CLIFT : Analysing Natural Distribution Shift on Question Answering Models in Clinical Domain

Anonymous Author(s)
Affiliation
Address
email

Abstract

This paper introduces a new testbed CLIFT (**Cli**nical Sh**ift**) for the clinical domain Question Answering task. The testbed includes 25k high-quality question-answering samples to provide a diverse and reliable benchmark. We performed a comprehensive experimental study and evaluated several deep-learning models under the proposed testbed. Despite impressive results on the original test set with no adaptive overfitting, the performance degrades when applied to new test sets, which leads to distribution shift. Our findings emphasise the need for and the potential for increasing the robustness of clinical domain models under distributional shift. The testbed offers one way to track progress in that direction. It also highlights the necessity of adopting evaluation metrics that consider robustness to natural distribution shift. The test sets and codes to reproduce the experiments and evaluate new models against Clift are available at anonymous.github.io

1 Introduction

Digital healthcare data has significantly expanded during the last several years. Question Answering (QA) in the Clinical domain is a particular task involved with clinical texts in Electronic Medical Records (EMR). Recent studies using neural QA models Rawat et al. [2020], Chen et al. [2017], Devlin et al. [2019] have shown promising outcomes. As a result, significant effort has been invested in developing question-and-answer algorithms & datasets in this field. In statistical learning theory, it is often assumed that test data points originate from the same distributional underpinnings as training data for a machine learning model. However, in reality, data may be sampled from various distributions with considerable domain changes; therefore, this assumption is seldom valid. Despite impressive results of QA models trained in the clinical domain, when data distribution unexpectedly changes, models that seem to be doing well in test set accuracy fail when deployed in real-time.Yue et al. [2020]

We perform a comprehensive experimental study on the emrQA dataset Pampari et al. [2018] in this paper. We introduce a new testbed derived from the MIMIC-III database. Our results show that there is no adaptive overfitting on the emrQA dataset. Moreover, we also demonstrate that emrQA models exhibit robustness to some of the natural distribution shift. However, Despite impressive results on the training dataset, when applied to new test sets, the performance degrades, and Models in our testbed experience a drop in F1 and EM metrics. Our experiments underline the need for and the potential for improving the robustness of clinical domain models under distributional shift. The results also highlight the necessity of adopting evaluation metrics that consider robustness to natural distribution shift.

In brief, the contributions of this study are as follows. (i) We are proposing five new test datasets. The testbed covers 5+ clinical domain diseases and 30+ sub-topics. These questions are from MIMIC-III

Submitted to Workshop on Robustness in Sequence Modeling at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Johnson et al. [2016], and human experts validated all the questions, answers and evidence triples. (ii) This dataset offers several advantages over existing datasets. They evaluate the diverse reasoning abilities of clinical models over a vast array of disease topics and subtopics. The testbed is a comprehensive evaluation of clinical domain data. (iii) Detailed statistics and fine-grained evaluation of natural distribution shift are provided in this study. We examine and quantify the shortcomings of the clinical deep learning models trained on emrQA when evaluated with new test data. Moreover, We also analysed the dataset on difficulty metrics such as answer diversity and Syntactic divergence. (iv) We conduct our experiments with the help of the HuggingFace Transformers library Wolf et al. [2020]. Furthermore, We also open-source the code and model checkpoints to reproduce the results and test with new models in the future. (v) We provide a new QA tool to tag the Question answering dataset and validate the samples. We hope this will accelerate the tagging process and robust ML research in this domain.

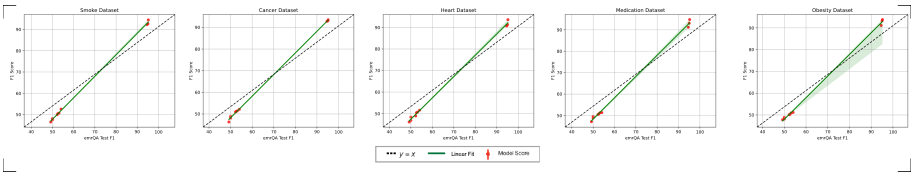

Figure 1: Comparison of different models' F1 scores on emrQA and our proposed test datasets. Our results show there is no adaptive overfitting on the Emrwa dataset. Moreover, it also demonstrates that Despite impressive results on the training dataset, when applied to new test sets, the model's performance degrades, which leads to distribution shift.

