Faithful AI in Healthcare and Medicine

Qianqian Xie[1] **and Fei Wang**[1]

[1]Department of Population Health Science, Weill Cornell Medicine, Cornell University

ABSTRACT

Artificial intelligence (AI) holds great promise in healthcare and medicine on being able to help with many aspects, from biological scientific discovery, to clinical patient care, to public health policy making. However, the potential risk of AI methods for generating factually incorrect or unfaithful information is a big concern, which could result in serious consequences. This review aims to provide a comprehensive overview of the faithfulness problem in existing research on AI in healthcare and medicine, including analysis of the cause of unfaithful results, evaluation metrics, and mitigation methods. We will systematically review the recent progress in optimizing the factuality in various generative medical AI methods, including knowledge grounded large language models, text-to-text generation tasks such as medical text summarization, medical text simplification, multimodality-to-text generation tasks such as radiology report generation, and automatic medical fact-checking. The challenges and limitations of ensuring the faithfulness of AI-generated information in these applications, as well as forthcoming opportunities will be discussed. We expect this review to help researchers and practitioners understand the faithfulness problem in AI-generated information in healthcare and medicine, as well as the recent progress and challenges on related research. Our review can also serve as a guide for researchers and practitioners who are interested in applying AI in medicine and healthcare.

1 Introduction

Artificial intelligence (AI) has been gradually applied in different aspects of healthcare and medicine[1–3]. It has demonstrated a positive impact on a wide range of areas including accelerating medical research, aiding disease detection and diagnosis, providing personalized health recommendations, and so on (as shown in Figure 2). The success of Medical AI has been closely tied to the development of fundamental AI algorithms and the availability of various biomedical data. For example, deep learning [4] has shown great performance on various medical image analysis tasks[5].

Recently, pre-trained language models (PLMs)[6,7], which are pre-trained on a huge amount of unlabelled language texts in a self-supervised learning manner (typically based on the new backbone neural network architecture called Transformer[8]), has become the new paradigm of natural language processing (NLP) and demonstrated strong potential in Medical AI. PLMs have shown high accuracy and strong generalization ability in few-shot learning and even zero-shot learning scenarios in various medical tasks[9,10], which went beyond NLP. For example, AlphaFold[11], a deep learning computational method based on Transformer developed by DeepMind, has been shown to be able to predict the protein structure with near experimental accuracy. More recently, large generative language models[12], such as ChatGPT[13] and GPT-4[14] developed by OpenAI, have exhibited strong capabilities of understanding complex input in natural language and produce near human-writing text contents. ChatGPT has been reported to be close to pass the threshold of the United States Medical Licensing Exam (USMLE) and showed great potentials on assisting clinical decision-making[15].

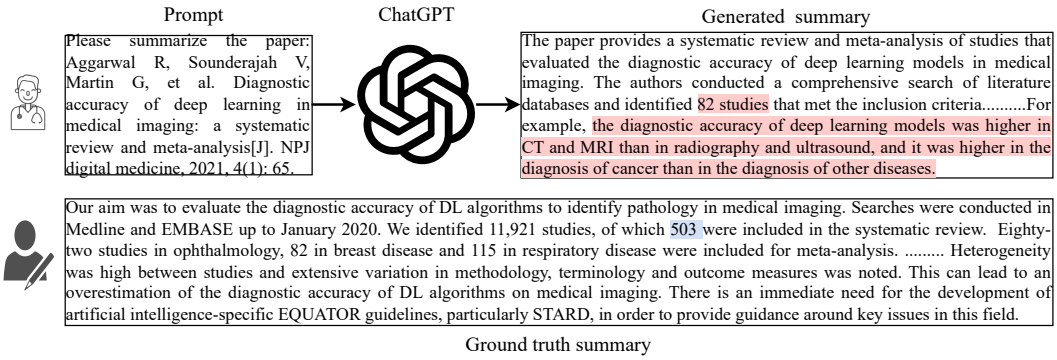

Figure 1. The example of the generated content with factual errors from ChatGPT.

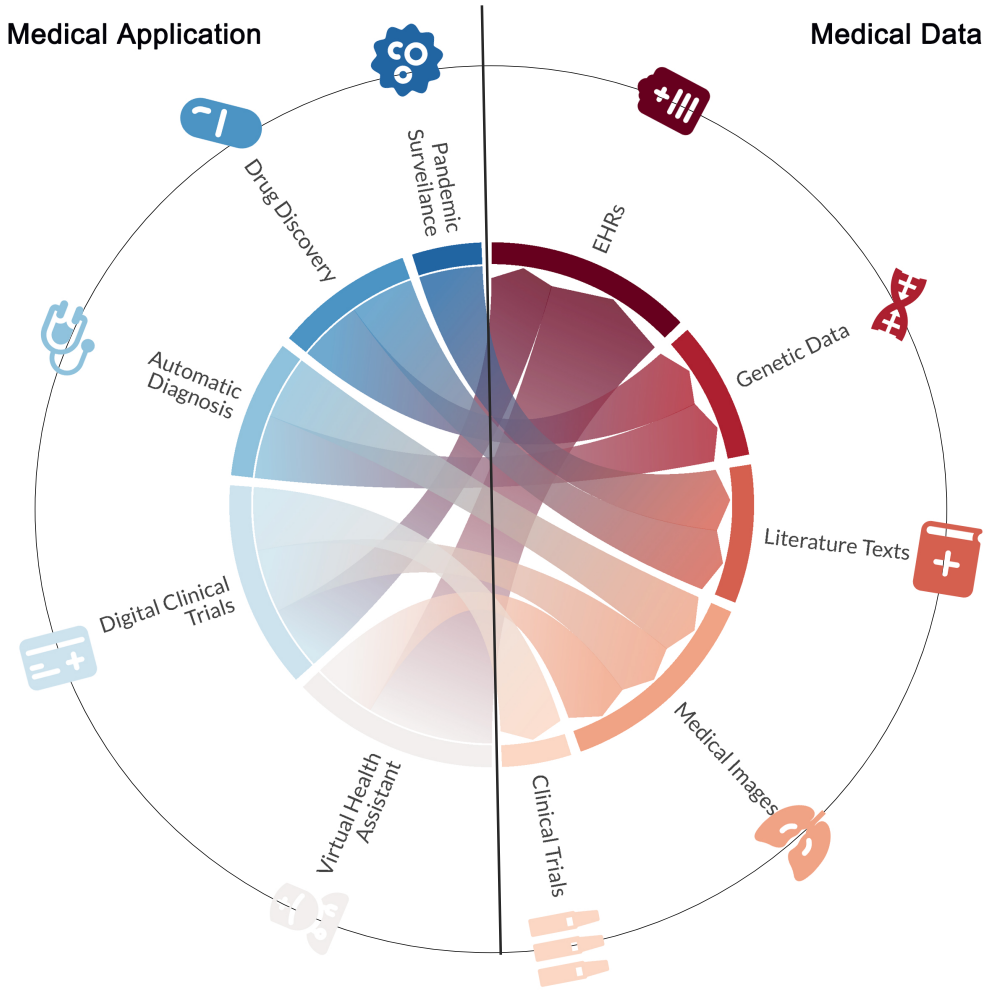

Figure 2. The example of medical data and realistic medical applications with AI.

Despite the promises of medical AI[16], one major concern that has attracted a lot of attention is its potential risk of generating non-factual or unfaithful information[17], which is usually referred to as the faithfulness problem[18]. Specifically, generative AI methods can generate contents that are factually inaccurate or biased. For example, in Figure 1, we input the natural language texts to ask ChatGPT to summarize a systematic review paper named "Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis" published in npj Digital Medicine 2021[19]. ChatGPT generates the summary with the intrinsic factual error that contradicts the content of the review, such as it showing 82 studies met the inclusion criteria while the review is actually based on 503 studies. The generated summary also has the extrinsic factual error that can't be supported by the review, such as "the diagnostic accuracy of deep learning models was higher in computed tomography (CT) and magnetic resonance imaging (MRI) than in radiography and ultrasound, and it was higher in the diagnosis of cancer than in the diagnosis of other diseases". This could mislead researchers and practitioners and lead to unintended consequences[20].

In this review, we provide an overview of the research on faithfulness problem in existing medical AI studies, including cause analysis, evaluation metrics, and mitigation methods. We comprehensively summarize the recent progress of maintaining factual correctness in various generative medical AI models including knowledge grounded large language models, text-to-text generation tasks such as medical text summarization and simplification, multimodality-to-text generation tasks such as radiology report generation, and automatic medical fact-checking. We discuss the challenges of maintaining the faithfulness of AI-generated information in the medical domain, as well as the forthcoming opportunities. The goal of this review is to provide researchers and practitioners with a blueprint of the research progress on the faithfulness problem in medical AI, help them understand the importance and challenges of the faithfulness problem, and offer them guidance for using AI methods in medical practice and future research.

