This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

DermaVQA: A Multilingual Visual Question Answering Dataset for Dermatology

Wen-wai Yim[1], Yujuan Fu[2], Zhaoyi Sun[2], Asma Ben Abacha[1], Meliha Yetisgen[2], and Fei Xia[2]

[1] Microsoft Health AI, Redmond, WA 98052, USA
{yimwenwai,abenabacha}@microsoft.com
[2] University of Washington, Seattle, WA 98195, USA
{velvinfu,zhaoyis,melihay,fxia}@uw.edu

Abstract. Remote medical care has become commonplace with the establishment of patient portals, the maturation of web technologies, and the proliferation of personal devices. However, though on-demand care provides convenience and expands patient access, this same phenomenon may lead to increased workload for healthcare providers. Drafting candidate responses may help speed up physician workflows answering electronic messages. One specialty that may benefit from the latest multi-modal vision-language foundational models is dermatology. However, there is no existing dataset that incorporate dermatological health queries along with user-generated images. In this work, we contribute a new dataset, DermaVQA(https://osf.io/72rp3/), for the task of dermatology question answering and we benchmark the performance of state-of-the-art multi-modal models on multilingual response generation using relevant multi-reference metrics. The dataset and corresponding code are available on our project's GitHub repository [3].

Keywords: VQA · Response Generation · Dermatology

1 Introduction

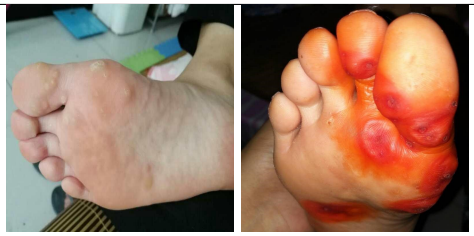
Rapid development of telecommunication technologies, increased demands for healthcare services, and recent COVID-19 pandemic needs, have accelerated the adoption of remote clinical diagnosis and treatment. In addition to live meetings with doctors – which may be conducted through telephone or video – asynchronous options such as e-visits, emails, and messaging chats have also been proven to be safe, effective, and convenient [5]. However, the burden of electronic messaging may unintentionally adversely impact physician workload by introducing a new and "never-ending" source of additional work items [19]. Automatic response generation can be used to produce draft responses as suggestions to increase doctor efficiency. In this task, we focus on the problem of dermatology multi-modal query response generation for Chinese and English. Input

[3] https://github.com/velvinnn/DermaVQA

includes textual context which provides clinical context and natural language questions, as well as one or more images (cf. Table 1). Although there is prior work in textual consumer health question answering, visual question answering, and dermatology image classification, to the best of our knowledge, this is the first work to study the problem of automatic clinical response generation, given user-generated images and queries on dermatology. Our contributions include the following:

- A new dataset of dermatology consumer health questions and image inputs as well as multiple doctor responses for each query.
- Our annotation methodology for processing multi-threaded consumer health platform conversation threads to question-answer pairs.
- Code and reported baselines using state-of-the-art models for this task.

Table 1. Example from the new DermaVQA dataset

Query	Responses
 帮忙诊断一下:三个月前出现如下图，自己用达克能宁喷雾两个月无明显效果，之后去乡村诊所，医生指导用鸡眼膏，之后出现变红变多，请帮忙诊断下 Three months ago, the condition shown in the picture below appeared. The patient used Daknening spray for two months without any noticeable effect. Afterwards, they went to a rural clinic, where the doctor advised them to use corn ointment. Subsequently, the condition turned red and worsened. Please help with the diagnosis.	RESPONSE1: 是鸡眼。 It's a corn. RESPONSE2: 考虑：跖疣 Consideration: Plantar wart RESPONSE3: 是跖疣，不是鸡眼，激光治疗。 It's a plantar wart, not a corn. Laser treatment is recommended.

