AI and Data Management in the Medical Field

Damien Beltran

Started 8/29/25

Finished 12/12/25

Dr. Yang

CS-4913-003

Introduction

Electronic Health Records (EHRs) provide a rich but challenging source of clinical information. Artificial Intelligence (AI) , particularly large language models (LLMs) has rapidly advanced the ability to summarize, interpret, and predict patient outcomes from EHR data. However, AI systems can amplify existing demographic biases or introduce new ones if not paired with external knowledge sources or interpretability tools.

This research focuses on the following central question:

How can Retrieval-Augmented Generation (RAG) be integrated into machine learning workflows using EHR data to mitigate bias, and how do text-based versus graphical output representations affect bias visibility and interpretability?

To answer this question, this project:

1. Builds upon and extends the RAM-EHR research framework.
2. Implements a custom QA Agent Pipeline using demographic-filtered patient cohorts.
3. Integrates local RAG knowledge retrieval to reduce hallucinations and improve grounding.
4. Designs graphical analyses to visualize bias across demographic subgroups.

RAG + Bias Connection in EHRs

RAM-EHR

https://arxiv.org/abs/2403.00815

Direct relevance

1. **RAM-EHR**
    a. Focuses on, on a pipeline to improve clinical predictions on Electronic Health Records (EHRs). For this, the pipeline collects knowledge sources, converts data into text format, and uses dense retrieval to obtain information related to medical concepts, *(Xu et al., 2024)*.

    b. Encompasses detailed information about patients such as symptoms, diagnosis, and medication, are widely used by physicians to be delivered, *(Jensen et al., 2012; Cowie et al., 2017; Shi et al., 2024)*.

2. **Data Driven (graphical context)**
    a. RAM-EHR over-passed knowledge baselines between datasets (3.4% gain in AUROC and 7.2% gain AUPAR), emphasizes the effectiveness of the summarized knowledge from RAM-EHR for clinical prediction tasks, *(Xu et al., 2024)*.
3. **Improvement Focuses**
    a. For further improvement, RAM-EHR is set to focus on predictive performance. There have been several works to attempt augmenting EHR visits with external knowledge, *(van Aken et al., Naik et al. 2022)*.
        i. Examples for this execution include incorporating additional clinical notes, but can be closely viewed as noisy and unrelated to predictions, *(van Aken et al., Naik et al. 2022)*.
        ii. Another example is informative knowledge summaries relevant to downstream tasks for each medical code, *(van Aken et al., Naik et al. 2022)*.
            1. Possible outcome is enhanced processes in relevancy and utility of the retrieved knowledge for clinical tasks, *(van Aken et al., Naik et al. 2022)*.

Applying RAG to EHR data

https://www.sciencedirect.com/science/article/pii/S1532046424000807

Direct relevance

Known Issues

- Malnutrition is a prevalent issue in aged care facilities that lead to adverse health outcomes, (*Alkhalaf et al., 2024a)*.
    - A practiced solution is the ability to extract key clinical information from a large volume of data in EHRs that can improve understanding of specific problems and developing interventions,(*Alkhalaf et al., 2024a)*.

Method of Approach:

- This specific research uses zero-shot prompt engineering that's applied to AI models with a combination of RAG, for the automating tasks of summarizing both structured and unstructured data in EHR and extracting important malnutrition data, (*Alkhalaf et al., 2024a)*.

Evidence / Evaluation:

- Summarization accuracy: 93.25% (zero-shot) -> 99.25% (RAG), (*Alkhalaf et al., 2024a).*

- Risk factor extraction: ~90% accuracy, but RAG did not improve further, (*Alkhalaf et al., 2024a)*.
- Shows task-specific limitations of RAG in bias mitigation.

Connections Particularly in Bias Mitigation

Zero-shot prompting with LLMs:

- Demonstrates effectiveness of general models with minimal task-specific tuning; but prone to variability and bias.

Hallucinations as Bias:

- Identifies hallucination explicitly as a bias category.
- Provides concrete examples of systemic bias in clinical text processing (e.g., fabricated demographics).

RAG:

- Used as a bias control mechanism, grounding outputs in evidence.
- Improves factual accuracy for summarization, but is less effective for implicit risk extraction.

Leveraging LLMs and evaluating their outputs using EHR data

https://www.proquest.com/openview/5059413b6f9e046d74398aa4706e4026/1?pq-origsite=gscholar&cbl=18750&diss=y

Direct relevance

Bias and Reliability in Clinical Unstructured Data

- Using RAG, the dissertation aims to explore advanced techniques to improve factual accuracy and reliability of clinical information extraction from unstructured texts (*Bhattarai, 2024*).

Challenges

- The use of ambiguous abbreviations, the presence of domain-specific jargon, and a limited amount of clinical data during pre-training processes for AI models to extract accurate data (*Bhattarai, 2024*).
- Clinical texts often contain complex relationships between entities that are difficult for models to capture without domain-specific knowledge (*Bhattarai, 2024*).

3 Aim Focuses

- Evaluating bias with clinical information extraction involves exploring the effectiveness of that purpose, optimizing outputs, and empirically evaluating their performance (*Bhattarai, 2024*).

