Question Answering for Electronic Health Records: A Scoping Review of datasets and models

Jayetri Bardhan, Kirk Roberts, Daisy Zhe Wang

Submitted to: Journal of Medical Internet Research
on: October 16, 2023

Table of Contents

Question Answering for Electronic Health Records: A Scoping Review of datasets and models

Jayetri Bardhan[1] BE, MS; Kirk Roberts[2] BS, MS, PhD; Daisy Zhe Wang[1] MS, PhD

[1]Department of Computer and Information Science and Engineering University of Florida Gainesville US
[2]School of Biomedical Informatics The University of Texas Health Science Center at Houston Houston US

Corresponding Author:
Kirk Roberts BS, MS, PhD
School of Biomedical Informatics
The University of Texas Health Science Center at Houston
7000 Fannin St #600
Houston
US

Abstract

Background: Question Answering (QA) systems on patient-related data can assist both clinicians and patients. They can, for example, assist clinicians in decision-making and enable patients to have a better understanding of their medical history. Significant amounts of patient data are stored in Electronic Health Records (EHRs), making EHR QA an important research area. In EHR QA, the answer is obtained from the patient's medical record. Because of the differences in data format and modality, this differs greatly from other medical QA tasks that employ medical websites or scientific papers to retrieve answers, making it critical to research EHR question answering.

Objective: This study aimed to provide a methodological review of existing works on QA over EHRs. The objectives of this study were to (i) identify the existing EHR QA datasets and analyze them, (ii) study the state-of-the-art methodologies used in this task, (iii) compare the different evaluation metrics used by these state-of-the-art models, and finally (iv) elicit the various challenges and the ongoing issues in EHR QA.

Methods: We searched for articles from January 1st, 2005 to September 30th, 2023 in four digital sources including Google Scholar, ACL Anthology, ACM Digital Library, and PubMed to collect relevant publications on EHR QA. Our systematic screening process followed PRISMA guidelines. 4111 papers were identified for our study, and after screening based on our inclusion criteria, we obtained a total of 47 papers for further study. The selected studies were then classified into two non-mutually exclusive categories depending on their scope: 'EHR QA datasets' and 'EHR QA Models'.

Results: A systematic screening process obtained a total of 47 papers on EHR QA for final review. Out of the 47 papers, 25 papers were about EHR QA datasets, and 37 papers were about EHR QA models. It was observed that QA on EHRs is relatively new and unexplored. Most of the works are fairly recent. Also, it was observed that emrQA is by far the most popular EHR QA dataset, both in terms of citations and usage in other papers. We have classified the EHR QA datasets based on their modality, and we have inferred that MIMIC-III and the n2c2 datasets are the most popular EHR database/corpus used in EHR QA. Furthermore, we identified the different models used in EHR QA along with the evaluation metrics used for these models.

Conclusions: EHR QA research faces multiple challenges such as the limited availability of clinical annotations, concept normalization in EHR QA, as well as challenges faced in generating realistic EHR QA datasets. There are still many gaps in research that motivate further work. This study will assist future researchers in focusing on areas of EHR QA that have possible future research directions.

(JMIR Preprints 16/10/2023:53636)
DOI: https://doi.org/10.2196/preprints.53636

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Question Answering for Electronic Health Records: A Scoping Review of datasets and models

Jayetri Bardhan[a], Kirk Roberts[b,1], Daisy Zhe Wang[a]
[a]*Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL, USA*
[b]*School of Biomedical Informatics, The University of Texas Health Science Center at Houston, TX, USA*

Abstract

Background: Question Answering (QA) systems on patient-related data can assist both clinicians and patients. They can, for example, assist clinicians in decision-making and enable patients to have a better understanding of their medical history. Significant amounts of patient data are stored in Electronic Health Records (EHRs), making EHR QA an important research area. In EHR QA, the answer is obtained from the patient's medical record. Because of the differences in data format and modality, this differs greatly from other medical QA tasks that employ medical websites or scientific papers to retrieve answers, making it critical to research EHR question answering.
Objective: This study aimed to provide a methodological review of existing works on QA over EHRs. The objectives of this study were to (i) identify the existing EHR QA datasets and analyze them, (ii) study the state-of-the-art methodologies used in this task, (iii) compare the different evaluation metrics used by these state-of-the-art models, and finally (iv) elicit the various challenges and the ongoing issues in EHR QA.
Methods: We searched for articles from January 1st, 2005 to September 30th, 2023 in four digital sources including Google Scholar, ACL Anthology, ACM Digital Library, and PubMed to collect relevant publications on EHR QA. Our systematic screening process followed PRISMA guidelines. 4111 papers were identified for our study, and after screening based on our inclusion criteria, we obtained a total of 47 papers for further study. The selected studies were then classified into two non-mutually exclusive categories depending on their scope: 'EHR QA datasets' and 'EHR QA Models'.
Results: A systematic screening process obtained a total of 47 papers on EHR QA for final review. Out of the 47 papers, 25 papers were about EHR QA datasets, and 37 papers were about EHR QA models. It was observed that QA on EHRs is relatively new and unexplored. Most of the works are fairly recent. Also, it was observed that emrQA is by far the most popular EHR QA dataset, both in terms of citations and usage in other papers. We have classified the EHR QA datasets based on their modality, and we have inferred that MIMIC-III and the n2c2 datasets are the most popular EHR database/corpus used in EHR QA. Furthermore, we identified the different models used in EHR QA along with the evaluation metrics used for these models.
Conclusions: EHR QA research faces multiple challenges such as the limited availability of clinical annotations, concept normalization in EHR QA, as well as challenges faced in generating realistic EHR QA datasets. There are still many gaps in research that motivate further work. This study will assist future researchers in focusing on areas of EHR QA that have possible future research directions.
Keywords: Medical Question Answering; Electronic Health Records (EHRs); Electronic Medical Records (EMRs); relational databases; Knowledge Graphs

[1] Corresponding author. Address: School of Biomedical Informatics, The University of Texas Health Science Center at Houston, 7000 Fannin St #600, Houston, TX 77030, USA
Email addresses: jayetri.bardhan@ufl.edu (Jayetri Bardhan), kirk.roberts@uth.tmc.edu (Kirk Roberts), daisyw@cise.ufl.edu (Daisy Zhe Wang)

Introduction

Motivation

Medical QA may use biomedical journals, internet articles, as well as patient-specific data such as that stored in the Electronic Health Record (EHR) for QA. While there has been a lot of work in medical QA [CITATION Dem20 \l 1033 \m Rob14 \m Cai11 \m Lee85 \m Wan20]12345, most of these works do not help to answer patient-specific questions. In patient-specific QA, the answer is obtained from the patient's medical record (i.e., the EHR). This differs from other medical QA tasks due to linguistic issues (for example, EHR notes are very different in terminology, grammar, style, and structure from biomedical articles) and privacy limitations (for example, most biomedical articles have a publicly available abstract while there are laws in most countries limiting the sharing of patient records). Additionally, patient-specific QA also prevents the use of many common QA techniques (such as aggregating answers from different biomedical articles to give weight to a consensus opinion). All this merits the review of EHR QA separate from other medical QA approaches to properly scope its data and methods. In this review paper, our aim is to discuss all the recent approaches and methodologies used for question answering on electronic health records. There have been some reviews on medical question answering [CITATION Mut21 \l 1033 \m Ath]67, but none of the previous review papers have focused solely on EHR QA. To the best of our knowledge, this is the first work that does a scoping review of QA on EHRs and examines the various datasets and methodologies utilized in EHR QA. There are several aspects of EHR QA that merit analysis of scope.

One such aspect is data modality and the variety of methodological approaches available for EHR QA. The methodological approach used is determined by the format of the EHR data. EHRs contain structured data and unstructured data. Structured EHR data is based on standardized terminologies and ontologies and is often available in the form of relational databases. On the other hand, unstructured EHR data has minimal standardization and includes data types such as textual notes and clinical imaging studies. Two kinds of approaches are used for QA on structured EHR data. In the first approach [CITATION Wan20 \l 1033]5, the natural language questions are converted into structured queries (such as SQL). These queries are used to retrieve answers from the database. In the second approach [CITATION Rag17 \l 1033]8, the structured EHR tables are converted into knowledge graphs, following which the natural language questions are converted into graph queries (such as SPARQL) in order to extract answers from the database. QA on unstructured clinical EHR notes is mostly performed as a reading comprehension task, where given a question and clinical notes as context, a span of text from the notes is returned as the answer. There can also be multimodal EHR QA which can use both structured and unstructured EHR data for QA. The aim of this study is to identify the studies that use EHR QA. We have further narrowed our search to EHR QA studies that use natural language processing (NLP) techniques on the questions, but may or may not use NLP on the answers. We have excluded studies in which questions are asked about images (e.g., radiology scans) as these questions and datasets have an entirely different focus. While QA over medical images is also a critical area of research, focusing a systematic review specifically on QA over EHR text (i.e. structured EHR and unstructured EHR containing textual information) allows for a more detailed, manageable, and methodologically consistent study. This focused approach can yield deeper insights and more practical recommendations for improving QA systems on structured and unstructured data in healthcare settings.

The second aspect of EHR QA is the access to raw medical data. Due to privacy restrictions on clinical data, the replication and sharing of methods have been reduced compared to QA in other domains. This has led to the emphasis on sharable EHR datasets on which QA benchmarks can be

made. MIMIC-IV [CITATION Joh23 \l 1033]9 and the eICU Database [CITATION Pol18 \l 1033]10 are large publicly available EHR databases for patients admitted to intensive care units. The MIMIC-III [CITATION Joh16 \l 1033]11 database provides the foundation for a lot of the existing QA studies on EHRs. MIMIC-IV introduced in the year 2020, is a recent update to the MIMIC-III database. Finally, the n2c2 datasets (previously known as i2b2 datasets) is another repository of clinical notes, which have been used by the clinical QA community to develop EHR QA datasets.

Another aspect that warrants a scoping review of EHR QA is to study its different applications, including information extraction, cohort selection, and risk score calculation. For instance, Datta et al. (2020) [CITATION Dat22 \l 1033]12 used a two-turn question answering approach to extract spatial relations from radiology reports. Similarly, Xiong et al. (2021) [CITATION Xio21 \l 1033]13 used a QA approach with the help of a machine reading comprehension framework for cohort selection, where every selection criterion is converted into questions using simple rules. For example, the selection criteria "ALCOHOL-ABUSE" is converted to the question - "Current alcohol use over weekly recommended limits?". Following this, using state-of-the-art machine-reading comprehension models like BERT [CITATION Devdf \l 1033]14, BiDAF [CITATION Seo03 \l 1033]15, BioBERT [CITATION Lee20 \l 1033]16, NCBI-BERT, RoBERTa [CITATION Liu08 \l 1033]17, and BIMPM [CITATION Wan14 \l 1033]18, are used to match question and passage pairs in order to select cohorts. Furthermore, Liang et al. (2022) [CITATION Lia42 \l 1033]19 demonstrates that QA over EHR data can improve risk score calculation.

Lastly, EHR QA systems face a variety of challenges ranging from parsing natural language questions to retrieving answers. In the case of structured data, the natural language question needs to be parsed and converted to a structured query which can be used to query the database. Medical terms from the queries, such as "blood pressure" and "leukemia," must be normalized into standard ontologies. Clinical text frequently uses acronyms for medical concepts. These abbreviations are often ambiguous (for example, "pt" can refer to the patient or physical therapy)[CITATION New21 \l 1033]20 and so must be identified and standardized by the QA system before querying over the EHR database or clinical data. These problems are exacerbated by the fact that the standard NLP approaches to such issues require large amounts of labeled data from the domain of interest. Few such labeled EHR datasets exist. This is because annotating EHR QA datasets requires clinical expertise and is time-consuming. Existing general-domain QA systems provide erroneous results when they are not trained on clinical QA datasets. Additionally, the majority of the data found in EHRs is complex and contains both missing and inconsistent information [CITATION Wel35 \l 1033 \m Han21]2122, which adds to the difficulty of performing QA on EHRs. In the Discussion section, we have provided more detailed explanations of the various challenges of using QA on EHRs.

The wide variety of challenges and barriers discussed above motivates the need for a systematic scoping review of EHR QA literature. This paper identifies the articles that fall under the scope of EHR QA, identifies the difficult challenges faced in the task, then enumerates both the data sources and QA methods that have been used to overcome such challenges. Finally, this paper also highlights the open issues in this field that demand future work in EHR QA.

Template-based dataset generation

Prior to diving into the methodology and results of this review, it is helpful to introduce a common semi-automated approach for building EHR QA datasets as essentially all large EHR QA datasets utilize this approach. This also impacts the screening process described in the next section. While other methods, such as semantic parsing with grammar-based techniques, exist for generating EHR QA datasets [CITATION Rob98 \l 1033]23[CITATION Rob48 \l 1033]24, template-based dataset generation remains the most widely used approach. In general, large EHR QA datasets are often

required to increase the performance of EHR QA models. However, the creation of these datasets necessitates subject expertise. The slot-filling approach to generate template-based datasets is a semi-automated process, and hence very popular. The majority of the EHR QA datasets are template-based [CITATION Rag17 \m Wan20 \m Bar17 \m Pam58 \m Moo17 \l 1033]85252627. The steps to construct template-based QA datasets are illustrated using a flowchart in .

