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

Sarvesh Soni, Meghana Gudala, Atieh Pajouhi, Kirk Roberts

School of Biomedical Informatics | The University of Texas Health Science Center at Houston (US)

{sarvesh.soni, kirk.roberts}@uth.tmc.edu

LREC 2022

Download RadQA now!

https://github.com/krobertslab/datasets/tree/master/radqa

INTRODUCTION – **WHY RadQA?**

- Machine reading comprehension (MRC) is widely explored to better comprehend unstructured text, by enabling machines to answer specific questions given a textual passage[1].
- Much attention in MRC drawn toward biomedical scientific articles[2].
- Limited work toward building a challenging MRC dataset for electronic health record (EHR) data[3].
  - Most existing datasets are small and/or publicly unavailable to build advanced models.
  - Question collection in most datasets incudes bias and does not reflect real-world user needs.
  - Almost all datasets use discharge summaries.
- Thus, we propose RadQA, a new EHR MRC dataset.

RadQA DATASET

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)

Tab 1. RadQA example. Fndg – Findings. Imp – Impression.

- Questions reflect true information needs of clinicians (inspired from the clinical referral section of radiology reports).
- Contains 3074 unique question-report pairs for 1009 radiology reports
- Each question has two answers for a report (in its Findings and Impressions sections), resulting in 6148 distinct question-answer evidence pairs (including unanswerable questions)
- Answers are oftentimes phrases or span multiple lines
- Questions require wide variety of reasoning & domain knowledge to answer

[**2101-4-7**] 7:57 PM
CHEST (PORTABLE AP) Clip # [**11462**]
Reason: ? CHF, effusions

[**Hospital 2**] MEDICAL CONDITION:
64 M s/p recent STEMI now with CHF (EF 10%) here with increasing edema, SOB, and for ICD placement.
REASON FOR THIS EXAMINATION:
? CHF, effusions

FINAL REPORT
INDICATION: 64 y … with status post …

Not shown to annotators

Annotator 1
i. Are there any **infiltrates** in the **lung**?
ii. Did the **cardiac silhouette** enlarge?
iii. Is there any **pleural or pericardial effusion** seen?

Annotator 2
i. Are any **abnormalities** seen in the **heart**?
ii. Is there any sign of **pleural effusion**?

Fig 1. Clinical referral section with constructed questions.

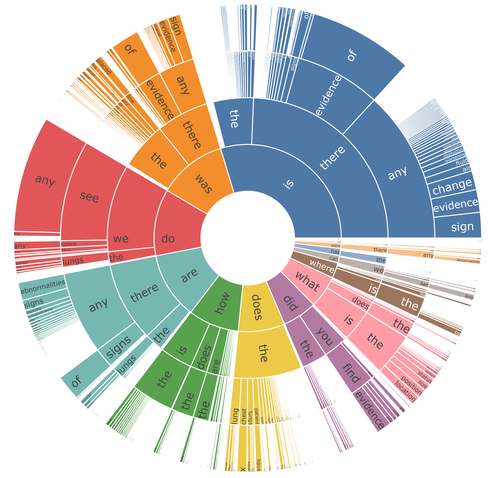

Fig 2. Types of questions in RadQA.

RELATED WORK – **EXISTING DATASETS**

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)**	✓

Tab 2. Existing MRC datasets. UN-Q – Unanswerable questions.

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%	–

Tab 3. Reasoning categories in RadQA.

BASELINES

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

	Dev		Test	
	EM	F1	EM	F1
Annotator 1	85.02	92.07	81.40	90.31
Annotator 2	71.66	81.41	69.36	78.72
Avg	78.34	86.74	75.38	84.52

Tab 4. Human performance on RadQA.

CONCLUSION

- 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 demonstrates the challenging nature of RadQA that leaves ample scope for future method research.

ACKNOWLEDGEMENTS

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

REFERENCES

1. Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ Questions for Machine Comprehension of Text. In EMNLP, pages 2383–2392.
2. 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].
3. 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.