Capstone JnJ Midterm report

Kevin Cai, Ruibin Lyu, Matt Xi, Yueming Xu

March 2025

Contents

1 Introduction

1.1 Background

Health Technology Assessment (HTA) decisions are critical to healthcare access and policy development, particularly in the area of drug reimbursement. In the UK, the National Institute for Health and Care Excellence (NICE) evaluates new medical technologies and treatments, offering guidance to the National Health Service (NHS) on their clinical effectiveness, cost-efficiency, and overall impact. These decisions influence which treatments become available to patients and how healthcare resources are allocated.

Each NICE decision involves a detailed review of clinical trials, cost-effectiveness analysis, and broader societal factors. This project uses a large dataset of over 800 existing NICE decisions on new medical technologies and treatments. The project leverages on machine learning and deep learning methods to improve predictive methods and guide medical researchers. We will be collaborating with Johnson & Johnson (JnJ), approaching the HTA decisions with the aim of understanding the decision making processes of HTA using Machine Learning and predicting NICE's HTA decisions with a set of given features. These objectives will enable JnJ to automate the evaluation process of new medical technologies and treatments, saving time and cost on NICE's evaluations.

1.2 Objectives

This project will use Natural Language Processing (NLP) and Machine Learning (ML) to develop tools to classify past HTA decisions and predict future outcomes based on raw text submissions. The project goals include:

1. Identifying key decision factors: Extracting important features and trends that impact NICE evaluations, such as specific clinical effectiveness and cost-effectiveness features.

2. Predicting HTA outcomes: Training predictive models on HTA submissions and estimating the likelihood of approval based on textual patterns.

3. Optimizing HTA submission strategies: Giving healthcare providers advice on improving their submissions to meet NICE evaluation standards and get more approvals.

With this project, we hope to improve the use of evidence in healthcare, decrease uncertainty in drug approvals, and increase efficiency.

2 Data Overview

2.1 Raw Data Description

The raw data currently consist of websites for 810 different types of health technologies and treatments, published by NICE on its website. Each website contains overview of recommendation made by NICE, committee papers, and other supporting documents.

2.2 Final Dataset For Feature Extraction

The committee papers contain all the values of the features we need to extract (more specifically, the information are within the company evidence submissions section of the committee papers). By web scraping, we have downloaded 476 committee papers, each representing a specific health technology. Among all these 476 health technologies, 59 of them got rejected, 119 of them got accepted, and 298 got conditionally accepted by NICE. Most health technologies with no published committee papers are health technologies that NICE is unable to make a recommendation for. The label of our final models, which is named "decision", is a categorical variable having three different categories ("accept", "conditionally accept", and "reject") as we mentioned above. It is obtained via web scraping from the section "Recommendation Overview" on the website. Hence, our final dataset for feature extraction is composed of 476 committee papers in PDF format to some degree of imbalance.

The features we are extracting from the committee papers are a mixture of numerical and text data, which have been identified as key parameters NICE likely use (e.g. clinical effectiveness, cost-efficiency, and other key information) determined through literature review and advice from our domain expert collaborators in JnJ.

3 Methodology

The project's methodology can be broken down into two main steps. The first step aims to extract features that are potentially valued by NICE. Once these features are obtained from the committee papers, we will deploy classical machine learning models to predict decisions NICE would like to make.

3.1 Feature Extraction Using Keyword Search, NLP, and LLM

The features we intend to extract can be classified into three different categories: basic information, clinical-related features, and cost-related features.

The basic information of interest includes the date of application, the date of decision, the type of disease, and the type of health technology. The two types of dates can be directly scraped from the website and the committee papers since they have exact values in fixed locations. For the type of disease, we consider it most effective to obtain it by using LLM. Although this feature has objective values, is not always fully mentioned in the papers. Therefore, in order to guarantee the correctness of values for this feature, LLM can be conveniently applied. For the type of health technology, keyword search should be suitable since a section in the committee paper indicates the type of health technology using words that are only suitable for some specific type of health technology. For example, in that section, many documents use words like "dose of injection," which represents this health technology must be some drug instead of some device.