To minimize the need for clinical experts' involvement in the dataset generation process, existing annotations of other non-QA clinical tasks (such as entity recognition and relations learning) are used for generating EHR question-answer pairs. The existing clinical annotations are used as proxy-expert in the dataset generation process [CITATION Pam58 \l 1033]26. In the first step, template questions containing placeholders (in place of entities) are constructed. An example of a question template is – "Has this patient ever been treated with |medication|?". Here, |medication|, |problem|, and |treatment| are some commonly used placeholders. These placeholders in the questions are then slot-filled to obtain QA pairs using the entities in the EHR data and database schema (for structured EHR database) with the help of the existing annotations from the clinical NLP datasets. So, in a question template such as - "Has this patient ever been treated with |medication|?", entities like "insulin" and "Tylenol" from the EHR database/clinical notes (sharing the same entity type as |medication|) are slot-filled in the question template to obtain questions such as - "Has this patient ever been treated with insulin?" and "Has this patient ever been treated with Tylenol?". Following this approach, the RxWhyQA [CITATION Moo17 \l 1033]27 and DrugEHRQA [CITATION Bar17 \l 1033]25 datasets use the existing annotations from the 2018 National NLP Clinical Challenges (n2c2) corpus, and the emrQA and emrKBQA datasets use annotations from six clinical tasks from the n2c2 repository [CITATION Uzu10 \m Uzu101 \l 1033 \m Uzu08 \m Uzu09 \m Stu15 \m Uzu12]282930313233

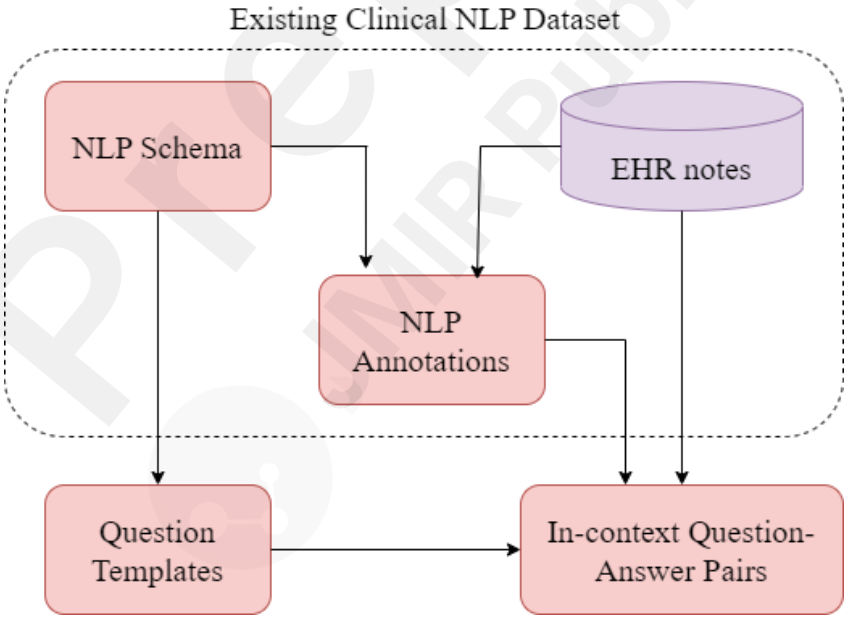


Figure 1. Flowchart showing the process of template-based dataset generation. The dotted boundary shows the existing non-QA NLP dataset along with the EHR data. Question templates (and logical form templates) are constructed based on the EHR data. Clinical expert annotations of non-QA tasks are used to slot-fill placeholders in question templates and generate question-answer pairs.

Some EHR QA datasets, such as emrQA and emrKBQA, have used logical form templates in their template-based generation methods. Logical form templates are predefined structured representations of questions that provide a human-comprehensible symbolic representation, linking questions to answers. These are used to map EHR schema or ontology to represent relations in the questions.

While generating these datasets, logical form templates are annotated by clinical experts for different question templates. For example, for the question template – “What is the dosage of |medication| ?”, the annotated logical form template for emrQA is – “MedicationEvent (|medication|) [dosage=x]”. If more than one question template maps to the same logical form template, then they are considered paraphrases of each other. In the emrQA dataset, clinical expert annotations of non-QA tasks such as entity recognition, relation learning, coreference, and medication challenge annotations (in the n2c2 repository) were used to slot-fill placeholders in question and logical form templates, which in turn were used to generate answers. This is shown in Figure 1. For example, the medication challenge in the n2c2 repository has annotations for medication and their corresponding dosage (for example, medication=Nitroglycerin, dosage=40mg). This was used to generate instances of the question “What is the dosage of |medication|?", along with instances of its corresponding logical form MedicationEvent(|medication|)[dosage=x]. The dosage value, i.e. 40mg is the answer to the question. Similarly, the heart disease challenge dataset contains temporal information, was used to derive temporal-reasoning related question-answer pairs. The emrKBQA dataset used the same question templates and logical form templates of emrQA, which were then slot-filled using entities from the MIMIC-III KB [CITATION Joh16 \l 1033]11. The answers of the emrKBQA dataset are present in the table cells of the MIMIC-III KB. The entity types used in the placeholders are test, problem, treatment, medication, and mode. So far, the slot-filling QA dataset generation process has proven to be the most common method of generating EHR QA datasets. This is because, while some manual annotation from domain experts is necessary, most of the process is automated.

Methods

Search Process

This study aims to review existing research on question answering over electronic health records. This includes papers on EHR QA datasets, QA models, and various approaches proposed over the years. We included papers related to question answering in the clinical domain, specifically in EHRs. Papers in which EHRs are not used have been excluded. In this review, we define 'Question Answering' (QA) as the task of automatically providing precise, relevant answers to user queries from EHR data. This involves understanding and processing EHR data to extract and deliver specific information. We distinguish QA from broader interactive systems such as conversational agents, chatbots, and general information retrieval systems, which may involve multi-turn dialogue and do not focus solely on providing direct answers to questions. The scope of this review is specifically on structured and unstructured data within EHRs due to the unique challenges and methodologies involved in processing natural language and structured information. While medical images (e.g., CT, MRI, X-ray) and physical signals (e.g., ECG, PPG) are critical components of EHRs, the techniques required to analyze these data types differ significantly from those used in structured and unstructured EHR data. Thus, studies focused on these modalities are excluded to maintain a clear and manageable focus on text-focused QA over structured and unstructured EHR data. We have fulfilled all PRISMA scoping review requirements and have attached a completed copy of the PRISMA checklist in Multimedia Appendix 1. The flowchart for conducting this study is shown in Figure 2.

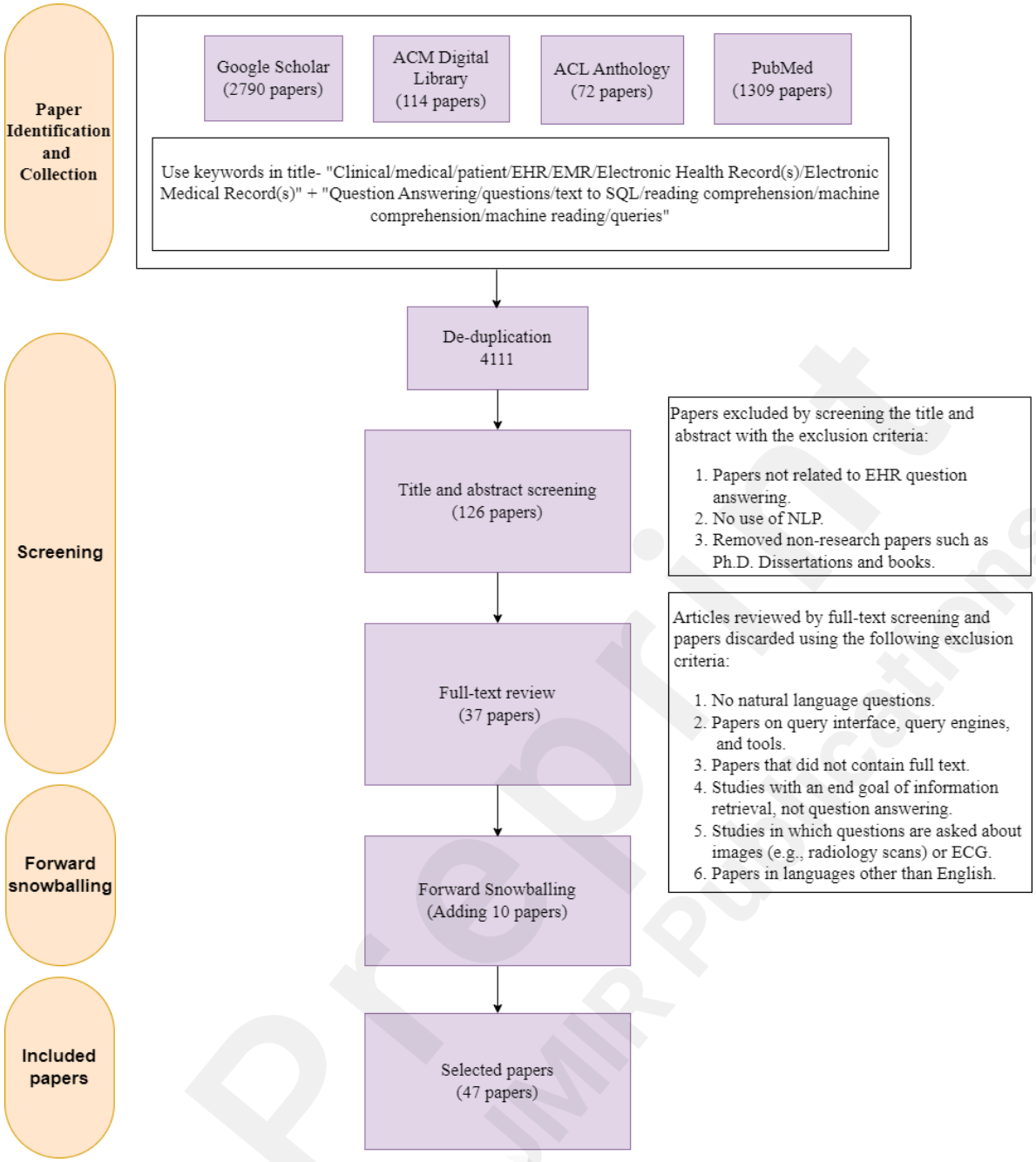


Figure 2. PRISMA diagram for study on QA over EHRs.

Each of the data sources has been queried to search for papers with the title having at least one of the following keywords: "clinical", "medical", "patient", "EHR", EMR", "Electronic Health Record(s)", or "Electronic Medical Record(s)". This should be used in combination with one or more of the keywords: "question answering", "questions", "text to SQL", "reading comprehension", “machine comprehension", "machine reading", or "queries". The search was limited to the period from January 1st, 2005 to September 30th, 2023, to review only recent works. In this record identification and collection step, a total of 4285 papers were collected (i.e. 2790 from Google Scholar, 114 papers from ACM Digital Library, 72 papers from ACL Anthology, and 1309 papers from PubMed). Following this, we removed the duplicate papers and obtained 4111 papers.

Screening Process

We used a two-step screening process. The first step involved reading the abstracts and titles of all the papers and including only papers that were about EHR question answering. This step obtained

126 papers. We also removed many irrelevant papers that focused on "clinical questions" and "patient questions" but did not use NLP. We also removed non-research papers (such as PhD dissertations and books).

In the final stage of screening, a full-text review was used to screen the papers further. Papers that were about query engines and tools and which did not use natural language questions were removed. We excluded papers in languages other than English. We also removed papers that just had an abstract and did not contain full text. There were some papers that were about information retrieval systems, and not specifically QA. These were also excluded. Furthermore, we have excluded studies in which questions are asked about images or ECG as these studies have an entirely different focus. At the end of this step, we obtained a total of 37 papers. After the two-stage screening process, we performed forward snowballing, adding ten more papers after a full-text review that cited the 37 previously included papers, according to Google Scholar. Finally, we obtained a total of 47 papers at the end of our study. Then, we conducted an in-depth review of the final 47 papers which are discussed in the Results section.

For this study, all the authors (JB, KR, DZW) jointly made the rules for inclusion and exclusion criteria, that were used during paper collection and screening process. Based on the rules decided, JB collected the papers and worked on the overall screening process. For papers that were borderline for inclusion were independently screened by KR and then were resolved after discussion. The final list made during the full-text review process was again independently screened and reviewed by JB and KR, with conflicts resolved after discussion.

Results

Classification of Selected Papers

This section presents the findings of our study about existing EHR QA papers.

lists our final list of selected publications post-screening and then classified the papers based on their scope - ’EHR QA Datasets’ and ’EHR QA Models’. We have further classified the studies on EHR QA models based on their function in the QA pipeline. ’Full QA’ denotes the papers on EHR QA models which are about end-to-end EHR QA systems. In the remaining part of the paper, we have provided our in-depth analysis of studies on QA using EHRs. In Multimedia Appendix 2, we have summarized our final list of selected papers.

Table 1. List of included papers in the systematic review

Type of study	References
EHR QA datasets	[CITATION Wan20 \m Rag17 \m Bar17 \m Pam58 \m Moo17 \m Rob98 \m Rob48 \l 1033]582526272324 [CITATION Predf \l 1033 \m Jun13 \m Gyu22 \m Son72 \m Oli10 \m Kim22 \m Wan18]34353637383940 [CITATION Pardf \l 1033 \m Yue00 \m Leh18 \m Son18 \m Misv1 \m Xia10 \m Palqm]41424344454647 [CITATION Ham49 \l 1033 \m Fle89 \m Mah77 \m Dad73]48495051
EHR QA models	

	Question generation	[CITATION Leh18 \l 1033]43
	Question paraphrasing	[CITATION Son03 \l 1033 \m Moo13 \m Son85]525354
	Question classification	[CITATION Pat12 \l 1033 \m Rob97]5556
	Full QA	[CITATION Wan20 \l 1033 \m Rag17 \m Bar17 \m Pam58 \m Moo17 \m Oli10 \m Kim22 \m Wan18 \m Pardf]5825262738394041 [CITATION Yue00 \l 1033 \m Ham49 \m Fle89 \m Mah77 \m Dad73 \m Rob17 \m Sch20 \m Raw52]4248495051575859 [CITATION Rawdf \l 1033 \m Bae03 \m Pan21 \m Son06 \m Wen20 \m Son79 \m Mai36 \m Moo88]6061626364656667 [CITATION LiY23 \l 1033 \m Yan22 \m Kan55 \m Tar98 \m Sar23 \m Leh91]686970717273

Figure 3 illustrates the number of publications on EHR QA over the years. From Figure 3, it can be observed that this is a relatively very new field and most of the publications in this domain are fairly recent. Note that since this systematic review is conducted based on studies published before September 30th, 2023, hence the number of studies shown for the year 2023 is recorded only for a period of 9 months. In the following subsections, we discuss our findings on existing EHR QA datasets, the various models used for questioning over EHRs; and also, the different evaluation metrics used.

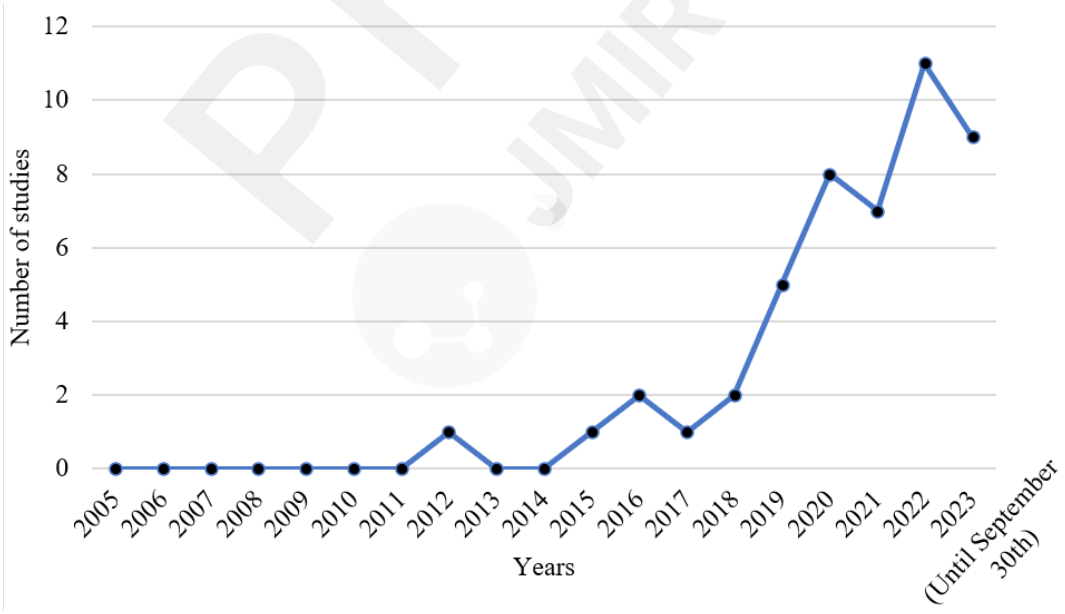


Figure 3. Number of studies on EHR QA over the years.