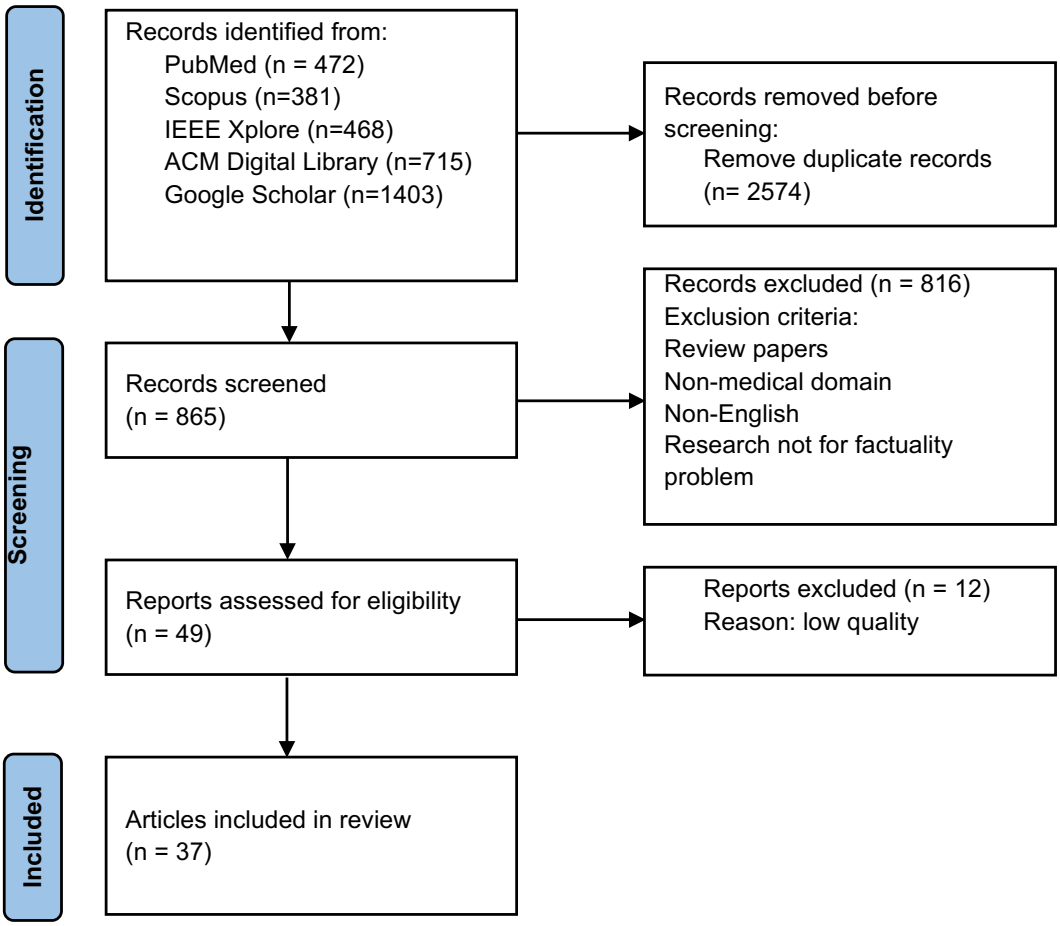

Figure 3. The process of article selection.

2 Search Methodology

2.1 Search Strategy

We conducted a comprehensive search for articles published between January 2018 to March 2023 from multiple databases including PubMed, Scopus, IEEE Xplore, ACM Digital Library, and Google Scholar. We used two groups of search queries: 1) faithful biomedical language models: factuality/faithfulness/hallucination, biomedical/medical/clinical language models, biomedical/medical/clinical knowledge, 2) mitigation methods and evaluation metrics: factuality/faithfulness/hallucination, evaluation metrics, biomedical/medical/clinical summarization, biomedical/medical/clinical text simplification; radiology report summarization, radiology report generation, medical fact-checking.

2.2 Filtering Strategy

A total of 3,439 records were retrieved from five databases. 865 articles were kept after removing duplicate records, which were further screened based on title and abstract. We applied the exclusion criteria during the filtering process (also shown in Figure 3): 1) the article is a review paper, 2) the main content of the article is not in English, 3) the article is not related to the medical domain, 4) the article is not relevant to the factuality problem. As a result, 49 articles were retained after the screening process. Then, we conducted a full-text review of these articles and further excluded 12 articles due to the low quality of the content.

3 What is Faithfulness Problem ?

3.1 Definition and Categorization

Faithfulness generally means being loyal to something or some person. In the context of AI or NLP, faithful AI means the algorithm can produce contents that are factually correct, namely staying faithful to facts[18,21]. Generative medical AI systems[11,22,23] learn to map from various types of medical data such as electronic health records (EHRs), medical images or protein sequences, to desired output such as the summarization or explanation of medical scans, radiology reports, and

Table 1. The reference of different exemplar generative medical tasks for evaluating the factual consistency of the generated output by AI systems.

Task	Input	Output	Reference
medical text summarization	medical texts	short summarization	input
radiology report summarization	radiology report	impression	impression from radiologist
medical text simplification	medical texts	simplified texts	input
medical dialogue generation	dialogue from patients	response	dialogue history
medical question answering	medical question	answering	clinician-generated answer
radiology report generation	Chest X-ray image	radiology report	radiology report from radiologist

three-dimensional (3D) protein structures. We refer a medical AI systems to be unfaithful or has a factual inconsistency issue (which is also called "hallucination" in some studies[21]) if it produces content that is not supported by existing knowledge, reference or data. Similar to the general generative AI methods, the factual inconsistency issue in generative medical AI generally can also be categorized into the following two major types[24]:

- **Intrinsic Error:** the generated output contradicts with existing knowledge, reference or data. For example, if there is a sentence in the radiology report summary generated by the AI system "The left-sided pleural effusion has increased in size", but the actual sentence in the radiology report is "There is no left pleural effusion". Then the summary contradicts facts contained in the input data.

- **Extrinsic Error:** the generated output cannot be confirmed (either supported or contradicted) by existing knowledge, reference or data. For example, if the content of the radiology report includes: "There is associated right basilar atelectasis/scarring, also stable. Healed right rib fractures are noted. On the left, there is persistent apical pleural thickening and apical scarring. Linear opacities projecting over the lower lobe are also compatible with scarring, unchanged. There is no left pleural effusion", and the generated summary includes: "Large right pleural effusion is unchanged in size". Then this piece of information cannot be verified by the given radiology report, since the information about "right pleural effusion" is not mentioned there.

Different from general applications, the tolerance of factual incorrectness should be relatively low for medical and healthcare tasks to due to their high-stake nature. Different medical tasks have different references on accessing the factuality of the generated output. For the AI based text summarization[22] or simplification systems[25] that summarize or simplify the input medical texts such as study protocols, biomedical literature, or clinical notes, the reference is usually the input medical document itself, which ensures the AI generated content is faithful to its original information. For AI systems that automatically generate radiology reports with the input Chest X-ray image[26], the reference is the radiology report of the input Chest X-ray image written by the radiologist. In table 4, we summarize the reference of different exemplar generative medical tasks for evaluating the factual consistency of the generated output by AI systems.

3.2 Why There is the Faithfulness Problem?

The factual inconsistency of medical AI systems can be attributed to a wide range of reasons.

3.2.1 Limitation of backbone language model

Pre-trained language models (PLMs), which have been the dominant backbone model for representing medical data in various types and modalities, have limitations on generalization and representation abilities for medical data. It has been shown that the large language models (LLMs) such as GPT-3[12] has poor performance of few-shot learning on information extraction (IE) tasks of biomedical texts[27,28]. This is because these LLMs were not pre-trained on medical data, thus they lack the representability there and cannot capture medical knowledge effectively. There have been domain-specific language models trained with biomedical data such as BioBERT[29], PubMedBERT[30], BioGPT[23] et al. either with fine-tuning or from scratch, but the scale of these data are much smaller than those used for training PLMs in general domains. For example, it has been shown that domain-specific language models such BioBERT[27] has better performance than GPT-3[12] on the few-shot learning of IE tasks, but it still has a large gap with the performance of fine-tuning based methods. In addition, most biomedical texts such as EHRs, and radiology report texts have limited availability due to privacy concerns. As a result, existing domain-specific language models were mainly pre-trained with biomedical literature texts, that are easy to be accessed on large scale. Therefore, it is challenging to pre-train LLMs with biomedical domain specific data that can generalize well and encode comprehensive medical knowledge. Overall, both existing biomedical PLMs and powerful LLMs in the general domain including ChatGPT and GPT-4 are not effective enough in encoding biomedical knowledge[31]. This inevitably leads to factual inconsistencies in various medical AI systems that uses these PLMs and LLMs as the backbone model.

