

Syntactic Patterns Improve Information Extraction for Medical Search

Roma Patel^{1,2}, Yinfei Yang*, Iain Marshall³, Ani Nenkova¹ and Byron C. Wallace⁴

¹Department of Computer and Information Science, University of Pennsylvania

²Department of Computer Science, Rutgers University-Camden

³Department of Primary Care and Public Health Sciences, Kings College London

⁴College of Computer and Information Science, Northeastern University

Abstract

Medical professionals search the published literature by specifying the type of *patients*, the medical *intervention(s)* and the *outcome* measure(s) of interest. In this paper we demonstrate how features encoding syntactic patterns improve the performance of state-of-the-art sequence tagging models (both linear and neural) for information extraction of these medically relevant categories. We present an analysis of the type of patterns exploited, and the semantic space induced for these, i.e., the distributed representations learned for identified multi-token patterns. We show that these learned representations differ substantially from those of the constituent unigrams, suggesting that the patterns capture contextual information that is otherwise lost.

1 Introduction

The efficacy of medical treatments depends on patient characteristics, treatment administration details (e.g., dosage) and the measures or outcomes used to quantify treatment success. These criteria should be precisely defined when searching the medical literature (Richardson et al., 1995; Heneghan and Badenoch, 2013; Miller and Forrest, 2001). Unfortunately, these aspects are not usually described in a structured way. Abstracts with explicit category headings (Nakayama et al., 2005) partially address this, but these are not standardized nor uniform. Automated solutions are thus emerging to better support medical search, including methods for: identifying sentences containing key pieces of clinical information (Wallace et al., 2016); summarization (Sarker et al., 2016); identifying contradictory claims in medical articles (Alamri and Stevenson, 2016); and information retrieval system prototypes that harness this type of information (Boudin et al., 2010a,b).

Several studies have assessed the use of the PICO framework (Huang et al., 2006; Demner-Fushman and Lin, 2007). Our task is also to identify spans of text describing *PICO elements* i.e., the participants (P), interventions (I)/comparators (C), and outcomes (O) in the abstracts of articles reporting findings from randomized controlled trials (RCTs). We exploit the availability of structured abstracts in the medical domain: from these coarse (multi-)sentence labels we derive patterns typically used in bootstrap methods for entity recognition and relation extraction (Carlson et al., 2010). We incorporate these patterns into supervised sequence labeling models to improve the identification of P, I and O spans in new texts. Below we show examples of each extraction type: patterns are bolded and target PICO description spans italicized. The extracted patterns disambiguate fairly well the type of information expressed in the segment when individual words (e.g., “children”), do not. (P) The trial included 230 **children with** *Stage-IV lymphoblastic leukemia* ...

(I) In Group I, the **children were treated** with *prednisone* ...

(O) .. reported that Group 2 **children underwent** fewer *isolated bone marrow relapses* ..

We explore three strategies for exploiting extracted patterns in a state-of-the-art LSTM-CRF sequence tagging model (Lample et al., 2016; Ma and Hovy, 2016): as additional features at the CRF layer; as one-hot indicators concatenated to distributed representations of words; and as individual units embedded in a semantic space shared with words. The second representation improves recall for two extraction tasks, and the third improves precision for all three tasks. We analyze the induced semantic space to show that patterns capture contextual information that is otherwise lost.

* now at Google Inc.

2 Data

For training sequence tagging models we use a corpus of 4,741 medical article abstracts with manual crowd-sourced annotations for P, I, and O sequences. For testing we use a set of 191 abstracts annotated for P, I, and O by medical professionals. There are 18,849 (831), 44,329 (1,808), 41,454 (1,711) variable length sequences for P, I, and O in the training (testing) data.¹

For minimally supervised extraction of n -gram patterns, we use structured abstracts in which the authors describe different aspects of their work under targeted headings. We retrieved the headings and associated sections automatically from abstracts in XML format (downloaded from PubMed²). In general abstracts are structured idiosyncratically (often as Introduction, Methods, Results, Discussion). We capitalized on the minority of abstracts that used the explicit Participants, Intervention and Outcome headings. We obtained 50,000 segments for each of these three categories.

