What Does the Evidence Say?
Models to Help Make Sense of the Biomedical Literature

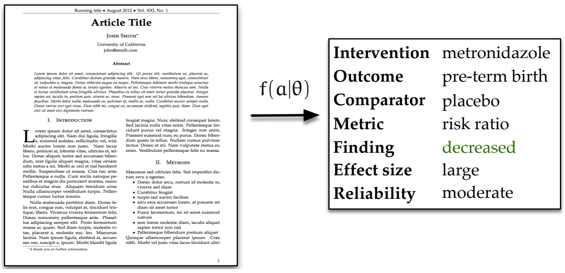
Byron C. Wallace
Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA
b.wallace@northeastern.edu

Abstract

Ideally decisions regarding medical treatments would be informed by the totality of the available evidence. The best evidence we currently have is in published natural language articles describing the conduct and results of clinical trials. Because these are unstructured, it is difficult for domain experts (e.g., physicians) to sort through and appraise the evidence pertaining to a given clinical question. Natural language technologies have the potential to improve access to the evidence via semi-automated processing of the biomedical literature. In this brief paper I highlight work on developing tasks, corpora, and models to support semi-automated evidence retrieval and extraction. The aim is to design models that can consume articles describing clinical trials and automatically extract from these key clinical variables and findings, and estimate their reliability. Completely automating 'machine reading' of evidence remains a distant aim given current technologies; the more immediate hope is to use such technologies to help domain experts access and make sense of unstructured biomedical evidence more efficiently, with the ultimate aim of improving patient care. Aside from their practical importance, these tasks pose core NLP challenges that directly motivate methodological innovation.

1 Introduction

Randomized Controlled Trials (RCTs) are at present the best tool we have to reliably measure the causal effects of alternative treatments. Unfortunately, results from RCTs are reported predominantly in unstructured (free-text) journal articles, which makes it onerous to sort through findings to assess interventions and ultimately make evidence-based decisions. This problem has been exacerbated by the rapid expansion of the biomedical evidence base. In 2012, about 75 articles describing clinical trials were published every day, on average [Bastian *et al.*, 2010]. At the time of this writing, more like 100 trial reports are published daily. Domain experts (e.g., physicians) cannot keep up with this torrent of unstructured evidence, hindering the practice of evidence-based care.



Intervention	metronidazole
Outcome	pre-term birth
Comparator	placebo
Metric	risk ratio
Finding	decreased
Effect size	large
Reliability	moderate

Figure 1: Schematic of a model that can map from an unstructured article describing an RCT to structured data codifying the evidence that it reports. We envision such models primarily being used to *help*, not *replace*, domain experts.

Researchers in ML and NLP can play an important role in addressing this issue by designing models that aid those trying to navigate and make sense of the evidence base. For example, interactive information retrieval and question answering systems may facilitate faster identification of relevant evidence. Realizing such aims motivates models for deep processing of trial reports to extract structured representations of reported findings [Blake and Lucic, 2015; Lehman *et al.*, 2019]. Figure 1 schematizes one such potential model f parameterized by θ that consumes an article a describing an RCT and yields a structured 'evidence frame' codifying the results that it reports. Figure 2 provides a higher-level view encompassing the entire process, from retrieval to extraction. Critically, the ML components of this system need not be perfect in order to meaningfully aid domain experts in searching and synthesizing evidence.

In addition to being an important practical application, evidence extraction poses a compelling set of challenges for NLP that push the boundaries of existing language technologies. For example, systems must process lengthy, technical articles to extract clinical entities (e.g., PICO elements) and infer the reported results concerning these. Because trials will often report results comparing multiple interventions across several different outcomes, realizing this aim will require some degree of reasoning. Furthermore, for most sub-tasks of interest, models must be able to provide *rationales* supporting decisions, which is an important general consideration for NLP [Wang *et al.*, 2019]. Finally, supervision is limited in this domain, as experts are overburdened and expensive. This motivates the need for efficient train-

