Browsing Health: Information Extraction to Support New Interfaces for Accessing Medical Evidence

Soham Parikh[1], **Elizabeth Conrad**[2], **Oshin Agarwal**[1],
Iain J. Marshall[3], **Byron C. Wallace**[4], **Ani Nenkova**[1]
[1] University of Pennsylvania, [2] University of Alabama,
[3] King's College London, [4] Northeastern University
{sohamp,oagarwal,nenkova}@seas.upenn.edu
ecconrad1@crimson.ua.edu
iain.marshall@kcl.ac.uk, b.wallace@northeastern.edu

Abstract

Standard paradigms for search do not work well in the medical context. Typical information needs, such as retrieving a full list of medical interventions for a given condition, or finding the reported efficacy of a particular treatment with respect to a specific outcome of interest cannot be straightforwardly posed in typical text-box search. Instead, we propose faceted-search in which a user specifies a condition and then can browse treatments and outcomes that have been evaluated. Choosing from these, they can access randomized control trials (RCTs) describing individual studies. Realizing such a view of the medical evidence requires information extraction techniques to identify the population, interventions, and outcome measures in an RCT. Patients, health practitioners, and biomedical librarians all stand to benefit from such innovation in search of medical evidence. We present an initial prototype of such an interface applied to pre-registered clinical studies. We also discuss pilot studies into the applicability of information extraction methods to allow for similar access to all published trial results.

1 Introduction

The most authoritative evidence regarding the efficacy of medical treatments is contained in papers describing results from randomized control trials (RCTs) (Byar et al., 1976). Evidence-based approaches to deciding standards of care require effective access to this literature, which may entail searching for information that the user does not have at the outset of their search (Relevo, 2012). Medical librarians (Crum and Cooper, 2013), practitioners, and patients would all benefit from a system that makes access to RCTs faster and more intuitive *via browsing capabilities*.

One of the obstacles to accessing RCT papers is that users may not begin with a well-formulated information need. For example a user may want to see what treatments have been studied for a given condition. Perhaps more importantly, individuals will value various health outcomes differently: some will have more interest in studies that used a particular criterion (outcome) to measure treatment effectiveness than in other studies.

For example, someone searching for treatments to control diabetes may be interested in knowing the extent to which treatments might prevent vision problems. But many trials studying diabetes use as the primary outcome measure changes in A1c, i.e. measurements indicative of average blood sugar levels over a couple of months. There is no correlation between A1c and retinopathy at least at diagnosis time (Maa and Sullivan, 2007). Being able to see a list of outcomes and selecting those of highest interest to preform a search for RCTs that talk about vision problems as well would be likely appreciated by users. Using surrogate outcome measures like A1c is considered as one of the core reasons ineffective or even harmful medical practices get adopted as standards of care (Chapter 3, (Prasad and Cifu, 2015)).

Here we present: *(i)* a faceted-search view to browse and search for medical literature based on the condition being studied (and other participant characteristics) in the study, the interventions used, and the outcomes measured; *(ii)* a prototype for the search of clinical studies on clinicaltrials.gov using study metadata; *(iii)* a study to determine the feasibility of using information extraction systems to extend this search to papers.

2 Browsing `ClinicalTrials.gov`

`ClinicalTrials.gov` is a centralized repository of clinical studies conducted around the world. Studies are registered by researchers who populate a number of required fields, such as the

Proceedings of the Workshop on Extracting Structured Knowledge from Scientific Publications, pages 43–47
Minneapolis, USA, June 6, 2019.

medical condition being studied, demographic information pertaining to the patients to be enrolled in the study (e.g., women, men, children), the medical interventions under consideration (e.g., specific drugs) and the outcomes that will be measured to determine success (or failure) of the medical intervention (such as the retinopathy and A1c example just discussed). The search interface provides a limited faceted-search ability[1] and a preview of interventions. It however does not provide capabilities to preview and select studies by type of intervention/outcome.

We provide a sense of how faceted search interface would work generally for RCT papers by initially providing this view over trials contained within ClinicalTrials.gov. The demo can be accessed here: https://browsing-health.herokuapp.com/.

