Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6250–6259
Marseille, 20-25 June 2022

RadQA: A Question Answering Dataset to Improve Comprehension of Radiology Reports

Sarvesh Soni[1], **Meghana Gudala, Atieh Pajouhi, Kirk Roberts**[2]
School of Biomedical Informatics, The University of Texas Health Science Center at Houston
{[1]sarvesh.soni, [2]kirk.roberts}@uth.tmc.edu

Abstract

We present a radiology question answering dataset, RadQA, with 3074 questions posed against radiology reports and annotated with their corresponding answer spans (resulting in a total of 6148 question-answer evidence pairs) by physicians. The questions are manually created using the clinical referral section of the reports that take into account the actual information needs of ordering physicians and eliminate bias from seeing the answer context (and, further, organically create unanswerable questions). The answer spans are marked within the Findings and Impressions sections of a report. The dataset aims to satisfy the complex clinical requirements by including complete (yet concise) answer phrases (which are not just entities) that can span multiple lines. We conduct a thorough analysis of the proposed dataset by examining the broad categories of disagreement in annotation (providing insights on the errors made by humans) and the reasoning requirements to answer a question (uncovering the huge dependence on medical knowledge for answering the questions). The advanced transformer language models achieve the best F1 score of 63.55 on the test set, however, the best human performance is 90.31 (with an average of 84.52). This demonstrates the challenging nature of RadQA that leaves ample scope for future method research.

Keywords: question answering, machine reading comprehension, radiology reports, clinical notes

1. Introduction

Question answering (QA) is an intuitive means to query text data and it is especially helpful in the case of large and complex documents. Machine reading comprehension (MRC) has been widely explored to this end in order to better comprehend unstructured text, by enabling machines to answer specific questions given a textual passage (Zeng et al., 2020). Much of these pursuits are powered by neural models and since a well-constructed dataset is pivotal to building a suitable model (for a given requirement, domain, or task) there is an explosion of MRC datasets in recent years (Dzendzik et al., 2021). However, little work is drawn toward the clinical domain to build challenging MRC datasets for improving comprehension of the semantically complex and diverse electronic health record (EHR) data.

In medicine, much work in MRC is targeted toward biomedical scientific articles, which comes under the umbrella of biomedical QA (Athenikos and Han, 2010; Jin et al., 2021). But, owing to the fundamental differences between the text data present in scientific articles and EHRs (Friedman et al., 2002), the datasets (and models) for the former cannot be directly used for the latter. Moreover, between the two types of EHR data, adequately-sized QA datasets are constructed for the structured part (e.g., semantic parsing) (Roberts and Demner-Fushman, 2016; Pampari et al., 2018; Wang et al., 2020; Raghavan et al., 2021) more so than that for the unstructured part. Since the information in both the structured and the unstructured parts of EHR are different and offer unique perspectives about care provision (Tayefi et al., 2021), it is crucial to dig deeper into the unstructured data as they include richer information.

FINAL REPORT
INDICATION: 64 year old male with status post recent STE MI. Now with increasing edema and shortness of breath.
FINDINGS: The heart is ***(enlarged in size)*** but stable in the interval. Mediastinal contour is unchanged. There is upper zone redistribution of the pulmonary artery vasculature. Perihilar haziness as well as ***diffuse bilateral pulmonary opacities***. These findings are consistent with acute CHF. There are also ***bilateral pleural effusions***. There is barium in the left colon from previous study.
IMPRESSION: 1. Findings consistent with ***pulmonary edema*** due to CHF. 2. ***Bilateral pleural effusions***.
Q – Are there any infiltrates in the lung? **A –** ***diffuse bilateral pulmonary opacities*** (Fndg), ***pulmonary edema*** (Imp)
Q – Did the cardiac silhouette enlarge? **A –** ***(enlarged in size)*** (Fndg)
Q – Is there any sign of pleural effusion? **A –** ***[b/B]ilateral pleural effusions*** (Fndg and Imp)

Table 1: An example from RadQA. The answers are ***italicized only***, ***underlined***, or ***(in parentheses)***. **Fndg –** Findings. **Imp –** Impression.

The current MRC datasets for unstructured EHR data fall short of many important considerations in order to build an advanced model for the task. Most of these datasets are too small (to build advanced models) (Fan, 2019) or publicly unavailable (making them

Dataset	Size		Annotation				Docs Source	Available
	# Ques	# Docs	Source	Ques Prompt	Ans Selection	UN-Q		
Raghavan et al. (2018)	1747	71	Medical students	patient summary, clinical note, reference questions	clinical note	✗	Cleveland Clinic (medical records)	✗
Pampari et al. (2018)	73111 (from 680 templates)	303	Automatically generated	question template	automatically using NLP annotations on clinical note	✗	n2c2 datasets (mostly discharge summaries)	✓
Fan (2019)	245	138	Author	candidate sentence with 'because' and/or 'due to'	candidate sentence	✗	2010 i2b2/VA NLP challenge (discharge summaries)	✓
Yue et al. (2020a)	50	–	Medical experts	–	clinical note	✗	MIMIC-III (clinical notes)	✗
Yue et al. (2020b)	1287	36	Medical experts	clinical note, candidate questions	clinical note, answers for candidate questions	✗	MIMIC-III (clinical notes)	✓
Oliveira et al. (2021)	18	9	Authors	nursing diagnosis, risk factors, defining characteristics	nursing/medical note	✗	SemClinBr corpus (Portuguese nursing and medical notes)	✗
RadQA (this work)	**3074 (6148 QA pairs)**	**1009**	**Physicians**	**clinical referral section of radiology report**	**whole radiology report**	✓	**MIMIC-III (radiology reports)**	✓