2 Background

2.1 Natural Distribution shift

For a prediction algorithm $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{Y}$ We assume a clinical domain dataset from a Hospital A where input space is X and label space is Y. To evaluate M's performance on a test dataset $Q = \{x_i | i = 1, \ldots, n\}$ drawn from a data distribution P, we will assume that algorithm M is fixed. The initial steps would be choosing a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ and calculating the loss $\ell_Q(M) = \sum_{i=1}^{n} \ell\left(M\left(x_i\right), y_i\right)$ and population loss $\ell_{\mathcal{P}}(M) = \mathbb{E}\mathcal{P}[\ell(M(x), y)]$ under the test distribution.While the input X and label Y spaces are fixed in this case, there may be many distinct joint distributions. Furthermore, Besides the dataset from Hospital A, we collect another clinical domain dataset from Hospital B where the test set is Q' from a new distribution P'

In addition to the test loss $l_Q(M)$, we are interested to know the loss $l_{Q'}(M)$ under real-time applications, known as distribution shift. The observations of models' expected loss to changes or differences would be helpful to improve the model's robustness, eliminate the biases and understand the dataset better. This difference may be broken down into three parts.Miller et al. [2020]

$$l_Q(M) - l_{Q'}(M) = \underbrace{(l_Q(M) - l_{\mathcal{P}}(M))}_{\text{Adaptivity gap}} + \underbrace{(l_{\mathcal{P}}(M) - l_{\mathcal{P}'}(M))}_{\text{Distribution gap}} + \underbrace{(l_{\mathcal{P}'}(M) - l_{Q'}(M))}_{\text{Generalization gap}} \tag{1}$$

The adaptivity gap quantifies the degree to which adjusting the model to the test set affects the estimate of population loss. While The distribution gap indicates how much a change in the sample's distribution impacts the model's performance, the difference between the sample and population losses as a result of Q's random sampling is captured by the generalization gap.

2.2 emrQA

emrQA Pampari et al. [2018] is a large-scale extractive question-answering dataset generated semi-automatically from question templates, and human experts annotated the templates. The proposed task was to extract a continuous text span from the clinical note while giving a clinical note and a question as input to the QA model. The corpus contains almost 400,000 triplets of a question, logical form, and evidence/answer.

3 Proposed Testbed

The MIMIC-III Johnson et al. [2016] is the most popular dataset for Electronic health record(EHR) related tasks. In the dataset, A clinical record is maintained for each patient, which contains information such as ECGs, physician notes, discharge summaries, nurse progress reports, and more. Due to its comprehensive collection of unstructured clinical notes, MIMIC-III has been extensively used in clinical NLP tasks.Park et al. [2021]

In order to evaluate the ability of healthcare domain QA systems to generalize to new clinical data, we created five new test sets from MIMIC-III. We utilized the same question templates provided by emrQA to create the question and answer pairs, along with a few more templates annotated by experts.

(i) The Heart disease test subset contains 5k+ annotated questions and over 3k hospital admission records related to heart disease and issues such as Heart Attack, Heart Failure, Coronary artery disease, Ischemia and more. (ii) The medication subset contains around 5k samples. It includes the question and answers related to Medical prescriptions, Drugs Name, Drug dosages, history, and & Status of Medications; it also contains questions related to the number of doses, side effects of medications, Reasons for medications, and more. (iii) Similar to the above datasets, we extracted 3k documents from the MIMIC-III, which contained around 4k questions and clinical notes related to obesity, such as comorbidities associated with the current issue, hypertension, Sleep Apnea, Diabetes, and more. The test set's diversity helps to quantify the degree to which current systems overfit the original test set. (iv) For the smoking subset, 5k questions were selected related to the current status & history of smoking, Quantity & Type of smoking, issues, and more. (v) We selected 3k reports from the MIMIC-III corpus with the regex search of cancer-related keywords. The subset includes 5k questions and answers related to different cancers, such as Breast Cancer, Colon cancer, Rectal Cancer, Thyroid Cancer, and more. Moreover, the set also contains samples mentioning the Cancer cause, side effects, medications for current cancer, stage of cancer, and more.

