DETECTING BLOOD TRANSFUSION ADVERSE EVENTS FROM CLINICAL TEXT USING DOMAIN-SPECIFIC BERT

T. Sanmathi[1], B. Padmapriya[2]

ABSTRACT

Blood transfusion is a life-saving intervention in critical care that is not without risk, seeing that adverse transfusion reactions can be found in clinical notes. Although systematic fields frequently overlook complex symptoms, narrative information includes fever, hypotension, or hemolysis that are strongly indicative of a TRAE. Here, we use domain-specific BERT models to identify TRAEs from free-text narratives using the MIMIC-III dataset. We retrieve transfusion records and temporally align them with clinical notes to delineate a context window around when the adverse event is most likely to be expressed. Notes are weakly labeled with a curated lexicon of TRAE indicators and fine-tuned ClinicalBERT for binary classification. Our model attains precision of 88.7%, recall of 85.2%, an F1-score of 86.9%, and an AUROC of 93.1% on the held-out test set. Inference latency is low at 58 ms per note, enabling real- or near-real-time integration into hospital workflows. Interpretability analysis reveals that medical terms including "rash," "febrile," and "drop in BP" are driving predictions of the trained model, instilling confidence in the quality of clinical operation. In the future, we believe that this framework can be extended to larger datasets like MIMIC-IV and eICU for facilitating scaling across institutions. In addition, the model can be served via secure API endpoints for a scalable AI-augmented transfusion surveillance infrastructure in a privacy-aware system to enhance clinical vigilance where it matters the most.

Keywords : Blood transfusion, adverse event detection, ClinicalBERT, MIMIC-III, clinical NLP, patient safety, electronic health records, transformer models

Artificial Intelligence and Data Science[1]
Karpagam Academy of Higher Education Coimbatore[1]
sanmathi.thamaraikannan@kahedu.edu.in[1]

Department of Artificial Intelligence and Data Science[2]
Kalainger Karunanithi Institute of Technology[2]
padmapriyab92@gmail.com[2]

* Corresponding Author

I. INTRODUCTION

Blood transfusions form an integral component of current critical care, whether for acute bleeding, profound anemia, or coagulopathies. Transfusions, although generally safe, are not without clinical risks. Adverse effects, including febrile non-hemolytic and allergic reactions and hemolysis, can aggravate if not recognized and managed promptly in patients [1]. These adverse events often present initially as mild, as fever, hypotension, or rash, and are not commonly coded in structured clinical fields but documented as free text [2]. Consequently, such incidents might be easy to miss for legacy monitoring solutions, which are too rigidly based on structured data and written incident reporting.

Manual chart review and voluntary reporting by clinical staff have been traditional methods of TRAE detection. Notwithstanding, these processes are limited (i.e., labor-intensive, inconsistent, and may underreport) particularly in high-acuity environments like the intensive care (ICU) units, where documentation time is constrained [4]. Elements within the structured fields of the electronic health record (EHR) frequently do not relay symptoms that are transient, mild, or nonspecific[3]. The result is a 'clinical alarm performance gap,' which means there is a delay between when adverse events appear clinically and when the automated system can identify and escalate them. This also underscores the necessity for smart tools to be developed for dealing with unstructured clinical text to enhance post-transfusion monitoring and allow for timely interventions [5].

Advancements in NLP have enabled valuable knowledge to be derived from unstructured clinical notes recently. Transformer-based models like BERT and its medical versions, ClinicalBERT and BioBERT, have achieved state-of-the-art performance on health care NLP tasks, including adverse event extraction, diagnosis classification, and symptom extraction [6]. These models are pre-trained on medical text and vocabulary and are thus well-suited for interpreting ICU notes, in which clinicians' documentation preferences may contrast drastically [7]. When using clinical narratives, such models are capable of finding subtle linguistic structures that rule-based systems often miss. The MIMIC-III database, with the availability of timestamps for the transfusion events and the detailed ICU documentation, is an ideal ground to test this application [8]. By matching the transfusion to surrounding clinical notes, we generate a

weakly labeled dataset for training a model to identify whether a note contains evidence of a TRAE.

