Journal of the American Medical Informatics Association, 2024, **31(4)**, 1009–1024
https://doi.org/10.1093/jamia/ocae015
Advance access publication 16 February 2024
Review

Review

Question answering systems for health professionals at the point of care—a systematic review

Gregory Kell, MPhil*[,1], Angus Roberts, PhD[2], Serge Umansky, PhD[3], Linglong Qian, MSc[2], Davide Ferrari, MSc[1], Frank Soboczenski, PhD[1], Byron C. Wallace, PhD[4], Nikhil Patel, BSc[1], Iain J. Marshall, PhD[1]

[1]Department of Population Health Sciences, King's College London, London, Greater London, SE1 1UL, United Kingdom, [2]Department of Biostatistics and Health Informatics, King's College London, London, Greater London, SE5 8AB, United Kingdom, [3]Metadvice Ltd, London, Greater London, SW1Y 5JG, United Kingdom, [4]Khoury College of Computer Sciences, Northeastern University, Boston, MA 02115, United States

*Corresponding author: Gregory Kell, MPhil, Department of Population Health Sciences, King's College London, Addison House, Guy's Campus, London, Greater London, SE1 1UL, United Kingdom (gregory.kell@kcl.ac.uk)

Abstract

Objectives: Question answering (QA) systems have the potential to improve the quality of clinical care by providing health professionals with the latest and most relevant evidence. However, QA systems have not been widely adopted. This systematic review aims to characterize current medical QA systems, assess their suitability for healthcare, and identify areas of improvement.

Materials and methods: We searched PubMed, IEEE Xplore, ACM Digital Library, ACL Anthology, and forward and backward citations on February 7, 2023. We included peer-reviewed journal and conference papers describing the design and evaluation of biomedical QA systems. Two reviewers screened titles, abstracts, and full-text articles. We conducted a narrative synthesis and risk of bias assessment for each study. We assessed the utility of biomedical QA systems.

Results: We included 79 studies and identified themes, including question realism, answer reliability, answer utility, clinical specialism, systems, usability, and evaluation methods. Clinicians' questions used to train and evaluate QA systems were restricted to certain sources, types and complexity levels. No system communicated confidence levels in the answers or sources. Many studies suffered from high risks of bias and applicability concerns. Only 8 studies completely satisfied any criterion for clinical utility, and only 7 reported user evaluations. Most systems were built with limited input from clinicians.

Discussion: While machine learning methods have led to increased accuracy, most studies imperfectly reflected real-world healthcare information needs. Key research priorities include developing more realistic healthcare QA datasets and considering the reliability of answer sources, rather than merely focusing on accuracy.

Key words: clinical decision support; question answering; evidence-based medicine; natural language processing; artificial intelligence.

Background and significance

Despite a plethora of available evidence, health professionals find answers to only half of their questions, due to time constraints.[1–3] This has motivated the development of online resources to answer clinicians' questions based on the latest evidence. While scientifically rigorous information resources such as UpToDate, Cochrane, and PubMed exist, Google search remains the most popular resource used in practice.[4] General-purpose search engines like Google offer ease-of-use, but rank results according to criteria that differ from evidence-based medicine principles of rigor, comprehensiveness, and reliability.[4]

To address these issues, there is burgeoning research into biomedical question answering (QA) systems.[5–13] These could rival the accessibility and speed of Google or "curbside consultations" with colleagues, while providing answers based on reliable, up-to-date evidence. Moreover, Google is free to access, while services such as UpToDate charge for access and require manual updates. On the other hand, biomedical QA systems could be updated automatically. More recently, rapid advances in language modeling (particularly large language models [LLMs] such as GPT,[14] and Galactica[15]) could allow healthcare professionals to request and receive natural language guidance summarizing evidence directly.

Many papers (eg, Refs. 5, 6, 8, 10, 16, 17) have described the development and evaluation of biomedical QA systems. However, the majority have not seen use in practice. We explored this problem previously,[18] and argue that key reasons for non-uptake include answers which are not useful in real-life clinical practice (eg, yes/no, factoids, or answers not applicable to the locality or setting); systems that do not justify answers, communicate uncertainties, or resolve contradictions.[5,6,10,16,17] Some existing papers have surveyed the literature on biomedical QA (eg, Refs. 19, 20) and found that few systems explain the reasoning for the returned answers, use all available domain knowledge, generate

Received: August 22, 2023; **Revised:** January 11, 2024; **Editorial Decision:** January 13, 2024; **Accepted:** January 15, 2024

answers that reflect conflicting sources and are able to answer non-English questions.

Our contributions are to comprehensively characterize existing systems and their limitations, with the hope of identifying key issues whose resolution would allow for QA systems to be used in practice. We focus on complete QA systems as opposed to subcomponents.

Methods

We conducted a systematic review and narrative synthesis of biomedical QA research, focusing on studies describing the development and evaluation of such systems. The protocol for this review is registered in PROSPERO (PROSPERO registration ID: CRD42021266053) and the Open Science Framework (OSF registration DOI: 10.17605/OSF.IO/4AM8D).

Studies were eligible if they were: (1) published in peer-reviewed conference proceedings and journals, (2) in English language, (3) described complete QA systems (ie, papers describing only subcomponent methods were excluded), evaluated the QA system (either based on a dataset of questions and answers, or a user study), (5) focused on biomedical QA for healthcare professionals. We excluded studies: (1) of QA systems for consumers/patients and (2) using modalities other than text, for example, vision. We searched PubMed, IEEE Xplore, ACM Digital Library, ACL Anthology, and forward and backward citations on February 7, 2023, using the following search strategy adapted for each database's syntax:

> *("question answering" OR "question-answering") AND (clinic* OR medic* OR biomedic* OR health*)*

Deduplicated titles and abstracts were double screened by G.K. (all) and D.F. and L.Q.A. (50% each). Disagreements were resolved via discussion, adjudicated by IJM. The same process was followed for full texts.

We used a structured data collection form which we refined after piloting (Appendix SA). We conducted a narrative synthesis following the steps recommended by Popay et al.[21] Specifically, we conducted an initial synthesis by creating textual descriptions of each study and tabulating data on methods, datasets, evaluation methods, and findings, and creating conceptual maps. We assessed the robustness of findings via a risk of bias assessment, and by evaluating QA systems' suitability for real-world use.

We evaluated the suitability of QA systems for use in practice, via criteria we developed previously and introduced in our position paper.[18] This paper described how problems with transparency, trustworthiness, and provenance of health information contribute to the non-adoption of QA systems in real-world use. We proposed the following markers of high-quality QA systems. (1) Answers should come from reliable sources; (2) systems should provide guidance where possible; (3) answers should be relevant to the clinician's setting; (4) sufficient rationale should accompany the answers; (5) conflicting evidence should be resolved appropriately; and (6) systems should consider and communicate uncertainties. We rated each system as completely, partially, or not meeting these criteria. We provide more detail regarding the application of these criteria in Appendix SB. Quality assessments were done in duplicate by G.K. (all papers), and L.Q. and D.F. (half of all papers each). Final assessments were decided through discussion and adjudicated by I.J.M.

In the absence of a directly relevant bias tool, we adapted PROBAST for use with QA studies.[22] PROBAST evaluates study design, conduct, or analysis which can lead to biases in clinical predictive modeling studies. QA systems are like predictive models, but rather than predicting a diagnosis (based on some clinical criteria), they predict the best answer for a given question.

We adapted PROBAST to consider the quality of studies' (1) questions (analogous to *population* in the original PROBAST), (2) input features (eg, bag-of-words, neural embeddings, etc., analogous to *predictors*), and (3) answers (analogous to *outcomes*). For each criterion, we assessed whether design problems led to *risk of bias*. We then assessed the studies for *applicability* concerns (ie, relevance of questions, models, and answers to general clinical practice). Risks of bias and applicability concerns were rated as high, low, or unclear for each paper. We provide the modified PROBAST in the Supplementary Materials; this may be useful to other researchers assessing QA systems. Other AI-focused tools (eg, APPRAISE-AI[23]) are rapidly becoming available; they cover similar aspects of bias to PROBAST.

We report our review according to the PRISMA[24] and SwiM guidance.[25] We provide raw data in the Supplementary Materials and present the final narrative synthesis below.

Results

The flow of studies, and reasons for inclusion/exclusion are shown in Figure 1. We included 79 of 7506 records identified in the searches in the final synthesis. Characteristics of included studies are described in Table 1 and Figure 2.

Risk of bias, applicability, and utility

We summarize the risks of bias in Figure 3; individual study assessments are in the Supplementary Materials. 85% of systems had a high risk of bias overall; primarily driven by problems in the questions used to develop and evaluate the systems. Many studies used unrealistically simple questions or covered too few information needs for a general biomedical QA system. Most questions were hypothetical, and not generated by health professionals.

Most systems were at low risk of bias for defining and extracting machine learning (ML) features (eg, deciding on predictive features without reference to the reference answers). Most studies did not provide clear descriptions of answer data or evaluation methodology (eg, details about the source of answers) which led to unclear risk of bias assessments for most papers' answers. Additionally, no answer was relevant to the biomedical QA domain. This led to high applicability concerns for most papers.

We present utility scores in Figure 4. Few systems completely met any criterion. Two systems[26,27] provided rationales (ie, justifications and sources) for their answers; 5 systems were judged to use reliable sources[11,28–31]; one system resolved conflicting information[26] and one system communicated uncertainties.[26] Very few systems provided contextually relevant answers (ie, locality-specific information, or specialty), while most systems provided clinical guidance at least partially (rather than basic science or less actionable information).

