Redefining Health Care Data Interoperability:

Empirical Exploration of Large Language Models in Information Exchange

Journal Club

21 Feb 2025

Presenter: Mr. Fookty PHENGPHAKEO

MSc. Student

Outline

- Introduction
  - Objective
- Method
  - Data Sources (MIMIC and UK Biobank)
  - Evaluation Metrics
- Results
- Discussion
  - Error Analysis
- Conclusion

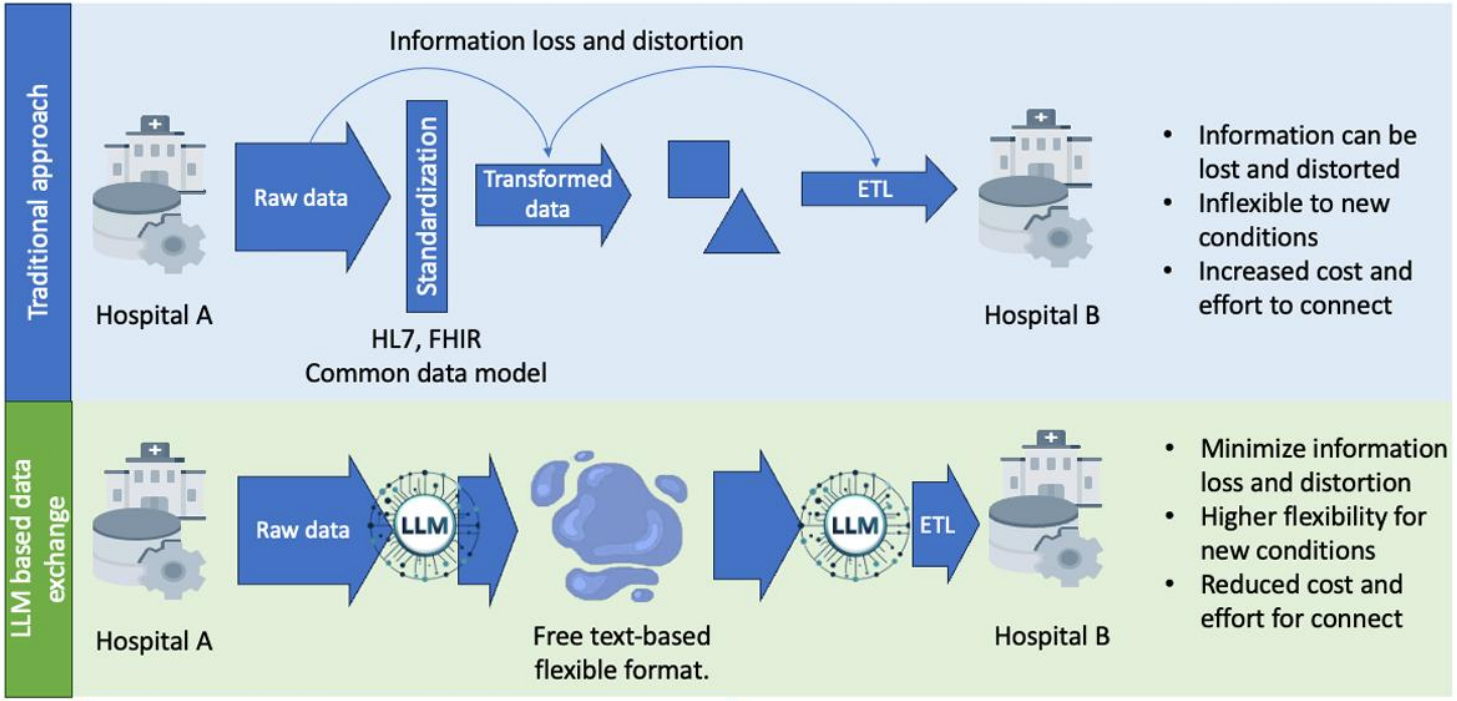
Introduction

- Efficient healthcare data exchange is crucial, especially with the rise of **personalized medicine** and **patient-generated health data**.
- The global healthcare systems struggle with **inconsistent** medical record formats and coding systems leading to **data integration challenges** and information loss.
- This study explores **Large Language Models (LLMs)** like **ChatGPT** as a solution to improve **data interoperability** by **flexibly converting structured data into natural language and** reducing reliance on strict standardization.

