A Dataset for Grounded Question Answering from Electronic Health Records to Relieve Clinician Burden

Sarvesh Soni, PhD, Dina Demner-Fushman, MD, PhD
National Library of Medicine, National Institutes of Health, Bethesda, MD, USA
Contact: sarvesh.soni@nih.gov

U.S. National Library of Medicine

INTRODUCTION

- Patient-portal inbox messages are a major driver of clinician burden[1].
- Tools that help clinicians draft replies can reduce this burden.
- Generating answers using patients' medical data is a key opportunity.
- Prior work targets general consumer health question answering (QA)[2].
- To our knowledge, no existing work aimed at answering patients' questions using their electronic health record (EHR) data.
- We propose a novel dataset, ArchEHR-QA, for EHR QA, with patient questions, clinician-interpreted questions, EHR note excerpts, and clinician answers (Figures 1 & 2).

DATA

Component	ICU	ED	All
Total Cases	104	30	134
Patient Question WC	90.2	94.8	91.3
Clinician Question WC	10.6	10.2	10.5
Answer WC	72.6	72.3	72.6
Note Excerpt WC	410.2	280.7	381.2
Note Sentences	27.6	19.3	25.7
Essential	7.0 (25.5%)	5.2 (26.7%)	6.6 (25.7%)
Supplementary	5.9 (21.4%)	2.6 (13.4%)	5.2 (20.1%)
Not Relevant	14.7 (53.1%)	11.6 (59.8%)	14.0 (54.3%)

Table 1: Descriptive statistics. All values are means. WC: word count.

EXAMPLE

Patient Question *(underlined are the areas of focus)*

Took my 59 yo father to ER ultrasound discovered he had an aortic aneurysm. He had a salvage repair (tube graft). Long surgery / recovery for couple hours then removed packs. why did they do this surgery????? After this time he spent 1 month in hospital now sent home.

Clinician Question *(interpreted from the patient question)*

Why did they perform the emergency salvage repair on him?

Clinical Note Excerpt *(sentences numbered for grounding)*

1: He was transferred to the hospital on 2025-1-20 for emergent repair of his ruptured thoracoabdominal aortic aneurysm. **2:** He was immediately taken to the operating room where he underwent an emergent salvage repair of ruptured thoracoabdominal aortic aneurysm with a 34-mm Dacron tube graft using deep hypothermic circulatory arrest. **3:** Please see operative note for details which included cardiac arrest x2. **4:** Postoperatively he was taken to the intensive care unit for monitoring with an open chest. **5:** He remained intubated and sedated on pressors and inotropes. **6:** On 2025-1-22, he returned to the operating room where he underwent exploration and chest closure. **7:** On 1-25 he returned to the OR for abd closure JP/ drain placement/ feeding jejunostomy placed at that time for nutritional support.

8: Thoracoabdominal wound healing well with exception of very small open area mid wound that is @1cm around and 1/2cm deep, no surrounding erythema. **9:** Packed with dry gauze and covered w/DSD.

Clinician Answer *(with citations to note sentences)*

The patient needed emergency salvage repair for his aortic aneurysm because the aorta had ruptured **[1]**. This rupture is something that needs to be repaired immediately or the patient will die. The patient needed a tube graft for the repair **[2]**. An aortic aneurysm is a very serious diagnosis and repair surgery is not something that can be put on hold. Additionally, it appears this surgery was absolutely necessary because the patient arrested twice during the operation **[3]**.

Figure 1: Example case. Sentences *1 and 2 are essential; 3 is supplementary; others not-relevant*.