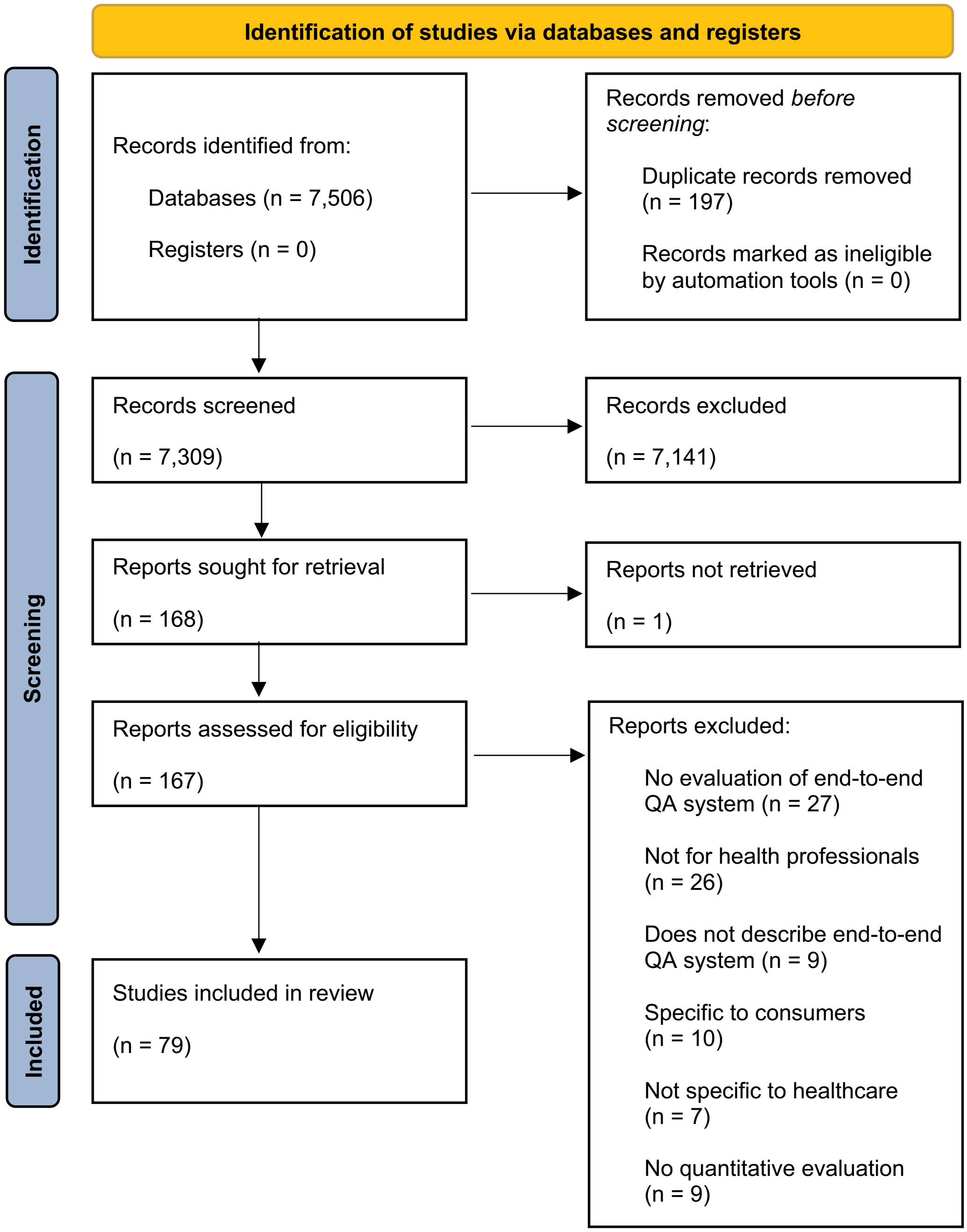
Identification of studies via databases and registers
Identification
Screening
Included
Records identified from:
Databases (n = 7,506)
Registers (n = 0)
Records removed *before screening*:
Duplicate records removed (n = 197)
Records marked as ineligible by automation tools (n = 0)
Records screened (n = 7,309)
Records excluded (n = 7,141)
Reports sought for retrieval (n = 168)
Reports not retrieved (n = 1)
Reports assessed for eligibility (n = 167)
Reports excluded:
No evaluation of end-to-end QA system (n = 27)
Not for health professionals (n = 26)
Does not describe end-to-end QA system (n = 9)
Specific to consumers (n = 10)
Not specific to healthcare (n = 7)
No quantitative evaluation (n = 9)
Studies included in review (n = 79)

Figure 1. PRISMA flow diagram.

Computational methods

Most QA systems used a knowledge base (ie, database of answers) that was created using documents from PubMed or other medical information sources (see Figure 5 for a typical example, from Alzubi et al[32]). Documents were either stored in structured form (knowledge graphs [KGs] or RDF triples) or as unstructured texts.

For a given user query, the system would retrieve the most relevant answer(s) from the knowledge base. KG-based (1 study), neural (24 studies) and modular systems (39 studies) were evaluated in the included studies (see Figure 2 and Appendix SC). KG-based systems accept natural language questions and convert them to KG-specific queries (eg, Cypher queries[33]) Modular systems comprise several distinct components (eg, question analysis, document retrieval, answer generation) designed separately and combined to form a QA system. Neural systems can be modular or monolithic.

All studies made use of datasets of questions with known answers. These datasets were used to train ML models (eg, document retrieval and answer extraction) and evaluate system performance. The topic focus of these datasets dictates the area(s) for which the QA can be successfully used; the quality of these datasets impacts both the accuracy of trained models and the reliability of the evaluations.

With regards to neural systems, 9 studies[32,34–40] incorporated pretrained LLMs (eg, BERT,[41] BioBERT,[42] and GPT[43]) in their QA pipelines for text span extraction, sentence reranking and integrating sentiment information. These models were used to find potential answer text spans given

Table 1. Characteristics of included studies.

Study	Model/method	Evaluation question sources	Evaluation answer sources
Demner-Fushman et al (2006a)[31]	Semantic type classifier (UMLS, MeSH) PICO classifier Rule-based system Machine learning system	Physicians	PubMed
Demner-Fushman et al (2006b)[53]	Semantic-type classifier (UMLS) Clustering	Authors	PubMed
Lee et al (2006)[56]	Question classification Query term generation TF-IDF Document retrieval Lexico-syntactic patterns	Physicians	PubMed World Wide Web
Weiming et al (2007)[54]	Semantic-type classifier (UMLS) Semantic relation extraction BM25 TF-IDF Boolean search	Unclear	Medical documents
Demner-Fushman et al (2007)[11]	Semantic-type classifier (UMLS, MeSH) PICO classifier Rule-based system Machine learning system	Physicians	PubMed
Sondhi et al (2007)[69]	Semantic-type classifier (UMLS, ICD-9) Document ranking Clustering	Physicians	PubMed
Yu et al (2007) a[47]	User study of different systems	Physicians in practice	World Wide Web Online dictionaries PubMed
Yu et al (2007) b[17]	Naïve Bayes Lexico-syntactic patterns TF-IDF Information retrieval	Physicians in practice	World Wide Web PubMed
Makar et al (2008)[50]	Bayesian classifier Part of speech tagger Text extractor Summarizer	Physicians in practice	Wikipedia Google
Cao et al (2009)[60]	BM25 Term frequency Unique term frequency Longest common subsequence	Physicians	MEDLINE eMedicine documents Clinical guidelines PubMed Central Wikipedia
Gobeil et al (2009)[68]	MeSH descriptors Information retrieval Information extraction	Authors	PubMed
Pasche et al (2009a)[81]	Logical rules Information retrieval	Authors	PubMed
Pasche et al (2009b)[55]	Logical rules Information retrieval	Authors	PubMed
Xu et al (2009)[71]	Semantic-type classifier (UMLS) Question-type classifier Keyword extractor Passage retrieval Answer extraction	Unclear	Unclear
Olvera-Lobo et al (2010)[70]	START: open-domain QA system MedQA: restricted-domain QA system	Health website	START: Wikipedia Merriam-Webster Dictionary American Medical Association I MDB Yahoo Webopedia.com MedQA: MEDLINE Dictionary of Cancer Terms Wikipedia Google Dorland's Illustrated Medical Dictionary Medline Plus Technical and Popular Medical Terms National Immunization Program Glossary

(continued)

Table 1. (continued)

Study	Model/method	Evaluation question sources	Evaluation answer sources
Tutos et al (2010)[28]	User study on different systems	Physicians	PubMed World Wide Web Brainboost
Cairns et al (2011)[12]	UMLS Rule-based algorithms Support vector machine	Physicians in practice	Medical wiki curated by approved physicians and doctoral-degreed biomedical students
Cao et al (2011)[5]	Semantic-type classifier (UMLS) Related questions extraction Information retrieval Information extraction Summarization	Unclear	Medical documents
Cruchet et al (2012)[72]	Semantic-type classifier (UMLS) Medical-term classifier Keyword-based retrieval	Physicians in practice	HONcode certified sites, for example, WebMD, Everyday Health, Drugs.com, and Healthline
Doucette et al (2012)[48]	Inference rules Semantic reasoner	Synthetic patient data	Synthetic patient data
Ni et al (2012)[29]	PICO classifier Rules-based system Template/pattern matching Information retrieval Machine learning system Answer candidate scoring	Medical health website	Medical health website
Ben Abacha and Zweigenbaum (2015)[6]	Semantic Web SPARQL Semantic graphs UMLS concepts UMLS semantic type Support vector machines Conditional random fields Rule-based methods	Physicians	PubMed
Gobeill et al (2015)[82]	Gene Ontology concepts Lazy pattern matching KNN BM25 Information retrieval	Authors	PubMed
Hristovski et al (2015)[57]	Semantic relation extraction (UMLS) Semantic relation retrieval	Authors	PubMed
Li et al (2015)[49]	Word2Vec Markov random field	Expert panel	PubMed
Tsatsaronis et al (2015)[52]	Comparison of different systems on the BioASQ dataset	Expert panel	PubMed
Vong et al (2015)[30]	PICO classifier Clustering	Authors	PubMed
Goodwin et al (2016)[8]	Knowledge graph Conditional random fields Bayesian inference	Unclear	Electronic health records PubMed
Yang et al (2016)[93]	Logistic Regression	Expert panel	PubMed
Brokos et al (2016)[103]	TF-IDF Word mover's distance	Expert panel	PubMed
Krithara et al (2016)[94]	Comparison of different systems on the BioASQ dataset	Expert panel	PubMed
Sarrouti and El Alaoui (2017) a[102]	UMLS concepts BM25	Expert panel	PubMed
Sarrouti and El Alaoui (2017) b[95]	UMLS concepts BM25	Expert panel	PubMed
Jin et al (2017)[113]	Bag of words Term frequency Collection frequency Sequential dependence models Divergence from randomness models Multimodal strategies	Expert panel	PubMed
Neves et al (2017)[108]	Question processing (regular expressions, semantic types, named entities, keywords) Document/passage retrieval Answer extraction	Expert panel	PubMed

