Journal of the American Medical Informatics Association, 28(8), 2021, 1703–1711
doi: 10.1093/jamia/ocab077
Advance Access Publication Date: 6 May 2021
Research and Applications

Research and Applications

A neuro-symbolic method for understanding free-text medical evidence

Tian Kang,[1] Ali Turfah,[2] Jaehyun Kim,[1] Adler Perotte ,[1] and Chunhua Weng [1]

[1]Department of Biomedical Informatics, Columbia University, New York, USA, and [2]Department of Statistics, Columbia University, New York, USA

Corresponding Author: Chunhua Weng, PhD, Department of Biomedical Informatics, Columbia University, 622 W 168 Street, PH-20, Room 407, New York, NY 10032, USA; chunhua@columbia.edu

Received 1 February 2021; Revised 18 March 2021; Editorial Decision 9 April 2021; Accepted 9 April 2021

ABSTRACT

Objective: We introduce Medical evidence Dependency (*MD*)–informed attention, a novel neuro-symbolic model for understanding free-text clinical trial publications with generalizability and interpretability.
Materials and Methods: We trained one head in the multi-head self-attention model to attend to the Medical evidence Ddependency (*MD*) and to pass linguistic and domain knowledge on to later layers (*MD informed).* This *MD-informed* attention model was integrated into BioBERT and tested on 2 public machine reading comprehension benchmarks for clinical trial publications: Evidence Inference 2.0 and PubMedQA. We also curated a small set of recently published articles reporting randomized controlled trials on COVID-19 (coronavirus disease 2019) following the Evidence Inference 2.0 guidelines to evaluate the model's robustness to unseen data.
Results: The integration of *MD-informed* attention head improves BioBERT substantially in both benchmark tasks—as large as an increase of +30% in the F1 score—and achieves the new state-of-the-art performance on the Evidence Inference 2.0. It achieves 84% and 82% in overall accuracy and F1 score, respectively, on the unseen COVID-19 data.
Conclusions: *MD-informed* attention empowers neural reading comprehension models with interpretability and generalizability via reusable domain knowledge. Its compositionality can benefit any transformer-based architecture for machine reading comprehension of free-text medical evidence.

Key words: natural language understanding, machine reading comprehension, transformer, medical evidence computing

INTRODUCTION

Evidence-based medicine (EBM) calls for the incorporation of the best available medical evidence from systematic research into clinical decision making for principled patient care.[1] Much medical evidence is locked in free-text randomized control trial (RCT) publications.[2] As vast evidence bases such as PubMed grows exponentially and rapidly, evidence retrieval and appraisal become extremely difficult due to information overload.[3] It usually takes more than 30 minutes for a clinician to search for evidence needed to answer one clinical question encountered during patient care. In practice, however, their busy clinical routines can only spare less than 2 minutes for such laborious searches,[4] resulting in limited translation of evidence from research to practice. Therefore, it is imperative to develop scalable and automated medical evidence extraction and comprehension methods. Methods have been developed for evidence retrieval,[5–8] data elements extraction,[9–13] automated systematic review,[14–16] and clinical question answering (QA).[4,17–21] In this study, we focus on machine reading comprehension (MRC).

MRC is the technology that teaches a machine to read unstructured text, mimic the inference process of human readers, and then answer questions about it. Efficient comprehension and synthesis of medical evidence in the literature is no trivial task—even for medical

experts. An example abstract from[22] and a related clinical question are shown in Figure 1. The abstract reports an interventional study that assessed the effectiveness of respiratory rehabilitation for elderly coronavirus disease 2019 (COVID-19) patients. The clinical question asks whether respiratory rehabilitation can significantly improve 2 outcomes: anxiety and depression. We highlight the text in which inference is made to answer "yes" for anxiety and "no" for depression based on the following rationale: (1) the answer comes from the conclusion about the interventional group (not control); (2) anxiety and depression are measured by Self-Rating Depression Scale (SDS) and Self Rating Anxiety Scale (SAS) scores; and (3) in the interventional group, both scores decreased but only the difference in SAS is statistically significant.

Early QA systems for improving patientcare relied heavily on biomedical ontologies—such as the UMLS Metathesaurus[23]—and lexico-syntactic patterns to extract biomedical concepts as candidate answers, followed by a scoring function (eg, TF-IDF, LexRank) for answer ranking.[17–19] Generating answers from automatically constructed knowledge graphs is another technique for answering clinical questions. A factorized Markov network was used to construct a clinical knowledge base from clinical notes.[20] Recent breakthroughs of pretrained language models such as ELMo[24] and BERT[25] show significant performance improvement on multiple tasks including QA and MRC. Neural approaches in the biomedical domain have benefited from these advances. It is common practice in biomedical MRC to introduce attention variants from general NLP applications, followed by domain adaptation by fine tuning on a biomedical corpus. For example, Du et al[26] used biomedical word embedding and a hierarchical multilayer transfer learning model with a co-attention mechanism by Xiong et al[27] and Wiese et al[28] to concatenate general word embeddings with biomedical embeddings and adopt FastQA[29] in the attention layer to develop an extractive QA system. While most of the prior work in the biomedical domain only incorporates biomedical concepts through concept embeddings or general transfer learning from large biomedical corpora, we dedicate our efforts to design an efficient neural approach to make use of relevant domain knowledge and improve the model's reasoning capability over medical evidence text.

All previous work can be categorized as either symbolic or statistical. The idea behind a symbolic approach is to teach machines to understand language in the same top-down manner that humans do—learning and using rules as well as symbolic representations of knowledge—which is explainable and offers good performance in reasoning tasks as expert systems do. However, this technique relies heavily on human-driven knowledge engineering and has had limited success in understanding and deciphering contextual information.[30] Recent state-of-the-art results in natural language processing (NLP) have been achieved predominantly by statistical methods, particularly the deep learning models. These bottom-up data-driven approaches have shown significant advantages in learning latent and sophisticated representations probabilistically. However, their reasoning capabilities are still rather limited when compared with symbolic AI.[31] In addition, the lack of transparency and the requirement for extensive training data to fit these models become 2 severe draw-

ABSTRACT

Background Different degrees of disorders are reported in respiratory function, …

Objective To investigate the effects of 6-week respiratory rehabilitation training on respiratory function, QoL, mobility and psychological function in elderly patients with COVID-19.