Table 2: Existing EHR MC datasets alongside our proposed RadQA dataset. **#** – Count. **Ques** – Questions. **Ans** – Answers. **Docs** – Documents. **UN-Q** – Unanswerable questions. **n2c2** – formerly i2b2.

nonexistent for building any models) (Raghavan et al., 2018) or both (Yue et al., 2020a; Oliveira et al., 2021). Additionally, the questions for most of these datasets are collected in a manner that induces bias and does not reflect the real-world user needs, including for an available dataset where the users are shown candidate questions (with answers) for reference (Yue et al., 2020b). Lastly, one of the "large" EHR MRC datasets, emrQA (Pampari et al., 2018), has gained much traction. However, the variety in emrQA is severely limited due to templatization, as is also found in a separate systematic analysis of emrQA's MRC data, where they achieved about the same model performance when trained on 5-20% of the dataset versus on the entire training data (Yue et al., 2020a). Thus, there is a need to build a representative dataset for the task of EHR MRC that encompasses user needs, is adequately sized to train advanced models, and is publicly available to push research in model development forward.

Furthermore, almost all existing datasets for EHR MRC use discharge summaries as documents. However, other types of clinical texts such as radiology reports (that have vastly different semantic content and vocabulary) are markedly underrepresented in the MRC task. In a recent study, radiology information extraction task is framed as QA to extract radiological entities from report text (Datta and Roberts, 2021), still, it is not illustrative of an actual MRC task because the predefined question templates are limited (only to the specific types of entities extracted) and the queries themselves are not natural (lacks variation). To our knowledge, no existing work focused on the task of radiology MRC in its true sense.

In this work, we propose RadQA[1], a new EHR MRC dataset, that aims to overcome the issues with the existing resources for the MRC task in clinical domain (an example from the dataset is shown in Table 1). The following are the main characteristics of RadQA:

- The questions reflect true information needs of clinicians ordering radiology reports (as the queries are inspired from the clinical referral section of the radiology reports).
- The corpus contains 3074 unique question-report pairs encompassing 1009 radiology reports from 100 patients.
- Each question has two answers for a radiology report (in its *Findings* and *Impressions* sections), resulting in a total of 6148 distinct question-answer evidence pairs (including unanswerable questions, that no available MRC dataset includes).
- The answers are oftentimes present in the form of phrases or span multiple lines (as opposed to only multi-word answer entities in available MRC datasets), fulfilling the clinical information needs.
- The questions require a wide variety of reasoning and domain knowledge to answer, that makes it a challenging dataset for advanced models.
- The distribution of the sampled radiology reports is similar to that in the MIMIC-III database.
- The dataset is publicly available (as the radiology reports come from the publicly available MIMIC-III database).

2. Related Work

The current datasets for the task of EHR MRC are summarized in Table 2, along with our proposed dataset, RadQA, for comparison. Raghavan et al. (2018) described a dataset for EHR MRC, where they had medical students create questions while reviewing a clinical

[1] https://github.com/krobertslab/datasets/tree/master/radqa

summary or the latest clinical note of a patient alongside a set of reference questions. Since the annotators were shown a set of candidate questions for reference, they may be more likely to ask questions that look similar to the referred questions (and thus the actual information need may not be met by the created questions). Moreover, the dataset is unavailable.

Pampari et al. (2018) employed a template-based approach to build a large dataset of question-logical form pairs by leveraging an existing set of natural language processing (NLP) annotations. The dataset claims over $400k$ question-answer evidence pairs, however, this number reduces to $73k$ after mapping the dataset to a span-extraction MRC task (where each question has a definite span as an answer in the associated evidence) (Soni and Roberts, 2020). Regardless, the variety in the dataset (of questions etc.) is severely limited due to templatization. This is also found in a separate systematic analysis of emrQA's MRC data, where Yue et al. (2020a) achieved about the same model performance when trained on 5-20% of the dataset versus on the entire training data. For evaluating emrQA on unseen data, Yue et al. (2020a) also created a small MRC dataset (limited information is available about the dataset creation).

In another work, Yue et al. (2020b) created a test set for evaluating their proposed framework for EHR MRC. During annotation, annotators view a clinical note along with candidate question-answer pairs (on this note) and are asked to create new questions (encouraged) and/or select some from the shown candidates. In their final set, over 75% of the questions are selected from the automatically-generated candidates.

Fan (2019) focused on why-question answering where the question-answer pairs were cultivated from sentences containing 'because' and/or 'due to'. The representation of this dataset is limited to these sentences and thus do not reflect the actual information needs. Further, the questions created do not involve cross-sentence synthesis and may be biased (both in content and style) and thus lack diversity. Oliveira et al. (2021) explored the EHR MRC task for Portuguese, building a small Portuguese dataset in order to evaluate a transfer learning model.