2 Related Work

2.1 Consumer Health Question Answering

The task of consumer health question answering involves generating responses to questions posted by health consumers (anyone with health information needs).
Health FAQs. Taking advantage of naturally occurring consumer health queries (CHQ) and quality answer responses from health professionals, several sources of CHQ datasets have been created. One of these works was the focus of the 2017

TREC LiveQA shared task [2], wherewithin 634 training and 104 test pairs of CHQ and associated answers were made available. During the challenge, system candidate responses were judged manually by experts according to the following levels: (a) Correct and Complete Answer, (b) Correct but Incomplete, (c) Incorrect but Related, and (d) Incorrect. In [3], the authors extended this data by adding entailed versions of questions and answers; as well as experimenting with their information retrieval system.

Online Doctor-Patient Conversations. With advances of remote communication technologies and the proliferation of personal devices, online medical consultation is an attractive viable solution. Typical data in this domain includes primarily two speakers exchanging textual messages over multiple turns. MedDialog [22] includes a Chinese dataset with 3.4 million online conversations between patients and doctors from Haodf.com; and an English dataset with 0.26 million conversations from iclinic.com and healthcaremagic.com. MedDG [16] includes 17k conversations related to gastrointestinal diseases from the Chunyu Yisheng platform.

Similar to previous work, our text comes from real user-generated questions which will contain informal text such as ungrammatical language and misspellings. Unlike previous work, our data includes images with the questions and are adapted from multi-party online conversations into two-turn question-answer pairs.

2.2 Visual Question Answering

The task of visual question answering (VQA) involves an image and a textual question. Datasets in this domain are constructed by employing freelance workers to generate questions related to a given image. One of the most widely used datasets is the VQA dataset [1] which includes 200,000 real images from the Microsoft Common Objects in Context (MSCOCO) dataset, labeled with three questions per image, and 10 ground truth answers. In the medical domain, several VQA datasets exist in the field of radiology. Here questions are produced by recruiting medical professionals; related works include VQA-RAD[13], VQA-Med[4], and SLAKE[15]. Other medical VQA datasets include PathVQA [10] with 32,795 questions asked from 4,998 pathology images, and MedVQA-GI [11] with 4k images and 80k question-answer pairs for GI tract images. VQA datasets traditionally have limited question topics and constrained answers, as shown by the answer length and the use of Accuracy as a common metric (cf. Table 2). Different from previous works, we focus on dermatological images; our data includes natural questions asked from real health consumers; our responses are natural response free-text with average answer lengths of 12+ words.

2.3 Dermatology Image Classification

Dermatology image classification includes disease classification (e.g., actinic keratoses) as well as attribute classification (e.g., scaly appearance, color). In ideal conditions, skin problems are photographed in controlled lighting, resolution,

Table 2. Visual Question Answering Corpora

Corpus	#images	#QA	Image Type	Answer Length	# Query Types	Metrics
MedVQA-GI	4k	80k	GI	–	text;binary; number	Accuracy
PathVQA	5k	32k	Pathology	2.5	wh-	Accuracy; BLEU;F1
SLAKE	642	14k	Radiology	–	closed; open-ended	Accuracy
Visual Genome	108k	1.7M	Photos	1.8	wh-	Accuracy
VQA	614k	7984k	Photos	1(89%);2(7%); 3(3%)	open-ended; multiple-choice	Accuracy
VQA-MED	4k	15k	Radiology	1(74%);2(2%); 3(20%);4(4%)	yes/no;wh-	Accuracy; BLEU
VQA-RAD	315	3515	Radiology	1.6	open-ended	BLEU
This Work	3434	1488	Dermatology	11.9(iiyi); 94.6(reddit)	open-ended	BLEU; Bertscore; MEDCON

and focal distances. Special devices such as dermatoscopes allow controlled images to be captured and saved. However both educational and textbook material, as well as personal consumer devices, may include images produced with varying lighting conditions, device resolution, and image clarity.

