Reasoning-Enhanced Healthcare Predictions with Knowledge Graph Retrieval

Pengcheng Jiang[1], Cao Xiao[2], Minhao Jiang[2], Parminder Bhatia[2], Taha Kass-Hout[2], Jimeng Sun[1], Jiawei Han[1]

[1]University of Illinois at Urbana-Champaign [2]GE HealthCare

Presenter: Pengcheng (Patrick) Jiang

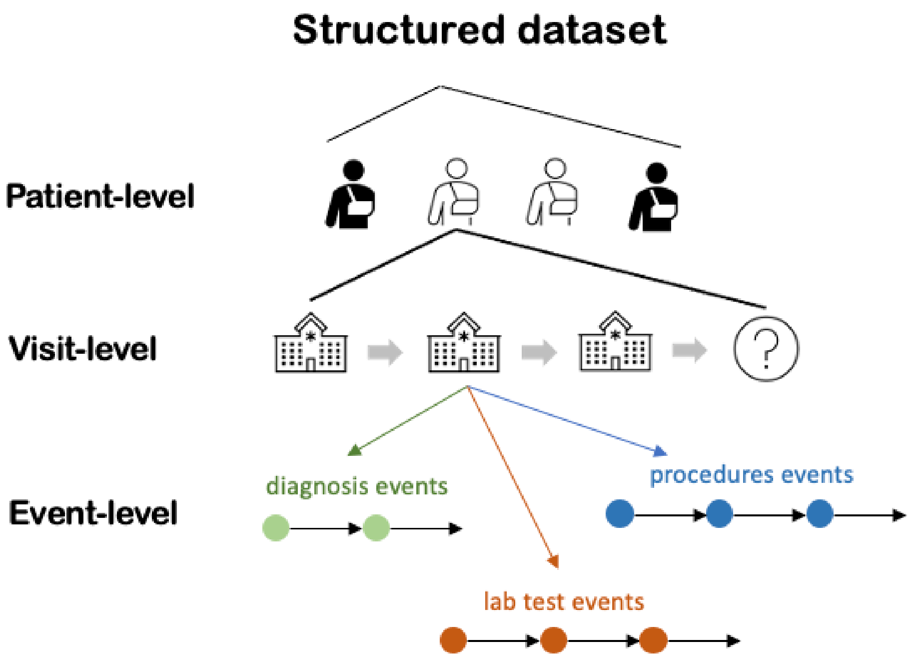
CS Ph.D. Student at UIUC

Overview

- Background
- Motivation
- Methodology
- Experiments
- Conclusion

Background

Electronic Healthcare Records (EHR) data



EHR-based Healthcare Prediction

- Using patient's history data to predict clinical outcome (e.g., mortality)

Data sample:

```
{
    "patient_id": "p001",
    "visit_id": "v001",
    "diagnoses": [...],
    "labs": [...],
    "procedures": [...],
    "label": 1,
}
```

Mortality prediction: "Will this patient die during next visit? (0/1)"

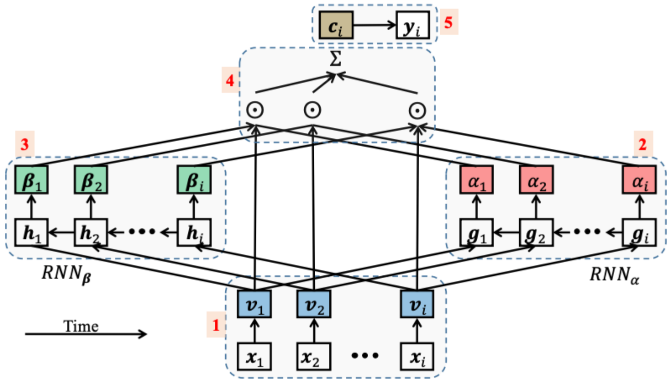
Readmission prediction: "Will this patient be readmitted within n days? (0/1)"

… …

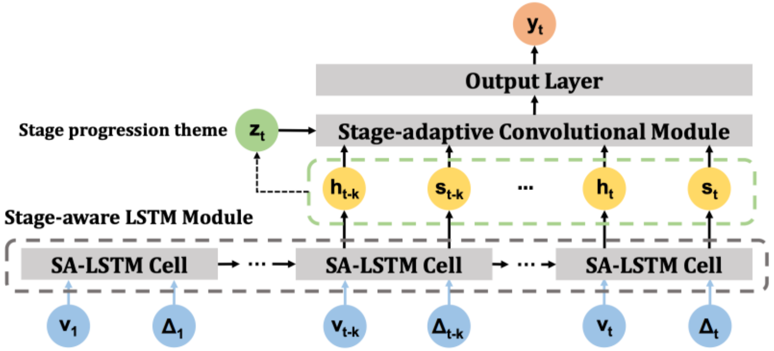
Background

ML-based Methods

RETAIN (Choi et al. NIPS'16)



StageNet (Gao et al. WWW'20)



…

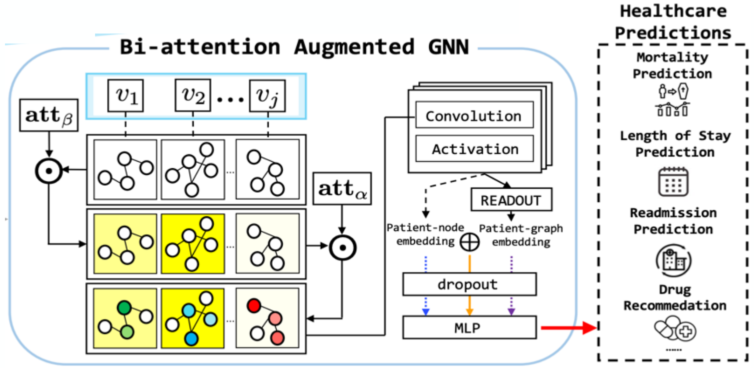
Limitations:
1. Lack interpretability; Limited ability to handle unstructured or complex medical information
2. Struggle to generalize beyond their original scope

Choi, Edward, et al. "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism." *NIPS 2016.*
Gao, Junyi, et al. "Stagenet: Stage-aware neural networks for health risk prediction." *WWW 2020.*

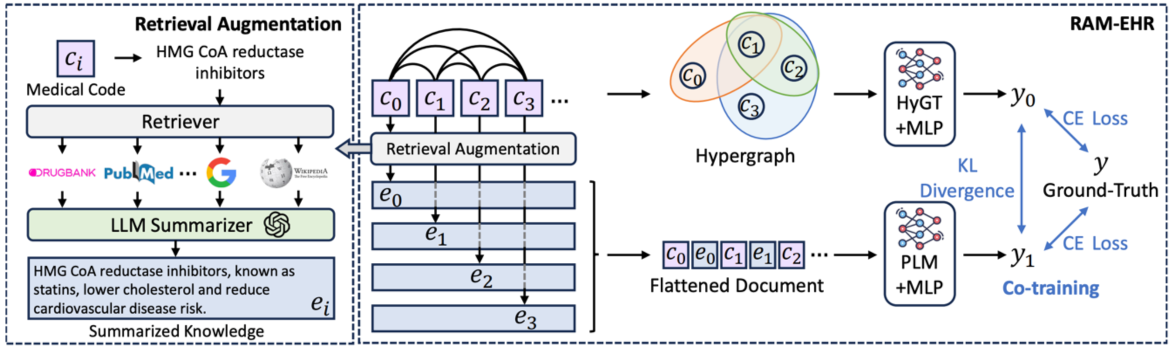
Background

LM+ML-based Methods

GraphCare (Jiang et al. ICLR'24)



RAM-EHR (Xu et al. ACL'24)



…

Limitations:

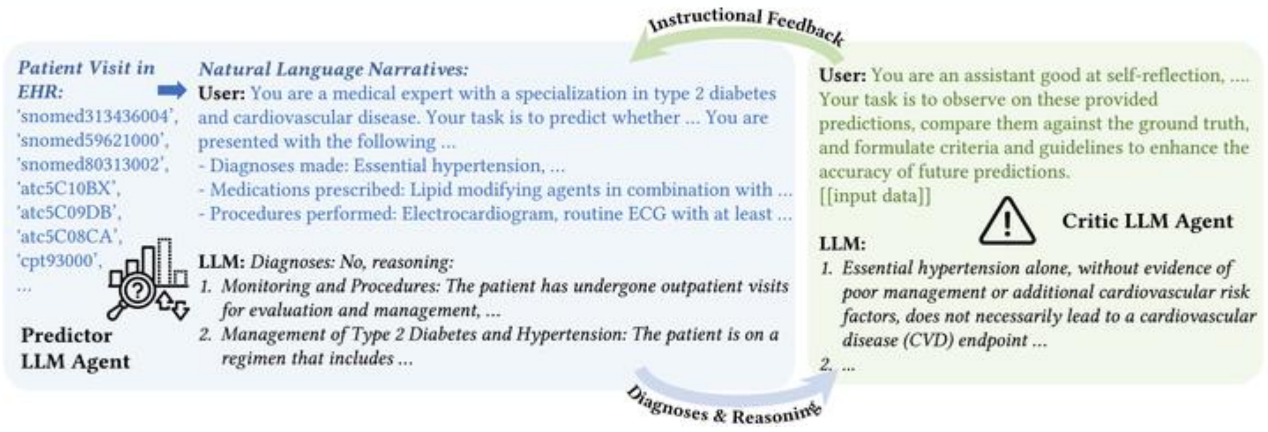
1. Lack interpretability; Not exploit reasoning capabilities of LMs
2. Retrieved information may be sparse or irrelevant

Jiang, Pengcheng, et al. "Graphcare: Enhancing healthcare predictions with personalized knowledge graphs." *ICLR 2024.*
Xu, Ran, et al. "Ram-ehr: Retrieval augmentation meets clinical predictions on electronic health records." *ACL 2024.*

Background

LLM-based Methods

EHR-CoAgent (Cui et al. 2024)



- *A framework combines the strengths of predictive agent reasoning and critical agent instruction*

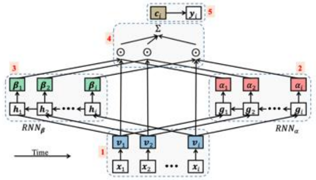
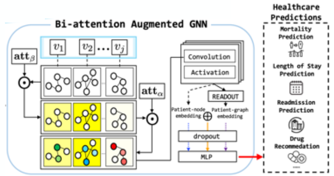
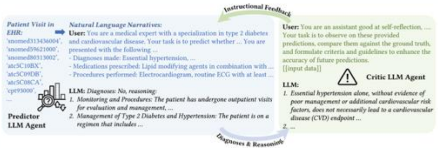
Limitations:

1. Lack of specialized medical knowledge → High risk of hallucination
2. Pure prompting-based → Cannot learn underlying EHR pattern
3. Performance highly dependent on quality of prompting and context

Cui, Hejie, et al. "LLMs-based Few-Shot Disease Predictions using EHR: A Novel Approach Combining Predictive Agent Reasoning and Critical Agent Instruction." *AMIA Annual Symposium 2024*