3.2.2 Lacking of data sources

Fine-tuning has been the main paradigm for adapting PLMs to various biomedical tasks[7]. It transfers the semantic information and knowledge encoded in PLMs and to the targeted tasks via supervised learning with labeled data. Similar to conventional deep learning methods[4], it relies on a sufficient amount of annotated medical data to achieve desirable performance. However, high-quality annotated medical data are both expensive and time consuming to obtain. This is another reason why medical AI systems can be vulnerable to factual errors.

3.2.3 Data discrepancy

The factual inconsistency problem also arises from the data discrepancy between the ground truth output used for training and the reference in most medical tasks. For example, in the task of generating lay summaries for technical biomedical scientific papers, the ground truth outputs namely lay summaries usually include additional information for ease the understanding of lay people, which is not mentioned in the reference namely the original biomedical scientific papers. For the medical dialogue generation tasks, the ground truth output responses don't always faithful to the references namely dialogue histories, since the topic of responses can be transferred and diverse according to input dialogues from patients. Thus, the medical AI systems that are trained by generating the output with minimized divergence with the ground truth output, can produce an output that is not faithful with the reference.

3.2.4 Limitations of decoding

Another cause for the factual inconsistency problem is the limitation of the decoding strategy used by medical AI systems for text generation. There exists the exposure bias problem[32,33], which is the discrepancy between training and inference on the decoding process based on the decoder. Specifically, during training, the decoder is encouraged to generate the next token on the conditional of the previous ground truth tokens. While during inference, the ground truth tokens are unobservable, and thus the decoder can only be conditioned on the previously generated tokens by itself to predict the next token. Such discrepancy between training and inference has been shown to potentially result in factual errors of outputs[24]. Moreover, existing methods usually use the sampling-based decoding strategy such as beam search and greedy search to improve the diversity of outputs, and they have not considered the factual consistency when selecting tokens and candidate outputs. The randomness of selecting tokens and candidate outputs can increase the probability of generating outputs with factual errors[34].

4 Evaluating and Optimizing Faithfulness

To improve the faithfulness of LLMs, many efforts have been focusing on improving the backbone model with medical knowledge, optimizing the factual correctness of AI medical systems for various generative medical tasks, and developing fact-checking methods, which will be introduced in the following subsections.

4.1 Faithfulness in Large Language Models (LLMs)

Although domain-specific language models such as BioBERT, PubMedBERT, and BioGPT et al, have shown the effectiveness in encoding medical texts through pre-training with unlabeled medical texts, their ability in understanding medical knowledge and texts is challenging for them to generate factually correct contents. Some efforts have focused on explicitly incorporating extra medical knowledge to address this challenge. Yuan et al[35] proposed to train knowledge-aware language model by infusing entity representations, as well as the entity detection and entity linking pre-training tasks based on the Unified Medical Language System (UMLS) knowledge base. The proposed method improves the performance of a series of biomedical language models such as BioBERT and PubMedBERT, on the named entity recognition and relation extraction tasks. Jha et al[36] proposed to prob diverse medical knowledge bases into language models with the continual knowledge infusion mechanism to avoid forgetting encoded knowledge previously when injecting multiple knowledge bases, which improves several biomedical language models such as BioBERT and PubMedBERT in medical question answering, named entity recognition and relation extraction. Singhal et al[37] investigated the ability of LLMs on capturing clinical knowledge, based on a 540-billion parameter LLM PaLM[38] and its instruction-tuned variant Flan-PaLM[39], on the medical question answering (Q&A) task. Flan-PaLM achieves state-of-the-art (SOTA) performance on medical Q&A benchmark datasets including MedQA[40], MedMCQA[41], PubMedQA[42], and MMLU clinical topics[43], by using the few-shot prompting, chain-of-thought prompting, and self-consistency prompting techniques and outperforms other domain-specific language models such as BioGPT with the 355 million parameters, PubMedGPT[44] with the 2.7 billion parameters and Galactica[45] with the 120 billion parameters. They further proposed the Med-PaLM that aligns PaLM into the medical domain by instruction prompt tuning, which greatly alleviates errors of PaLM on scientific grounding, harm, and bias from low-quality feedback. Their experiments show that current LLMs such as Flan-PaLM are not ready to be used in areas such as medicine where satefy is the first priority, and the instruction prompt tuning used in Med-PaLM is indeed useful in improving its factuality, consistency, and safety. Zakka et al[46] proposed the knowledge-grounded language model Almanac, which uses the LLMs as the clinical knowledge base to retrieve and distill information from medical databases for replying to clinical queries, rather than directly generate content with LLMs. Almanac has shown better performance

than ChatGPT on medical Q&A based on the 20 questions derived from their ClinicalQA dataset and mitigates the factual inconsistency problem by grounding LLMs with factually correct information retrieved from predefined knowledge repositories. Nori et al[47] conducted a comprehensive study on the ability of GPT-4 on medical competency examinations and medical Q&A benchmarks recently[37]. They evaluated the performance of GPT-4 on steps 1-3 of the United States Medical Licensing Examination (USMLE). On the USMLE self-assessment and sample exam, GPT-4 achieves the average accuracy score of 86.65% and 86.70% and outperforms GPT-3.5 by around 30%. On the multiple-choice medical Q&A benchmark with four datasets, GPT-4 also significantly outperforms GPT-3.5 and the Flan-PaLM 540B[39] as introduced before by a large margin. On the USMLE self-assessment and sample exam, they further evaluated the calibration of GPT-4, which measures the alignment of the predicted probability of the answer's correctness and the true probability. GPT-4 shows much better calibration than GPT-3.5. More concretely, on the multiple-choice question-answering task, the data-points with predicted probability of 0.96 assigned by GPT-4 can be correct at 93% of the time, while only 55% correctness of data-points with the similar predicted probability from GPT-3.5. However, they suggested GPT-4 still has a large gap on safe adoption in the medical domain like prior LLMs.

Discussion. Although the above studies have shown that the incorporation of medical knowledge into LLMs can potentially help improve their faithfulness. They showed that even SOTA LLMs such as Med-PaLM and GPT-4 cannot satisfy the need for safe use in medicine and healthcare. To bridge the gap, there are several limitations to be addressed and promising future directions.

- Systematic evaluation benchmark: existing methods only assess the factuality of LLMs in limited tasks such as question answering and radiology report summarization, there is no systematic evaluation of LLMs in many other critical generative tasks such as medical text generation. We believe future efforts should be spent on evaluating and improving LLMs in diverse medical tasks to fill the gap with domain experts.

- Multimodal and multilingual: most existing methods can only process the medical texts and the language in English. Future efforts are encouraged to build LLMs in the medical domain with the ability to tackle inputs with multi-modalities and multiple languages, for example adapt the recently released multi-modal large language models such as GPT-4 and Kosmos[48] into the medical domain.

- Unified automatic evaluation method: evaluating the performance of LLMs in factuality is especially challenging in the medical domain and existing methods rely on human evaluation, which is expensive and hard to be on large scale. The unified automatic factuality evaluation method should be proposed for supporting the effective evaluate the factual correctness of LLMs on various medical tasks.

The above methods only take the initial step, and we consider more efforts should be proposed in the future to make AI methods closer to real-world medical applications.

4.2 Faithfulness of AI Models in Different Medical Tasks

Many efforts have been devoted to optimize the factuality of generative methods in medicine and healthcare, as well as their factuality evaluation for a specific task with various techniques such as incorporating medical knowledge, reinforcement learning and prompt learning.

4.2.1 Medical Text Summarization

Medical text summarization[49] is an important generative medical task, with the goal of condensing medical texts such as scientific articles, clinical notes, or radiology reports into short summaries. Medical text summarization supports many applications in medicine and healthcare, such as assisting researchers and clinicians to quickly access important information from a large amount of medical literature and patient records and identify key medical evidence for clinical decisions. It is important to ensure the generated summaries by AI methods are factually consistent with the input or reference medical texts. In the following we briefly overview the recent efforts on studying the factual inconsistency problem in medical text summarization.

Optimization Methods The factual inconsistency problem was first explored in the radiology report summarization. Specifically, Zhang et al[50] found that nearly 30% of radiology report summaries generated from the neural sequence-to-sequence models contained factual errors. To deal with the problem, Zhang et al[22] proposed to optimize the factual correctness of the radiology report summarization methods with reinforcement learning. They evaluated the factual correctness of the generated summary with the CheXpert F1 score[51], which calculates the overlap of 14 clinical observations between the generated summary and the reference summary. They optimized such factual correctness with policy learning, by taking the factual correctness score of the generated summary as the reward and the summarizer as the agent. They demonstrated that such training strategy could improve the CheXpert F1 score by 10% when compared with the baseline method. However, this evaluation metric was

only limited to chest X-rays. Delbrouck et al[52] further released the new dataset based on MIMIC-III[53] with new modalities including MRI and CT, as well as anatomies including chest, head, neck, sinus, spine, abdomen, and pelvis. They proposed the new factual correctness evaluation metric RadGraph score that could be used for various modalities and anatomies and designed a summarization method that optimized the RadGraph score-based reward with reinforcement learning. Their experiment results showed that optimizing the RadGraph score as the reward could consistently improve the factual correctness and quality of the generated summary from the summarizer, where the RadGraph score, F1CheXbert, ROUGE-L[54] (a commonly used metric in text generation, that calculates the longest common sub-sequence between the generated summary and reference summary) are improved by 2.28%-4.96%, 3.61%-5.1%, and 0.28%-0.5%. Xie et al[55] proposed the two-stage summarization method FactReranker, which aims to select the best summary from all candidates based on their factual correctness scores. They proposed to incorporate the medical factual knowledge based on the RadGraph Schema to guide the selection of FactReranker. FactReranker achieves the new SOTA on the MIMIC-CXR dataset and improves the RadGraph score, F1CheXbert, and ROUGE-L by 4.84%, 4.75%, and 1.5%.