Dermatoscopic images. Corpora including dermatoscopic images are often collected in connection to clinical malignancy testing. Some of datasets including imaging and malignancy classifications include: HAM10000 [21] with 10,015 images and confirmed histopathologic diagnoses, the PH2 dataset[17], the derm7pt dataset with 2k diagnosis and seven-point checklist criteria labels [12], and the ISIC 2021 challenge set [18] with 13k images.

Table 3. Dermatology Diagnoses Classification Datasets

Corpus	Size	Source	#Disease	#Classes	Performance
DermNetNZ	19k	dermnetnz.org	23	23	–
Fitzpatrick17k [9]	16.6k	DermaAmin; Atlas Dermatologico	114	114	20 Acc
DDI[6]	656	Stanford Hospital	–	2	0.78 ROC-AUC
SD-128[20]	5.6k	DermQuest	729	128	52 Acc
SD-198[20]	6.5k	DermQuest	729	198	53 Acc
Glock et al[8]	1,316	Bing Web Search	13	2	95 Acc

Natural images. Use of natural images are common place for health consumers but also as a resource in textbooks and educational websites for healthcare professionals. In fact, with the exception of the DDI corpus[6], most corpora of this

nature arises from web platforms with images, voluntarily-uploaded or sourced from educational websites. Table 3 provides size and number of diseases and classification classes per dataset. In this work, outputs are free text responses to open-ended questions, rather than a closed classification task.

3 Corpus Creation

3.1 Data Collections

In order to evaluate multi-modal capabilities on dermatology images and queries, we collected and labeled data from two different platforms: IIYI (爱爱医) and Reddit.

IIYI Dataset. IIYI.com is a Chinese online platform which allows users to ask medical questions that can be answered by medical health professionals. A user can post their question along with relevant files and images. Multiple users may respond. User profiles include validation badges including medical license certification, as well as user level rankings based on past up-votes. The site is used for educational purposes. Posters include both patients and doctors. We crawled the threads from the forum 皮肤及性传播疾病讨论版. The final exact post ids can be found in our released dataset.

Reddit Dataset. Reddit.com similarly is an online platform where users may post queries and receive responses from other netizens. We used the subreddit *r/DermatologyQuestions*, written in English, and retrieved the posts with images, using PRAW (github.com/praw-dev/praw). The exact post ids can be found in our dataset.

3.2 Filtering and Processing

Threads with images of genitalia, identifiable features (e.g. special tattoos), image annotations (e.g. drawn arrows) or full-face photos were removed. For the IIYI dataset, post response without meaningful value (e.g. "I also want to know the answer to this") were removed; threads without images, responses, or had no meaningful responses were excluded. Detailed annotations guidelines are included in our data release.

3.3 Gold Standard Response Curation and Creation

As the IIYI threads originally include multi-party multi-turn conversations, both original poster-queries and post responses had to be modified if referring to other posts or if adding information. For example, posts that "I support {Post X}" would be replaced with the content from "{Post X}". If impossible to make a post context-free, these posts are excluded. If the original poster adds information or asks additional information, their posts would be concatenated and edited to a final query. Only responses appearing after the final query posts considered are kept. The test set required at least two valid responses. Validation and test

Table 4. DermaVQA: Characteristics of the IIYI and Reddit subsets

	IIYI	Reddit
SPLIT		
train	842	347
valid	56	50
test	100	93
total	998	490
SEX		
F	218	94
M	239	82
UNK	541	314
AGE		
mentions	398	50
mean	31	25
min	0.08	0.5
max	92	70
IMAGES		
total count	2944	490
mean count	2.9	1
encs-1 img	225	490
encs-2 img	285	0
encs-3 img	200	0
encs->=4 img	288	0

	IIYI	Reddit
RESPONSES		
total count	7214	583
mean count	7.2	1.2
encs-1 resp	66	397
encs-2 resp	91	93
encs-3 resp	110	0
encs->=4 resp	731	0
LENGTH		
query(en/zh)	79.6/87.6	30.1/–
responses(en/zh)	11.9/16.3	94.6/–
MEDICAL TOPICS		
Diagnosis (en)	704	383
Tests (en)	46	25
Treatments (en)	567	318
LOCATIONS		
Arm region	187	64
Back region	104	35
Chest/Abdomen region	124	40
Foot region	152	44
Hand region	271	74
Head region	203	175
Leg region	231	53

sets were translated to English by a medical translator. Because Reddit post responses may not include responses from medical doctors, dermatologists were hired to write answers. 93 encounters are double annotated while the rest are single annotated.
Queries were found to cover 1 or more question types including:

- **diagnosis**: open diagnosis query, e.g. *"What is this?"*
- **diagnosis_binary**: yes/no diagnosis query, e.g. *"Is this athlete's foot?"*
- **diagnosis_mc**: multiple choice diagnosis query, e.g. *"Perioral dermatitis or seborrheic dermatitis?"*
- **treatment**: treatment query, e.g. *"How do I treat this?"*
- **open-advice**: open-ended advice requests, e.g. *"What should I do now?"*
- **other**: specific questions not included in above, e.g. *"Does this look like it's healing?"*
- **no_explicit_question**: no requests/questions asked, e.g. *"I've had this for a year."*

We present various characteristics and statistics of the two datasets in Table 4. In addition to the query-answer format gold standard, annotators manually reviewed and collected the subject of the posts' age and sex when available. During the Reddit dataset labeling, dermatologists also rated the Fitzpatrick Skin Type, resulting in final frequencies of I(44), II(323), III(179), IV(33), V(6), VI(0). A

formal internal review related to data collection, usage, and release for legal and ethical compliance was conducted and approved. The dataset and annotation guidelines can found here: https://osf.io/72rp3/.

4 Methodology

4.1 Baseline Systems

We measure the performance of several state-of-the-art systems on this dataset. Exact prompts and algorithms are included in our code-base.

- ***Gemini-Pro Vision.*** The latest Google LLM allowing both textual and image inputs at the time of experimentation. We accessed this model at cloud.google.com/vertex-ai/docs/gen erative-ai/multimodal/overview.
- ***GPT-4-vision-preview.*** The latest GPT model allowing both textual and image inputs from openai.com at the time of experimentation.
- ***LLaVA-FT+GPT4.*** We built a VQA system for dermatology questions and images that includes two components: (a) an image-to-text diagnosis text generator and (b) a text-to-text response generator, given the diagnosis from (a) and the user query. The image-to-text model, LLaVA-Med [14], was first fine-tuned on a subset of Fitzpatrick[9] with diagnoses contained in the training and validation sets, then on image-diagnosis pairs extracted from the training sets. GPT-4 was used to generate the final response.

4.2 Evaluation Metrics

We evaluate the system responses by comparing with the multiple gold standard responses per query. We used relevant multi-reference metrics/variants including:

deltaBLEU. deltaBLEU is a variant of SacreBLEU developed for response generation, a case in which many diverse gold standard responses are possible [7]. It takes into account a human-annotated quality rating and gives higher weights to n-grams from responses rated to be of higher quality. The authors have shown this method produces higher correlation with human rankings compared to previous BLEU metrics. In our system, we assign response weights according to four criteria: (a) if user expertise level is 4 or above (out of 9), (b) if user is formally validated as a medical doctor by the platform, (c) if the response answer is the most frequent answer, and (d) if the response answers the query completely. The former two were manually assigned to the validation and test sets by two NLP scientists. The test set was double-reviewed. Out of a 0.0-1.0 scale, if (d) is not met, the score is discounted to 0.9; for the other 3 criteria, 0.1 is discounted for every missing element to reach the final weight.

BERTScore. BERTScore[4] averages the maximum word embedding similarity scores between two texts based on BERT embeddings. It has been shown to work

[4] github.com/Tiiiger/bert_score

well on a variety of tasks, including image captioning and machine translation. The maximum was taken over multiple pairwise scores when multiple references were present.