(continued)

Table 1. (continued)

Study	Model/method	Evaluation question sources	Evaluation answer sources
Wiese et al (2017a)[109]	RNN Domain adaptation	Expert panel	PubMed
Wiese et al (2017b)[110]	RNN Domain adaptation	Expert panel	PubMed
Nentidis et al (2017)[96]	Comparison of different systems on the BioASQ dataset	Expert panel	PubMed
Du et al (2018)[62]	GloVe LSTM Self-attention	Expert panel	PubMed
Eckert et al (2018)[98]	Semantic role labelling	Expert panel	PubMed
Papagiannopoulou et al (2016)[97]	Binary relevance models Linear SVMs Labeled LDA variant Prior LDA Fast XML HOMER-BR Multilabel ensemble	Expert panel	PubMed
Dimitriadis et al (2019)[73]	Word2Vec WordNet Custom textual features Logistic regression Support vector machine XGBoost	Expert panel	PubMed
Du et al (2019)[63]	GloVe LSTM Self-attention Cross-attention	Expert panel	PubMed
Jin Q et al (2019)[101]	BioBERT	Titles of papers	PubMed
Ozyurt et al (2019)[37]	GloVe BERT Inverse document frequency Relaxed word mover's distance	Expert panel	PubMed
Jin ZX et al (2019)[61]	TF-IDF Noun extraction Part of speech tagger Semantic-type classifier (UMLS) Query expansion (MeSH) Markov random field Divergence from randomness Model ensemble	Expert panel	PubMed
Wasim et al (2019)[65]	Rules-based system Semantic-type classifier (UMLS) Logistic regression	Expert panel	PubMed
Oita et al (2020)[90]	Dynamic memory networks Bidirectional attention flow Transfer learning Biomedical named entity recognition Corroboration of semantic evidence	Expert panel	PubMed
Du et al (2020)[35]	BERT BiLSTM Self-attention	Expert panel	PubMed
Yan et al (2020)[112]	Binary classification RNNs Semi-supervised learning Recursive autoencoders	Expert panel	PubMed
Kaddari et al (2020)[58]	Survey of existing models	Expert panel	PubMed
Nishida et al (2020)[111]	BERT Domain adaptation Multitask learning	Expert panel Crowdworkers	PubMed Wikipedia
Omar et al (2020)[46]	Convolutional neural networks Attention Gated convolutions Gated attention	PubMed	PubMed
Ozyurt et al (2020a)[34]	GloVe BERT Inverse document frequency Relaxed word mover's distance	Expert panel	PubMed

(continued)

Table 1. (continued)

Study	Model/method	Evaluation question sources	Evaluation answer sources
Ozyurt et al (2020b)[104]	ELECTRA	Expert panel	PubMed
Sarrouti et al (2020)[16]	Lexico-syntactic patterns Support vector machine Semantic-type classifier (UMLS) TF-IDF Semantic similarity-based retrieval BM25 Sentiment analysis	Expert panel	PubMed
Shin et al (2020)[99]	BioMegatron	Expert panel	PubMed
Wang et al (2020)[107]	Event extraction SciBERT	Authors	PubMed
Alzubi et al (2021)[32]	TF-IDF BERT	Authors	PubMed
Du et al (2021)[76]	QANet BERT GloVe Model weighting	Expert panel	PubMed
Nishida et al (2021)[100]	BERT fastText	Expert panel Crowdworkers	PubMed Wikipedia
Peng et al (2021)[39]	BERT BiLSTM Bagging	Expert panel	PubMed
Pergola et al (2021)[92]	BERT Masking strategies	Epidemiologists Medical doctors Medical students Expert panel	PubMed World Health Organization's Covid-19 Database Preprint servers
Wu et al (2021)[27]	BERT Numerical encodings	Expert panel PubMed	PubMed
Xu et al (2021)[36]	BERT Syntactic and lexical features Feature fusion Transformer	Expert panel	PubMed
Bai et al (2022) a[79]	Dual encoder BioBERT	Expert panel	PubMed
Bai et al (2022) b[74]	Knowledge distillation Adversarial learning BioBERT	Expert panel	PubMed
Du et al (2022)[78]	QANet BERT GloVe Model weighting	Expert panel	PubMed
Kia et al (2022)[84]	Convolution neural network Attention	Authors	PubMed
Naseem et al (2022)[75]	ALBERT	Expert panel	PubMed
Pappas et al (2022)[105]	ALBERT-XL	Expert panel	PubMed
Raza et al (2022)[83]	BM25 MPNet	Expert panel	PubMed
Rakotoson et al (2022)[26]	BERT RoBERTa T5 Boolean classifier	Expert panel PubMed	PubMed
Wang et al (2022)[106]	Event extraction SciBERT Domain adaptation	Authors	PubMed
Weinzierl et al (2022)[77]	BERT BM25 Question generation Question entailment recognition	Expert panel	PubMed
Yoon et al (2022)[80]	BERT Sequence tagging BiLSTM-CRF	Expert panel	PubMed
Zhang et al (2022)[38]	BERT BM25	Expert panel	PubMed
Zhu et al (2022)[40]	BERT RoBERTa T5 XGBoost	PubMed	PubMed
Raza et al (2022)[85]	BM25 MPNet	Expert panel	PubMed

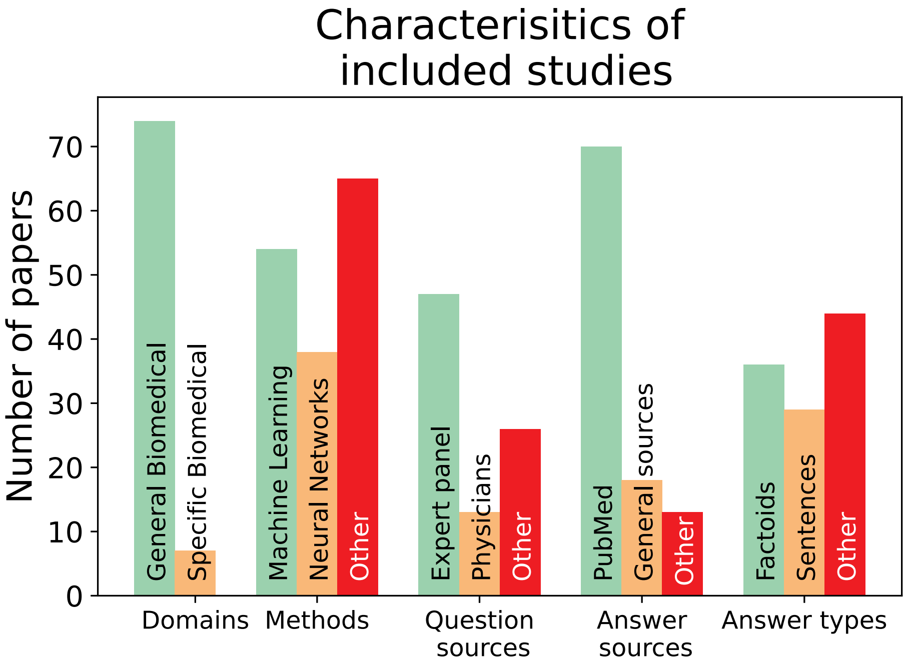

Figure 2. Number of papers with each category of domain, method, question, answer source, and answer type. The distinction was made between a major category and all the others, as one main category tended to dominate several smaller others. Table 1 contains more detail on the specifics of each paper.

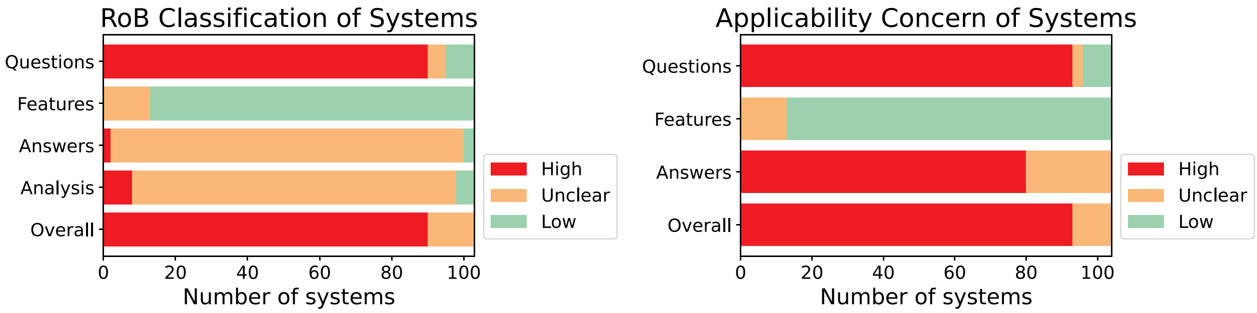

Figure 3. Number of papers achieving each risk of bias and applicability concern classification. Risk of bias refers to the risk of a divergence between the stated problem the paper tries to solve and the execution for reasons such as an unrealistic dataset or failing to split data for training and evaluation. Applicability refers to how applicable the system is to the review.

