Predicting Patient Readmission Causes using Large Language Models

Amogh Patankar
University of California, San Diego

1 Abstract

Throughout the history of medicine, especially in more recent times, hospital readmissions have consistently been an unresolved, prevalent issue in healthcare systems for a plethora of reasons. While many prior studies use advanced machine learning techniques to accurately predict hospital readmissions, there exists scarce literature on using natural language processing (NLP) techniques to identify causes of readmission. Using data from the MIMIC-III database, this project uses NLP techniques, namely large language models and transformers, to accurately identify the causes for readmission for patients. We use text summarization models such as BART-Large-CNN to generate concise report summaries from MIMIC-III raw reports. We then used Llama-2-7b, OpenAI's GPT-3.5-turbo and the Retrieval Augmented generation to generate the causes of readmission and found the OpenAI's GPT-3.5-turbo to be most suitable and cost efficient for our work.

2 Introduction and Prior Work

Hospital readmissions, for dozens of years, are a pressing issue in healthcare systems as well as patient lives for a vast variety of reasons. Reasons such as financial impact, patient experience, and resource allocation are the most prevalent metrics that provide uncertainty surrounding the readmission of patients. Efforts to address readmissions require more communication and resources allocated towards finding causes of readmissions. Readmissions result in strain with respect to resources and finances when it comes to institutions and decreases in the quality of life for the patients themselves. As such, identifying possible causes of readmission is difficult, and causes for readmission may include factors such as patient case complexity and social factors. While identifying these causes is complex, it presents an opportunity to accurately and rapidly unearth specific factors and improve the readmission process for both institutions and patients. Using reports from patients' first hospital visit/stay from the MIMIC-III database, and models like Llama-2-7b, GPT-3.5-turbo and Retrieval Augmented Generation (RAG), we demonstrate significant causes for readmission. Overall, our study attempts to identify causes for readmissions for patients, while also contrasting the efficacies of different NLP models.

Related Work Hospital readmission within 90-day of an index hospitalization may result from actions or inactions taken during the initial hospital stay [Friedman and Basu(2004)]. Since Centers for Medicare & Medicaid began publishing readmission data in 2009, this metric quickly became viewed as an indicator of health care quality and cost provided by hospitals [Jamei et al.(2017)Jamei, Nisnevich, Wetchler, Sudat, and Liu] [Low et al.(2015)Low, Lee, Hock Ong, Wang, Tan, Thumboo, Liu et al.]. Accordingly, reducing unnecessary readmissions became a key concern for healthcare providers and payers [Low et al.(2015)Low, Lee, Hock Ong, Wang, Tan, Thumboo, Liu et al.] [Shams et al.(2015)Shams, Ajorlou, and Yang]. The majority of a hospital's total expenditures are attributed to inpatient curative costs. According to the 2019 OECD survey, around 65% of a hospital's total expenditures in the 36 OECD countries are spent on inpatient care services [Indicators and Hagvísar(2019)]. The term "readmission" is defined variably in the relevant literature, ranging from 1 [Cardiff et al.(1995)Cardiff, Anderson, and Sheps], 2

[Anderson and Steinberg(1985)], 3 [Tabak et al.(2017)Tabak, Sun, Nunez, Gupta, and Johannes] to 12 months [Kelly et al.(1992)Kelly, McDowell, Crawford, and Stout] after the initial admission. In the U.S., nearly 20% of Medicare-covered patients are readmitted within 30 days of their initial discharge, incurring an estimated annual cost of $17 billion [Jencks et al.(2009)Jencks, Williams, and Coleman].

The rate of readmissions serves as a quality measure for healthcare services and has been extensively studied for decades [Ashton et al.(1997)Ashton, Del Junco, Souchek, Wray, and Mansyur]. Strategies such as identifying high-risk patients for readmissions, managing medication reconciliation, optimizing technology use, and implementing a transitional care model are among the approaches that can be employed to reduce preventable readmissions. Machine learning and statistical approaches have been employed in the past mainly to predict the readmission, but most don't focus on the causes of readmission.