3 Patterns extraction and analysis

We extract syntactic patterns associated with each of the extraction types using AutoSlog-TS (Riloff, 1996), which consumes two sets of text: one relevant to an extraction domain and one irrelevant. In our case the relevant sets are the 50K P, I, and O sections, respectively, from the structured abstracts described above. The irrelevant set is a mix of 25K of the other two categories.

AutoSlog-TS generates n -gram patterns from input texts that capture the context of all noun phrases appearing as subject, direct and indirect object, or in a prepositional phrase. Each of these patterns is scored with the estimated probability that it occurred in an instance from the relevant set (out of all occurrences of the pattern), scaled by the number of times the pattern occurs (Riloff and Phillips, 2004). Common patterns that tend to occur in relevant sentences thus receive relatively high scores. We filter out patterns that contain digits, and those that occur fewer than 10 times in the structured abstract texts. Of the remaining patterns, we preserve those with probability 0.8 or

¹The complete details of the corpus, along with inter-annotator analysis and links for download of the full corpus will be described in a forthcoming paper and eventually made available here: http://www.byronwallace.com/EBM_abstracts_data.

²https://www.nlm.nih.gov/databases/download/pubmed_medline.html

higher of occurring with the relevant class. This yields 3,499, 3,898 and 2,386 patterns associated with P, I and O, respectively.

The vast majority of patterns are bigrams: 90% for P, 81% for I and 86% O. Fewer than 0.5% of the n -grams for each type are trigrams, and the remaining are unigrams. Examples of extracted patterns include: *women_who, years_of* and *diagnosed_with* for P; *patients_received* and *performed_after* for I; and *scale_of, patients_reported* and *rate_of* for O.

The majority (82.86%) of the extracted n -gram patterns comprise a combination of a content word and a function/stopword token.³ For example, the patterns *patients_with, patients_who* or *patients_from* are associated with the condition that a patient had, while *patients_were, patients_in* or *patients_received* describe the treatment they received. Function words provide disambiguating context for otherwise ambiguous words; this aids text classification and information retrieval (Riloff, 1995), and here we use them to improve sequence tagging models.

4 Patterns + linear CRF

For supervised IE models, we first consider including n -gram patterns as features in a linear-chain CRF (Lafferty et al., 2001). The standard set of token-level features used in the model include word identity, POS tag (from CoreNLP), and a list of binary features indicating whether the token is a digit, title (i.e., the first token only is uppercase), uppercase word, hyphenated word, or if the token is a punctuation mark (colon, fullstop or another symbol). In addition, features for the current token include the identity of the previous and next words, and the immediately preceding and following bi- and trigrams.

For the pattern-augmented CRF (**CRF-Pattern**), we add nine binary features that indicate if the current token and the immediately preceding/following bigrams are one of the AutoSlog-TS patterns associated with a given extraction type.⁴ There are three indicators, for P, I and O respectively. For the context bigrams, a feature is 1 if the bigram is one of the bigram patterns associated with this extraction type, 0 otherwise. The remaining three indicators have value 1 if the current token is one of the

³We use stopwords from CoreNLP (Manning et al., 2014).

⁴We ignore trigram patterns as they constitute <0.5% of identified patterns.