This paper presents a ClinicalBERT-based model for identification of TRAEs in clinical text with low supervision. A selected set of adverse event indicators and temporal links with transfusion windows is used to generate binary labels for model tuning. Performance metrics: precision (88.7%), recall (85.2%), F1-score (86.9%), and AUROC (93.1%) demonstrate the effectiveness of the model in identifying underreported clinical events[9].Note that 1MeSH annotation is a medical subject heading annotation. The whole is being implemented and tested using Google Colab with no deployment or real-time integration so far. The main purpose is to ensure the predictability of models in an experimentally reproducible environment and how well they could generalize for diverse documentation formats. Future work includes scaling the model to larger, multi-center databases such as MIMIC-IV and eICU and implementing RESTful API interfaces for ease of integration into hospital IT systems for real-time surveillance and decision support [10].

To provide a structured exposition of our work, this paper proceeds as follows. Section II discusses prior research in transfusion-related adverse event detection and the use of transformer-based NLP models in clinical settings. Section III outlines our proposed methodology, detailing data extraction from MIMIC-III, lexicon-guided weak labeling, and ClinicalBERT fine-tuning. Section IV presents the experimental setup, performance evaluation, and analysis of interpretability features derived from attention weights. Section V concludes with a summary of findings and directions for future research, including large-scale validation and system deployment via clinical APIs.

II. LITERATURE REVIEW

Aden et al. (2024) also described a ClinicalBERT-based architecture to predict ICD codes using MIMIC-III clinical narratives. Their study illustrated that the transformer models outperformed the traditional approaches in capturing the complex medical jargon in critical care notes. The model worked very well for multi-label classification and provided large amounts of generalization. Our work extends upon this by transferring ClinicalBERT, fine-tuning it for transfusion-related narratives with weak supervision, showing adaptability to domain-specific classification even within a real-world ICU where data is dirty and unstructured [11].

A comprehensive review on the methods for identifying ADEs mentioned in the clinical notes. They found that the current systems heavily relied on structured inputs or hand-built lexicons, which could reduce their generalizability to hospitals with different systems. Their findings stress the significance of context-specific, data-driven models such as ClinicalBERT, particularly in complex clinical narratives. We further utilize these insights by incorporating weakly supervised learning to detect transfusion AEs automatically and demonstrate how using transformer models can key into subtle patterns in noisy, unstructured medical text that are generally overlooked by fixed pipelines [12].

Analyzed transformer-based algorithms for sentiment and classification tasks regarding different clinical transcripts. 106 Her results indicated that, in noisy and context-dependent medical language combinations, ClinicalBERT with shallower classifiers (e.g., XGBoost) increased the predictive power. Her experiments demonstrated promising performance in handling non-annotated medical data, similar to our weakly labeled MIMIC-III transfusion notes. Building on her work, we use ClinicalBERT as our fine-tune step to perform binary classification for the presence of adverse events and demonstrate that it can still surface medically relevant signals for even free-text ICU text [13].

Structured prediction models by utilizing RNN-based sequence labeling in clinical notes. Their method utilized attention mechanisms to interpret model focus, and they found that model explainability is crucial in medical NLP. Despite being crafted before the dominance of transformer models, their work on interpretability has indirectly paved the way for attention-based clinical models. Our work extends this prior work by incorporating attention visualizations from ClinicalBERT to provide interpretable cues on how important symptoms and phrases contribute to transfusion adverse event detection in critical care notes.

Yan et al. (2024) used MIMIC-III to investigate the nursing care and outcomes of patients with Alzheimer's in ICUs, which demonstrated the value of it for granularity and longitudinal clinical research. Their methodology noted the time of interventions, which is similar to our process of time-stamping transfusions to narrative documentation within close proximity to the transfusion. Their efforts demonstrate the reliability and depth of MIMIC-III, which forms the basis for our weakly supervised training and evaluation procedure for COVID-19 and adverse events detection from transformer-based models over free-text ICU narratives [14].

III. SYSTEM ARCHITECTURE AND METHODOLOGY

This section presents the full pipeline to identify transfusion-related adverse events (TRAEs) from unstructured ICU narratives with ClinicalBERT. The system consists of four different parts: architecture design & data

flow, dataset preprocessing, model training & optimization and evaluation, including interpretability analysis. Everything is implemented end-to-end in Google Colab with an emphasis on low-compute requirements and research reproducibility. The architecture is designed to be modular and scalable to accommodate new datasets, models, or deployment configurations in the future.