In a recent study, [Zhou et al.(2022)Zhou, Li, Wang, Chai, and Zhang] focused on elderly hospital readmission prediction, proposing a fuzzy partition enhanced weighted factorization machine (WFM) model. In a meta-research, [Mahmoudi et al.(2020)Mahmoudi, Kamdar, Kim, Gonzales, Singh, and Waljee] collected 41 studies during the period from 2015 to 2019, focusing on the comparison of traditional and ML forecasting models. They found that on average, ML models outperformed statistical ones. In our project we use the advanced deep learning based solutions and leverage the power of large language models for finding the causes of readmissions for a 30 day period.

3 Methods

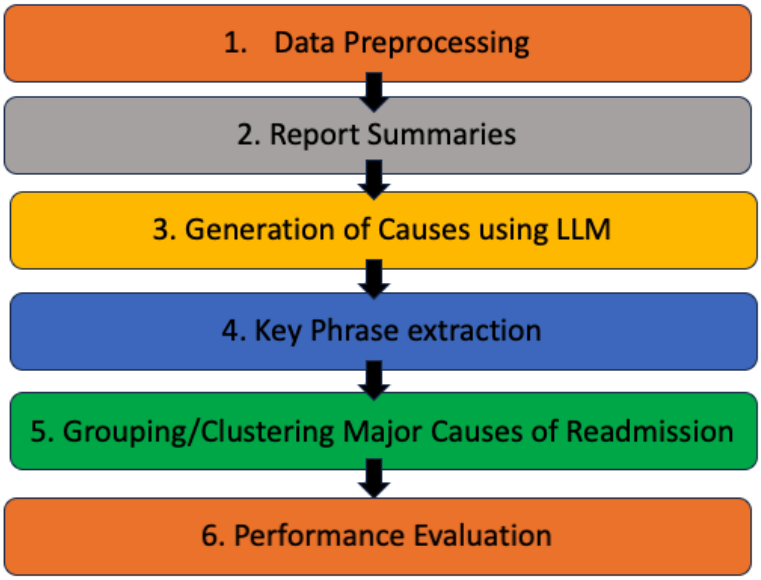


Figure 1: Flowchart for our method

The first step in our methodology is data preprocessing, involving cleaning and transforming the dataset to generate data for readmitted patients, which includes their demographic information, discharge summary, and report summary from the next visit to the hospital. The next step is to create concise summary reports from the raw reports using a text summarization model, specifically bart-large-cnn. Following that, we generate causes of readmission using large language models such as LLM-7B and OpenAI.

Subsequently, we extract key phrases from the generated causes to gain insights into the distribution of diseases across gender, ethnicity, etc., and perform cluster analysis. We then manually evaluate the performance of our generated causes by randomly sampling approximately 20 generated causes

and reviewing them for coherence with the raw summary reports, with the assistance of a medical practitioner.

3.1 Dataset

The Medical Information Mart for Intensive Care III (MIMIC-III) v1.4 Database is a large, publicly available, de-identified database, available via PhysioNet. Data is recorded from the Beth Israel Deaconess Medical Center in Boston, Massachusetts, across a 11 year span from 2001-2012. Amongst other forms of data, it includes clinical data, physiological data, and various reports. Specifically, we looked at patient demographic information and reports from their visit to the hospital. Looking through the data, many patients have multiple different hospital visits (three or more). Given our approach, we focused on patients who had discharge summaries and the earliest report summary from the next visit, qualifying as a 30-day readmission. Discharge summaries provide a holistic view of patient's journey during their stay in the hospital. It contains information on on-going care, follow-ups, previous medical history, reason for examination, discharge date, final report etc. The dataset, limited by this idea, contains data for over 11K patients. For identifying readmission causes, we are restricting our analysis to 425 patients.

Figure 3 provides insights into the most common values for the patients' demographic information from the dataset. It's evident that most of the patients fall into the white male category and undergo a radiology imaging procedure in their initial visit during readmission.

Entity	Most common value	% Count
INSURANCE	Medicare	42.6
RELIGION	CATHOLIC	32.5
MARITAL_STATUS	MARRIED	49.6
ETHNICITY	WHITE	65.4
GENDER	Male	57.6
Report Category	Radiology	77.2

Figure 2: Most common values

Data Preprocessing

For data preprocessing, we initially combine patients' demographic data and discharge/report summaries from the from the admission.csv file and noteevents.csv file (respectively) in the raw MIMIC III dataset. Subsequently, we clean the dataset by removing entries with missing dates and NaN values. We then filter patients who have a discharge summary and the earliest report summary from the index admission, qualifying as a 30-day readmission. Our next step involves condensing the dataset by retaining only patients' demographic information, discharge summary from the previous visit, and report summary from the index admission for our analysis.