Introduction (Cont')

- LLMs enhance data exchange by:
  - **Generating coherent summaries** of structured data.
  - **Accurately mapping diagnostic codes** across different systems.
  - **Extracting critical information** from unstructured medical records.
- The rise of personalized health care necessitates the integration of personal health records from **various institutions** showing the complicated data exchange processes;
- Health care systems worldwide face challenges due to varying medical record formats and coding systems:
  - International Classification of Diseases (ICD);
  - Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT);
  - FHIR (Fast Healthcare Interoperability Resources)…

Comparison of traditional standardization-based and proposed text-based flexible data exchange processes



Objective

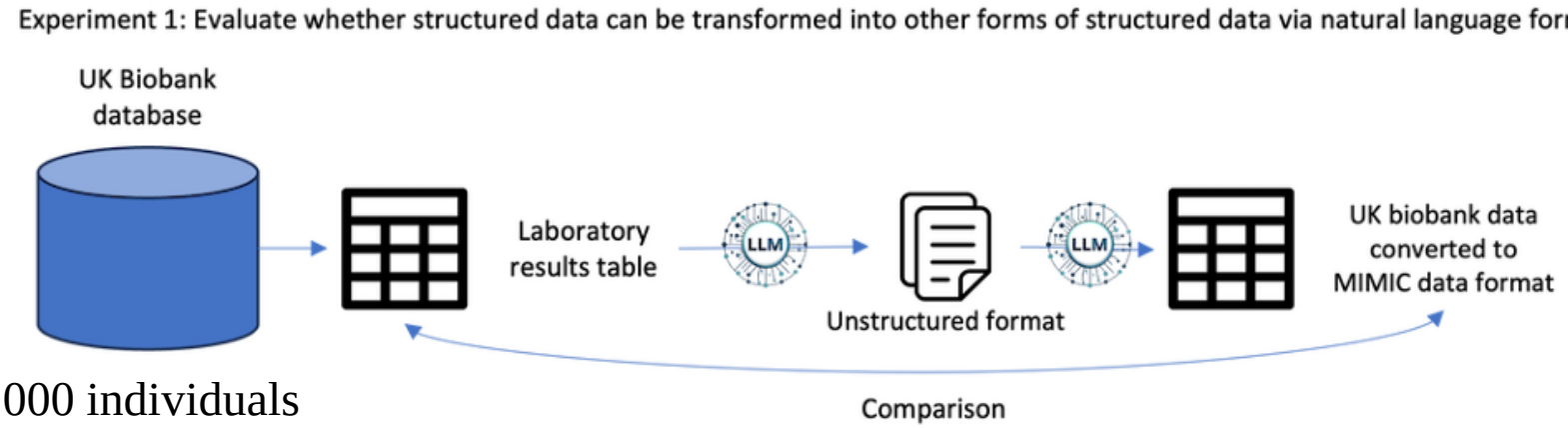
- The study aimed to evaluate LLMs (i.e., ChatGPT3.5) in transforming and transferring healthcare data to support interoperability; Specifically,
  - **Experiment 1**: Evaluate the accuracy of numerical data transformation into text and back
  - **Experiment 2**: Fidelity of text-based transformation for semantic data using ICD codes
  - **Experiment 3**: Effectiveness of extracting specific information during the transfer .of text-format data

Three experiments

1	2	3
Accuracy of transforming structured **lab** results into unstructured format	Conversion of diagnostic codes between ICD-9-CM and SNOMED-CT	Extracting targeted information from unstructured discharge notes
		
Transforming structured lab results into an unstructured format.	Converting diagnostic codes between ICD-9-CM and SNOMED-CT using an LLM and traditional mapping	Extracting clinical information (drug names) from unstructured records.

Overview of experimental approaches to evaluate LLM performance in data extraction and transformation.