Motivation

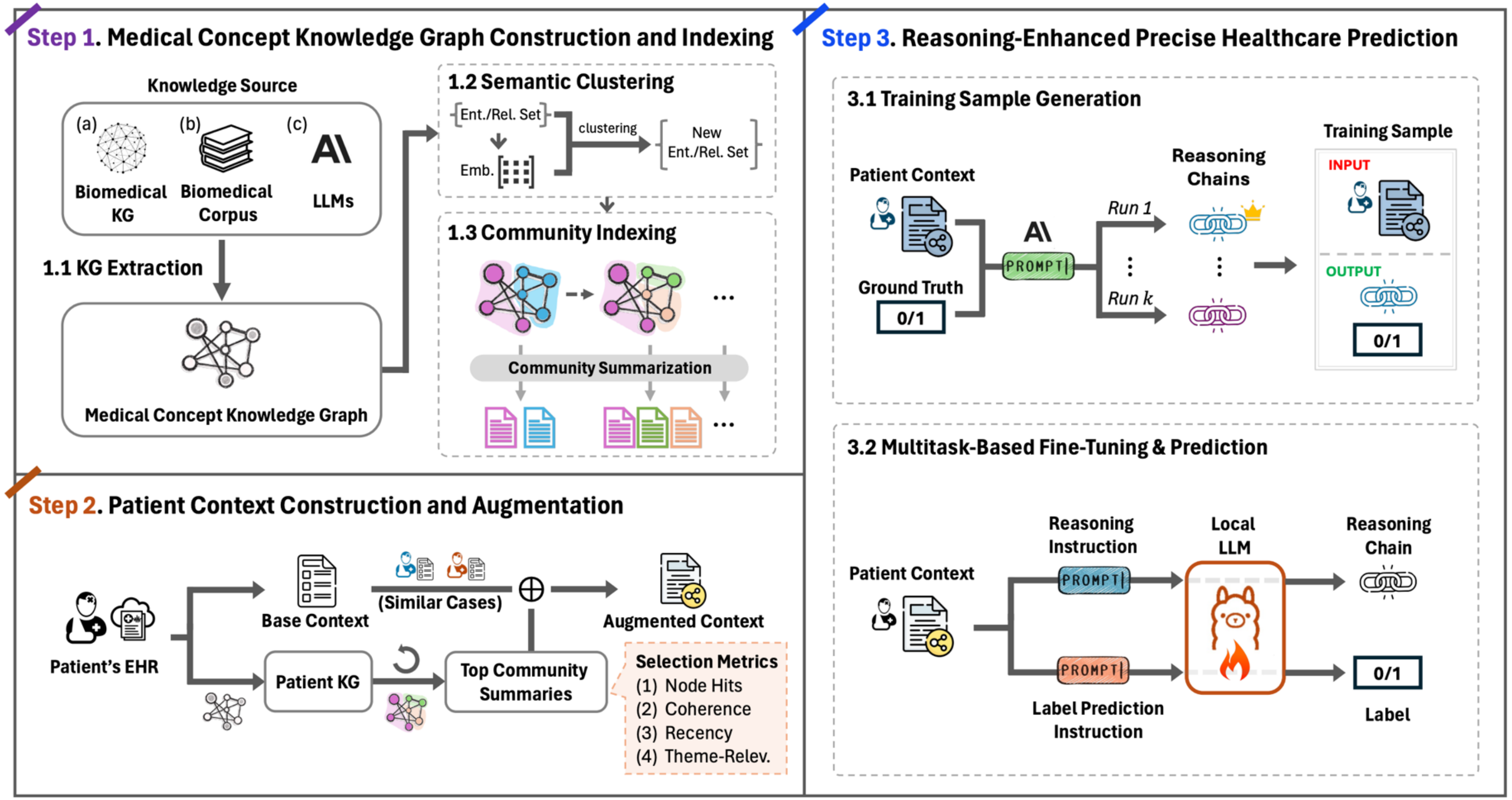
Summarization of Limitations

Methods	Limitations
ML-based 	1. Lack interpretability; Limited ability to handle unstructured or complex medical information 2. Struggle to generalize beyond their original scope
LM+ML -based 	1. Lack interpretability; Not exploit reasoning capabilities of LMs 2. Retrieved information may be sparse or irrelevant
LLM-based (Untrained) 	1. Lack of specialized medical knowledge 2. Cannot learn underlying EHR pattern 3. Performance highly dependent on quality of prompting and context

*"Can we design a framework integrating **high-utility knowledge retrieval** and **reliable reasoning** for interpretable & precise prediction?"*

Methodology

KARE – **K**nowledge **A**ware **R**easoning-**E**nhanced Health<u>Care</u> Prediction



Methodology – KARE

*"Can we design a framework integrating **high-utility knowledge retrieval** and **reliable reasoning** for interpretable & precise prediction?"*

Challenge 1: How to enable high-utility knowledge retrieval?

Classic RAG?

EHR example

Patient Context:
Patient ID: 29488

Visit 0:
Conditions:
- Deficiency and other anemia
- Essential hypertension
- Complication of device; implant or graft
- Congestive heart failure; nonhypertensive - Cancer of prostate
- Anxiety disorders
- Thyroid disorders
- Disorders of lipid metabolism
- Conduction disorders
- Mycoses
- Other diseases of kidney and ureters
- Cancer of esophagus

Procedures:
- Diagnostic cardiac catheterization; coronary arteriography
- Other or procedures on vessels other than head and neck
- Colorectal resection

Drugs:
- Other drugs for obstructive airway diseases, inhalants in ATC
- Lipid modifying agents, plain
- Antithrombotic agents
- Angiotensin II receptor blockers (ARBs), plain

Visit 1:
Conditions:
- Congestive heart failure; nonhypertensive
- Cardiac dysrhythmias
- Shock
- Alcohol-related disorders
- Diabetes mellitus with complications
- E codes: adverse effects of medical care
- Disorders of teeth and jaw

Procedures:
- Other vascular catheterization; not heart
- Blood transfusion Drugs:
- Other drugs for obstructive airway diseases, inhalants in ATC
- Other analgesics and antipyretics in ATC
- Drugs for peptic ulcer and gastro-oesophageal reflux disease (GORD)
- Beta blocking agents
- Potassium supplements

Retrieval Result from PubMed

- Rank 1, Similarity: 0.7204270362854004 PMID: 314160 Title: Resection of posterior ventricular aneurysm, replacement of mitral valve, and coronary bypass. Authors: C M Chavez, G Rodriquez, J H Conn Abstract: A 58-year-old man had sudden and progressive heart failure after a severe myocardial infarction. Aggressive medical treatment consisting of diuretics, vasopressors, and digitalis failed to improve his condition significantly. Cardiac catheterization disclosed a critical stenosis in the left anterior descending branch of the left coronary artery, a large posterior left ventricul aneurysm, and severe mitral insufficiency. Intermittent third degree heart block developed after admission. Surgical correction resulted in a dramatic recovery, and three years after operation he is fully recovered and asymptomatic. Publication Year: 1979
- Rank 2, Similarity: 0.7123403549194336 PMID: 804881 Title: [Acute coronary thrombosis in a 28 year-old woman]. Authors: G Motté, M Vogel, G Coatantiec, P Mariette Abstract: A 28-year-old woman, with no previous cardiovascular history, was hospitalized for myocardial infarction complicated by bifascicular block followed by complete atrio-ventricular block with a regressive course. A coronary arteriography performed on the 10th day demonstrated a thrombosis of the anterior interventricular artery, the rest of the coronary network being normal. The influence of a dyslipidaemia and the taking of oral contraceptives was discussed as an aetiology. Publication Year: 1975

Unwanted Information

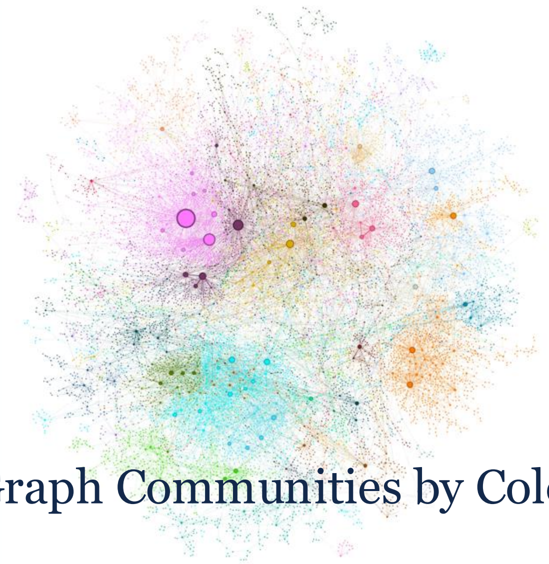
Methodology – KARE

*“Can we design a framework integrating **high-utility knowledge retrieval** and **reliable reasoning** for interpretable & precise prediction?”*

Challenge 1: How to enable high-utility knowledge retrieval? → We need **high-quality knowledge source**!

Information Re-indexing: “Text → Graph → Text” by *GraphRAG**

Corpus → (Relation Extraction) → Knowledge Graph → (Community Detection) →



→ (Community Summarization) → New Corpus

(Graph Communities by Colors)

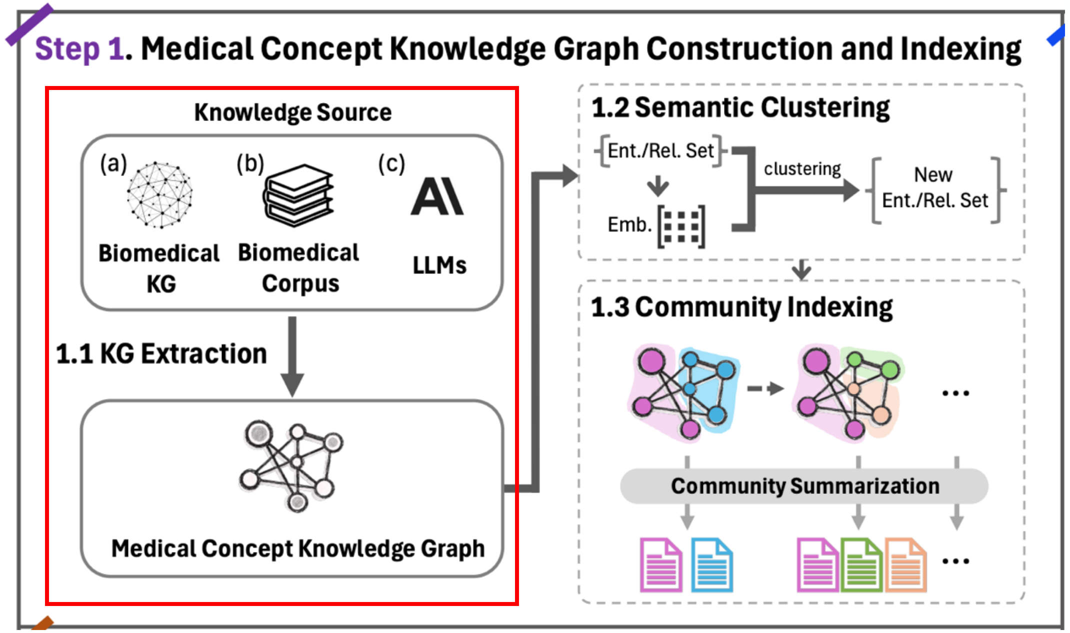
*Edge, Darren, et al. "From local to global: A graph rag approach to query-focused summarization." *arXiv preprint arXiv:2404.16130* (2024).

Methodology – KAREI

*"Can we design a framework integrating **high-utility knowledge retrieval** and **reliable reasoning** for interpretable & precise prediction?"*

Challenge 1: How to enable high-utility knowledge retrieval? → We need **high-quality knowledge source**!

Information Re-indexing: "Text → Graph → Text"



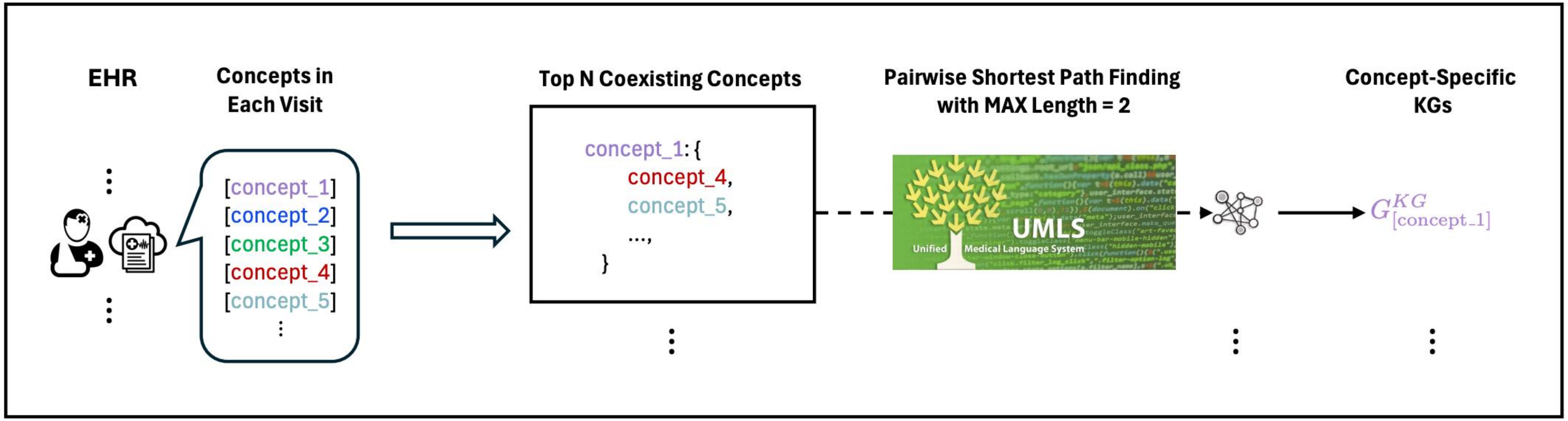
However, we don't need all the information in the available data sources!