Datasets

Dataset Classification and Analysis

Table 2 displays the total number of citations of all the EHR QA datasets (as noted on September 30th, 2023). It also lists the number of studies included in our review that have used these datasets. Moreover, Table 2 classifies the EHR-QA based on the accessibility of the datasets. Please note that the information presented here is based on the data available at the time of submission. We can observe from the figures that emrQA [CITATION Pam58 \l 1033]26 is the most popular out of all the other EHR QA datasets. This is likely due to emrQA's size (1,295,814 question-logical forms and 455,837 question-answer pairs) and similarity to the SQuAD-QA format.

Table 2. Popularity and accessibility of EHR QA datasets

Datasets	No. of citations	No. of studies on EHR QA using the datasets	Publicly available?
emrQA [CITATION Pam58 \l 1033]26	151	11	Yes
MIMICSQL [CITATION Wan20 \l 1033]5	51	3	Yes
Yue et al. (2020) [CITATION Xia10 \l 1033]46	40	0	No
MIMICSPARQL* [CITATION Pardf \l 1033]41	27	2	Yes
Yue et al. (2021) [CITATION Yue00 \l 1033]42	18	0	Yes
Roberts et al. (2016) [CITATION Rob98 \l 1033]23	18	3	No
emrKBQA [CITATION Rag17 \l 1033]8	15	0	No
Raghavan et al. (2018) [CITATION Predf \l 1033]34	13	0	No
Roberts et al. (2015) [CITATION Rob48 \l 1033]24	10	1	No
Soni et al. (2019) [CITATION Son18 \l 1033]44	7	3	No
Fan et al. (2019) [CITATION Jun13 \l 1033]35	7	0	Yes
DrugEHRQA [CITATION Bar17 \l 1033]25	5	0	Yes
DiSCQ [CITATION Leh18 \l 1033]43	6	0	Yes
Oliveira et al. (2021)	3	0	No

[CITATION Oli10 \l 1033]38			
RadQA [CITATION Son72 \l 1033]37	3	1	Yes
EHRSQL [CITATION Gyu22 \l 1033]36	3	0	Yes
Kim et al. (2022) [CITATION Kim22 \l 1033]39	2	0	Yes
ClinicalKBQA [CITATION Wan18 \l 1033]40	2	0	No
Hamidi et al. (2023) [CITATION Ham49 \l 1033]48	1	0	No
MedAlign [CITATION Fle89 \l 1033]49	1	0	No
RxWhyQA [CITATION Moo17 \l 1033]27	0	0	Yes
Mishra et al. (2021) [CITATION Misv1 \l 1033]45	0	0	No
CLIFT [CITATION Palqm \l 1033]47	0	0	No
Mahbub et al. (2023) [CITATION Mah77 \l 1033]50	0	0	No
Dada et al. (2023) [CITATION Dad73 \l 1033]51	0	0	No

The classification of EHR QA datasets is shown in Figure 4. EHR QA datasets can be unimodal or multimodal. Unimodal EHR QA datasets are based on question answering over one modality which can be in the form of structured EHR data or unstructured EHR clinical notes. Multimodal EHR QA datasets use both modalities for QA over EHRs. The DrugEHRQA [CITATION Bar17 \l 1033]25 and MedAlign [CITATION Fle89 \l 1033]49 datasets are examples of multimodal EHR QA datasets that use structured and unstructured EHR data for QA. Figure 5 shows the size and modalities of the different EHR QA datasets.

It is to be noted that the dataset introduced in Mishra et al. (2021) [CITATION Misv1 \l 1033]45 uses six key questions (as can be observed from Figure 5), i.e. the same six questions have been re-used for all the articles. Multimedia Appendix 3 summarizes the existing EHR QA datasets. It should be noted that in Multimedia Appendix 3, "Database/Corpus" refers to the EHR database or clinical annotations from which the QA datasets are based on. These EHR databases or corpora contain answers to the questions. From the table in Multimedia Appendix 3, we can infer that most of the EHR QA datasets on structured EHR data use the MIMIC-III database [CITATION Wan20 \l 1033 \m Rag17 \m Gyu22 \m Kim22 \m Pardf]58363941, while most of the QA datasets on unstructured data use the n2c2 repository [CITATION Pam58 \l 1033 \m Moo17 \m Jun13]262735 or the clinical notes of MIMIC-III [CITATION Son72 \l 1033 \m Yue00 \m Leh18 \m Misv1 \m Xia10 \m Palqm \m Ham49]37424345464748.

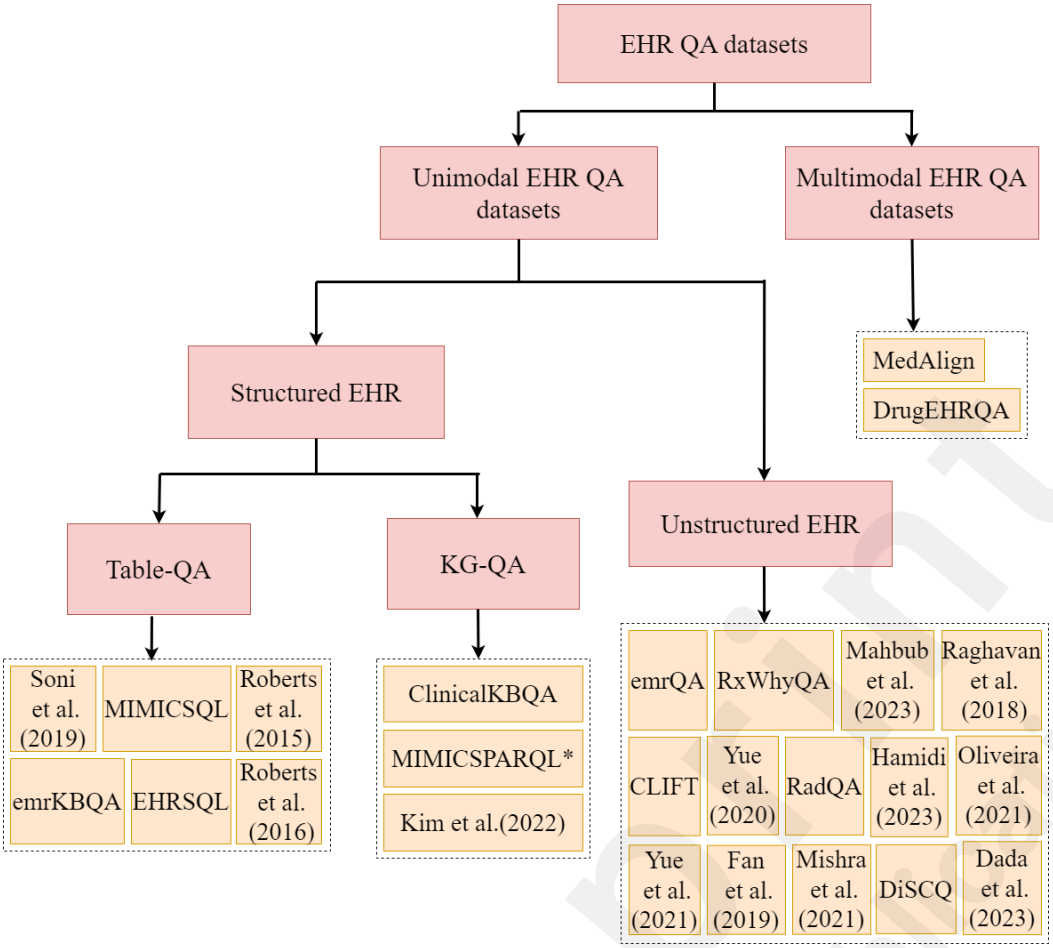


Figure 4. Classification of EHR QA datasets based on modality.

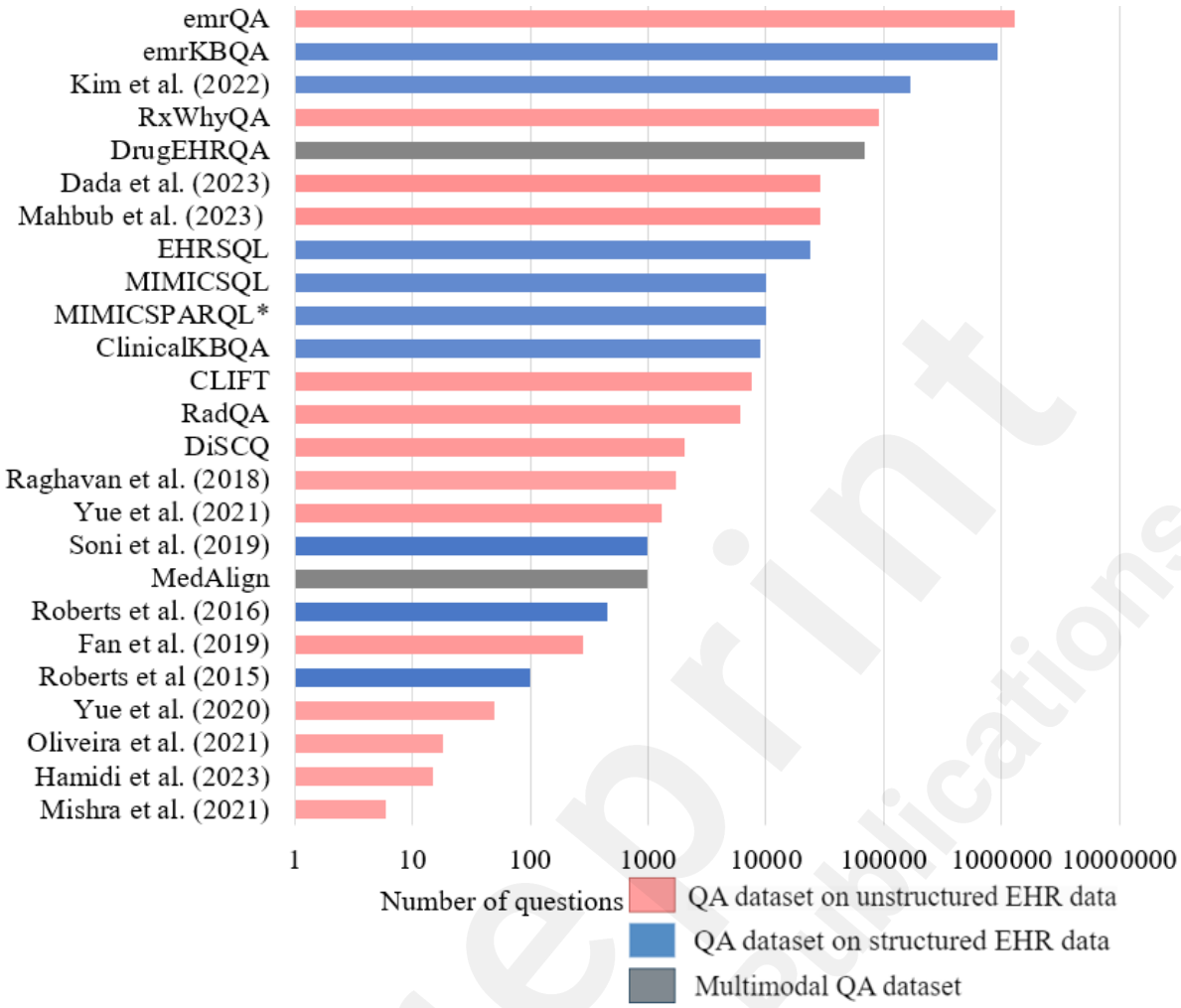


Figure 5. Plot shows the total number of questions included in various EHR QA datasets and classifies them into unstructured, structured, and multimodal EHR QA datasets.

The following sections describe the QA datasets based on unimodal (structured or unstructured) and multimodal EHR data in detail.

QA datasets based on unstructured EHR data

Unstructured free text EHR data comprises discharge summaries, radiology reports, lab reports, medical images, progress notes, and many more note types. It accounts for roughly 80% of all EHR data [CITATION Tsu21 \l 1033]74. One way to make use of this is to create a QA system that can extract answers from unstructured EHR data. Most of the QA datasets on unstructured clinical data are designed for the task of machine comprehension. Given clinical notes (containing patient information) and natural language questions, the objective of these tasks is to retrieve a span of text from the clinical notes as the answer.

emrQA [CITATION Pam58 \l 1033]26, the most popular among the EHR QA datasets, contains 455,837 question-answer samples along with 1,295,814 question-logical form pairs. It relies on expert annotated n2c2 datasets [CITATION Uzu10 \l 1033 \m Uzu101 \m Uzu08 \m Uzu09 \m Stu15 \m Uzu12]282930313233. A semi-automatic template-based process was used to generate the dataset. From Figure 5, we can observe that the emrQA is the largest EHR QA dataset overall.

In spite of emrQA's popularity, it has some flaws. The emrQA dataset has attempted to simulate clinicians' questions using pre-defined templates and generating QA datasets by slot-filling with entities. As a result, the questions in the emrQA dataset are not very realistic or relevant to the medical community. They are also highly repetitive. For example, it was shown in Yue et al. (2020) [CITATION Xia10 \l 1033]46 that the same model performance was obtained by sampling 5-20% of the dataset as with the entire dataset. This makes it necessary to create datasets that are more realistic and closer to real physicians' questions. Later, Yue et al. (2021) [CITATION Yue00 \l 1033]42 developed 975 human-verified questions along with 312 human-generated questions based on 36 discharge summaries from MIMIC-III's clinical notes. After randomly sampling 100 questions individually the 975 human-verified questions and 312 human-generated questions, it was learned that 96% of the [CITATION Yue00 \l 1033]42's human-verified questions were obtained from the emrQA's templates, and 54% of the human-generated questions of Yue et al. (2021) [CITATION Yue00 \l 1033]42 used the same templates from emrQA.

The RxWhyQA dataset [CITATION Moo17 \l 1033]27 and Fan et al. (2019) dataset [CITATION Jun13 \l 1033]35 have reasoning-based questions. The RxWhyQA dataset contains a combination of reasoning-based unanswerable and multi-answer questions. Similar to the emrQA dataset, RxWhyQA is also a template-based dataset, and hence not very realistic. This made it necessary to create datasets that are more realistic and closer to real physicians' questions. The DiSCQ dataset [CITATION Leh18 \l 1033]43 was created to address this issue and included questions about clinically relevant problems by gathering questions that clinicians could ask. It includes 2029 questions and over 1000 triggers based on MIMIC-III discharge reports. Each question in the DiSCQ dataset is paired with a segment of text that prompted the query. While annotating the DiSCQ dataset, the annotators read the discharge summaries and made a note of all questions that might be useful to the patient, while also recording the span of text that prompted the question.