Methods This paper reported the findings of an observational, prospective, quasi-experimental study, which totally recruited 72 participants, of which 36 patients underwent respiratory rehabilitation and the rest without any rehabilitation intervention. The following outcomes were measured: pulmonary function tests including plethysmography and diffusing lung capacity for carbon monoxide (DLCO), functional tests (6-min walk distance test), Quality of life (QoL) assessments (SF-36 scores), activities of daily living (Functional Independence Measure, FIM scores), and mental status tests (SAS anxiety and SDS depression scores).

Results After 6 weeks of respiratory rehabilitation in the intervention group, there disclosed significant differences in FEV1(L), FVC(L), FEV1/FVC%, DLCO% and 6-min walk test. The SF-36 scores, in 8 dimensions, were statistically significant within the intervention group and between the two groups. SAS and SDS scores in the intervention group decreased after the intervention, but only anxiety had significant statistical significance within and between the two groups.

Conclusions Six-week respiratory rehabilitation can improve respiratory function, QoL and anxiety of elderly patients with COVID-19, but it has little significant improvement on depression in the elderly.

CLINICAL QUESTION

Can respiratory rehabilitation significantly improve the anxiety and depression of the elderly patient with COVID-19?

ANSWER

Respiratory rehabilitation can significantly improve anxiety of elderly patients with COVID-19, but not depression.

Figure 1. An example clinical question and answer for a study formulated from Liu et al.[22]

backs. These challenges are exacerbated in the healthcare domain by the lack of trust in machines among clinicians. Other challenges facing text comprehension for medical literature have also been identified, such as that (1) models suffer from lengthy text and the long distance dependencies throughout the articles[32,33] and (2) the complexities in clinical studies limit the neural models' ability to efficiently incorporate domain knowledge and develop clear intuitions around strong patterns denoting complex concepts.[33] The attention mechanism[34] and its variants conditioned on question text have been applied to such problems and only achieve modest predictive gains.[32]

Therefore, in this study, we aim to design a neuro-symbolic MRC model to understand free-text medical evidence (eg, RCT publications) by leveraging both the high capacity of neural networks as well as the expressiveness of symbolic methods. The traditional technique used to combine the 2 approaches is multitask learning with hard parameter sharing between symbolic knowledge representations for medical evidence and neural reading comprehension models. Their potential shortcomings include overfitting and dependency on the quality of the parser. By synergizing neural and symbolic methods, our goal is to improve the interpretability, reasoning ability, and task generalizability of neural networks by adding reusable domain knowledge. Our contributions are 3-fold: (1) we propose a symbolic representation, called medical evidence dependency (MD), to represent the compositional elements of medical evidence; (2) we propose a novel attention mechanism, *MD-informed* attention, which provides compositional submodel for any Transformer-based language models and is able to pass linguistic and domain knowledge onto later layers; and (3) we integrate *MD-informed* attention into BioBERT to evaluate the model's ability to understand and synthesize unstructured medical evidence on 2 public benchmarks. *MD-informed* attention substantially improves BioBERT performance and achieves new state-of-the-art performance.

MATERIALS AND METHODS

Models

Medical evidence dependency

First, we define a simple and computable representation for medical evidence, which represents compositional evidence elements and relations among them. A medical evidence element is an atomic entity in a finding. We adopt the PICO framework developed for formulating clinical questions to retrieve evidence from literature[1] to define 4 types elements (P, I, C, and O):

Population the characteristics of the study population

Intervention the primary intervention considered

Comparator comparison for the intervention

Outcome the anticipated measures, improvements, or effects

Additionally, we define 2 new attribute classes to represent necessary context: observation (quantitative or qualitative results with respect to an outcome measure) and count (the count of participants observed to have the same result for an outcome measure) Then we define the directional relationships between a pair of evidence elements, called MD, with one element being the governor and the other being the dependent. The directions are fixed to `Intervention(Comparator)`→`Observation` →`Outcome`. Example MD-structured text is shown in Figure 2. Using the elements and the dependency, we can construct a "medical evidence proposition," a compositional unit of medical evidence. In the example, 2 medical evidence propositions are formulated from the extracted intervention, outcome, and observation elements. Both represent an observed clinical fact with respect to the outcomes (cardiac index became higher; vascular resistance was decreased) after the intervention is applied.

MD-informed self-attention

Most of the current neural NLP models use the Transformer introduced by Vaswani et al[35] as their backbone, such as BERT,[25] XL-Net,[36] and GPT-2.[37] The multihead attention mechanism is used to capture global interactions across the text in multiple "representation subspaces." Such an architecture offers flexibility and potential to teach the model to learn a "subspace" in the medical domain. The conventional neural attention mechanism is unsupervised when learning to attend to relevant inputs. In this study, we train the self-attention to attend to the MD as a mechanism for passing both linguistic and domain knowledge to subsequent layers, and we hypothesize that our model can better attend to relevant text and improve reasoning capability over long-distance evidence for clinical questions (Figure 3).

Inspired by Strubell et al,[38] in which syntactic dependency is integrated into attention for semantic role labeling, we design an MD matrix, a specialized adjacency matrix for the directed graph induced by MDs from text. The MD matrix, like self-attention, captures global dependencies within text segments (Figure 4). When a MD is identified between a pair of terms, 1 is assigned to the corresponding slot in the matrix; otherwise, 0 is assigned. In addition, because intervention elements are in the top hierarchy in the MDs among all others, we define every recognized intervention element as dependent on itself and assign 1 to the corresponding slot in the matrix. Figure 4 shows an MD matrix for the example text.