Model	Precision			Recall			F1		
	P	I	O	P	I	O	P	I	O
CRF	70.29	47.01	64.78	38.75	43.89	10.47	49.95	45.39	18.02
CRF-Pattern	73.3	52.1	66.37	40.62	45.41	44.07	52.27	48.52	52.96
LSTM-CRF	62.27	52.37	47.91	49.48	40.49	36.16	55.14	45.67	41.21
LSTM-CRF-Pattern (best)	76.10	58.25	44.66	64.75	43.39	35.20	69.97	49.74	39.69
Before CRF	61.87	38.65	46.27	41.45	23.8	37.27	49.64	29.55	41.28
Before BiLSTM	76.10	58.25	44.66	64.75	43.39	35.20	69.97	49.74	39.69
Embedding	55.18	51.07	44.30	54.24	47.41	41.60	54.71	49.17	42.91

Table 1: Models for extracting **P**articipants, **I**ntervention and **O**utcomes with and without pattern features, evaluated via token-level precision, recall and F1 scores. The first and second groups of rows report results for CRF and LSTM-CRF models without and with pattern features. The bottom group reports results achieved using different means of incorporating pattern features in neural models.

unigram patterns associated with a given type. For example, the nine features for the token “chronic” in the sequence *patients with chronic sinus issues* will be [1,0,0—0,0,0—0,0,0] because *patients_with* is one of the bigrams associated with the P type, the word “chronic” does not match any of the unigram patterns and “sinus issues” does not match any of the bigram patterns. Table 1 (top) reports the performance on the test data of the original CRF model, and the one augmented with pattern features. Including patterns yields consistent and considerable improvements in both precision and recall.

5 Patterns + LSTM-CRF

LSTM-CRF models (Lample et al., 2016; Ma and Hovy, 2016) for sequence tagging are general in that they do not require feature engineering. Instead, the features representing each token in the CRF are generated by a bi-directional LSTM. To generate this representation the LSTM consumes distributed word representations as input and outputs vector representations describing words *in context* (the bi-LSTM runs one LSTM in each direction, concatenating outputs). This vector is passed to a CRF layer for prediction. Character-level information for each word is incorporated by running a bi-LSTM over the characters of each word (Lample et al., 2016). We used the IO tagging scheme. We set the hidden state dimensions to 200 and dropout to 0.5. We did not perform gradient clipping. We used the Adam optimizer (Kingma and Ba, 2014) with learning rate = 0.001.

We consider three alternatives for extending this model with patterns. The first two use the indicator features describing the presence of patterns in the context, similar to those we described above for the linear CRF model. The difference is where these features are introduced: immediately before the CRF layer, concatenated with the output of the

LSTMs (**Before CRF**), or as part of the input to the LSTM, concatenated to the distributed word and character representations (**Before LSTM**). We use Moen and Ananiadou (2013)’s release of 200 dimensional word vectors trained over 5.5 billion words from medical articles as pre-trained word embeddings as input to the LSTM. We use the same set of hyperparameters for the LSTM as used in Lample et al. (2016), and do not optimize these for the present extraction tasks. The third alternative (**Embedding**) treats the patterns as collocations; we derive embedded representations for them as a unit, the way collocations are treated in Mikolov et al. (2013b). In training and during prediction each occurrence of a pattern in the input is treated as a single token with a corresponding distributed representation. Character-level representations are concatenated to word representations and the output of the LSTM cells is passed to the CRF to make predictions (as above).

For these embeddings, we collected 6 million PubMed abstracts (~1.4 billion words) filtering for only Human RCTs and used this to train word vectors using the Word2Vec tool (Mikolov et al., 2013a), inducing 200-dimensional vectors using the Skip-Gram model, where our vocabulary now consists of the learned n -gram patterns as single units, along with other unigrams. We then test these embedding representations by using them as input to our neural model for the structured prediction task.

6 Discussion of results

Table 1 reports the performance of the LSTM-CRF model achieved using each of the three strategies for incorporating pattern features discussed above. Inserting the pattern indicator features before the CRF layer yields the worst performance. Compared to the generic LSTM-CRF model, its F -measure is lower or the same for all three ex-

