Medical Question Answering for Critical Care Medicine

Results and Issues

Mar.18 2025

Overview

- **Outline**
  - Methods (Proposed solutions and Issues).
  - Results (Retrieval, Synthesis)
  - Issues (hallucination, performance)
- **Progress**

Item	Description	Note
Results	Retrieval, Stance, Summary	See below
Model buildup	In progress (with issues)	
Paper	Introduction and Methods	Draft
Dataset	Paper and Question collection, Answers Annotation	Done

Objective

We aim to design and execute an **experiment to demonstrate our proposed methodology** for medical question answering for critical care medicine.

Example Question/ Claim:

Steroids improves survival and reversal of shock **in patients with septic shock**.

Supporting the claim:

Paper 1
Study Design and Methodology:
Study Population:
Interventions:
Comparator:
Outcomes:
Strengths and Weaknesses:
Key Findings and Conclusion:

Paper 2
…

Against the claim:

Paper 3
Study Design and Methodology:
Study Population:
…

Dataset

- **Research Papers**
- 112 research papers related to any medical conditions are collected for model build up and evaluation. (34 papers from WikiJournal, 86 from PubMed, Google Scholar)
- **Medical Questions**
- 36 yes-no medical questions (12 for ARDS, 10 for sepsis, 6 for cardiac arrest, 7 for delirium, 1 for Sepsis *and* delirium).

Annotation

- **Relevance, Stances**
  - Each research paper was manually annotated against all in medical questions (yes-no) or claim. (For relevance, 1 indicates relevant and 0 indicates irrelevant. For stance, 1 indicates supporting, 0 indicates neutral, and -1 indicates against the claim).
- **PICO format summary**
  - Claude from AWS Bedrock was leveraged to generate a PICO format summary (Study Design and Methodology, Study Population, Interventions, Comparator, Outcomes, Strengths and Weaknesses, Key Findings and Conclusion). The generated summary was edited and proofread KN and served as ground truth.

Evaluation (code demo)

- **Relevance**
  - Relevance of retrieved papers measured by precision and recall (aggregated with all medical questions).
- **Factuality**
  - Stances evaluated using overall accuracy.
  - Synthesis (PICO summary) evaluated human rating (subset), Ragas
- **Consistency**
  - All experiments will be executed 5 times at temperature 0.6.

Model and Issue

- **GraphRAG**: modified prompt from the Edge et al.’s solution

  → Issue: black box, hallucination (retrieval + synthesis)

- **Naïve RAG + Knowledge Graph**
  - Retrieval: separate text into 100 chunk, use query-chunk embedding similarity.
  - Synthesis: Hybrid

  → Issue: performance, hallucination (synthesis), incomplete summary

Results - Retrieval

	Recall	Precision	F-1	Accuracy
Naïve RAG (Suim)	.476	.560	.444	.896
Naïve RAG (Bob)	.516	.396	.390	.927

- Mean value of all metrics from 36 questions
- Database: 112 papers (ARDS, Sepsis, Cardiac Arrest, and Delirium)
- **Different retrieval results from different knowledge graph under the same condition**

Issue - Hallucination

Supporting the Claim (**Does early administration of neuromuscular blocking agents increases the ventilator free days?**)

*2. **Impact of Early Neuromuscular Blockade on ARDS Mortality*** ← **NON EXISTING PAPER**

- *Authors*: Moss et al. *Year*: 2016
- *Journal*: Critical Care Medicine
- *Research Question*: Does early NMBA administration affect mortality in ARDS patients?
- *Study Design and Methodology*: Cohort study analyzing data from a large database.
- *Study Population*: 1,000+ ARDS patients.
- *Interventions & Comparators*: Early NMBA administration vs. standard care.
- *Outcomes*: Decreased mortality and increased ventilator-free days.
- *Strengths and Weaknesses*: Large sample size is a strength; retrospective design is a weakness.
- *Key Findings and Conclusion*: Supports the claim with significant findings on mortality and ventilator-free days.