Users can see at a glance typical outcome measures used in studies, and they can access studies that considered specific outcomes of interest. For example a search for 'asthma' reveals that the most commonly used outcome is *time to first severe asthma exacerbation*, a direct measure of effectiveness, while the second most used is 'fev1', a measurement of lung function which is a convenient but indirect surrogate measure – lung function can improve without affecting the number of severe exacerbations. Overall, the most common outcome measures across all registered studies were *overall survival, progression free survival, response rate* and *quality of life*.

Patient advocates, medical researchers and policy makers may benefit from this view of interventions and outcomes data, namely by using it to inform care and plan future studies. However, this search prototype was created using the metadata manually provided by researchers at the time of registration. This does not scale to handle the entire corpus of published evidence.

3 IE for RCTs

To organize all medical papers describing RCTs under a similar view, we need automated methods for extracting patient, intervention, and outcome descriptions from the abstracts (or full-texts) of articles describing trials. In this section we use pre-trained models for sequence labeling for these three aspects of RCTs (Nye et al., 2018). These are standard LSTM-CRF models (Huang et al., 2015; Lample et al., 2016) trained on crowdsourced annotations of ∼5000 abstracts of papers from MEDLINE (via PubMed) that describe RCTs with human subjects. We use the publicly released pre-trained models for sequence labeling from https://ebm-nlp.herokuapp.com/.

In the prior evaluation of these models, token-level precision and recall for coarse annotation of spans is reasonably good[2]. Spans describing participants are marked well in terms of both precision (75%) and recall (80%). Outcomes have good precision (80%) but lower recall and intervention spans have the lowest accuracy for automatic tagging. Here we explore the feasibility of using automated extraction to provide access to the medical literature via a browsing interface.

3.1 Complete label set

First, we ask whether the automatic span tagging can identify at least one span for each for patient, intervention, and outcome descriptors in (most) papers. This is a minimum requirement for being able to display the article via a faceted view. Note that this concern is independent of whether spans are *accurately* marked; a bare necessity prior to this is that any spans are marked at all.

We sampled thousands of abstracts of medical papers from MEDLINE (Greenhalgh, 1997). We used the associated metadata to identify a subset of abstracts for RCTs with human subjects. We extracted patient, intervention, and outcome spans using the pre-trained models mentioned above. Table 1 shows the percentage of articles for which at least one instance of each information type was labeled. Nearly 80% of articles had all three labels. Further, there were almost no human RCT abstracts that did not have any label (less than 1%). On inspection, we noticed that most of the abstracts without any automatically extracted study descriptors were either not actually descriptions of RCTs, or they were RCTs for diagnostic tests, not treatments for medical conditions.

The contrast with the coverage of extracted snippets in non-RCT human studies is reassuring. Only about 15 percent of such studies had all three study aspects labeled. On inspection, these tended to be RCTs in animals or observational studies.

We tested the coverage of automated extrac-

[1] https://clinicaltrials.gov/ct2/results?cond=diabetes

[2] See the leaderboard at http://www.ebm-nlp.com/#Leaderboard

Type of Article	% with 3 labels	% with no labels
Human RCT	76.72	0.77
Other abstracts	14.42	21.00

Table 1: Percentage of abstracts of papers describing human RCTs (337k) with all three study elements marked and no study element marked. This is contrasted with extracts from other papers (106k), either not RCTs or not with human subjects.

Type of Article	% with 3 labels	% with no labels
Structured	78.45	0.27
Unstructured	74.12	1.50

Table 2: The percentage of structured (176k) and unstructured (161k) abstracts of RCT humans studies for which all three/no descriptors are extracted.

tors on structured and unstructured abstracts, respectively. In unstructured abstracts authors decide what information to include in the abstracts of their paper. Structured abstracts were introduced to ensure that important information is included under an explicit heading, i.e. BACKGROUND, PARTICIPANTS, METHOD, OUTCOME. Different journals require their idiosyncratic structure for abstract but in general these have become the norm in the medical literature. The motivation for requiring structured abstracts is that they are more likely to explicitly and clearly describe important aspect of the described research (Sharma and Harrison, 2006). Here we use this expectation of better coverage on structured abstract as indirect measure of the abilities of automatic sequence tagging.

Here again we use meta-data to consider only human RCTs. Structured abstracts have been found to be more accessible and informative (Huth, 1987), so we expected that an automated extractor would similarly have different coverage of extracted information for the two types of abstracts. As Table 2 shows, this is indeed the case. A larger percentage of structured abstracts have all three study elements marked automatically, with 4% absolute difference in coverage between the two types of abstracts. Even in unstructured abstracts, there is virtually no abstracts from which not a single RCT aspect is extracted.