4 Preprocessing & Quality Checks

The following procedures were used to sanitize the raw data and to ensure that all questions could be answered using textual input. We lowercase the text in the subset test sets, remove punctuation marks and tokenize the text using whitespace. We set the min length to 100 and split the long documents into sub-documents. We also converted a few abbreviations to complete forms such as y.o., y/o, to a year old and yr, yr's, yrs to years. Moreover, additional data cleaning procedures were performed to ensure that the information gathered from the question, answer, and evidence triples met the data quality objectives. The combined dataset contains around 24k questions.

5 Dataset Statistics & Analysis

This dataset covers different diseases and medical conditions based on patient information.

Table 6 in Appendix B summarises the general statistics of preprocessed data. Moreover, Table 3 in Appendix B depicts an additional useful statistic: the number of unique tokens in the dataset. The performance of the models is influenced by the amount of vocabulary, which is a good indicator of the linguistic and domain complexity associated with a text corpus. The Obesity & Heart subset was observed to include more extended questions and a broader vocabulary than the others. As a result, it may be concluded that the former subsets question are more complex than the rest.

In addition, We use the two difficulty metrics provided by Rajpurkar et al. [2016] to compare the original emrQA test set to our four new test sets. Figure 2 in Appendix B shows Syntactic divergence histograms for the proposed datasets.

Model	EM Score			F1 Score		
	emrQA EM	Clift EM	Gap(δ)	emrQA F1	Clift F1	Gap (δ)
BioBERT Lee et al. [2019]	46.40	42.90	3.5	49.20	47.10	2.1
EHReaderZhang et al. [2021]	50.00	46.50	3.5	50.00	49.60	0.4
DistilBERT (Frozen)	50.00	48.10	1.9	50.00	48.80	1.2
BiDAF Seo et al. [2017]	50.74	50.34	0.4	52.27	50.37	1.9
QANet Yu et al. [2018]	50.68	48.58	2.1	52.56	51.36	1.2
BioBERT(+TAV)	51.20	50.80	0.4	53.10	51.20	1.9
ClinicalBERT Huang et al. [2019]	53.33	49.83	3.5	54.06	50.66	3.4
DistilBERT Sanh et al. [2019]	94.00	92.10	1.9	94.60	91.20	3.4
DistilBERT(+TAV)	94.40	90.90	3.5	95.00	93.10	1.9
QuickReader-distilBERT Zhang et al. [2021]	94.60	91.20	3.4	95.20	94.80	0.4

Table 1: Performance of different models on the emrQA Testset and new Medication Testset

6 Results

6.1 Adaptive Overfitting

The objective of the proposed testbed is to test the performance of a clinical model when evaluated on data relevant to the same Question-Answering task but from multiple data sources/hospitals and environments. We ran a battery of experiments and evaluated a total of ten models on proposed datasets. In addition to testing for adaptive overfitting on the emrQA dataset, we examine the robustness of models on the natural distribution shift in our proposed test sets. More information about baseline models is provided in Appendix A

To Answer the question that if adaptive overfitting occurred on the original test set, We compare the F1 test scores from the original test with our new five-test set. The F1 results on the few test set closely resemble the old F1 scores for the models trained on the emrQA dataset. We can see the results of this analysis in Figure 1 supports the argument and proves that the dataset has no adaptive overfitting.

6.2 Natural Distribution Shift in emrQA

To test the hypothesis that the original test set performance is better than new test sets, we compare the F1 scores for all the models trained on the emrQA to the F1 scores on each of our additional test sets.Our findings showed that, when applied to proposed test datasets, All the models indicate a performance loss ranging from 0.4% to 3.5%. Table 1 show the performance of different models on Medication testset.