A

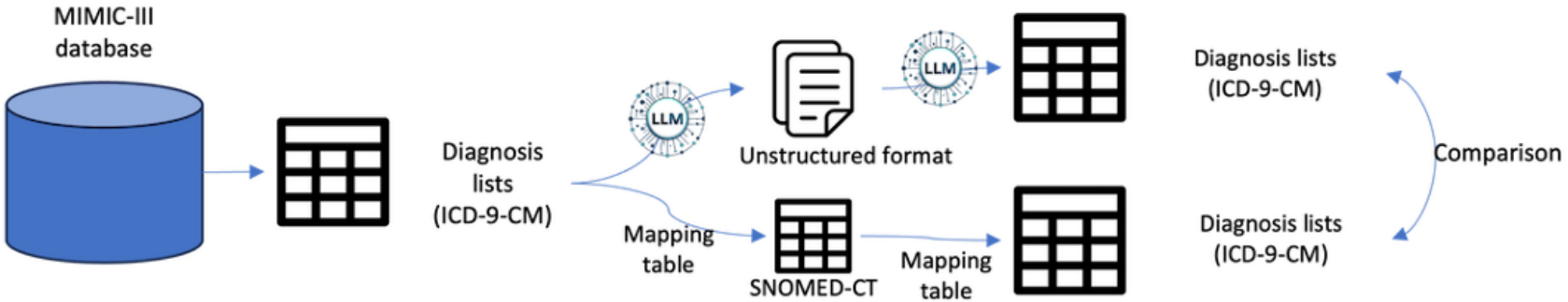


- This experiment assessed the accuracy of transforming structured **laboratory** results into an unstructured format using LLMs.
- The data were restructured to comply with the **MIMIC-III data architecture**.

Overview of experimental approaches to evaluate LLM performance in data extraction and transformation.

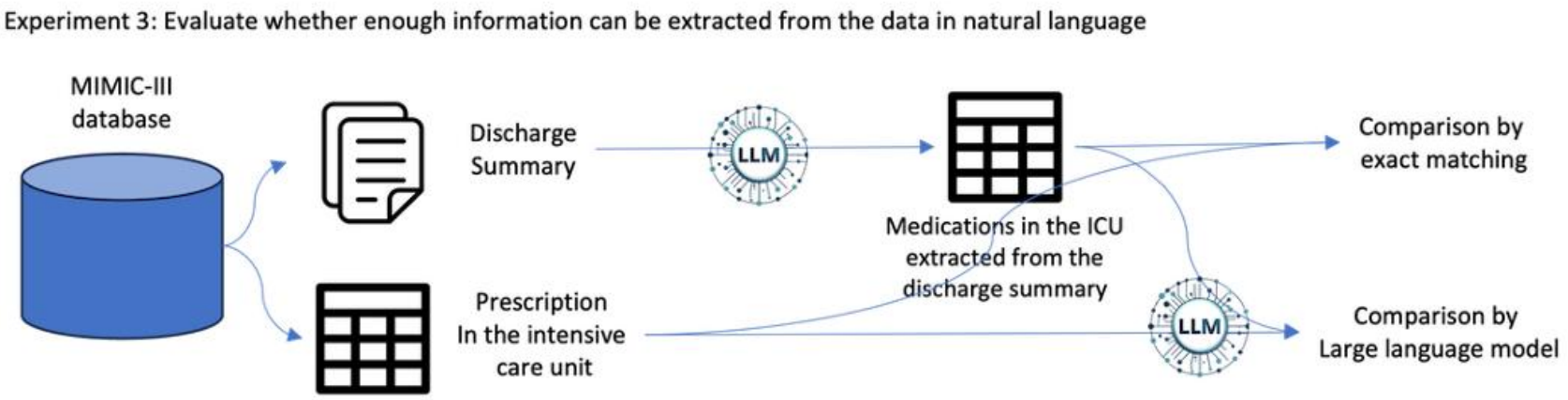
B

Experiment 2: Evaluate whether the diagnosis can be accurately conveyed in natural language form



Overview of experimental approaches to evaluate LLM performance in data extraction and transformation.