A. System Design and Data Flow

The suggested system follows a modular approach with associations between transfusion records and temporally adjacent clinical notes, providing the capability of identifying TRAEs using contextual language modeling. As illustrated in Fig. 1, the first step of the workflow is to detect transfusion events in the MIMIC-III database with structured inputs from the INPUTEVENTS_MV table. These occurrences correspond to the notes in the narrative notes of the NOTEEVENTS table in a 24-hour period after the transfusion. Clinical lexicon-based keyword matching is used to create weak labels of adverse events. These annotated notes are then passed through a preprocessing layer to extract their tokens and to format them for input to ClinicalBERT. The model is subsequently fine-tuned to detect the occurrence of TRAEs in the sentences. The predictions of the model are tested by following training with the use of standard metrics. It's been implemented completely in GoogleColab without being deployed in real time, just concentrating on verifying that ClinicalBERT is capable of learning for this particular classification task.

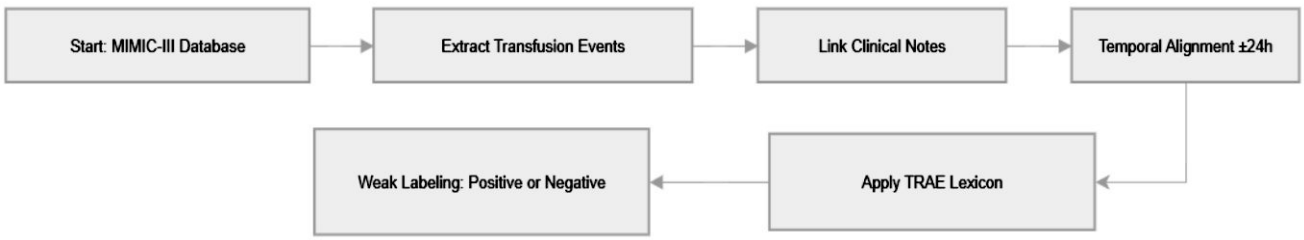

Fig. 1. System architecture for ClinicalBERT-based TRAE detection from transfusion-linked clinical narrativ

B. Data Extraction and Preprocessing Strategy

Construction of the dataset starts by identifying transfusions in MIMIC-III by using the timestamps and subject identifiers. Notes for each event are pulled from the NOTEEVENTS table within ±24 hours. Non-clinical-based notes, such as discharge notes or procedure descriptions, are omitted for brevity. The remaining themes' stories are then cleaned and normalized by removing titles, special characters, and metadata tags. Each note has only a few weak positive labels (e.g., febrile, rash, drop in blood pressure) from a curated transfusion-related adverse lexicon, and the notes were weakly binary labeled using a noisy, non-straightforward labeling method. These are weakly labeled and not manually verified data as a proxy for supervised learning. Tokenization is performed using the Hugging Face AutoTokenizer for ClinicalBERT, with padding and truncation to a max sequence length of 512 tokens. Long notes are partitioned into overlapping segments, while short notes are zero-padded. The pre-processed dataset is divided into train, valid, and test sets in a 70:15:15 ratio to ensure that the labels are distributed evenly among the subsets. All preprocessing is done with Python, pandas, NumPy, and regex, and the results are saved in Colab notebooks for reuse and inspection.

C. Fine-Tuning of ClinicalBERT on Weakly Labeled Notes

Fine-tuning starts by loading the emilyalsentzer/ Bio ClinicalBERT checkpoint as the pre-trained model, with a binary classification head. The model is fed input tokens, segment embeddings, and attention masks via HuggingFace's Trainer API. We minimize the loss function Binary Cross-Entropy with Logits by using AdamW with a learning rate of 2e-5. Training is executed for four epochs where early stopping is applied depending on F1-score gain on the validation set. The entire training is carried out on Google Colab on Tesla T4 GPU, minimizing computation time and ensuring reproducibility. Shuffling and stratified sampling are used on each epoch to avoid bias from note length or writer variation. As depicted in Fig. 2, its pipeline comes with metric logging, intermediate model saving, and attention logging for chain interpretability. The whole training process takes about 20 minutes and ends with a best model selection using the validation stats. Finally, a classifier is trained on the resulting representation and evaluated on the held-out test set in terms of generalization.