There are several efforts toward EHR QA for collecting patient-specific questions (that can be answered using EHR data) (Patrick and Li, 2012) and focusing on retrieving information from the structured part of EHRs by creating datasets (Roberts and Demner-Fushman, 2016; Wang et al., 2020; Raghavan et al., 2021) and/or building models (Roberts and Patra, 2017; Pan et al., 2021). However, because both the structure and the information content vary significantly between the two types of EHR data (structured and unstructured), the models and datasets built for the structured data cannot be readily applied to structured EHR data. Thus, we do not compare the other methods and datasets built for structured data as it is not the focus of this work.

Measure	**RadQA**	**MIMIC-III**
# of patients	100	34,325
# of reports	1009	332,922
Avg reports per patient	10.09	9.7
Std of reports per patient	8.15	8.33
Median reports per patient	7.5	7
Top five modalities	*(proportion in %)*	
X-ray	59.76	55.32
Computed Tomography (CT)	14.37	16.23
Ultrasound (US)	5.15	4.61
Magnetic Resonance (MR)	3.87	4.31
CT Angiography (CTA)	2.38	2.43

Table 3: Descriptive statistics of the sampled reports and MIMIC-III data (after removing outlier patients). Top modalities determined separately after filtering out the report types with proportions less than 0.1%.

3. RadQA Dataset

3.1. Document Sampling

We source the radiology reports (used as documents) for our dataset from the MIMIC-III database (Johnson et al., 2016), which is a publicly available resource containing information on intensive care unit patients. MIMIC-III includes over $2M$ clinical notes, out of which more than a quarter are radiology reports (over $500k$) covering a wide variety of modalities such as chest x-ray, computed tomography (CT), and magnetic resonance imaging (MRI).

We sample a realistic set of reports by selecting the documents at the patient level, i.e., we first sampled patients and then included all the associated radiology reports in our dataset for annotations. Specifically, we randomly sample 100 patients from the set of patients with at least 1 and at most 36 (to remove outliers) radiology reports, thereby resulting in a total of 1009 reports in our final set. We further divide our data of 100 patients into training, development, and testing splits in the ratio of 8:1:1, respectively. The descriptive statistics of our sampled reports are shown in Table 3 alongside the MIMIC-III statistics for comparison.

3.2. Question Creation

We take a novel approach to create the set of questions that satisfy the true information needs of the readers of radiology reports, i.e., the physicians who order radiology exams. The ordering physicians include their requirements in the form of a clinical referral that is sent to the radiologist along with the radiographs. The radiologists refer to these expressed requirements while writing their interpretation of the radiology images in the form of a radiology report, the final version of which includes the referral section at the beginning. We harness the clinical referral section to capture the actual information needs of the ordering physicians.

Figure 1 illustrates our question creation process. In order to align well to the information needs mentioned in the referral portion, we hide the other contents of the report from annotators in the question creation

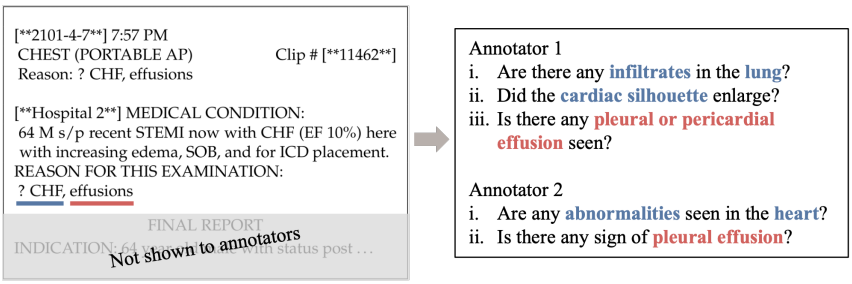


Figure 1: An example of clinical referral section with the corresponding constructed questions. Only the referral section is shown to the annotators in the question creation phase, not the whole report.

phase. This forces the annotators to focus only on the aspects related to conducting the exam (as provided in referral), and not diverge into creating biased questions based on the remaining full report content. The referral contains, along with the examination type, mainly two sections of our interest, namely, *Medical Condition* (gives a brief overview about the condition of the patient) and *Reason for this Examination* (provides the reasons for ordering the radiology study). The referral shown in Figure 1 is for a *Chest (Portable AP)* study (a type of *X-ray*), where the patient is 64 year old male with certain conditions and the exam is ordered to evaluate for *CHF (Congestive Heart Failure)* and *effusions*.

The annotators were asked to create questions thinking about both implicit and explicit information needs expressed by the referral section. E.g., *"Is there any sign of pleural effusion?"* asks about the explicit needs conveyed by the reason *effusions* while the question *"Did the cardiac silhouette enlarge?"*, instead of directly asking about the reason *CHF*, asks about an implicit detail associated with it, i.e., the *enlargement of heart* due to *CHF*. Our annotators possess the medical knowledge to understand both the types of information needs in the referral and thus were able to incorporate those needs into their created questions.

To create specific questions, we instructed the annotators to ask separate questions for each *reason* mentioned in the referral. In other words, we refrained from creating questions encompassing all the *sub-reasons* present in the *reason* section of the referral. E.g., we asked annotators to create separate questions for reasons *CHF* and *effusions*, instead of asking about both in the same question. Further, annotators were asked to create at least one question for each *reason*. This was helpful to create individual questions capturing the distinct information needs.