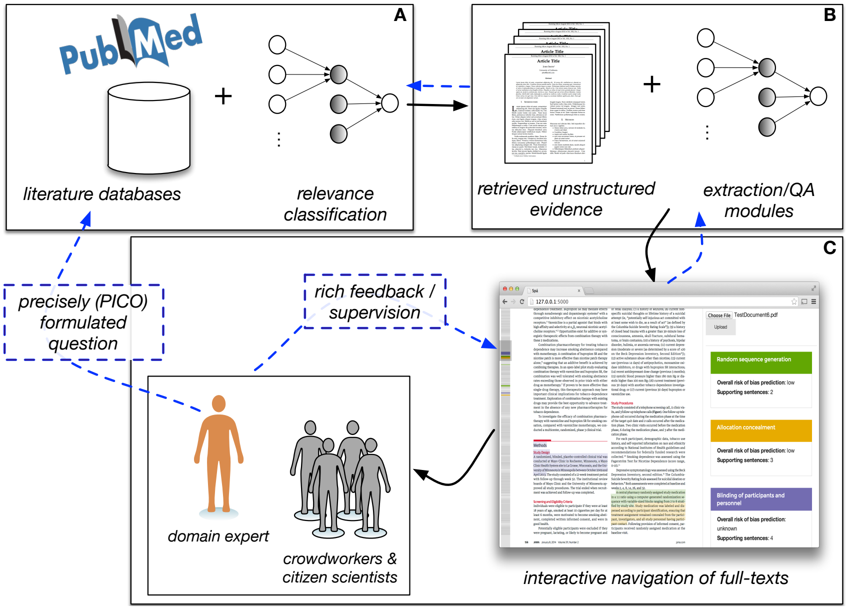

A
PubMed
literature databases
+
relevance classification
B
retrieved unstructured evidence
+
extraction/QA modules
C
precisely (PICO) formulated question
rich feedback / supervision
domain expert
crowdworkers & citizen scientists
Choose File TestDocument6.pdf
Upload
Random sequence generation
Overall risk of bias prediction: low
Supporting sentences: 2
Allocation concealment
Overall risk of bias prediction: low
Supporting sentences: 3
Blinding of participants and personnel
Overall risk of bias prediction: unknown
Supporting sentences: 4
interactive navigation of full-texts

Figure 2: High-level overview of an envisioned semi-automated system for evidence retrieval and extraction. (**A**) An interactively trained classification/retrieval model facilitates rapid identification of trials relevant to a given clinical question, specified as a PICO frame. (**B**) Models next extract clinically salient information from relevant articles. (**C**) Domain experts then interact with the extracted evidence extracted by models and the underlying documents to find the data they are after; the idea is to make this process faster and less tedious.

ing regimes, including exploiting *distant* [Mintz *et al.*, 2009; Wallace *et al.*, 2016] and *active* [Settles, 2009; Wallace *et al.*, 2010a] supervision, and harnessing crowd and citizen-science worker annotations to mitigate the problem of limited domain expert availability [Wallace *et al.*, 2017].

In the remainder of this *Early Career Spotlight* paper I outline some of the key sub-tasks that must be performed by systems intended to automatically process and — to some degre — make sense of articles describing clinical trials, or at least help domain experts do so. I elaborate on the core ML and NLP challenges these tasks entail, which are situated at the intersection of information extraction and 'machine reading' [Peng *et al.*, 2017], and I discuss progress made on them so far. I also highlight recently developed publicly available resources (datasets) to support future work on these problems.

2 Finding Relevant Evidence

Evidence retrieval concerns finding all evidence that addresses a particular clinical question. Well-formed clinical questions typically specify — at a minimum — a Population, Intervention, Comparator, and Outcome of interest [Huang *et al.*, 2006]. These are collectively referred to as PICO elements. For example, one might be interested in the efficacy of *ACE inhibitors* (I) relative to *placebo* (C), in adult patients with type-2 diabetes (P), with respect to blood pressure measurements (O). Identifying evidence that aligns with a given PICO criteria typically entails searching databases of published literature such as PubMed.[1] To ensure comprehensive (unbiased) results, search strategies in the space tend to be systematic and recall-centric [Dickersin *et al.*, 1994].