3.2 Report Summaries

After preprocessing the dataset, We generate summaries for each of the patient's notes , the main idea being to maintain the data confidentiality and make the input reasonable for the model. Goals being as follows -

1. **Maintaining data confidentiality:** Since we downloaded the dataset from Physionet, and we cannot directly upload that to public API (due to our Data Use Agreement), we summarize the data in a concise and meaningful format using other means.
2. **Reducing Input Length:** Large Language Models often have limitations on the maximum input length they can handle. Summarizing the text that is passed as input helps to condense the content, ensuring that it fits within the model's input constraints.
3. **Enhancing Relevance:** Summarization helps in identifying and preserving the most relevant information from a document. Removing unnecessary info (in this case, patient name, doctor's name, other identifiers) was quite helpful.

As for the models, we initially tried spacy, but it was highly ineffective; we then used different Facebook Bart models which are transformer based models. Finally, after testing the various models, we used the bart-large-cnn for our implementation.

3.3 Cause(s) Generation

For the purpose of our analysis, as already mentioned, we generate causes for a pair of hospital encounters for a patient. The first of these, is the discharge summary and the other report containing the notes from the index admission within 30 days for a total of 425 patients.

Llama-2-7b: Llama 2 is an auto-regressive language model that uses an optimized transformer architecture. The tuned versions use supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align to human preferences for helpfulness and safety. Here we used a context length of 2048 for the purpose of our assignment. Since we summarized the documents we were able to fit them within the context.

GPT-3.5-turbo: Generative Pre Trained (GPT) models can understand and generate natural language or code. The most capable and cost effective model in the OpenAI GPT family that we experimented with was gpt-3.5-turbo which has been optimized for chat using the Chat Completions API but works well for traditional completions tasks as well. Hence we used the GPT-3.5-turbo for our experiments. It has a token length of 4096; it was approximately $0.0010/1K tokens for input, and $0.002/1K tokens for the output.

Retrieval Augmented Generation (RAG) Retrieval Augmented Generation (RAG) represents a strategic approach to address the limitations of standalone pre-trained models. It introduces a hybrid architecture that combines pre-trained parametric and non-parametric memories, enhancing language generation efficiency. By integrating a pre-trained seq2seq model with a dense vector index from a structured knowledge base like the Medical Question and Answering Dataset (MedQuAD)[Ben Abacha and Demner-Fushman(2019)], RAGs ensure the generation of contextually rich and factually corroborated responses. This integration elevates the specificity and reliability of generated outputs, particularly in knowledge-dense applications within Natural Language Processing (NLP). The patients' first and second reports are utilized as prompts, while MedQuAD serves as the vector store for RAG. This tailored approach enhances the generation of responses with a heightened degree of relevance and accuracy, contributing to the effectiveness of the AI-driven medical query responses.

3.4 Key Phrase Generation

BERT is a bidirectional transformer model designed to convert phrases and documents into vectors that effectively capture their semantic meaning. An KeyBERT is a Python package available as open-source, that leverages BERT to simplify the process of extracting keywords for a given piece of text.

Steps in Key Phrase Extraction:

1. **Candidate Keywords/Key phrases**: For this step, we used Scikit-Learn CountVectorizer. This allows us to specify the length of the keywords and convert them into key phrases. This step includes generating n-grams and removing stopwords. After analyzing the results, specific to our dataset, we remove additional words such as:

   [”patient”, ”cause”, ”readmission”, ”patients”, ”causes”, ”readmissions”, ”complications”, ”complication”, ”information”, ”provided”].

   This step removes the redundant words that come up frequently in the potential causes for readmission.
2. **Embedding:** The original causes for readmission and the candidate key phrases into word vectors. BERT is used for this task due to its notable performance in similarity and paraphrasing tasks. KeyBERT uses the SentenceTransformers package under the hood to create encoding for every candidate phrase.
3. **Cosine Similarity**: Finally, we find the candidates that are most similar to the readmission causes. To measure the similarity between candidates and the causes, we'll utilize cosine similarity between vectors, known for its robust performance in high-dimensional spaces.