Conventional self-attention adopted the scaled dot-product attention, in which the attention is weighted sum of the values (Value). The weight assigned to each value is determined by the dot-product of the query (Query, ie, the information we are looking for) with all the keys (Key [ie, the relevance to the query]). The detailed explanation is given in Vaswani et al.[35]

$$Attention(Query,\ Key,\ Value) = Softmax\left(\frac{Query \cdot Key}{\sqrt{d_k}}\right) \cdot Value$$

Here, we modify the weights (scaled dot-product of Query and Key) to make it relevant to medical evidence. In one self-attention head from the Transformer, we drop in the MD matrix to replace the scaled attention score generated from the dot product of Query and Key (Figure 4), and take its Softmax to compute new weights. Then by computing a new weighted sum of Value, we get a context representation Z_{MD} specialized to attend to medical evidence:

$$Attention(MD\ Matrix,\ Value) = Softmax(MD\ Matrix) \cdot Value$$

Figure 3 depicts the overall architecture of one multihead self-attention module with *MD-informed* attention. The attention heads within black boxes in Figure 3 contain the *MD-informed* attention values. We leave the other attention heads in multihead self-attention as default to learn their own attention representation Z from their Query, Value, and Key, and concatenate the learned Z_{MD} from *MD-informed* attention head with the rest of the conventional context layers to obtain Z as the final output of one attention module in the Transformer (the top layers in Figure 3). By introducing *MD-informed* attention, the neural reading comprehension model

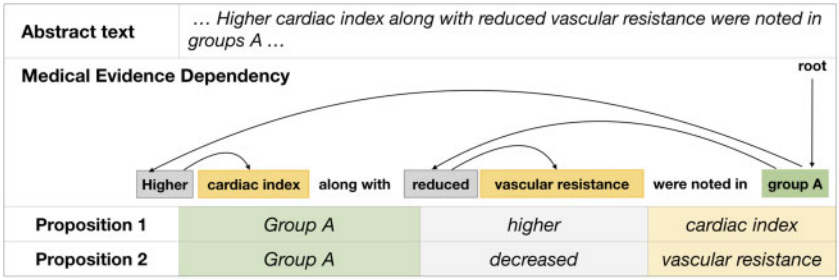


Figure 2. Example medical evidence dependency (MD) for unstructured medical evidence. Two medical evidence propositions are extracted from this example sentence.

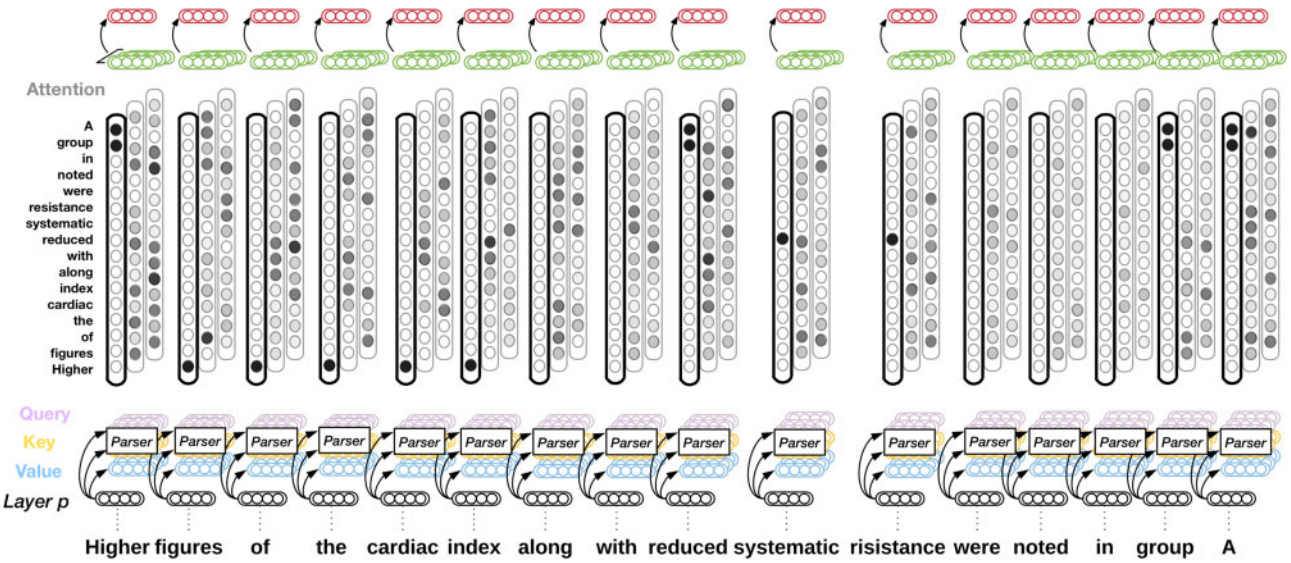


Figure 3. Overall architecture for multihead self-attention with medical evidence dependency (MD)–informed attention. It visualizes 3 heads (vertical dot lines) in the figure, but it is worth noting that there are usually 12 heads in BERT base architecture. The head with bold border represent the *MD-informed* attention head. The tokens that relate in MD are assigned higher weights.

can make efficient end-to-end use of domain knowledge. In addition, because MD is global, the model can efficiently capture and reason over long-distance evidence relations.

MD parser

The parser extends our previous work and annotated dataset.[10] We module the task of extracting Medical Evidence Elements as named entity recognition, and parsing Medical evidence Dependency as relation extraction. Both named entity recognition and relation extraction models are trained by modifying the last layer and fine-tuning a biomedical version of BERT[39] on the dataset. It achieves a micro-F1 score of 0.72 for 5-class named entity recognition and 0.92 for extracting MDs among PICO elements (details in Supplementary Appendix). We apply this parser to construct the MD matrix and *MD-informed* attention head. It is worth noting that this can be replaced when a more advanced method or tool is available.