Once a set of potentially relevant articles is retrieved, one must 'screen' these to identify those that are in fact relevant to one's clinical question. For comprehensive (high-recall) initial search strategies, this will involve screening out many irrelevant articles, imposing substantial burden on domain experts. Machine learning methods can play a key role in expediting this process by semi-automating screening [Cohen *et al.*, 2006; Wallace *et al.*, 2010b]. *Interactive* methods, in which human experts actively engage with the model to train it to identify studies pertinent to their specific clinical question, are a particularly natural fit here; empirical results suggest that methods can substantially reduce the workload involved in comprehensive retrieval of literature relevant to a particular clinical question [Wallace *et al.*, 2010a; Przybyła *et al.*, 2018; O'Mara-Eves *et al.*, 2015].

An alternative strategy to designing models that can automatically screen articles for pertinence to a given question is to instead design classifiers that infer more general characteristics of trials on the basis of papers describing them. Such models can then be combined to realize a specific search strategy. For example, models that can reliably identify reports of RCTs have been developed and validated [Cohen *et al.*, 2015; Marshall *et al.*, 2018]. Applying such models to an initial set of candidate articles retrieved using a highly-sensitive search strategy may be an efficient means of reducing workload.

3 PICO Tagging

As mentioned above, clinical questions typically specify PICO elements of trials, but these are not consistently available in a structured format for articles. This makes it difficult to facilitate faceted search (i.e., retrieval with respect to constituent PICO elements, specified individually), which may

[1] https://www.ncbi.nlm.nih.gov/pubmed/

yield improved search results [Scells *et al.*, 2017]. To address this challenge, there has been work on designing models to extract descriptions of the PICO elements from abstracts of RCT reports automatically [Boudin *et al.*, 2010; Kiritchenko *et al.*, 2010].

Earlier work on automating PICO tagging was hindered by a dearth of annotated corpora. Some more recent efforts have considered *distant supervision* (i.e., automatically derived, large-scale but potentially noisy annotations) to inducing larger training datasets, e.g., by exploiting structured abstracts [Jin and Szolovits, 2018] or deriving PICO annotations on sentences informed by existing, manually generated descriptions of these elements [Wallace *et al.*, 2016].

However, to more directly address the data paucity issue, colleagues and I recently introduced *EBM-NLP* [Nye *et al.*, 2018]. This is a corpus of ∼5,000 abstracts of articles describing RCTs with detailed PICO annotations. Specifically, spans in abstracts have been explicitly marked as describing the respective PICO elements. The dataset also includes more granular annotations within these spans. For details and for the dataset itself, refer to the corresponding paper [Nye *et al.*, 2018] and website,[2] respectively.

Initial results using this corpus as training data using an LSTM-CRF tagging model coupled with pre-trained word vectors induced over a large set of biomedical articles demonstrated promising performance [Nye *et al.*, 2018]. Further progress has since been realized by using (semi-supervised) data augmentation methods [Patel *et al.*, 2018] and by exploiting a neural language model (i.e., BERT [Devlin *et al.*, 2018]) pre-trained on a large corpus of scientific manuscripts [Beltagy *et al.*, 2019].

4 Appraising Reliability: Risk of Bias

A key component of evidence curation is assessing the reliability of the findings reported in articles describing individual trials. The Cochrane "risk of bias" tool [Higgins *et al.*, 2011] has formalized this for RCTs by codifying different types of statistical biases that might be introduced into a trial due to poor design or execution. For sake of transparency, risk of bias judgments are usually accompanied by snippets that support them extracted from corresponding articles. For example, if a study is deemed at *low* risk of bias due to shoddy randomization, the supporting snippet might read "Patients were randomized to groups according to a computer-generated sequence". Unfortunately, assessing risks of bias manually is time-consuming, often requiring more than 20 minutes per paper [Higgins *et al.*, 2011].