- This has a framework for addressing key challenges in clinical data extraction (*Bhattarai, 2024*).
    - Aims related to this research:
        - Aim 1: Extracting clinical phenotypes from EHR data
        - Aim 2: UMLS integration in GPT models
        - Aim 3: Impact of model parameters and complexity of patient texts on feature extraction results

Bias in Patient Data

Bias in patient demographic data arises because EHR systems collect, represent, and document demographic groups unequally. EHRs contain incomplete or poor-quality demographic fields, inconsistent measurement of social demographic fields, inconsistent measurement of social determinants of health, and significant underrepresentation of medically underreserved or structurally marginalized populations. Structural inequities influence who is represented in EHR datasets, while provider documentation practices, institutional policies, and implicit biases introduce further measurement and labeling differences, (*Siber-Sanderowitz et al., 2022*).

As a result, demographic groups such as males vs. female, racial minorities, or low-literacy patients are effectively defined differently through patterns of missingness, documentation quality, and unequal representation within the data, which can ultimately limit generalizability and contribute to health inequities, (*Siber-Sanderowitz et al., 2022*).

Codebase Relevant to this Research

RAM-EHR is improving clinical predictions on EHRs

https://github.com/ritaranx/RAM-EHR?tab=readme-ov-file

- Retrieval augmentation pipeline for improving clinical predictions on EHR data. Focuses on a collection of multiple knowledge sources, converts them into text format, and uses dense retrieval to obtain information related to medical concepts.

Understanding RAM-EHR

Using RAM-EHR

Focuses on, on a pipeline to improve clinical predictions on Electronic Health Records (EHRs). For this, the pipeline collects knowledge sources, converts data into text format, and uses dense retrieval to obtain information related to medical concepts.

General Questions

What knowledge sources does RAM-EHR use, and how are they structured?
RAM-EHR leverages multiple structured and textual knowledge sources:

- **Structured biomedical knowledge bases**: UMLS (Unified Medical Language System)
- **Textual definitions**: Disease definitions from multiple sources, including:
    - MeSH (Medical Subject Headings) definitions (mesh_def)
    - Medical textbook definitions (medical_text_def)
    - Wikipedia medical definitions (wiki_def)
    -

How are these knowledge sources integrated?

Information on diseases, procedures, and prescriptions.

These sources are integrated through the mimic_disease_id_name_gpt-summary file, where each disease concept includes definitions from all three textual sources. This multi-sourced approach provides:

1. Contextual definitions for medical concepts
2. Input for prompt construction with Azure OpenAI's GPT models
3. Knowledge-grounded summarization capabilities for downstream EHR phenotyping tasks

"model_1_summary": "The disease_id 'Malignant neoplasm of liver and intrahepatic bile ducts' refers to a type of cancer that affects the liver and bile ducts, which is important information for health phenotyping tasks.",

Before execution

- The command "--data mimic --domain disease-id":
    - Loads disease or concept names from the dataset (like the mimic_disease_id_name_merge.json) that is asked by gpt-3.5-turbo for gathering a summary of data.

- Retrieves textual definitions or ontology-based descriptions for each disease.
- When using the GPT model, it summarizes or synthesizes knowledge for that disease.

Testing RAM-EHR with a Model to Model Comparison

Purpose

Understand how different AI models summarize medical diagnosis with definitions from a knowledge source. Analyze how precise and accurate the definitions are presented to the user.

- Models used for this comparison are GPT-3.5-turbo and Grok-3
- The types of data being pulled from the generated JSON files are
  - Keyword presence like (cancer, infection, heart, kidney, blood, lung, sepsis, and hypertension).
  - Words overlap similarity in regards to predictions med generating medical summaries using the knowledge sources.
  - Jaccard similarity when the models output overlapping summaries explaining health conditions and diseases.
  - Output length between the models.

A create_merge_file.py file is created to load training data, read as a simple gpt structure with the name and disease to lastly then be saved as a merge JSON file.

- From the command line, a ″Created mimic_disease_id_name_merge.json″ statement is created with a loadable file to read the output of summaries.

When getting the produced data from the analysis a “mimic_disease_id_bias_analysis.json” is created from the command line.

What was the Outcome of the Textual output?

Variation in Output Detail

- Model_1 tends to produce **concise, straightforward summaries** (shorter lengths, fewer keywords).
- Model_2 often provides **longer, more detailed explanations**, sometimes adding risk factors, clinical context, or synonyms.
  - E.g.

- Malignant neoplasm of liver:
    - gpt-35-turbo: “is a type of cancer that originates in the liver of bile ducts” (24 words)
    - grok-3: Adds risk factors like hepatitis, cirrhosis, toxins (37 words)
    -

Implications: Models differ in **information richness**, which may influence how biased or incomplete outputs appear in downstream phenotyping or clinical interpretations.