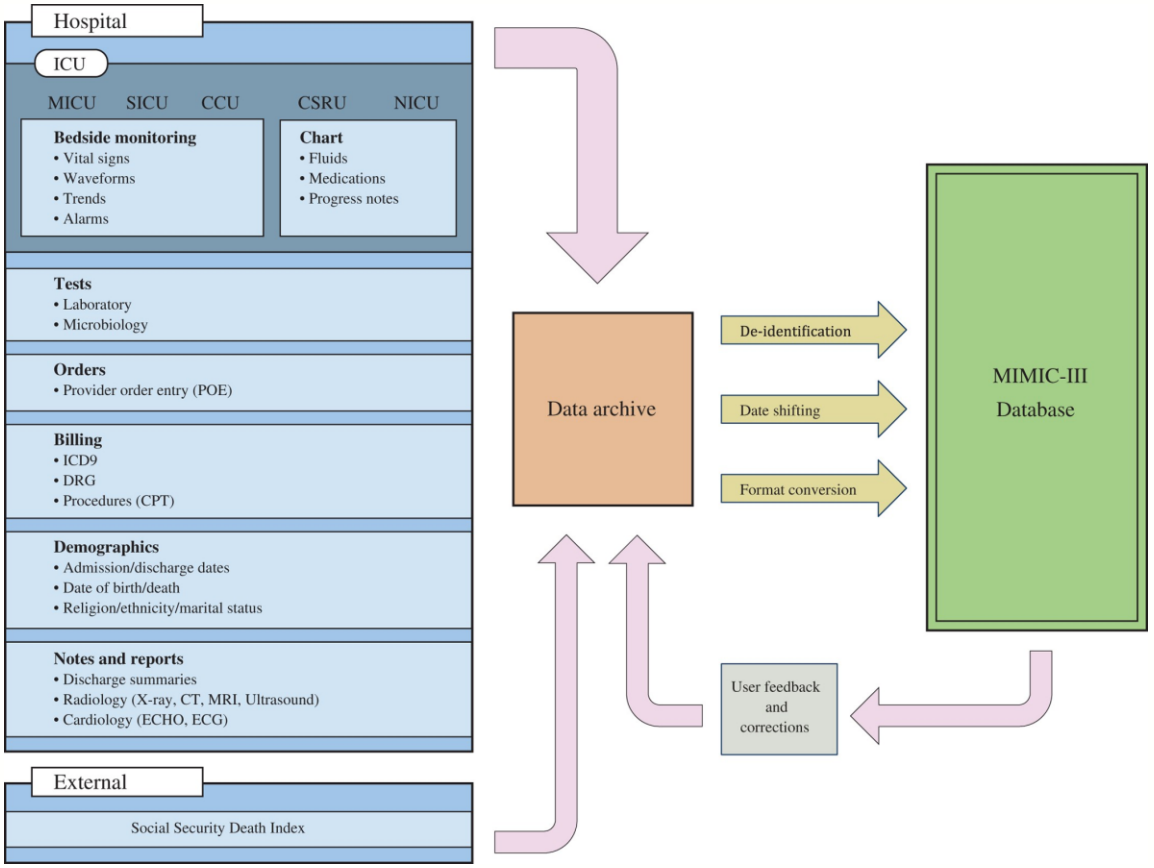
C



Data sources

- The public health care data sets were used:
  - The Medical Information Mart for Intensive Care III (**MIMIC-III**)
    - Contains lab for Computational Physiology and publicly available resource containing the deidentified health data of approximately 40,000 critical care patients.
  - UK Biobank.
    - is a large-scale biomedical database and research resource containing **de-identified genetic**, **lifestyle** and health information and **biological** samples;
    - Monitoring the lives of 500,000 voluntary participants aged between 40 and 69 years in the UK from 2006 to 2010.

Overview of the MIMIC-III critical care database.



https://www.nature.com/articles/sdata201635/figures/1

Summary of data used in each experiment.