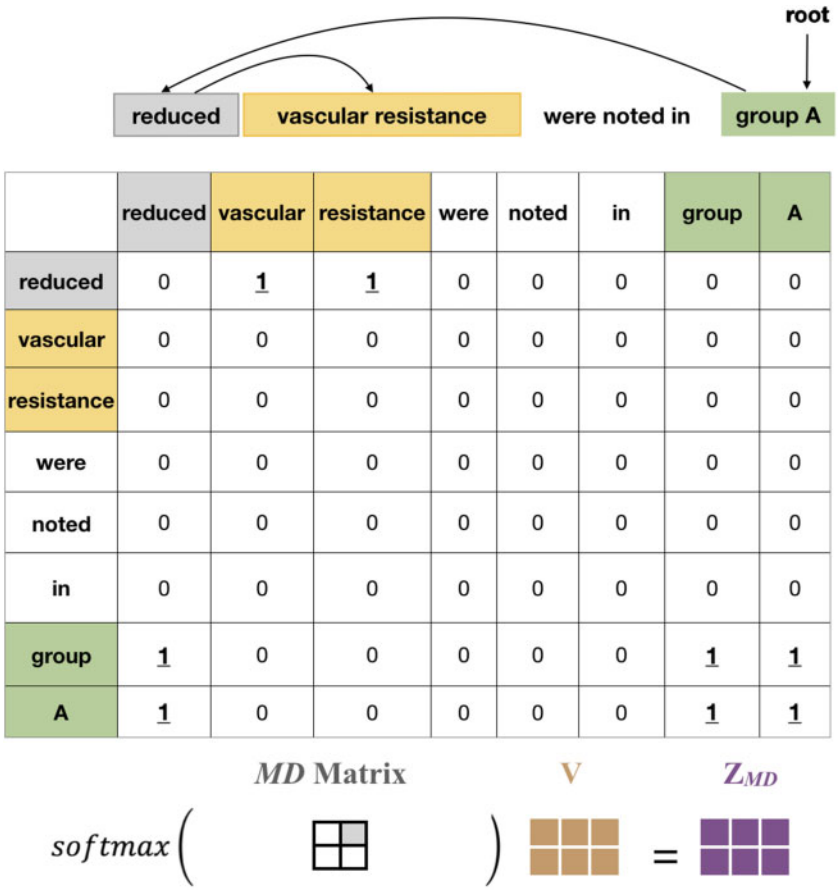
EXPERIMENTS

MD-informed self-attention is compatible with any Transformer-based model and can support various natural language understanding tasks on unstructured medical literature. In this study, we evaluate its effectiveness under the BioBERT architecture and present results on 2 shared benchmark datasets for text comprehension for the medical literature, Evidence Inference 2.0 and PubMedQA.

Benchmark datasets

Evidence Inference 2.0[2]***:*** Evidence inference and synthesis is a key task in practicing EBM. Entries in this dataset consist of an intervention (eg, chemotherapy), a comparator (eg, surgery), and an outcome (eg, 5-year survival rate of operable cancers), along with an associated article. The task is to infer the comparative performance of the 2 treatments with respect to the outcome based on the article to tell if there was a significant increase, a significant decrease, or no significant change between the intervention and comparator. The prompts labeled as invalid or whose answers cannot be found in the article abstract are filtered out before training.

PubMedQA[40]***:*** This is a machine reading comprehension dataset for biomedical research questions. The task is, given a question and a relevant piece of medical literature (a context), predict an answer of yes, no, or maybe. The questions in the dataset are constructed from the titles of PubMed articles, while the context is a structured abstract with the Conclusion sentences omitted. No filtering is done on this dataset. During preprocessing, each question-context pair is separated by the special token `[SEP]`. Particularly in Evidence Inference 2.0, questions are given in terms of "prompts," each specifying



	reduced	vascular	resistance	were	noted	in	group	A
reduced	0	1	1	0	0	0	0	0
vascular	0	0	0	0	0	0	0	0
resistance	0	0	0	0	0	0	0	0
were	0	0	0	0	0	0	0	0
noted	0	0	0	0	0	0	0	0
in	0	0	0	0	0	0	0	0
group	1	0	0	0	0	0	1	1
A	1	0	0	0	0	0	1	1

Figure 4. Medical evidence dependency (*MD*) matrix and *MD-informed* self-attention function, ZMD. Conventionally, the attention function Z is learned from Query, Value, and Key: Z = Softmax(Q · KT) V. To adapt it to medical evidence, we drop in the MD matrix to replace the scaled attention score generated from the dot product of Query and Key, and take its Softmax.

an Intervention, a Comparator, and an Outcome. Every prompt and context are processed as:

[O] Outcome [I] Intervention [C] Comparator [SEP] abstract text

Basic statistics and examples for 2 benchmark datasets after preparation are provided in Figure 5.

Experiments of *MD-informed* attention

We tested *MD-informed* attention on 2 benchmarks. When constructing the MD matrix for the prompts or questions, the questions from PubMedQA are processed as the same way as the abstracts: first we identify the Medical evidence Dependencies and then construct the MD matrix accordingly. For Evidence Inference prompts, because the element types are given, we assign 1 to all pairs of words in Intervention and Comparator. Special tokens like `[O]`, `[SEP]` are left as 0 in the matrix. The model is then trained to select one correct answer from multiple choice options (Evidence Inference 2.0: "*significantly increased*," "*significantly decreased*," and "*no significant difference*"; PubMedQA: "*yes*," "*no*," and "*maybe*").

MD-informed attention is integrated into BioBERT and pretrained on biomedical corpus and SQUAD 2.0 for biomedical QA,[21] by replacing one conventional Self-Attention head in the Transformer Encoder (henceforth, such systems referred as BioBERT-MDAtt). To evaluate the robustness of *MD-informed* attention, we also apply an attention mask on this attention head to randomly remove part of learned dependencies (BioBERT-MDAtt-masked), by setting each pair of words in MD matrix assigned 1 to 0 with a probability of p.