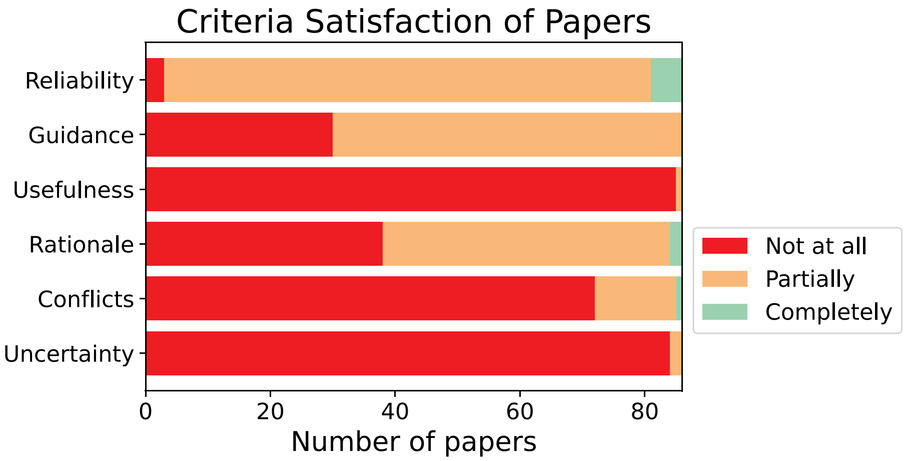

Figure 4. Number of papers achieving each satisfaction classification for each criterion.

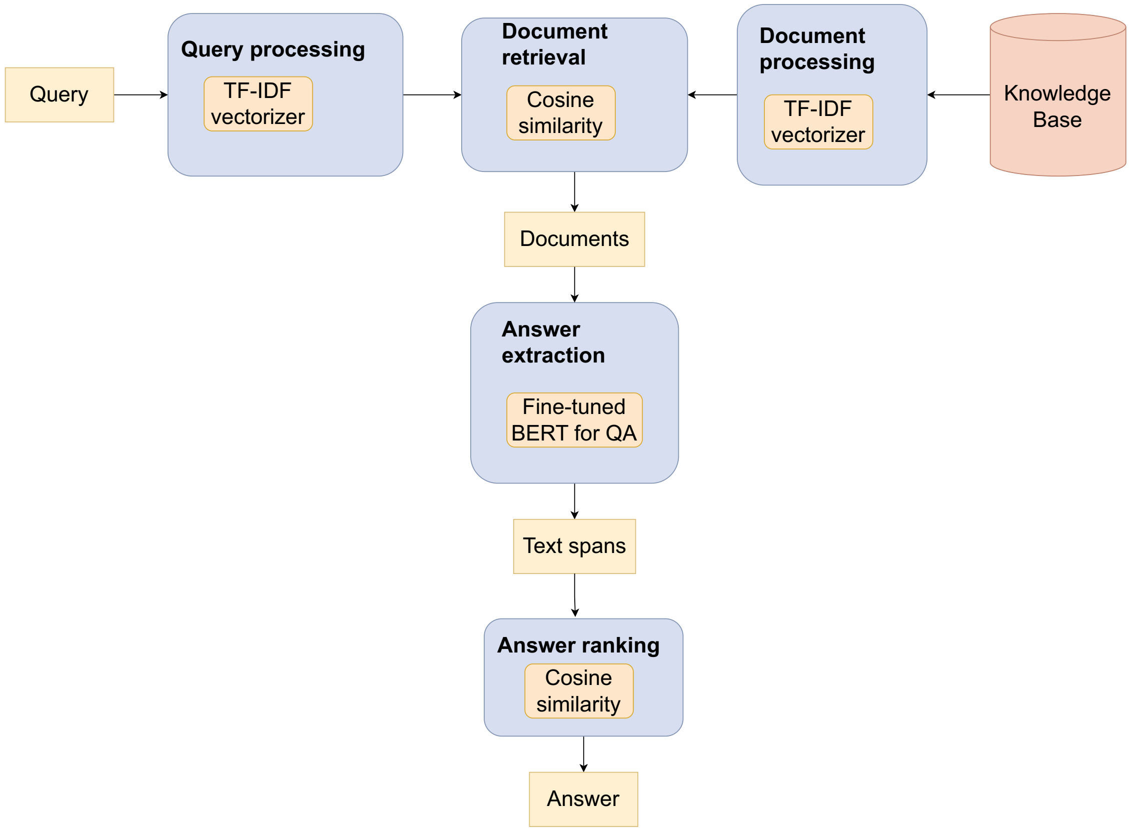

Figure 5. Typical QA architecture as used by Alzubi et al.[32]

questions and passages. Four studies[27,35,36,39] found fine-tuning pretrained LLMs on biomedical data led to improvements in performance compared with only using only a general-domain LLM. No experiments were conducted on LLMs that were trained only on biomedical data.

Few studies used common datasets for training or evaluation. However, several of the included studies arose from the BioASQ 5b[44] and 6b[45] shared tasks, which aimed to answer 4 types of questions (yes/no, factoids, list, and summary questions) and had 2 phases: information retrieval and exact answer production. Three studies arising from BioASQ[53,54,65] evaluated QA systems with a neural component, while 5 studies[52–54,57,65] evaluated QA systems that relied only on rule-based or classical ML components (eg, support vector machines). The neural components encoded questions and passages with a recurrent neural network (RNN) that were then used to create intermediate representations before answers were generated with additional layers. Comparing results across the BioASQ studies suggests generally that QA systems employing ML components outperformed those that relied solely on rule-based components (see Figure 6 and Appendix SC).

Two papers included a numerical component in their QA pipelines. For example, one paper[27] used numerical results (eg, odds ratios from clinical trial reports) to generate answers either to answer statistical questions (ie, “Do preoperative statins reduce atrial fibrillation after coronary artery bypass grafting?”). One study[27] generated BERT-style embeddings using both textual and numerical encodings, leading to improved performance compared with using text alone.

Topic areas

Fifty-three studies[5,8,11,16,17,26–31,34,35–40,46–80] described QA systems covering a wide breadth of biomedical topics (Figure 2). These systems typically sourced answers from the unfiltered medical literature (eg, PubMed, covering both clinical practice guidelines and primary studies, including laboratory science and epidemiology). Eight studies examined specific specialties: one study focused on bacteriotherapy,[81] 2 focused on genetics/genomics,[71,82] and 5 on Covid-19.[32,77,83–85] The genomics and Covid-19 systems were designed for specialists, while the bacteriotherapy system generated rules for managing antibiotic prescribing via a QA interface.

Question datasets

Studies used several sources to generate question datasets (see Figure 2 and Appendix SD). We group these into questions collected from health professionals (either collected in the course of work or elicited as generate hypothetical questions; 14 studies), those generated by topic experts (13 studies), people without direct healthcare experience (eg, crowd-workers; 3 studies), and automatically/algorithmically derived (scraped from health websites, or generated from abstract titles; 2 studies). In 9 papers, questions were written by study authors.

Only 5[17,47,50,60,80] studies used genuine questions posed by clinicians during consultations. Two studies[11,28] used either simple or simplified questions. Examples of simple questions include “How to beat recurrent UTIs?”[28] and “What is the best treatment for analgesic rebound headaches?”[11] Questions the BioASQ challenge questions[52] were

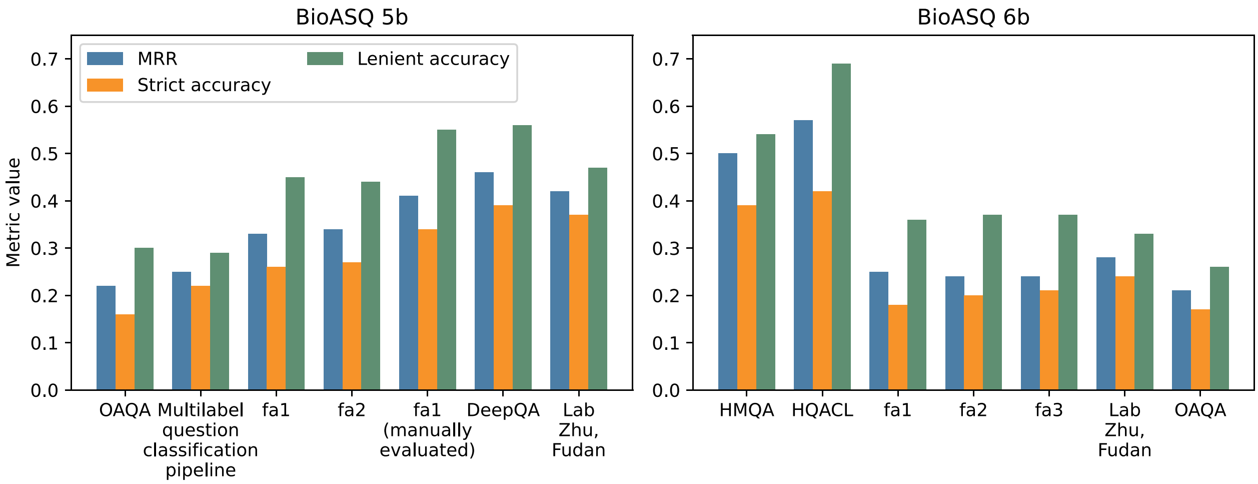

Figure 6. Results of the BioASQ 5b and 6b challenges for factoid-type answers.

created by an expert panel. BioASQ questions were restricted to yes/no, factoid and summary-type questions, and tended to have a highly technical focus. For example, the question "Which is the most common disease attributed to malfunction or absence of primary cilia?" could be answered with a factoid: "autosomal recessive polycystic kidney disease." Alternatively, it could be answered with a summary (see Appendix SE for example). One study included definition questions created by the authors,[70] while another[32] included author-created factoid-style questions about a particular topic. Two studies[28,29] utilized questions derived from health websites: one included questions generated by physicians[28] and one[29] used questions that were of unclear provenance.