	Experiment 1	Experiment 2	Experiment 3
Database	UK Biobank	MIMIC-III[a]	MIMIC-III
Data type	Laboratory test results	Diagnosis code (*ICD-9-CM*[b])	Discharge summary
Number of records	502,396	651,047	59,652
Number of patients	502,396	46,520	41,127
Age (years), mean (SD)	56.53 (8.09)	64.43 (57.20)	58.35 (53.63)
Sex (male), n (%)	229,079 (45.6)	26,121 (56.2)	23,199 (56.4)
Length of text (number. of characters), mean (SD)	N/A[c]	N/A	9618.92 (5539.64)
Number of tests	11,973	N/A	N/A
Number of diagnosis codes	N/A	6984	N/A

[a]MIMIC-III: Medical Information Mart for Intensive Care III.

[b]ICD-9-CM: International Classification of Diseases, Ninth Revision, Clinical Modification.

[c]N/A: not applicable.

Prompt Engineering:

- Conducted multiple trials featuring a range of prompts to avoid overfitting bias;
- They differentiated between the data used for prompt engineering and those used to assess the performance
  - No mention of how much data used to develop prompt
- Used a manual trial-and-error approach to develop the prompt.

Textbox 1. Summary of prompts used in experiments 1

Step 1: Translating laboratory test results into free text

"I have the following patient. Based on this information, summarize the patient's condition in natural language. Make sure to include all the information presented. The values of the lab results should remain numerical. (For the Sex variable, 0 = female and 1 = male.)"

{List of lab results}

Step 2: Transforming free text data into the structured format "I have the following patient." {Generated text from the above step} "Extract and organize information on the following items." (Add the value next to the variable name with no further explanation.) {Defined result extraction format}

Summary of prompts used in experiments 2

Step 1: Translating diagnosis codes to natural language text

"I have a diagnosis called {Diagnosis code}.

Describe it in natural language used by doctors and other health care professionals. Write it as a single phrase of only a few words (less than 15 words but do not use abbreviations).

All semantics must be included."

Step 2: Translating natural language text to diagnosis codes "Where does {Descriptions on diagnosis} fit in the following categories?

Categories according to International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM)}

Provide the most appropriate ICD-9-CM code directly or choose one of the categories above.

Choose only one answer that seems the most relevant and answer in the following format.

The corresponding code: [Code (without periods): a description of the code]."

Summary of prompts used in experiments 3

Step 1: Extracting medication list from discharge summary

role: “system,” “content:” Your role is to interpret medical records.

role: “assistant,” “content:” I only need prescriptions from the ICU, not from the general ward or not from outside our hospital.

Organize by ingredient name, not generic name.

Never include medications on admission and discharge medications.

Exclude information before ICU admission or after ICU discharge, even if it is for a hospital stay.

In other words, exclude prescriptions that were written in a regular ward or emergency room.

Exclude any medications that may not have been prescribed in the ICU.

Finally, exclude all prescriptions for procedures, and tests. that are not prescriptions for medication.

role: “user,” “content:” Observing the following patient record, organize a list of medications prescribed during the ICU visit.

Organize them in the following format (Provide only the name, not the dose)

drug name 1

drug name 2

If any information on the medications prescribed in the ICU is unavailable, simply answer “None.”

Summary of prompts used in experiments 3

Step 2: Converting drug names to ingredient names

{extracted drug list from the above step}

Organize the above medications by ingredient name.

If the drug is recorded by trade name, replace it with the ingredient name.

In the case of multiple ingredient names, record a representative one.

The format should be a single line of ingredient names with no further explanation, like this

List: Ingredient 1, Ingredient 2,...

Summary of prompts used in experiments 3

Step 3: Comparing extracted drug information with actual prescription records

Here is the medication information extracted from the discharge summary.

{extracted drug list from the above step}

These are the medication details actually recorded in the prescription record.

{Ingredient list from the above step}

Organize the medication information extracted from the discharge summary by its actual inclusion in the prescription record.

Medications not mentioned in the discharge summary should not be listed.

The exact name of the medication may not be recorded, or a synonym may be used.

In these cases, mark the medication as actually prescribed.

For example, warfarin might be described as coumadin.