As we can see that ClinicalBERT Huang et al. [2019] drops around 3.4 F1 points while evaluating the Medication dataset, compared to a drop of about 1.5 F1 points when evaluating the Smoke dataset and 1.9 F1 points when testing on Cancer. The outcome demonstrates that both the distributional and adaptivity gaps are modest. However, These results suggest that the models are sensitive to distribution shift. Real-world deployments of health machine learning systems may encounter these failure situations. Therefore, depending on the entity of interest, performance for downstream applications would vary greatly. Table 3, 4, 5 and 2 in Appendix B show the results of different models on Cancer, Obesity, Heart and Smoke dataset respectively.

These results imply the potential for further study handling distribution shift in clinical datasets. In addition, future research should look at specific model training parameters and the datasets' size, quality, and quantity that contribute to these variations.

7 Conclusion

Robust machine learning aims to create techniques that consistently work in different environments.To this end, we propose a testbed called CLIFT to evaluate the clinical deep learning model's performance under the distributional shift of the different test sets. We evaluate this testbed on several algorithms trained on the emrQA dataset and find that despite impressive results on the original test set with no adaptive overfitting, the performance degrades when applied to new test sets. Moreover, In this study, we performed an empirical analysis on emrQA and analyse the proposed test sets on difficulty metrics such as answer diversity and Syntactic divergence.We anticipate that our extensive testbed will indicate progress towards trustworthy machine learning in the healthcare area.

References

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. pages 1870–1879. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-1171.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. pages 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423. URL `https://aclanthology.org/N19-1423/`.

Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *ArXiv*, abs/1904.05342, 2019.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035, 12 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.35.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 9 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/btz682.

John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. The effect of natural distribution shift on question answering models. volume 119, pages 6905–6916. PMLR, 10 2020. URL `https://proceedings.mlr.press/v119/miller20a.html`.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. emrqa: A large corpus for question answering on electronic medical records. pages 2357–2368. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1258.

Junwoo Park, Youngwoo Cho, Haneol Lee, Jaegul Choo, and E. Choi. Knowledge graph-based question answering with electronic health records. In *MLHC*, 2021.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 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, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL `https://aclanthology.org/D16-1264`.

Bhanu Pratap Singh Rawat, Wei-Hung Weng, So Yeon Min, Preethi Raghavan, and Peter Szolovits. Entity-enriched neural models for clinical question answering. pages 112–122. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.bionlp-1.12.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019. URL `http://arxiv.org/abs/1910.01108`.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. *ArXiv*, abs/1611.01603, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural

language processing. pages 38–45. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-demos.6.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *ArXiv*, abs/1804.09541, 2018.

Xiang Yue, Bernal Jimenez Gutierrez, and Huan Sun. Clinical reading comprehension: A thorough analysis of the emrqa dataset. *ArXiv*, abs/2005.00574, 2020.

Zhuosheng Zhang, Junjie Yang, and Hai Zhao. Retrospective reader for machine reading comprehension. 2021.

A Baseline Models

In our baseline studies, we take into account ten existing models. These are developed on several pre-trained language models employing the Transformers architecture Vaswani et al. [2017]. Notably, BioBERT Lee et al. [2019], and variant BioBERT(+TAV) where TAV stands for Threshold-based Answerable Verification. The objective of verification is to determine the deep reader's verification score. EHReaderZhang et al. [2021], DistilBERT Sanh et al. [2019] and it's variants DistilBERT (Frozen) , DistilBERT(+TAV) and QuickReader-distilBERT Zhang et al. [2021] and a bert model trained on clinical domain, ClinicalBERT Huang et al. [2019]. Moreover, we also evaluated the datasets on BiDAF Seo et al. [2017], and QANet Yu et al. [2018].

We finetuned the models on the training dataset of emrQA and evaluated the proposed test sets.

B In-depth Analysis

Based on the outcomes of our experiments and model evaluation, we can say that the proposed testbed is more challenging & complex than the original emrQA test set. In order to support this argument & take a closer look at the proposed test sets, We use the three difficulty metrics provided by Rajpurkar et al. [2016] to compare the original emrQA test set to our four new test sets.

B.1 Answer Diversity

We used SpacyHonnibal and Montani [2017] to classify each answer into numerical and non-numerical replies, clinical named entities, and constituents. Figure 4 shows the Histograms of different reasoning types required to answer the questions. Interestingly, the histogram of the old test set and the new smoke dataset exhibit quite similar properties.