For clinical effectiveness features from past NICE reports, features such as "paediatric population," "How many RCTs," and "Meta-Analysis performed" can be detected using keyword search in the PDFs. These direct matches provide structured data for model training. Other features like "clinical uncertainties," "Adverse reactions," "quality of evidence," and "treating unmet needs" usually have long sentences as results, and NLP models will be needed to process those long sentences in the

PDFs. We extract relevant sentences and use an AI chatbot (LLM) to analyze and rank them. This will convert the underlying contextual meaning into numerical data, which can be more easily used for model development later.

For cost-related features, many key values, such as QALY (Quality-adjusted life year) and ICER (Incremental Cost-Effectiveness Ratio), are intentionally obscured by black boxes due to them being sensitive information. However, we are able to leverage on LLM to obtain a general understanding of the cost-effectiveness and uncertainty of the test methods. Since these features are mainly in the section "Interpretation and conclusions of economic evidence", we used the entire section to generate a variable cost-effectiveness to evaluate the cost-effectiveness of medicines. We extract the text in the section and input them to models such as OpenAI API, Microsoft Phi-2, and Medalpaca-7B to interpret and analyze the information in the section, before giving a score to medicine's cost-effectiveness and the uncertainty of the cost-effective evaluation methodology proposed by the company. Understandably, every institution is likely to present its cost-effectiveness analysis in a more favorable way to increase the chances of NICE approval. In response to this bias that may be overlooked by the LLM, we are currently assessing various ways to counteract this bias, such as setting a baseline or expanding the scoring range to capture the difference between the cost-effectiveness of the medicines.

Overall, we combine keyword extraction with NLP and AI chatbot ranking to improve the accuracy of extracting features from PDFs. This ensures that all features (and the underlying context) can be effectively captured. The extracted data will train machine learning models to classify past decisions and predict future HTA outcomes.

3.2 NICE Decision Prediction with Machine Learning

Once we extract the features from the dataset, we will use machine learning models to predict NICE's HTA decisions. The goal is to develop a predictive model that can estimate whether a given medicine will receive NICE approval or not based on its submission.

Model Selection:

We intend use logistic regression, K-nearest neighbors (KNN), tree-based models including boosting trees, and deep learning models. We will perform hyperparameter tuning with stratified K-fold cross-validation on logistic regression and K-nearest neighbors to optimize model performance.

Evaluation Metrics:

- F1-score: measure overall predictive performance.
- Precision and Recall: balance false positives and false negatives, as different incorrect NICE decision predictions have different implications.
- AUC-ROC: evaluate the models' ability to distinguish between approved and rejected medicines.
- Intuition: Make use of domain knowledge to choose the best set of decision boundaries.

4 Results

4.1 Feature Extraction

We have experimented with various LLM for text summarization and feature extraction. In this process a large chunk of work has been dedicated to prompt engineering, for instance, by optimizing the instructions to the LLM and including relevant examples so the AI understands how to effectively summarize, analyze and classify the information into a feature.

For basic information, the date of application and the date of decision for each health technology have been obtained via scraping. The type of disease has already derived using Openai API. For the type of health technology, we currently are trying to extract it using keyword search and should be able to get the result within a week.

For cost-effectiveness, the Openai API is the best LLM so far that effectively summarizes the cost-effectiveness and uncertainty of the cost-effectiveness evaluation methodology performed by the company. As expected, the LLM is often unable to detect 'sugar-coating' performed by the

company to make the medicine more cost-effective than it actually is. For example, the company may conduct the test in a specific way such that the reported ICER value is below $30,000 per QALY gained in compliance with NICE's regulations. To account for this bias, we used a skewed range of to measure the cost-effectiveness of the medicine. Since all cost-effective values are inflated due to sugar-coating, while the instructions provided to the LLM is to rate the cost-effectiveness of the from 1 to 10, almost no medicine was given a value of less than 5. Hence, we skew the scale such at values such as 5 or 6 is considered less cost-effective. Moreover, we have also individually investigated cases where low cost-effective scores are given to medicines that are recommended and vice versa. Some of these instances provide important insights into the nuances the LLM cannot capture, and we included that directly as part of the prompt to enhance LLM's ability to evaluate the text.

4.2 NICE Decision Prediction

We are still in the process of extracting all the features we need and ensuring that the use of LLM can sufficiently capture the information and summarize them into a categorical/numerical feature. Since this first step is the hardest part of the project, we have allocated additional time for this part. We will begin second part (NICE Decision Prediction) in 2 weeks.