We advised the annotators to employ different variations in phrasing the questions and to not merely fill in reason information in a set of templates. E.g., for the sample in Figure 1, we asked not to create all questions like *"Does the patient have CHF?"* and *"Does the patient have effusions?"*. Several other examples were also provided to help them understand the task. This helped us create a syntactically diverse set of questions.

Two annotators independently constructed questions for all the reports in our dataset. This further improved the heterogeneity of our set of questions. We assessed the created questions at regular intervals through the annotation process to reinforce all the instructions (provided to the annotators under annotation guidelines). All the questions are associated with the report whose clinical referral section was used while constructing them. Finally, the questions are deduplicated at report level, i.e., if the two annotators ended up asking the exact same question (in terms of its text) about a report, we remove the duplicate.

3.3. Answer Annotation

For marking answers, the annotators were shown the whole radiology report (including the referral) along with the corresponding set of questions. There are two main sections in radiology reports, namely, *Findings* and *Impressions*, where the former describes the characteristics of underlying medical image(s) in detail while the latter summarizes the findings (largely inspired by the requirements mentioned in the referral section). We tasked the annotators to annotate answer spans in the report text, at most one span each in *Findings* and *Impressions* sections. We instructed them to annotate the shortest possible span that answers a question to the point. Specifically, the selected answer span should be sufficient by itself to answer the question but, simultaneously, it should not contain any additional information that is not required by the question.

In the example from Table 1, the answer span for the question *"Did the cardiac silhouette enlarge?"* is annotated as *"enlarged in size"* because this is the shortest span in the *Findings* that can sufficiently to answer the question. Note that we did not include in the answer span any other portion of the sentence *"heart is enlarged in size but stable in the interval"*. Though other information in the whole sentence (or even in the other parts of the report) may be relevant to the question at hand, we do not include it because it is extraneous to the exact information needs of the question. In other words, we asked the annotators to keep this in mind while selecting the answers – if a clinician asks the question at hand, would the selected span be just enough to answer it (given that the clinician can view the origin of the returned answer from report text).

Again, the medical knowledge of our annotators is vital in this phase. Because the questions are not con-

Category	Description	Example	%
Additional info	Disagreement in keeping extra information related to question	**Q:** Is there any lung consolidation? **A1:** new region of consolidation around the right hilum **A2:** consolidation around the right hilum (✓)	21%
Additional info (unrelated)	Disagreement in keeping extra information not directly related to question	**Q:** Is the vagal nerve stimulator intact on the left side of chest? **A1:** Presence of vagal nerve stimulator (✓) **A2:** Presence of vagal nerve stimulator, old healed rib fractures on the left side	16%
Answerable/ Unanswerable	One annotator selected a span while the other marked unanswerable	**Q:** Do we see any fluid in the pulmonary interstitial spaces? **A1:** pleural fluid layering on the right **A2:** *No Answer* (✓)	27%
		Q: Does the chest x ray show any consolidated lungs? **A1:** new apparent patchy opacities at both lung bases **A2:** *No Answer* **RC:** patchy opacities at both lung bases (✓)	
Missing key clinical information	At least one annotator missed relevant clinical information for the question	**Q:** Is there any mass obstructing the upper GI tract? **A1:** No structural abnormalities are detected **A2:** Barium passes freely through the esophagus (✓)	25%
		Q: What do the intrapleural space look like? **A1:** no pleural effusion or pneumothorax (✓) **A2:** no pleural effusion	
Missing clinical specificity	At least one annotator did not mark all the required specific answer details	**Q:** What is the position of central line? **A1:** line at the brachiocephalic vein junction **A2:** Left internal jugular line at the brachiocephalic vein junction (✓)	18%
Other	Error in selecting annotation span	**Q:** Are there any fractures in the right hip joint? **A1:** Unremarkable right hip radiograp **A2:** Unremarkable right hip radiograph (✓)	1%

Table 4: Common disagreement categories with examples from manual evaluation of 100 randomly sampled questions with any disagreement. The main differences between the answer spans are underlined. Percentages do not add to 100% as some questions fall into multiple categories. Final reconciled answers are marked with a checkmark (✓). **%** – Percentage. **Q** – Questions. **A1** – Annotator 1. **A2** – Annotator 2. **RC** – Reconciled.

structed after viewing the full report text or deciding an answer in advance, they may not have direct answers in the report. E.g., in *"Are there any infiltrates in the lung?"* (Table 1), both the medical keywords *infiltrates* and *lung* are not even present in report text. But the annotators used their medical expertise to annotate *"diffuse bilateral pulmonary opacities"* (as they represent *infiltrates* on a *X-ray*) and *"pulmonary edema"* (characterized by *infiltrates*) (Tuddenham, 1984; Hansell et al., 2008). This creates a unique challenging aspect in RadQA for next generation clinical models.

