Large Language Models Enhance the Identification of Emergency Department Visits for Symptomatic Kidney Stones

Cosmin A. Bejan, PhD[1], Amy M Reed, MD[1], Natalie M Pace, MD[1], Siwei Zhang, MS[1], Yaomin Xu, PhD[1], Daniel Fabbri, PhD[1], Ryan S. Hsi, MD[1]
[1]Vanderbilt University Medical Center, Nashville, TN, USA

Introduction

Correct classification of a symptomatic kidney stone event is critically important for understanding disease burden and epidemiology. The use of International Classification of Diseases (ICD) codes derived from electronic health records (EHRs) is the common strategy to identify kidney stones. However, when used in the context of emergency department (ED) visits, this approach shows limitations. This study explores the capabilities of large language models (LLMs) and traditional machine learning methods to enhance symptomatic kidney stone detection in ED settings.

Methods

The study (IRB# 182297) included 500 randomly selected ED reports from the Vanderbilt EHR. Of these, 450 (90%) reports corresponded to ED visits where an ICD code for kidney stone was also assigned. Each report was manually labeled as positive/negative for the description of a symptomatic kidney stone as the primary cause of its ED visit. Stratified sampling was used to split the labeled ED reports into training (80%) and test (20%) sets. Logistic regression, XGBoost, and LightGBM models were trained to classify ED reports as follows. First, data processing consisted of lower-case conversion, lemmatization, and removal of punctuations, numbers, stop words, and tokens of length 1. Next, a TF-IDF weighting scheme and a bag-of-words approach were used to extract features from each ED report. Hyperparameter tuning using 10-fold cross validation with stratified sampling was employed over the training set. Evaluation was performed on the test set using the retrained models with the best performing parameters. LLMs selected for this study included Llama-2-70B, GPT-3.5, and GPT-4. Nine zero-shot prompts were crafted in total, each with varying degrees of specification. For classifying an ED report, each prompt consisted of specific classification instructions (eg, “*The only answer choices are 'Yes' or 'No'*”), additional phenotypic information (eg, “*the stone is in the kidney or ureter*”), and the entire content of the report. All the prompts used in this study are available at https://github.com/bejanlab/LLMs4KS-ED.git. The same 10-fold cross validation with stratified sampling approach was used for prompt optimization on the training set and the best performing prompts were evaluated on the test set. To generate predictions with the highest degree of confidence, each LLM was run with the temperature parameter set to 0. Due to the stochastic nature of LLMs, each experiment was repeated 10 times and their results were macro-averaged. All the models were evaluated with standard metrics: precision, recall, specificity, and F1.

Results

The manual review resulted in identifying 260 (52%) ED visits due to kidney stones and the remaining 240 (48%) visits due to other reasons. Table 1 lists the evaluation results on the test set by the LLMs, traditional machine learning models, and a baseline system using ICD codes for kidney stones. While all the implemented models yielded better F1 scores than the ICD-based baseline, the GPT models obtained the best F1 results (0.80 by GPT-3.5 and 0.84 by GPT-4). Of note, due to its limitation on prompts exceeding 4096 tokens, Llama-2 achieved a response rate of 91%.

Discussion and Conclusions

This study demonstrates the potential of large language models to analyze clinical text to improve the identification of healthcare encounters for symptomatic kidney stone disease in the acute care setting. Future directions include exploration of additional LLM prompt engineering strategies, including few-shot prompts and prompt augmentation methods to allow the integration of patient specific data such as demographic information.

Table 1 Evaluation of methods for identifying kidney stone disease in emergency department reports.

Model		**P**	**R**	**S**	**F1**	**AR**
ICD codes		0.560	0.981	0.167	0.713	1
Logistic regression		0.809	0.731	0.813	0.768	1
XGBoost		0.818	0.692	0.833	0.750	1
LightGBM		0.844	0.731	0.854	0.784	1
LLM	**Prompt**	**Macro P (95% CI)**	**Macro R (95% CI)**	**Macro S (95% CI)**	**Macro F1 (95% CI)**	**AR**
Llama-2-70B	P9	0.616 (0.616-0.616)	0.918 (0.918-0.918)	0.333 (0.333-0.333)	0.738 (0.738-0.738)	0.91
GPT-3.5	P7	0.867 (0.867-0.867)	0.750 (0.750-0.750)	0.875 (0.875-0.875)	0.804 (0.804-0.804)	1
GPT-4	P2	0.925 (0.917-0.933)	0.777 (0.769-0.785)	0.931 (0.923-0.940)	0.844 (0.840-0.849)	1

P precision, R recall, S specificity, F1 F1-measure, AR answer rate, CI confidence interval.