Issue: GraphRAG

Equivocal or Inconclusive:
1. **Study 1**
   - **Authors:** Annane et al.
   - **Year:** Not specified
   - **Title:** "Effect of Treatment With Low Doses of Hydrocortisone and Fludrocortisone on Mortality in Patients With Septic Shock"
   - **Journal:** Not specified
   - **Research Question:** Examined the impact of low-dose corticosteroids on 28-day survival in patients with septic shock and adrenal insufficiency.
   - **Type of Study:** Placebo-controlled, randomized trial
   - **Methodology:**
     - Randomized patients to hydrocortisone and fludrocortisone vs. placebos.
     - Main outcome: 28-day survival in nonresponders to corticotropin.
   - **Key Findings:**
     - No significant difference in 28-day mortality between hydrocortisone and placebo.
     - No difference in mortality between responders and nonresponders to corticotropin.
   - **Conclusion:** Low-dose corticosteroids did not improve 28-day survival in septic shock patients with adrenal insufficiency.

Question: Does early application of high frequency oscillatory ventilation compared with ventilation strategy of low tidal volume decrease mortality?

Ground Truth: Supporting (1) / **Synthesis**: Equivocal or Inconclusive (0)

Issue: GraphRAG

Refuting the Claim:
2. **Study 2: CORTICUS**
   - **Authors:** Charles L. Sprung et al.
   - **Year:** 2008
   - **Title:** Hydrocortisone Therapy for Patients with Septic Shock
   - **Journal:** New England Journal of Medicine
   - **Research Question:** To assess the impact of hydrocortisone therapy on mortality in septic shock patients.
   - **Study Design:** Randomized, double-blind, placebo-controlled trial.
   - **Population Size:** 499 patients with septic shock.
   - **Intervention:** Hydrocortisone infusion.
   - **Primary Outcome:** No significant difference in mortality between hydrocortisone and placebo groups.
   - **Key Findings:** Hydrocortisone did not improve survival or reversal of septic shock.
   - **Conclusion:** The study did not find evidence supporting a reduction in mortality with hydrocortisone in septic shock patients.

Equivocal or Inconclusive:
3. **Study 3: APROCCHSS**
   - This study supports the claim that hydrocortisone plus fludrocortisone reduced 90-day mortality in septic shock patients.

Question: Does early application of high frequency oscillatory ventilation compared with ventilation strategy of low tidal volume decrease mortality?

Insufficient Information: Methodology, Comparator, Strength and Weakness

Synthesis and Evaluation

	Study Design and Methodology
Ground Truth (ACURASYS)	- **Multicenter, randomized, double-blind**, placebo-controlled trial - March 2006 through March 2008 - 20 Intensive Care Units (ICUs) in France - Patients **randomly assigned to cisatracurium or placebo** groups - Double-blind design - Independent data and safety monitoring board

JSON:	{
	"Study Design": "Placebo trial",
	Intervention: “cisatracurium”,
	Population: “340”
	}

Generated Summary	- **Multicenter, double-blind trial** - **Randomly assigned patients to receive cisatracurium besylate or placebo.**
Score	Factual Correctness (Recall): **0.25** Factual Correctness (Precision): **0.67**

	Study Population

Ground Truth (ACURASYS)	The study enrolled **340 patients** from 20 ICUs in France. The patients were randomly assigned to either receive cisatracurium besylate (178 patients) or a placebo (162 patients) for 48 hours. The study population consisted of patients with severe ARDS, defined as a ratio of partial pressure of arterial oxygen (PaO2) to the fraction of inspired oxygen (FIO2) of less than 150, with a positive end-expiratory pressure of 5 cm or more of water, and a tidal volume of 6 to 8 ml per kilogram of predicted body weight.- Total Participants: 340 patients
	- Inclusion Criteria:
	  - Receiving endotracheal mechanical ventilation
	  - Acute hypoxemic respiratory failure
	  - Severe ARDS (PaO2:FIO2 < 150)
	  - ARDS onset within previous 48 hours
	- Cisatracurium Group: 178 patients
	- Placebo Group: 162 patients