Further, owing to the question creation phase, all the questions in our dataset are not required to have answers in the report. Thus, there was also an option for the annotators to mark a question as *unanswerable*, in case they are unable to find its answer in any section of the given report. All the answers are marked by two annotators independently and reconciled at regular intervals. We use the Haystack annotation tool [2] (modified to our needs) for annotating answers. Note that the *Impressions* section is sometimes also present with heading *"Conclusion"* and *Findings* are oftentimes included under *"Procedure and Findings"*.

[2] https://github.com/deepset-ai/haystack

Split	Docs	Ques	EM	F1
First	100	296	52.40	68.02
Remaining	909	5522	50.16	69.48
All	1009	6148	50.39	69.34

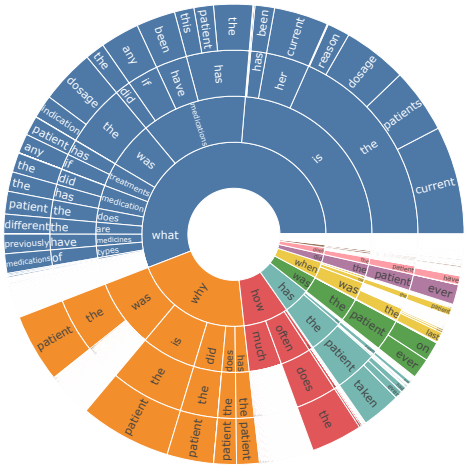
Table 5: Inter-annotator agreement.

3.4. Reconciliation Process and Challenges

We adjudicated the annotated answers frequently in order to ensure the quality of our dataset. For the first 100 reports, we reconciled in the batches of 10, giving both the annotators sufficient time to ramp up on the annotation scheme. Afterward, we reconciled in the batches of 100 reports. We calculate inter-annotator agreement using F1 measure as used in an earlier study for span-based MRC (Rajpurkar et al., 2016). We manually reviewed the disagreements and characterized the challenges encountered during reconciliation (Table 4). The agreement statistics are reported in Table 5.

We reconciled all the answer spans down to one unique answer for our training split. For the dev and test splits, however, as long as both the annotated spans answered the question at hand, we kept both the answers in the dataset. This facilitates a natural development and evaluation of models which respects the presence of more than one ways of answering a question.

(a) RadQA



(b) emrQA

Figure 2: Types of questions in RadQA and emrQA.

3.5. Dataset Analysis

The descriptive statistics of our proposed dataset are shown in Table 6. We note that the RadQA dataset is built using more than 3 times the number of reports used to generate emrQA and the average number of questions generated per paragraph in the emrQA data is significantly higher than those created in the RadQA dataset. These numbers underline a greater variation of passage data in the RadQA data as compared to emrQA. The paragraph length is a characteristic of the type of reports used in both the datasets, radiology reports in RadQA versus discharge summaries (or similar notes) in emrQA. Both the average and median answer lengths of the RadQA data (16.21 and 7, respectively) is significantly higher than that for emrQA (1.88 and 2, respectively). This emphasizes the wide variety of answers (both structurally and semantically) available in the RadQA dataset. The answers are usually present as phrases (offering a complete answer satisfying the clinical information needs of ordering physicians) in contrast to emrQA (and most other available EHR MRC datasets), which only includes answers as clinical entities such as *"aspirin"* and *"50 MG"*.

We illustrate the structural variety of questions by plotting 4-grams prefixes in Figure 2. The sunburst graph for the RadQA dataset is well-distributed while for the emrQA dataset it is skewed (having more than half of the questions beginning with *"what"*). This can be attributed to the mechanism of creating questions in the dataset, i.e., manual for the RadQA versus automatic template filling for the emrQA.

Measure		RadQA	emrQA
# of paragraphs		1009	303
# of questions	*Total*	6148	73,111
	UnAns	1754	–
Ques per para	*Avg*	6.09	241.29
	Med	6	215
Paragraph len	*Avg*	274.49	1394.29
	Med	207	1208
Question len	*Avg*	8.56	9.40
	Med	8	9
Answer len	*Avg*	16.21	1.88
	Med	7	2

Table 6: Descriptive statistics of RadQA (with emrQA for comparison). Lengths are in tokens. **#** – Count. **UnAns** – Unanswerable. **Med** – Median.

To analyze the types of reasoning that are required to answer the questions in the RadQA dataset, we perform a human evaluation by randomly sampling 100 answerable questions (see Table 7). We report the emrQA statistics directly from (Pampari et al., 2018). A majority of the questions in RadQA data require medical knowledge (73%) to answer, which justifiably compensates the lower proportion of questions with coreference reasoning. We also characterize the questions based on an additional set of reasoning categories that are peculiar to radiology (or clinical) domain.

4. Baselines

Deep learning models based on transformer architectures are shown to achieve state-of-the-art performance for MRC, where a model extracts specific answer spans from the context paragraph given a question (Liu et al., 2019). In order to understand the current level of comprehension of the advanced transformer language models, we gauged their performance on RadQA. To identify the effect of transfer learning and domain language information, we employed different variations of fine-tuning strategies and used models that are pre-trained on various open-domain and domain-specific datasets.