While biomedical question sources enabled training of models, general domain QA datasets created using crowdworkers (eg, SQuAD[86,87]) were used to pretrain QA models in 3 studies.[35,62,63] These pretrained models were then fine-tuned on biomedical QA datasets (eg, BioASQ[52]) Pretraining QA models on general crowdworker-created datasets prior to fine-tuning on biomedical datasets led to overall improvements in model performance in all 3 studies that explored this approach. In other words, pretraining on general-domain data led to an improvement in performance compared with training only on biomedical data.

Reliability of answer sources

The answer sources used by the studies are summarized in Figure 2 and Appendix SG. Two studies[11,30] found ranking biomedical articles by strength of evidence (based on publication types, source journals and study design) improved accuracy (eg, precision at 10 documents, mean average precision, mean reciprocal rank). None of the other studies accounted for differences in answer reliability within datasets (ie, information from major guidelines was treated equally to a letter to the editor).

Several studies included answers derived from health websites such as Trip Answers,[29] WebMD,[70] HON-certified websites,[72] clinical guidelines and eMedicine documents (OSF registration DOI: 10.17605/OSF.IO/4AM8D). These answers were created by qualified physicians and underwent a review process. On the other hand, 3 studies[47,56,70] explored systems that provided only term definitions from medical dictionaries. One study derived answers entirely from general domain sources,[28] while another generated answers from a combination of medical and general sources. In the case of the latter, only the medical sources had a rigorous validation process.[70] Two QA systems[29,72] only derived the answers from health websites containing information that was vetted by the administrators. One study found that restricting the QA document collection based on trustworthiness increased the relevance of answers.[72]

Detail of answers

Systems we reviewed varied in terms of what they produced as an "answer" (Figure 2 and Appendix SH). Answers consisting of only of one word (ie, cloze-style QA), factoids (a word or phrase, eg, aspirin 3g), list of factoids, or definitions were absolute in nature and therefore did not contain guidance (Appendix SH). On the other hand, contextual texts (eg, ideal answers[52] and document summaries[16]) that accompanied absolute answers (eg, factoids) may have contained guidance. Similarly, biomedical articles accompanying answers consisting of medical concepts may have also included guidance, along with the sentences accompanying yes/no/unclear answers (see Appendix SH).

Several systems used a *clustered* approach to display answers. These systems grouped several candidate answers either by keyword or topics, eg, articles/sentences about heart conditions as one cluster. Clustered answers returned by the systems in 6 studies[5,30,53,54,60,69] may contain guidance as the clusters are based around sentences, extracts of documents, or conclusions of abstracts. Other types of answers included abstracts and single/multiple sentences, documents/webpages, and URL-based answers (Appendix SH).

Evaluation

Most studies (188) considered the accuracy of answers provided (see Table 2). Some assessed the degree to which the words in the answer match the reference, that is accuracy, precision, recall, F1 with respect to words (eg, ROUGE) or correct entire answers (eg, yes/no or factoids), numbers of answers/questions, exact matches. While ROUGE[88] or BLEU[89] may quantify the degree of similarity between candidate answers and the reference sentence, they are unable to account for, for example, negation or re-phrasings. Other systems were retrieval-based and so evaluated using the position of the correct answers in the returned list (ie, reciprocal rank, MAP, normalized discounted cumulative gain). Of the

Table 2. Grouping of papers according to accuracy metric.

Metric	Metric type	Papers	Number of papers
Accuracy	Accuracy/correctness	[16, 35, 36, 38–40, 46, 48, 50, 57, 58, 63, 65, 69, 73–75, 90, 92, 94–96, 98, 103–106, 109, 113]	29
Precision	Accuracy/correctness	[6, 11, 12, 16, 26, 29, 34, 40, 52, 54, 55, 57, 58, 65, 79, 80, 93–98, 103, 104, 107, 110–112]	28
Recall	Accuracy/correctness	[16, 26, 40, 52, 54, 58, 65, 70, 79–82, 93–96, 98, 103, 104, 107, 110–112]	24
Reciprocal rank	Accuracy/correctness	[6, 8, 12, 12, 16, 34, 35–39, 58, 62, 63, 65, 70, 71, 73, 74, 79, 94–96, 98, 100–107, 109]	32
F1	Accuracy/correctness	[16, 26, 29, 40, 52, 58, 62, 63, 65, 76, 78, 80, 83–85, 90, 93, 95, 96, 98, 100–105, 107–109, 111–113]	32
ROUGE	Accuracy/correctness	[16, 26, 31, 52, 90, 95, 96, 98, 100, 103]	10
Time taken to find answer	Usability	[5, 17, 26, 28, 47, 50	6
Likert score	Usability	5, 17, 28, 30, 47, 56, 60	7
Action frequency	Usability	17]	1
MAP	Accuracy/correctness	[52, 55, 61, 71, 91, 93, 99, 100, 104]	9
Numbers of queries/answers	Accuracy/correctness	[69–71, 100, 111]	5
Exact matches	Accuracy/correctness	[26, 32, 76, 78, 83–85, 107–109	10
Normalized discounted cumulative gain	Accuracy/correctness	77, 97]	2
AUC ROC	Accuracy/correctness	12	1

models that assessed accuracy/correctness, 31 used internal cross-validation, while 17 were evaluated on an independent dataset. Only 7 studies evaluated their design, system usability, or the relevance of the answer to the question as assessed by users. The most popular answer source was PubMed; most systems used a single source of answers.

Presentation and usability

Only 13 studies evaluated 7 systems that provided a user interface for user queries. These systems were MedQA,[17,28,47,70,114] Omed,[48] the system introduced in,[50] EAGLi,[55,81,82] AskHERMES,[5,30,60] CQA-1.0,[30] and CliniCluster.[30] User interfaces are essential for assessing the performance of the systems with genuine users.

The only usability study[30] assessed the effectiveness of a system that clustered answers to drug questions by I (intervention) and C (comparator) elements. The answers were tagged with P-O (patient-outcome) and I/C (intervention/comparator) elements (see Appendix SI for details). The participants agreed that the clustering of the answers helped them find answers more effectively, while more of the older participants found the P-O and I/C useful for finding relevant documents. Additionally, possessing prior knowledge about a given subject assisted with additional learning.

The ease of use of QA and IR systems was assessed in 3 studies.[5,48,17] The systems evaluated included Google,[5,17,48] MedQA,[17,48] Onelook,[17,48] PubMed,[17,48] UpToDate,[5] and AskHermes.[5] Both Doucette et al[48] and Yu et al[17] rated Google as being the easiest to use, followed by MedQA, Onelook, and PubMed. On the other hand, Cao et al[5] rated Google, UpToDate, and AskHermes equally in terms of ease of use.

None of the included systems presented any information about the certainty of answers; although nearly all systems used quantitative answer scoring to select the chosen answer. One study[60] evaluated 2 approaches to presenting answers on the AskHermes system:[5] passage-based (collection of several sentences) and sentence-based. The study found that passage-based approaches produced more relevant answers as rated by clinicians.

Discussion

We systematically reviewed studies of the development and evaluation of biomedical QA systems, focusing on their merits and drawbacks, evaluation and analysis, and the overall state of biomedical QA. Most of the included studies had high overall risks of bias and applicability concerns. Few of the papers satisfied any utility criterion.[18]

Several studies highlight obstacles that should be overcome and measures that should be taken before deploying biomedical QA systems. For example, one general-domain QA user study[115] found that users tended to prefer conventional search engines as they "felt less cognitive load" and "were more effective with it" than when they queried QA systems.

We note that commercial search engines are likely to benefit from comparatively vast development resources, and a focus on user experience. By contrast, the academic research we found tended to focus on the underlying computational methodology/models, with little attention to the user interface or experience—aspects which are likely highly influential in how QA systems are used.

Law et al[116] found that presenting users with causal claims and scatterplots could lead users to accept unfounded claims. Nonetheless, warning users that "correlation is not causation" led to more cautious treatment of reasonable claims. Additionally, Schuff et al[118] and Yang et al[117] explored metrics for assessing the quality of the explanations: answer location score (LOCA) and the Fact-Removal score (FARM), F1 score, and exact matches.

More recently, there has been rapid development in LLMs, such as GPT,[14] PaLM,[119] and Med-PaLM,[120] which are the current state of the art in natural language processing. There were 9 studies included that used LLMs, but they were used for text span extraction, sentence reranking and integrating sentiment information. A nascent application of LLMs is direct summarization of one or more sources. While LLMs can produce fluent answers to any given question,[121] they are vulnerable to "hallucinating" plausible but fabricated information.[122–124] This may be especially risky in healthcare due to the potentially life-threatening ramifications. One solution might be retrieval-augmented methods (where LLMS only use documents of known provenance). LLMs should be rigorously assessed before deployment in biomedical QA pipelines. This would ensure that the references provided by LLMs are genuine and that information is faithfully reproduced.

Barriers to adoption have been studied in detail in related technologies (eg, Clinical Decision Support Systems [CDSS]). Greenhalgh et al[125,126] introduced the NASSS framework to characterize the complex reasons why technologies succeed (or fail) in practice; finding that aspects such as the dependability of the underlying technology and organizations' readiness to adopt the new systems are critical. Similarly, Cimino and colleagues[127] found that design issues (eg, time taken to answer each question, or the number of times a given link is clicked) were critical. We would argue that future QA research should take a broader view of evaluation if QA is to move from an academic computer science challenge to real-world benefit.

To our knowledge, this is the first systematic review of QA systems in healthcare. While other (non-systematic) reviews provide an overview of the biomedical QA field,[19,20] we have evaluated existing systems and datasets for their utility in clinical practice. Furthermore, the inclusion of quantitative evaluations allowed for comparisons between different system types. Examination of questions, information sources, and answer types has allowed identification of factors that affect adherence to the criteria defined in.[18]

Most of the included studies were method papers describing systems that were built by computer scientists with limited input from clinicians. These systems were designed to perform well on benchmark datasets, such as BioASQ. While the studies were rigorous in their evaluation, they did consider how the systems could be used in practice. Future work should focus on translating biomedical QA research into practice.