Similarity Metrics Highlight Differences

- Word overlap and Jaccard similarity are often moderate to low:
    - Highest word overlap: 0.6 (*Malignant neoplasm of colon*)
    - Lowest word overlap: 0.19 (*Septicaemia*)
- This shows that even when both models capture the same core concepts (e.g., “cancer”, or “heart”), **the framing and additional context differ**, which could propagate subtle bias in automated analyses.

Keyword Analysis

- Core medical terms like *“cancer”, “heart”, “blood”, “infection”* are mostly consistent across models.
- Model 2 sometimes adds extra terms or context (e.g., “infection” in endocardium diseases)

Takeaways

- Differences in **keyword extraction and emphasis** may affect phenotyping pipelines, especially if models highlight certain risk factors disproportionately.

Length Differences

- grok-3 outputs are often **longer**, potentially adding nuances but also **introducing bias risk** emphasis on certain conditions or risk factors.
- Length disparity is especially notable in non-specific disease categories (e.g., “Other diseases of endocardium”).

Bias Related Observations

- grok-3 sometimes **introduces extra casual context or epidemiological factors**. While medically valid, these might **implicitly highlight or downplay certain populations or risk factors**, which is relevant for this bias mitigation research.

- Low similarity scores in some conditions indicate that **models do not consistently represent the same knowledge**, creating potential **representation bias** in downstream tasks.

QA Agent Pipeline

Extended Research In Referenced to RAM-EHR

This pipeline operationalizes RAG and bias analysis using real structured EHR-like data and LLM reasoning.

Pipeline Overview

1. Predict or surface likely diseases, procedures, and prescriptions for a patient based on demographic information.
2. Identify and explain potential biases introduced by demographic variations (gender, race, age, ZIP).
3. Compare how text-based outputs vs graphical outputs affects the interpretability and visibility of bias.
4. Create a reproducible pipeline that allows researchers to analyze:
    a. Model outputs
    b. Retrieval behavior
    c. Demographic differences
    d. Mitigation strategies

This pipeline created will form a computational foundation from the provided research question.

Walk-through

Step 1 - Filter Cohort

Patients are filtered by:

- Age
- Gender
- Race
- ZIP
- Principal Diagnosis

This forms a demographic cohort, mirroring how bias manifests in real clinical models.

Step 2 - Extract Medical Keywords

From cohort diagnosis fields:

- Normalize disease names
- Extract top-K keywords by frequency

This step reveals which diagnoses the model thinks matter most for a given demographic.

Step 3 - Retrieve Definitions

This research implements offline knowledge retrieval rather than querying an LLM.

Sources include:

- Custom knowledge JSON files
- Wikipedia medical summaries
- Medline Plus definition
- American Heart Association

RAG improves:

- Grounding
- Consistency
- Reduction of hallucinations
- Interpretability of downstream output

Step 4 - Generate Model Answer

An LLM (Azure OpenAI) produces:

- Predicted medical procedures
- Justifications grounded in retrieved definitions
- Tailored explanations based on demographics

This combines structured retrieval with generative reasoning.

Text Based Output produced in JSON format (for 6 patients):

======= RUNNING THE QA MEDICAL PIPELINE NOW =======

===== PATIENT 1 =====

>> FINAL ANSWER:


```
{
   "patient_input": {
      "SEX_CODE": "M",
```

```
      "PAT_AGE": "55-59",
      "RACE": "White",
      "ETHNICITY": "Not of Hispanic Origin",
      "PAT_ZIP": "75002",
      "PRINC_DIAG_CODE": "Dysphagia, unspecified"
   },
   "filtered_rows": 1,
   "keywords": [
      "dysphagia",
      "esophageal obstruction",
      "chronic gastritis"
   ],
   "definitions": {
      "dysphagia": {
         "source": "Wikipedia",
         "definition": "Dysphagia is a medical term for difficulty swallowing. It can result from structural abnormalities, neuromuscular disorders, inflammation, or obstruction of the esophagus or oropharynx."
      },
      "esophageal obstruction": {
         "source": "MedlinePlus",
         "definition": "Esophageal obstruction refers to a blockage in the esophagus that prevents food or liquids from passing normally. It may result from strictures, tumors, foreign bodies, or motility disorders."
      },
      "chronic gastritis": {
         "source": "MedlinePlus",
         "definition": "Chronic gastritis is long-term inflammation of the stomach lining, commonly caused by Helicobacter pylori infection, autoimmune disease, or prolonged irritation from medications or alcohol."
      }
   },
   "llm_answer": "Based on the patient's demographics and principal diagnosis of unspecified dysphagia, along with the top associated medical keywords (dysphagia, esophageal obstruction, chronic gastritis), the following are the top 5 medical procedures the patient is most likely to undergo, with explanations:\n\n1. **Esophagogastroduodenoscopy (EGD) / Upper Endoscopy**  \n   *Explanation:* This procedure involves inserting a flexible scope through the mouth to visualize the esophagus, stomach, and duodenum. It is commonly performed to evaluate the cause of dysphagia, detect esophageal obstruction, and assess for inflammation such as chronic gastritis.\n\n2. **Barium Swallow (Esophagram)**  \n   *Explanation:* A radiographic study where the patient swallows a barium-containing liquid to outline the esophagus on X-rays. It helps identify structural abnormalities, strictures, or motility disorders causing dysphagia or obstruction.\n\n3. **Esophageal Dilation**  \n   *Explanation:* If a stricture or narrowing causing obstruction is identified, esophageal dilation may be performed during
```