We use BERT (Devlin et al., 2019) and BERT-MIMIC (Si et al., 2019) as our baseline models. Devlin et al. (2019) employ masked language modelling to learn (pre-train) Bidirectional Encoder Representations from Transformers (BERT), the knowledge from which can be easily transferred to downstream tasks (such as MRC) by further fine-tuning these models. This helps transfer learning from large unstructured data to specific tasks without building the models from scratch each time and is especially helpful in the specific domains (such as clinical) with limited availability of datasets. The BERT model is pre-trained on massive textual corpora from BooksCorpus and English Wikipedia (3.3B words) for 1M steps. Si et al. (2019) further pre-train BERT on MIMIC-III notes (786M

Reasoning	Description	Example	RadQA	emrQA
Lexical Variation (Synonym)	Key links between ques and ans sentences are synonyms	**Q:** Was the PICC line **placed correctly**? **S:** **Malposition** of right sided PICC line with tip in the right internal jugular vein.	37%	15.2%
Lexical Variation (world/medical knowledge)	Key links between ques and ans sentences demand world or medical knowledge	**Q:** Is there any **obstruction in the lungs**? **S:** There has been some interval improvement of the left basilar **opacity**, consistent with atelectasis**/pneumonia**.	73%	39.0%
Syntactic Variation	Declarative form of ques does not syntactically match the ans sentences	**Q:** **Are there any fractures** in the pelvis? **S:** AP PELVIS: trauma board limits fine osseous evaluation. **No overt fractures** are seen.	66%	60.0%
Coreference	Anaphora or intra-sentence fusion	**Q:** I: Was the **PICC** placed? **S:** **PICC line** placement via … internal length is 55 cm with the tip of the **catheter** positioned in SVC. The line is ready to use.	7%	23.8%
Incomplete Context	Missing contextual information in ans sentences	**Q:** I: Do we find any stenosis in the carotid arteries that require grafting during/after CABG? **S:** Right ICA stenosis 40-59%.	16%	13.3%
Change information	Ques related to interval changes	**Q:** Has thyroid cancer **progressed**? **S:** The right neck mass appears to have significantly **increased** in size and surrounding **mass effect** compared with the prior …	18%	–
Diagnosis knowledge	Ques require diagnosis understanding to ans	**Q:** Are there signs of **pneumonia**? **S:** Marked improvement in left **perihilar alveolar process** with residual well-marginated mass-like opacity …	26%	–
Anatomy knowledge	Ques require anatomy understanding to ans	**Q:** Did the gastric cancer metastasize to **chest**? **S:** There are no **lung** nodules or masses. No destructive lytic or blastic lesions are seen in the osseous structures of the **torso**.	21%	–
Require specification	Ques require specific information in ans	**Q:** What is the **status** of the skull fracture through midface? **S:** 5. Possible **nondisplaced** fracture of the anterior wall of the right maxillary sinus. 6. **Displaced** fracture of the right nasal bone.	13%	–
Negative answer	Ans is present but negated	**Q:** I: Is there any **mediastinal shift due to** pneumothorax? **S:** **No pneumothorax**.	23%	–

Table 7: Reasoning categories with examples from a manual evaluation of 100 randomly sampled answerable questions from RadQA. Words relevant to the reasoning are **bolded**; the ground truth answers are underlined. % do not add to 1 as questions fall into multiple categories. **Q** – Question. **S** – Sentence. Hyphen **(–)** – unavailable.

words) for 300k steps, to build BERT-MIMIC. Both the models achieved state-of-the-art performances on challenge NLP datasets, that resulted in a wide adoption of transformer models in NLP. We choose the same base cased variant of both the models for a fair comparison.

5. Evaluation

We formulate the task to, given a question and a paragraph, either extract a single answer span or mark it unanswerable. We feed into the models the amalgam of *Findings* and *Impressions* sections as paragraphs along with questions and their corresponding answers. Inspired from our previous work on evaluating the fine-tuning variations of the transformer models (Soni and Roberts, 2020), we fine-tune the baseline models on different combinations of MRC datasets, both from general and clinical domains. Specifically, we explore fine-tuning on a single dataset along with a combination of two and three datasets. Here, the model is fine-tuned on each of the involved dataset for 2 epochs. Thus, the single-dataset variation is fine-tuned for 2 epochs while the double- and triple-dataset variations are fine-tuned on a total of 4 and 6 epochs, respectively.

We use SQuAD 2.0 (Rajpurkar et al., 2016; Rajpurkar et al., 2018) and emrQA (Pampari et al., 2018) for additional fine-tuning, aside from RadQA. SQuAD 2.0 is a large open-domain MRC dataset (over 150k questions), built from Wikipedia paragraphs through crowdsourcing, containing both answerable (single-span) and unanswerable questions.

The dataset is split into *training*, *development*, and *testing* sets at patient level (in the ratio of 8:1:1) for a realistic evaluation. The training set is used to train the models while the development and testing sets are used to tune the models and evaluate the final models, respectively. We calculate standard evaluation metrics for MRC, i.e., exact match or EM (strict metric that matches predicted answer phrases exactly with ground truth) and F1 (calculates F1 at word level matches between the prediction and ground truth). We tune the models on our development set. The maximum sequence length is 384; document stride is 128; maximum query length is 128; learning rate is $3e$-5.