Record the same medication under different names as the one that was prescribed.

Match the same ingredient even if the added bases differ.

For example, the ingredient name of Lopressor is Metoprolol tartrate, but the ingredient must be confirmed as “true” even if it is Metoprolol.

Ingredient names may be written as abbreviations. For example, acetaminophen may be written as APAP.

Exclude P.R.N. prescriptions.

Exclude simple fluid prescriptions.

Provide only “true” or “false” information for each drug.

Do not provide Python code. Provide only the results in an array.

Fill in the blanks with a “true” or “false” result in the following format

{Defined result extraction format}

Evaluation metrics:

- Positive predictive value (PPV)
  - For instance, in Experiments 1 and 3, PPV was used to measure the accuracy;
  - **Additionally**, the study assessed the absence or presence of data omissions using sensitivity and specificity in experiment 1.

The Positive Predictive Value (PPV) (Example Experiment 3)

$$PPV = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

- Total Medications Extracted by the LLM: **5604**
- True Positives (Exact Match): **2483**
  (medications correctly matched the structured prescription table)
- False Positives:
  - 5604−2483=3121 (medications extracted but did not match the prescription record)

$$PPV = \frac{2483}{2483 + 3121} = \frac{2483}{5604} \approx 44.3\%$$

Results

Results of Experiment 1: Efficiency of the LLM in Data Transformation and Retrieval

- This resulted in 11,996 data points spanning 13 distinct test items;
- During the transformation process 23 items were lost, **with 11,973 (99.8%) being successfully converted;** 24 items did not match their original values.

Summary of experimental results from data transformation and extraction using LLM in experiment 1.

Variable	Raw data		After transformation		Number of data not transferred during the transformation process	Number of data with changed values during the transformation process	MSE[a]
	n, (%)	Mean (SD)	n, (%)	Mean (SD)			
Age	1000 (100)	56.94 (8.03)	1000 (100)	56.94 (8.03)	0	0	0
Sex	1000 (100)	0.47 (0.5)	1000 (100)	0.47 (0.5)	0	0	0
BMI	994 (100)	27.04 (4.78)	994 (100)	27.04 (4.78)	0	24	2.12×10^{-6}
ALT[b]	919 (100)	23.55 (15.32)	919 (100)	23.55 (15.32)	0	0	0
AST[c]	918 (100)	26.13 (11.21)	918 (100)	26.13 (11.21)	0	0	0
Bilirubin	772 (100)	1.84 (0.81)	772 (100)	1.84 (0.81)	0	0	0
Creatinine	920 (100)	72.94 (18.65)	907 (98.6)	72.85 (18.69)	13	0	0
GGT[d]	920 (100)	39.06 (46.96)	920 (100)	39.06 (46.96)	0	0	0
HbA_{1c}[e]	930 (100)	35.88 (5.66)	930 (100)	35.88 (5.66)	0	0	0
HDL[f]	846 (100)	1.46 (0.38)	846 (100)	1.46 (0.38)	0	0	0
LDL[g]	915 (100)	3.54 (0.88)	915 (100)	3.54 (0.88)	0	0	0
Platelet count	943 (100)	255.39 (59.79)	943 (100)	255.39 (59.79)	0	0	0
Triglycerides	919 (100)	1.73 (1.04)	909 (98.9)	1.74 (1.05)	10	0	0
Total	11996 (100)	42.9 (69.66)	11973 (99.8)	42.9 (69.71)	23	24	1.76×10^{-7}

[a]MSE: mean squared error.
[b]ALT: alanine transaminase.
[c]AST: aspartate transaminase.
[d]GGT: gamma-glutamyl transferase.
[e]HbA_{1c}: hemoglobin A_{1c}.
[f]HDL: high-density lipoprotein.
[g]LDL: low-density lipoprotein.