Baseline models

We compare BioBERT-MDAtt results to the 2 baselines on both tasks.

State-of-the-art performance: For the Evidence Inference 2.0 dataset, we compare our results to the best performance reported in,[2] and the top system on the leaderboard. In,[2] the best model predicts answers using a BERT to BERT, 2-stage pipeline. A variant of RoBERTa[41] pretrained over scientific corpora serves as the base model. The first BERT identifies evidence bearing sentences within an article for given PICO elements. The second then classifies the answer using the evidence extracted from the first stage. In the up-to-date leaderboard, the top system applies a similar strategy and outperforms the original system by 2%. The state-of-the-art system for PubMedQA, reported in Jin et al[40]—which is also the top performer on their leaderboard—adopts a multiphase fine-tuning of BioBERT on both labeled and unlabeled data collections. In our experiments on PubMedQA, only labeled QA pairs are used.

BioBERT for QA: Additionally, we implement BioBERT for biomedical QA[21,42] as another strong baseline, with all attention heads left on their own to learn. The last layer is modified to adapt to our

Benchmark	# Train	# Dev	# Test	Example instance
PubMedQA (PQA-L set)	450	50	500	**Question** *Do preoperative statins reduce atrial fibrillation after coronary artery bypass grafting?* **Abstract** *…… The overall incidence of postoperative AF was 26%. Postoperative AF was significantly lower in the Statin group compared with the Non-statin group (16% versus 33%, p=0.005). Multivariate analysis demonstrated that independent predictors of AF……* **Answer** yes
Evidence Inference 2.0	4993	625	615	**Prompt** *With respect to Cephalea relief at 24 hours, what is reported difference between patients receiving Rizatriptan and those receiving Ibuprofen and placebo?* **Abstract** *… Efficacy was assessed by headache relief, and headache freedom at 2 h and 24 h. Two-hour headache relief, was noted in 73% in rizatriptan, 53.8% in ibuprofen and 8% in placebo groups. Headache freedom was achieved in 37.7% in rizatriptan, 30.8% in ibuprofen and 2% in placebo groups. Rizatriptan was superior to ibuprofen and placebo in relieving headache at 2 h but not at 24 h. Side effects were noted in 9 patients in rizatriptan, 8 in ibuprofen and 3 in placebo, all of which were nonsignificant. Rizatriptan and ibuprofen are superior to placebo. Rizatriptan is superior to ibuprofen in relieving headache, associated symptoms and functional disability.…* **Answer** No significant difference

Figure 5. Dataset statistics and example instance for the 2 benchmarks. Both are formulated as a machine reading comprehension task, in which the model is trained to predict the answer by giving a question (in Evidence Inference 2.0's case, a prompt) and a related abstract from randomized controlled trial reports (context). Partial relevant context to answer the example question is highlighted in the figure. It is worth noting that, for PubMedQA dataset, we only use the part that is labeled with gold standard answer (PQA-L set).

question type and fine-tuned on both datasets (referred as BioBERT).

Given that the information necessary to answer the question might be scattered throughout the abstract, we fix a large number, 384, as the maximum sequence length while training all the models. All BERT models deployed in this study are BERT-Base, with 12 attention heads. If *MD-informed* attention is applied, one head will be replaced. We fine-tuned all other underlying parameters. We trained all models using the Adam optimizer[43] with a learning rate 2e-5. All systems are implemented *in* TensorFlow 1.14.0 and trained on 4x NVIDIA GeForce RTX 2080 Ti GPUs.

RESULTS AND DISCUSSION

Evidence inference 2.0

Table 1 lists the main results on the Evidence Inference 2.0 test set. Our proposed BioBERT-MDAtt model achieves the new state of the art (macro-F1: 0.843, micro-F1: 0.844, accuracy: 0.84), over 4% absolute macro-F1 higher than previously reported best models.[2] The baseline model that fine-tunes on BioBERT achieves 0.55 macro-F1, comparable to the reported performance (0.51 macro-F1) of the BERT Pipeline without conditioning on recognized PICO elements in DeYoung et al.[2] We report per-class performance from BioBERT and BioBERT-MDAtt in Table 2. Simple addition of *MD-informed* attention brings substantial improvement—almost a +0.30 increase in macro-F1 score and accuracy.

Table 1 shows the performance of the model with $P = .4$. The performance drops slightly compared with the model with the attention mask, but *MD-informed* attention still outperforms the previous state of the art. Compared with the prior models, in which the final label is predicted based upon evidence sentence extracted in the prior stages, BioBERT-MDAtt is conducted as a completely end-to-end pipeline and leverages knowledge in a domain-agnostic way rather than running the risk of overfitting to the training data.

Table 1. Accuracy, macro-F1 score, precision, and recall on Evidence Inference 2.0 test set.

Model	Accuracy	F1 Score	Precision	Recall
DeYoung et al[2]	/	0.780	0.784	0.777
Leaderboard	/	0.797	0.796	0.797
BioBERT	0.56	0.551	0.551	0.551
+ MDAtt	**0.84**	**0.843**	**0.850**	**0.841**
+ MDAtt-masked	0.82	0.819	0.823	0.817

The first 2 rows are retrieved from the original publication and the leaderboard for the benchmark.

Table 2. Per-class and overall performance by BioBERT and BioBERT-MDAtt on 2 benchmarks.