<i>n</i> -gram	similar to <i>n</i> -gram	similar to unigram
have children	1: <i>marry</i> 2: <i>conceive</i> 3: <i>breast-feed</i> 4: <i>be_pregnant</i> 5: <i>have_surgery</i>	1: <i>adults</i> 2: <i>adolescents</i> 3: <i>toddlers</i> 4: <i>youngsters</i> 5: <i>school-age</i>
condition _at	1: <i>status_at</i> 2: <i>features_at</i> 3: <i>outcome_at</i> 4: <i>qol_at</i> 5: <i>outcomes</i>	1: <i>circumstance</i> 2: <i>conditions</i> 3: <i>malady</i> 4: <i>ailment</i> 5: <i>situation</i>
filled _with	1: <i>covered_with</i> 2: <i>mixed_with</i> 3: <i>sealed_with</i> 4: <i>suspended</i> 5: <i>immersed_in</i>	1: <i>sealed</i> 2: <i>obturated</i> 3: <i>enclosed</i> 4: <i>enclosing</i> 5: <i>fill</i>
side_effects	1: <i>toxicities</i> 2: <i>side-effect</i> 3: <i>complications</i> 4: <i>AEs</i> 5: <i>nausea</i>	1: <i>effect</i> 2: <i>Effects</i> 3: <i>action</i> 4: <i>impact</i> 5: <i>influence</i>

Table 2: Example illustrating the shift in semantic space realized using pattern embeddings. For each of the listed *n*-grams, we report the top 5 most similar words to: (1) the *n*-gram pattern embedding, and, (2) the most relevant constituent *n*-gram i.e., the word in bold font.

traction categories, P, I, O.

Including the pattern features as input to the LSTM or as part of the embedding leads to substantial improvements over the baseline model, and this despite the smaller dataset over which pattern embeddings were learned: compared to the LSTM-CRF without pattern features, the former markedly improves precision for P and I, while the latter improves the recall for all three types. In terms of *F*-measure, best results for P and I are achieved by inserting the pattern features as input to the LSTM, with about 15% and 4% absolute improvement. For O, the best *F*-measure is achieved by incorporating patterns as part of the embeddings, yielding 1% absolute improvement.

The linear CRF and its variant enriched with pattern feature has the best precision, outperforming the LSTM-CRF models, but worse recall. It may still be useful for scenarios in which high precision extraction is needed.

7 Semantics of pattern embeddings

We established that syntactic patterns can markedly improve the extraction of patient, intervention and outcome descriptions in medical abstracts. We now turn to an analysis of how the patterns fit into the semantic space of word embeddings. Our goal is to quantify the extent to which including pattern representations changes which words will be considered similar to the pattern, but not to the words that compose it.

To this end, we find the ten words most similar (under cosine similarity) to each pattern, and those most similar to the individual words these comprise, in the embedding space. We analyze the size of the intersection of these two sets for all patterns (~10,000). To simplify the comparison we consider only the constituent word that has the largest intersection of similar words with the pattern of interest. The size of the intersection theoretically ranges from 0 to 10, but on average there is only

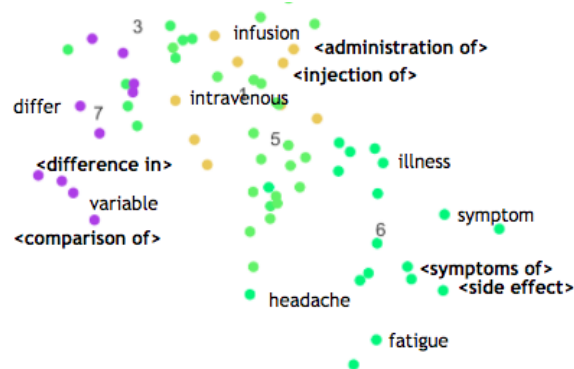


Figure 1: Scatter of PCA-reduced embeddings clustered using K-means. <> brackets show the syntactic pattern *n*-grams given by Autoslog-TS that are embedding in the same space as unigrams.

one word overlap between the words most similar with the pattern and those most similar with the constituent word. For the majority (61%) of the pattern–constituent word pairs, there is no overlap between the top 10 most similar words. To make this discussion more concrete, Table 2 provides examples of the top 5 most similar words to select bigram patterns and the constituent unigram with greatest overlap, shown in italics. The patterns encode disambiguating context that was previously lost in unigram representations.