We need information tailored to the EHR data.

→ Construct the KG based on the co-existence of medical concepts in EHR

Methodology – KARE

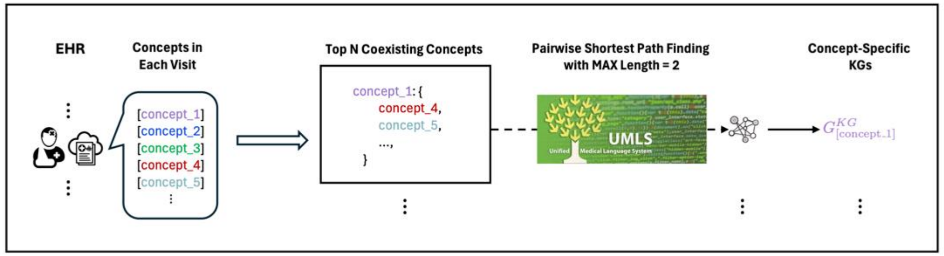
1.1.1 EHR Concpet-specific KG Construction from Large Bio KG (UMLS)



Construction Pipeline

Methodology

1.1.1 EHR Concpet-specific KG Construction from Large Bio KG (UMLS)



Construction Pipeline

(Top-20 Co-existing Concepts for Each Concept)

⋮

```
"spinal fusion": [
        "opioid analgesics",
        "other beta-lactam antibacterials
        in atc",
        "antithrombotic agents",
        "potassium supplements",
        "drugs for constipation",
        "other analgesics and
        antipyretics in atc",
        "i.v. solution additives",
        "drugs for peptic ulcer and
        gastro-oesophageal reflux disease
        (gord)",
        "iv solutions used in parenteral
        administration of fluids,
        electrolytes and nutrients",
        "spondylosis; intervertebral disc
        disorders; other back problems",
        "antiemetics and antinauseants",
        "anxiolytics",
        "other mineral supplements in
        atc",
        "antiinfectives and antiseptics,
        excl. combinations with
        corticosteroids",
        "antiepileptics",
        "other nutrients in atc",
        "antivaricose therapy drugs",
        "beta blocking agents",
        "lipid modifying agents, plain",
        "decompression peripheral nerve"
]
```

⋮

(Alg. 2)

==

Target Concept: opioid analgesics

Relationship Triples:
- [aspects of adverse effects, Allowed qualifier, opioid analgesics]
- [spinal fusion, Allowed qualifier, aspects of adverse effects]

Connection Paths:
- spinal fusion → (Allowed qualifier) → aspects of adverse effects → (Allowed qualifier) → opioid analgesics

==

Target Concept: other beta-lactam antibacterials in atc

Relationship Triples:
- [aspects of adverse effects, Allowed qualifier, Carbapenem-containing product]
- [Carbapenem-containing product, isa, other beta-lactam antibacterials in atc]
- [spinal fusion, Allowed qualifier, aspects of adverse effects]

Connection Paths:
- spinal fusion → (Allowed qualifier) → aspects of adverse effects → (Allowed qualifier) → Carbapenem-containing product → (isa) → other beta-lactam antibacterials in atc

==

Target Concept: antithrombotic agents

Relationship Triples:
- [aspects of adverse effects, Allowed qualifier, Fibrinolytic Agents]
- [Fibrinolytic Agents, has relationship, antithrombotic agents]
- [spinal fusion, Allowed qualifier, aspects of adverse effects]

Connection Paths:
- spinal fusion → (Allowed qualifier) → aspects of adverse effects → (Allowed qualifier) → Fibrinolytic Agents → (has relationship) → antithrombotic agents

==

Target Concept: potassium supplements

Relationship Triples:
- [POTASSIUM CITRATE, member_of, potassium supplements]
- [POTASSIUM CHLORIDE, member_of, potassium supplements]
- [aspects of adverse effects, can be qualified by., POTASSIUM CITRATE]
- [aspects of adverse effects, Allowed qualifier, POTASSIUM CHLORIDE]
- [spinal fusion, Allowed qualifier, aspects of adverse effects]

Connection Paths:
- spinal fusion → (Allowed qualifier) → aspects of adverse effects → (Allowed qualifier) → POTASSIUM CHLORIDE → (member_of) → potassium supplements
- spinal fusion → (Allowed qualifier) → aspects of adverse effects → (can be qualified by) → POTASSIUM CITRATE → (member_of) → potassium supplements

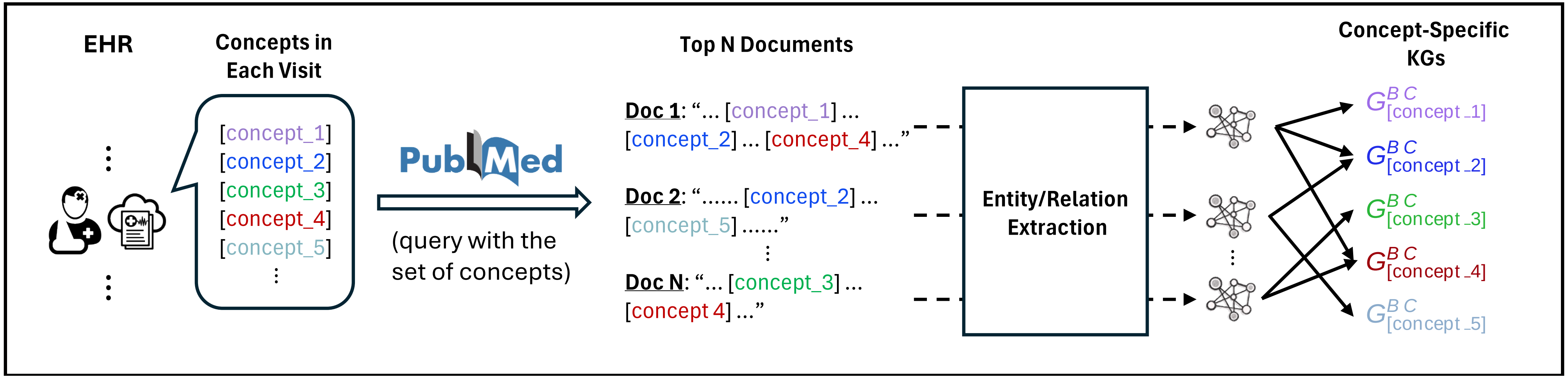
==

⋮

Example

Methodology – KARE

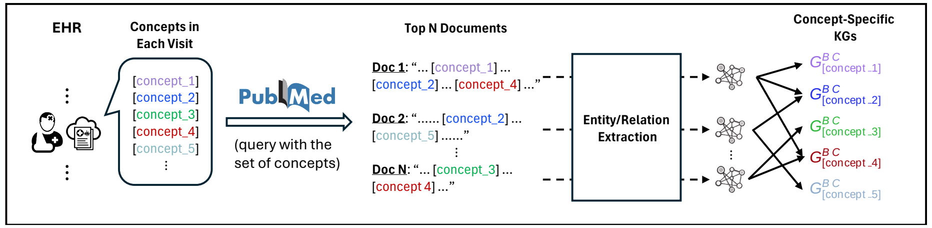
1.1.2 EHR Concpet-specific KG Construction from Large Bio Corpus (PubMed)



Construction Pipeline

Methodology – KARE

1.1.2 EHR Concpet-specific KG Construction from Large Bio Corpus (PubMed)



Construction Pipeline

Example

(Concept set in a patient visit)

```
[
"pleurisy; pneumothorax; pulmonary collapse",
"coronary atherosclerosis and other heart disease",
"potassium supplements",
"other fractures",
"anxiolytics",
"opioid analgesics",
"other gastrointestinal disorders",
"epilepsy; convulsions",
"i.v. solution additives",
"antibiotics for topical use",
"e codes: fall",
"adrenergics, inhalants",
"antiepileptics",
"other diagnostic agents in atc",
"diabetes mellitus without complication",
"drugs for peptic ulcer and gastro-oesophageal reflux disease (gord)",
"other mineral supplements in atc",
"spinal fusion",
"beta blocking agents",
"other analgesics and antipyretics in atc",
"laminectomy; excision intervertebral disc",
"intracranial injury",
"other beta-lactam antibacterials in atc"
]
```

Doc. Retrieval

(Retrieved Top-10 PubMed Abstracts)

Rank 1, Similarity: 0.7296074032783508
PMID: 310377
Title: Management of acute pain in trauma.
Authors: C N Shealy
Abstract: In acute pain, TENS, ice packs, and a calm, reassuring attitude and voice are useful in reducing pain. Narcotic requirements can be reduced, and chronic pain may be prevented. Complications such as paralytic ileus and atelactasis can also be reduced. These techniques can be used in the emergency ward, the recovery room, and the doctor's office.
Publication Year: 1979

Rank 2, Similarity: 0.7247765064239502
PMID: 235156
Title: Interaction between local anesthetics and analeptic drugs.
Authors: R G Thompson, J A Aldrete
Abstract: Although clinically undesirable, the fortuitous pharmacologic interactions between local anesthetic agents and analeptic drugs may be protective when large doses of both agents are used. Mice pretreated with procaine, lidocaine, and tetracaine had a lower incidence of seizures when convulsive doses of either nikethamide or doxapram hydrochloride were given intraperitoneally. Mortality was also decreased in the groups given nikethamide and was zero in the animals treated with doxapram. All animals treated with pentylenetetrazol convulsed and only lidocaine (also used as an anticonvulsant) was able to reduce mortality in this group. Of the local anesthetic agents, tetracaine afforded the least protection from death, whereas lidocaine seemed to be most effective.
Publication Year: 1975

⋮

Triple Extraction

(Triples)

[opioid analgesics, can be reduced by, TENS]

[opioid analgesics, can be reduced by, ice packs]

Triple Extraction

[epilepsy; convulsions, can be caused by, large doses of analeptic drugs]

[local anesthetic agents, may reduce, epilepsy; convulsions]

[lidocaine, used as, antiepileptics]

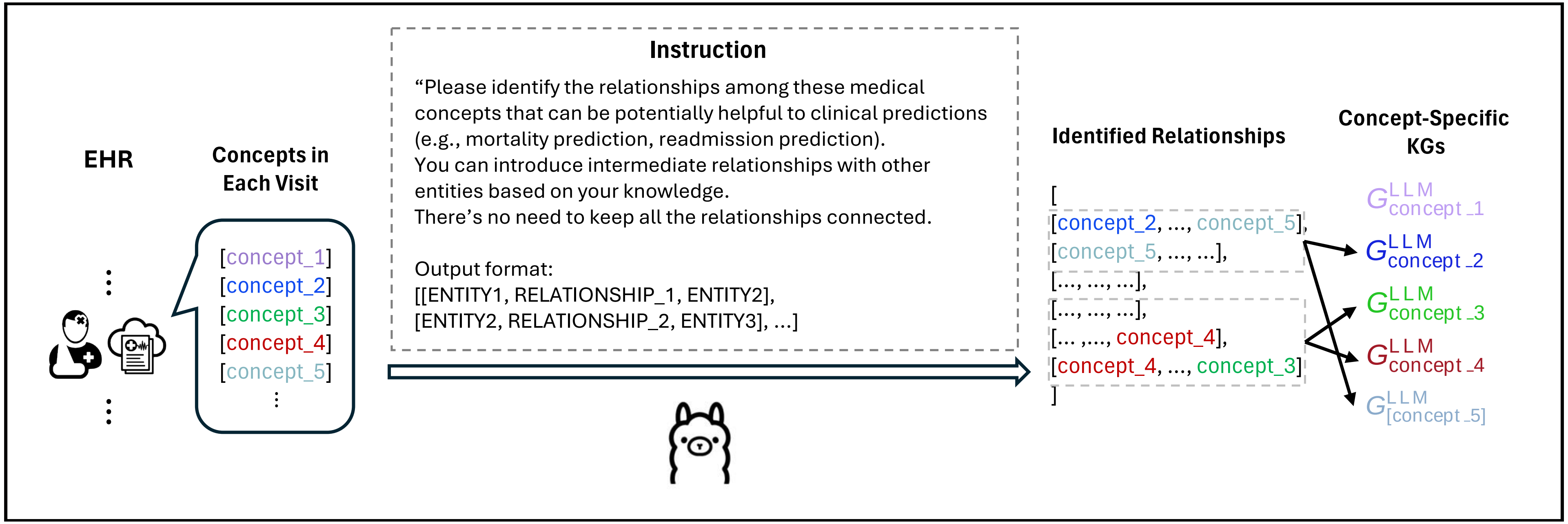
[lidocaine, most effective in reducing, mortality]