MEDCON. MEDCON uses QuickUMLS[5] to identify medical concepts in conjunction with an in-house llama-based assertion classifier. Concepts identified by QuickUMLS are normalized according to a curated concept map. Precision, recall, and F1 were calculated based on combined concept and assertion statuses. The maximum was taken over multiple pairwise scores when multiple references were present.

5 Results

Table 5 shows the results of the three models on the iiyi and Reddit test sets. We found that the Gemini-ProVision model had the highest performance across deltaBLEU and MEDCON metrics for the iiyi Chinese and English datasets (9.3 and 1.1 deltaBLEU scores). Performance on the Reddit dataset across systems seemed comparable without noticeably large score differences across all metrics. GPT-4-vision achieved the highest BERTScore scores on the iiyi Chinese and English test sets. Our third system, LLaVA-FT+GPT, achieved the highest BERTScore on the Reddit dataset.

Table 5. Baseline Results

		CHINESE			ENGLISH	
model	deltaBLEU	BERTScore	medcon	deltaBLEU	BERTScore	medcon
iiyi						
Gemini-ProVision	**9.311**	0.685	**0.107**	**1.094**	0.800	**0.157**
gpt-4-vision	7.025	**0.735**	0.016	0.813	**0.867**	0.083
LLaVA-FT+GPT	2.185	0.654	0.068	0.385	0.840	0.120
reddit						
Gemini-ProVision	–	–	–	**3.399**	0.766	0.132
gpt-4-vision	–	–	–	2.709	0.854	**0.147**
LLaVA-FT+GPT	–	–	–	3.390	**0.856**	0.133

Although the results are modest, these scores are consistent with previous tasks. For example in dialogue generation, in [22] the authors report a BERT-GPT BLEU-4 score of 1.8%. As shown in Table 2, VQA task performances vary widely depending on the question type, method of evaluation, and corpus. For example, within the VQA-RAD dataset, some question types received accuracy scores ranging 0-81% depending on question type; while BLEU scores ranged from modest 0.0058 to 0.1047. Compared to other dermatology classification results (Table 3), performances vary depending on the number of classification categories used. In fact in [9], the authors found 20% accuracy when classifying

[5] github.com/Georgetown-IR-Lab/QuickUMLS

against 114 skin conditions; which jumps up to 62.4% if simplifying to three categories (non-neoplastic, benign, and malignant).

6 Conclusions and Future Works

Despite a plethora of past-related work, the area of user-generated dermatological VQA remains a challenging task, requiring much further research. Future extensions of this work may incorporate intermediate steps such as: image segmentation and normalization, characteristic extraction, and text re-ranking steps. Additional possible methods of improving model results include experimenting with other foundation models, such as BiomedCLIP, and fine-tuning synthetically-produced dermatology VQA data. To improve evaluation, future metrics may utilize automatic ways to measure question completeness, improve medical term normalization, and weigh answers according to similarity with other answers or consistency with external knowledge bases. Real-world testing of recommendation response drafts for health applications by domain experts is also a critical future work. As the technology matures, conducting usability studies will be paramount to measuring impact in real scenarios. We hope that the work here can be used to develop and benchmark new medical systems for multi-modal response generation that can alleviate doctor burden.

Acknowledgments. We would like to thank Thomas Lin from Microsoft Health AI and the ClinicalNLP and ImageCLEF organizers for their feedback and support for the MEDIQA-M3G and MEDIQA-MAGIC 2024 shared tasks, which utilized this dataset. We also thank our diverse annotation team for preparing the data in time for the challenge and all the participating teams who contributed to the success of these shared tasks through their interesting approaches and experiments and strong engagement.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C.L., Parikh, D., Batra, D.: Vqa: Visual question answering. International Journal of Computer Vision **123**, 4 – 31 (2015)
2. Ben Abacha, A., Agichtein, E., Pinter, Y., Demner-Fushman, D.: Overview of the medical question answering task at trec 2017 liveqa. In: TREC 2017 (2017)
3. Ben Abacha, A., Demner-Fushman, D.: A question-entailment approach to question answering. BMC Bioinformatics **20** (2019)
4. Ben Abacha, A., Hasan, S.A., Datla, V.V., Demner-Fushman, D., Müller, H.: Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In: Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes. 9-12 September 2019 (2019)
5. Bishop, T.F., Press, M.J., Mendelsohn, J.L., Casalino, L.P.: Electronic communication improves access, but barriers to its widespread adoption remain **32**(8), 10.1377/hlthaff.2012.1151