	Class	Precision		Recall		F1 score		Support
		BioBERT	+MDAtt	BioBERT	+MDAtt	BioBERT	+MDAtt	
Evidence Inference 2.0	Significantly increased	0.63	0.80	0.64	0.88	0.63	0.84	227
	No significant change	0.52	0.88	0.53	0.87	0.52	0.87	208
	Significantly decreased	0.53	0.86	0.51	0.77	0.52	0.82	180
	Macro-average	0.55	0.85	0.55	0.84	0.55	0.84	615
	Micro-average	0.55	0.84	0.55	0.84	0.55	0.84	615
PubMedQA	Yes	0.58	0.67	0.85	0.73	0.69	0.70	276
	Maybe	0.00	0.18	0.00	0.11	0.00	0.14	55
	No	0.40	0.57	0.22	0.57	0.29	0.57	169
	Macro-average	0.32	0.48	0.35	0.48	0.31	0.48	500
	Micro-average	0.53	0.61	0.57	0.61	0.53	0.61	500

PubMedQA

There are multiple data collections in PubMedQA, and only PQA-L(abeled) includes human-curated answers. However, it is an extremely low resource setting in which there are 1000 abstracts and only 450 training question-answer pairs in each fold of cross-validation. All evaluations in Tables 2 and 3 are carried out on the PQA-L test set of 500 QA pairs by 10-fold cross validation. Two systems from Jin et al[40] are compared against ours. The multiphase system in PubMedQA paper achieves the state-of-the-art performance by multiphase fine-tuning BioBERT, first on a large unlabeled corpus and then on PQA-L (over 200 000 abstracts in total). The Final Phase system in Table 3 is trained by fine-tuning BioBERT only on PQA-L (1000 abstracts). To evaluate the effectiveness of the *MD-informed* attention, we also only use PQA-L data to train our MRC models in this study, thus our models are comparable to the Final Phase Only model.

Our baseline model, fine-tuning BioBERT on PQA-L, achieves comparable results to Final Phase Only system from the PubMedQA article. The effects of incorporating the *MD-informed* attention head into BioBERT are reported in Table 2. The BioBERT-MDAtt model achieves near state-of-the-art performance with substantially less data (macro-F1: SOTA 0.527 using 200 000 abstracts vs BioBERT-MDAtt 0.482 on 1000 abstracts) and showed considerable improvement upon the counterpart models (+0.17 in macro-F1 to the BioBERT baseline, and +0.19 to Final Phase Only model). Of note, the strategy adopted in the state-of-the-art (SOTA) system does not contradict with ours. It is easy to combine the two (ie, training *MD-informed* attention on PQA-L data after multiphase fine-tuning on large unlabeled corpus) and benefit from both strategies. This combination theoretically can achieve better results than individual approaches. Additionally, we notice both models have low performance for this class given the inherent ambiguity of "maybe" class (Table 2). Consistent with what we observe in the Evidence Inference 2.0 task, when masking is applied at $P = .4$ the performance drops slightly, but the addition of *MD-informed* attention head still results in a significant improvement in the model's performance. The results on PubMedQA task show that, by applying neuro-symbolic approach, the model can generalize over tasks via reusable knowledge and achieve better results with less data. We believe that our model has great potential to excel when a larger dataset is available.

Table 3. Macro-averaged performance from 10-fold cross-validation on PubMedQA test set

Model	Accuracy	F1 Score	Precision	Recall
Jin et al. (2019)[40]				
Multiphase (state of the art)	0.68	0.527	/	/
Final phase only	0.57	0.287	/	/
BioBERT	0.53	0.311	0.315	0.34
+ MDAtt	0.61	0.482	0.482	0.483
+ MDAtt-masked	0.60	0.463	0.469	0.463

For both tasks, the evaluations in Table 2 reveal that, replacing one conventional attention head with *MD-informed* attention in BioBERT results in extensive improvement in all measures. The *MD-informed* attention helps BioBERT further generalize over different tasks by reusable domain knowledge. More importantly, this improvement is understandable via human-readable symbolic form introduced by Medical evidence Dependency. In addition, because *MD-informed* attention is adaptable to any Transformer-based model (ie, most of the state-of-the-art language models), it provides a beneficial feature as being compositional and easy to be integrated. Therefore, *MD-informed* attention can serve as a reusable submodel to benefit any Transformer-based architecture and improve their abilities in understanding free-text medical evidence.

COVID-19 clinical trials case study

To further evaluate the robustness of *MD-informed* attention, we curate a small set of recently published PubMed abstracts reporting clinical trials on COVID-19. We selected this disease domain for evaluation because the studies in this domain have only started to accumulate recently, which provides us unseen examples for both the MD parser and the MRC model. Following the annotation guidelines from Evidence Inference 2.0, we create 50 "prompt-abstract" pairs from 10 abstracts that report RCTs of COVID-19 and make it available in the Supplementary Appendix. BioBERT-MDAtt trained on Evidence Inference 2.0 (performance reported in Tables 1 and 2) is applied to predict the 50 pairs.

We evaluate the model from 3 aspects: (1) performance on unseen data, (2) reasoning capabilities over variance of the expressions for `Intervention/Comparator/Outcome`, and (3) reasoning capabilities over long-distance evidence relationships. To do so, while creating prompt-abstract pairs, we intentionally replicate the original pair and replace elements in the prompts with their variants occurring in the other sections in abstract—a model with good reasoning capability should predict the same results for the pairs. For instance, consider the 2 pairs of prompts created from the article shown in Figure 1:

Table 4. Per-class and overall performance on new COVID-19 dataset by applying BioBERT-MDAtt model trained on Evidence Inference 2.0

	Accuracy	F1 Score	Precision	Recall	Support
Significantly decreased	0.67	0.74	0.83	0.67	15
No difference	0.92	0.89	0.86	0.92	26
Significantly increased	0.89	0.84	0.80	0.89	9
Macro average	0.84	0.82	0.83	0.83	50

COVID-19: coronavirus disease 2019.