[pentylenetetrazol, causes, epilepsy; convulsions]

[lidocaine, reduces mortality from, epilepsy; convulsions]

Methodology – KARE

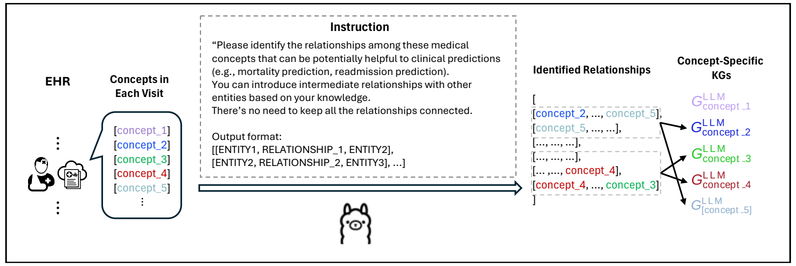
1.1.3 EHR Concpet-specific KG Construction from Large Language Model



Construction Pipeline

Methodology – KARE

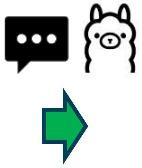
1.1.3 EHR Concpet-specific KG Construction from Large Language Model



Construction Pipeline

(Concept set in a patient visit)

```
[
"pleurisy; pneumothorax; pulmonary
collapse",
"coronary atherosclerosis and other
heart disease",
"potassium supplements",
"other fractures",
"anxiolytics",
"opioid analgesics",
"other gastrointestinal disorders",
"epilepsy; convulsions",
"i.v. solution additives",
"antibiotics for topical use",
"e codes: fall",
"adrenergics, inhalants",
"antiepileptics",
"other diagnostic agents in atc",
"diabetes mellitus without
complication",
"drugs for peptic ulcer and gastro-
oesophageal reflux disease (gord)",
"other mineral supplements in atc",
"spinal fusion",
"beta blocking agents",
"other analgesics and antipyretics
in atc",
"laminectomy; excision
intervertebral disc",
"intracranial injury",
"other beta-lactam antibacterials
in atc"
]
```



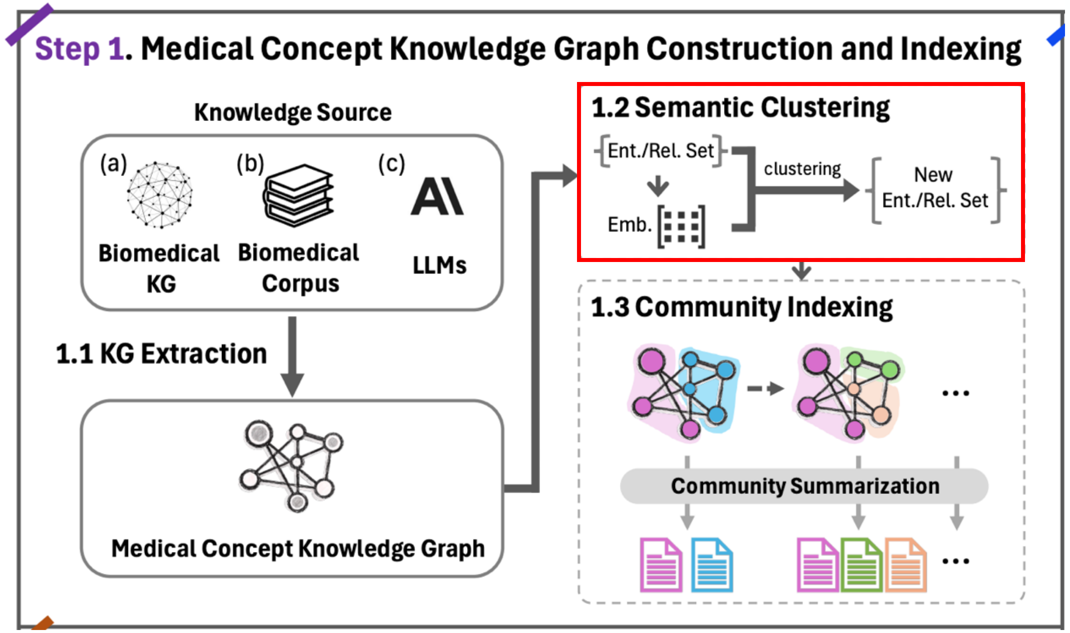
Triple Extraction

(Triples)

[pleurisy; pneumothorax; pulmonary collapse, can be treated with, adrenergics, inhalants],
[diabetes mellitus without complication, may require monitoring of, potassium supplements],
[epilepsy; convulsions, treated with, antiepileptics],
[epilepsy; convulsions, may be treated with, anxiolytics],
[other gastrointestinal disorders, treated with, drugs for peptic ulcer and gastro-oesophageal reflux disease (gord)],
[e codes: fall, commonly causes, other fractures],
[e codes: fall, can result in, intracranial injury],
[other fractures, often treated with, opioid analgesics],
[intracranial injury, may be managed with, opioid analgesics],
[spinal fusion, surgical alternative to, laminectomy; excision intervertebral disc],
[coronary atherosclerosis and other heart disease, commonly treated with, beta blocking agents],
[other beta-lactam antibacterials in atc, belongs to same class as, antibiotics for topical use]

Example

Methodology – KARE



Step 1.2 Semantic Clustering

The same entity/relation from different knowledge sources may have different names.

- We apply embedding clustering to form new entity/relation set in the concept KG

Entity clustering example

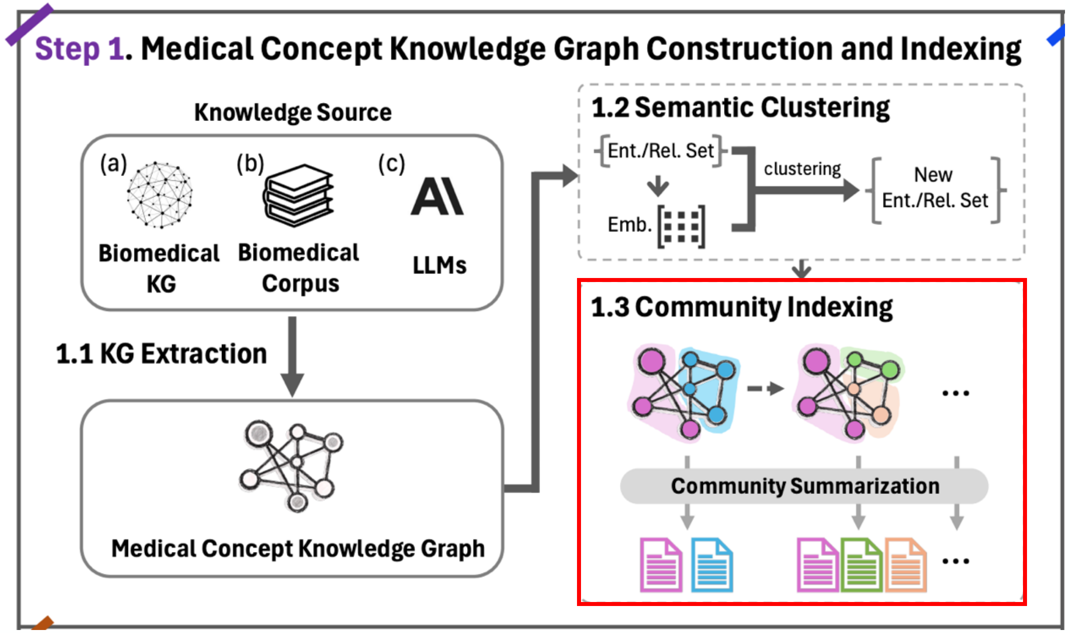
"continuation of treatment": [
"continuation of treatment",
"continued treatment",
"continuation of treatments"
],

"underlying heart condition": [
"underlying cardiac condition",
"underlying heart conditions",
"underlying heart disease",
"underlying cardiac disease",
"underlying heart condition"
],
⋮

Relation clustering example

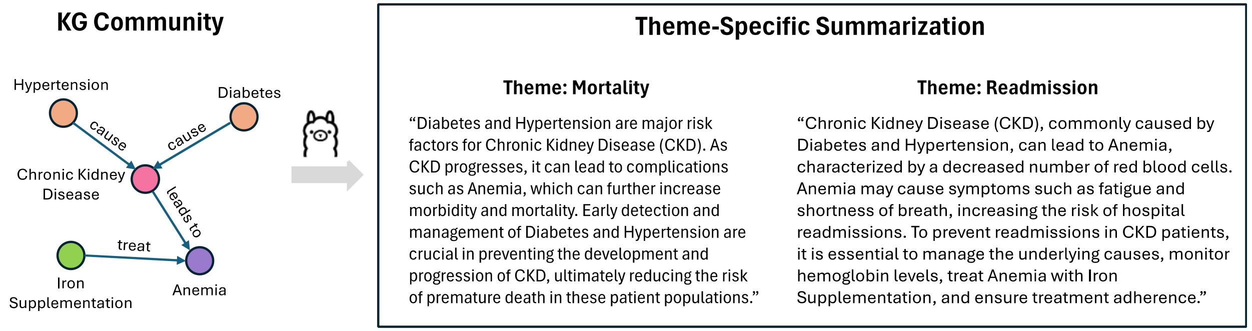
"does not significantly impact": [
"not significantly impacts",
"does not substantially impact",
"does not significantly impact",
"do not substantially impact",
"do not significantly impact",
"not significantly impacting",
"not significantly impact",
"doesn't significantly impact",
"don't significantly impact"
],
⋮

Methodology – KARE



Step 1.3 KG Community Indexing

- Use Leiden* to partition the concept KG into communities, with n runs to get a diverse set
- Prompt an LLM to summarize each community regarding specific themes.



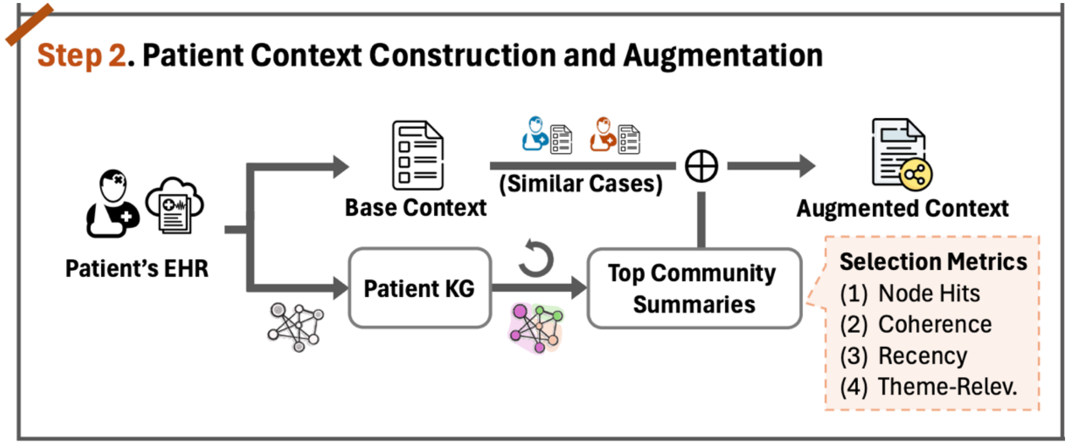
*Traag, Vincent A., Ludo Waltman, and Nees Jan Van Eck. "From Louvain to Leiden: guaranteeing well-connected communities." *Scientific reports* 9.1 (2019): 1-12.

Methodology – KARE

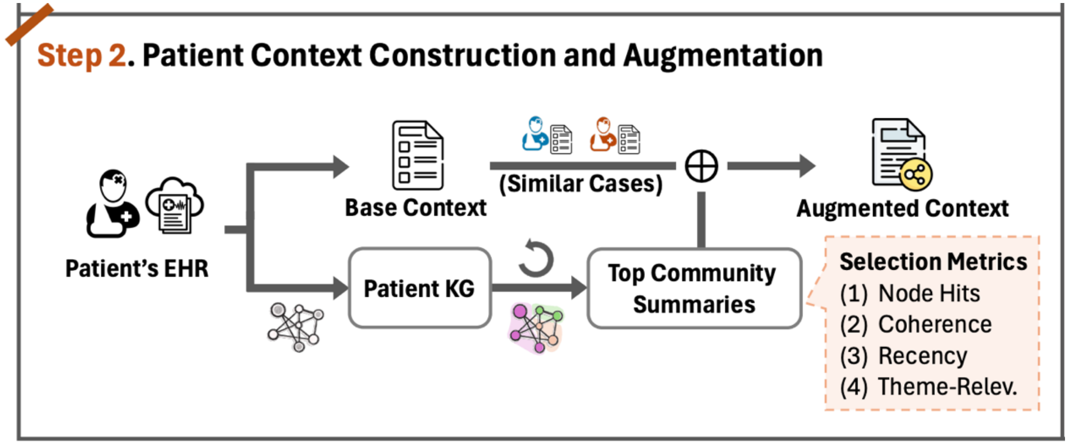
*"Can we design a framework integrating **high-utility knowledge retrieval** and **reliable reasoning** for interpretable & precise prediction?"*

Challenge 1: How to enable high-utility knowledge retrieval? → We need **high-quality knowledge source**!