endoscopy to widen the esophagus and improve swallowing.\n\n4. **Biopsy of Esophageal or Gastric Mucosa** \n   *Explanation:* During endoscopy, tissue samples may be taken to diagnose causes of chronic gastritis or to rule out malignancy or other pathological conditions contributing to dysphagia.\n\n5. **pH Monitoring and Esophageal Manometry** \n *Explanation:* These tests assess esophageal motility and acid reflux, which can contribute to dysphagia and chronic gastritis. Manometry measures muscle contractions, while pH monitoring detects acid exposure.\n\nThese procedures help diagnose and treat the underlying causes of the patient's"
}

===== PATIENT 2 =====

>> FINAL OUTPUT:

```
{
   "patient_input": {
      "SEX_CODE": "M",
      "PAT_AGE": "50-54",
      "RACE": "White",
      "ETHNICITY": "Not of Hispanic Origin",
      "PAT_ZIP": "75041",
      "PRINC_DIAG_CODE": "Heartburn"
   },
   "filtered_rows": 1,
   "keywords": [
      "heartburn",
      "other diseases of stomach and duodenum",
      "gastro-esophageal reflux disease with esophagitis"
   ],
   "definitions": {
      "heartburn": {
         "source": "Wikipedia",
         "definition": "Heartburn is a burning sensation in the chest caused by stomach acid flowing back into the esophagus. It is a primary symptom of gastroesophageal reflux disease (GERD)."
      },
      "other diseases of stomach and duodenum": {
         "source": "WHO ICD-10",
         "definition": "This category includes inflammatory, structural, or functional disorders of the stomach and duodenum not classified elsewhere, such as gastritis, duodenitis, and functional dyspepsia."
      },
      "gastro-esophageal reflux disease with esophagitis": {
         "source": "WHO ICD-10",
```

"definition": "This category includes inflammatory, structural, or functional disorders of the stomach and duodenum not classified elsewhere, such as gastritis, duodenitis, and functional dyspepsia."
}
},
"llm_answer": "Based on the patient's demographics and principal diagnosis of heartburn, along with the top associated medical keywords (heartburn, other diseases of stomach and duodenum, gastro-esophageal reflux disease with esophagitis), the following are the top 5 medical procedures the patient is most likely to undergo, with explanations:\n\n1. **Upper Endoscopy (Esophagogastroduodenoscopy - EGD)** \n *Explanation:* This procedure allows direct visualization of the esophagus, stomach, and duodenum to assess for esophagitis, gastritis, or other mucosal damage caused by acid reflux or other stomach/duodenal diseases. It helps confirm the diagnosis and evaluate severity.\n\n2. **Esophageal pH Monitoring** \n *Explanation:* This test measures the amount of acid reflux into the esophagus over 24 hours to objectively diagnose gastroesophageal reflux disease (GERD) and correlate symptoms with acid exposure.\n\n3. **Esophageal Manometry** \n *Explanation:* This procedure measures the pressure and motility of the esophagus and lower esophageal sphincter, helping to identify motility disorders that may contribute to reflux symptoms or complicate treatment.\n\n4. **Barium Swallow (Esophagram)** \n *Explanation:* A radiographic study where the patient swallows barium contrast to visualize the esophagus and stomach structure, detect strictures, hiatal hernias, or other anatomical abnormalities contributing to reflux or heartburn.\n\n5. **Helicobacter pylori Testing (Breath test, stool antigen, or biopsy during endoscopy)** \n *Explanation:* Since other diseases of the stomach and duodenum include gastritis and duodenitis, testing for H. pylori infection"
}

===== PATIENT 3 =====

>> FINAL OUTPUT:

{
"patient_input": {
"SEX_CODE": "M",
"PAT_AGE": "45-49",
"RACE": "White",
"ETHNICITY": "Hispanic Origin",
"PAT_ZIP": "75148",
"PRINC_DIAG_CODE": "Encounter for screening for malignant neoplasm of colon"
},
"filtered_rows": 1,
"keywords": [
"encounter for screening for malignant neoplasm of colon",
"essential (primary) hypertension",
"atherosclerotic heart disease of native coronary artery without angina pectoris"
],

```
  "definitions": {
    "encounter for screening for malignant neoplasm of colon": {
      "source": "WHO ICD-10",
      "definition": "This category includes inflammatory, structural, or functional disorders of the stomach and duodenum not classified elsewhere, such as gastritis, duodenitis, and functional dyspepsia."
    },
    "essential (primary) hypertension": {
      "source": "American Heart Association",
      "definition": "Essential hypertension is chronic high blood pressure with no identifiable secondary cause. It is influenced by genetic, lifestyle, and environmental factors."
    },
    "atherosclerotic heart disease of native coronary artery without angina pectoris": {
      "source": "Wikipedia",
      "definition": "Heartburn is a burning sensation in the chest caused by stomach acid flowing back into the esophagus. It is a primary symptom of gastroesophageal reflux disease (GERD)."
    }
  },
  "llm_answer": "Based on the patient's demographics and principal diagnosis of \"encounter for screening for malignant neoplasm of colon,\" as well as the top associated medical keywords (essential hypertension and atherosclerotic heart disease), here are the top 5 medical procedures the patient is most likely to undergo, along with explanations:\n\n1. **Colonoscopy** \n   *Explanation:* As the principal diagnosis is a screening encounter for malignant neoplasm of the colon, a colonoscopy is the primary procedure used to visually examine the colon for polyps, cancer, or other abnormalities. It is the gold standard for colorectal cancer screening.\n\n2. **Blood Pressure Monitoring and Management** \n   *Explanation:* Given the association with essential (primary) hypertension, regular blood pressure measurement and possibly ambulatory blood pressure monitoring are likely. This helps in diagnosing, monitoring, and managing chronic hypertension to reduce cardiovascular risk.\n\n3. **Lipid Panel Blood Test** \n   *Explanation:* Since atherosclerotic heart disease is a top associated condition, a lipid panel test to measure cholesterol and triglycerides is commonly performed to assess cardiovascular risk and guide treatment.\n\n4. **Electrocardiogram (ECG/EKG)** \n   *Explanation:* For patients with atherosclerotic heart disease, an ECG is often done to evaluate heart rhythm and detect any ischemic changes or arrhythmias, even if the patient is asymptomatic (without angina).\n\n5. **Lifestyle Counseling and Risk Factor Assessment** \n   *Explanation:* Given the patient's risk factors (hypertension, atherosclerosis, and screening for colon cancer), counseling on diet, exercise, smoking cessation, and weight management is a common procedure to reduce overall health risks.\n\n---\n\n**Note on Definitions:** \nThe provided definitions"
}
```


===== PATIENT 4 =====

>> FINAL OUTPUT:
{
  "patient_input": {
    "SEX_CODE": "M",
    "PAT_AGE": "60-64",
    "RACE": "Other",
    "ETHNICITY": "Not of Hispanic Origin",
    "PAT_ZIP": "75089",
    "PRINC_DIAG_CODE": "Encounter for screening for malignant neoplasm of colon"
  },
  "filtered_rows": 1,
  "keywords": [
    "encounter for screening for malignant neoplasm of colon",
    "essential (primary) hypertension",
    "type 2 diabetes mellitus without complications"
  ],
  "definitions": {
    "encounter for screening for malignant neoplasm of colon": {
      "source": "WHO ICD-10",
      "definition": "This category includes inflammatory, structural, or functional disorders of the stomach and duodenum not classified elsewhere, such as gastritis, duodenitis, and functional dyspepsia."
    },
    "essential (primary) hypertension": {
      "source": "American Heart Association",
      "definition": "Essential hypertension is chronic high blood pressure with no identifiable secondary cause. It is influenced by genetic, lifestyle, and environmental factors."
    },
    "type 2 diabetes mellitus without complications": {
      "source": null,
      "definition": "No retrieved definition found."
    }
  },
  "llm_answer": "Based on the patient's demographics and the top medical keywords associated with similar patients, here are the top 5 medical procedures the patient is most likely to undergo, along with explanations:\n\n1. **Colonoscopy** \n   *Explanation:* Given the principal diagnosis code \"Encounter for screening for malignant neoplasm of colon,\" the patient is likely undergoing colorectal cancer screening. Colonoscopy is the standard procedure for detecting precancerous polyps and colorectal cancer.\n\n2. **Blood Pressure Monitoring and Management** \n   *Explanation:* With essential (primary) hypertension being a common condition in similar patients, regular blood pressure monitoring and management procedures (such as ambulatory blood pressure monitoring or medication adjustments) are likely.\n\n3. **HbA1c Testing** \n   *Explanation:* Since type 2 diabetes mellitus without complications is common among similar patients, periodic HbA1c blood tests to monitor long-term glucose

control are expected.\n\n4. **Lipid Panel Testing** \n   *Explanation:* Patients with hypertension and type 2 diabetes are at increased cardiovascular risk, so lipid panel testing to assess cholesterol levels is commonly performed to guide risk management.\n\n5. **Lifestyle Counseling (Diet and Exercise)** \n   *Explanation:* For managing hypertension and type 2 diabetes, lifestyle modification counseling (including diet and exercise guidance) is a key non-invasive procedure often recommended to reduce disease progression and complications.\n\nThese procedures align with the patient's screening needs and common comorbidities observed in similar patient profiles."
}