`[O]` *anxiety* `[I]` *respiratory rehabilitation* `[C]` *without any rehabilitation intervention*

`[O]` *SAS score* `[I]` *respiratory rehabilitation* `[C]` *without any rehabilitation intervention*

The 2 are asking the same question: if the intervention has significant effect on anxiety compared with the comparator, which we should infer from the abstract, then it should significantly affect the SAS score (which is used to quantify anxieties). We report the BioBERT-MDAtt model performance in Table 4. Overall, even though this is an unseen dataset for both the parser and MRC model, the F1 score only drops slightly compared with original evaluation on the Evidence Inference 2.0 test set (from 0.84 to 0.82). A total of 42 of 50 pairs are answered correctly, indicating that our proposed model is robust. From examining the detailed results, we find that the model can answer both variants correctly for the created prompt pairs. The most common error that it makes is to misclassify "no significant difference" as 1 of other 2 labels. For example, from the example in Figure 1, "SAS and SDS scores in the intervention group decreased after the intervention, but only anxiety had significant statistical significance within and between the 2 groups," the model misclassified "depression" as significantly decreased instead of correctly reasoning over the adversative transition.

By visualizing the *MD-informed* attention head for the example text just mentioned (Figure 6), the isolated attention is visualized, showing connecting from the word "score" to all the words or tokens in a separated sentence that generates separated medical evidence propositions. The *MD-informed* attention head learns to highlight the relevant evidence components like "anxiety," "statistical significance" and "groups," and the highest weight comes the pair "score" to "anxiety," congruent with the facts that they both belong to outcome class and "(SAS) score" is the quantified measure for "anxiety." This shows that *MD-informed* attention is able to capture clinically meaningful or understandable interactions across different medical evidence propositions, instead of being a "black box" for practitioners. In future work, we would like to incorporate *MD-informed* attention into more advanced models to further test its effectiveness.

CONCLUSIONS

In this study, we present and evaluate a novel attention mechanism, *MD-informed* self-attention, for understanding and reasoning over free-text medical evidence such as RCT publications. By integrating *MD-informed* self-attention into BioBERT, and evaluating on 2 benchmarking tasks, we gain substantial improvement over BioBERT with the conventional multihead attention. We also outperform the prior state of the art on one task, and achieve near state-of-the-art performance with considerably less data on the other. By synergizing neural and symbolic methods, we introduce reusable knowledge and empower existing neural reading comprehension models with better understandability, reasoning ability, and task generalizability. In addition, because *MD-informed* attention is adaptable to any Transformer-based model (ie, most of the state-of-the-art language models), its compositionality is a beneficial feature to any Transformer-based architecture and can improve their abilities in understanding free-text medical evidence.

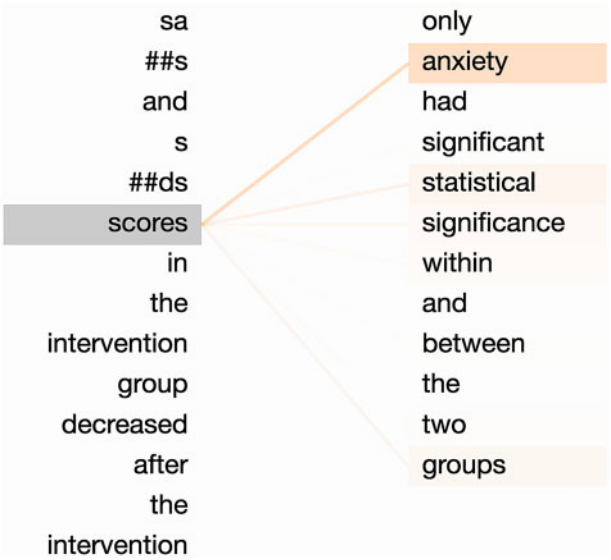


Figure 6. Medical evidence dependency (MD)–informed attention head visualization. The weights for the word "scores" learned from the *MD-informed* head is visualized between 2 sentences.

FUNDING

This work was supported by 5R01LM009886-11 (Bridging the semantic gap between research eligibility criteria and clinical data; PI: CW).

AUTHOR CONTRIBUTIONS

TK designed and carried out the experiments and drafted the manuscript. AT participated in the study design, experiments, and manuscript writing. JK participated in the data generation and reviewed the manuscript. AP participated in the study design and manuscript writing. CW supervised the research and participated in study design and manuscript writing.

SUPPLEMENTARY MATERIAL

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

CONFLICT OF INTEREST STATEMENT

None.