There are also efforts investigating the factual inconsistency problem in the automatic summarization of other medical texts such as biomedical literature, medical Q&A, and medical dialogues. Deyoung et al[56] found that the summarizer based on language models such as BART[57] could produce fluent summaries for medical studies, but the faithfulness problem remains an outstanding challenge. For example, the generated summary from the summarizer only had 54% agree on the direction of the intervention's effect with that of the input systematic review. Wallace et al[58] proposed the decoration and sorting strategy that explicitly informed the model of the position of inputs conveying key findings, to improve the factual correctness of the generated summary from the BART-based summarizer for published reports of randomized controlled trials (RCTs). Alambo[59] studied the factual inconsistency problem of a transformer-based encoder decoder summarization method. They proposed to integrate the biomedical named entities detected in input articles and medical facts retrieved from the biomedical knowledge base to improve the model faithfulness. Yadav et al[60] proposed to improve the factual correctness of the generated summary for medical questions, via maximizing the question type identification reward and question focus recognition reward with the policy gradient approach. Chintagunta et al[61] investigated using GPT-3 as the labeled data generator, and incorporated the medical concepts identified from the input medical dialogues as the guidance for generating labeled data with higher quality, which has proven to be able to train the summarizer with better factual correctness when compared with the human-labeled training data that is thirty times larger. Liu et al[62] proposed the task of automatic generating discharge instruction based on the patients' electronic health records and the Re^{3}Writer method for the task, which retrieved related information from discharge instructions of previous patients and medical knowledge to generate faithful patient instructions. The human evaluation results showed the patient instructions generated by proposed method had better faithfulness and comprehensiveness than those generated from baseline sequence-to-sequence methods.

Evaluation Metrics Derivation of effective automatic factual correctness evaluation metrics is critical for evaluating the quality of the generated summary and summarization method development. Existing commonly used evaluation metrics in text summarization such as ROUGE[54] and BERTScore[63] have been proven to be ineffective in evaluating factual correctness, especially in medical domain. Existing evaluation metrics for factual consistency such as FactCC[64] and BARTScore[65], are also designed for the general domain. Recently, efforts have been conducted to develop new metrics for evaluating the factual consistency of the summarizer for different medical texts.

For radiology report summarization, Zhang et al[22] proposed the CheXpert F1 score that calculates the overlap of 14 clinical observations such as "enlarged cardiom" and "cardiomegaly", between the generated summary and the reference summary. However, it is limited to the reports of chest X-rays. Delbrouck et al[52] further proposed RadGraph score that calculates the overlap of medical entities and relations based on the RadGraph[66] (a dataset with annotations of medical entities and relations for radiology reports) between the generated summary and the gold summary, and can be used for various modalities and anatomies. For the biomedical literature summarization, Wallace et al[58] proposed findings-Jensen-Shannon Distance (JSD) calculating the agreement of evidence directions (including significant difference, or no significant difference) of the generated summary and reference summary of the systematic review, according to JSD. Based on findings-JSD, Deyoung et al[56] further proposed the improved metric ΔEI that calculates the agreement of the intervention, outcome and evidence direction based on the Jensen-Shannon Distance, between the generated summary and the input medical studies. Otmakhova et al[67] proposed the human evaluation approach for medical study summarization, where they defined several quality dimensions including PICO correctness, evidence direction correctness, and modality to evaluate the factuality of the generated summary. Based on the human evaluation protocol, Otmakhova et al[68] further developed the Δloss to evaluate the factual correctness of the generated summary on different aspects such as strong claim, no evidence, no claim, etc., and evidence directions. It calculates the difference of negative log-likelihood loss with the summarization model, between the generated summary and the counterfactual summary (the corrupted target summary that has different modality and polarity with the target summary). Adams et al[69] did a meta-evaluation on existing automatic evaluation metrics including BARTScore, BERTScore, CTC^2 and SummaC[70] (two SOTA metrics based on natural language inference) on assessing long-form hospital-course summarization. They first

created a human-annotated dataset with the fine-grained faithfulness annotations of generated hospital-course summaries with Longformer Encoder-Decoder (LED)[71], and then measure the correlation between the assessment of automatic evaluation metrics and human annotations. Although without aligning to the biomedical domain, they found the evaluation results of these automatic evaluation metrics have a good correlation with that of human annotations. However, the calculated correlation is found to has bias and not correct enough, since the generated summaries from LED tend to have a high overlap with original sentences of original inputs rather than regenerate new sentences, which is not applicable in a realistic situation. Moreover, these metrics have limitations on assessing factual errors which require deep clinical knowledge such as missingness, incorrectness and not in notes.

Discussion To understand how existing methods perform on the factuality of medical text summarization, in Table 2 and Table 3, we show the performance of SOTA methods on medical study summarization and radiology report summarization with human evaluation and automatic evaluation metrics introduced above. We can see that existing SOTA methods have relatively good performance on the grammar and lexical of generated summaries for medical studies. Only 8%-9% of the generated summaries are completely factually correct. Less than 50% of the generated summaries are with correct PICO[72] (Patient, Intervention, Comparison, Outcome) and modality indicating the certainty level of evidence claim (such as strong claim, moderate claim, and weak claim). For radiology report summarization, we can find that although most methods[22,52] use reinforcement learning for optimizing the factuality, the method proposed in Xie et al[55] incorporating medical knowledge achieves the best performance, which shows the importance of using medical knowledge to improve factuality. Similarly, the factuality performance of SOTA methods was not high either. For example, the RadGraph F1 scores of SOTA methods were only around 50%.

Table 2. The performance of human evaluation of SOTA methods on medical study summarization in MS2 dataset[56] from the paper[67], which evaluates generated summaries from the factual correctness of PICO, evidence direction, and modality. Factually correct means the composition performance of PICO, evidence direction, and modality.

	PICO	Direction	Modality	Factually correct	Grammar	Lexical
BART[56]	45%	77%	45%	9%	75%	69%
LED[56]	40%	75%	44%	8%	73%	73%

Table 3. The performance of SOTA methods for radiology reports summarization on MIMIC-CXR dataset.

	ROUGE-1	ROUGE-2	ROUGE-L	F1CheXbert	RadGraph
L4-RL[52]	51.96	35.65	47.10	74.86	48.23
FactReranker[55]	55.94	40.63	51.85	76.36	53.17

For future work, we believe the following aspects are important to focus on.

- *Creation of benchmark data sets*. There are limited open source datasets for medical text summarization, due to the high cost of expert annotation and other issues such as privacy. In fact, there are only two public datasets for medical study and radiology report summarization, which covers limited modalities, anatomies and languages. Moreover, the sizes of these datasets are much smaller when compared with datasets in the general domain. High-quality large-scale benchmark datasets should be created in the future for facilitating the development of medical summarization methods.
- *Derivation of unified automatic evaluation metrics*. Existing evaluation metrics for assessing the factuality of methods in different medical texts are specific, and there are even no such metrics for methods on certain types of medical texts such as medical dialogue and medical question summarization. It is important to develop unified automatic evaluation metrics for supporting the assessment of summarization methods across different types of medical text.
- *Development of better optimization methods*. Most existing methods employ reinforcement learning to improve the factuality of generated summaries, while little attention has been paid to incorporating medical knowledge. As mentioned before, the factual correctness of SOTA methods is not good enough for reliable use in realistic medicine and healthcare. Therefore, more advanced methods are needed, for example, by using more powerful backbone language models and designing a more effective decoding strategy guided by medical facts.

4.2.2 Medical Text Simplification

Medical text simplification aims to simplify highly technical medical texts to plain texts that are easier to understand by non-experts such as patients. It can greatly improve the accessibility of medical information.