One weakness is that we did not include purely qualitative evaluations. This might be a worthwhile SR to do in the future. We limited our search to published systems; therefore, this review would not have included any deployed systems which were not published; or systems described only in the "gray" literature (eg, pre-prints, PhD theses, etc.). We also did not search all the CDSS literature for pipelines incorporating QA systems. Deployment of such systems might not be described in the literature, as health providers may not have provided the results. Although we would expect most relevant papers to be published in English, there may have been pertinent non-English language papers that were missed.

Implications for research

Studies to date have too often used datasets of factoids/multiple choice questions, which do not resemble real-life queries. There is a need for high-quality datasets derived from real clinical queries, and actionable high-quality clinical guidance.

Future research should move beyond maximizing accuracy of a model alone, and include aspects of transparency, answer certainty, and information provenance (is the reliability and source of answers understood by users?). These aspects will only become more important with the advent of LLMs, which tend to generate highly plausible and fluent answers, but are not always correct.

Implications for practice

The performance of QA systems on biomedical tasks has increased over time, but the QA datasets were either unrepresentative of real-world information needs or were unrealistically simple. We recommend that practitioners exercise caution with any QA system which advertises accuracy only. Instead, systems should produce verifiable answers of known provenance, which make use of high-quality clinical guidelines and research.

Conclusions

In this review, we reviewed the literature on QA systems for health professionals. Most studies assessed the accuracy of the systems on various datasets; only 13 evaluated the usability of the systems. Few studies explored practical usage of the systems, opting to compare them using QA benchmarks instead. Although none of the included studies described systems that completely satisfied our utility criteria, they discussed several characteristics that could be appropriate for future systems. These included, limiting the document collection to reliable sources, providing more verbose answers, clustering answers according to themes/categories and employing methodologies for numerical reasoning. Most of the papers used QA datasets which either were unrepresentative of real-world information needs, or were unrealistically simple (eg, factoids, yes/no). Thus, more realistic and complex datasets should be developed.

Author contributions

Gregory Kell (Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Visualization, Writing—original draft), Linglong Qian (Formal Analysis, Investigation, Writing—review and editing), Davide Ferrari (Formal Analysis, Investigation, Writing—review and editing), Frank Soboczenski (Formal Analysis, Investigation, Writing—review and editing), Byron Wallace (Writing—review and editing), Angus Roberts (Writing—review and editing), Serge Umansky (Writing—review and editing), Nikhil Patel (Formal Analysis, Investigation, Writing—review and editing), and Iain J. Marshall (Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Visualization, Writing—review and editing, Supervision)

Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

Funding

This work was supported by the National Institutes of Health (NIH) under the National Library of Medicine (R01-LM012086-01A1), "Semi-Automating Data Extraction for Systematic Reviews." G.K. was supported by a studentship funded by King's College London and Metadvice Ltd; there is no grant number associated with this funding.

Conflicts of interest

None declared.

Data availability

The data underlying this article are available in the article and in its online supplementary material.

References

1. Del Fiol G, Workman TE, Gorman PN. Clinical questions raised by clinicians at the point of care: a systematic review. *JAMA Intern Med*. 2014;174(5):710-718.
2. Bastian H, Glasziou P, Chalmers I. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS Med*. 2010;7(9):e1000326.
3. Hoogendam A, Stalenhoef AFH, Robbé PFdV, Overbeke AJPM. Answers to questions posed during daily patient care are more likely to be answered by UpToDate than PubMed. *J Med Internet Res*. 2008;10(4):e29.
4. Hider P, Griffin G, Walker M, Coughlan E. The information-seeking behavior of clinical staff in a large health care organization. *J Med Libr Assoc*. 2009;97(1):47-50.
5. Cao Y, Liu F, Simpson P, et al. AskHERMES: an online question answering system for complex clinical questions. *J Biomed Inform*. 2011;44(2):277-288.
6. Ben Abacha A, Zweigenbaum P. MEANS: a medical question-answering system combining NLP techniques and semantic web technologies. *Inform Procss Manage*. 2015;51(5):570-594.
7. Terol RM, Martínez-Barco P, Palomar M. A knowledge based method for the medical question answering problem. *Comput Biol Med*. 2007;37(10):1511-1521.
8. Goodwin TR, Harabagiu SM. Medical question answering for clinical decision support. In: *Proceedings of the ACM International Conference on Information & Knowledge Management*. 2016:297-306.
9. Ben Abacha A, Demner-Fushman D. A question-entailment approach to question answering. *BMC Bioinform*. 2019;20 (1):511.
10. Zahid MAH, Mittal A, Joshi RC, Atluri G. CLINIQA: A Machine Intelligence Based Clinical Question Answering System. ArXiv [Internet]. 2018;abs/1805.05927. https://api.semanticscholar.org/CorpusID:21689474
11. Demner-Fushman D, Lin J. Answering clinical questions with knowledge-based and statistical techniques. *Comput Linguist*. 2007;33(1):63-103.
12. Cairns B, Nielsen RD, Masanz JJ, et al. The MiPACQ clinical question answering system. AMIA. *Annu Sympos Proc AMIA Symp*. 2011;2011:171-180.
13. Niu Y, Hirst G, McArthur G, Rodriguez-Gianolli P. Answering clinical questions with role identification. In: *Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine—Volume 13*. Association for Computational Linguistics; 2003:73-80. https://doi.org/10.3115/1118958.1118968
14. Brown TB, Mann B, Ryder N, et al. Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems *(NIPS'20)*. Curran Associates Inc; 2020.
15. Taylor R, Kardas M, Cucurull G, et al. *Galactica: A Large Language Model for Science*. ArXiv[Internet]. 2022:abs/2211.09085. https://arxiv.org/abs/2211.09085
16. Sarrouti M, Ouatik El Alaoui S. SemBioNLQA: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions. *Artif Intell Med*. 2020;102:101767.
17. Yu H, Lee M, Kaufman D, et al. Development, implementation, and a cognitive evaluation of a definitional question answering system for physicians. *J Biomed Inform*. 2007;40(3):236-251.
18. Kell G, Marshall I, Wallace B, Jaun A. What would it take to get biomedical QA systems into practice? In: *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*. Association for Computational Linguistics; 2021:28-41. https://aclanthology.org/2021.mrqa-1.3
19. Athenikos SJ, Han H. Biomedical question answering: a survey. *Comput Methods Programs Biomed*. 2010;99(1):1-24.
20. Jin Q, Yuan Z, Xiong G, et al. Biomedical question answering: a survey of approaches and challenges. *ACM Comput Surv*. 2022;55(2):1-36. https://doi.org/10.1145/3490238
21. Popay J, Roberts H, Sowden A, et al. *Guidance on the Conduct of Narrative Synthesis in Systematic Reviews: A Product from the ESRC Methods Programme*. *J Epidemiol Community Health*. 2006;59(Suppl 1):A7. https://doi.org/10.13140/2.1.1018.4643
22. Wolff RF, Moons KGM, Riley RD, et al.; PROBAST Group. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. 2019;170(1):51-58.
23. Kwong JCC, Khondker A, Lajkosz K, et al. APPRAISE-AI Tool for quantitative evaluation of AI studies for clinical decision support. *JAMA Netw Open*. 2023;6(9):e2335377.
24. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71.
25. Campbell M, McKenzie JE, Sowden A, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. *BMJ*. 2020;368:I6890. https://www.bmj.com/content/368/bmj.l6890
26. Rakotoson L, Letaillieur C, Massip S, Laleye FAA. Extractive-boolean question answering for scientific fact checking. In: *Proceedings of the 1st International Workshop on Multimedia AI against Disinformation (MAD '22)*. Association for Computing Machinery; 2022:27-34. https://doi.org/10.1145/3512732.3533580
27. Wu Y, Ting HF, Lam TW, Luo R. BioNumQA-BERT: answering biomedical questions using numerical facts with a deep language representation model. In: *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB '21)*. Association for Computing Machinery; 2021. https://doi.org/10.1145/3459930.3469557
28. Tutos A, Mollá D. A study on the use of search engines for answering clinical questions. In: *Proceedings of the Fourth Australasian Workshop on Health Informatics and Knowledge Management (HIKM '10)*, Vol. 108. Australian Computer Society, Inc.; 2010:61-68.
29. Ni Y, Zhu H, Cai P, Zhang LEI, Qui Z, Cao F. CliniQA : highly reliable clinical question answering system. *Stud Health Technol Inform*. 2012;180:215-219.
30. Vong W, Then PHH. Information seeking features of a PICO-based medical question-answering system. In: *2015 9th International Conference on IT in Asia (CITA)*. 2015;1-7.
31. Demner-Fushman D, Lin J. Situated question answering in the clinical domain: selecting the best drug treatment for diseases. In: *Proceedings of the Workshop on Task-Focused Summarization and Question Answering (SumQA '06)*. Association for Computational Linguistics; 2006:24-31.
32. Alzubi JA, Jain R, Singh A, Parwekar P, Gupta M. COBERT: COVID-19 question answering system using BERT. *Arab J Sci Eng*. 2021;48:11003–11013. https://doi.org/10.1007/s13369-021-05810-5