6. Results

The evaluation results from baseline models on RadQA is shown in Table 8. Unsurprisingly, among the single-dataset variations, the best model performances come from fine-tuning on the RadQA dataset. Fine-tuning on emrQA by itself did not generalize well and essen-

Fine-tuned on	BERT				BERT-MIMIC			
	Dev		Test		Dev		Test	
	EM	F1	EM	F1	EM	F1	EM	F1
emrQA	25.08	25.08	35.21	35.21	24.92	24.92	35.21	35.21
SQuAD	25.41	36.73	30.79	42.92	25.57	42.81	24.39	40.37
RadQA	42.02	58.67	40.09	55.04	48.05	65.85	45.73	60.08
emrQA ⇒ RadQA	43.16	59.75	41.92	57.60	50.65	67.97	47.71	61.60
SQuAD ⇒ RadQA	**49.51**	**65.80**	46.04	60.71	52.28	**69.42**	**49.39**	**63.55**
SQuAD ⇒ emrQA ⇒ RadQA	48.53	63.01	**46.65**	**60.98**	**53.26**	67.79	48.32	62.29

Table 8: Model performances on the RadQA dev and test sets when fine-tuned on different data combinations.

	Dev		Test		Train		All	
	EM	F1	EM	F1	EM	F1	EM	F1
Annotator 1	85.02	92.07	81.40	90.31	66.97	81.11	70.32	83.18
Annotator 2	71.66	81.41	69.36	78.72	72.76	84.73	72.28	83.76
Avg	78.34	86.74	75.38	84.52	69.87	82.92	71.30	83.47

Table 9: Human performance on RadQA.

tially marked all the questions as unanswerable. Note that the evaluated models are trained the same way for answerable and unanswerable questions and the main difference lies in the prediction phase where a post-processing pipeline decides, on the basis of output model probabilities, whether to mark a question unanswerable or not. The predictions from the emrQA-only model are in contrast to the dataset characteristic of having only answerable questions. However, after turning off the option to mark a question unanswerable, the emrQA-only model performed even worse.

Almost all the BERT-MIMIC models performed better than their BERT equivalent variations. This echoes the usefulness of injecting clinical text information in the language models. The model variant fine-tuned on SQuAD and RadQA performed the best in a majority of cases. Thus, the additional fine-tuning on SQuAD before tuning the model on the RadQA dataset was helpful. Note that the performance jump, when compared to the RadQA-only variant, after an additional round of fine-tuning on SQuAD is higher than that with the emrQA. Note that there is a significant gap between the best baseline model performance and the average human performance (Table 9) on development and testing sets (a difference of 25 and 26 points in exact match for dev and test, respectively).

7. Discussion

We propose the RadQA (Radiology Question Answering) dataset encapsulating the actual information needs of clinicians who order the radiology exams. We present a thorough analysis of our dataset, highlighting its complexity and the reasoning required to answer (or not) the questions. The substantial gap between the baseline model and human performance presents an ample opportunity to implement sophisticated models to better comprehend radiology report text.

The evaluation results from the baseline models are consistent with our prior work (Soni and Roberts, 2020) on evaluating the task of clinical MRC, where we saw a similar trend in performance improvements when the models were fine-tuned on the different variation of open- and specific-domain datasets. Notably, the current findings reiterate that additional fine-tuning on an open-domain dataset, SQuAD, results in better performance gains as compared to additionally fine-tuning on a different clinical-domain dataset, emrQA. This may be attributed to the quality of a manually-created dataset over an automatically-generated corpus.

Injecting additional information while training the models may be helpful in improving their performance. We use the *Findings* and *Impressions* sections from radiology reports to train and test the baseline models, in order to understand the performance under standard settings. Besides, it will be interesting to prepend clinical referral section during evaluation to understand how these models use this additional information to their advantage. Moreover, as radiology reports are oftentimes long (average length of 274 tokens), a good future direction will be to explore the recent efforts toward improving the comprehension of long documents (Beltagy et al., 2020; Zaheer et al., 2020). Further, as the RadQA questions generally require medical knowledge (73% in our analysis) to answer, incorporating such knowledge into the models, such as in Hao et al. (2020), will be another appealing avenue of research for improving comprehension of radiology reports.

8. Conclusion

With the aim to improve the comprehension of radiology reports, we propose the RadQA dataset that encapsulates the information needs of ordering physicians in its questions and includes complete answers in the form of phrases. The exhaustive analysis of RadQA uncovers the common disagreements and reasoning requirements while answering the questions. The performance of the best transformer language model, MIMIC-BERT, is 63.55 (F1), which falls significantly short of the best human performance of 90.31. This indicates that the RadQA dataset is challenging and provides scope for future research in EHR MRC.

Acknowledgements This work was supported by the U.S. National Library of Medicine, National Institutes of Health, (R00LM012104); the National Institute of Biomedical Imaging and Bioengineering (R21EB029575); and UTHealth Innovation for Cancer Prevention Research Training Program Predoctoral Fellowship (CPRIT RP210042).

9. Bibliographical References

Athenikos, S. J. and Han, H. (2010). Biomedical question answering: A survey. *Computer Methods and Programs in Biomedicine*, 99:1–24.

Beltagy, I., Peters, M. E., and Cohan, A. (2020). Longformer: The Long-Document Transformer. *arXiv:2004.05150 [cs]*.