Optimization Methods Most existing work focused on creating data resources for supporting the development and evaluation of medical text simplification. For example, Trienes et al[73] created a dataset with 851 pathology reports for document-level simplification, Devaraj et al[25] constructed a dataset with the biomedical systematic reviews, including both their technical summaries and plain language summaries, to support the paragraph-level simplification. Lu et al[74] proposed the summarize-then-simplify method for paragraph-level medical text simplification, where they designed the narrative prompt with key phrases to encourage the factual consistency between the input and the output. The human evaluation showed that their proposed method significantly outperforms the BART-based simplification method proposed in[25] by 0.49 in the 5-point scale on the factuality of outputs. Jeblick et al[75] utilized ChatGPT to simplify 45 radiology reports and asked 15 radiologists to evaluate outputs from factual correctness, completeness, and potential harm. They found the radiologists agreed that most of the outputs were factually correct, but there were still some errors such as misinterpretation of medical terms, imprecise language, and extrinsic factual errors. Lyu et al[76] also evaluated the performance of ChatGPT on simplifying radiology reports to plain language. The evaluation dataset consisted of 62 chest CT screening reports and 76 brain MRI screening reports. ChatGPT showed good performance with an overall score of 4.268 in the five-point system assessed by radiologists. It has an averaged of 0.08 and 0.07 places of information missing (assessing the number of places with information lost) and inaccurate information (assessing the number of places with inaccurate information). The suggestions for patients and healthcare provider generated by ChatGPT tended to be general, such as closely observing any symptoms, and only 37% provides specific suggestions based on the findings of reports. ChatGPT also showed instability in generating over-simplified reports, which can be alleviated by designing better prompts. Overall, the factuality problem in automatic medical text simplification methods has rarely been explored and we believe more efforts should be devoted in this area.

Evaluation Metrics Similar to the medical text summarization, automatic similarity-based evaluation metrics such as ROUGE and BERTScore were also used for evaluating the semantic similarity between outputs and references in the medical text simplification. Moreover, other important aspects in the evaluation are the readability and simplicity of outputs, for which commonly used metrics include the Flesch-Kincaid grade level (FKGL)[77], the automated readability index (ARI)[78], and SARI[79]. Intuitively, the language model pre-trained on the technical corpus can assign higher likelihoods to the technical terms than the language models pre-trained on the general corpus. Based on this intuition, Devaraj et al.[25] proposed a new readability evaluation metric calculating the likelihood scores of input texts with a masked language model trained on the technical corpus. Devaraj et al[80] further proposed a method for evaluating factuality in text simplification, in which they trained a RoBERTa-based classification model to classify the factuality level according to different types of factual errors including insertion, deletion, and substitution errors based on human-annotated data. They found that existing metrics such as ROUGE and SARI cannot effectively capture factual errors, and it is also challenging to evaluate the factual errors for their proposed method, which is even trained with extra training data generated by data augmentation. As introduced above, only one automatic factuality evaluation metric is proposed by Devaraj et al[80], which is however not effective in capturing factual errors. We believe more efforts on factuality evaluation metrics should be developed in this area.

Discussion We can see that there is limited research studying the factuality problem in medical text simplification. Moreover, the assessment of the factuality of existing methods have been largely relying on human evaluation. Therefore, it is important to have more efforts to study the factuality problem in medical text simplification by developing better optimization methods, automatic evaluation metrics, and creating more data resources in the future.

4.2.3 Radiology Report Generation

Radiology report generation aims to automatically generate radiology reports illustrating clinical observations and findings with the input medical images such as chest X-rays and MRI scans. It can help to reduce the workload of radiologists and improve the quality of healthcare.

Optimization Methods Most existing efforts adopted reinforcement learning (RL) to optimize the factual correctness of radiology report generation methods. Nishino et al[81] proposed the RL-based method by optimizing the clinical reconstruction score calculating the correctness of predicting finding labels with the generated report, to improve the factual correctness of generated reports. Their experiments showed that the model improved the performance on the F1 score of the factuality metric CheXpert (it assesses the clinical correctness of the generated report that calculates the overlap of finding labels between generated and reference report) by 5.4% compared with the model without the RL-based optimization. Miura[26] proposed to use reinforcement learning to optimize the entailing entity match reward assessing the inferentially consistent (such as entailment, neutral, contradiction) between entities of the generated report and the reference report, and exact entity match reward evaluating the consistent of disease and anatomical entities between the generated report and reference report, which encourages the model to generate key medical entities that are consistent with references. The proposed method greatly improves the F1 CheXpert score by 22.1% compared with the baselines. However, it relied on the named entity recognition methods that are not trained with annotated data of the Chest X-ray domain. Delbrouck et al[82] further designed the RadGraph

reward calculating the overlap of entities and relations between the generated report and the reference, based on the RadGraph dataset including annotated entities and relations of the Chest X-ray reports. The proposed method improves the factuality evaluation metric F1 RadGraph score (calculating the overlap of entities and relations between the generated report and the reference) by 5.5% on the MIMIC-CXR dataset when compared with baselines including[26]. Nishino et al[83] further proposed an RL-based method Coordinated Planning (CoPlan) with the fact-based evaluator and description-order-based evaluator, to encourage the model to generate radiology reports that are not only factually consistent but also chronologically consistent with reference reports. Their method outperforms the baseline T5 model on clinical factual accuracy by 9.1%.

Evaluation Metrics Similar to medical text summarization, to evaluate the clinical correctness of the generated radiology reports, some efforts[26,81,84] proposed to use the CheXpert-based metric to evaluate the overlap of 14 clinical observations between generated reports and references annotated by the CheXpert, for which they calculated the precision, recall, and F1 score. Delbrouck et al[82] proposed the RadGraph-based metric to calculate the overlap of the clinical entities and relations between generated reports and references annotated by the RadGraph schema. Recently, Yu et al[85] examined the correlation between existing automatic evaluation metrics including BLEU[86], BERTScore, F1 CheXpert, and RadGraph F1, and the score given by radiologists on evaluating the factuality of the generated reports. They found that the evaluation results of F1 CheXpert and BLEU were not aligned with that of radiologists, and BERTScore and RadGraph F1 were more reliable. BERTScore and RadGraph F1 outperformed BLEU in evaluating the false prediction of finding, while F1 CheXpert has worse performance than BLEU in evaluating the incorrect location of finding. The authors further proposed a new evaluation metric RadCliQ, which trained an regression model to predict the number of errors in generated radiology reports assigned by radiologists based on the combination score with BLEU and RadGraph F1. RadCliQ is thus the weighted sum of the score from BLEU and RadGraph F1 based on their optimized coefficients. It showed better alignment with the evaluation of radiologists than the above four metrics and had the Kendall-tau b correlation coefficient of 0.522 for the number of errors annotated by radiologists.

Discussion Existing methods have demonstrated the effectiveness of using RL and medical knowledge in improving the factuality of generated radiology reports. However, the medical knowledge are typically incorporated in implicit ways based on RL, we consider that future efforts should pay more attention to explicitly incorporating medical knowledge on improving the encoder and decoder of the LLMs. For example, investigating the use knowledge grounded backbone language models as encoder[37], and developing decoding strategy guided by medical facts[55,87]. Moreover, the radiology report generation task requires the combination of information both radiology images and the associated text reports, we believe cross-modality vision-language foundation models[88] should be explored to improve the faithfulness of radiology report generation methods in the future. For the evaluation metrics, there is only one paper[85] as we described above on analyzing the correlation between automatic factuality evaluation metrics and scores of experts based on human annotation. It is necessary to have more efforts on developing automatic factuality evaluation metrics, and creating public benchmark datasets to help the meta-evaluation of these metrics.

4.3 Medical Fact Checking

Automatic medical fact-checking aims to detect whether the claim made in certain medical text is true, which is a promising method for assisting in detecting factual errors and improving the factuality of medical generative methods. Many existing efforts have contributed to the creation of medical data resources to facilitate the development of automatic medical fact-checking methods and their evaluation. Kotonya et al.[89] built the PUBHEALTH dataset with 11.8K public health-related claims with gold-standard fact-checking explanations by journalists. Wadden et al.[90] created the SCI-FACT dataset with 1.4K clinical medicine-related scientific claims paired with abstracts including their corresponding evidence, and the annotated labels (including supports, refutes, and neutral) as well as rationales. Poliak et al[91] collected over 2.1K verified question-answer pairs related to COVID-19 from over 40 trusted websites. Sarrouti et al[92] developed HEALTHVER for evidence-based fact-checking of health related claims, with 14,330 claim-evidence pairs from online questions on COVID-19, and their label including supports, refutes, and neutral. On a relevant effort, Saakyan et al[93] created COVID-Fact with 4,086 real-world claims related to the COVID-19 pandemic, sentence-level evidence for these claims, and their counter claims. Mohr et al[94] built a fact-checked corpus with 300 COVID-19-related tweets annotated with their verdict label, biomedical named entities, and supporting evidence. Srba et al[95] developed a new medical misinformation dataset with 573 manually and more than 51k automatically annotated relations between verified claims and 317k news articles, including the claim presence label indicating whether a claim is included in the article and article stance label indicating whether a claim is supported by the article such as supports, refutes, and neutral. Wadhwa et al[96] constructed the RedHOT corpus with 22,000 health-related social media posts from Reddit annotated with claims, questions, and personal experiences, which can support for identifying health claims and retrieving related medical literature.