Other libraries for keyphrase extraction:

RAKE (Rapid Automatic Keyword Extraction) is a keyword extraction algorithm that focuses on identifying keywords or key phrases from text documents. However, similar to some other traditional keyword extraction methods, RAKE is primarily based on statistical patterns and co-occurrence of words rather than deep contextual understanding.

spaCy's part-of-speech tagging and noun phrase extraction can be used to identify potential keywords or key phrases. However, these methods are primarily based on syntactic structures rather than deep contextual understanding.

Hence, given the use case, KeyBERT was the most suitable approach.

3.5 Grouping/Clustering Major Causes of Readmission

Once we generate the causes using the OpenAI's GPT-3.5-turbo and extract keyphrases from each causes, we do various demographical analysis to gain further insights. We hope these insights discussed in later sections can be a useful tool for framing global health policies and outcomes.

4 Results

4.1 Sample cause generation and comparison

For the cause generation, we first compare which method (Llama-2-7b, OpenAI GPT-3.5 or RAG model) is most suitable for our task. Starting with Llama-2-7b with a context length of 2048, we ran it locally, we observed significant inefficiency, taking approximately 300 seconds for each request. In contrast, OpenAI's GPT-3.5-turbo demonstrated a remarkable improvement during the human evaluation, reducing response times to around 20 seconds for each request. We also explored the efficiency of the Retrieval-Augmented Generation (RAG) model, which showed promise. However, the response time was approximately 200 seconds for each request. Used the MedQuAD as the vector store. We found the OpenAI's GPT-3.5-turbo responses to be most aligned to actual causes found during the manual evaluation discussed in the next section.

Figure 3 illustrates the sample results we obtained for a patient.

4.2 Grouping/Clustering Major Causes of Readmission

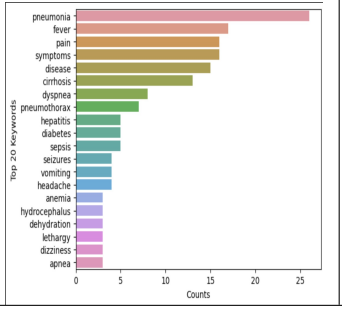
Overall plot	Overall word cloud
	

Table 1 : Overall causes insights

- text1 = """The patient is an 87-year-old Russian speaking female. She was admitted to the hospital on to the Vascular Service with a nonhealing ulcer on her fourth right toe x3 months. She underwent a bypass by Vascular Surgery. Postoperatively in the PACU, the patient developed some respiratory distress and had a PEA arrest with very difficult intubation."""
- text2 = """An 87-year-old woman with pelvic pain was seen by a radiologist. The patient had multiple chronic medical issues. She had a large hiatal hernia, a small pleural effusion, and a large stomach tube."""
- prompt = "You're a medical expert; here are summaries for one patient: {}\n{}\n Can you determine the cause of readmission. Be very concise, try to be as technical as possible, use one sentence:".format(r1, r2)

Llama-2-7B	OpenAI gpt-3.5-turbo	Retrieval Augmented Generation (RAG)
After reading these notes, I believe that the cause of readmission for this patient is related to the complications from the femoral bypass surgery. Specifically, the patient experienced respiratory distress and difficulty Intubation after the procedure, which may have contributed to her readmission to the hospital.	The cause of readmission was likely respiratory distress and a PEA arrest following a difficult intubation post fem-bypass surgery.	Based on the provided medical reports, the cause of readmission for the patient is: Respiratory distress after bypass surgery in the postoperative period. The patient was readmitted to the hospital due to respiratory distress developed after the surgery, which suggests that there may have been some complications during or after the procedure.