Datta, S. and Roberts, K. (2021). Fine-grained spatial information extraction in radiology as two-turn question answering. *International Journal of Medical Informatics*, 158:104628.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*, pages 4171–4186.

Dzendzik, D., Vogel, C., and Foster, J. (2021). English Machine Reading Comprehension Datasets: A Survey. *arXiv:2101.10421 [cs]*.

Fan, J. (2019). Annotating and Characterizing Clinical Sentences with Explicit Why-QA Cues. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 101–106.

Friedman, C., Kra, P., and Rzhetsky, A. (2002). Two biomedical sublanguages: A description based on the theories of Zellig Harris. *Journal of Biomedical Informatics*, 35:222–235.

Hansell, D. M., Bankier, A. A., MacMahon, H., McLoud, T. C., Müller, N. L., and Remy, J. (2008). Fleischner Society: Glossary of Terms for Thoracic Imaging. *Radiology*, 246:697–722.

Hao, B., Zhu, H., and Paschalidis, I. (2020). Enhancing Clinical BERT Embedding using a Biomedical Knowledge Base. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 657–661.

Jin, Q., Yuan, Z., Xiong, G., Yu, Q., Ying, H., Tan, C., Chen, M., Huang, S., Liu, X., and Yu, S. (2021). Biomedical Question Answering: A Survey of Approaches and Challenges. *arXiv:2102.05281 [cs]*.

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035.

Liu, S., Zhang, X., Zhang, S., Wang, H., and Zhang, W. (2019). Neural Machine Reading Comprehension: Methods and Trends. *Applied Sciences*, 9:3698.

Oliveira, L. E. S., Schneider, E. T. R., Gumiel, Y. B., da Luz, M. A. P., Paraiso, E. C., and Moro, C. (2021). Experiments on Portuguese Clinical Question Answering. In André Britto et al., editors, *Intelligent Systems*, pages 133–145.

Pampari, A., Raghavan, P., Liang, J., and Peng, J. (2018). emrQA: A Large Corpus for Question Answering on Electronic Medical Records. In *EMNLP*, pages 2357–2368.

Pan, Y., Wang, C., Hu, B., Xiang, Y., Wang, X., Chen, Q., Chen, J., and Du, J. (2021). A BERT-Based Generation Model to Transform Medical Texts to SQL Queries for Electronic Medical Records: Model Development and Validation. *JMIR Med Inform*, 9:e32698.

Patrick, J. and Li, M. (2012). An ontology for clinical questions about the contents of patient notes. *Journal of Biomedical Informatics*, 45:292–306.

Raghavan, P., Patwardhan, S., Liang, J. J., and Devarakonda, M. V. (2018). Annotating Electronic Medical Records for Question Answering. *arXiv:1805.06816*.

Raghavan, P., Liang, J. J., Mahajan, D., Chandra, R., and Szolovits, P. (2021). emrKBQA: A Clinical Knowledge-Base Question Answering Dataset. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 64–73.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392.

Rajpurkar, P., Jia, R., and Liang, P. (2018). Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Roberts, K. and Demner-Fushman, D. (2016). Annotating logical forms for EHR questions. In *LREC*, pages 3772–3778.

Roberts, K. and Patra, B. G. (2017). A Semantic Parsing Method for Mapping Clinical Questions to Logical Forms. In *AMIA Annual Symposium Proceedings*, volume 2017, pages 1478–1487.

Si, Y., Wang, J., Xu, H., and Roberts, K. (2019). Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, 26:1297–1304.

Soni, S. and Roberts, K. (2020). Evaluation of Dataset Selection for Pre-Training and Fine-Tuning Transformer Language Models for Clinical Question Answering. In *LREC*, pages 5534–5540.

Tayefi, M., Ngo, P., Chomutare, T., Dalianis, H., Salvi, E., Budrionis, A., and Godtliebsen, F. (2021). Challenges and opportunities beyond structured data in analysis of electronic health records. *WIREs Computational Statistics*, 13:e1549.

Tuddenham, W. (1984). Glossary of terms for thoracic radiology: Recommendations of the Nomenclature Committee of the Fleischner Society. *American Journal of Roentgenology*, 143:509–517.

Wang, P., Shi, T., and Reddy, C. K. (2020). Text-to-SQL Generation for Question Answering on Electronic Medical Records. In *Proceedings of The Web Conference*, pages 350–361.

Yue, X., Jimenez Gutierrez, B., and Sun, H. (2020a). Clinical Reading Comprehension: A Thorough Analysis of the emrQA Dataset. In *Proceedings of*

the 58th Annual Meeting of the Association for Computational Linguistics, pages 4474–4486.

Yue, X., Zhang, X. F., Yao, Z., Lin, S., and Sun, H. (2020b). CliniQG4QA: Generating Diverse Questions for Domain Adaptation of Clinical Question Answering. *arXiv:2010.16021 [cs]*.

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., and Ahmed, A. (2020). Big bird: Transformers for longer sequences. In H. Larochelle, et al., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297.

Zeng, C., Li, S., Li, Q., Hu, J., and Hu, J. (2020). A Survey on Machine Reading Comprehension—Tasks, Evaluation Metrics and Benchmark Datasets. *Applied Sciences*, 10:7640.