Results of Experiment 2: Analysis of Diagnostic Code Conversion (Mapping Table vs Text-Based Methods)

- Diagnostic codes were adapted based on a mapping table
  - 1:n relationship owing to challenges in semantic translation.
    - 5748 diagnostic codes expanded to 218,088 codes
  - 1:1 ratio
    - 5748 original codes corresponded to 5748 records
  - Example: The ICD-9-CM code was mapped as 2 distinct codes in SNOMED-CT

ICD9-CM	SNOMED CT
1. Malignant pleural effusion (51181)	1. Malignant pleural effusion (363346000) 2. Pleural effusion owing to malignant neoplastic disease (disorder) (860792009)

The results before and after the conversion, we found that the mapping table achieved the following consistency values:

	consistency values Of 1:n relationship	Consistency of text-based methods 1:1	Top 1000 diagnostic frequency used
Level 1	0.626 (136,431/218,088)	0.904 (5197/5748)	0.918
Level 2	0.248 (54,068/218,088)	0.844 (4850/5,748)	0.896
Level 3	0.096 (21,000/218,088)	0.597 (3430/5748)	0.733

These results suggested that the frequent use of diagnostic names may provide better precision when shared between different databases.

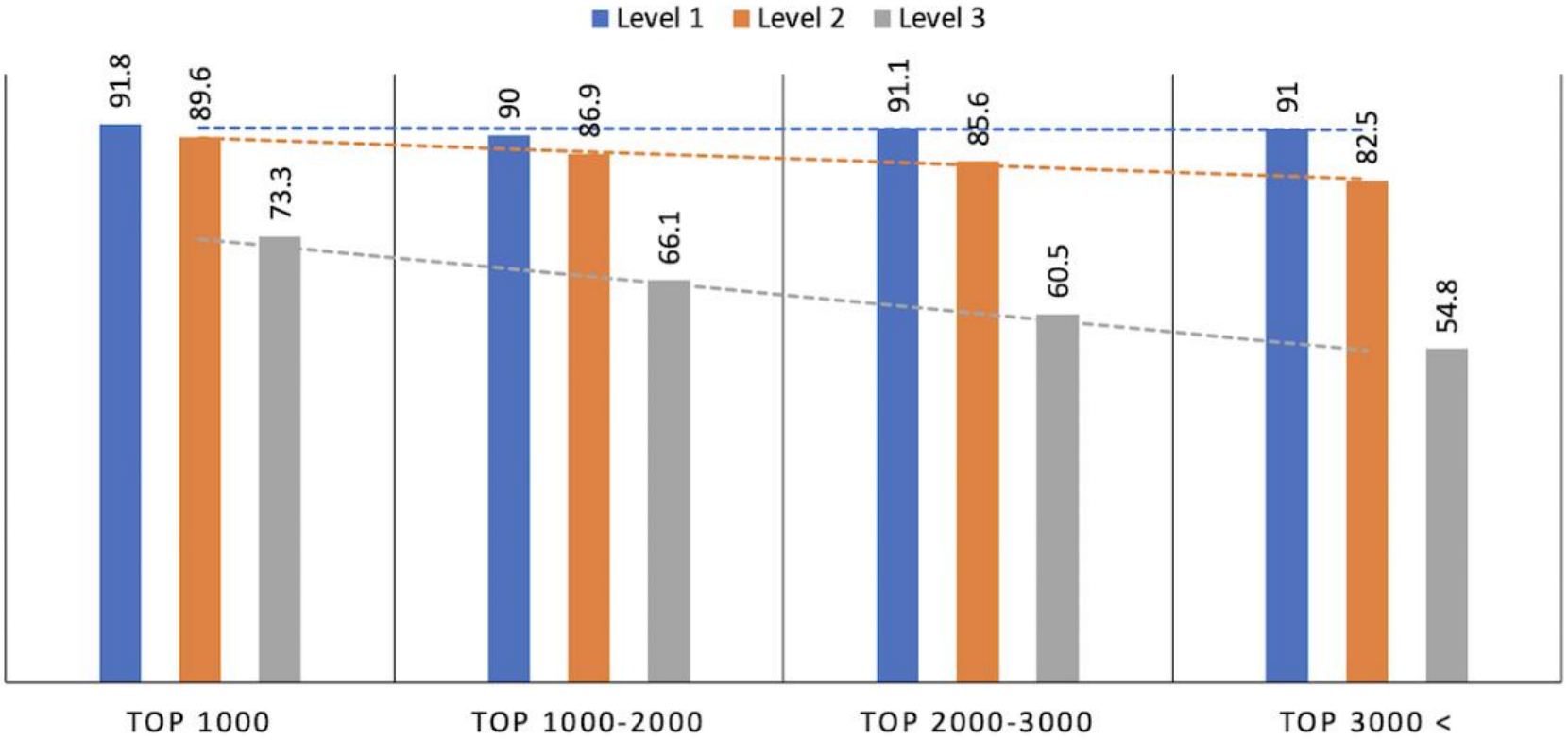
Experiment 2: Diagnostic Code Conversion