Figure 3: Results for a sample patient

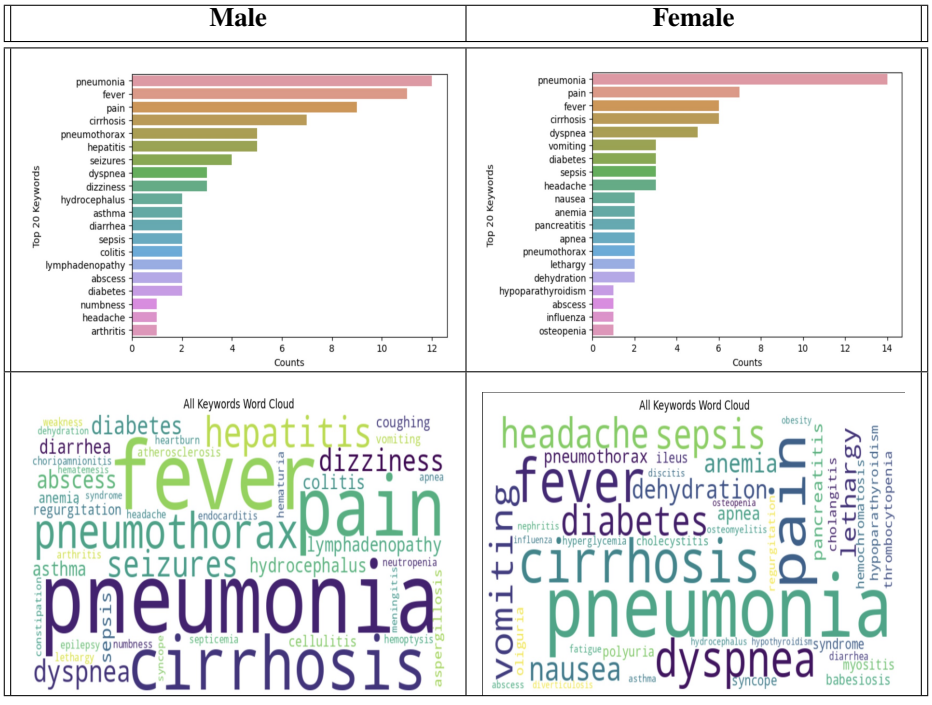


Table 2 : Gender comparison for causes

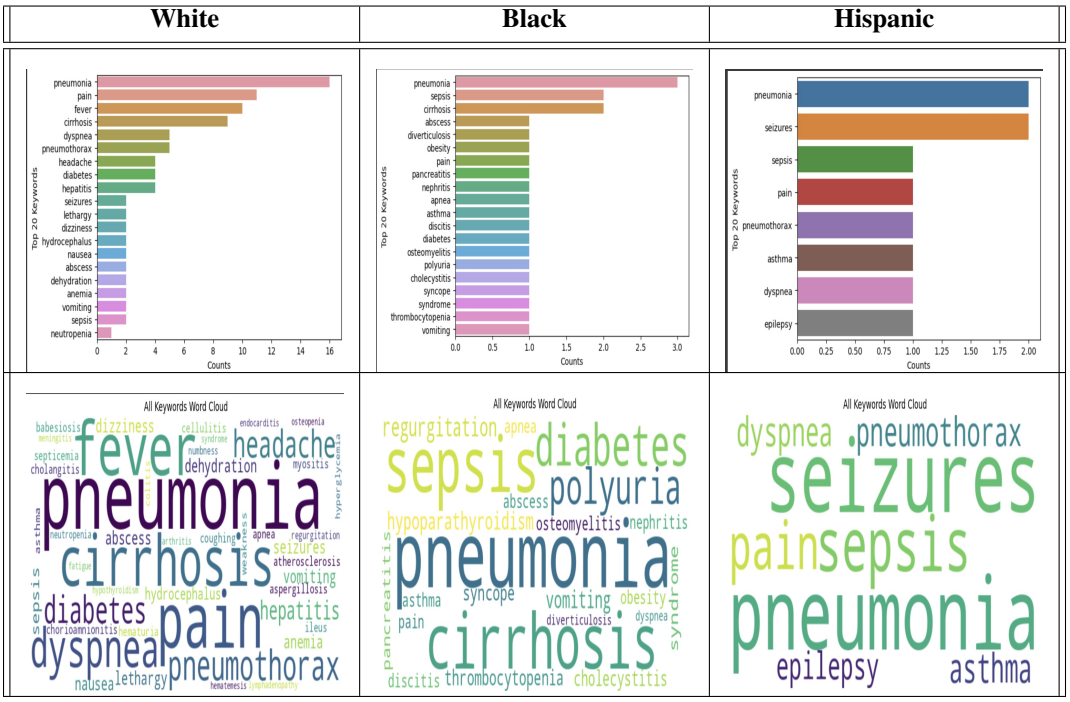


Table 3: Ethnicity comparison for causes

5 Discussion

Previous works have only focused on predicting whether a patient would be admitted or not by employing various ML techniques. Our work employs novel deep learning-based LLM techniques to identify the tangible causes behind patient readmissions. We also generate intuitive insights by summarizing our findings through various visualizations.