The majority of the QA on unstructured EHR datasets are based on discharge summaries [CITATION Pam58 \l 1033 \m Moo17 \m Jun13 \m Leh18 \m Misv1 \m Yue05]262735434575. RadQA [CITATION Son72 \l 1033]37 and Dada et al. (2023) [CITATION Dad73 \l 1033]51 are the only two QA dataset that uses radiology reports for question answering. The types of questions used in the EHR QA datasets vary greatly from one another. emrQA covers different types of questions including Factual ("What"/"Show me"), reasoning ("How"/"Why"), and class prediction ("Is"/"has"). But the distribution of questions for the emrQA dataset is skewed, i.e., a majority of the questions in the emrQA dataset start with "what". In comparison, the authors of RadQA claim that the questions in their dataset are more evenly distributed than emrQA. The RxWhyQA dataset [CITATION Moo17 \l 1033]27 and Fan et al.(2019) [CITATION Jun13 \l 1033]35 are reasoning-based questions and hence their questions have "why-cues". Raghavan et al. (2018) [CITATION Predf \l 1033]34 predominantly has temporal questions along with questions on presence/absence (i.e. "yes" or "no" questions) as well as questions on medications, tests, and procedures. Mishra et al. [CITATION Misv1 \l 1033]45 on the other hand restricts itself to diagnosis-related questions. Table 3 compares some of the EHR QA datasets using unstructured EHR data for QA. Note that in Table 3, the length of questions and articles are expressed in terms of the number of tokens. Out of the fourteen QA datasets on unstructured EHR notes, only four of them (RadQA [CITATION Son72 \l 1033]37, RxWhyQA [CITATION Moo17 \l 1033]27, Hamidi et al. (2023) [CITATION Ham49 \l 1033]48, and Dada et al. (2023) [CITATION Dad73 \l 1033]51) contain unanswered questions.

Table 3. Comparison of different EHR QA datasets on unstructured data.

Dataset	Database/ Corpus	Mode of dataset generation	Total questions	Unanswered questions	Average question length	Total articles	Average article length
emrQA [CITATION Pam58 \l 1033][26]	n2c2 annotations (mostly discharge summaries)	Semi-automatically generated	1,295,814	0	8.6	2425	3825
RxWhyQA [CITATION Moo17 \l 1033][27]	MIMIC-III (discharge summaries)	Automatically derived from the n2c2 2018 ADEs NLP challenge	96,939	46,278	-	505	-
Raghavan et al. (2018) [CITATION Predf \l 1033][34]	Cleveland Clinic (medical records)	Human-generated (Medical students)	1747	0	-	71	-
Fan et al. (2019) [CITATION Jun13 \l 1033][35]	2010 n2c2/VA NLP challenge (discharge summaries)	Human-generated (Author)	245	0	-	138	-
RadQA [CITATION Son72 \l 1033][37]	MIMIC-III (radiology reports)	Human-generated (Physicians)	6148	1754	8.56	1009	274.49
Oliveira et al. (2021) [CITATION Oli10 \l 1033][38]	SemClinBr corpus (Portuguese nursing and medical notes)	Human-generated (Author)	18	0	-	9	-
Yue et al. (2021) [CITATION Yue00 \l 1033 \m Yue05][4275]	MIMIC-III (Clinical notes)	Trained question generation model paired with a human-in-the-loop	1287	0	8.7	36	2644
DiSCQ [CITATION Leh18 \l 1033][43]	MIMIC-III (discharge summaries)	Human-generated (medical experts)	2029	0	4.4	114	1481

Mishra et al. (2021) [CITATION Misv1 \l 1033]45	MIMIC-III (discharge summaries)	Semi-automatically generated	6 questions/ article	-	-	568	-
Yue et al. (2020) [CITATION Xia10 \l 1033]46	MIMIC-III (Clinical notes)	Human-generated (medical experts)	50	0	-	-	-
CLIFT [CITATION Palqm \l 1033]47	MIMIC-III	Validated by human experts	7500	0	6.42, 8.31, 7.61, 7.19, and 8.40 for Smoke, Heart, Medication, Obesity, and Cancer datasets	-	217.33, 234.18, 215.49, 212.88, and 210.16 for Smoke, Heart, Medication, Obesity, and Cancer datasets respectively
Hamidi et al. (2023) [CITATION Ham49 \l 1033]48	MIMIC-III (Clinical notes)	Human-generated	15	5	-	-	-
Mahbub et et al. (2023) [CITATION Mah77 \l 1033]50	VA Corporate Data Warehouse (CDW) (Clinical progress notes)	Combination of manual exploration and rule-based NLP methods	28855	-	6.22	2336	1003.98
Dada et al. (2023) [CITATION Dad73 \l 1033]51	Radiology reports related to brain CT scans	Human-generated (medical student assistants)	29,273	- (Yes)	-	1223	-

QA datasets based on structured EHR data

EHR tables contain patient information such as diagnoses, medications prescribed, treatments, procedures recommended, lab results details, and so on. It also includes a lot of temporal information, such as the date of admission, the date of discharge, and the duration of certain medications. The goal of QA tasks over structured databases is to translate the user's natural language question into a form that can be used to query the database.

The QA task on structured EHRs can be classified into two types based on the two most common

forms of structured data: relational databases and knowledge graphs. The first type of QA task entails converting natural language questions into SQL (or logical form) queries that can be used to query the database. In the other type of approach, the EHR data exists in the form of knowledge graphs containing patient information, and the natural language questions are often converted into SPARQL queries to retrieve the answer. MIMICSQL, emrKBQA, and EHRSQL are examples of datasets that use table-based QA approach, whereas datasets like ClinicalKBQA and MIMIC-SPARQL* use KG-based QA approach.

MIMICSQL [CITATION Wan20 \l 1033]5 is a large dataset used for question-to-SQL query generation tasks in the clinical domain. The MIMICSQL dataset is based on the tables of the MIMIC-III database. emrKBQA [CITATION Rag17 \l 1033]8 is the counterpart of the emrQA dataset, for question answering on structured EHRs. It is the largest QA dataset on structured EHR data (shown in Figure 5) and contains 940,000 samples of questions, logical forms, and answers. EHRSQL [CITATION Gyu22 \l 1033]36 is a text-to-SQL dataset for two publicly available EHR databases - MIMIC-III [CITATION Joh16 \l 1033]11 and eICU [CITATION Pol18 \l 1033]10. It is the only QA dataset on structured EHR data that contains unanswerable questions. Other QA datasets for structured EHR databases include MIMIC-SPARQL* [CITATION Pardf \l 1033]41 and ClinicalKBQA [CITATION Wan18 \l 1033]40. However, unlike previous table-based QA datasets, these are knowledge graph-based QA datasets.

The MIMICSQL dataset [CITATION Wan20 \l 1033]5 was created by making changes to the MIMIC-III database's original schema. Nine tables from the MIMIC-III database were merged into five tables in order to simplify the data structure. The derived tables and schemas were not the same as those found in actual hospitals and nursing homes. As a result, a model trained on the MIMICSQL dataset will not be able to generalize to a real-world hospital setting. To address this issue, Park et al. (2021) [CITATION Pardf \l 1033]41 introduced two new datasets - a graph-based EHR QA dataset (MIMIC-SPARQL*) and a table-based EHRQA dataset (MIMICSQL*). This was done to improve the analysis of EHR QA systems and to investigate the performance of each of these datasets. MIMICSQL [CITATION Wan20 \l 1033]5 was modified to create MIMICSQL* in order to comply with the original MIMIC-III database schema [CITATION Joh16 \l 1033]11. The graph counterpart of the MIMICSQL* dataset is MIMIC-SPARQL*. Figure 6 compares the two datasets - MIMICSQL and MIMICSPARQL* based on the length of the questions and the length of SQL/SPARQL queries.

Wang et al. (2021) [CITATION Wan18 \l 1033]40 generated a clinical knowledge graph (ClinicalKB) with the help of clinical notes of n2c2 annotations and linked different patient information in order to perform knowledge base QA. The ClinicalKB links clinical notes of patients, allowing questions about different patients to be answered. At the same time, ClinicalKB contains clinical notes which allow questions not present in the database.

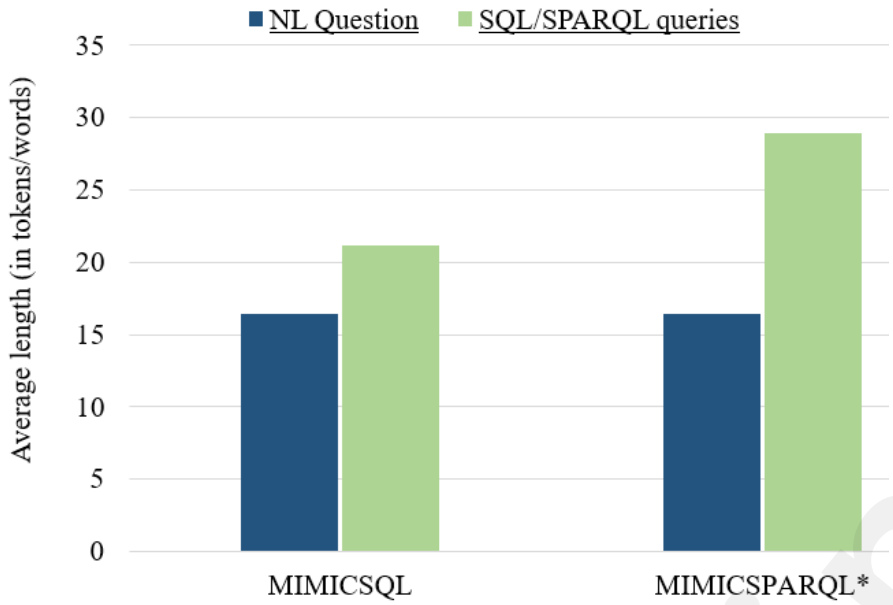


Figure 6. Average length of questions and SQL/SPARQL queries (in tokens or words) for MIMICSQL and MIMICSPARQL* datasets.

Furthermore, Wang et al. (2021) [CITATION Wan18 \l 1033]40 generated the ClinicalKBQA dataset that can answer statistics-related questions about different patients as well as questions specific to individual patient records.

Roberts et al. [CITATION Rob98 \l 1033 \m Rob48]2324 and Soni et al. (2019) [CITATION Son18 \l 1033]44 introduced datasets where logical forms (based on lambda calculus expressions) were created for questions in order to perform QA on EHR data (known as semantic parsing). Roberts et al. [CITATION Rob98 \l 1033 \m Rob48]2324 generated a bottom-up grammar-based method that generates logical forms for question phrases. Soni et al. (2019) [CITATION Son18 \l 1033]44 constructed the question-logical form dataset with the help of Fast Healthcare Interoperability Resources (FHIR) server. Roberts et al. (2015) [CITATION Rob48 \l 1033]24 annotated logical forms for 100 EHR questions, and Roberts et al. (2016) [CITATION Rob98 \l 1033]23 extended this to 446 patient-related questions.

QA datasets based on multimodal EHR data

Multimodal question answering is QA over more than one modality. QA over more than one modality can help in seeking more accurate answers while taking advantage of more than one source for QA. DrugEHRQA [CITATION Bar17 \l 1033]25 is the first multimodal EHR QA dataset. It uses both structured tables of MIMIC-III and unstructured clinical notes for QA. The DrugEHRQA dataset is a template-based dataset containing medicine-related queries, its corresponding SQL queries for querying over multi-relational EHR tables, the retrieved answer from one or both modalities, as well as the final multi-modal answer. The MedAlign dataset [CITATION Fle89 \l 1033]49 also utilizes structured and unstructured EHR data for QA, but indirectly. The instructions and response pairs of the MedAlign dataset are based on XML markup documents that are derived from structured and unstructured EHR data.

Models and Approaches for QA on EHRs

This section describes the various QA models used in EHRs. QA tasks vary depending on the EHR modality since different information is found in different modalities. Most QA models on clinical notes use a machine reading comprehension approach (MRC), i.e. for a given question, the QA

model is trained to predict the span of text containing the answer from the clinical note [CITATION Pam58 \l 1033 \m Moo17 \m Oli10 \m Yue00 \m Ham49 \m Mah77 \m Dad73 \m Raw52 \m Son79 \m Mai36]26273842485051596566. For question answering over EHR tables, translating questions to SQL queries is one of the major approaches used to retrieve answers from the EHR tables [CITATION Wan20 \l 1033 \m Bae03 \m Pan21 \m Tar98]5616271. The other approach is to transform the EHR relational database into a knowledge graph and perform a knowledge-graph QA task [CITATION Kim22 \l 1033 \m Pardf \m Bae03]394161. Table 4 summarizes all the QA models (Full QA) used for EHRs.

Table 4. Summary of models for QA over EHRs.