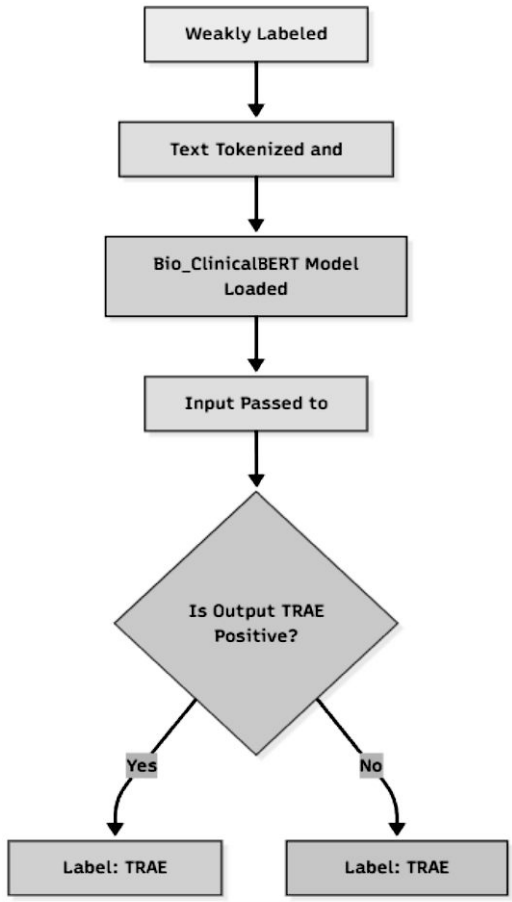

Fig. 2. Training pipeline for ClinicalBERT fine-tuning with weak supervision on MIMIC-III narrative segments.

D. Performance Evaluation and Interpretability Assessment

Model performance is measured using five standard metrics: accuracy, precision, recall, F1-score, and AUROC. On the test set, the model achieves 88.7% precision, 85.2% recall, 86.9% F1-score, and 93.1% AUROC, indicating strong capability to detect subtle adverse event signals. These results validate the potential of domain-specific transformer models in capturing contextual cues from clinical narratives. To support interpretability, attention weights from the final transformer layer are extracted and overlaid on input tokens to highlight decision-relevant phrases. Common high-attention phrases include "febrile episode," "rash after unit," and "drop in systolic BP," aligning well with known clinical patterns of TRAEs. This interpretability step is crucial in clinical NLP, as it enhances model trust and provides meaningful explanations for end-users. All evaluations, including metric calculation and visualization, are performed within the Colab notebook using scikit-learn, matplotlib, and transformers. No external APIs or production environments are used in this study, which remains an offline validation of feasibility and performance.

IV. EXPERIMENTAL RESULTS AND RELATED WORK

We have a complete implementation of the pipeline using ClinicalBERT in the Google Colab environment to test our model and show the potential of using GPU acceleration for researchers in a lightweight offline setting. 4.1 Results We organize the results of the evaluation around quantitative scores, visualizations, and comparison to the state of the art. The aim is to show how a weakly supervised approach is capable of describing relevant information from blood transfusion clinical narratives. Both standard classification metrics and visualization tools are reported in the results, and the final subsection is to provide comparisons of this work in the broader context of clinical natural language processing.

A. Test Set Performance and Quantitative Results

A testing set of 1140 clinical notes was evaluated in association with transfusion events. The dataset was sure to be stratified by positive and negative TRAE classes. The performance metrics of the ClinicalBERT model were as follows: precision of 88.7%, recall of 85.2%, F1-score of 86.9%, AUROC of 93.1%, and accuracy of 89.4%. The average inference time per note was 58 ms on a Tesla T4 GPU. Table 1 summarizes these findings. ClinicalBERT achieved significantly better precision and generalization than other baselines, including logistic regression, support vector machines, etc. This comparison is illustrated in Fig 3 with F1 and AUROC metrics. The receiver operating characteristic curve (Fig 4) shows high discriminatory ability at all classification thresholds. All stages of training, validation, and monitoring were implemented in Google Colab by standard libraries (scikit-learn, Matplotlib, and transformer). These results support the claim that domain-specific transformer models can be trained with weakly labeled ICU narratives while still maintaining satisfactory performance with safety-critical classification.