There were also efforts on developing automatic medical fact-checking methods with these data resources. Specifically, Kotonya et al[89] proposed an explainable automatic fact-checking method with a classifier for predicting label of the given

Table 4. Summary of datasets.

Data	Domain	Claim Source	Negation Method	Size
PUBHEALTH[89]	Publich health	Fact checking and News websites	Natural	11,832
HEALTHVER[92]	COVID	TREC-COVID[97]	Natural	14,330
SCI-FACT[90]	Biomed	S2ORC[98]	Human	1,409
COVID-Fact[93]	COVID	Reddit	Automatic	4,086
Misinformation[95]	Medical	Fact checking and News websites	Human	573
Covert[94]	COVID	Twitter	Human	300
RedHOT[96]	Health	Reddit	Human	22,000

clain based on pre-trained language models such as BERT, SciBERT[99], BioBERT, and a summarization model based on the BERTSum summarizer[100] for generating fact-checking explanations. On the PUBHEALTH dataset, the SciBERT-based prediction method achieved the highest macro F1, precision, and accuracy scores, and fact-checking explanation model fine-tuned on the PUBHEALTH dataset achieved promising performance. Wadden et al[90] proposed the automatic fact-checking pipeline with the SCI-FACT dataset that retrieves abstracts based on input claims according to the TD-IDF similarity, selects rationale sentences and then predicts the labels (SUPPORTS, REFUTES, or NOINFO) of abstracts regarding the given claims with BERT based related language models. They investigated using SciBERT, BioMedRoBERTa, RoBERTa-base, and RoBERTa-large as the sentence encoder, where the RoBERTa-large achieves the best performance on label prediction. Wadden et al[101] proposed MULTIVERS for predicting the fact-checking label for a claim and the given evidence abstract, which uses Long-former encoder[71] to encode the long sequence from the claim and the evidence abstract, and predicts the abstract-level fact-checking label aligned to the sentence-level label by multi-task learning. On three medical fact-checking datasets including HEALTHVER, COVID-Fact, and SCI-FACT, MULTIVERS showed better performance on the zero-shot and few-shot settings compared with existing methods, due to the weak supervision by the multi-task learning.

Discussion Although fact-checking is a promising way for detecting and mitigating the hallucination problem of AI methods, we can see that existing fact-checked datasets only covered limited medical facts in specific domains such as COVID and public health. Therefore, future efforts can also be spent on developing effective automatic fact-checking methods based on other resources with rich medical facts such as medical knowledge bases and plain medical texts. Moreover, there are no methods exploring fact-checking on LLMs, which should be a research focus as well given the popularity of LLMs.

5 Limitations and Future Directions

With all the reviews and discussions of existing works on faithful AI in healthcare and medicine above, we will summarize the overall limitations of existing studies and discuss future research directions.

5.1 Limitations

5.1.1 Datasets

Unlabeled Data for Self-supervised Learning Many large AI models including LLMs are typically trained with self-supervised learning, which relies on the availability of large-scale unlabeled medical data. However, it is challenging to collect large amounts of biomedical data in many scenarios due to the various considerations such as privacy and cost, which makes the volume of publicly available biomedical to be much smaller than that of unlabeled data in other application domains such as computer vision and natural language processing. For example, the unlabeled clinical data used to train the clinical language models such as ClinicalBERT[102] is 3.7GB, while the data size of unlabeled data used to train the large language models in the general domain can be up to 45TB.

Annotated Data for Supervised Learning and Evaluation Developing faithful medical AI methods and evaluating their factuality also relies on high-quality annotated medical data. However, collecting large-scale high-quality annotated medical data is even more challenging, due to the high cost on both time and expertise. The sizes of existing annotated datasets are mostly small (e.g., the sample sizes of existing datasets for medical question summarization are around 1,000[103]). In addition, there is no publicly available annotated data for supporting the meta-evaluation of automatic evaluation metrics in most medical tasks, which makes it challenging to verify the reliability of different automatic evaluation metrics.

Data in Multimodal and Multilingual Many existing medical datasets consist of only a single data modality such as the UK Biobank[104]. There are few multimodal datasets with texts and images, such as MIMIC-CXR consisting of medical radiographs and text reports. However, there are many other data modalities such as biosensors, genetic, epigenetic, proteomic, and microbiome, that have been rarely considered in existing datasets. Moreover, most existing datasets are limited to a single

language, where English is the predominantly used language. This can hinder the development of faithful medical AI methods in low-resources and rural areas.

5.1.2 Backbone Models

Biomedical Domain-Specific Language Models Many existing medical AI methods use biomedical domain-specific language models as the backbone and fine-tune them with task-specific datasets for various downstream tasks. However, these models are typically pre-trained with the biomedical literature texts and a few other types of medical texts such as EHRs and radiology reports. Therefore, they can only capture limited medical knowledge and cannot be effectively generalized to tasks involving other types of medical data. Moreover, these domain-specific language models are typically small in their sizes (measured by the number of parameters, usually less than 1B parameters), which further impacts their performance and generalization abilities. Up to now, the largest generative language model in the medical domain PubMedGPT has 2.7B parameters, which is far smaller than the scale of large language models in the general domain such as GPT-3 with 175B parameters, PaLM with 540B parameters, and the most recent GPT-4 with 100T parameters.

Large Generative Language Models Recently, large generative language models in the general domain such as GPT-3, PaLM, ChatGPT, GPT-4 et al, have shown amazing abilities in natural language understanding and generation. However, they are not trained with data in medical domain, which limits their ability in understanding medical data. Moreover, none of them are made publicly available, which hinders the research in using these LLMs for developing faithful medical AI methods. There is a recent work[37] that aligns the LLM PaLM with the medical domain. Unfortunately, it is still not publicly available.

5.1.3 Faithful Medical AI Methodologies

Mitigation Methods Although factuality is a critical issue in existing medical AI methods, little efforts has been devoted to the improvement of the faithfulness of backbone language models and medical AI methods for downstream tasks. There has been no research investigating the factuality in medical tasks including medical dialogue generation, medical question answering and drug discovery et al. It is clear the factuality problem should attract more attention in future efforts.

Incorporating Medical Knowledge Existing efforts have proven the effectiveness and importance of improving the factuality in both backbone language models and medical AI methods for specific tasks by incorporating medical knowledge. Most of them focused on extracting medical knowledge from external biomedical knowledge bases. Recently, there is an effort[37] investigating the efficiency of instruction prompt tuning on injecting medical knowledge into LLMs, which relies on human feedback and thus can be expensive and time consuming. Effective incorporation of medical knowledge in efficient and scalable ways remains a critical challenge.

5.1.4 Evaluations

Automatic Evaluation Metrics Existing automatic evaluation metrics usually calculate the overlap of medical facts such as medical entities and relations between outputs generated by the algorithm and references. They cannot assess and distinguish different types of factual errors such as intrinsic and extrinsic errors as introduced in the Section 3.1, which is essential to analyze the cause of factual errors and mitigate them. In addition, the assessment of factuality relies on human evaluations for many tasks such as medical question answering and dialogue generation, where there are no automatic factuality evaluation metrics.

Meta-evaluation Assessing the effectiveness of automatic evaluation metrics is critical for correctly evaluating the factuality of methods. Otherwise, the ineffective automatic evaluation metrics can misguide the optimization and evaluation of methods. There is rare work[85] investigating the meta-evaluation of automatic factuality metrics used in various medical tasks, and analyzing their alignment with domain experts.

5.2 Future Directions

With the above limitations we summarized, in the following We outline promising directions for future research to develop faithful medical AI methods.

5.2.1 Datasets

Large Scale Datasets for Pre-training Large scale public multimodal and multilingual medical datasets should be developed in the future for facilitating the pre-training of reliable biomedical language models. We believe it is important to improve the generalization ability and faithfulness of language models in the medical domain, by pre-training them with various languages and modalities of medical data such as text, image, clinical, genetic et al.

Benchmark Datasets for Factuality Evaluation More public benchmark datasets for various medical tasks should be constructed to support the development of reliable medical AI methods. Moreover, future efforts should pay attention to developing benchmark factuality evaluation datasets in the medical domain to assess the factuality of LLMs, like the TruthfulQA dataset[105] for evaluating the factuality of GPT-4 in the general domain.

Annotated Datasets for Meta-evaluation It is urgent to build domain expert annotated datasets of various medical tasks for supporting analysis of the alignment between automatic factuality evaluation metrics and preferences of domain experts. They are critical for future research of developing reliable automatic evaluation metrics and developing effective mitigation methods.

5.2.2 Backbone Models

Faithful Biomedical Language Models Reliable backbone models, mainly the biomedical language models, are crucial to enhance the factuality of medical AI methods. Different strategies can be explored in the future, such as training with multimodal data, incorporating medical knowledge, and prompt-based pre-training.