METHODS

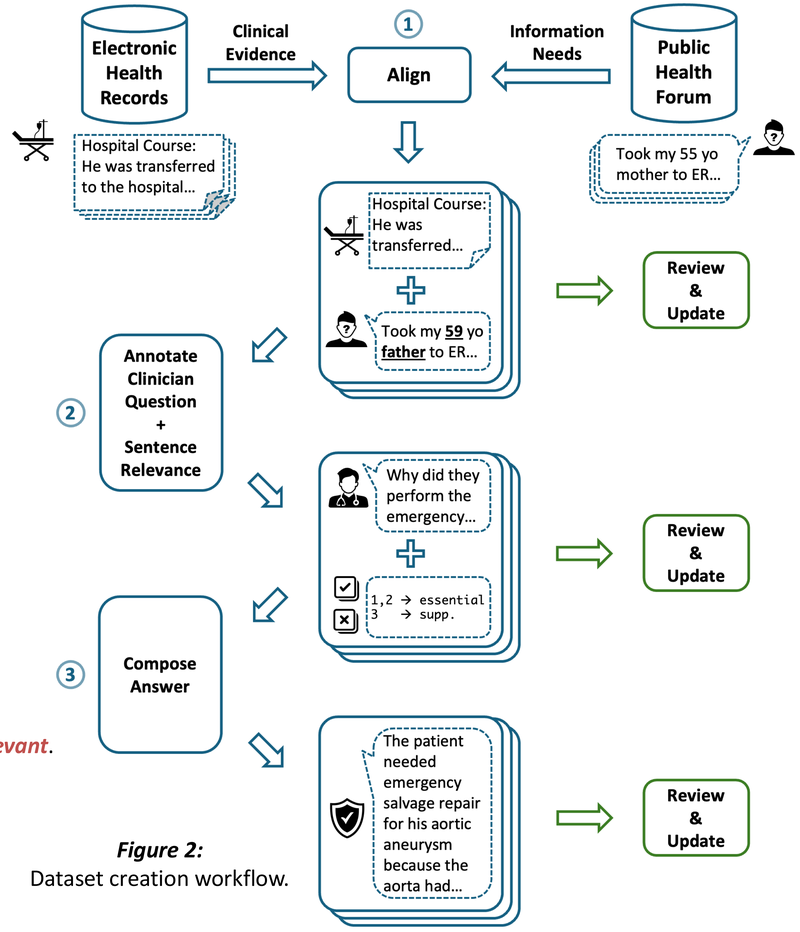

Figure 2: Dataset creation workflow.

RESULTS

Model	Factuality			Relevance						
	P	R	F1	BLEU	ROUGE	SARI	BERTScore	MEDCON	AlignScore	Avg
Llama 3	**59.5**	39.8	47.7	3.4	22.7	51.0	40.4	38.9	**41.5**	33.0
Llama 4	56.9	**47.6**	**51.8**	6.7	23.6	51.7	40.3	38.1	36.0	32.7
Mixtral	49.7	45.9	47.7	**6.8**	**24.2**	**53.1**	**42.5**	**40.2**	35.8	**33.8**

Table 2: Benchmarking EHR QA with sentence-level citations to input clinical note. Factuality: F1 Score between cited and ground truth sentences. Relevance: Text and semantics-based metrics against clinician answer. Best scores are **bolded**. P: Precision, R: Recall, F1: F1 Score.

DISCUSSION

- Manual error analysis of model responses:
  - Unsupported attribution: Correctly inferred antibiotics were for pneumonia but invented the rationale ("CXR showed low O_2 levels").
  - Prompt echoing over evidence: Disproportionately reproduced patient's question instead of grounding answer in clinical evidence.
- ArchEHR-QA will serve as a strong benchmark for evaluating automated patient EHR QA systems.

ACKNOWLEDGMENTS

This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health, and utilized the computational resources of the NIH HPC Biowulf cluster.

REFERENCES

1. Martinez KA, et al. Patient Portal Message Volume and Time Spent on the EHR: An Observational Study of Primary Care Clinicians. J Gen Intern Med. 2024.
2. Welivita A, Pu P. A Survey of Consumer Health Question Answering Systems. AI Magazine. 2023.