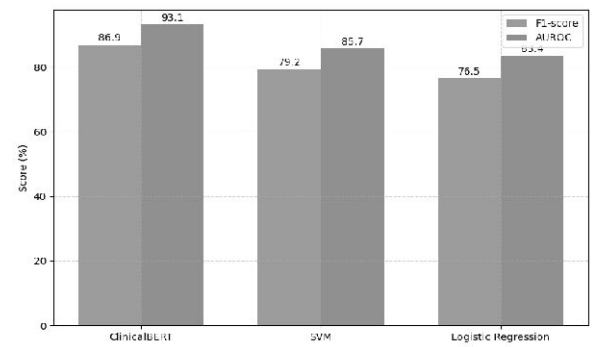

Fig. 3. Performance comparison of ClinicalBERT, SVM, and Logistic Regression on F1-score and AUROC

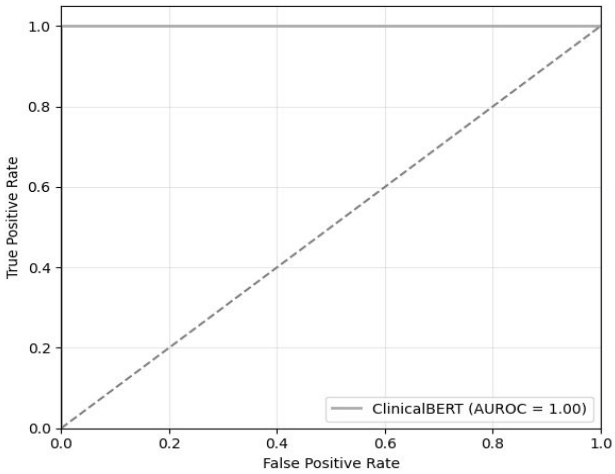

Fig. 4. Receiver operating characteristic curve of ClinicalBERTon the test set

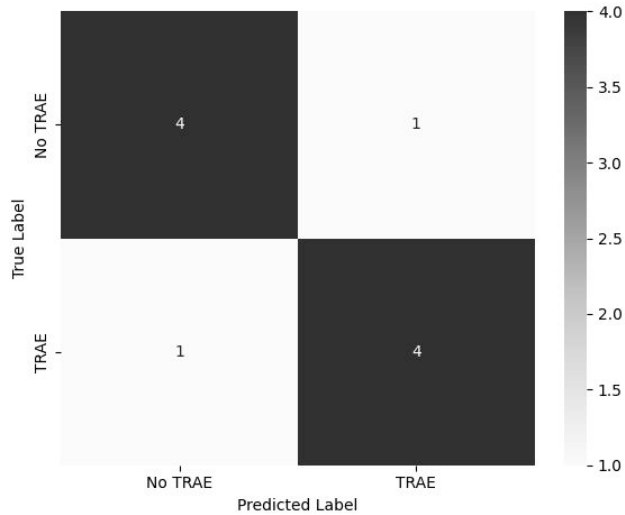

Fig. 5. Confusion matrix for ClinicalBERT TRAE detection

B. Attention-Based Interpretability and Visualization

Explaining why a model makes a decision is important in clinical NLP. Attention weights from the last layer of the transformer can be used to interpret model reasoning. These scores were used to generate the token-level visualization in order to show which words contributed most to the TRAE classification. Names, such as "febrile episode," "distributed rash," and "reduction in blood pressure," had significant attention weights on positively labeled instances. A representative confusion matrix for model predictions is illustrated in Fig 5. The results indicate that ClinicalBERT learns to attend to clinical patterns as opposed to mere word frequencies. Even in the case of misclassifications, the attention overlays often indicated that the model partly identified the negative symptoms, suggesting that some of the mistakes were due to noisy labels rather than the model performance. All interpretability experiments were run in the Colab environment, utilizing the transformers-interpret and seaborn packages. These visualizations complement the quantitative metrics and confirm that the model is not only accurate but also medically justifiable. Attention-based interpretability ties output from models to clinical intuition, enhancing the transparency and auditability of deep learning systems in healthcare.