These results are encouraging. The sequence labeling models behave intuitively and do not mark spans in abstracts where the presence of spans is not expected (as in non-RCT/human study abstracts) or is expected to be harder to find, either because of wording or because it is not included (as in unstructured abstracts).

		Unseen		Seen	
	N	Unique	Total	Unique	Total
P	1	13.88%	407k	0.31%	575k
	2	33.10%	822k	3.70%	66k
	3	61.33%	783k	10.50%	12k
	4	80.27%	708k	17.81%	1.8k
I	1	15.22%	432k	0.42%	818k
	2	36.39%	796k	3.40%	107k
	3	64.92%	704k	6.99%	27k
	4	80.55%	595k	12.71%	5k
O	1	9.00%	808k	0.16%	1888k
	2	23.45%	1980k	1.63%	222k
	3	52.72%	1681k	3.72%	61k
	4	73.69%	1387k	6.44%	15k

Table 3: The number of N-grams (N=1,2,3,4) seen during training and marked during inference as well for each label. P stands for Population, I stands for Intervention and O stands for Outcomes

3.2 Do the models generalize?

Another important question is whether IE models generalize, that is, whether such models mark phrases not seen in the training data (Augenstein et al., 2017). To investigate this, we classify the extracted snippets from MEDLINE data into 'seen' (those that match exactly with or that appear as a substring of an annotated span in the training data) and 'unseen', i.e., snippets that do not appear as a (sub)unit in the training data.

Table 3 provides the number and percentage of extracted spans that do not occur in the training data, broken down by the length of the extracted span. The results are encouraging: even for unigrams, a large fraction of marked snippets are unseen and hence are generalized from the context. As expected, the longer the snippet, the larger the proportion of uniquely marked phrases, as longer phrases are unlikely to be repeated verbatim.

These results suggest that the models generalize well, and can identity novel snippets. This finding is promising in its implications for using IE to power a browseable view of trial data.

3.3 Impressions of Extraction Quality

In this section, we discuss a few qualitative observations related to automated extraction of patient, intervention and outcome information and the implications these have for further computational work on the extraction task.

Figures 1 and 2 show two abstracts with automatic annotations of participants, interventions and outcomes. Overall, the mark-up looks good, with all three RCT aspects covered. For the abstract in Figure 1, the interventions are accurately

This study analyzed the effectiveness of suprascapular nerve block under ultrasonographic guidance in patients with perishoulder pain. Patients with perishoulder pain were enrolled in the study and were randomly divided into 2 groups. In the first group of 25 patients (12 men and 13 women), nerve block was applied under ultrasonographic guidance. Mean patient age in this group was 55.1 years. In the control group, 25 patients (11 men and 14 women) underwent nerve block without ultrasonographic guidance; mean patient age was 51.6 years. Degree of pain was assessed using a visual analog scale (VAS) and shoulder function was evaluated using the Constant shoulder score (CSS) before the nerve block, immediately following the procedure, and 1 month after the procedure. There was no statistically significant difference between the 2 groups in VAS score and CSS before the procedure ($P > .05$). Immediately after the procedure, both the study and control groups revealed significantly improved VAS and CSS patterns ($P < .05$). However , the study group showed better VAS and CSS patterns than the control group at 1-month follow-up ($P < .05$). No complications occurred in the study group. In the control group , there were 2 cases of arterial punctures and 3 cases of direct nerve injury with neurological deficit for 2 months. Ultrasonography-guided suprascapular nerve injection is a safe, accurate, and useful procedure compared to the blind technique.

Figure 1: Example of a Human RCT abstract with the predicted spans for Participants (red), Intervention (blue) and Outcome (orange)

identified in the first and last sentence but in addition, a number of mentions of outcomes are erroneously marked as interventions. Importantly, the (same) outcomes are mentioned four times in the abstract. Some mentions are missed by the system, others are mistyped (recognized as interventions) and others are correctly identified. There is a similar problem where the unusual and unseen in training intervention *yogic package* is correctly marked but one of the subsequent mentions towards the end of the abstract is not detected. This observation implies that typical use of precision and recall, either token- or span-level, for evaluation of the sequence labeling may be misleading. Instead, an evaluation would need to capture the degree to which at least one instance of each aspect was captured correctly. Matching variants of the same aspect, such as 'Constant shoulder score' and 'CSS' will also be needed in order to support indexing and search over the extracted elements.