We wanted to explore the advanced capabilities of the Llama-2–7b model with a Retrieval Augmented Generation (RAG) approach and compare it against OpenAI's GPT-3.5-turbo for readmission cause generation. To achieve this we followed the manual evaluation approach following the insights from our advisor, Shannon Cotton who helped us evaluate a sample cause pair. We first generate the causes in a human-friendly concise format. Once we have descriptive causes we extract key phrases from the causes using KeyBERT.

Manual Evaluation: We generated causes using all three methods (Llama-2-7b, GPT-3.5, and RAG model) for a sample of 20 patients to individually evaluate the efficiency of the three methods. We gave a ranking for each cause from 1-3 for the three methods and combined our findings. We found that OpenAI's GPT-3.5 responses were most relevant.

Insights from demographic analysis on readmission causes: Here we filtered the top keywords from causes to only include the UMLS concepts that correspond to Disease or syndrome or signs and symptoms category. In the results obtained in Table 1, we see that overall the top 3 readmission reasons have come up as, "Pneumonia", "Fever", and "Pain".

To gain more insights we distributed the results based on gender in Table 2.
For females, the top 5 causes are: "Pneumonia", "Fever", "Pain", "Cirrhosis", "Dyspnea".
For males: "Pneumonia", "Fever", "Pain", "Cirrhosis", "Pneumothorax"

A second demographic-based analysis was conducted on Ethnicity as shown in table 3. The top 3 ethnicities in the dataset were White, African American/Black, and Hispanic or Latino.
For White ethnicity, the top 5 causes are: "Pneumonia", "Fever", "Pain", "Cirrhosis", "Dyspnea".
For African American/Black ethnicity, the top 5 causes are: "Pneumonia", "Sepsis", "Abscess", "Diverticulosis".

For Hispanic or Latino ethnicity, the top 5 causes are: "Pneumonia", "Seizures", "Sepsis", "Pain", and "Pneumothorax".

Limitations and comments from the presentation: Our work focuses on patients who have a discharge summary, but this can be extended to limit it solely to ICU-related discharge summaries. Additionally, we can consider patients who were transferred from the ICU to LL wards during their readmission. This way, we can concentrate more on ICU-related readmission causes rather than general readmission causes. We have also incorporated feedback to filter out non-disease-related keywords while doing cluster analysis using UMLS. We have removed the mention of concepts and only focussed on diseases, signs, and symptoms.

6 Conclusion and Future Work

In summary, this project establishes a foundation for accurately predicting causes of readmission, employing various transformer models and large language models for text summarization and generation tasks. Our findings, based on the MIMIC-III reports dataset, demonstrate the feasibility of generating meaningful causes for patient readmission through the application of NLP techniques, particularly leveraging large language models. Moving forward, our endeavors will extend to ICU patients, exploring alternative model architectures and types. We also hope to compare our findings from the social media platform which contain invaluable sources of first hand experiences from people suffering with these diseases.

The assessment of model performance involved a comprehensive clinical validation by advisors combined with manual evaluation on a sample of 20 randomly selected patients by our group. A crucial aspect underscored during discussions and presentations was the practicality of deploying these techniques in real-world scenarios. Recognizing the significance of factors such as cost and latency, we concluded that although RAG models and larger counterparts might exhibit superior performance, their high latency and cost render them impractical for real-world use. Thus, the decision to adopt OpenAI's GPT-3.5-turbo aligns with our prioritization of practicality.

In closing, our study validates the rationale behind integrating machine learning techniques in the medical field, offering a reasonable threshold. We express sincere gratitude to our professors, Dr. Mike Hogarth and Dr. Shamim Nemati, as well as our TA, Aaron Boussina, for their invaluable support. Special thanks to Shannon Cotton for her insightful advice.