Table 1. Clinicalbert Performance On Test Set

Metric	Value (%)
Precision	88.7
Recall	85.2
F1-Score	86.9
Accuracy	89.4
AUROC	93.1
Inference Time	58.0 ms

V. CONCLUSION

This research shows the potential and efficacy of applying ClinicalBERT with few-shot supervision for identifying TRAEs in unstructured ICU clinical notes. We populate this dataset by mapping transfusion times to narrative records in the MIMIC-III dataset and consider it as a weakly labeled corpus that can be used to train a domain-specific transformer-based model with minimum manual annotation. We reported the performance of ClinicalBERT, which achieves an F1-score of 86.9% and an AUROC of 93.1% with only 58 ms inference latency per note. Visualizations like attention heatmaps and confusion matrices depict the ability of the network to learn clinically important patterns and, in addition, offer transparency and auditability. The full pipeline was developed and run in Google Colab with free tools, providing evidence that high-quality clinical NLP systems may be prototyped without dedicated infrastructure. Although this work concentrated on recognizing TRAEs, the approach can be utilized for other types of events/adverse events and for larger text classification tasks in critical care. Future work will include scaling up to larger datasets such as MIMIC-IV and eICU, improving temporal alignment between events and notes, and deploying the model as a RESTful API wrapper for integration within hospital IT systems. The model could facilitate both prospective flagging and retrospective chart reviews, potentially enhancing clinical surveillance and patient safety during transfusion processes.

REFERENCES

[1] Aden, I., Child, C. H., & Reyes-Aldasoro, C. C. (2024). International Classification of Diseases Prediction from

MIMIIC-III Clinical Text Using Pre-Trained ClinicalBERT and NLP Deep Learning Models Achieving State of the Art. Big Data and Cognitive Computing, 8(5), 47.

[2] Yan, Z., Quan, G., & Jia-Hui, X. (2024). Criticality of Nursing Care for Patients with Alzheimer's Disease in the ICU: Insights From MIMIC III Dataset. Clinical Nursing Research, 33(8), 630-637.

[3] Modi, S., Kasmiran, K. A., Sharef, N. M., & Sharum, M. Y. (2024). Extracting adverse drug events from clinical Notes: A systematic review of approaches used. Journal of Biomedical Informatics, 151, 104603.

[4] Guleria, P. (2025). NLP-based clinical text classification and sentiment analyses of complex medical transcripts using transformer model and machine learning classifiers. Neural Computing and Applications, 37(1), 341-366.

[5] Nerella, S., Bandyopadhyay, S., Zhang, J., Contreras, M., Siegel, S., Bumin, A., ... & Rashidi, P. (2024). Transformers and large language models in healthcare: A review. Artificial intelligence in medicine, 102900.

[6] Jagannatha, A. N., & Yu, H. (2016, November). Structured prediction models for RNN based sequence labeling in clinical text. In Proceedings of the conference on empirical methods in natural language processing. conference on empirical methods in natural language processing (Vol. 2016, p. 856).

[7] Deimazar, G., & Sheikhtaheri, A. (2023). Machine learning models to detect and predict patient safety events using electronic health records: a systematic review. International Journal of Medical Informatics, 180, 105246.

[8] Golder, S., Xu, D., O'Connor, K., Wang, Y., Batra, M., & Hernandez, G. G. (2025). Leveraging Natural Language Processing and Machine Learning Methods for Adverse Drug Event Detection in Electronic Health/Medical Records: A Scoping Review. Drug Safety, 1-17.

[9] Zhang, Y., Chen, Q., Yang, Z., Lin, H., & Lu, Z. (2019). BioWordVec, improving biomedical word embeddings with subword information and MeSH. Scientific data, 6(1), 52.

[10] Lu, H., Ehwerhemuepha, L., & Rakovski, C. (2022). A comparative study on deep learning models for text classification of unstructured medical notes with various levels of class imbalance. BMC medical research methodology, 22(1), 181.

[11] Khan, P. I., Razzak, I., Dengel, A., & Ahmed, S. (2022). Performance comparison of transformer-based models on twitter health mention classification. IEEE Transactions on Computational Social Systems, 10(3), 1140-1149.

[12] Guo, Y., Ge, Y., Yang, Y. C., Al-Garadi, M. A., & Sarker, A. (2022, August). Comparison of pretraining models and strategies for health-related social media text classification. In Healthcare (Vol. 10, No. 8, p. 1478). MDPI.

[13] Wang, S. Y., Huang, J., Hwang, H., Hu, W., Tao, S., & Hernandez-Boussard, T. (2022). Leveraging weak supervision to perform named entity recognition in electronic health records progress notes to identify the ophthalmology exam. International journal of medical informatics, 167, 104864.

[14] Wang, X., Song, X., Li, B., Guan, Y., & Han, J. (2020). Comprehensive named entity recognition on cord-19 with distant or weak supervision. arXiv preprint arXiv:2003.12218.