Another possible issue is the need to chunk more complex marked spans, particularly the conjunction of outcomes in Figure 1 and the list of outcomes in Figure 2. Similar need arises in getting the *medical condition* for which treatment is studied, by separating that string from the overall span including 'patients with/subjects with'.

This study aimed at studying the effect of yogic package (YP) with some selected pranayama, cleansing practices and meditation on pain intensity, inflammation, stiffness, pulse rate (PR), blood pressure (BP), lymphocyte count (LC), C-reactive protein (CRP) and serum uric acid (UA) level among subjects of rheumatoid arthritis (RA). Randomized control group design was employed to generate pre and post data on participants and controls. Repealed Measure ANOVAs with Bonferroni adjustment were applied to check significant overall difference among pre and post means of participants and controls by using PASW (SPSS Inc. 18th Version). Observed result favored statistically significant positive effect of YP on selected RA parameters and symptoms under study at $P < 0.05$, 0.01 and 0.001 respectively that showed remarkable improvement in RA severity after 40-day practice of YP. It concluded that YP is a significant means to reduce intensity of RA .

Figure 2: Example of a Human RCT abstract with the predicted spans for Participants (red), Intervention (blue) and Outcome (orange)

Such granular spans were annotated in the original EBM-NLP corpus (Nye et al., 2018), along with a detailed types of interventions and outcomes. Performance for labeling these details and granular spans however is much lower than that for the original high-level spans that we examine here. An alternative would be to learn chunking rules to identify the condition, individual interventions and individual outcomes in an unsupervised manner, by collocation analysis of the thousands of extracted snippets from the MEDLINE corpus.

In sum, progress on IE to aid browsing of the medical literature would require several modifications to track meaningful progress. Evaluation should be on exact spans that can serve directly as indexing terms for abstracts, and these should measure the ability of the system to find at least one mention of each RCT aspect.

4 Conclusion

We presented a proposal for an alternative mode of access to papers describing randomized control trials. We present a crude example of the browsing capabilities that can be built upon information extraction results from the medical literature. The initial prototype is powered by RCT descriptors written by a person during the registration of the study. We then present some preliminary experiments on applying existing sequence labeling methods for extracting RCT descriptors from the free text of paper abstracts. Results are promising, showing good coverage and reasonable activation of the extraction. We identify aspects in which the information extraction tasks ought to be adjusted

in order to better serve indexing needs.

Biomedical librarians are increasingly asked to identify medical evidence in preparation of future randomized control trials and questions regarding patient care. The browsing interface we envision will likely facilitate their work.

References

Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. 2017. Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language*, 44:61–83.

David P Byar, Richard M Simon, William T Friedewald, James J Schlesselman, David L DeMets, Jonas H Ellenberg, Mitchell H Gail, and James H Ware. 1976. Randomized clinical trials: perspectives on some recent ideas. *New England Journal of Medicine*, 295(2):74–80.

Janet A Crum and I Diane Cooper. 2013. Emerging roles for biomedical librarians: a survey of current practice, challenges, and changes. *Journal of the Medical Library Association: JMLA*, 101(4):278.

Trisha Greenhalgh. 1997. How to read a paper: the medline database. *Bmj*, 315(7101):180–183.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Edward J Huth. 1987. Structured abstracts for papers reporting clinical trials. *Annals of Internal Medicine*, 106(4):626–627.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 260–270.

April Y Maa and Brian R Sullivan. 2007. Relationship of hemoglobin a1c with the presence and severity of retinopathy upon initial screening of type ii diabetes mellitus. *American journal of ophthalmology*, 144(3):456–457.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain James Marshall, Ani Nenkova, and Byron C. Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 197–207.

Vinayak K Prasad and Adam S Cifu. 2015. *Ending medical reversal: improving outcomes, saving lives*. JHU Press.

Rose Relevo. 2012. Effective search strategies for systematic reviews of medical tests. *Journal of general internal medicine*, 27(1):28–32.

Sandeep Sharma and Jayne E Harrison. 2006. Structured abstracts: do they improve the quality of information in abstracts? *American journal of orthodontics and dentofacial orthopedics*, 130(4):523–530.