6. Daneshjou, R., Vodrahalli, K., Liang, W., Novoa, R.A., Jenkins, M., Rotemberg, V., Ko, J.M., Swetter, S.M., Bailey, E.E., Gevaert, O., Mukherjee, P., Phung, M., Yekrang, K., Fong, B., Sahasrabudhe, R., Zou, J., Chiou, A.S.: Disparities in dermatology ai performance on a diverse, curated clinical image set. Science Advances **8** (2021)
7. Galley, M., Brockett, C., Sordoni, A., Ji, Y., Auli, M., Quirk, C., Mitchell, M., Gao, J., Dolan, B.: deltaBLEU: A discriminative metric for generation tasks with intrinsically diverse targets. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 445–450. Association for Computational Linguistics, Beijing, China (Jul 2015)
8. Glock, K., Napier, C., Gary, T., Gupta, V., Gigante, J., Schaffner, W., Wang, Q.: Measles rash identification using transfer learning and deep convolutional neural networks. In: 2021 IEEE International Conference on Big Data (Big Data). pp. 3905–3910 (2021)
9. Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., Badri, O.: Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1820–1828 (2021)
10. He, X., Cai, Z., Wei, W., Zhang, Y., Mou, L., Xing, E., Xie, P.: Towards visual question answering on pathology images. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 708–718. Association for Computational Linguistics, Online (Aug 2021)
11. Hicks, S., Storås, A.M., Halvorsen, P., de Lange, T., Riegler, M., Thambawita, V.L.: Overview of imageclefmedical 2023 - medical visual question answering for gastrointestinal tract. In: Conference and Labs of the Evaluation Forum (2023)
12. Kawahara, J., Daneshvar, S., Argenziano, G., Hamarneh, G.: Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE Journal of Biomedical and Health Informatics **23**(2), 538–546 (2019)
13. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific data **5**(1), 1–10 (2018)
14. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. arXiv preprint arXiv:2306.00890 (2023)
15. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y.F., Wu, X.M.: Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) pp. 1650–1654 (2021)
16. Liu, W., Tang, J., Cheng, Y., Li, W., Zheng, Y., Liang, X.: Meddg: An entity-centric medical consultation dataset for entity-aware medical dialogue generation. In: Natural Language Processing and Chinese Computing (2020)
17. Mendonça, T., Ferreira, P.M., Marques, J.S., Marcal, A.R., Rozeira, J.: Ph 2-a dermoscopic image database for research and benchmarking. In: 2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC). pp. 5437–5440. IEEE (2013)
18. Rotemberg, V., Kurtansky, N., Betz-Stablein, B., Caffery, L., Chousakos, E., Codella, N., Combalia, M., Dusza, S., Guitera, P., Gutman, D., et al.: A patient-centric dataset of images and metadata for identifying melanomas using clinical context. Scientific data **8**(1), 34 (2021)

19. Sinsky, C.A., Shanafelt, T.D., Ripp, J.A.: The electronic health record inbox: Recommendations for relief **37**(15), 4002–4003
20. Sun, X., Yang, J., Sun, M., Wang, K.: A benchmark for automatic visual classification of clinical skin disease images. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI 14. pp. 206–222. Springer (2016)
21. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**(1), 1–9 (2018)
22. Zeng, G., Yang, W., Ju, Z., Yang, Y., Wang, S., Zhang, R., Zhou, M., Zeng, J., Dong, X., Zhang, R., Fang, H., Zhu, P., Chen, S., Xie, P.: MedDialog: Large-scale medical dialogue datasets. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 9241–9250. Association for Computational Linguistics, Online (Nov 2020)