Aligning Large Language Models to Medical Domain Large language models have shown predominantly superior performance than pre-trained language models with smaller scales, and have been the new backbone model in the general domain. It is crucial to adapt them to the medical domain for their reliable use. Several adaptation strategies can be explored, such as fine-tuning with domain-specific data and instruction prompt tuning with human feedback. Med-PaLM[37] is a pioneer work in this direction, however it is not publicly accessible.

Evaluating the Faithfulness of Backbone Models Future efforts should also focus on evaluating and analyzing the performance of LLMs in factuality, which can provide crucial insights on how to improve their faithfulness. To support the research in this direction, it is important to build a standardized benchmark that can cover a broad variety of critical medical tasks and datasets.

5.2.3 Methodologies

Medical Knowledge Medical knowledge is critical for alleviating the hallucination of medical AI methods for various tasks. Future efforts should explore effective ways of injecting medical knowledge into medical AI methods, such as encoding medical knowledge based on prompt learning[106], and explicitly leveraging the medical knowledge to guide the text generation[87].

Reinforcement Learning with Human Feedback Reinforcement learning with human feedback (RLHF) has exhibited a strong potential on improving the factuality of LLMs, and GPT-4 has greatly improved its safety and faithfulness by the RLHF post-training, when compared to GPT-3.5. More efforts should be devoted to explore the mitigation methods with RLHF in the medical domain. One potential challenge of using RLHF is its reliance on human feedback, which comes with high cost in the medical domain. Therefore, it is an open question on how to apply RLHF in a cost-effective and low-resource manner in medicine and healthcare.

Model Explainability It is also important to develop explainable medical AI methods, especially LLMs. The explanation can play an important role in alleviating the hallucination problem since it helps to understand and trace causes of factual errors, and also makes it easier to assess factual errors and enhance the medical AI faithfulness.

5.2.4 Evaluation Methods

Evaluation Guideline It has been observed that the criteria for factuality evaluation with different automatic evaluation metrics and human evaluation used in different research often differ. For example, in radiology report summarization and generation, two automatic metrics the F1 ChexPert and RadGraph Score were used. F1 ChexPert calculates the overlap of 14 clinical observations such as "enlarged cardiom" and "cardiomegaly", between the generated summary and the reference summary, while F1 RadGraph calculates their overlap on medical entities and relations. They focused more on evaluating the alignment of different medical facts between outputs and references. It is necessary to have a unified evaluation guideline to define the detailed standardized evaluation criteria for various medical tasks, which can support the development of unified automatic evaluation metrics and fair comparison of different research studies.

Fine-grained Automatic Evaluation Metrics It is expected that the fine-grained automatic evaluation metrics that can clarify and assess different medical factual errors can be explored in future efforts. Assessing the fine-grained factual errors such as intrinsic and extrinsic errors of medical AI methods plays an important role in understanding and alleviating their factuality problem. One promising direction is to explore developing fine-grained evaluation metrics with the RLHF.

6 Conclusions

The progress of fundamental AI methods, especially the most recent large language models, provides great opportunities for medical AI, but there is a severe concern about the reliability, safety, and factuality of generated content by medical AI methods.

In this review, we provide the first comprehensive overview of the faithfulness problem in medical AI, analyzing causes, summarizing mitigation methods and evaluation metrics, discussing challenges and limitations, and outlooking future directions. Existing research on investigating the factuality problem in medical AI remains in the initial phase, and there are several significant challenges to data resources, backbone models, mitigation methods, and evaluation metrics in this research direction. It is clear more future research efforts should be conducted and there are significant opportunities for novel faithful medical AI research involving adapting large language models, prompt learning, and reinforcement learning from human feedback et al. We hope this review can inspire further research efforts in this direction, as well as serve as a guide for researchers and practitioners on the safe use of AI methods in realistic medical practice.

Competing Interests

The authors declare no competing interests.

References

1. Yu, K.-H., Beam, A. L. & Kohane, I. S. Artificial intelligence in healthcare. *Nat. biomedical engineering* **2**, 719–731 (2018).
2. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. medicine* **25**, 44–56 (2019).
3. Rajpurkar, P., Chen, E., Banerjee, O. & Topol, E. J. Ai in health and medicine. *Nat. medicine* **28**, 31–38 (2022).
4. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *nature* **521**, 436–444 (2015).
5. Shen, D., Wu, G. & Suk, H.-I. Deep learning in medical image analysis. *Annu. review biomedical engineering* **19**, 221–248 (2017).
6. Kenton, J. D. M.-W. C. & Toutanova, L. K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 4171–4186 (2019).
7. Wang, B. *et al.* Pre-trained language models in biomedical domain: A systematic survey. *arXiv preprint arXiv:2110.05006* (2021).
8. Vaswani, A. *et al.* Attention is all you need. *Adv. neural information processing systems* **30** (2017).
9. Agrawal, M., Hegselmann, S., Lang, H., Kim, Y. & Sontag, D. Large language models are zero-shot clinical information extractors. *arXiv preprint arXiv:2205.12689* (2022).
10. Tiu, E. *et al.* Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nat. Biomed. Eng.* 1–8 (2022).
11. Jumper, J. *et al.* Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589 (2021).
12. Brown, T. *et al.* Language models are few-shot learners. *Adv. neural information processing systems* **33**, 1877–1901 (2020).
13. OpenAI. Chatgpt. *https://openai.com/blog/chatgpt* (2022).
14. OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
15. Kung, T. H. *et al.* Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLOS Digit. Heal.* **2**, e0000198 (2023).
16. Moor, M. *et al.* Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
17. van Dis, E. A., Bollen, J., Zuidema, W., van Rooij, R. & Bockting, C. L. Chatgpt: five priorities for research. *Nature* **614**, 224–226 (2023).
18. Li, W. *et al.* Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods. *arXiv preprint arXiv:2203.05227* (2022).
19. Aggarwal, R. *et al.* Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *NPJ digital medicine* **4**, 65 (2021).
20. Weidinger, L. *et al.* Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359* (2021).
21. Ji, Z. *et al.* Survey of hallucination in natural language generation. *ACM Comput. Surv.* (2022).

22. Zhang, Y., Merck, D., Tsai, E., Manning, C. D. & Langlotz, C. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5108–5120 (2020).

23. Luo, R. *et al.* Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings Bioinforma.* **23** (2022).

24. Maynez, J., Narayan, S., Bohnet, B. & McDonald, R. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1906–1919 (2020).

25. Devaraj, A., Wallace, B. C., Marshall, I. J. & Li, J. J. Paragraph-level simplification of medical texts. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, vol. 2021, 4972 (NIH Public Access, 2021).

26. Miura, Y., Zhang, Y., Tsai, E., Langlotz, C. & Jurafsky, D. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5288–5304 (2021).

27. Moradi, M., Blagec, K., Haberl, F. & Samwald, M. Gpt-3 models are poor few-shot learners in the biomedical domain. *arXiv preprint arXiv:2109.02555* (2021).

28. Gutiérrez, B. J. *et al.* Thinking about gpt-3 in-context learning for biomedical ie? think again. *arXiv preprint arXiv:2203.08410* (2022).

29. Lee, J. *et al.* Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).

30. Gu, Y. *et al.* Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Comput. for Healthc. (HEALTH)* **3**, 1–23 (2021).

31. Antaki, F., Touma, S., Milad, D., El-Khoury, J. & Duval, R. Evaluating the performance of chatgpt in ophthalmology: An analysis of its successes and shortcomings. *medRxiv* 2023–01 (2023).

32. Bengio, S., Vinyals, O., Jaitly, N. & Shazeer, N. Scheduled sampling for sequence prediction with recurrent neural networks. *Adv. neural information processing systems* **28** (2015).

33. Wang, C. & Sennrich, R. On exposure bias, hallucination and domain shift in neural machine translation. In *2020 Annual Conference of the Association for Computational Linguistics*, 3544–3552 (Association for Computational Linguistics (ACL), 2020).

34. Lee, N. *et al.* Factuality enhanced language models for open-ended text generation. In *Advances in Neural Information Processing Systems*.

35. Yuan, Z., Liu, Y., Tan, C., Huang, S. & Huang, F. Improving biomedical pretrained language models with knowledge. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, 180–190 (2021).

36. Jha, K. & Zhang, A. Continual knowledge infusion into pre-trained biomedical language models. *Bioinformatics* **38**, 494–502 (2022).

37. Singhal, K. *et al.* Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138* (2022).

38. Chowdhery, A. *et al.* Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311* (2022).

39. Chung, H. W. *et al.* Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416* (2022).

40. Jin, D. *et al.* What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Appl. Sci.* **11**, 6421 (2021).

41. Pal, A., Umapathi, L. K. & Sankarasubbu, M. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*, 248–260 (PMLR, 2022).