This has motivated work on semi-automating bias assessments [Millard *et al.*, 2015; Marshall *et al.*, 2015; Marshall *et al.*, 2015]. Direct supervision for risk of bias assessment is prohibitively expensive to acquire. We have therefore instead relied on *distant supervision* [Craven *et al.*, 1999; Mintz *et al.*, 2009], which refers to deriving (potentially noisy) annotations from existing resources, often via string matching and other heuristics. We acquired a database of previously conducted evidence syntheses,[3] which included risk of bias assessments. We then matched these assessments to corresponding trial reports (articles), and aligned supporting snippets retrieved from the database to sentences in articles via simple string matching. This process introduced noise, as alignments were imperfect. We mitigated this noise by exploiting a small amount of direct supervision to learn to derive the annotations on full-texts from matched database records, a novel strategy we have termed *supervised distant supervision* (SDS) [Wallace *et al.*, 2016].

This distant supervision strategy yielded a corpus of over 12,000 full-text articles with (derived) labels for the overall assessments concerning subsets of the four aforementioned risk of bias criteria (not all syntheses assess all of these criteria), along with the snippets in the texts that supported these assessments. In machine learning, text snippets that explicitly support document categorization are often referred to as *rationales* [Zaidan *et al.*, 2007; Strout *et al.*, 2019]. Models that exploit rationales have been proposed in previous work [Zaidan *et al.*, 2007; Small *et al.*, 2011], but these pre-dated the re-emergence of neural networks as the dominant class of models in NLP. We therefore extended convolutional neural network (CNN) text classification architectures to capitalize on rationales, which improved predictive performance over baseline models [Zhang *et al.*, 2016].

5 Inferring Results

Ultimately, the goal of processing medical literature is to help domain experts ascertain which treatments the evidence supports. An audacious NLP aim is then to build models that aim to directly facilitate this. As a first step toward this, we have recently proposed a new task and dataset that we call *evidence inference* [Lehman *et al.*, 2019].[4]

The idea is to build models that consume a full-text article describing an RCT along with an "ICO" triplet specifying an Intervention, a Comparator, and an Outcome (the Population is specified implicitly by the article). The model is then to infer whether the article provides evidence that suggests the Intervention *significantly increased*, *significantly decreased*, or *had no significant effect* relative to the Comparator, with respect to the Outcome. Note that most RCT reports will describe results for multiple interventions, comparators, and outcomes — hence one can assess multiple ICO triplets for the same articles, and the answer will very likely be different for these. We have collected a corpus (annotated by medical doctors) of about 10,000 ICO triplets coupled with ∼2,400 unique full-text articles.

This is a difficult problem that poses core NLP problems related to "machine reading" and question answering. Our initial results confirm this difficulty but also establish the feasibility of the task. In particular we have shown that if a model can reliably identify the snippets in the text that support inference concerning a given ICO frame the task is substantially easier [Lehman *et al.*, 2019]. Ultimately, we envision a system that can jointly extract the ICO elements *and* infer the reported findings concerning these, highlighting the relevant supporting snippets for the domain expert to inspect.

[2] http://pico-extraction.ebm-nlp.com

[3] The Cochrane Database of Systematic Reviews.

[4] http://evidence-inference.ebm-nlp.com/.

6 Hybrid Crowd and Expert Systems

As mentioned above, subject matter experts in this domain are expensive and already overburdened, making it difficult (and expensive) to collect direct supervision in general. Crowdsourcing is a popular strategy to mitigate the cost of acquiring training data in general, but the specialized nature of this domain (and the inherently somewhat technical nature of the literature) may preclude straightforward *crowdsourcing* of annotations to lay workers. We have investigated the use of (lay) annotators for evidence retrieval [Mortensen *et al.*, 2017] and annotation [Nye *et al.*, 2018]. In both cases we have found that if care is taken in aggregating redundant annotations provided by independent workers, one can derive reasonably good labels for training [Nguyen *et al.*, 2017].