B.2 Syntactic divergence

Another metric of example difficulty is syntactic divergence, which assesses how similar the syntactic dependency tree structures of the question and response phrases are. Figure 2 shows the Syntactic divergence histograms for our new test sets and the emrQA test set.

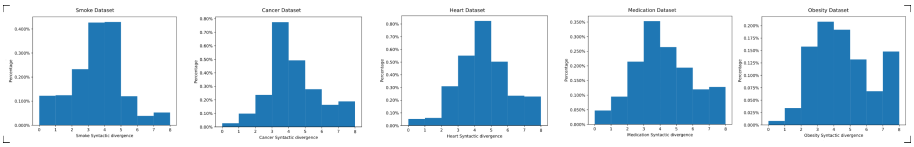

Figure 2: Syntactic divergence histograms for the proposed datasets' question and response phrases. Syntactic divergence distributions in Medication and Heart are quite similar. However, the Smoke and Cancer datasets have a more significant number of question-answer combinations with minor syntactic differences.

	EM Score			F1 Score		
Model	**emrQA EM**	**Clift EM**	**Gap(δ)**	**emrQA F1**	**Clift F1**	**Gap (δ)**
BioBERT	46.40	44.90	1.5	49.20	46.50	2.7
EHReader	50.00	48.50	1.5	50.00	47.50	2.5
DistilBERT (Frozen)	50.00	49.30	0.7	50.00	48.10	1.9
BiDAF	50.74	48.44	2.3	52.27	49.97	2.3
QANet	50.68	48.18	2.5	52.56	50.66	1.9
BioBERT(+TAV)	51.20	50.50	0.7	53.10	50.40	2.7
ClinicalBERT	53.33	52.63	0.7	54.06	52.56	1.5
DistilBERT	94.00	92.50	1.5	94.60	92.30	2.3
DistilBERT(+TAV)	94.40	91.70	2.7	95.00	92.70	2.3
QuickReader-distilBERT	94.60	93.10	1.5	95.20	94.50	0.7

Table 2: Performance of different models on the emrQA Testset and new Smoke Testset

	EM Score			F1 Score		
Model	**emrQA EM**	**Clift EM**	**Gap(δ)**	**emrQA F1**	**Clift F1**	**Gap (δ)**
BioBERT	46.40	43.50	2.9	49.20	46.30	2.9
EHReader	50.00	47.80	2.2	50.00	48.90	1.1
DistilBERT (Frozen)	50.00	48.10	1.9	50.00	48.10	1.9
BiDAF	50.74	47.84	2.9	52.27	51.17	1.1
QANet	50.68	48.48	2.2	52.56	51.46	1.1
BioBERT(+TAV)	51.20	50.10	1.1	53.10	50.90	2.2
ClinicalBERT	53.33	50.93	2.4	54.06	52.16	1.9
DistilBERT	94.00	92.10	1.9	94.60	93.50	1.1
DistilBERT(+TAV)	94.40	92.20	2.2	95.00	92.80	2.2
QuickReader-distilBERT	94.60	91.70	2.9	95.20	93.00	2.2

Table 3: Performance of different models on the emrQA Testset and new Cancer Testset

	EM Score			F1 Score		
Model	**emrQA EM**	**Clift EM**	**Gap(δ)**	**emrQA F1**	**Clift F1**	**Gap (δ)**
BioBERT	46.40	44.20	2.2	49.20	48.10	1.1
EHReader	50.00	47.50	2.5	50.00	48.90	1.1
DistilBERT (Frozen)	50.00	47.60	2.4	50.00	47.80	2.2
BiDAF	50.74	48.34	2.4	52.27	51.17	1.1
QANet	50.68	46.78	3.9	52.56	50.16	2.4
BioBERT(+TAV)	51.20	48.70	2.5	53.10	50.90	2.2
ClinicalBERT	53.33	51.13	2.2	54.06	50.16	3.9
DistilBERT	94.00	92.90	1.1	94.60	93.20	1.4
DistilBERT(+TAV)	94.40	93.00	1.4	95.00	91.10	3.9
QuickReader-distilBERT	94.60	92.40	2.2	95.20	93.80	1.4

