Project Proposal: Patient Risk Prediction from Clinical Notes

Team: Transformers MD

Aryan Chaudhary (2023114015)
Ayush Kumar Gupta (2023114001)
Divyash Pipersaniya (2023111006)

1 Introduction and Problem Definition

Thirty-day hospital readmission is a major concern in healthcare. For patients, it often means a decline in health and quality of life. For hospitals, it leads to significant financial penalties under programs like the Hospital Readmissions Reduction Program (HRRP). The core challenge is identifying at-risk patients **before** they are discharged so that targeted interventions can be applied.

Current predictive models often rely on structured data like lab results and diagnosis codes. While useful, these models achieve only moderate success, typically with AUROC scores between 0.6 and 0.8. They miss the rich, nuanced information hidden in unstructured clinical notes—the narratives where clinicians document their reasoning, observations, and concerns. Our project hypothesizes that by applying advanced Natural Language Processing (NLP) techniques to these notes, we can build a more accurate and clinically insightful readmission prediction model.

2 Literature Review and Project Scope

Our initial research shows a clear path for improving readmission prediction.

- **Traditional Baselines:** Simpler models like Logistic Regression using TF-IDF text features establish a performance benchmark, often achieving an AUROC around 0.62 - 0.75. This will be our starting point.
- **Advanced NLP Models:** The literature strongly supports using transformer-based models. **ClinicalBERT**, a model pre-trained specifically on medical notes, has been shown to significantly outperform generic models by better understanding clinical jargon and context, pushing AUROC scores above 0.71.
- **Multimodal Approach:** The most successful models are **hybrid systems** that combine the insights from clinical notes (via NLP) with structured data (demographics, diagnoses, etc.). This fusion of data provides a more complete picture of the patient.

Our project scope is to develop and compare these models in a phased approach:

1. Establish a strong **TF-IDF baseline**.
2. Implement and fine-tune a **ClinicalBERT model** to leverage the unstructured text.
3. Develop a final **hybrid model** that fuses text embeddings with structured data to achieve state-of-the-art performance.

3 Dataset: MIMIC-IV

We will use the **MIMIC-IV dataset**, a large, de-identified public database from the Beth Israel Deaconess Medical Center. It contains comprehensive data for hundreds of thousands of patients from 2008-2019.

- **Key Tables:** We will primarily use the `admissions` table to define our patient cohort and generate readmission labels, and the `noteevents` table, which contains millions of free-text clinical notes (e.g., discharge summaries).
- **Challenges:** The data is de-identified, meaning dates are shifted and personal information is replaced with placeholders like `[**...**]`. Our preprocessing pipeline will need to handle these artifacts to avoid introducing noise into the models.

4 Proposed Methodology

Our methodology is designed to be systematic and robust, progressing from a simple baseline to a complex, state-of-the-art model.

4.1 Step 1: Data Preparation and Cohort Construction

- **Label Generation:** We will define our target label by calculating the time between a patient's discharge and their next admission. If it's ≤ 30 days, the label is 1 (readmitted), otherwise 0. Patients who died in the hospital or their final admission record will be excluded.
- **Text Preprocessing:** Clinical notes will be cleaned by converting text to lowercase, removing de-identification artifacts, and standardizing section headers. The text will then be tokenized using the specific tokenizer for our chosen model (e.g., ClinicalBERT's tokenizer).

4.2 Step 2: Predictive Modeling

- **Phase 1 (Baseline):** We will build a **Logistic Regression** model using **TF-IDF** vectors created from the clinical notes.
- **Phase 2 (Core NLP Model):** We will fine-tune a pre-trained **ClinicalBERT** model. This involves adding a classification layer to the model and training it for a few epochs on our labeled dataset to adapt it for the readmission task.
- **Phase 3 (Hybrid Model):** We will extract text embeddings from the fine-tuned ClinicalBERT model and concatenate them with structured features (age, gender, ICD codes, etc.). This combined feature set will be fed into a powerful final classifier like **XGBoost** or **LightGBM**.

4.3 Step 3: Handling Class Imbalance

Readmission is a rare event, leading to a highly imbalanced dataset. Instead of using techniques like SMOTE which risk generating clinically invalid synthetic data, we will use **class weights** in the model's loss function. This robust method forces the model to pay more attention to the minority (readmission) class without altering the original data distribution.

5 Evaluation Strategy

To ensure our model is clinically useful, we will use appropriate evaluation metrics and focus on interpretability.

- **Metrics:** We will avoid using simple **accuracy**, as it is misleading on imbalanced datasets. Our primary metrics will be the **Area Under the ROC Curve (AUC-ROC)** and the **Area Under the Precision-Recall Curve (AUPRC)**, which are standard for clinical prediction tasks and provide a better measure of a model's discriminative power.
- **Interpretability:** A "black box" model is not useful in a clinical setting. We will use **SHAP (SHapley Additive exPlanations)** to explain our model's predictions. This will allow us to visualize which specific words or phrases in a clinical note (e.g., "patient confused," "limited family support") are pushing the model to flag a patient as high-risk, building trust and utility for clinicians.

6 Project Timeline

Week(s)	**Task**	**Deliverable**
1-2	Data Exploration & Preprocessing Pipeline	A clean, analysis-ready dataset and cohort.
3	Baseline Model (TF-IDF + Logistic Regression)	Initial benchmark performance metrics (AUROC/AUPRC).
4-5	Fine-Tuning ClinicalBERT Model	A trained NLP model for text classification.
6-7	Develop Hybrid Model (ClinicalBERT + Structured Data + XGBoost)	The final, highest-performing predictive model.
8	Model Evaluation & Interpretability Analysis	Performance comparison and SHAP visualizations.
9-10	Final Report Preparation & Presentation	A comprehensive project report and presentation.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). *Attention is all you need*. Advances in neural information processing systems, 30.

[2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv preprint arXiv:1810.04805.

[3] Alsentzer, E., Murphy, J., Boag, W., Weng, W. H., Jindi, D., Naumann, T., & McDermott, M. (2019). *ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission*. arXiv preprint arXiv:1904.05342.

[4] Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., & Mark, R. (2023). *MIMIC-IV (version 2.2)*. PhysioNet. https://doi.org/10.13026/s6n6-xd98.

[5] Johnson, A., Pollard, T., & Mark, R. (2023). *MIMIC-IV-Note: Deidentified free-text clinical notes (version 2.2)*. PhysioNet. https://doi.org/10.13026/1n74-cg13.

[6] Zhu, L., Chen, Y. (2024). *Prompting Large Language Models for Clinical Outcome Prediction using In-Context Learning. Nature Digital Medicine.*