33. Francis N, Green A, Guagliardo P, et al. Cypher: an evolving query language for property graphs. In: *Proceedings of the 2018 International Conference on Management of Data (SIGMOD '18)*. Association for Computing Machinery; 2018:1433-1445. https://doi.org/10.1145/3183713.3190657
34. Ozyurt IB, Bandrowski A, Grethe JS. Bio-AnswerFinder: a system to find answers to questions from biomedical texts. *Database*. 2020;2020:baz137.
35. Du Y, Pei B, Zhao X, Ji J. Deep scaled dot-product attention based domain adaptation model for biomedical question answering. *Methods*. 2020;173:69-74. https://doi.org/10.1016/j.ymeth.2019.06.024
36. Xu G, Rong W, Wang Y, Ouyang Y, Xiong Z. External features enriched model for biomedical question answering. *BMC Bioinform*. 2021;22(1):272.
37. Ozyurt IB, Grethe J. Iterative document retrieval via deep learning approaches for biomedical question answering. In: *2019 15th International Conference on eScience (eScience)*. 2019:533-538.
38. Zhang X, Jia Y, Zhang Z, Kang Q, Zhang Y, Jia H. Improving end-to-end biomedical question answering system. In: *Proceedings of the 8th International Conference on Computing and Artificial Intelligence (ICCAI '22)*. Association for Computing Machinery; 2022:274-279. https://doi.org/10.1145/3532213.3532254
39. Peng K, Yin C, Rong W, Lin C, Zhou D, Xiong Z. Named entity aware transfer learning for biomedical factoid question answering. *IEEE/ACM Trans Comput Biol Bioinform*. 2022;19(4):2365-2376. https://doi.org/10.1109/TCBB.2021.3079339
40. Zhu X, Chen Y, Gu Y, Xiao Z. SentiMedQAer: a transfer learning-based sentiment-aware model for biomedical question answering. *Front Neurorobot*. 2022;16:773329. https://doi.org/10.3389/fnbot.2022.773329
41. Devlin J, Chang MW, Lee K, Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding [Internet]. North American Chapter of the Association for Computational Linguistics; 2019. Accessed August 23, 2022. https://doi.org/10.18653/v1/N19-1423
42. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*. 2020;36(4):1234-1240.
43. Radford A, Narasimhan K. Improving language understanding by generative pre-training [Internet]. 2018. Accessed August 22, 2022. https://api.semanticscholar.org/CorpusID:49313245
44. Krallinger M, Krithara A, Nentidis A, Paliouras G, Villegas M. BioASQ at CLEF2020: large-scale biomedical semantic indexing and question answering. In: Jose JM, Yilmaz E, Magalhães J, Castells P, Ferro N, Silva MJ, et al., eds. *Advances in Information Retrieval*. Springer International Publishing; 2020:550-556.
45. Nentidis A, Krithara A, Bougiatiotis K, Paliouras G, Kakadiaris I. Results of the sixth edition of the BioASQ Challenge. In: *Proceedings of the 6th BioASQ Workshop A Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering*. Association for Computational Linguistics; 2018:1-10. https://www.aclweb.org/anthology/W18-5301
46. Omar R, El-Makky N, Torki M. A character aware gated convolution model for cloze-style medical machine comprehension. In: *2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA)*. 2020:1-7.
47. Yu H, Kaufman D. A cognitive evaluation of four online search engines for answering definitional questions posed by physicians. *Pac Symp Biocomput*. 2007:328-339.
48. Doucette JA, Khan A, Cohen R. A comparative evaluation of an ontological medical decision support system (OMeD) for critical environments. In: *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium (IHI '12)*. Association for Computing Machinery; 2012:703-708. https://doi.org/10.1145/2110363.2110444
49. Li Y, Yin X, Zhang B, Liu T, Zhang Z, Hao H. A generic framework for biomedical snippet retrieval. In: *2015 3rd International Conference on Artificial Intelligence, Modelling and Simulation (AIMS)*. 2015:91-95.
50. Makar R, Kouta M, Badr AA. Service oriented architecture for biomedical question answering system. In: *2008 IEEE Congress on Services Part II (Services-2 2008)*. 2008:73-80.
51. Wen A, Elwazir MY, Moon S, Fan J. Adapting and evaluating a deep learning language model for clinical why-question answering. *JAMIA Open*. 2020;3(1):16-20.
52. Tsatsaronis G, Balikas G, Malakasiotis P, et al. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*. 2015;16:138. https://doi.org/10.1186/s12859-015-0564-6
53. Demner-Fushman D, Lin J. Answer extraction, semantic clustering, and extractive summarization for clinical question answering. In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (ACL-44)*. Association for Computational Linguistics; 2006:841-848. https://doi.org/10.3115/1220175.1220281
54. Weiming W, Hu D, Feng M, Wenyin L. Automatic clinical question answering based on UMLS relations. In: *Third International Conference on Semantics, Knowledge and Grid (SKG 2007)*. 2007:495–498. https://doi.org/10.1109/SKG.2007.126
55. Pasche E, Teodoro D, Gobeill J, Ruch P, Lovis C. Automatic medical knowledge acquisition using question-answering. *Stud Health Technol Inform*. 2009;150:569-573.
56. Lee M, Cimino J, Zhu H, et al. Beyond information retrieval—medical question answering. *AMIA Annu Symp Proc*. 2006;2006:469-473.
57. Hristovski D, Dinevski D, Kastrin A, Rindflesch TC. Biomedical question answering using semantic relations. *BMC Bioinform*. 2015;16(1):6.
58. Kaddari Z, Mellah Y, Berrich J, Bouchentouf T, Belkasmi MG. Biomedical question answering: a survey of methods and datasets. In: *2020 Fourth International Conference on Intelligent Computing in Data Sciences (ICDS)*. 2020:1-8.
59. Singh Rawat BP, Li F, Yu H. Clinical judgement study using question answering from electronic health records. *Proc Mach Learn Res*. 2019;106:216-229.
60. Cao YG, Ely J, Antieau L, Yu H. Evaluation of the clinical question answering presentation. In: *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing (BioNLP '09)*. Association for Computational Linguistics; 2009:171-178.
61. Jin ZX, Zhang BW, Fang F, Zhang LL, Yin XC. Health assistant: answering your questions anytime from biomedical literature. *Bioinformatics*. 2019;35(20):4129-4139.
62. Du Y, Pei B, Zhao X, Ji J. Hierarchical multi-layer transfer learning model for biomedical question answering. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2018;362–367.
63. Du Y, Guo W, Zhao Y. Hierarchical question-aware context learning with augmented data for biomedical question answering. In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2019:370-375.
64. Mairittha T, Mairittha N, Inoue S. Improving fine-tuned question answering models for electronic health records. In: *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers (UbiComp-ISWC '20)*. Association for Computing Machinery; 2020:688-691. https://doi.org/10.1145/3410530.3414436
65. Wasim M, Mahmood W, Asim MN, Khan MU. Multi-label question classification for factoid and list type questions in biomedical question answering. *IEEE Access*. 2019;7:3882-3896.
66. Ruan T, Huang Y, Liu X, Xia Y, Gao J. QAnalysis: a question-answer driven analytic tool on knowledge graphs for leveraging electronic medical records for clinical research. *BMC Med Inform Decis Mak*. 2019;19(1):82.