42. Jin, Q., Dhingra, B., Liu, Z., Cohen, W. & Lu, X. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2567–2577 (2019).

43. Hendrycks, D. *et al.* Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300* (2020).

44. Bolton, E. *et al.* Pubmedgpt 2.7b. *https://github.com/stanford-crfm/BioMedLM* (2022).

45. Taylor, R. *et al.* Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085* (2022).

46. Zakka, C., Chaurasia, A., Shad, R. & Hiesinger, W. Almanac: Knowledge-grounded language models for clinical medicine. *arXiv preprint arXiv:2303.01229* (2023).

47. Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375* (2023).

48. Huang, S. *et al.* Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045* (2023).

49. Afantenos, S., Karkaletsis, V. & Stamatopoulos, P. Summarization from medical documents: a survey. *Artif. intelligence medicine* **33**, 157–177 (2005).

50. Zhang, Y., Ding, D. Y., Qian, T., Manning, C. D. & Langlotz, C. P. Learning to summarize radiology findings. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, 204–213 (2018).

51. Irvin, J. *et al.* Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, 590–597 (2019).

52. Delbrouck, J.-B., Varma, M. & Langlotz, C. P. Toward expanding the scope of radiology report summarization to multiple anatomies and modalities. *arXiv preprint arXiv:2211.08584* (2022).

53. Johnson, A. E. *et al.* Mimic-iii, a freely accessible critical care database. *Sci. data* **3**, 1–9 (2016).

54. Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81 (2004).

55. Xie, Q., Zhou, J., Peng, Y. & Wang, F. Factreranker: Fact-guided reranker for faithful radiology report summarization. *arXiv preprint arXiv:2303.08335* (2023).

56. DeYoung, J., Beltagy, I., van Zuylen, M., Kuehl, B. & Wang, L. Ms^2: Multi-document summarization of medical studies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 7494–7513 (2021).

57. Lewis, M. *et al.* Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880 (2020).

58. Wallace, B. C., Saha, S., Soboczenski, F. & Marshall, I. J. Generating (factual?) narrative summaries of rcts: Experiments with neural multi-document summarization. *AMIA Summits on Transl. Sci. Proc.* **2021**, 605 (2021).

59. Alambo, A., Banerjee, T., Thirunarayan, K. & Raymer, M. Entity-driven fact-aware abstractive summarization of biomedical literature. In *2022 26th International Conference on Pattern Recognition (ICPR)*, 613–620 (IEEE, 2022).

60. Yadav, S., Gupta, D., Abacha, A. B. & Demner-Fushman, D. Reinforcement learning for abstractive question summarization with question-aware semantic rewards. 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)*, 249–255 (2021).

61. Chintagunta, B., Katariya, N., Amatriain, X. & Kannan, A. Medically aware gpt-3 as a data generator for medical dialogue summarization. In *Machine Learning for Healthcare Conference*, 354–372 (PMLR, 2021).

62. Liu, F. *et al.* Retrieve, reason, and refine: Generating accurate and faithful patient instructions. In *Advances in Neural Information Processing Systems*.

63. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

64. Kryściński, W., McCann, B., Xiong, C. & Socher, R. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9332–9346 (2020).

65. Yuan, W., Neubig, G. & Liu, P. Bartscore: Evaluating generated text as text generation. *Adv. Neural Inf. Process. Syst.* **34**, 27263–27277 (2021).

66. Jain, S. *et al.* Radgraph: Extracting clinical entities and relations from radiology reports. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

67. Otmakhova, J., Verspoor, K., Baldwin, T. & Lau, J. H. The patient is more dead than alive: exploring the current state of the multi-document summarisation of the biomedical literature. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5098–5111 (2022).

68. Otmakhova, J., Verspoor, K., Baldwin, T., Yepes, A. J. & Lau, J. H. M3: Multi-level dataset for multi-document summarisation of medical studies. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 3887–3901 (2022).

69. Adams, G., Zucker, J. & Elhadad, N. A meta-evaluation of faithfulness metrics for long-form hospital-course summarization. *arXiv preprint arXiv:2303.03948* (2023).

70. Laban, P., Schnabel, T., Bennett, P. N. & Hearst, M. A. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions Assoc. for Comput. Linguist.* **10**, 163–177 (2022).

71. Beltagy, I., Peters, M. E. & Cohan, A. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150* (2020).

72. Schardt, C., Adams, M. B., Owens, T., Keitz, S. & Fontelo, P. Utilization of the pico framework to improve searching pubmed for clinical questions. *BMC medical informatics decision making* **7**, 1–6 (2007).

73. Trienes, J., Schlötterer, J., Schildhaus, H.-U. & Seifert, C. Patient-friendly clinical notes: Towards a new text simplification dataset. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, 19–27 (2022).

74. Lu, J., Li, J., Wallace, B. C., He, Y. & Pergola, G. Napss: Paragraph-level medical text simplification via narrative prompting and sentence-matching summarization. *arXiv preprint arXiv:2302.05574* (2023).

75. Jeblick, K. *et al.* Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports. *arXiv preprint arXiv:2212.14882* (2022).

76. Lyu, Q. *et al.* Translating radiology reports into plain language using chatgpt and gpt-4 with prompt learning: Promising results, limitations, and potential. *arXiv preprint arXiv:2303.09038* (2023).

77. Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L. & Chissom, B. S. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Tech. Rep., Naval Technical Training Command Millington TN Research Branch (1975).

78. Senter, R. & Smith, E. A. Automated readability index. Tech. Rep., Cincinnati Univ OH (1967).

79. Xu, W., Napoles, C., Pavlick, E., Chen, Q. & Callison-Burch, C. Optimizing statistical machine translation for text simplification. *Transactions Assoc. for Comput. Linguist.* **4**, 401–415 (2016).

80. Devaraj, A., Sheffield, W., Wallace, B. C. & Li, J. J. Evaluating factuality in text simplification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7331–7345 (2022).

81. Nishino, T. *et al.* Reinforcement learning with imbalanced dataset for data-to-text medical report generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2223–2236 (2020).

82. Delbrouck, J.-B. *et al.* Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 4348–4360 (2022).

83. Nishino, T. *et al.* Factual accuracy is not enough: Planning consistent description order for radiology report generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 7123–7138 (2022).

84. Liu, G. *et al.* Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*, 249–269 (PMLR, 2019).

85. Yu, F. *et al.* Evaluating progress in automatic chest x-ray radiology report generation. *medRxiv* 2022–08 (2022).

86. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318 (2002).

87. Wan, D., Liu, M., McKeown, K., Dreyer, M. & Bansal, M. Faithfulness-aware decoding strategies for abstractive summarization. *arXiv preprint arXiv:2303.03278* (2023).

88. Radford, A., Sutskever, I., Kim, J. W., Krueger, G. & Agarwal, S. Clip: Learning an image classifier from language and pixels. https://openai.com/research/clip (2021).

89. Kotonya, N. & Toni, F. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7740–7754 (2020).

90. Wadden, D. *et al.* Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7534–7550 (2020).

91. Poliak, A. *et al.* Collecting verified covid-19 question answer pairs. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020* (2020).

92. Sarrouti, M., Abacha, A. B., M'rabet, Y. & Demner-Fushman, D. Evidence-based fact-checking of health-related claims. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3499–3512 (2021).

93. Saakyan, A., Chakrabarty, T. & Muresan, S. Covid-fact: Fact extraction and verification of real-world claims on covid-19 pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2116–2129 (2021).

94. Mohr, I., Wührl, A. & Klinger, R. Covert: A corpus of fact-checked biomedical covid-19 tweets. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 244–257 (2022).

95. Srba, I. *et al.* Monant medical misinformation dataset: Mapping articles to fact-checked claims. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2949–2959 (2022).

96. Wadhwa, S., Khetan, V., Amir, S. & Wallace, B. Redhot: A corpus of annotated medical questions, experiences, and claims on social media. *arXiv preprint arXiv:2210.06331* (2022).

97. Voorhees, E. *et al.* Trec-covid: constructing a pandemic information retrieval test collection. In *ACM SIGIR Forum*, vol. 54, 1–12 (ACM New York, NY, USA, 2021).

98. Lo, K., Wang, L. L., Neumann, M., Kinney, R. & Weld, D. S. S2orc: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4969–4983 (2020).

99. Beltagy, I., Lo, K. & Cohan, A. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3615–3620 (2019).

100. Liu, Y. & Lapata, M. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3730–3740 (2019).

101. Wadden, D. *et al.* Multivers: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, 61–76 (2022).

102. Alsentzer, E. *et al.* Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 72–78 (2019).

103. Abacha, A. B. & Demner-Fushman, D. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2228–2234 (2019).

104. Sudlow, C. *et al.* Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine* **12**, e1001779 (2015).

105. Lin, S., Hilton, J. & Evans, O. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3214–3252 (2022).

106. Wang, J. *et al.* Knowledge prompting in pre-trained language model for natural language understanding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 3164–3177 (Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022).