"Now we have high-quality knowledge source, what's next?" → We need to retrieve knowledge with **high coverage** of patient's EHR, **low repetition** …



Methodology – KARE

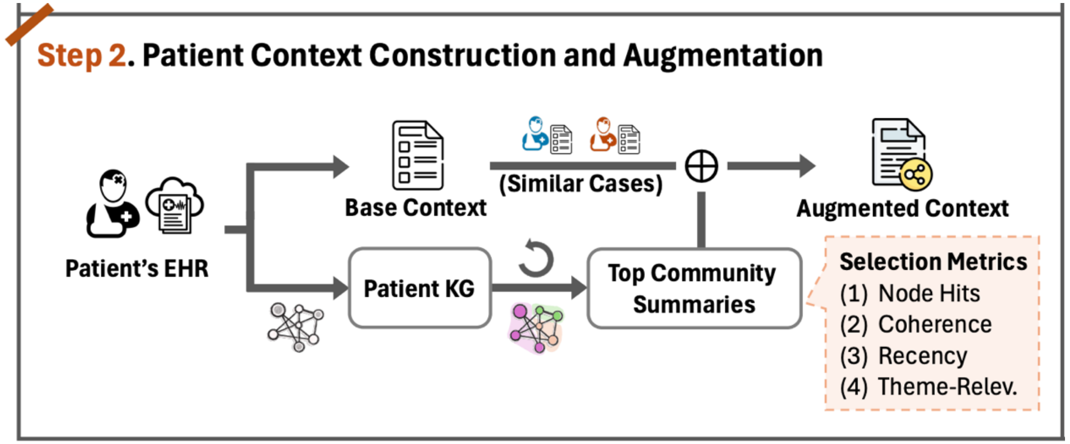


Step 2.1 Patient Base Context Construction

Example:

```
Patient ID: 10088

Visit 0:
Conditions:
1. Septicemia
2. Shock
3. Urinary tract infections

Procedures:
1. Enteral and parenteral nutrition
2. Blood transfusion

Medications:
1. Beta blocking agents
2. Antithrombotic agents

Visit 1:
Conditions:
1. Septicemia (continued from previous visit)
2. Acute myocardial infarction (new)
3. Respiratory failure (new)

Procedures:
1. Respiratory intubation
2. Mechanical ventilation

Medications:
1. Antithrombotic agents (continued from previous visit)
2. Beta blocking agents (continued from previous visit)


Similar Patients:

Patient ID …
…
Label: 1

Patient ID: …
…
Label: 0
```

Methodology – KARE



Step 2.2 Patient Context Augmentation

- We dynamically retrieve knowledge (community summaries) that maximize the relevance score considering **node hits, coherence, recency,** and **theme relevance.**

$$\text{Relevance}(C_k) = (\mathcal{H}_{\text{direct}}(C_k, G_p) + \alpha \cdot \mathcal{H}_{\text{indirect}}(C_k, G_p)) \times \text{Decay}(C_k, G_p) \times \text{Coherence}(S_{C_k}, \mathcal{B}_p) \times \text{Recency}(C_k, G_p) \times \text{ThemeRel}_\tau(C_k)$$

Algorithm 1 Dynamic Graph Retrieval and Augmentation

Input: Set of communities $\mathcal{C}$, patient graph G_p, base context $\mathcal{B}_p$, desired number of summaries N
Output: Augmented patient context $\mathcal{A}_p$
Initialize $S_p \leftarrow \emptyset$
Initialize hit counts $H(v) \leftarrow 0$ for each node $v \in V_{G_p}$
while $|S_p| < N$ **do**
  Compute Relevance(C_k) for all $C_k \in \mathcal{C}$ using Eq. 3
  Select $C_{\text{best}} \leftarrow \arg\max_{C_k \in \mathcal{C}} \text{Relevance}(C_k)$
  Add $S_{C_{\text{best}}}$ to S_p: $S_p \leftarrow S_p \cup \{S_{C_{\text{best}}}\}$
  For each $v \in V_{C_{\text{best}}}$, $H(v) \leftarrow H(v) + 1$
  Remove C_{best} from $\mathcal{C}$: $\mathcal{C} \leftarrow \mathcal{C} \setminus \{C_{\text{best}}\}$
end
Augment patient context: $\mathcal{A}_p = \mathcal{B}_p \oplus \left(\bigcup_{S_{C_k} \in S_p} S_{C_k}\right)$
return $\mathcal{A}_p$

Methodology – KARE

*“Can we design a framework integrating **high-utility knowledge retrieval** and **reliable reasoning** for interpretable & precise prediction?”*

Challenge 1: How to enable high-utility knowledge retrieval?

→ We need **high-quality knowledge source**!

→ We need to retrieve knowledge with **high coverage** of patient’s EHR, **low repetition** …

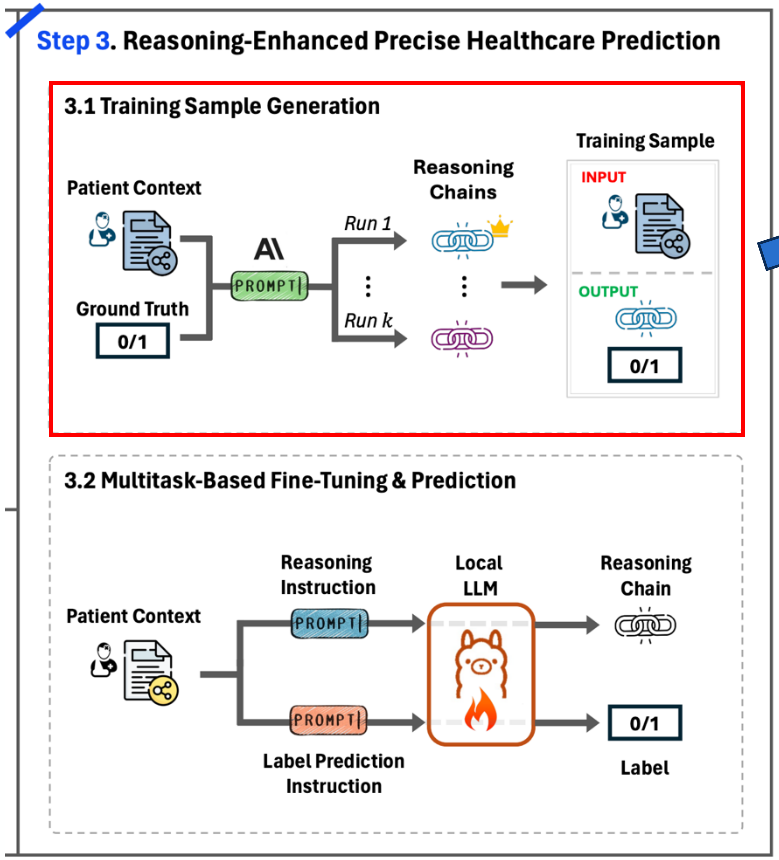
Challenge 2: How to enable reliable reasoning by LLM?

→ Need to fine-tune an LLM with EHR data

→ Easier to learn if the reasoning chain (rationale) is in a uniform format.

Methodology – KARE

Challenge 2: How to enable reliable reasoning by LLM?



Given the following task description, patient EHR context, similar patients, retrieved medical knowledge, and ground truth label, provide a step-by-step reasoning process that leads to the correct prediction:

==
Task
{task_description}
==
Patient EHR Context
{context}
==
Similar Patients
{similar_patients}
==
Retrieved Medical Knowledge
{medical_knowledge}
==
Ground Truth
{ground_truth}
==

→ Augmented patient context

Please provide a step-by-step reasoning process that leads to the correct prediction based on the patient's context, similar patients, and the retrieved relevant medical knowledge.

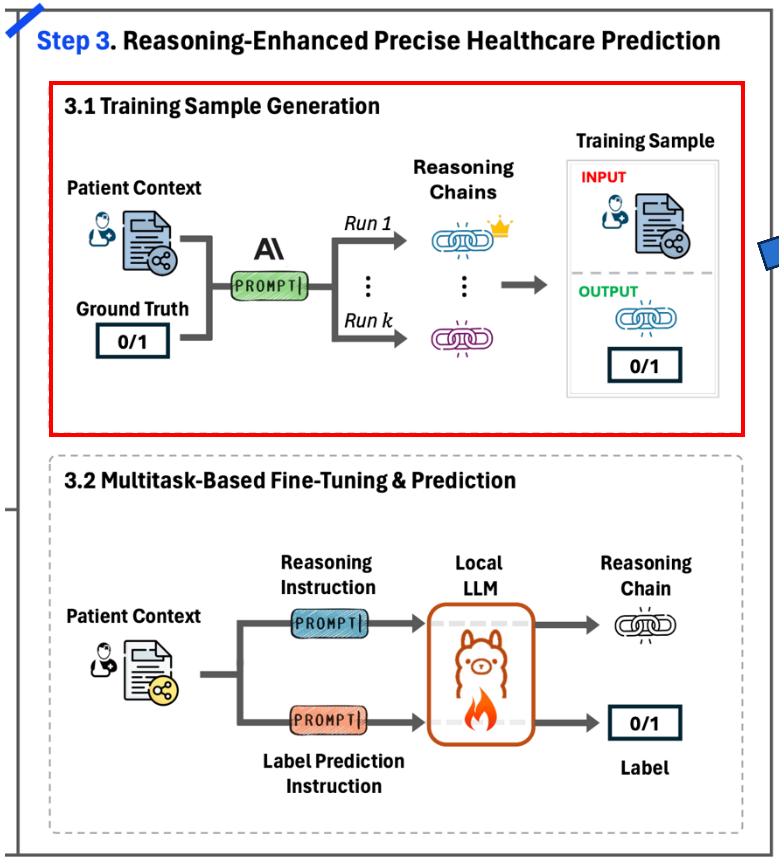
The reasoning chain should follow this structured format:

1. **Patient Overview**: Check the key information in the patient's context, with the Key Considerations from the task description in mind.
2. **Relevant Retrieved Medical Knowledge**: Highlight the retrieved medical knowledge pertinent to the patient's condition.
3. **Comparison with Similar Patients**: Analyze the similarities and differences between the patient and similar patients, explaining how these factors influence the prediction.
4. **Reasoning Towards Prediction**: Integrate the above information to logically reason towards the predicted outcome.
5. **Conclusion**: Summarize the reasoning and state the prediction without mentioning the ground truth.

The reasoning should be comprehensive, medically sound, and clearly explain how the patient's information leads to the predicted outcome.

Methodology – KARE

Challenge 2: How to enable reliable reasoning by LLM?



Important Notes:
- Do not mention the ground truth label in the reasoning process.
- Use the relevant knowledge as needed.
- Analyze the similarities and differences between the patient and similar patients to justify the prediction.

After generating the reasoning chain, please review it and indicate your confidence in the reasoning chain at the end.

Options of confidence: [Very Confident, Confident, Neutral, Not Confident, Very Not Confident.]