Generated Summary	- **340 patients**
Score	Factual Correctness (Recall): **0.08** Factual Correctness (Precision): **1.0** Semantic Similarity : **0.45**

	Intervention	Comparator

Ground Truth (ACURASYS)	- Cisatracurium Group:
	  - 3-ml rapid IV infusion of 15 mg
	  - Continuous infusion of 37.5 mg per hour for 48 hours
	- Ventilation Protocol:
	  - Volume assist-control mode
	  - Tidal volume: 6-8 ml per kg of predicted body weight
	  - Target SpO2: 88-95%
		- Placebo group receiving identical IV infusion without neuromuscular blocking agent

Generated Summary	- Cisatracurium besylate or placebo for 48 hours.	N/A
Score	Factual Correctness (Recall): **0.17** Factual Correctness (Precision): **0.5** Semantic Similarity : **0.65**	Factual Correctness (Recall): **0** Factual Correctness (Precision): **0** Semantic Similarity : **0.06**

	Outcome	Strengths and Weaknesses
Ground Truth (ACURASYS)	Primary Outcome: - 90-day in-hospital mortality rate Secondary Outcomes: - 28-day mortality - Ventilator-free days - Days outside ICU - Days without organ failure - Barotrauma rate - ICU-acquired paresis	Strengths: - Multicenter design - Blinded randomization - Comprehensive follow-up - Intention-to-treat analysis Weaknesses: - Limited to cisatracurium - Underpowered study - No assessment of late-stage ARDS intervention - Lack of data on conditions affecting neuromuscular blockade
Generated Summary	- Reduced 90-day in-hospital mortality with cisatracurium.	N/A
Score	Factual Correctness (Recall): **0** Factual Correctness (Precision): **0** Semantic Similarity : **0.49**	Factual Correctness (Recall): **0** Factual Correctness (Precision): **0** Semantic Similarity : **0.05**

	Key Findings and Conclusion
Ground Truth (ACURASYS)	Key Findings: - **Adjusted 90-day survival improved** (hazard ratio 0.68) - 90-day mortality: 31.6% (cisatracurium) vs. 40.7% (placebo) - More ventilator-free days - Reduced barotrauma - No significant increase in muscle weakness Conclusion: **Early administration of neuromuscular blocking agent in severe ARDS**: - Improved adjusted 90-day survival - Increased ventilator-free days - Did not increase muscle weakness - Potentially beneficial in patients with PaO2:FIO2 ratio < 120
Generated Summary	Key Findings: **Cisatracurium group had lower mortality at 90 days**. Conclusion: **Cisatracurium reduced mortality in early, severe ARDS patients.**
Score	Factual Correctness (Recall): **0.18** Factual Correctness (Precision): **1.0** Semantic Similarity : **0.70**

Next Steps

Date	Description
Mar. 19	JAMIA deadline
Mar. 20	In Class Presentation (4:40 pm)
Mar. 21	Model refinement and troubleshooting. Result write up. Draft symposium presentation.
Mar. 28	Finalize symposium presentation. Discussion write up.
Apr. 4	Capstone Symposium

Backup

Ground Truth Annotation I

- Each research paper was **annotated** in terms of **relevance and stance against *all*** in medical questions (yes-no) or claim. The annotations can be used for **automatic output evaluation**

Relevance Annotation

	Q1	Q2	…
Paper 1	1 (relevant)	0 (irrelevant)	…
Paper 2	…	1 (relevant)	…
…	…	…	…

Stance Annotation

	Q1	Q2	…
Paper 1	1 (supporting)	0 (neutral)	…
Paper 2	…	-1 (against)	…
…	…	…	…

BACK

Ground Truth Annotation II

- For each research paper, we leveraged large language models (specifically, Claude from AWS Bedrock), to **generate summaries**. **After editing and proofreading by physicians** (KN, MA), the summaries will serve as **ground truth** for automatic/ human evaluation.

	Study Design and Methodology	Study Population	Interventions	…	Key Findings and Conclusion
Paper 1			…		
Paper 2			…		
…	…	…	…		