Papers	Task	Answer Type	Dataset	Model
Pampari et al. (2018) [CITATION Pam58 \l 1033]26	MRC	Text span	emrQA	For question answering task: DrQA's document reader and a multi-class logistic regression model for predicting class. For question-to-logical form task: A sequence-to-sequence model is used with attention paradigm.
Moon et al. (2023) [CITATION Moo17 \l 1033]27	MRC	Text span	RxWhyQA	Clinical BERT model with incremental masking to generate multiple answers.
Oliveira et al. (2021) [CITATION Oli10 \l 1033]38	MRC	Text span	SQUAD dataset [CITATION Raj64 \l 1033]76 in Portuguese and another QA dataset developed in Portuguese from SemClinBr corpus	Used BioBERTpt (A deep contextual embedding model to support Portuguese language for biomedical data).
Yue et al. (2021) [CITATION Yue00 \l 1033]42	MRC	Text span	emrQA and [CITATION Yue00 \l 1033 \m Yue05]4275 as test set.	For question answering task: DrQA's DocReader and ClinicalBERT are used. For question generation task: Question Phrase Prediction (QPP) module is used along

				with the base question generation models (NQG, NQG++, and BERT-SQG).
Hamidi et al. (2023) [CITATION Ham49 \l 1033]48	MRC	Text span	QA dataset constructed based on TREC 2016 Clinical Decision Support Track [CITATION Robdf \l 1033]77	ChatGPT (versions 3.5 and 4), Google Bard, and Claude
Fleming et al. (2023) [CITATION Fle89 \l 1033]49	Multi-step refinement approach using standard prompt template	Response based on XML markup derived from EHR data	MedAlign dataset	Six language models – GPT-4 (32K tokens+ multi-step refinement), GPT-4 (32K tokens), GPT4 (2K tokens), Vicuña-13B (2K tokens), Vicuña-7B (2K tokens), and Vicuña-7B (2K tokens)
Mahbub et et al. (2023) [CITATION Mah77 \l 1033]50	MRC	Text span	QA dataset on Injection Drug use constructed	Baseline models: Four state-of-the-art pre-trained language models - BERT, BioBERT, BlueBERT, and ClinicalBERT used for QA. Modeling with transfer learning: Sequential learning and Adversarial learning
Dada et al. (2023) [CITATION Dad73 \l 1033]51	MRC	Text span	Reading comprehension question answering constructed based on radiology report	G-BERT and GM-BERT (G-BERT further pre-trained on German medical articles)
Roberts et al. (2017) [CITATION Rob17 \l 1033]57	question to logical form	Text span	Annotation of 446 questions in [CITATION Rob98 \l 1033]23	Hybrid semantic parsing method, uses rule-based methods along with a machine learning-based classifier.
Rawat et al.	MRC	Text span	Naranjo Scale	Employs multi-level

(2019) [CITATION Raw52 \l 1033]59			Questionnaire	attention layers along with local and global context while answering questions.
Rawat et al. (2020) [CITATION Rawdf \l 1033]60	MRC	Text span	emrQA and MADE-QA dataset	Multitask learning with BERT and ERNIE [CITATION Zha39 \l 1033]78 as the base model.
Wen et al. (2020) [CITATION Wen20 \l 1033]64	MRC	Text span	n2c2 notes, $emrQA_{why}$, and $SQuAD_{why}$	BERT model trained on different data sources.
Soni et al. (2020) [CITATION Son79 \l 1033]65	MRC	Text span	CliCR [CITATION Sus40 \l 1033]79 and emrQA dataset	BERT, BioBERT, Clinical BERT, XLNet
Mairittha et al. (2020) [CITATION Mai36 \l 1033]66	MRC	Text span	why-question answering (why-QA) dataset developed based on 2010 n2c2/VA Workshop on Natural Language Processing Challenges for Clinical Records	Makes use of BERT (Large, Uncased, Whole Word Masking), BERT fine-tuned on Stanford Question Answering Dataset (SQuAD) benchmark, BioBERT, and an extended BioBERT fine-tuned on unstructured EHR data
Moon et al. (2022) [CITATION Moo88 \l 1033]67	MRC	Text span	Why-QAs from the n2c2 ADE Challenge and Medication Why-QAs from the emrQA	ClinicalBERT model fine-tuned on SQuAD-why dataset
Li et al. (2023) [CITATION LiY23 \l 1033]68	MRC	Text span	emrQA	Clinical-Longformer and Clinical-BigBird language model
Yang et al. (2022) [CITATION Yan22 \l 1033]69	MRC	Text span	emrQA	GatorTron language model

Lehman et al. (2023) [CITATION Leh91 \l 1033]73	MRC	Text span	RadQA	Evaluates 12 different language models (T5-Base, Clinical-T5-Base-Ckpt, Clinical-T5-Base, RoBERTa-Large, BioClinRoBERTa, GatorTron, T5-Large, Clinical-T5-Large, PubMedGPT, T5-XL, Flan-T5-XXL, GPT-3) ranging from 220M to 175B parameters on three tasks including MRC task on EHR QA dataset.
Kang and Baek et al. (2022) [CITATION Kan55 \l 1033]70	Knowledge conditioned Feature Modulation on Transformer for MRC	Text span	emrQA	Knowledge-Augmented Language model Adaptation (KALA)
Wang et al. (2020) [CITATION Wan20 \l 1033]5	question to SQL query	Table content	MIMICSQL	TRanslate-Edit Model for Question-to-SQL (TREQS)
Raghavan et al. (2021) [CITATION Rag17 \l 1033]8	question to logical forms	Table content	emrKBQA	Min et al. (2020) [CITATION Min30 \l 1033]80 for sequence-to-sequence task along with ParaGen and ParaDetect model.
Pan et al. (2021) [CITATION Pan21 \l 1033]62	question to SQL query	Table content	MIMICSQL	Medical text–to-SQL model (MedTS) model
Soni et al. (2022) [CITATION Son06 \l 1033]63	question to logical forms	Table content	ICU_{data} [CITATION Rob98 \l 1033]23 and $FHIR_{data}$ [CITATION Son18 \l 1033]44	Tranx, Coarse2Fine, Transformer, Lexicon-based.
Tarbell et al. (2023) [CITATIO	question to SQL query	Table content	MIMICSQL 2.0 split	T5 language model for question-to-SQL task, along with data

N Tar98 \l 1033]71				augmentation method for back-translation. Then, the model is trained on both the MIMICSQL and Spider datasets to improve generalizability.
quEHRy [CITATION Sar23 \l 1033]72	question to answer extraction pipeline	Table content	$FHIR_{data}$ [CITATION Son18 \l 1033]44 and ICU_{data} [CITATION Rob98 \l 1033]23	End-to-end EHR QA pipeline with concept normalization (MetaMap), time frame classification, semantic parsing, visualization with question understanding, and query module for FHIR mapping/processing
Kim et al. (2022) [CITATION Kim22 \l 1033]39	question to Program	Element from knowledge graph	MIMICSPARQL*	Program-based model
Wang et al. (2021) [CITATION Wan18 \l 1033]40	KBQA	Element from knowledge graph	ClinicalKBQA	Attention-based aspect reasoning (AAR) method for KBQA
Park et al. (2021) [CITATION Pardf \l 1033]41	question to SPARQL query	Element from knowledge graph	MIMICSPARQL*	Seq2Seq model [CITATION Luo66 \l 1033]81 and TREQS [CITATION Wan20 \l 1033]5
Schwertner et al. (2019) [CITATION Sch20 \l 1033]58	question to SPARQL query	Element from knowledge graph	QA dataset developed on Oncology XML EHR notes in Portuguese	Developed a framework ENSEPRO, which supported question answering from Knowledge Bases.
Bae et al. (2021) [CITATION Bae03 \l 1033]61	question to query (SQL/ SPARQL)	Table content or element from knowledge graph	MIMICSQL* and MIMICSPARQL*	Unified encoder-decoder architecture that uses input masking (UniQA)
Bardhan et al. (2022) [CITATION Bar17 \l 1033]25	Multimodal QA	Text span or Table content	DrugEHRQA	MultimodalEHRQA

In Table 4, the ‘Answer Type’ column specifies the expected type of the EHR-QA model, indicating whether the answer is derived from tables, text notes, a knowledge graph, or a combination of sources. The input for all models consists of natural language questions along with relevant EHR data (table, clinical text, or knowledge graph) based on the modality. For ‘text span’ answers, the input is questions and clinical notes. For ‘table content’ answers, the input of the model includes questions and EHR tables, and so on. Table 4 also describes the type of QA task: the output of the MRC task is a text span from paragraphs, KBQA outputs an element from the knowledge graph, 'question to SQL query' outputs SQL queries for input questions, and 'question to SPARQL query' outputs SPARQL queries.

We can observe from Table 4 that over the years, DrQA’s document reader, BERT and ClinicalBERT are some of the most popular QA models used for unstructured clinical notes [CITATION Pam58 \l 1033 \m Moo17 \m Yue00 \m Mah77 \m Rawdf \m Wen20 \m Son79 \m Mai36 \m Moo88]2627425 06064656667. But since the year 2022 there has been a sharp rise in the number of studies introducing new large language models (besides BERT and other variants of BERT) for machine reading comprehension tasks [CITATION Ham49 \l 1033 \m LiY23 \m Yan22 \m Leh91]48686973. For example, Clinical-Longformer and Clinical-BigBird [CITATION LiY23 \l 1033]68 and GatorTron [CITATION Yan22 \l 1033]69 language models were proposed for various tasks including EHR QA. Hamidi et al. (2023) [CITATION Ham49 \l 1033]48 also evaluated the performance of ChatGPT, Google Bard, and Claude for EHR QA. Lehman et al. (2023) [CITATION Leh91 \l 1033]73 is another study introduced in the year 2023 that evaluated different language models (T5-Base, Clinical-T5-Base-Ckpt, Clinical-T5-Base, RoBERTa-Large, BioClinRoBERTa, GatorTron, T5-Large, Clinical-T5-Large, PubMedGPT, T5-XL, Flan-T5-XXL, GPT-3) for machine reading comprehension task on EHR notes.

For QA over structured EHR tables, TREQS [CITATION Wan20 \l 1033]5, MedTS [CITATION Pan21 \l 1033]62, and T5 [CITATION Tar98 \l 1033]71 models are used. The TRanslate-Edit Model for Question-to-SQL (TREQS) [CITATION Wan20 \l 1033]5 is a sequence-to-sequence model that uses a question encoder to convert the questions into vector representations, which are then decoded into SQL queries by the decoder. The generated SQL queries are further edited using an attentive-copying mechanism and recovery mechanism. The generated query’s condition values are compared to the closest value in the database, and the condition value is replaced by this value in the database. Medical text-to-SQL model (MedTS) [CITATION Pan21 \l 1033]62 is another text-to-SQL model that uses a pre-trained BERT model as an encoder and a grammar-based LSTM decoder to obtain an intermediate sequence. Experiments on the MIMICSQL dataset have shown that the MedTS model outperforms the TREQS model by 22.8% logical form accuracy and by 24.5% execution accuracy. Note that logical form accuracy and execution accuracy are some common evaluation metrics in text-to-SQL tasks. They are explained in detail in the subsection Evaluation Metrics. Some other examples of table-based QA methods include Tranx [CITATION Yin02 \l 1033]82, Coarse2Fine [CITATION Don68 \l 1033]83, transformer-based model [CITATION Son06 \l 1033]63, lexicon-based models [CITATION Son06 \l 1033]63, quEHRy [CITATION Sar23 \l 1033]72, and sequence-to-sequence task used with ParaGen and ParaDetect models [CITATION Rag17 \l 1033]8.

Some models for QA over graph-based EHR are the sequence-to-sequence model[CITATION Pardf \l 1033]41, TREQS model [CITATION Pardf \l 1033]41, UniQA model[CITATION Bae03 \l 1033]61, and attention-based aspect reasoning (AAR) method for KBQA [CITATION Wan18 \l 1033]40. For majority of these models [CITATION Pardf \l 1033 \m Bae03]4161, the EHR

relational database (like MIMIC-III) is converted into a knowledge graph and a question-to-SPARQL task is performed in order to retrieve answers from the knowledge graph. The sequence-to sequence model [CITATION Luo66 \l 1033]81 uses a bidirectional LSTM as the encoder and uses an LSTM decoder while having an attention paradigm. Unlike the TREQS model [CITATION Wan20 \l 1033]5, the sequence-to-sequence model cannot handle out-of-vocabulary words. The UniQA model [CITATION Bae03 \l 1033]61 uses unified encoder-decoder architecture along with input-masking (IM) and value recovering technique, thus it is robust to typos and mistakes in questions. The UniQA model can be used for both table-based QA datasets (MIMICSQL*) and graph-based QA datasets (MIMIC-SPARQL*). The condition value of the query generated using the question-to-query model is compared with the values in the database. This is called the condition value recovery technique. ROUGE-L score [CITATION Lin13 \l 1033]84 is used to check the similarity between the values in the database to that of the condition values in the generated query. Then, the condition values are replaced with values most similar to those in the database. After applying the recovery technique, UniQA outperforms both the sequence-to-sequence model (by 301.35% logical form accuracy and 170.23% execution accuracy) and the TREQS model (by 46.06% logical form accuracy and 30.76% execution accuracy). Wang et al. (2021) [CITATION Wan18 \l 1033]40 developed a knowledge base by linking the clinical notes of patients and introduced an attention-based aspect reasoning (AAR) method for KBQA.

Most of the existing works discuss only QA on unimodal EHR data. Bardhan et al. (2022) [CITATION Bar17 \l 1033]25 has proposed a simple pipeline for multimodal QA on EHRs (called MultimodalEHRQA) that uses a modality selection network in order to choose the modality between structured and unstructured EHR as the preferred modality. If the selected modality obtained is "unstructured text", then QA is performed over the clinical notes using BERT or ClinicalBERT, and the span of text from the clinical notes is returned as the multimodal answer. Similarly, if the preferred modality selected is "structured tables", then a text-to-SQL task is performed using the TREQS model [CITATION Wan20 \l 1033]5. Further research is still needed to develop a multimodal QA model capable of handling the more challenging task of using answers from both structured and unstructured data to obtain a contextualized answer.

Evaluation Metrics

In this section, we discuss the different evaluation metrics used for EHR. Evaluation metrics are used to evaluate the efficacy of different models. Multimedia Appendix 4 lists the different evaluation metrics used in different EHR QA studies.

The type of QA task would determine the evaluation metrics used. For QA with machine reading comprehension tasks (for example, in QA over clinical notes), exact match and F1 score are the most popular metrics for evaluation [CITATION Pam58 \l 1033 \m Moo17 \m Yue00 \m Xia10 \m Mah77 \m Dad73 \m Rawdf \m Son79 \m Mai36 \m LiY23 \m Yan22 \m Leh18 \m Kan55]26274246505160656668. Exact match refers to the percentage of predictions that exactly match the ground truth answers. In [CITATION Pam58 \l 1033]26, an exact match is used to determine if the answer entity is included in the evidence. If not, it is determined whether the projected span of evidence is within a few characters of the actual evidence. Precision measures the number of tokens in a prediction that overlaps with the correct answer compared to the total number of tokens in the prediction. Recall calculates the proportion of tokens in the correct answer that are included in the prediction compared to the total number of tokens in the correct answer. Precision and recall are represented using equations 1 and 2.

$$\text{Precision} = TP / (TP + FP) \quad \textbf{(1)}$$

$$\text{Recall} = TP/ (TP + FN) \quad \textbf{(2)}$$

Where TP, FP, and FN represent true positives, false positives, and false negatives at the token level. The F1 measure is a broader metric that calculates the average overlap between the prediction and the correct answer [CITATION Mut21 \l 1033]6. It is defined as the harmonic mean of precision and recall. This is represented using equation 3. Wen et al. (2020) [CITATION Wen20 \l 1033]64 and Moon et al. (2022) [CITATION Moo88 \l 1033]67 used exact match and partial match to assess the QA models for answering questions based on patient-specific clinical text. F1 measure was used for weighing the partial match between the predicted token of words and the golden token of words. The F1 score is calculated using the following equation:

$$F1 = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad \textbf{(3)}$$

Evaluation metrics such as logical form accuracy and execution accuracy are commonly used for evaluating models responsible for table-based QA that use a question-to-SQL query-based approach [CITATION Wan20 \l 1033 \m Pan21 \m Tar98]56271. They are also used for graph-based QA that use a question-to-SPARQL query-based approach [CITATION Pardf \l 1033 \m Bae03]4161. The logical form accuracy is calculated by doing a string comparison between the predicted SQL/SPARQL queries and the ground truth queries, and execution accuracy is calculated by obtaining the ratio of the number of generated queries that produce correct answers to the total number of queries [CITATION Wan20 \l 1033]5. There are instances where execution accuracy might include questions where the generated SQL query is different from the ground truth query, but the returned answer is the same. Structural accuracy is another metric used to evaluate models used for question-to-SQL/question-to-SPARQL query tasks [CITATION Pardf \l 1033 \m Bae03]4161. Structural accuracy is similar to measuring logical form accuracy, except that it ignores the condition value tokens. Condition value refers to the string value or numeric value in the WHERE part of the SQL/SPARQL query. For example, in the SQL query - "SELECT MAX(age) from patients WHERE Gender = “F” and DoB > 2020", "F" and 2020 are the condition values. The objective of using structural accuracy is to evaluate the accuracy of converting questions to SQL/SPARQL query structures, by not giving importance to the condition values (like in Spider dataset [CITATION YuT25 \l 1033]85). Raghavan et al. (2021) [CITATION Rag17 \l 1033]8 uses exact match and denotation accuracy for evaluating clinical table-QA models. The framework involves two stages - (1) Predicting logical forms for questions, and (2) obtaining answers from the database with logical forms as input. Exact match is used for semantic parsing, while denotation accuracy is used to evaluate models for obtaining answers from logical forms. Denotation accuracy checks if the logical forms which are input to the model return the correct label answer, and the exact match is used to check if the logical forms generated are the same as the ground truth logical forms.