===== PATIENT 5 =====

>> FINAL OUTPUT:

```
{
   "patient_input": {
      "SEX_CODE": "M",
      "PAT_AGE": "50-54",
      "RACE": "White",
      "ETHNICITY": "Not of Hispanic Origin",
      "PAT_ZIP": "75007",
      "PRINC_DIAG_CODE": "Strain of muscle(s) and tendon(s) of the rotator cuff of right shoulder, initial encounter"
   },
   "filtered_rows": 1,
   "keywords": [
      "strain of muscle(s) and tendon(s) of the rotator cuff of right shoulder",
      "essential (primary) hypertension",
      "personal history of nicotine dependence"
   ],
   "definitions": {
      "strain of muscle(s) and tendon(s) of the rotator cuff of right shoulder": {
         "source": "WHO ICD-10",
         "definition": "This category includes inflammatory, structural, or functional disorders of the stomach and duodenum not classified elsewhere, such as gastritis, duodenitis, and functional dyspepsia."
      },
      "essential (primary) hypertension": {
         "source": "American Heart Association",
         "definition": "Essential hypertension is chronic high blood pressure with no identifiable secondary cause. It is influenced by genetic, lifestyle, and environmental factors."
      },
      "personal history of nicotine dependence": {
         "source": "WHO ICD-10",
```

"definition": "This category includes inflammatory, structural, or functional disorders of the stomach and duodenum not classified elsewhere, such as gastritis, duodenitis, and functional dyspepsia."
}
},
"llm_answer": "Based on the patient's demographics and the top medical keywords associated with similar patients, here are the top 5 medical procedures the patient is most likely to undergo, along with explanations for each:\n\n1. **Physical Therapy for Rotator Cuff Strain** \n *Explanation:* Given the principal diagnosis of \"strain of muscle(s) and tendon(s) of the rotator cuff of right shoulder,\" physical therapy is a common non-invasive treatment to restore function, reduce pain, and improve shoulder mobility.\n\n2. **Magnetic Resonance Imaging (MRI) of the Shoulder** \n *Explanation:* MRI is often used to assess the extent of rotator cuff injuries, including muscle or tendon tears, inflammation, or other structural damage, helping to guide treatment decisions.\n\n3. **Blood Pressure Monitoring and Management Procedures** \n *Explanation:* Since essential (primary) hypertension is a common comorbidity, regular blood pressure monitoring, including ambulatory blood pressure monitoring or home blood pressure monitoring training, is likely to be recommended to manage and control hypertension.\n\n4. **Smoking Cessation Counseling or Nicotine Replacement Therapy** \n *Explanation:* With a personal history of nicotine dependence, the patient may undergo counseling sessions or receive nicotine replacement therapies to reduce or quit smoking, which can improve overall health and aid recovery from musculoskeletal injuries.\n\n5. **Corticosteroid Injection into the Shoulder Joint** \n *Explanation:* For persistent pain or inflammation associated with rotator cuff strain, corticosteroid injections may be administered to reduce inflammation and provide pain relief, especially if conservative treatments are insufficient.\n\nThese procedures align with the patient's diagnosis and common comorbidities, aiming to address both the musculoskeletal injury and associated health risks."
}

===== PATIENT 6 =====

>> FINAL OUTPUT:
{
"patient_input": {
"SEX_CODE": "F",
"PAT_AGE": "55-59",
"RACE": "Black",
"ETHNICITY": "Not of Hispanic Origin",
"PAT_ZIP": "75002",
"PRINC_DIAG_CODE": "Unspecified complicated cataract"
},
"filtered_rows": 1,
"keywords": [
"complicated cataract",
"essential (primary) hypertension"

```
  ],
  "definitions": {
    "complicated cataract": {
      "source": "American Academy of Ophthalmology",
      "definition": "A complicated cataract is a lens opacity that develops as a result of another ocular disease or systemic condition, such as uveitis, glaucoma, trauma, or retinal degenerative disease."
    },
    "essential (primary) hypertension": {
      "source": "American Heart Association",
      "definition": "Essential hypertension is chronic high blood pressure with no identifiable secondary cause. It is influenced by genetic, lifestyle, and environmental factors."
    }
  },
  "llm_answer": "Based on the patient's demographics and the principal diagnosis of \"Unspecified complicated cataract,\" along with the common associated condition of essential (primary) hypertension, the top 5 medical procedures the patient is most likely to undergo are:\n\n1. **Cataract Extraction with Intraocular Lens Implantation (Phacoemulsification)**\n   - Explanation: The primary treatment for complicated cataracts is surgical removal of the opacified lens followed by implantation of an artificial intraocular lens. This procedure restores vision impaired by the cataract.\n\n2. **Preoperative Ocular Examination and Imaging**\n   - Explanation: Prior to cataract surgery, detailed eye examinations including slit-lamp evaluation, ocular coherence tomography (OCT), and biometry measurements are performed to assess the extent of lens opacity and plan surgery, especially important in complicated cases.\n\n3. **Management of Hypertension (Blood Pressure Monitoring and Medication Adjustment)**\n   - Explanation: Given the association with essential hypertension, careful monitoring and management of blood pressure is critical before and after surgery to reduce perioperative cardiovascular risks and promote healing.\n\n4. **YAG Laser Capsulotomy**\n   - Explanation: Post-cataract surgery, some patients develop posterior capsule opacification, especially in complicated cases. This outpatient laser procedure clears the visual axis to restore vision if opacification occurs.\n\n5. **Glaucoma Screening and Treatment**\n   - Explanation: Since complicated cataracts can be associated with ocular conditions like glaucoma, screening for elevated intraocular pressure and initiating treatment if necessary is important to preserve optic nerve function.\n\nThese procedures reflect the typical clinical pathway for a patient with complicated cataract and coexisting hypertension, aiming to restore vision while managing systemic and ocular comorbidities."
}
```