Output Format:

Reasoning Chain

1. Patient Overview:
[YOUR OUTPUT]

2. Relevant Retrieved Medical Knowledge:
[YOUR OUTPUT]

3. Comparison with Similar Patients:
[YOUR OUTPUT]

4. Reasoning Towards Prediction:
[YOUR OUTPUT]

5. Conclusion:
[YOUR OUTPUT]

Confidence
[CONFIDENCE ("Very Confident", "Confident", "Neutral", "Not Confident", "Very Not Confident")]

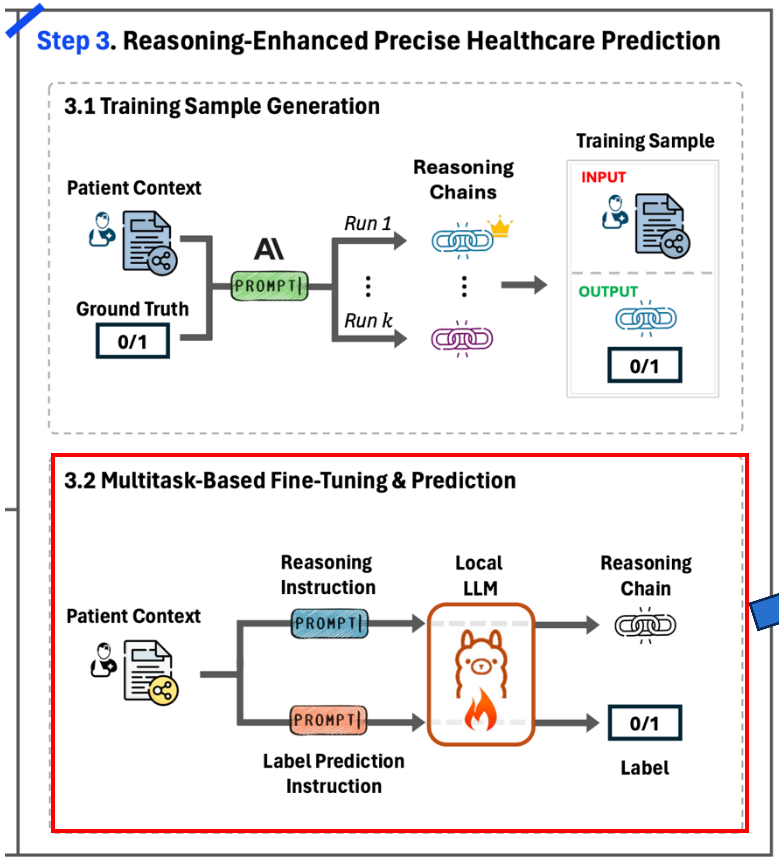
→ Uniform-formatted output

Methodology – KARE

Table 4: Comparison of two strategies for fine-tuning LLM with reasoning chain and label.

	MIMIC-IV-Mortality		MIMIC-IV-Readmission	
Strategy	Macro F1	Sensitivity	Accuracy	Macro F1
Multitask	90.4	73.2	73.9	73.8
“Two-In-One”	86.5	68.0	67.2	65.4

Challenge 2: How to enable reliable reasoning by LLM?



Templates used for fine-tuning

Fine-Tuning Input (Reasoning)

[Reasoning] Given the following task description, patient EHR context, similar patients, and retrieved medical knowledge, please provide a step-by-step reasoning process that leads to the prediction outcome based on the patient's context and relevant medical knowledge. After the reasoning process, provide the prediction label (0/1).

==
 # Task #
{task_description}
==
 # Patient EHR Context #
{context}
==
 # Similar Patients #
{similar_patients}
==
 # Retrieved Medical Knowledge #
{medical_knowledge}
==

Fine-Tuning Input (Label Prediction)

[Label Prediction] Given the following task description, patient EHR context, similar patients, and retrieved medical knowledge, please directly predict the label (0/1)

==
 # Task #
{task_description}
==
 # Patient EHR Context #
{context}
==
 # Similar Patients #
{similar_patients}
==
 # Retrieved Medical Knowledge #
{medical_knowledge}
==

Fine-Tuning Output (Reasoning)

 # Reasoning #
{reasoning}

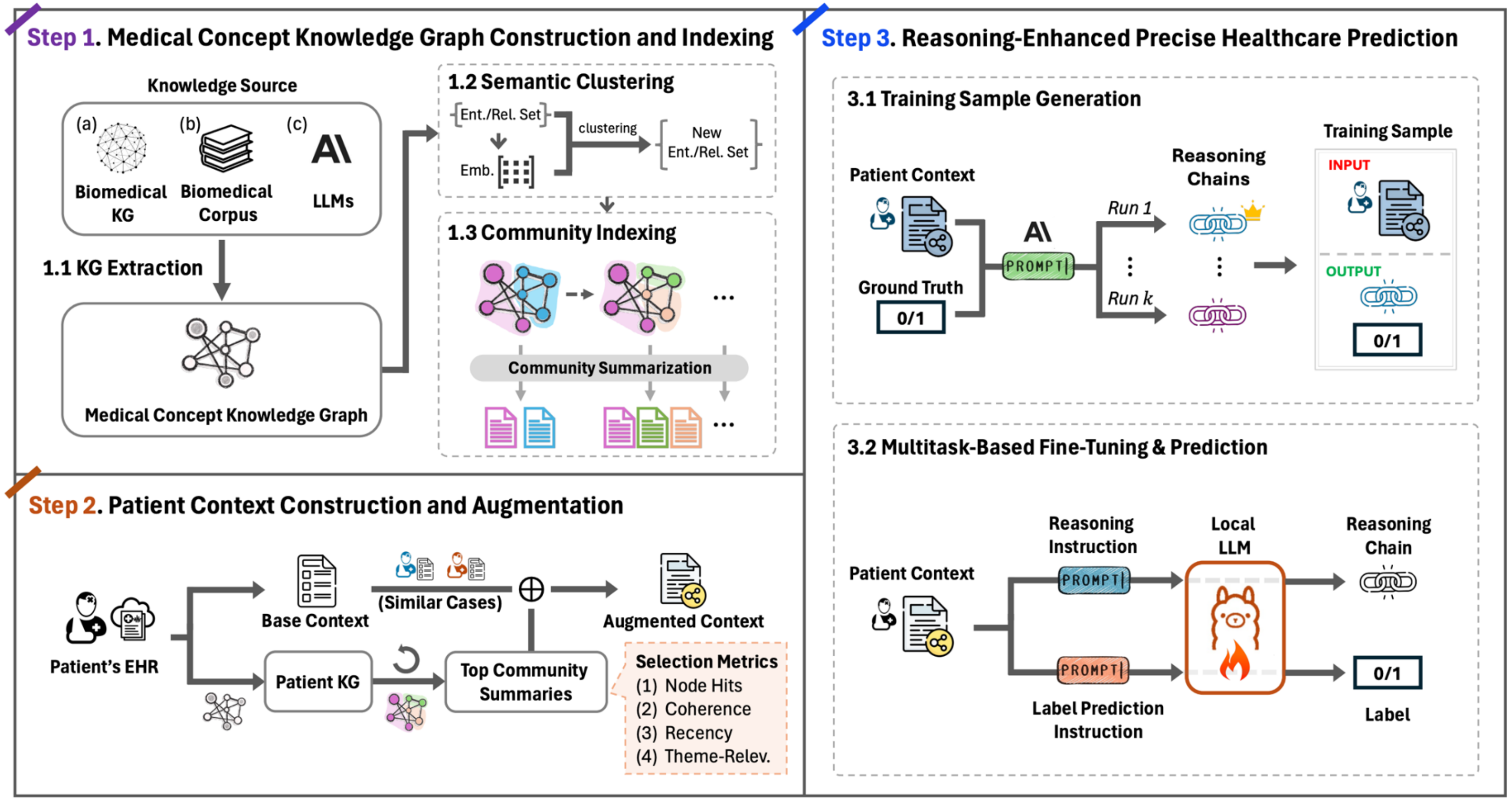
 # Prediction #
{Label (0/1)}

Fine-Tuning Output (Label Prediction)

{Label (0/1)}

Methodology – KARE

KARE – **K**nowledge **A**ware **R**easoning-**E**nhanced Health<u>Care</u> Prediction



Experiments

Datasets: MIMIC-III and MIMIC-IV

Data Descriptor | Open access | Published: 24 May 2016

MIMIC-III, a freely accessible critical care database

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi & Roger G. Mark

Scientific Data **3**, Article number: 160035 (2016) | Cite this article

Data Descriptor | Open access | Published: 03 January 2023

MIMIC-IV, a freely accessible electronic health record dataset

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi & Roger G. Mark

Scientific Data **10**, Article number: 1 (2023) | Cite this article

Table 1: Statistics of pre-processed EHR datasets. "#": "the number of", "/ patient": "per patient".

	MIMIC-III-Mort.			MIMIC-III-Read.			MIMIC-IV-Mort.			MIMIC-IV-Read.		
	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test	Train	Valid	Test
# Patients (Samples)	7730	991	996	7730	991	996	8018	996	986	8029	958	1013
# Visits / Patient	1.56	1.60	1.61	1.56	1.60	1.61	1.26	1.30	1.21	1.26	1.28	1.25
# Conditions / Patient	23.27	23.92	25.89	23.27	23.92	25.89	14.34	15.30	13.59	13.62	14.21	13.21
# Procedures / Patient	6.22	6.56	7.17	6.22	6.56	7.17	2.96	3.08	2.84	2.89	2.96	2.81
# Medications / Patient	54.79	55.77	63.73	54.79	55.77	63.73	30.66	32.86	28.40	28.74	30.61	27.59

Experiments

Tasks: (1) Mortality Prediction, (2) Readmission Prediction. Both are binary classification task.

- *Mortality Prediction.* This task estimates mortality outcome for next visit, defined as $f : (x_1, x_2, \ldots, x_{t-1}) \rightarrow y[x_t]$, where $y[x_t] \in \{0, 1\}$ is patient's survival status during visit x_t.

"Will this patient die in the next visit?"

- *Readmission Prediction.* This task predicts if patient will be readmitted within σ days, defined as $f : (x_1, x_2, \ldots, x_{t-1}) \rightarrow y[\varphi(x_t) - \varphi(x_{t-1})]$, where $y \in \{0, 1\}$, $\varphi(x_t)$ is timestamp of visit x_t, and $y[\varphi(x_t) - \varphi(x_{t-1})] = 1$ if $\varphi(x_t) - \varphi(x_{t-1}) \leq \sigma$, else 0. σ is set to 15 in this study.

"Will this patient be readmitted within 15 days?"

Experiments

Metrics: (1) Accuracy, (2) Macro-F1, (3) Sensitivity, (4) Specificity

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Accuracy measures the overall proportion of correct predictions

$$\text{Macro-F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Macro-F1 provides a balanced measure that is particularly important for imbalanced datasets

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Sensitivity quantifies the model's ability to correctly identify high-risk pateints (e.g., correctly predict the patient who will die in the next visit)

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

Specificity quantifies the model's ability to correctly identify low-risk pateints (e.g., correctly predict the patient who will not die in the next visit)

Experiments

Main Results (MIMIC-III)

1. For mortality prediction where the data is extremely imbalanced (5.42% positive labels), most ML models performed poor
2. LM+ML based methods improved the performance by leveraing external knowledge
3. Zero-shot, few-shot, and backbone-finetuned LLM-based methods perform worse than traditional ML methods in most cases
4. Classic RAG can even downgrade the performance in zero-shot setting
5. KARE significantly outperforms all the previous methods