A variety of text-generating metrics have been used to evaluate question paraphrasing. Soni et al. (2019) [CITATION Son03 \l 1033]52 used BLEU [CITATION Papdf \l 1033]86, METEOR [CITATION Aga34 \l 1033]87, and TER [CITATION Sno06 \l 1033]88 for evaluating paraphrasing models. The BLEU (or Bilingual Evaluation Understudy) score evaluates how closely generated paraphrases (or candidate translation) resemble those in the reference. This is done with exact token matching. The BLEU score is calculated as follows:

$$\text{BLEU} = \text{brevity penalty} * \exp\left(\sum w_n \log p_n\right) \quad \textbf{(4)}$$

where,

$$\text{Bravity penalty} = \min(1, \exp(1 - \text{reference length/output length})) \quad \textbf{(5)}$$

$$p_n = \text{total number of candidate n-grams / total number of matched n-grams} \quad \textbf{(6)}$$

In equation 4, w_n represents the weight for each n-gram. The METEOR score (Metric for Evaluation

of Translation with Explicit ORdering) on the other hand uses synonyms and word stems. This is represented using the following equations:

$$\text{Meteor} = F * (1 - \text{Penalty}) \quad \textbf{(7)}$$

$$\text{where, } F = (\text{Precision} * \text{Recall})/(\alpha * \text{Precision} + (1 - \alpha) * \text{Recall}) \quad \textbf{(8)}$$

$$\text{Penalty} = \gamma * (\text{ch}/\text{m})^{\beta} \quad \textbf{(9)}$$

In equation 9, 'ch' is the number of chunks that match, and m is the number of uniforms that match between the prediction and the reference. The parameters α, β, and γ are adjusted to maximize the correlation with human judgments. The edit distance (the number of edits necessary to change one sentence into another) between generated and reference paraphrases is measured by the TER score (Translation Error Rate). It is calculated by adding up all the edits, dividing that total by the number of words, and multiplying that result by 100, i.e.

$$\text{TER} = (\text{\# of edits} / \text{average \# of reference words}) * 100 \quad \textbf{(10)}$$

Discussion

Challenges and existing solutions

Limited number of clinical annotations for constructing EHR QA datasets

There are very few clinical EHR annotations that are publicly available. n2c2 repository is one of the very few public repositories that hosts EHR NLP datasets (that can be used to create template-based QA datasets). This is because creating these annotations requires a lot of manual work which can be time-consuming, and at the same time require domain knowledge [CITATION Bar17 \l 1033 \m Pam58]2526. For the same reasons, it was difficult to annotate EHR QA datasets. There are also some ethical issues and privacy concerns which need to be handled while constructing EHR QA datasets. This involves the de-identification of information related to patients.

Datasets like emrQA [CITATION Pam58 \l 1033]26 and ClinicalKBQA [CITATION Wan18 \l 1033]40 are examples of template-based datasets that have used the available expert annotations of the n2c2 repository to generate large-scale patient-specific QA datasets using semi-automated methods, taking advantage of the limited clinical annotations. While questions in these datasets do not represent the true distribution of questions one would ask to EHR, their scale makes them valuable for transfer learning and methods development.

Concept normalization in clinical QA

Question answering in any domain has its own challenges. But clinical QA has added challenges. One major challenge is when different phrases are used for the same medical concept in the question and the database. To deal with this issue, clinical normalization is used. Clinical normalization involves recognizing the medical entities and terminologies and converting them into a singular clinical terminology or language. A lot of EHR QA datasets such as emrQA have used MetaMap [CITATION Aro66 \l 1033]89 during the dataset generation process to map medical terminologies mentioned in the clinical text to the UMLS Metathesaurus. However, it has been argued that concept normalization for EHR QA is fundamentally different than the task on clinical notes [CITATION Sar23 \l 1033]72, so QA-specific datasets are clearly needed.

Generating realistic EHR QA datasets

It is necessary to make sure that questions in EHR QA datasets contain realistic questions that clinicians and patients would want answered from EHR data. In order to create realistic questions while constructing the EHRSQL dataset [CITATION Gyu22 \l 1033]36, a poll was created at a hospital to gather real-world questions that are frequently asked on the structured EHR data. The DiSCQ dataset [CITATION Leh18 \l 1033]43 also included clinically relevant questions by collecting questions that physicians could ask. This ensured the use of medically relevant questions in the EHR QA datasets.

Adding more paraphrases to the QA dataset is another way to make sure the questions are realistic. This is because, in a real-world scenario, the same question may be posed or stated in different ways. Generation of paraphrases may be machine-generated, human-generated [CITATION Pam58 \l 1033]26, or it could be a combination of both [CITATION Gyu22 \l 1033]36. Table 5 lists the number of paraphrases used per template in different EHR QA datasets.

Table 5. Summary of paraphrases used in various EHR QA datasets

Dataset	Average no. of paraphrases per question type	Method of generating paraphrases	Total number of questions
MIMICSQL [CITATION Wan20 \l 1033]5	1	Human labor (crowdsourcing)	10,000 questions
emrQA [CITATION Pam58 \l 1033]26	7	Human labor (Templates generated by physician were slot-filled)	1 million questions
emrKBQA [CITATION Rag17 \l 1033]8	7.5	Human labor (Templates generated by physicians were slot-filled)	940,173
EHRSQL [CITATION Gyu22 \l 1033]36	21	Human labor and machine learning	24,000

Open Issues and Future Work

Redundancy in the types of clinical questions

Most of the existing EHR QA datasets are template-based datasets that are obtained by slot-filling. These datasets have several instances of the same type of templates that are slot-filled with various entities. As a result, there is redundancy in the diversity of questions generated. This is still an ongoing issue that needs to be addressed.

Need for multimodal EHR QA systems

Clinical EHRs contain a vast amount of patient information. They contain information about the patient’s medical history, diagnosis information, discharge information, information related to medicine dosage, as well as medication allergies. Structured EHR data contains highly complementary data that may or may not be present in the clinical notes. The information in

structured and unstructured EHR data may contain information that is similar, may contradict, or can provide additional context between these sources. There is a clear need for EHR QA systems that reason across both types of data.

DrugEHRQA [CITATION Bar17 \l 1033]25 and MedAlign [CITATION Fle89 \l 1033]49 datasets are the only multimodal EHR QA datasets that uses both structured data and unstructured clinical notes to answer questions (though MedAlign dataset is technically a pseudo-multimodal EHR QA dataset since the QA pairs of the MedAlign dataset are based on an XML markup that are derived from structured and unstructured EHR data). Bardhan et al. (2022) [CITATION Bar17 \l 1033]25 introduced a simple baseline QA model for multimodal EHR data, and further research is needed to develop a multimodal QA model that unifies the EHR data modalities to obtain a contextualized answer.

QA of EHRs on unseen paraphrased questions

QA models trained on clinical question-answer pairs when tested on unseen paraphrased questions have historically produced poor results. There have been works that have tried to address this challenge. The model in Raghavan et al. (2021)[CITATION Rag17 \l 1033]8 uses paraphrasing detection and generation as a supplementary task to handle this issue. Another solution was discussed in Rawat et al. (2020) [CITATION Rawdf \l 1033]60. Rawat et al. (2020) [CITATION Rawdf \l 1033]60 introduced a multi-task learning approach where extractive QA and prediction of answer span was the primary task, with an auxiliary task of logical form prediction for the questions. But this is still an ongoing issue which needs further work.

QA of EHRs on unseen data

QA models should be able to generalize to new clinical contexts and EHR questions. In order to study generalization, Yue et al. (2020) [CITATION Xia10 \l 1033]46 evaluated the performance of a model trained on the emrQA dataset on a new set of questions based on clinical notes of MIMIC-III. The experiment proved that the accuracy of the QA model dropped down by 40% when tested on unseen data. The same research group later proposed a solution [CITATION Yue00 \l 1033]42. They developed the CliniQG4QA framework, which uses question generation to obtain QA pairs for unseen clinical notes and strengthen QA models without the need for manual annotations. This was done using a sequence-to-sequence-based question phrase prediction model (QPP).

This issue was also addressed in question-to-SQL tasks for table-based EHR QA. Tarbell et al. (2023) [CITATION Tar98 \l 1033]71 introduced the MIMICSQL 2.0 data split (derived from the existing MIMICSQL dataset[CITATION Wan20 \l 1033]5) to test generalizability of existing text-to-SQL models on EHRs. The performance of TREQS [CITATION Wan20 \l 1033]5 model on the MIMICSQL 2.0 data split was drastically poor (logical form accuracy of 0.068 and execution accuracy of 0.173 when trained on paraphrased questions and tested on paraphrased questions), thus showing the need for improvement. To improve generalizability of text-to-SQL tasks on EHR data, Tarbell et al. (2023) then introduced the use of T5 model with data augmentation method using back-translation and further adding out-of-domain training data to improve generalizability on text-to-SQL tasks. The proposed model even though outperformed the TREQS model (logical form accuracy of 0.233 and execution accuracy of 0.528 when trained on paraphrased questions and tested on paraphrased questions), still needs further improvement. More work is required in the future to overcome this challenge.

Progress of QA models in real clinical applications

Integrating QA systems into clinical workflows allows healthcare practitioners to access current medical information and recommendations, potentially lowering medical errors and improving

patient care. Studies are now being conducted on QA models to determine their accuracy, safety, and reliability in clinical settings. These studies are critical for establishing their usefulness in real-world settings [CITATION Joh231 \l 1033]90. Efforts are underway to create user-friendly interfaces that allow healthcare providers to communicate more easily. Some QA models are currently being tested in cohort selection studies [CITATION Xio21 \l 1033]13 and clinical trials to determine their efficacy and safety in real-world contexts. Deploying QA models in clinical contexts involves ethical problems about patient privacy, bias reduction, and transparency in decision-making. Addressing these concerns is critical for establishing acceptance among healthcare professionals and patients. To summarize, while QA models have considerable benefits for clinical practice and research, their implementation in real-world clinical applications necessitates resolving integration and ethical issues. To fully harness the power of QA models in healthcare, it is essential that AI researchers and physicians keep working together.

Strengths

In this study, we presented the first scoping review for QA on EHRs. We methodologically collected and screened papers related to EHR QA from January 1st, 2005 to September 30th, 2023, and performed a thorough review of the existing studies on EHR question answering. Then, we explored all the existing datasets, approaches, and evaluation metrics used in EHR QA. Furthermore, we identified the different modalities for QA over EHRs and described the approaches used for each. We have fulfilled all PRISMA scoping review requirements.

This review helps to identify the challenges faced in EHR QA. Also, this study sheds light on the problems that have been solved along with the additional gaps that are still remaining. This will encourage researchers in this domain to pursue these open problems that have not yet been solved.

Limitations

Despite the strengths of this study, we note a few limitations. First, the search process was limited to a handful of EHR and QA-related keywords. There is a long tail in how these types of systems are described in the literature, but there is a possibility that we might have missed relevant studies that did not match this initial search criteria. We used forward snowballing to partially resolve this issue. This helped us to identify ten additional papers that we had missed out on earlier. But in spite of this, there is still a slim chance that we might have missed a few relevant studies in our final list. Furthermore, given the current expansion of research into EHR QA, we predict that new studies would been added to this list since our search.

Conclusions

In recent years, question answering over EHRs has made significant progress. This is the first systematic/scoping review of QA over EHRs. In this paper, we have provided a detailed review of the different approaches and techniques used for EHR QA. The study began by discussing the need for large domain specific EHR QA datasets and then discussed the existing EHR QA datasets. We have reviewed the different unimodal EHR QA models used for both structured EHRs and unstructured EHRs, as well as QA models on multimodal EHRs. Then, we identified the major challenges in this field, such as the limited number of clinical annotations available for EHR QA dataset generation. We also talked about potential future directions in this field. It is a relatively new field with many unexplored challenges that require attention. This study should help future researchers to explore various research directions within EHR QA and expand the horizons of research areas in this field.

Acknowledgements

This project has been funded by NIH grants R00LM012104, R21EB029575, and R01LM011934.

Author's Contributions

For this study, JB, KR, and DZW proposed the idea of the study. All the authors (JB, KR, DZW) jointly made the rules for inclusion and exclusion criteria. JB and KR contributed towards paper collection and overall screening process. Both JB and KR classified the papers based on their scope. JB conducted the initial analysis and drafted the manuscript. The manuscript of the paper was then critically reviewed by KR and DZW. All authors approved the final version of the manuscript.

Conflicts of Interest

None declared.