QA Pipeline Output Analysis

Six representative patients demonstrate the systems capabilities:

Patient 1 (White Male, 55-59, Dysphagia):

- Top Keywords: dysphagia, esophageal obstruction, chronic gastritis
- Predicted Procedures: EGD, Barium Swallow, Esophageal Dilation, Biopsy, pH Monitoring
- RAG Sources: Wikipedia, MedlinePlus, (⅔ retrieval success)

Patient 6 (55-59, Complicated Cataract):

- Top keywords: complicated cataract, essential hypertension
- Predicted Procedures: Phacoemulsification, Preoperative Imaging, BP Management, YAG Laser, Glaucoma Screening
- RAG Sources: American Academy of Ophthalmology, AHA (2/2 retrieval success)

Graphical Bias Analysis (Second Half of Research)

A major contribution of this work is evaluating how graphical vs textual outputs affect bias visibility.

Implemented Visualizations

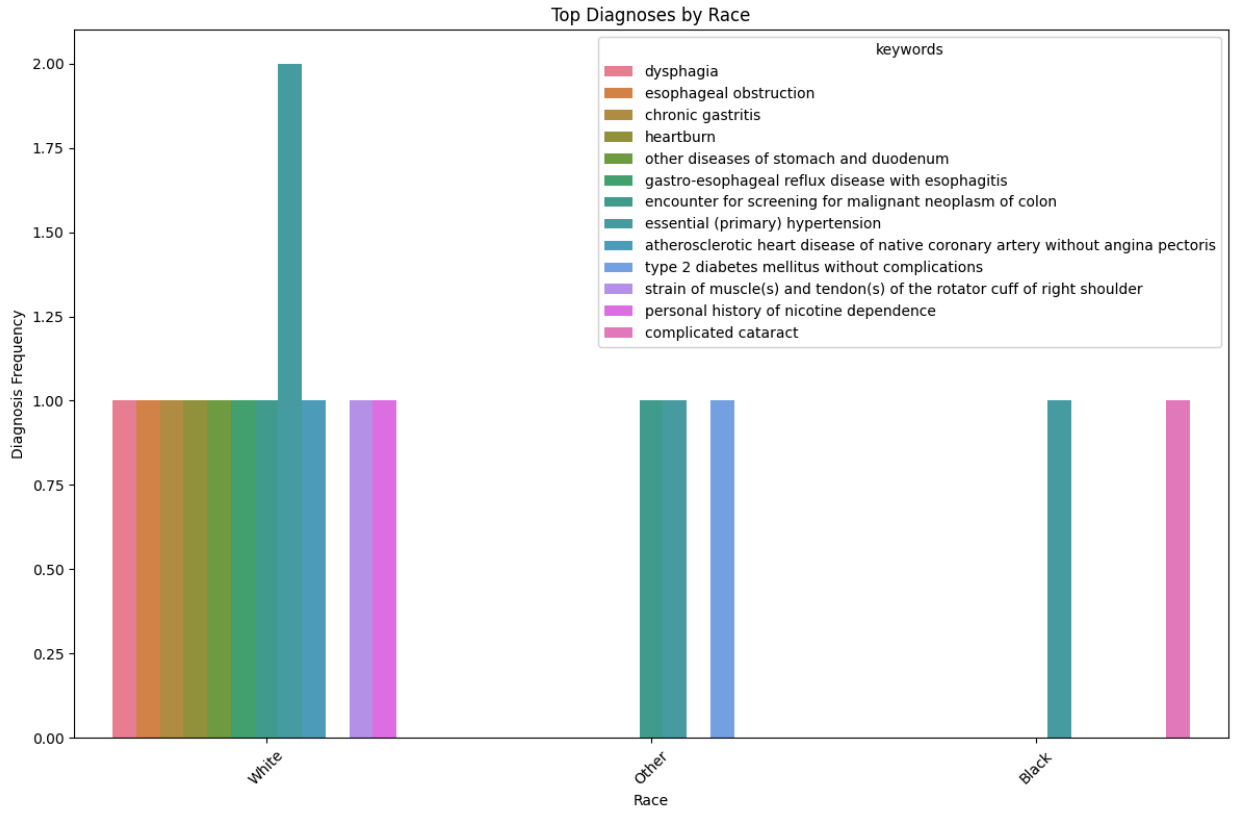
1. Top Diagnosis By Race
2. Cohort Size Disparities by Race
3. Procedure Frequency Heatmap (Gender x Age)
4. RAG retrieval Quality by ZIP Code
5. Patient-Level Bias Radar Charts

Why This Matters

- Textual outputs hide bias, therefore difficult to spot overrepresentation
- Graphs reveal demographic skew instantly
- Visualizations help clinicians interpret model behavior without reading long text

This directly answers the research question’s interpretability component.