Finally, we present a scatter of learned embeddings, reduced via IncrementalPCA⁵ in Figure 1. Embedded patterns cluster more intuitively than their content words alone. For example, the patterns *injection_of* and *administration_of* cluster together, along with other topically similar unigrams such as *infusion* and *intravenous* that may all correspond to Intervention terms. Similarly, *side_effect* is very different from its constituent words *side* or *effect*, and moreover, clusters with actual side effects like *headache* and *fatigue* that patients may suffer from in the course of a trial.

⁵We use the implementation in *scikit-learn* (Pedregosa et al., 2011).

Model	Precision			Recall			F1		
	P	I	O	P	I	O	P	I	O
LSTM-CRF	62.27	52.37	47.91	49.48	40.49	36.16	55.14	45.67	41.21
LSTM-CRF (Bigrams)	64.41	53.37	43.20	50.33	41.24	37.32	59.91	46.52	40.04
LSTM-CRF (Autoslog)	76.10	58.25	44.66	64.75	43.39	35.20	69.97	49.74	39.69

Table 3: Results to illustrate syntactic nature of Autoslog bigrams. Row 1 shows results of baseline model with no added features. Row 2 shows results of the model that uses all bigrams as features and Row 3 shows results of the model that uses only Autoslog extracted bigrams as features. Features are added before the LSTM, as incorporated in the best working model from Table 1.

8 Syntactic patterns vs bigrams

Our experiments show that using these bigram features extracted by AutoSlog improves model predictions. AutoSlog takes a fundamentally syntax-driven approach to identifying patterns, which suggests the discovered patterns (and associated performance boost) is due to exploiting syntax. However, the performance gains could also be due to additional contextual information that bigrams and larger n -grams provide over unigrams alone, rather than their syntactic properties.

We therefore performed an experiment to assess the influence of the syntactic AutoSlog bigrams, as compared to general bigram features. We consider the same data used as input to AutoSlog, i.e., 50,000 segments for the three categories P, I, and O. In the same setup, we decompose sentences within each category into bigrams, and collect bigram counts in the respective categories. We calculate precision for each category by collapsing the other two categories, similar to the AutoSlog procedure. We use the same threshold values as AutoSlog for filtering, i.e., we remove bigrams that occur fewer than 10 times or that have a score < 0.8 of occurring with the target class out of all occurrences. This procedure for identifying predictive bigrams yields a notably larger number of bigrams (30k) than AutoSlog (~ 10 K). Table 3 shows that while using generic bigrams as features sometimes leads to small improvements, the AutoSlog induced pattern bigrams result in substantially better performance. This suggests that the exploitation of syntactic structure in identifying patterns is indeed important. We also compare the performance of word2vec embeddings for unigrams and bigrams, and extended with collocations and syntactic patterns, trained on exactly the same data. In the experiments reported in Table 1, the unigram embeddings are trained on a larger dataset of generic medical text while the patterns are trained on a smaller set of medical ab-

stracts describing RCTs. In addition here we compare the AutoSlog patterns with collocations discovered by word2vec. Representing collocations leads to markedly lower F-score (Table 4). Representing bigrams leads to prediction performance better than that with collocations, but worse than unigrams.

Standard unigram representations that we trained work better than the off-the-shelf medical representations, possibly because they were trained specifically on abstracts of papers reporting the conduct and results of RCTs and thus better fit the abstracts we are analyzing. Most importantly, the LSTM-CRF with syntactic pattern embeddings results in the best observed performance.

Embedding	Vocabulary	P	I	O
Unigram	947,670	54.31	46.19	42.68
Bigram	9,326,144	52.01	43.71	38.77
collocation	1,254,863	50.31	40.21	40.21
Pattern	949,112	54.71	47.68	42.27

Table 4: LSTM-CRF predictions on word embeddings trained on the same 6 million documents. Column 1 shows the type of embedding, column 2 shows the size of the vocabulary and columns 3-5 show F1 score.