Abbreviations

EHR: Electronic Health Record
EMR: Electronic Medical Record
KG: Knowledge Graph
KB: Knowledge Base
NLP: Natural Language Processing
QA: Question Answering

References

CITATION Dem20 \l 1033 \m Rob14 \m Cai11 \m Lee85 \m Wan20: , ,
CITATION Mut21 \l 1033 \m Ath: , ,
CITATION Wan20 \l 1033 : , ,
CITATION Rag17 \l 1033 : , ,
CITATION Joh23 \l 1033 : , ,
CITATION Pol18 \l 1033 : , ,
CITATION Joh16 \l 1033 : , ,
CITATION Dat22 \l 1033 : , ,
CITATION Xio21 \l 1033 : , ,
CITATION Devdf \l 1033 : , ,
CITATION Seo03 \l 1033 : , ,
CITATION Lee20 \l 1033 : , ,
CITATION Liu08 \l 1033 : , ,
CITATION Wan14 \l 1033 : , ,
CITATION Lia42 \l 1033 : , ,
CITATION New21 \l 1033 : , ,
CITATION Wel35 \l 1033 \m Han21: , ,
CITATION Rob98 \l 1033 : , ,
CITATION Rob48 \l 1033 : , ,
CITATION Rag17 \m Wan20 \m Bar17 \m Pam58 \m Moo17 \l 1033 : , ,
CITATION Pam58 \l 1033 : , ,
CITATION Moo17 \l 1033 : , ,
CITATION Bar17 \l 1033 : , ,
CITATION Uzu10 \m Uzu101 \l 1033 \m Uzu08 \m Uzu09 \m Stu15 \m Uzu12: , ,
CITATION Wan20 \m Rag17 \m Bar17 \m Pam58 \m Moo17 \m Rob98 \m Rob48 \l 1033 : , ,
CITATION Predf \l 1033 \m Jun13 \m Gyu22 \m Son72 \m Oli10 \m Kim22 \m Wan18: , ,

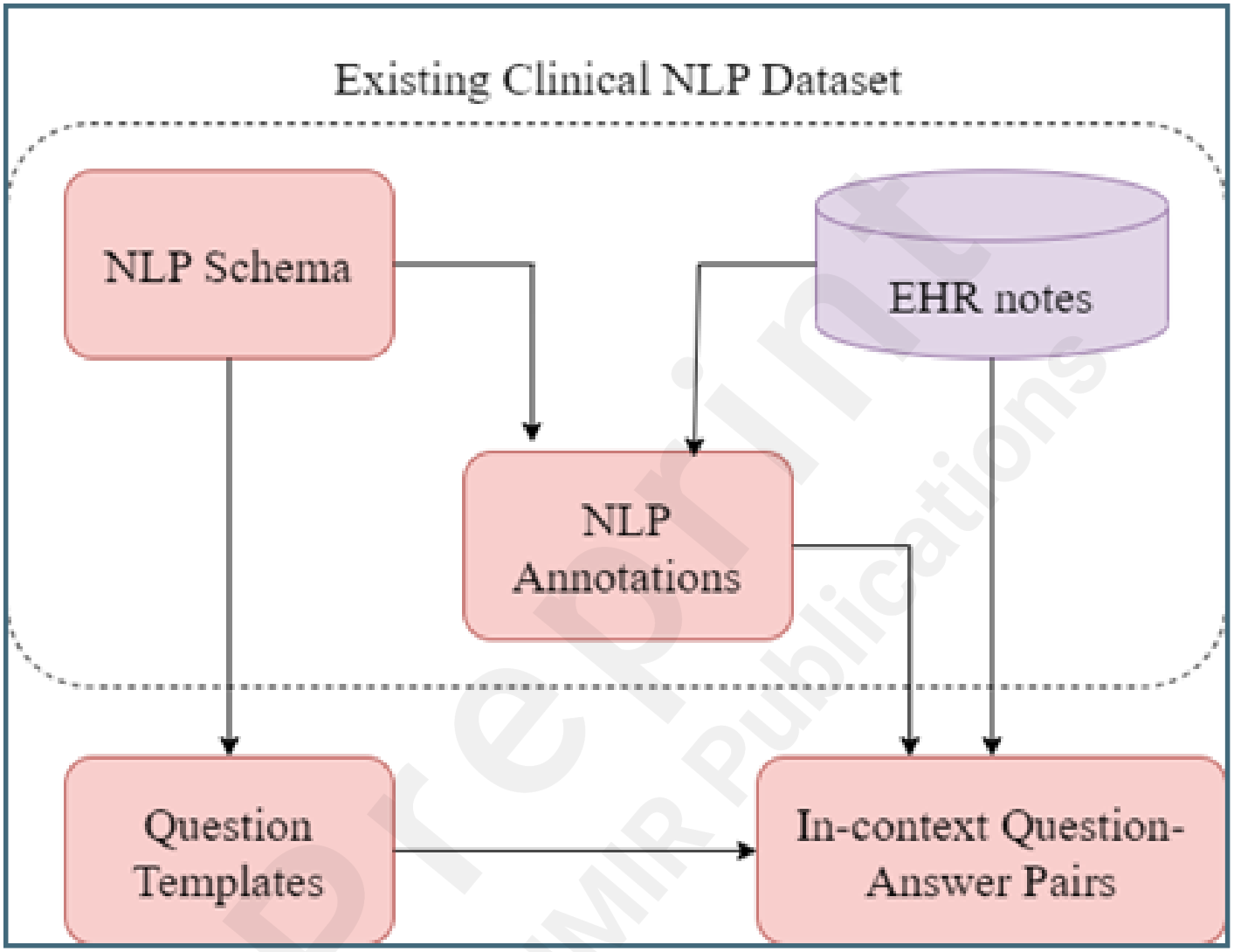
CITATION Pardf \l 1033 \m Yue00 \m Leh18 \m Son18 \m Misv1 \m Xia10 \m Palqm: , ,
CITATION Ham49 \l 1033 \m Fle89 \m Mah77 \m Dad73: , ,
CITATION Leh18 \l 1033 : , ,
CITATION Son03 \l 1033 \m Moo13 \m Son85: , ,
CITATION Pat12 \l 1033 \m Rob97: , ,
CITATION Wan20 \l 1033 \m Rag17 \m Bar17 \m Pam58 \m Moo17 \m Oli10 \m Kim22 \m Wan18 \m Pardf: , ,
CITATION Yue00 \l 1033 \m Ham49 \m Fle89 \m Mah77 \m Dad73 \m Rob17 \m Sch20 \m Raw52: , ,
CITATION Rawdf \l 1033 \m Bae03 \m Pan21 \m Son06 \m Wen20 \m Son79 \m Mai36 \m Moo88: , ,
CITATION LiY23 \l 1033 \m Yan22 \m Kan55 \m Tar98 \m Sar23 \m Leh91: , ,
CITATION Xia10 \l 1033 : , ,
CITATION Pardf \l 1033 : , ,
CITATION Yue00 \l 1033 : , ,
CITATION Predf \l 1033 : , ,
CITATION Son18 \l 1033 : , ,
CITATION Jun13 \l 1033 : , ,
CITATION Oli10 \l 1033 : , ,
CITATION Son72 \l 1033 : , ,
CITATION Gyu22 \l 1033 : , ,
CITATION Kim22 \l 1033 : , ,
CITATION Wan18 \l 1033 : , ,
CITATION Ham49 \l 1033 : , ,
CITATION Fle89 \l 1033 : , ,
CITATION Misv1 \l 1033 : , ,
CITATION Palqm \l 1033 : , ,
CITATION Mah77 \l 1033 : , ,
CITATION Dad73 \l 1033 : , ,
CITATION Wan20 \l 1033 \m Rag17 \m Gyu22 \m Kim22 \m Pardf: , ,
CITATION Pam58 \l 1033 \m Moo17 \m Jun13: , ,
CITATION Son72 \l 1033 \m Yue00 \m Leh18 \m Misv1 \m Xia10 \m Palqm \m Ham49: , ,
CITATION Tsu21 \l 1033 : , ,
CITATION Uzu10 \l 1033 \m Uzu101 \m Uzu08 \m Uzu09 \m Stu15 \m Uzu12: , ,
CITATION Pam58 \l 1033 \m Moo17 \m Jun13 \m Leh18 \m Misv1 \m Yue05: , ,
CITATION Moo17 \l 1033 : , ,
CITATION Yue00 \l 1033 \m Yue05: , ,
CITATION Pol18 \l 1033 : , ,
CITATION Rob98 \l 1033 \m Rob48: , ,
CITATION Pam58 \l 1033 \m Moo17 \m Oli10 \m Yue00 \m Ham49 \m Mah77 \m Dad73 \m Raw52 \m Son79 \m Mai36: , ,
CITATION Wan20 \l 1033 \m Bae03 \m Pan21 \m Tar98: , ,
CITATION Kim22 \l 1033 \m Pardf \m Bae03: , ,
CITATION Raj64 \l 1033 : , ,
CITATION Robdf \l 1033 : , ,
CITATION Rob17 \l 1033 : , ,
CITATION Raw52 \l 1033 : , ,
CITATION Rawdf \l 1033 : , ,
CITATION Zha39 \l 1033 : , ,
CITATION Wen20 \l 1033 : , ,

CITATION Son79 \l 1033 : , ,
CITATION Sus40 \l 1033 : , ,
CITATION Mai36 \l 1033 : , ,
CITATION Moo88 \l 1033 : , ,
CITATION LiY23 \l 1033 : , ,
CITATION Yan22 \l 1033 : , ,
CITATION Leh91 \l 1033 : , ,
CITATION Kan55 \l 1033 : , ,
CITATION Min30 \l 1033 : , ,
CITATION Pan21 \l 1033 : , ,
CITATION Son06 \l 1033 : , ,
CITATION Tar98 \l 1033 : , ,
CITATION Sar23 \l 1033 : , ,
CITATION Luo66 \l 1033 : , ,
CITATION Sch20 \l 1033 : , ,
CITATION Bae03 \l 1033 : , ,
CITATION Pam58 \l 1033 \m Moo17 \m Yue00 \m Mah77 \m Rawdf \m Wen20 \m Son79 \m Mai36 \m Moo88: , ,
CITATION Ham49 \l 1033 \m LiY23 \m Yan22 \m Leh91: , ,
CITATION Yin02 \l 1033 : , ,
CITATION Don68 \l 1033 : , ,
CITATION Pardf \l 1033 \m Bae03: , ,
CITATION Lin13 \l 1033 : , ,
CITATION Pam58 \l 1033 \m Moo17 \m Yue00 \m Xia10 \m Mah77 \m Dad73 \m Rawdf \m Son79 \m Mai36 \m LiY23 \m Yan22 \m Leh18 \m Kan55: , ,
CITATION Mut21 \l 1033 : , ,
CITATION Wan20 \l 1033 \m Pan21 \m Tar98: , ,
CITATION YuT25 \l 1033 : , ,
CITATION Son03 \l 1033 : , ,
CITATION Papdf \l 1033 : , ,
CITATION Aga34 \l 1033 : , ,
CITATION Sno06 \l 1033 : , ,
CITATION Bar17 \l 1033 \m Pam58: , ,
CITATION Aro66 \l 1033 : , ,
CITATION Sar23 \l 1033 : , ,
CITATION Joh231 \l 1033 : , ,