Perhaps a more exciting direction is to combine machine learning, crowdworkers, and domain experts to form an efficient curation system. We have made initial forays into this direction in partnership with Cochrane crowd (a network of volunteer 'citizen scientists' working to curate evidence) in an effort to exhaustively identify all reports of RCTs. To this end we trained a high precision RCT classifier and explored different strategies for using it together with crowdworkers; we found that this can reduce workload by 60% to 80% without sacrificing recall. This setting also motivates novel research directions, e.g., how should annotation tasks be routed to workers so as to maximize the predictive accuracy of the trained model while minimizing effort/cost [Nguyen *et al.*, 2015; Yang *et al.*, 2019]?

More generally, we envision all of the NLP methods developed for this domain being used as assistive technologies, rather than replacing domain experts. This motivates research into the usability of language technologies in practice — how much do they actually help (if at all)?

7 Putting Models into Practice

We have incorporated many of the models described above into a prototype open-source system that we call *RobotReviewer* [Marshall *et al.*, 2017]. This includes both a web-based front-end (demo accessible at: https://robotreviewer.vortext.systems/), and functionality to provide annotation-as-a-service via a RESTful API.

We used this prototype to conduct a RCT of our own to assess the degree to which semi-automation of risk-of-bias assessment (discussed above) was found useful by end-users [Soboczenski *et al.*, 2019]. We found that semi-automation of this task reduced their workload by about 25% on average. Perhaps more importantly, users enjoyed working with the system, and found the automated assessments and supporting extracted snippets helpful. We envision conducting additional such exercises going forward to assess the practical utility of NLP and ML technologies that help domain experts make sense of the evidence.

8 Conclusions

As the evidence base continues to grow rapidly, so too the need for technologies that can help domain experts make sense of it. Here I have highlighted efforts led by myself and my collaborators on designing new NLP methods that aspire to meet this need. I have argued that this important application motivates core methodological challenges in NLP, and also highlights the need for human/machine hybrid systems that use ML to make domain experts more efficient, rather than attempting to replace them.

Acknowledgements

This *Early Career Spotlight* paper describes work that has been conducted in collaboration with many wonderful colleagues and graduate students — too many to list exhaustively here. However, I acknowledge in particular Iain Marshall, Thomas Trikalinos, Matt Lease, and Ani Nenkova for their key contributions to these efforts.

Parts of this work have been supported by: An NSF CAREER award (NSF 1750978); A grant from the National Library of Medicine (NLM R01LM012086); and other support from the NIH (NIH UH2CA203711).

References

[Bastian *et al.*, 2010] Hilda Bastian, Paul Glasziou, and Iain Chalmers. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS Med*, 7(9):e1000326, 2010.

[Beltagy *et al.*, 2019] Iz Beltagy, Arman Cohan, and Kyle Lo. SciBERT: Pretrained contextualized embeddings for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

[Blake and Lucic, 2015] Catherine Blake and Ana Lucic. Automatic endpoint detection to support the systematic review process. *Journal of Biomedical Informatics*, 56:42–56, 2015.

[Boudin *et al.*, 2010] Florian Boudin, Jian-Yun Nie, and Martin Dawes. Positional language models for clinical information retrieval. In *EMNLP*, pages 108–115, 2010.

[Cohen *et al.*, 2006] Aaron M Cohen, William R Hersh, K Peterson, and Po-Yin Yen. Reducing workload in systematic review preparation using automated citation classification. *Journal of the American Medical Informatics Association*, 13(2):206–219, 2006.

[Cohen *et al.*, 2015] Aaron M Cohen, Neil R Smalheiser, Marian S McDonagh, Clement Yu, Clive E Adams, John M Davis, and Philip S Yu. Automated confidence ranked classification of randomized controlled trial articles: an aid to evidence-based medicine. *Journal of the American Medical Informatics Association*, 22(3):707–717, 2015.