References

[Anderson and Steinberg(1985)] Gerard F Anderson and Earl P Steinberg. 1985. Predicting hospital readmissions in the medicare population. *Inquiry*, pages 251–258.

[Ashton et al.(1997)Ashton, Del Junco, Souchek, Wray, and Mansyur] Carol M Ashton, Deborah J Del Junco, Julianne Souchek, Nelda P Wray, and Carol L Mansyur. 1997. The association between the quality of inpatient care and early readmission: a meta-analysis of the evidence. *Medical care*, pages 1044–1059.

[Ben Abacha and Demner-Fushman(2019)] Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC bioinformatics*, 20(1):1–23.

[Cardiff et al.(1995)Cardiff, Anderson, and Sheps] Karen Cardiff, Geoffrey Anderson, and Samuel Sheps. 1995. Evaluation of a hospital-based utilization management program. In *Healthcare management forum*, volume 8, pages 38–45. Elsevier.

[Friedman and Basu(2004)] Bernard Friedman and Jayasree Basu. 2004. The rate and cost of hospital readmissions for preventable conditions. *Medical Care Research and Review*, 61(2):225–240.

[Indicators and Hagvísar(2019)] OECD Indicators and OECD Hagvísar. 2019. *Health at a glance 2019: OECD indicators*. Paris: OECD Publishing.

[Jamei et al.(2017)Jamei, Nisnevich, Wetchler, Sudat, and Liu] Mehdi Jamei, Aleksandr Nisnevich, Everett Wetchler, Sylvia Sudat, and Eric Liu. 2017. Predicting all-cause risk of 30-day hospital readmission using artificial neural networks. *PloS one*, 12(7):e0181173.

[Jencks et al.(2009)Jencks, Williams, and Coleman] Stephen F Jencks, Mark V Williams, and Eric A Coleman. 2009. Rehospitalizations among patients in the medicare fee-for-service program. *New England Journal of Medicine*, 360(14):1418–1428.

[Kelly et al.(1992)Kelly, McDowell, Crawford, and Stout] James Francis Kelly, H McDowell, Vivienne Crawford, and RW Stout. 1992. Readmissions to a geriatric medical unit: is prevention possible? *Aging Clinical and Experimental Research*, 4:61–67.

[Low et al.(2015)Low, Lee, Hock Ong, Wang, Tan, Thumboo, Liu et al.] Lian Leng Low, Kheng Hock Lee, Marcus Eng Hock Ong, Sijia Wang, Shu Yun Tan, Julian Thumboo, Nan Liu, et al. 2015. Predicting 30-day readmissions: performance of the lace index compared with a regression model among general medicine patients in singapore. *BioMed research international*, 2015.

[Mahmoudi et al.(2020)Mahmoudi, Kamdar, Kim, Gonzales, Singh, and Waljee] Elham Mahmoudi, Neil Kamdar, Noa Kim, Gabriella Gonzales, Karandeep Singh, and Akbar K Waljee. 2020. Use of electronic medical records in development and validation of risk prediction models of hospital readmission: systematic review. *bmj*, 369.

[Shams et al.(2015)Shams, Ajorlou, and Yang] Issac Shams, Saeede Ajorlou, and Kai Yang. 2015. A predictive analytics approach to reducing 30-day avoidable readmissions among patients with heart failure, acute myocardial infarction, pneumonia, or copd. *Health care management science*, 18:19–34.

[Tabak et al.(2017)Tabak, Sun, Nunez, Gupta, and Johannes] Ying P Tabak, Xiaowu Sun, Carlos M Nunez, Vikas Gupta, and Richard S Johannes. 2017. Predicting readmission at early hospitalization using electronic clinical data: an early readmission risk score. *Medical care*, 55(3):267.

[Zhou et al.(2022)Zhou, Li, Wang, Chai, and Zhang] Jiandong Zhou, Xiang Li, Xin Wang, Yunpeng Chai, and Qingpeng Zhang. 2022. Locally weighted factorization machine with fuzzy partition for elderly readmission prediction. *Knowledge-Based Systems*, 242:108326.