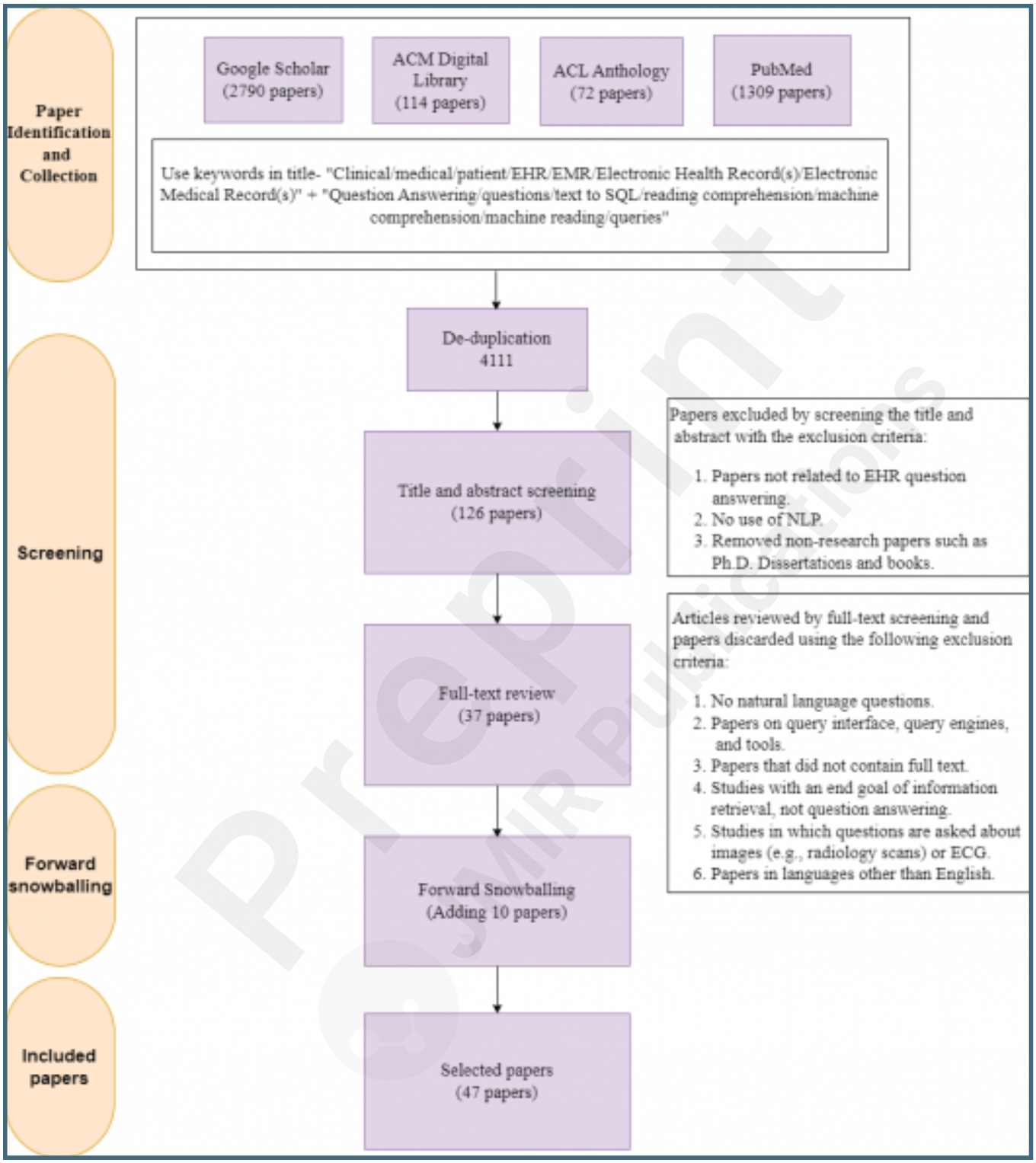
Supplementary Files

Figures

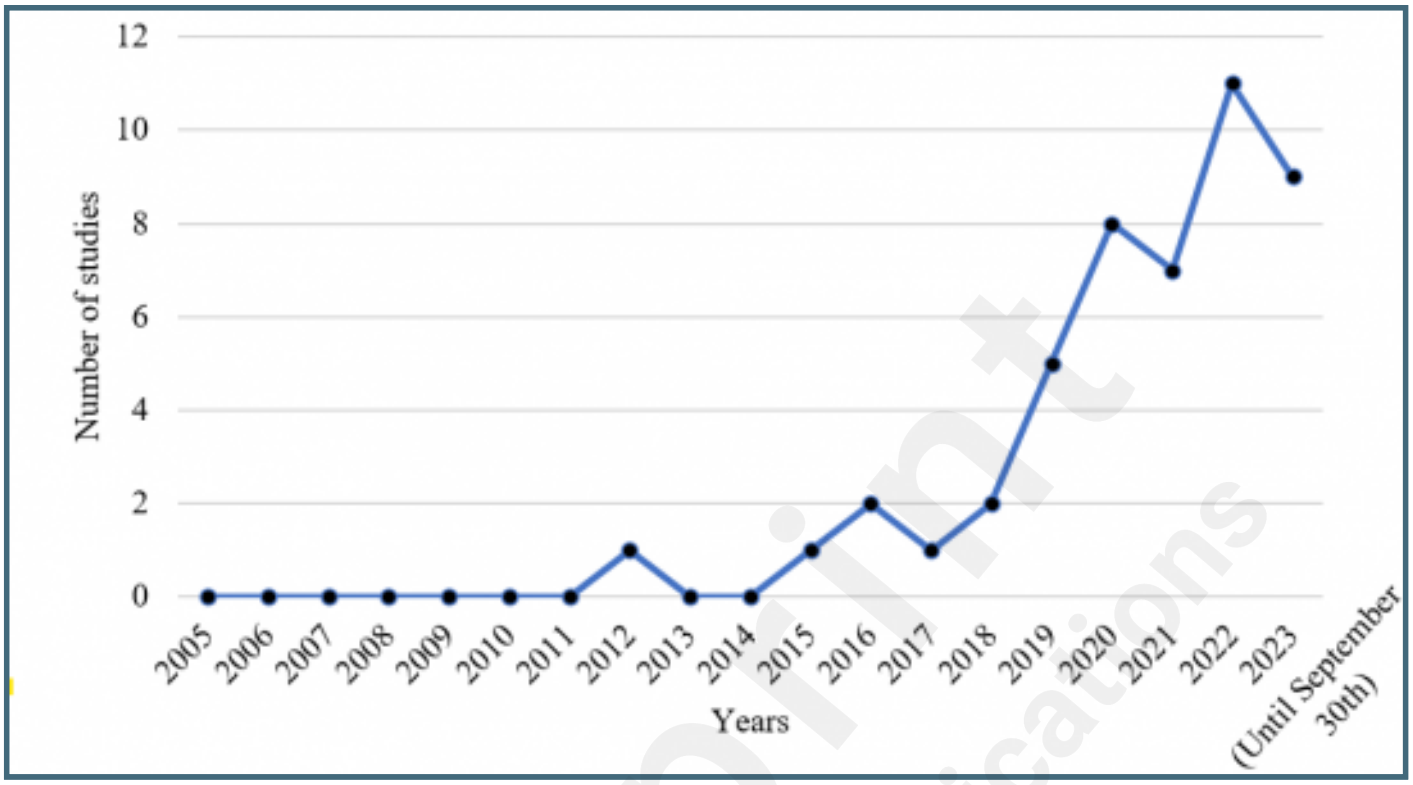
Flowchart showing the process of template-based dataset generation. The dotted boundary shows the existing non-QA NLP dataset along with the EHR data. Question templates (and logical form templates) are constructed based on the EHR data. Clinical expert annotations of non-QA tasks are used to slot-fill placeholders in question templates and generate question-answer pairs.



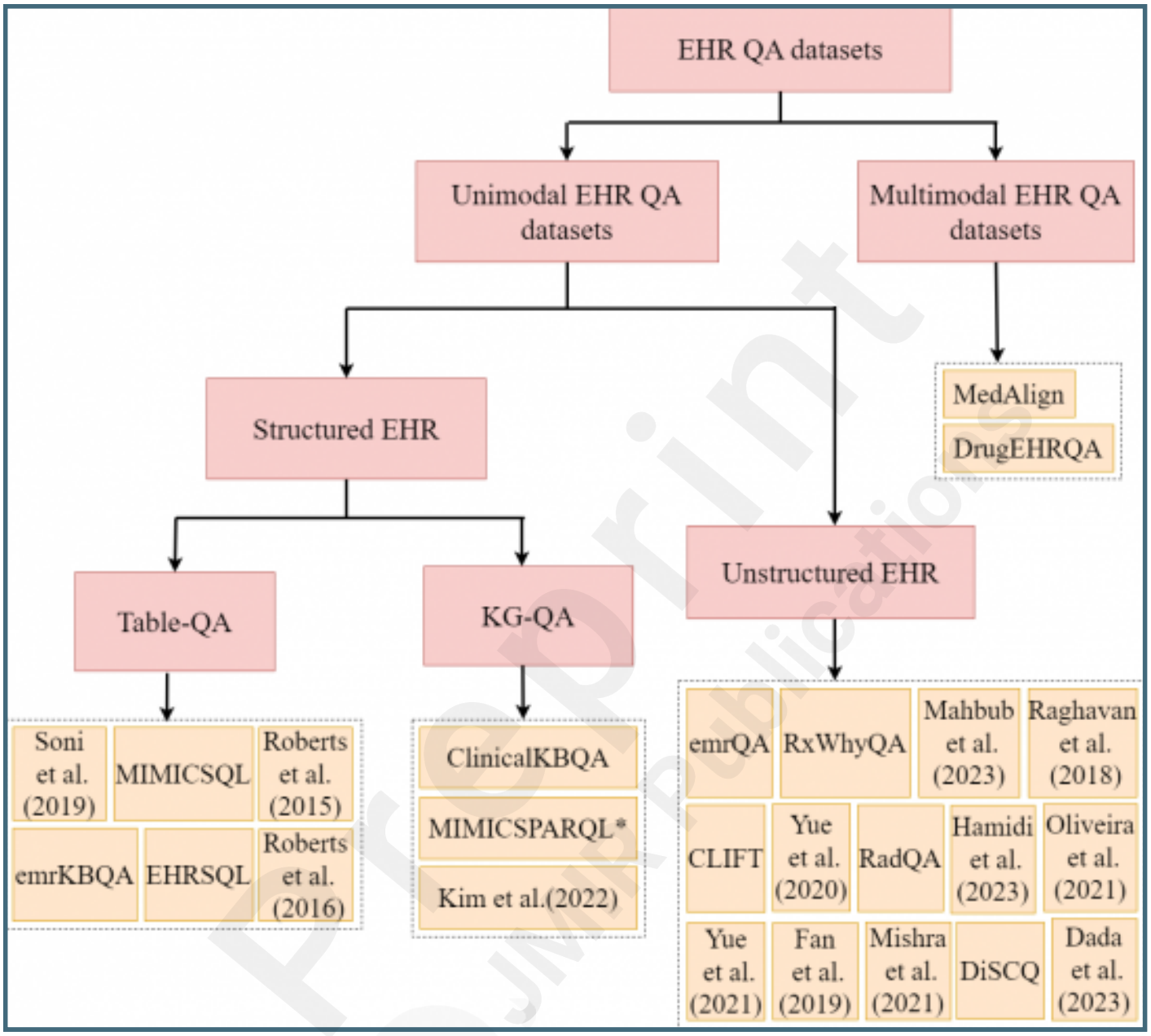
PRISMA diagram for study on QA over EHRs.



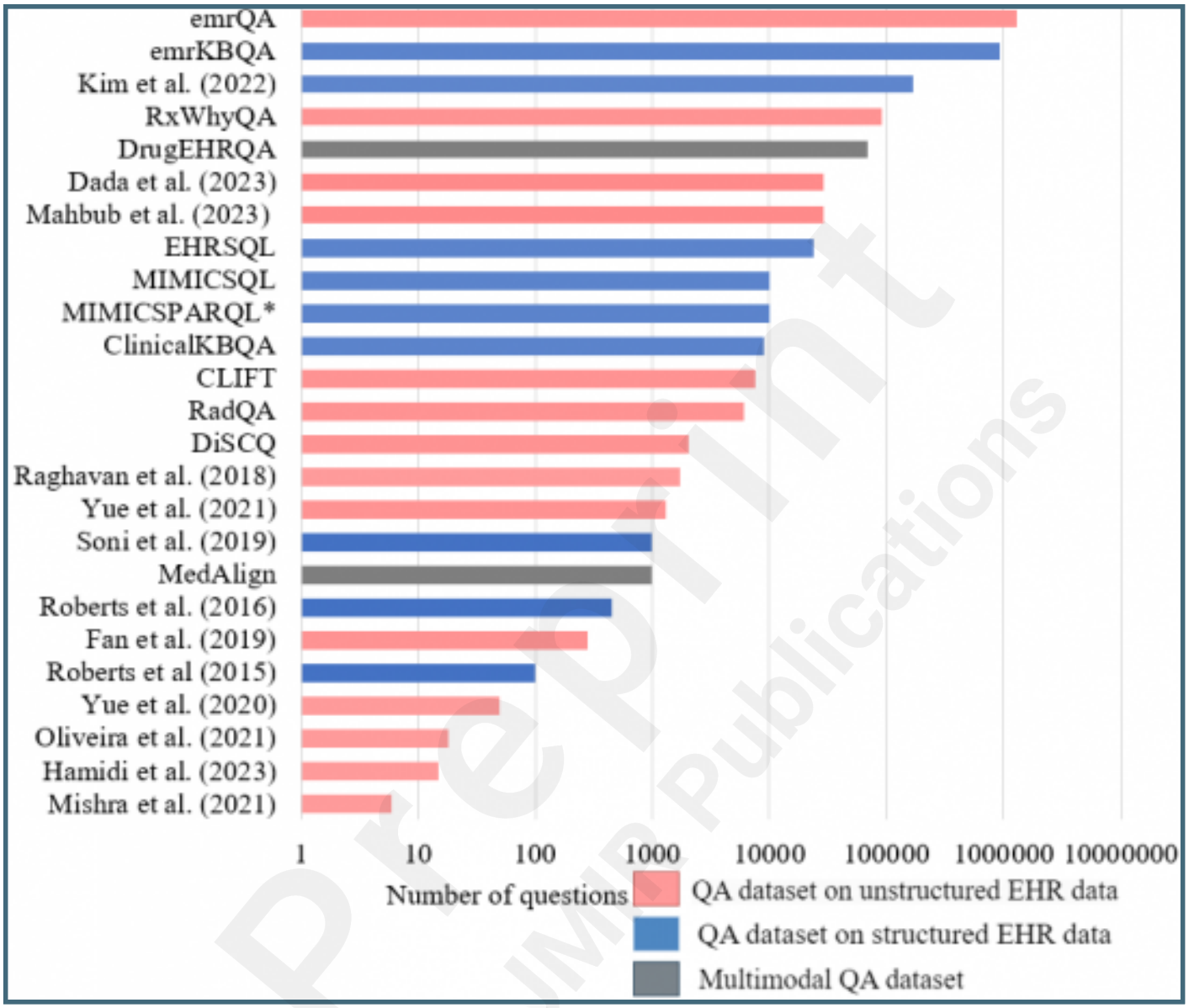
Number of studies on EHR QA over the years.



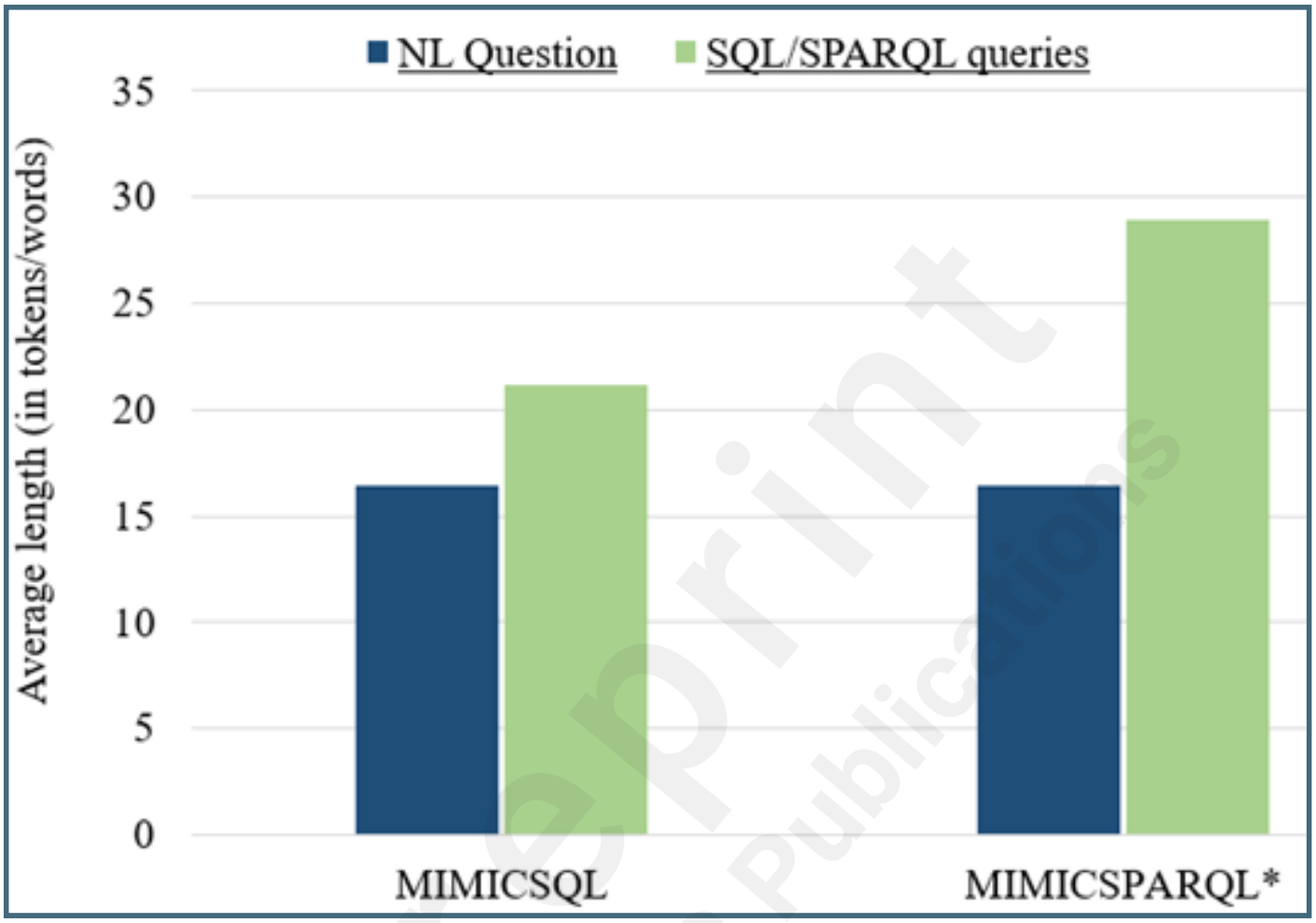
Classification of EHR QA datasets based on modality.



Plot shows the total number of questions included in various EHR QA datasets and classifies them into unstructured, structured, and multimodal EHR QA datasets.



Average length of questions and SQL/SPARQL queries (in tokens or words) for MIMICSQL and MIMICSPARQL* datasets.



Multimedia Appendixes

PRISMA Checklist.
URL: http://asset.jmir.pub/assets/6a03ddba8354da4be7959e2fcf0c3524.docx

Summaries of selected papers.
URL: http://asset.jmir.pub/assets/d0ae2f373ff0629cf83af11d7f52e935.xlsx

Comparison of different EHR QA datasets.
URL: http://asset.jmir.pub/assets/9f8f9ca8389bf93dff79bd6413764960.docx

Evaluation metrics.
URL: http://asset.jmir.pub/assets/a22126f86f950d2a451710ee8ccdd33e.docx