[Craven *et al.*, 1999] Mark Craven, Johan Kumlien, et al. Constructing biological knowledge bases by extracting information from text sources. In *ISMB*, volume 1999, pages 77–86, 1999.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Dickersin *et al.*, 1994] Kay Dickersin, Roberta Scherer, and Carol Lefebvre. Systematic reviews: identifying relevant studies for systematic reviews. *BMJ*, 309(6964):1286–1291, 1994.

[Higgins *et al.*, 2011] Julian PT Higgins, Douglas G Altman, Peter C Gøtzsche, Peter Jüni, David Moher, Andrew D Oxman, Jelena Savović, Kenneth F Schulz, Laura Weeks, and Jonathan AC Sterne. The Cochrane collaboration's tool for assessing risk of bias in randomised trials. *British Medical Journal*, 343:d5928, 2011.

[Huang *et al.*, 2006] Xiaoli Huang, Jimmy Lin, and Dina Demner-Fushman. Evaluation of PICO as a knowledge representation for clinical questions. In *AMIA*, volume 2006, page 359, 2006.

[Jin and Szolovits, 2018] Di Jin and Peter Szolovits. PICO element detection in medical text via long short-term memory neural networks. In *BioNLP*, pages 67–75, 2018.

[Kiritchenko *et al.*, 2010] Svetlana Kiritchenko, Berry de Bruijn, Simona Carini, Joel Martin, and Ida Sim. Exact: automatic extraction of clinical trial characteristics from journal publications. *BMC medical informatics and decision making*, 10(1):56, 2010.

[Lehman *et al.*, 2019] Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. Inferring Which Medical Treatments Work from Reports of Clinical Trials. In *NAACL*, 2019.

[Marshall *et al.*, 2015] Iain J Marshall, Joël Kuiper, and Byron C Wallace. Automating risk of bias assessment for clinical trials. *IEEE Journal of Biomedical and Health Informatics*, 19(4):1406–1412, 2015.

[Marshall *et al.*, 2017] Iain Marshall, Joël Kuiper, Edward Banner, and Byron C. Wallace. Automating Biomedical Evidence Synthesis: RobotReviewer. In *ACL*, pages 7–12, 2017.

[Marshall *et al.*, 2018] Iain J Marshall, Anna Noel-Storr, Joël Kuiper, James Thomas, and Byron C Wallace. Machine learning for identifying randomized controlled trials: an evaluation and practitioner's guide. *Research synthesis methods*, 9(4):602–614, 2018.

[Millard *et al.*, 2015] Louise AC Millard, Peter A Flach, and Julian PT Higgins. Machine learning to assist risk-of-bias assessments in systematic reviews. *International journal of epidemiology*, 45(1):266–277, 2015.

[Mintz *et al.*, 2009] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *ACL/AFNLP*, pages 1003–1011, 2009.

[Mortensen *et al.*, 2017] Michael L Mortensen, Gaelen P Adam, Thomas A Trikalinos, Tim Kraska, and Byron C Wallace. An exploration of crowdsourcing citation screening for systematic reviews. *Research synthesis methods*, 8(3):366–386, 2017.

[Nguyen *et al.*, 2015] An Thanh Nguyen, Byron C Wallace, and Matthew Lease. Combining crowd and expert labels using decision theoretic active learning. In *HCOMP*, 2015.

[Nguyen *et al.*, 2017] An T Nguyen, Byron C Wallace, Junyi Jessy Li, Ani Nenkova, and Matthew Lease. Aggregating and predicting sequence labels from crowd annotations. In *ACL*, volume 2017, pages 299–310, 2017.

[Nye *et al.*, 2018] Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain J Marshall, Ani Nenkova, and Byron C Wallace. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *ACL*, volume 2018, pages 197–207, 2018.

[O'Mara-Eves *et al.*, 2015] Alison O'Mara-Eves, James Thomas, John McNaught, Makoto Miwa, and Sophia Ananiadou. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *Systematic reviews*, 4(1):5, 2015.

[Patel *et al.*, 2018] Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron C Wallace. Syntactic patterns improve information extraction for medical search. In *NAACL*, volume 2018, pages 371–375, 2018.