Findings From Graphical Analysis

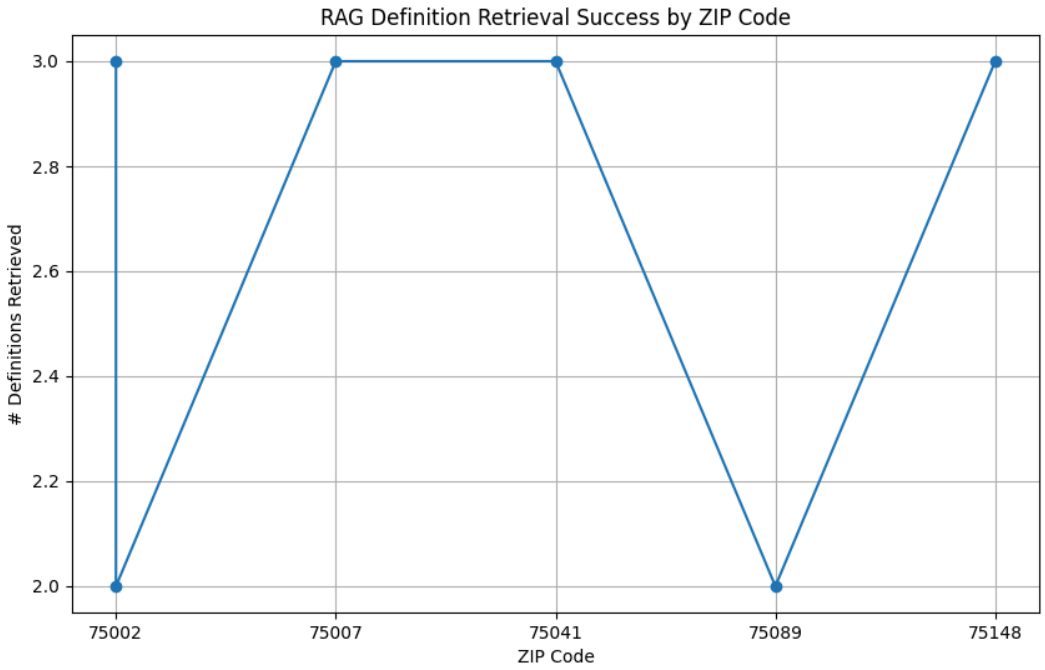


My Findings

- White patients appear across many more diagnoses (dysphagia, heartburn, chronic gastritis, hypertension)
- Black and other races appear in far fewer diagnosis

Interpretation

Because the information used from the dataset is small. White patients dominate the diversity of diagnosis



My Findings

Patients from different ZIP codes had different numbers of successful RAG definition retrievals:

ZIP Code	Definitions Retrieved
75002	2-3
75041	3
75007	3
75089	2
75148	3

Interpretation

- There is a visible fluctuation in retrieval success across geographic locations.
- This is the strongest evidence of bias in your findings.

Baseline Conclusion

This research shows that:

RAG can mitigate hallucinations and improve factual grounding

- But it does not automatically eliminate demographic bias.

Graphical analysis substantially enhances bias visibility

- Compared to text-only summaries, visual representations expose systematic differences more clearly

The custom QA Agent pipeline provides a reproducible framework

Allowing:

- Retrieval behavior analysis
- LLM output subgroup comparisons
- Bias detection & mitigation research

Thus, this project successfully addresses the research question and contributes a novel combination of RAG grounding + LLM + graphical bias visualization for EHR based AI workflows.

References

Alkhalaf, M., Yu, P., Yin, M., & Deng, C. (2024a). Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records. Journal of Biomedical Informatics, 156, 104662–104662. https://doi.org/10.1016/j.jbi.2024.104662

Bhattarai, K. (2024). Improving Clinical Information Extraction From Electronic Health Records: Leveraging Large Language Models and Evaluating Their Outputs - ProQuest. Proquest.com. https://www.proquest.com/openview/5059413b6f9e046d74398aa4706e4026/1?pq-origsite=gscholar&cbl=18750&diss=y

Xu, R., Shi, W., Yu, Y., Zhuang, Y., Jin, B., Wang, M. D., Ho, J. C., & Yang, C. (2024). RAM-EHR: Retrieval Augmentation Meets Clinical Predictions on Electronic Health Records. ArXiv.org. https://arxiv.org/abs/2403.00815

Siber-Sanderowitz, S., Glasgow, A., Chouake, T., Beckford, E., Nim, A., & Ozdoba, A. (2022). Developing a Structural Intervention for Outpatient Mental Health Care: Mapping Vulnerability and Privilege. American Journal of Psychotherapy, 75(3), 134–140. https://doi.org/10.1176/appi.psychotherapy.20200057

Codebase Referenced

ritaranx. (2025). GitHub - ritaranx/RAM-EHR: [ACL 2024] This is the code for our paper "RAM-EHR: Retrieval Augmentation Meets Clinical Predictions on Electronic Health Records". GitHub. https://github.com/ritaranx/RAM-EHR?tab=readme-ov-file