		MIMIC-III: Mortality Prediction (pos = 5.42%)				MIMIC-III: Readmission Prediction (pos = 54.82%)			
Type	Models	Accuracy	Macro F1*	Sensitivity*	Specificity	Accuracy	Macro F1	Sensitivity	Specificity
ML	GRU (Chung et al., 2014)	92.7	50.7	3.7	97.8	62.2	61.5	68.9	54.0
	Transformer (Vaswani et al., 2017)	92.7	51.9	5.6	97.6	58.8	58.2	65.0	51.3
	RETAIN (Choi et al., 2016)	92.4	50.6	3.7	97.6	59.1	56.9	74.9	40.0
	GRAM (Choi et al., 2017)	92.4	50.2	5.2	95.2	61.8	60.4	74.9	46.4
	Deepr (Nguyen et al., 2016)	91.9	51.0	3.7	98.2	62.6	62.1	66.7	57.6
	TCN (Bai et al., 2018)	91.6	53.2	9.3	96.4	63.4	62.7	70.7	54.7
	ConCare (Ma et al., 2020b)	94.6	48.6	0.0	**100.0**	59.2	59.0	61.5	56.4
	AdaCare (Ma et al., 2020a)	90.6	54.1	9.1	97.6	61.6	60.5	70.8	50.3
	GRASP (Zhang et al., 2021)	93.7	49.9	1.9	98.9	61.3	59.5	74.9	44.8
	StageNet (Gao et al., 2020)	90.5	50.5	5.6	95.4	60.5	60.0	65.1	54.9
	KerPrint (Yang et al., 2023b)	92.4	52.2	9.8	94.7	63.5	62.1	68.0	56.1
LM+ML	MedRetriever (Ye et al., 2021)	93.2	53.3	11.3	95.2	63.2	62.7	66.3	59.1
	GraphCare (Jiang et al., 2024a)	94.9	58.3	17.2	97.1	65.4	64.1	70.3	57.8
	RAM-EHR (Xu et al., 2024)	94.4	59.6	14.8	98.9	64.8	63.5	74.7	52.4
	EMERGE (Zhu et al., 2024a)	94.1	57.7	13.2	98.4	63.7	62.0	68.0	55.9
LLM	Zero-shot (LLM: Claude 3.5 Sonnet)								
	w/ EHR context only	89.5	50.4	6.4	94.4	54.3	35.4	98.9	0.2
	w/ Classic RAG[a]	89.9	51.2	10.2	92.8	53.2	34.6	91.2	1.4
	w/ KARE-augmented context[b]	92.3	54.6	14.2	94.6	56.3	43.8	93.9	10.6
	Few-Shot (LLM: Claude 3.5 Sonnet)								
	w/ exemplar only (N=2)[c]	88.7	49.5	5.6	93.4	52.7	42.2	87.0	11.1
	w/ exemplar only (N=4)	88.0	49.2	5.6	92.7	53.6	44.7	84.0	15.7
	w/ EHR-CoAgent[d] (Cui et al., 2024)	87.4	51.7	13.0	91.8	55.2	46.1	78.2	20.1
	w/ KARE-augmented context	91.5	53.5	13.7	94.0	57.1	49.3	75.5	27.2
	Fine-tuned (LLM: Mistral-7B-Instruct-v0.3)								
	Backbone	90.4	53.0	11.4	94.3	57.6	57.6	50.5	66.3
	w/ Classic RAG	90.1	51.4	12.5	91.6	60.2	59.9	56.1	64.5
	KARE (ours)	**95.3**	**64.6**	**24.7**	98.3	**73.9**	**73.7**	**76.7**	**70.7**

Experiments

Main Results (MIMIC-IV)

1. For mortality prediction where the data is extremely imbalanced (5.42% positive labels), most ML models performed poor
2. LM+ML based methods improved the performance by leveraing external knowledge
3. Zero-shot, few-shot, and backbone-finetuned LLM-based methods perform worse than traditional ML methods in most cases
4. Classic RAG can even downgrade the performance in zero-shot setting
5. **KARE significantly outperform all the previous methods**

		MIMIC-IV							
		Mortality Prediction (pos = 19.16%)				**Readmission Prediction** (pos = 46.50%)			
Type	**Models**	**Accuracy**	**Macro F1***	**Sensitivity***	**Specificity**	**Accuracy**	**Macro F1**	**Sensitivity**	**Specificity**
ML	GRU (Chung et al., 2014)	88.7	76.4	42.9	99.6	62.4	62.2	68.3	56.2
	Transformer (Vaswani et al., 2017)	83.7	71.3	47.1	92.3	61.3	61.3	63.0	59.5
	RETAIN (Choi et al., 2016)	84.8	73.8	52.4	92.4	62.8	62.6	68.7	56.6
	GRAM (Choi et al., 2017)	86.4	74.4	50.6	93.9	62.5	62.5	67.4	57.8
	Deepr (Nguyen et al., 2016)	89.2	78.9	50.8	98.2	59.2	59.2	57.0	61.5
	TCN (Bai et al., 2018)	89.9	79.2	47.6	**99.9**	63.6	63.5	72.2	56.1
	ConCare (Ma et al., 2020b)	89.8	78.9	47.1	**99.9**	59.8	59.8	63.5	56.6
	AdaCare (Ma et al., 2020a)	88.7	78.2	50.3	97.8	62.9	62.9	58.4	67.7
	GRASP (Zhang et al., 2021)	89.9	79.1	47.6	99.8	59.7	59.6	53.1	66.7
	StageNet (Gao et al., 2020)	88.1	77.8	51.9	96.7	62.8	62.7	62.6	62.9
	KerPrint (Yang et al., 2023b)	88.7	79.8	53.1	98.0	63.5	63.3	67.0	60.1
LM+ML	MedRetriever (Ye et al., 2021)	89.5	77.9	55.6	95.2	63.0	62.1	69.4	55.8
	GraphCare (Jiang et al., 2024a)	91.5	80.3	57.8	96.6	65.7	65.5	66.2	65.0
	RAM-EHR (Xu et al., 2024)	90.5	78.4	52.6	97.0	65.5	65.5	64.0	67.0
	EMERGE (Zhu et al., 2024a)	90.7	78.3	53.4	96.6	63.3	63.2	61.5	64.9
LLM	Zero-shot (LLM: Claude 3.5 Sonnet)								
	w/ EHR context only	80.5	47.0	2.7	98.7	49.4	45.7	81.8	21.5
	w/ Classic RAG[a]	81.0	49.9	8.1	94.6	49.0	44.2	83.2	18.8
	w/ KARE-augmented context[b]	83.2	54.3	12.7	96.3	52.3	49.7	80.6	27.7
	Few-Shot (LLM: Claude 3.5 Sonnet)								
	w/ exemplar only (N=2)[c]	80.8	46.7	2.1	99.5	49.3	44.7	84.0	19.1
	w/ exemplar only (N=4)	81.6	49.9	5.3	99.8	49.0	44.1	84.3	18.2
	w/ EHR-CoAgent[d] (Cui et al., 2024)	81.0	55.5	13.8	97.0	51.2	46.3	78.4	24.0
	w/ KARE-augmented context	84.5	57.4	15.8	97.6	54.1	51.9	75.2	34.1
	Fine-tuned (LLM: Mistral-7B-Instruct-v0.3)								
	Backbone	92.2	83.1	65.0	96.2	56.1	46.7	23.1	**76.2**
	w/ Classic RAG	92.5	83.8	63.2	97.6	58.8	52.1	46.7	57.5
	KARE (ours)	**94.1**	**90.4**	**73.2**	99.8	**73.9**	**73.8**	**85.6**	63.7

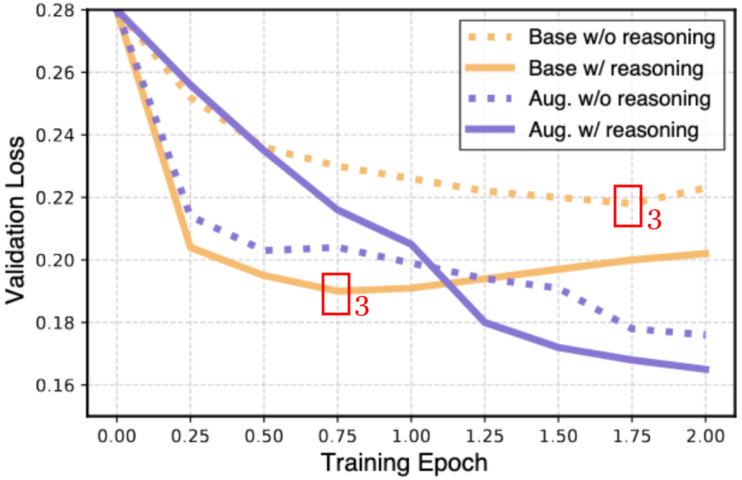
Experiments

Ablation Study

Study the components of fine-tuning

Similar Patients	Retrieved Knowledge	Reasoning	MIMIC-III-Mortality				MIMIC-III-Readmission			
			Accuracy	Macro F1	Sensitivity	Specificity	Accuracy	Macro F1	Sensitivity	Specificity
✗	✗	✗	90.4	53.0	11.4	94.3	57.6	57.6	50.5	66.3
✗	✗	✓	93.1	58.4	15.8	97.5	65.5	64.7	62.3	67.7
✗	✓	✓	95.3	64.6	24.7	98.3	72.8	72.6	74.7	70.6
✓	✓	✓	93.6	61.3	18.4	98.6	73.9	73.7	76.7	70.7

Similar Patients	Retrieved Knowledge	Reasoning	MIMIC-IV-Mortality				MIMIC-IV-Readmission			
			Accuracy	Macro F1	Sensitivity	Specificity	Accuracy	Macro F1	Sensitivity	Specificity
✗	✗	✗	92.2	83.1	65.0	96.2	56.1	46.7	23.1	76.2
✗	✗	✓	93.3	85.4	67.3	97.5	64.7	62.1	69.3	55.9
✗	✓	✓	93.8	89.6	74.5	98.8	72.2	71.9	81.1	64.0
✓	✓	✓	94.1	90.4	73.2	99.9	73.9	73.8	85.6	63.7



Key findings:

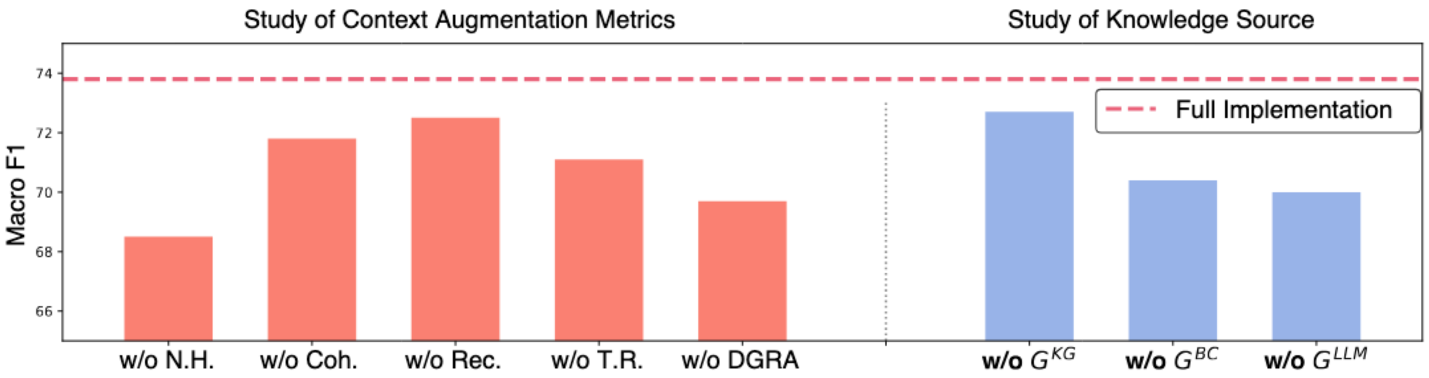
1. Both retrieved knowledge and reasoning chain significantly contribute to the performance gain
2. When the data is imbalanced (MIMIC-III-Mortality), similar pateint retrieval could hurt the performance
3. Without retrieved knowledge, the LLM could easily encounter overfitting issue

Experiments

Ablation Study

$$\text{Relevance}(C_k) = (\mathcal{H}_{\text{direct}}(C_k, G_p) + \alpha \cdot \mathcal{H}_{\text{indirect}}(C_k, G_p)) \times \text{Decay}(C_k, G_p) \times \text{Coherence}(S_{C_k}, \mathcal{B}_p) \times \text{Recency}(C_k, G_p) \times \text{ThemeRel}_\tau(C_k)$$

Study the components of retrieval algorithm and retrieval knowledge sources



(N.H.: node hits, Coh.: coherence, Rec.: recency, T.R.: theme relevance, DGRA: dynamic retrieval alg.)