Table 4: Performance of different models on the emrQA Testset and new Obesity Testset

	EM Score			F1 Score		
Model	**emrQA EM**	**Clift EM**	**Gap(δ)**	**emrQA F1**	**Clift F1**	**Gap (δ)**
BioBERT	46.40	45.00	1.4	49.20	46.80	2.4
EHReader	50.00	48.90	1.1	50.00	46.10	3.9
DistilBERT (Frozen)	50.00	48.90	1.1	50.00	48.90	1.1
BiDAF	50.74	48.34	2.4	52.27	48.37	3.9
QANet	50.68	49.58	1.1	52.56	50.36	2.2
BioBERT(+TAV)	51.20	47.30	3.9	53.10	50.70	2.4
ClinicalBERT	53.33	52.23	1.1	54.06	51.56	2.5
DistilBERT	94.00	91.50	2.5	94.60	90.70	3.9
DistilBERT(+TAV)	94.40	93.00	1.4	95.00	93.60	1.4
QuickReader-distilBERT	94.60	93.20	1.4	95.20	91.30	3.9

Table 5: Performance of different models on the emrQA Testset and new Heart Testset

	Smoke	Heart	Medication	Obesity	Cancer
Question #	5k	5k	5k	4k	5k
Vocab	53053	51810	44136	54429	51545
Min Q tokens	2	3	3	3	3
Min A tokens	1	2	1	1	1
Min C tokens	100	200	100	100	100
Avg Q tokens	6.42	8.31	7.61	7.19	8.40
Avg A tokens	4.59	4.15	3.93	4.04	4.12
Avg C tokens	217.33	234.18	215.49	212.88	210.16

Table 6: Proposed datasets' statistics, where Q, A, and C represent the Question, Answer, and Context, respectively.

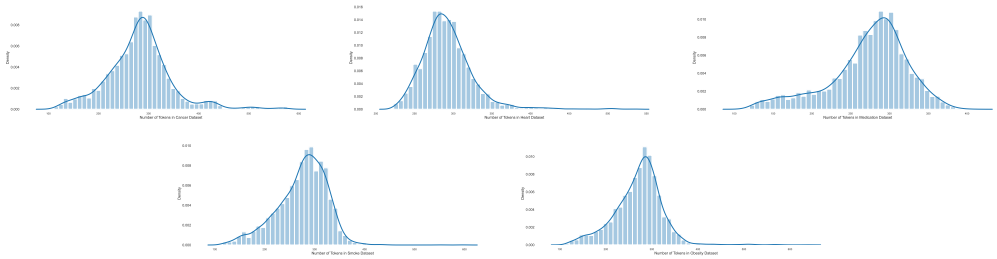

Figure 3: Subfigure (a), Subfigure (b), Subfigure (c), Subfigure (d), and Subfigure (e) shows the Distribution of the context length in the Cancer, Heart, Medication, Smoke and Obesity dataset, respectively

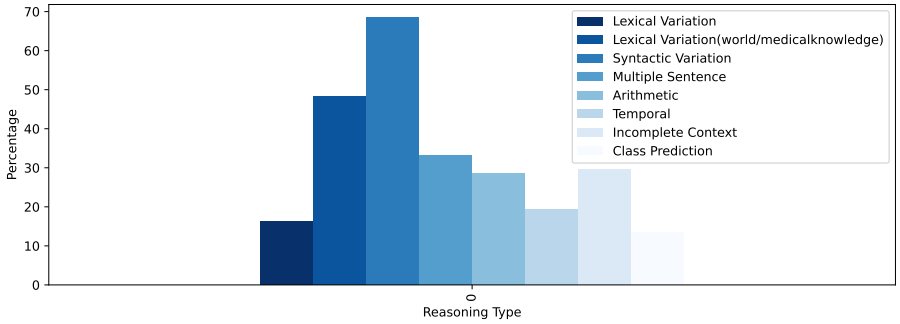

Figure 4: Relative sizes of Reasoning Types in the proposed datasets.