[Peng *et al.*, 2017] Nanyun Peng, Hoifung Poon, Chris Quirk, Kristina Toutanova, and Wen-tau Yih. Cross-sentence n-ary relation extraction with graph lstms. *TACL*, 5:101–115, 2017.

[Przybyła *et al.*, 2018] Piotr Przybyła, Austin J Brockmeier, Georgios Kontonatsios, Marie-Annick Le Pogam, John McNaught, Erik von Elm, Kay Nolan, and Sophia Ananiadou. Prioritising references for systematic reviews with robotanalyst: A user study. *Research synthesis methods*, 9(3):470–488, 2018.

[Scells *et al.*, 2017] Harrisen Scells, Guido Zuccon, Bevan Koopman, Anthony Deacon, Leif Azzopardi, and Shlomo Geva. Integrating the framing of clinical questions via pico into the retrieval of medical literature for systematic reviews. In *CIKM*, pages 2291–2294. ACM, 2017.

[Settles, 2009] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[Small *et al.*, 2011] Kevin Small, Byron C Wallace, Carla E Brodley, and Thomas A Trikalinos. The constrained weight space svm: learning with ranked features. In *ICML*, pages 865–872. Omnipress, 2011.

[Soboczenski *et al.*, 2019] Frank Soboczenski, Thomas A. Trikalinos, Joël Kuiper, Randolph G. Bias, Byron C. Wallace, and Iain J. Marshall. Machine learning to help researchers evaluate biases in clinical trials: a prospective, randomized user study. *BMC Medical Informatics and Decision Making*, 2019.

[Strout *et al.*, 2019] Julia Strout, Ye Zhang, and Raymond J Mooney. Do human rationales improve machine explanations? *arXiv preprint arXiv:1905.13714*, 2019.

[Wallace *et al.*, 2010a] Byron C Wallace, Kevin Small, Carla E Brodley, and Thomas A Trikalinos. Active learning for biomedical citation screening. In *KDD*, pages 173–182, 2010.

[Wallace *et al.*, 2010b] Byron C Wallace, Thomas A Trikalinos, Joseph Lau, Carla Brodley, and Christopher H Schmid. Semi-automated screening of biomedical citations for systematic reviews. *BMC bioinformatics*, 11(1):55, 2010.

[Wallace *et al.*, 2016] Byron C. Wallace, Joël Kuiper, Aakash Sharma, Mingxi (Brian) Zhu, and Iain Marshall. Extracting PICO Sentences from Clinical Trial Reports using Supervised Distant Supervision. *Journal of Machine Learning Research*, 17(132):1–25, 2016.

[Wallace *et al.*, 2017] Byron C Wallace, Anna Noel-Storr, Iain J Marshall, Aaron M Cohen, Neil R Smalheiser, and James Thomas. Identifying reports of randomized controlled trials (rcts) via a hybrid machine learning and crowdsourcing approach. *Journal of the American Medical Informatics Association*, 24(6):1165–1168, 2017.

[Wang *et al.*, 2019] Hai Wang, Dian Yu, Kai Sun, Jianshu Chen, Dong Yu, Dan Roth, and David McAllester. Evidence sentence extraction for machine reading comprehension. *arXiv preprint arXiv:1902.08852*, 2019.

[Yang *et al.*, 2019] Yinfei Yang, Oshin Agarwal, Chris Tar, Byron C. Wallace, and Ani Nenkova. Predicting annotation difficulty to improve task routing and model performance for biomedical information extraction. In *NAACL*, pages 1471–1480, June 2019.

[Zaidan *et al.*, 2007] Omar Zaidan, Jason Eisner, and Christine Piatko. Using "annotator rationales" to improve machine learning for text categorization. In *NAACL*, pages 260–267, 2007.

[Zhang *et al.*, 2016] Ye Zhang, Iain Marshall, and Byron C Wallace. Rationale-augmented convolutional neural networks for text classification. In *EMNLP*, page 795, 2016.