67. Qiu J, Zhou Y, Ma Z, Ruan T, Liu J, Sun J. Question answering based clinical text structuring using pre-trained language model. In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2019:1596-1600.
68. Gobeill J, Patsche E, Theodoro D, Veuthey A, Lovis C, Ruch P. Question answering for biology and medicine. In: *2009 9th International Conference on Information Technology and Applications in Biomedicine*. 2009:1-5.
69. Sondhi P, Raj P, Kumar VV, Mittal A. Question processing and clustering in INDOC: a biomedical question answering system. *EURASIP J Bioinform Syst Biol*. 2007;2007(1):28576.
70. Olvera-Lobo MD, Gutiérrez-Artacho J. Question-answering systems as efficient sources of terminological information: an evaluation. *Health Info Libr J*. 2010;27(4):268-276.
71. Xu B, Lin H, Liu B. Study on question answering system for biomedical domain. In: *2009 IEEE International Conference on Granular Computing*. 2009:626-629.
72. Cruchet S, Boyer C, van der Plas L. Trustworthiness and relevance in web-based clinical question answering. *Stud Health Technol Inform*. 2012;180:863-867.
73. Dimitriadis D, Tsoumakas G. Word embeddings and external resources for answer processing in biomedical factoid question answering. *J Biomed Inform*. 2019;92:103118. https://doi.org/10.1016/j.jbi.2019.103118
74. Bai J, Yin C, Zhang J, et al. Adversarial knowledge distillation based biomedical factoid question answering. *IEEE/ACM Trans Comput Biol Bioinform*. 2022. https://doi.org/10.1109/TCBB.2022.3161032
75. Naseem U, Dunn AG, Khushi M, Kim J. Benchmarking for biomedical natural language processing tasks with a domain specific ALBERT. *BMC Bioinform*. 2022;23(1):144.
76. Du Y, Yan J, Zhao Y, Lu Y, Jin X. Dual model weighting strategy and data augmentation in biomedical question answering. In: *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2021:659-662.
77. Weinzierl MA, Harabagiu SM. Epidemic question answering: question generation and entailment for Answer Nugget discovery. *J Am Med Inform Assoc*. 2023;30(2):329-339.
78. Du Y, Yan J, Lu Y, Zhao Y, Jin X. Improving biomedical question answering by data augmentation and model weighting. *IEEE/ACM Trans Comput Biol Bioinform*. 2022. https://doi.org/10.1109/TCBB.2022.3171388
79. Bai J, Yin C, Wu Z, et al. Improving biomedical ReQA with consistent NLI-transfer and post-whitening. *IEEE/ACM Trans Comput Biol Bioinform*. 2023;20(3):1864-1875.
80. Yoon W, Jackson R, Lagerberg A, Kang J. Sequence tagging for biomedical extractive question answering. *Bioinformatics*. 2022;38(15):3794-3801.
81. Pasche E, Teodoro D, Gobeill J, Ruch P, Lovis C. QA-driven guidelines generation for bacteriotherapy. *AMIA Annu Symp Proc*. 2009;2009:509-513.
82. Gobeill J, Gaudinat A, Pasche E, et al. Deep Question Answering for protein annotation. *Database*. 2015;2015:bav081. https://doi.org/10.1093/database/bav081
83. Raza S, Schwartz B, Ondrusek N. A question-answering system on COVID-19 scientific literature. In: *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*. 2022:1331-1336.
84. Kia MA, Garifullina A, Kern M, Chamberlain JON, Jameel S. Adaptable closed-domain question answering using contextualized CNN-attention models and question expansion. *IEEE Access*. 2022;10:45080-45092. https://doi.org/10.1109/ACCESS.2022.3170466
85. Raza S, Schwartz B, Rosella LC. CoQUAD: a COVID-19 question answering dataset system, facilitating research, benchmarking, and practice. *BMC Bioinform*. 2022;23(1):210.
86. Rajpurkar P, Zhang J, Lopyrev K, Liang P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2016;2016:2383–2392. https://doi.org/10.18653/v1/D16-1264
87. Rajpurkar P, Jia R, Liang P. Know What You Don't Know: Unanswerable Questions for SQuAD. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. 2018;2:784–789. https://doi.org/10.18653/v1/P18-2124
88. Lin CY. ROUGE: a package for automatic evaluation of summaries In: *Text Summarization Branches Out*. Association for Computational Linguistics; 2004;2004:74-81. https://aclanthology.org/W04-1013
89. Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics; 2002:311-318. https://aclanthology.org/P02-1040
90. Oita M, Vani K, Oezdemir-Zaech F. Semantically corroborating neural attention for biomedical question answering In: Cellier P, Driessens K, eds. *Machine Learning and Knowledge Discovery in Databases*. Springer International Publishing; 2020:670-685.
91. Arabzadeh N, Bagheri E. A self-supervised language model selection strategy for biomedical question answering. *J Biomed Inform*. 2023;146(1):104486. https://doi.org/10.1016/j.jbi.2023.104486
92. Pergola G, Kochkina E, Gui L, Liakata M, He Y. Boosting low-resource biomedical QA via entity-aware masking strategies. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics; 2021. https://doi.org/10.18653%2Fv1%2F2021.eacl-main.169
93. Yang Z, Zhou Y, Nyberg E. Learning to answer biomedical questions: OAQA at BioASQ 4B. In: *Proceedings of the Fourth BioASQ Workshop*. Association for Computational Linguistics; 2016. https://doi.org/10.18653%2Fv1%2Fw16-3104
94. Krithara A, Nentidis A, Kakadiaris I. Results of the 4th edition of BioASQ Challenge. In: *Proceedings of the Fourth BioASQ workshop*, 2016.
95. Sarrouti M, Alaoui SOE. A biomedical question answering system in BioASQ 2017. In: *BioNLP 2017*. Association for Computational Linguistics; 2017. https://doi.org/10.18653%2Fv1%2Fw17-2337
96. Nentidis A, Bougiatiotis K, Krithara A, Paliouras G, Kakadiaris I. Results of the fifth edition of the BioASQ Challenge—BioNLP 2017. In: *BioNLP 2017*. 2017. https://doi.org/10.18653%2Fv1%2Fw17-2306
97. Papagiannopoulou E, Papanikolaou Y, Dimitriadis D, et al. Large-scale semantic indexing and question answering in biomedicine. In: *Proceedings of the Fourth BioASQ Workshop*. Association for Computational Linguistics; 2016. https://doi.org/10.18653%2Fv1%2Fw16-3107
98. Eckert F, Neves M. Semantic role labeling tools for biomedical question answering: a study of selected tools on the BioASQ datasets. In: *Proceedings of the 6th BioASQ Workshop A Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering*. Association for Computational Linguistics; 2018. https://doi.org/10.18653%2Fv1%2Fw18-5302
99. Shin HC, Zhang Y, Bakhturina E, et al. BioMegatron: larger biomedical domain language model. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics; 2020. https://doi.org/10.18653%2Fv1%2F2020.emnlp-main.379
100. Nishida K, Nishida K, Yoshida S. Task-adaptive pre-training of language models with word embedding regularization. In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics; 2021. https://doi.org/10.18653%2Fv1%2F2021.findings-acl.398
101. 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). 2019:2567–2577. https://doi.org/10.18653/v1/D19-1259

102. Sarrouti M, Ouatik El Alaoui S. A passage retrieval method based on probabilistic information retrieval model and UMLS concepts in biomedical question answering. *J Biomed Inform*. 2017;68:96-103. https://doi.org/10.1016/j.jbi.2017.03.001
103. Brokos GI, Malakasiotis P, Androutsopoulos I. Using centroids of word embeddings and word mover's distance for biomedical document retrieval in question answering. In: *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics; 2016. https://doi.org/10.18653%2Fv1%2Fw16-2915
104. Ozyurt IB. On the effectiveness of small, discriminatively pre-trained language representation models for biomedical text mining. In: *Proceedings of the First Workshop on Scholarly Document Processing*. Association for Computational Linguistics; 2020. https://doi.org/10.18653%2Fv1%2F2020.sdp-1.12
105. Pappas D, Malakasiotis P, Androutsopoulos I. Data augmentation for biomedical factoid question answering. In: *Proceedings of the 21st Workshop on Biomedical Language Processing*. Association for Computational Linguistics; 2022. https://doi.org/10.18653%2Fv1%2F2022.bionlp-1.6
106. Wang XD, Leser U, Weber L. BEEDS: large-scale biomedical event extraction using distant supervision and question answering. In: *Proceedings of the 21st Workshop on Biomedical Language Processing*. Association for Computational Linguistics; 2022. https://doi.org/10.18653%2Fv1%2F2022.bionlp-1.28
107. Wang XD, Weber L, Leser U. Biomedical event extraction as multi-turn question answering. In: *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*. Association for Computational Linguistics; 2020. https://doi.org/10.18653%2Fv1%2F2020.louhi-1.10
108. Neves M, Eckert F, Folkerts H, Uflacker M. Assessing the performance of Olelo, a real-time biomedical question answering application—BioNLP 2017. In: *BioNLP 2017*. 2017. https://doi.org/10.18653%2Fv1%2Fw17-2344
109. Wiese G, Weissenborn D, Neves M. Neural domain adaptation for biomedical question answering. In: *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. 2017. https://doi.org/10.18653%2Fv1%2Fk17-1029
110. Wiese G, Weissenborn D, Neves M. Neural question answering at BioASQ 5B—BioNLP 2017. In: *BioNLP 2017*. 2017. https://doi.org/10.18653%2Fv1%2Fw17-2309
111. Nishida K, Nishida K, Saito I, Asano H, Tomita J, et al. Unsupervised domain adaptation of language models for reading comprehension. In: Calzolari N, Béchet F, Blache P, Choukri K, Cieri C, Declerck T, eds. *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 2020:5392-5399.
112. Yan Y, Zhang BW, Li XF, Liu Z. List-wise learning to rank biomedical question-answer pairs with deep ranking recursive autoencoders. *PLoS One*. 2020;15(11):e0242061.
113. Jin ZX, Zhang BW, Fang F, Zhang LL, Yin XC. A multi-strategy query processing approach for biomedical question answering: USTB_PRIR at BioASQ 2017 Task 5B. In: *BioNLP 2017*. Association for Computational Linguistics; 2017. https://doi.org/10.18653%2Fv1%2Fw17-2348
114. Lee M, Cimino J, Zhu HR, et al. Beyond information retrieval-medical question answering. *AMIA Annu Symp Proc*. 2006;2006:469-473.
115. Robles-Flores JA, Roussinov D. Examining question-answering technology from the task technology fit perspective. *Commun Assoc Inf Syst*. 2012;30:26. https://doi.org/10.17705/1cais.03026
116. Law PM, Lo LYH, Endert A, Stasko J, Qu H. Causal perception in question-answering systems. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery; 2021. https://doi.org/10.1145/3411764.3445444
117. Yang Z, Qi P, Zhang S, et al. HotpotQA: a dataset for diverse, explainable multi-hop question answering. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics; 2018:2369-2380. https://aclanthology.org/D18-1259
118. Schuff H, Adel H, Vu NT. F1 is not enough! models and evaluation towards user-centered explainable question answering. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics; 2020:7076-7095. https://aclanthology.org/2020.emnlp-main.575
119. Chowdhery A, Narang S, Devlin J, et al. PaLM: scaling language modeling with pathways. *J Mach Learn Res (JMLR)*. 2022:24(240):1–113.
120. Singhal K, Azizi S, Tu T, et al. 2022. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172–180. https://doi.org/10.1038/s41586-023-06291-2.
121. Shaib C, Li ML, Joseph S, Marshall IJ, Li JJ, Wallace BC. Summarizing, simplifying, and synthesizing medical evidence using GPT-3 (with varying success). In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. 2023;2:1387–1407. https://doi.org/10.18653/v1/2023.acl-short.119
122. Das S, Saha S, Srihari R. Diving deep into modes of fact hallucinations in dialogue systems. In: Goldberg Y, Kozareva Z, Zhang Y, eds. *Findings of the Association for Computational Linguistics: EMNLP 2022 [Internet]*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics; 2022:684-99. https://aclanthology.org/2022.findings-emnlp.48
123. Kuhn L, Gal Y, Farquhar S. CLAM: selective clarification for ambiguous questions with large language models [Internet]. arXiv; 2022. https://arxiv.org/abs/2212.07769
124. Weidinger L, Mellor J, Rauh M, et al. Ethical and social risks of harm from Language Models [Internet]. arXiv; 2021. https://arxiv.org/abs/2112.04359
125. Abimbola S, Patel B, Peiris D, et al. The NASSS framework for ex post theorisation of technology-supported change in healthcare: worked example of the TORPEDO programme. *BMC Med*. 2019;17(1):233.
126. Greenhalgh T, Wherton J, Papoutsi C, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. *J Med Internet Res*. 2017;19(11):e367.
127. Cimino J, Friedmann B, Jackson K, Li J, Pevzner J, Wrenn J. Redesign of the Columbia University Infobutton Manager. *AMIA Annu Symp Proc*. 2007;2007:135-139.