Specialist Clinical Services
Clinical Care Group

NHS
Nottingham
University Hospitals
NHS Trust

# Large Language Model for assessment of value in CTPA examinations – A Project Proposal

Siân Heath, Radiation Physics, Nottingham University Hospitals NHS Trust (NUH)

## Introduction

Radiology departments are facing increased demands, often operating at or beyond capacity. To optimise workflow and ensure timely patient care, it is essential to prioritise scans with the highest diagnostic value. This project focuses on Computed Tomography Pulmonary Angiograms (CTPAs), a key diagnostic tool for suspected pulmonary embolisms (PEs).

Current guidelines often recommend a CTPA for patients with a D-dimer (or age-related D-dimer) score above a threshold. The core objective of this research is to investigate if this threshold can be adjusted to prioritise high-value scans, defined as those that are more likely to lead to a positive PE diagnosis.

To achieve this, a large dataset of patient outcomes (PE or no PE) from CTPA reports is required. Manually extracting this data is a time-consuming and labour-intensive task. Therefore, this project will leverage a large language model (LLM) [1] to automate the extraction of PE outcomes from a substantial number of unstructured CTPA reports. The LLM-generated outcome data will then be used to perform retrospective analysis of the D-dimer threshold's effectiveness, allowing for a data-driven recommendation to optimise CTPA scan utilisation.



[2]

## Aims/Objectives

### Primary aim:

- To develop and validate a large language model (LLM) for the automated extraction of pulmonary embolism (PE) outcomes from a large cohort of CTPA radiology reports.

### Secondary aims:

- To create a gold-standard dataset of CTPA radiology reports, manually annotated by clinical experts to serve as training and validation data for the LLM.
- To train and fine-tune a pre-trained LLM (e.g. BioClinicalBERT or GatorTron) to accurately perform multiclass classification of PE outcomes from free-text radiology reports.
- To evaluate the LLM's performance using standard multiclass metrics including per-class precision, recall, and F1-score (harmonic mean of precision and recall). Compare performance to a manual annotation baseline.
- To utilise the LLM-generated outcome data to perform retrospective analysis and assess the D-dimer threshold's effectiveness in predicting PE diagnosis.
- To propose a data-driven d-dimer threshold that could improve the prioritisation of CTPA scans, thereby enhancing clinical workflow and resource management.

## Methods

### Data Acquisition and Annotation

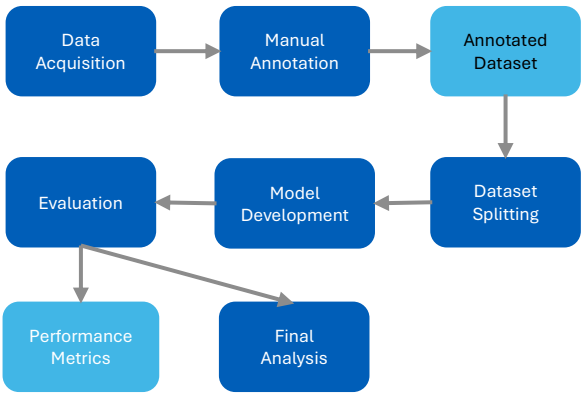
- Data Sourcing: Obtain a dataset of CTPA radiology reports for patients who had a D-dimer test up to 7 days prior.
- Manual Annotation: A subset of these reports will be manually annotated by a clinical expert, labelling each as positive, negative or uncertain for a PE. These labels will form our gold-standard dataset.
- Anonymisation: All patient identifiers will be removed to ensure privacy.
- Dataset Splitting: The gold-standard dataset will be split into a training set (~80%), a validation set (~10%), and a test set (~10%). This will be done on a per-patient basis to ensure no patient appears in more than one set. [3]

### Model Development

- Model Selection: Choose a pre-trained transformer-based LLM specifically designed for clinical text (e.g. GatorTron or BioClinicalBERT).
- Text Pre-processing: The text will be tokenised, broken up, using the model's native tokeniser. For longer reports, the beginning of the text will be truncated to ensure the model focusses on the most clinically relevant information, typically found in the final sections. [3]
- Fine-tuning: Add a new classification layer on top of the pre-trained model and train the entire network on the annotated dataset to perform multiclass classification.

### Evaluation

- Performance Metrics: Evaluate the model's performance on the unseen test set using standard classification metrics:
  - Precision and Recall and F1-score for each class.
  - Accuracy: Report overall and per-class.
  - Confusion Matrix: A table to visualise the performance of the classification model, detailing true/false positives and negatives for the three classes.
- Comparison: Compare the LLM's performance against a simpler baseline model, (e.g. TF-IDF and Logistic Regression) to demonstrate the value of the deep learning approach. [3]




## References

[1] Zhou, B., Yang, G., Shi, Z., & Ma, S. (2022). Natural Language Processing for Smart Healthcare. IEEE Reviews in Biomedical Engineering, 15, 1–17.[
2] Nishio, T., Sugiyama, O., Fujisawa, Y., et al. (2018). Use of pulmonary CT angiography with low tube voltage and low-iodine-concentration contrast agent to diagnose pulmonary embolism. Scientific Reports, 8, Article number: 13917. Figure 1a.
[3] Barlow, S. H. C., Chicklore, S., et al. (2024). Uncertainty-aware automatic TNM staging classification for [18F] Fluorodeoxyglucose PET-CT reports for lung cancer utilising transformer-based language models and multi-task learning. BMC Medical Informatics and Decision Making, 24, Article number: 301.

Kind Inclusive Ambitious One team