- The conversion of diagnostic codes between ICD-9-CM and SNOMED-CT using both a **traditional mapping table** and a **text-based approach** facilitated by the LLM (ChatGPT3.5)
- The text-based approach showed enhanced consistency in diagnostic code conversion, particularly for frequently used diagnostic names, compared with the traditional mapping approach.
  - **Enhanced Consistency**

    The text-based approach outperformed traditional mapping.
  - **Frequency Matters**

    More frequent diagnoses had higher conversion accuracy.

Figure 3. Results of converting diagnoses from ICD-9-CM (International Classification of Diseases, Ninth Revision, Clinical Modification) coding to natural language and back to ICD-9-CM. At the highest level, Level 1, most information aligned closely with the original data. However, accuracy decreased as the categories became more specific at levels 2 and 3. Notably, more frequently used diagnoses (toward the left on the x-axis) showed higher conversion accuracy.



Results of Experiment 3: Effectiveness of the LLM in Extracting Relevant Information From Medical Records

	Number of medications, n (%)
Medications in the intensive care unit extracted from the discharge summary	5604 (100)
Medications that exactly matched the prescription name	2483 (44.3)
Medications semantically matched by large language model, including synonyms	5055 (90.2)[a]

[a]A total of 2572 medications were described using different terminology than the prescription.

Comparison of drug information extracted from natural language discharge summaries with prescription records.

The LLM showed a PPV of **44.3%** in extracting **generic drug names**, demonstrating its potential for efficient information retrieval from unstructured medical text.

Discussion

LLMS improve accuracy in data exchange	The study demonstrated that LLMs can effectively transform structured data into text and back to structured data with minimal loss of information.
Text-based diagnostic code conversion is more accurate	**- Traditional mapping methods many-to-many mappings** **- LLM approach maintained significantly higher accuracy**
LLMs extract key medical information effectively	**LLM extracted ICU medication data from discharge summaries with 90.2% accuracy** **LLMs' ability to understand synonyms is crucial**
Implications for global health data exchange	**LLMs can facilitate international data sharing and improve patient care, clinical decision-making, and research collaborations**

Limitations and future directions

- The study used the GPT-3.5 model; newer models like GPT-4 may yield better results.
- The study relied on MIMIC-III and UK Biobank datasets, which may not fully represent global healthcare data.
- Future research should apply LLMs across more **varied datasets** to ensure generalizability and applicability to healthcare contexts.

Limitations and future directions

Security and privacy concerns

- Using **cloud-based LLMs** (like ChatGPT) may pose **data privacy risks**.
- Healthcare institutions should consider **on-premise LLMs** to ensure **data security and regulatory compliance**.

Error analysis

- All inconsistencies were found to stem from the rounding off of decimal values. For instance, an original BMI value of 24.4383 was translated as 24.44.
- Consequently, the calculated mean squared error was a minimal 1.76×10^{-7}

Conclusion

- LLMs have the potential to transform healthcare data exchange, enhancing precision and efficiency.
- Their role in minimizing errors and facilitating knowledge transfer among providers is significant.
- The study suggests a unified global healthcare landscape through improved data sharing.
- Using LLMs can enhance international health information exchanges, leading to better collaboration between countries and potentially benefiting patient care worldwide by ensuring that medical knowledge and practices are more consistently applied.

Thank you