9 Conclusions

We presented a method for exploiting abundant unlabeled biomedical texts to generate minimally supervised extraction patterns that improve generic supervised models for sequence tagging in this domain. We explored alternative ways to incorporating the patterns in both linear and neural tagging models. In the latter, we analyzed the changes in semantic space that likely explain the observed gains in predictive performance.

10 Acknowledgements

This work was supported in part by the National Cancer Institute (NCI) of the National Institutes of Health (NIH), award number UH2CA203711 and the National Science Foundation (NSF), award number CCF-1433220.

References

- Abdulaziz Alamri and Mark Stevenson. 2016. A corpus of potentially contradictory research claims from cardiovascular research abstracts. *Journal of biomedical semantics* 7(1):36.
- Florian Boudin, Jian-Yun Nie, and Martin Dawes. 2010a. Clinical information retrieval using document and PICO structure. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA*. pages 822–830.
- Florian Boudin, Lixin Shi, and Jian-Yun Nie. 2010b. Improving medical information retrieval with PICO element detection. In *Advances in Information Retrieval, 32nd European Conference on IR Research, ECIR 2010, Milton Keynes, UK, March 28-31, 2010. Proceedings*. pages 50–61.
- Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. 2010. Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*.
- Dina Demner-Fushman and Jimmy Lin. 2007. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics* 33(1):63–103.
- Carl Heneghan and Douglas Badenoch. 2013. *Evidence-based medicine toolkit*. John Wiley & Sons.
- Xiaoli Huang, Jimmy Lin, and Dina Demner-Fushman. 2006. Evaluation of pico as a knowledge representation for clinical questions. In *AMIA annual symposium proceedings*. American Medical Informatics Association, volume 2006, page 359.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of NAACL-HLT*. pages 260–270.
- Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 1064–1074. <http://www.aclweb.org/anthology/P16-1101>.
- Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, System Demonstrations*. pages 55–60.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. corr abs/1301.3781 (2013).
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositional-ity. In *Advances in neural information processing systems*. pages 3111–3119.
- Syrene A Miller and Jane L Forrest. 2001. Enhancing your practice through evidence-based decision making: PICO, learning how to ask good questions. *Journal of Evidence Based Dental Practice* 1(2):136–141.
- SPFGH Moen and Tapio Salakoski2 Sophia Ananiadou. 2013. Distributional semantics resources for biomedical text processing.
- Takeo Nakayama, Nobuko Hirai, Shigeaki Yamazaki, and Mariko Naito. 2005. Adoption of structured abstracts by general medical journals and format for a structured abstract. *Journal of the Medical Library Association* 93(2):237.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12:2825–2830.
- W Scott Richardson, Mark C Wilson, Jim Nishikawa, and Robert SA Hayward. 1995. The well-built clinical question: a key to evidence-based decisions. *ACP journal club* 123(3):A12–A12.
- Ellen Riloff. 1995. Little words can make a big difference for text classification. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, pages 130–136.
- Ellen Riloff. 1996. [Automatically generating extraction patterns from untagged text](#). In *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 2*. AAAI Press, AAAI'96, pages 1044–1049. <http://dl.acm.org/citation.cfm?id=1864519.1864542>.
- Ellen Riloff and William Phillips. 2004. An introduction to the sundance and autoslog systems. Technical report.

Abeed Sarker, Diego Mollá, and Cecile Paris. 2016. Query-oriented evidence extraction to support evidence-based medicine practice. *Journal of biomedical informatics* 59:169–184.

Byron C. Wallace, Joël Kuiper, Aakash Sharma, Mingxi (Brian) Zhu, and Iain James Marshall. 2016. Extracting PICO sentences from clinical trial reports using supervised distant supervision. *Journal of Machine Learning Research* 17:132:1–132:25.