center, Nashville, TN; [2]Dartmouth Center for Implementation Science Geisel School of Medicine, Lebanon, NH; [3]Department of Pulmonology, Dartmouth Hitchcock Medical Center, Lebanon, NH

INTRODUCTION

Timely prediction of adverse outcomes in patients with chronic obstructive pulmonary disease (COPD) is essential for effective clinical decision-making. Clinical notes contain valuable information for anticipating such events; however, their unstructured, high-volume nature presents significant challenges for predictive modeling. Directly applying advanced language models to large electronic health record (EHR) datasets is computationally intensive. To overcome this, we developed and evaluated a two-stage framework that first filters irrelevant documents and then leverages language model embeddings to efficiently predict multiple COPD-related outcomes.

METHODS

We developed COPD outcome prediction models using VUMC data (2022–2024) for six outcomes: exacerbation, ED visits, hospitalizations, readmissions, and home oxygen use. Outcomes were defined within 30 days of a non-urgent visit. Notes from the prior two years were used, with exacerbations modeled by prior history. Two pipelines:- (1) Clinical-Longformer embeddings of all notes and (2) filtered notes before embedding—were compared using XGBoost. Each task included 4,000 positive and 4,000 negative encounters (70/20/10 split), with 95% CIs from bootstrapping.



RESULTS

Across all six COPD-related outcomes, the filtered pipeline achieved AUC performance comparable to the unfiltered model. Document filtering reduced the volume of clinical notes by over 42%, markedly lowering computational demands. Although a modest AUC decline was observed in all but one outcome, the difference remained small and acceptable for practical implementation (Table 1).

Outcome (The event occurs within 30 days)	Without document filter (AUC, 95% CI)	With document filter (AUC, 95% CI)
Exacerbation among those with a history of exacerbation	0.758 [0.752, 0.761]	0.749 [0.748, 0.757]
Exacerbation among those without a history of exacerbation	0.708 [0.701, 0.712]	0.692 [0.687, 0.707]
Acute Admission	0.782 [0.771, 0.793]	0.772 [0.769, 0.789]
30-Day Acute Readmission	0.778 [0.764, 0.791]	0.770 [0.761, 0.790]
Emergency Department Visit	0.769 [0.762, 0.785]	0.777 [0.765, 0.789]
Time to Home Oxygen Initiation	0.670 [0.665, 0.684]	0.665 [0.661, 0.681]

Discussion

- Combining a lightweight supervised filter with Clinical-Longformer enabled accurate and efficient COPD outcome prediction.
- Non-inferior performance to pipeline integrating all text
- 42% reduction in volume for processing, lowering cost & time
- The two-stage design offers a scalable, clinically actionable blueprint for LLM use in healthcare.

Limitations	Future Work
• Single-site, small-scale study may not capture documentation variability.	• Conduct multi-site validation to assess generalizability.
• Manual labeling introduces potential subjectivity.	• Automate labeling and expand annotated datasets.
• Model used only text data, excluding structured EHR variables.	• Integrate structured data to improve prediction accuracy.
• Initial retrospective design limits real-world evaluation.	• Explore real-time deployment within clinical workflows.

CONCLUSION

The proposed framework demonstrates that selective note filtering can enhance computational efficiency while maintaining predictive accuracy, underscoring its potential for scalable LLM deployment in clinical settings. Further validation across diverse institutions and integration with structured data are warranted to ensure robustness and translational utility.

VANDERBILT UNIVERSITY MEDICAL CENTER

Department of Biomedical Informatics

Take a picture to download the full text

Acknowledgment: Supported by NIH (R01 HL157130-01).

Mohammed.a.al-garadi@vumc.org