(KGs constructed by bio-KG, bio-corpus, and LLM, respectively)

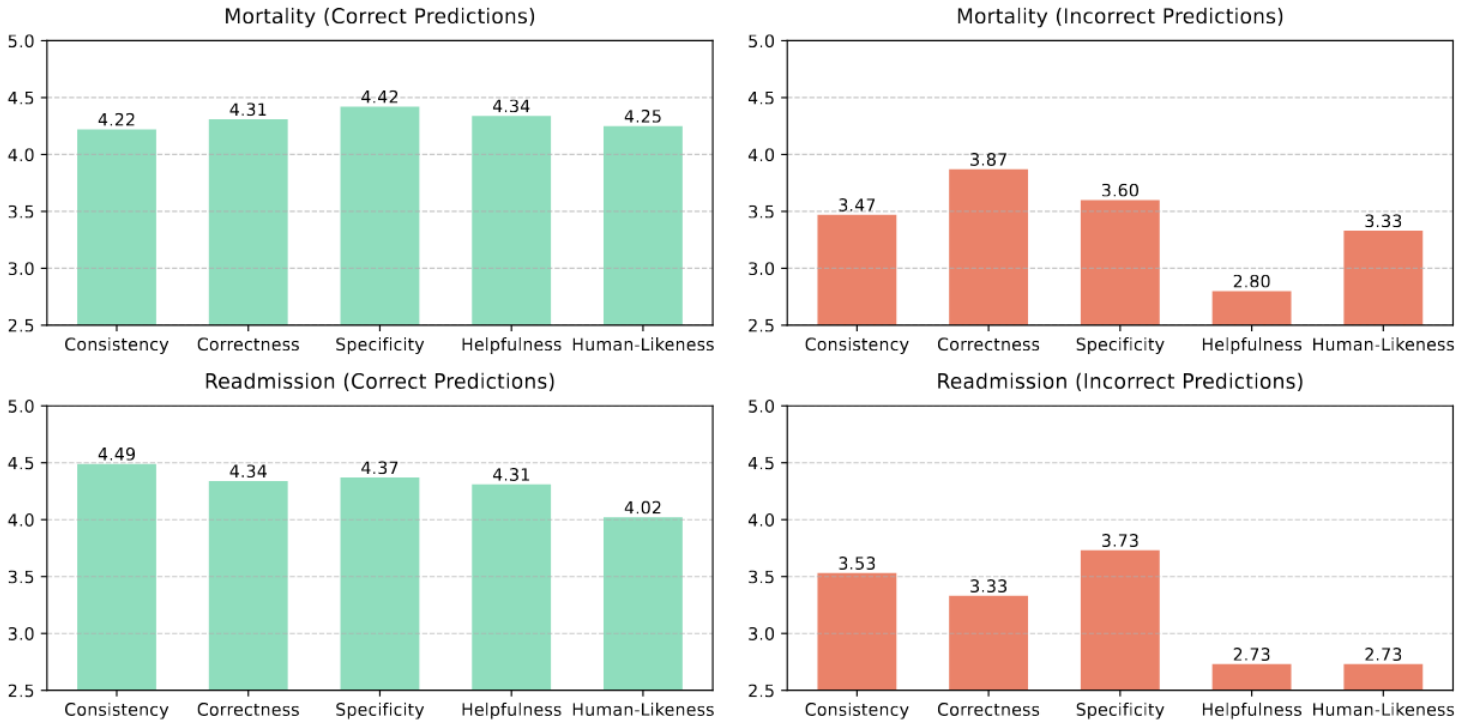
Key findings:

1. Node hits and DGRA are crucial factors to control the utility of retrieved knowledge, while other factors are also important
2. KGs constructed by bio-corpus (PubMed Abstract) and LLM contribute most, while the removal of KG constructed by UMLS has minor impact

Experiments

Human Evaluation

- We hired 3 MD students and 1 MD professional to evaluate 50 (35 correct and 15 incorrect predictions) randomly selected reasoning chains generated by KARE
- Five mertics: consistency, correctness, specificity, hepfulness, human (expert)-likeness

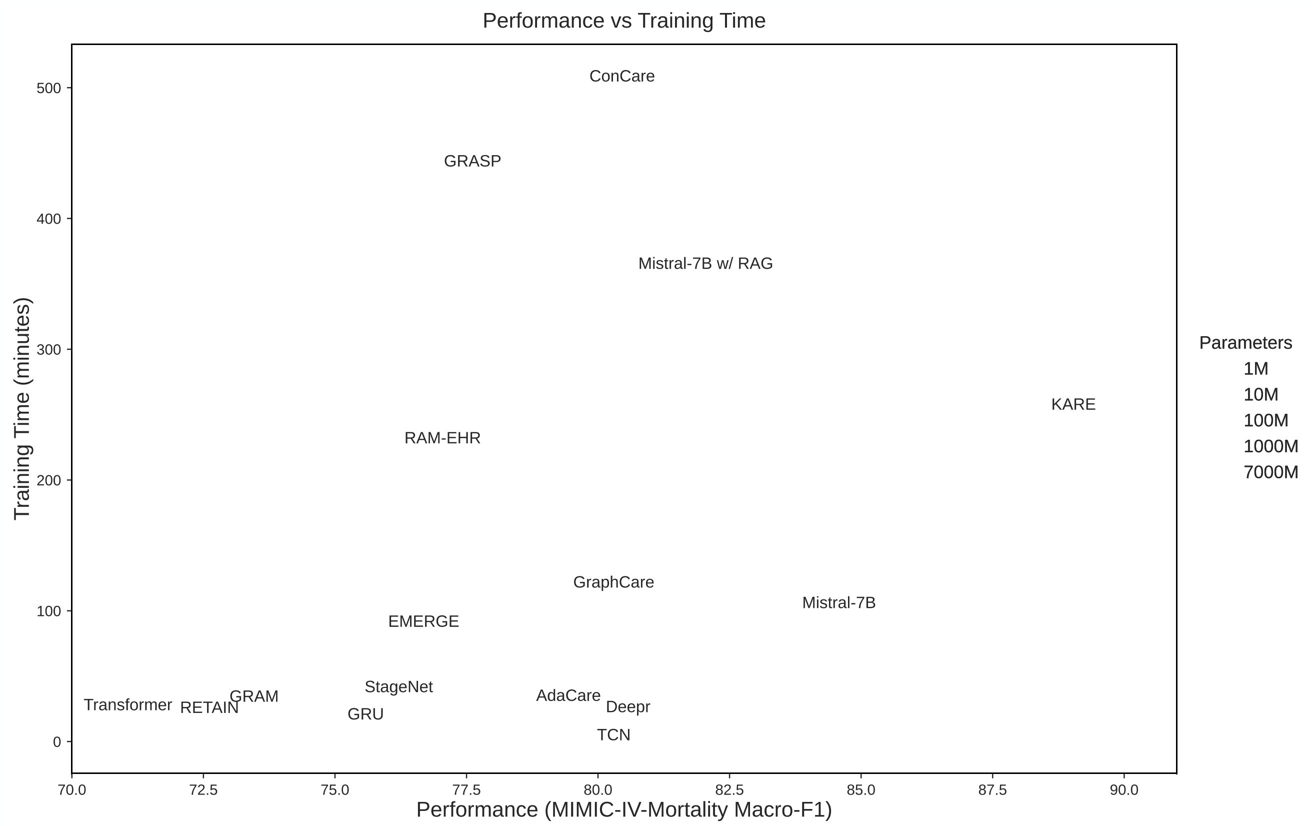


Discussions:

1. Reasoning chains leading to incorrect clinical predictions consistently score lower across all metrics
   - Highlights the critical role of high-qulaity reasoning chains
2. Human-likeness is notably lower for readmission prediction. This is because it is inherently difficult for clinicians gieven limited information (**KARE outperforms all the 4 experts on this task given these samples**)
3. Some conflictions between reasoning chains and the label predictions were observed, indicating a room for future improvement

Experiments

Efficiency Analysis



The training of KARE is more efficient than base LLM w/ Classic RAG, and is even faster than some ML models

Qualitative Study

Example of Knowledge Retrieval by KARE

EHR example

Patient Context:
Patient ID: 29488

Visit 0:
Conditions:
- Deficiency and other anemia
- Essential hypertension
- Complication of device; implant or graft
- Congestive heart failure; nonhypertensive - Cancer of prostate
- Anxiety disorders
- Thyroid disorders
- Disorders of lipid metabolism
- Conduction disorders
- Mycoses
- Other diseases of kidney and ureters
- Cancer of esophagus
Procedures:
- Diagnostic cardiac catheterization; coronary arteriography
- Other or procedures on vessels other than head and neck
- Colorectal resection
Drugs:
- Other drugs for obstructive airway diseases, inhalants in ATC
- Lipid modifying agents, plain
- Antithrombotic agents
- Angiotensin II receptor blockers (ARBs), plain

Visit 1:
Conditions:
- Congestive heart failure; nonhypertensive
- Cardiac dysrhythmias
- Shock
- Alcohol-related disorders
- Diabetes mellitus with complications
- E codes: adverse effects of medical care
- Disorders of teeth and jaw
Procedures:
- Other vascular catheterization; not heart
- Blood transfusion Drugs:
- Other drugs for obstructive airway diseases, inhalants in ATC
- Other analgesics and antipyretics in ATC
- Drugs for peptic ulcer and gastro-oesophageal reflux disease (GORD)
- Beta blocking agents
- Potassium supplements

Knowledge Retrieved by KARE

- Patients with congestive heart failure combined with cardiac dysrhythmias and shock have a significantly increased risk of mortality due to impaired cardiac output and hemodynamic instability.
- The presence of esophageal cancer, especially when co-occurring with prostate cancer, indicates a higher burden of malignancy, which can exacerbate overall health decline.
- Complications of devices, implants, or grafts can lead to severe infections or organ dysfunction, which is life-threatening, especially in patients with weakened cardiac function.
- The combination of anemia and thyroid disorders can exacerbate cardiovascular conditions, leading to poorer outcomes in heart failure patients.
- Alcohol-related disorders can negatively impact medication adherence and exacerbate conditions like hypertension and diabetes, increasing the risk of death.
- The use of blood transfusions and vascular catheterization procedures indicates severe clinical interventions often linked to critical conditions and a higher risk of death.
- Polypharmacy, including beta-blockers, potassium supplements, and medications for obstructive airway diseases, reflects complex medical management, which can be associated with increased risk of adverse events.
- Patients with diabetes mellitus with complications face a greater risk of serious outcomes due to potential organ damage and delayed healing processes.

Qualitative Study

Example of Inference by KARE

OUTPUT

Reasoning#
1. **Patient Overview**:
The patient has a history of polyhydramnios, other complications of pregnancy, thyroid disorders, and ob-related trauma to the perineum and vulva. The patient underwent repair of current obstetric laceration and other procedures to assist delivery. Postpartum medications include throat preparations, antiinflammatory and antirheumatic products, analgesics, antipyretics, agents for hemorrhoids and anal fissures, antipruritics, drugs for constipation, bacterial and viral vaccines, antacids, calcium supplements, and thyroid preparations.

2. **Relevant Retrieved Medical Knowledge**:
- Polyhydramnios can increase the risk of preterm labor and premature rupture of membranes, potentially leading to complications after delivery that require close monitoring and readmission.
- Thyroid disorders during pregnancy, if not properly managed, may result in adverse outcomes and increase the likelihood of postpartum complications necessitating readmission.
- Extensive lacerations or complications from assisted delivery may require additional care and monitoring, potentially leading to readmission.
- Proper thyroid hormone regulation is crucial postpartum, and imbalances may lead to complications requiring readmission if not closely monitored and adjusted.

3. **Comparison with Similar Patients**:
The patient shares several similarities with the two similar patients, including polyhydramnios, ob-related trauma to perineum and vulva, other complications of pregnancy, thyroid disorders, repair of current obstetric laceration, other procedures to assist delivery, and similar postpartum medications. However, the patient has an additional condition of "other complications of pregnancy" compared to the second similar patient. The first similar patient, who was readmitted within 15 days, also had this condition, along with "other complications of birth; puerperium affecting management of mother," which the current patient does not have.

4. **Reasoning Towards Prediction**:
The patient's history of polyhydramnios, thyroid disorders, and other complications of pregnancy, combined with the need for repair of obstetric laceration and assisted delivery procedures, suggests a higher risk of postpartum complications that may require readmission. The patient's similarity to the first patient, who was readmitted within 15 days and had the additional condition of "other complications of pregnancy," further supports this increased risk. While the patient is receiving appropriate postpartum medications and preventive measures, the presence of multiple risk factors and the need for close monitoring of thyroid function and potential complications from assisted delivery indicate a substantial likelihood of readmission within 15 days.

5. **Conclusion**:
Based on the patient's history of polyhydramnios, thyroid disorders, other complications of pregnancy, and the need for repair of obstetric laceration and assisted delivery procedures, along with the similarity to a patient who was readmitted within 15 days, there is a high likelihood that this patient will be readmitted to the hospital within 15 days of discharge.

Prediction
1

Conclusion

1. We proposed KARE, a novel framework integrating high-utility meidcal knowledge retrieval and LLM reasoning for interpretable and precise clinical prediction tasks.

2. KARE significantly outperforms previous methods on mortality and readmission prediction tasks on MIMIC-III and MIMIC-IV.

3. Without highly relevant medical knowledge and reasoning, LLMs perform bad on both tasks.

4. KARE performs even better than professinal clinicians on challenging tasks like readmission prediction given scarce information (e.g., no demographic information)

Thank you for your attention!
Please feel free to ask any questions.

Paper: https://arxiv.org/abs/2410.04585
Code: https://github.com/pat-jj/KARE

Thank you!

Patrick Jiang