DATA AVAILABILITY STATEMENT

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

REFERENCES

1. Sackett DL. Evidence-based medicine. *Semin Perinatol* 1997; 21 (1): 3–5.
2. DeYoung J, Lehman E, Nye B, et al. Evidence inference 2.0: more data, better models. arXiv, doi: https://arxiv.org/abs/04177.2005, 14 May 2020, preprint: not peer reviewed.
3. Goldstein A, Venker E, Weng C. Evidence appraisal: a scoping review, conceptual framework, and research agenda. *J Am Med Inform Assoc* 2017; 24 (6): 1192–203.
4. Ely JW, Osheroff JA, Ebell MH, *et al.* Analysis of questions asked by family doctors regarding patient care. *BMJ* 1999; 319 (7206): 358–61.
5. Sim I, Owens DK, Lavori PW, *et al.* Electronic trial banks: a complementary method for reporting randomized trials. *Med Decis Making* 2000; 20 (4): 440–50.
6. Sim I, Detmer DE. Beyond trial registration: a global trial bank for clinical trial reporting. *PLoS Med* 2005; 2 (11): e365.
7. Verspoor K, Suster S, Otmakhova Y, *et al.* COVID-SEE: Scientific Evidence Explorer for COVID-19 related research. arXiv, doi: https://arxiv.org/abs/07880.2008, 18 Aug 2020, preprint: not peer reviewed.
8. Marshall IJ, Nye B, Kuiper J, *et al.* Trialstreamer: A living, automatically updated database of clinical trial reports. *J Am Med Inform Assoc* 2020; 27 (12): 1903–12.
9. Nye B, Li JJ, Patel R, *et al.* A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. *Proc Conf Assoc Comput Linguist Meet* 2018; 2018: 197–207.
10. Kang T, Zou S, Weng C. Pretraining to recognize piCO elements from randomized controlled trial literature. *Stud Health Technol Inform* 2019; 264: 188–92.
11. Amini I, Martinez D, Molla D. Overview of the ALTA 2012 shared task. In: *Proceedings of Australasian Language Technology Association Workshop*; 2012: 124–9.
12. Jin D, Szolovits P. Pico element detection in medical text via long short-term memory neural networks. In: *Proceedings of the BioNLP 2018 Workshop*; 2018: 67–75.
13. Kim S, Martinez D, Cavedon L, *et al.* Automatic classification of sentences to support evidence based medicine. *BMC Bioinform* 2011; 12 (suppl 2): S5.
14. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Syst Rev* 2019; 8 (1): 163.
15. Marshall IJ, Kuiper J, Wallace BC. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *J Am Med Inform Assoc* 2016; 23 (1): 193–201.
16. Marshall IJ, Kuiper J, Banner E, et al. Automating biomedical evidence synthesis: robot reviewer. *Proc Conf Assoc Comput Linguist Meet* 2017; 2017: 7–12.
17. Lee M, Cimino J, Zhu HR, *et al.* Beyond information retrieval—medical question answering. *AMIA Ann Symp Proc* 2006; 2006: 469–73.
18. 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 44th Annual Meeting of the Association for Computational Linguistics*; 2006: 841–8.
19. Schulze F, Schüler R, Draeger T, *et al.* HPI question answering system in BioASQ 2016. In: *Proceedings of the Fourth BioASQ Workshop*; 2016: 38–44.
20. Goodwin TR, Harabagiu SM. Medical question answering for clinical decision support. *Proc ACM Int Conf Inf Knowl Manag* 2016; 2016: 297–306.
21. Yoon W, Lee J, Kim D, Jeong M, et al. Pre-trained language model for biomedical question answering. arXiv, doi: https://arxiv.org/abs/1909.08229, 18 Sep 2019, preprint: not peer reviewed.
22. Liu K, Zhang W, Yang Y, Zhang J, Li Y, et al. , Respiratory rehabilitation in elderly patients with COVID-19: a randomized controlled study. *Complement Ther Clin Pract* 2020; 39: 101166.
23. Schuyler PL, Hole WT, Tuttle MS, et al. The UMLS Metathesaurus: representing different views of biomedical concepts. *Bull Med Libr Assoc* 1993; 81 (2): 217.
24. Peters ME, Neumann M, Iyyer M, *et al.* Deep contextualized word representations. arXiv, doi: https://arxiv.org/abs/1802.05365, 22 Mar 2018, preprint: not peer reviewed.
25. Devlin J, Chang M-W, Lee K, et al. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv, doi: https://arxiv.org/abs/1810.04805, 24 May 2019, preprint: not peer reviewed.
26. Du Y, Pei B, Zhao X, et al. Hierarchical multi-layer transfer learning model for biomedical question answering. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*; 2018.
27. Xiong C, Zhong V, Socher R. Dynamic coattention networks for question answering. arXiv, doi: https://arxiv.org/abs/1611.01604, 6 Mar 2018, preprint: not peer reviewed.
28. Wiese G, Weissenborn D, Neves M. Neural domain adaptation for biomedical question answering. arXiv, doi: https://arxiv.org/abs/1706.03610, 8 Jun 2017, preprint: not peer reviewed.
29. Weissenborn D, Wiese G, Seiffe L. A simple and efficient neural architecture for question answering. arXiv, doi: https://arxiv.org/abs/1703.04816, 8 Jun 2017, preprint: not peer reviewed.
30. Kursuncu U, Gaur M, Sheth A. Knowledge infused learning (K-IL): Towards deep incorporation of knowledge in deep learning. arXiv, doi: https://arxiv.org/abs/1912.00512, 29 Feb 2020, preprint: not peer reviewed.
31. Chen D, Fisch A, Weston J, et al. Reading wikipedia to answer open-domain questions. arXiv, doi: https://arxiv.org/abs/1704.00051, 4 Apr 2019, preprint: not peer reviewed.
32. Lehman E, DeYoung J, Barzilay R, et al. Inferring which medical treatments work from reports of clinical trials. arXiv, doi: https://arxiv.org/abs/1904.01606, 4 Apr 2019, preprint: not peer reviewed.
33. Oita M, Vani, K, Oezdemir-Zaech F.. Semantically corroborating neural attention for biomedical question answering. In: *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*; 2019: 670–85.
34. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv, doi: https://arxiv.org/abs/1409.0473, 19 May 2016, preprint: not peer reviewed.
35. Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. In: *NIPS '17: Proceedings of the 31st International Conference on Neural Information Processing Systems*; 2017: 6000–10.
36. Yang Z, Dai Z, Yang Y, et al. Xlnet: generalized autoregressive pretraining for language understanding. In: *NIPS '19: Proceedings of the 33rd International Conference on Neural Information Processing Systems*; 2019.
37. Radford A, Wu J, Child R, *et al.* Language models are unsupervised multitask learners. *OpenAI Blog* 2019; 1 (8): 9.
38. Strubell E, Verga P, Andor D, et al. Linguistically-informed self-attention for semantic role labeling. arXiv, doi: https://arxiv.org/abs/1804.08199, 12 Nov 2018, preprint: not peer reviewed.
39. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. arXiv, doi: https://arxiv.org/abs/1906.05474, 18 Jun 2019, preprint: not peer reviewed.
40. Jin Q, Dhingra B, Liu Z, et al. PubMedQA: a dataset for biomedical research question answering. arXiv, doi: https://arxiv.org/abs/1909.06146, 13 Sep 2019, preprint: not peer reviewed.
41. Liu Y, Ott M, Goyal N, *et al.* RoBERTa: a robustly optimized BERT pretraining approach. arXiv, doi: https://arxiv.org/abs/1907.11692, 26 Jul 2019, preprint: not peer reviewed.
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): 1234–40.
43. Kingma DP, Ba J. Adam: a method for stochastic optimization. arXiv, doi: https://arxiv.org/abs/1412.6980, 2 Dec 2014, preprint: not peer reviewed.