MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220150

AUTOMETA: Automatic Meta-Analysis System Employing Natural Language Processing

Faith W. Mutinda, Shuntaro Yada Ph.D., Shoko Wakamiya Ph.D., Eiji Aramaki Ph.D.

Nara Institute of Science and Technology, Ikoma, Nara, Japan

Abstract

Meta-analyses examine the results of different clinical studies to determine whether a treatment is effective or not. Meta-analyses provide the gold standard for medical evidence. Despite their importance, meta-analyses are time-consuming and this poses a challenge where timeliness is important. Research articles are also increasing rapidly and most meta-analyses become outdated after publication since they have not incorporated new evidence. Therefore, there is increasing interest to automate meta-analysis so as to speed up the process and allow for automatic update when new results are available. In this preliminary study we present AUTOMETA, our proposed system for automating meta-analysis which employs existing natural language processing methods for identifying Participants, Intervention, Control, and Outcome (PICO) elements. We show that our system can perform advanced meta-analyses by parsing numeric outcomes to identify the number of patients having certain outcomes. We also present a new dataset which improves previous datasets by incorporating additional tags to identify detailed information.

Keywords:

Automatic Meta-analysis, Natural Language Processing (NLP)

Introduction

A meta-analysis is a type of quantitative study that collects and analyses the results of different studies that are all focused on the same disease, treatment, or outcome to ascertain if a treatment is effective or not. Meta-analyses provide the gold standard for medical evidence [1]. Regardless of their importance, meta-analyses tend to be labor-intensive, cost-, and time-consuming because they require comprehensive search and reading of hundreds of research articles written in unstructured natural language to find medical evidence [2]. These research articles are increasing rapidly and it is becoming difficult for researchers to keep up with new publications [3,4]. For instance, a recent study showed that on average 59 research articles related to the COVID-19 pandemic are published daily [5]. It takes more than 1 year (from registration to publication) to finalize a meta-analysis which is rarely updated [6,7]. This poses a challenge especially for practitioners in the infectious disease field where timeliness is important, and informed decisions need to be made promptly. Furthermore, most meta-analyses are quickly outdated after publication as they have not included new evidence which might change the primary results [6].

Automating the meta-analysis process including searching databases for relevant studies, screening the studies, data extraction, and statistical analysis, will improve the dissemination of medical evidence. Also, it allows for automatic updates when new results are available [8]. Surveys on automation of meta-analysis show that many methods have been proposed for automating the different stages for meta-analysis [2,3]. A survey by Marshall et al. [3] suggests that systems for searching literature, identifying randomized controlled trials (RCTs), and screening articles have achieved high performance and are ready for use. However, the systems for the data extraction step are still not readily available. This is because data extraction requires high accuracy which may be difficult for automated systems to achieve. A barrier to the development of high-performance models is the lack of training data for the data extraction task [3]. Although there are few high-quality training data, which are usually expensive to create, Nye et al. [9] developed the EBM-NLP corpus containing about 5000 abstracts of RCT articles annotated in detail. This corpus is helpful for the development of automatic models for data extraction for meta-analysis. A drawback of this corpus is that they do not annotate numbers which identify the outcome results (i.e., the number of the patients having certain outcomes).

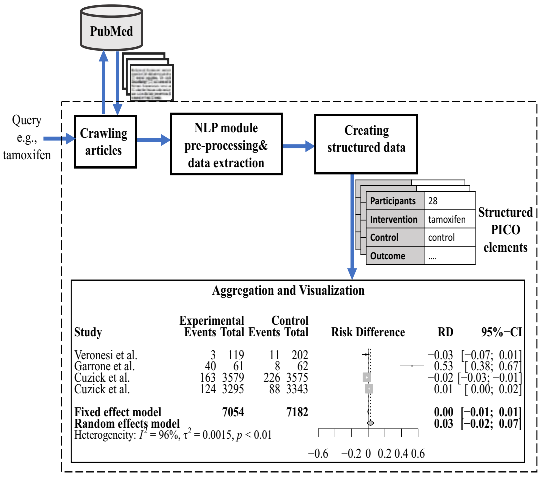

PubMed
Query e.g., tamoxifen
Crawling articles
NLP module pre-processing& data extraction
Creating structured data
Participants
28
Intervention
tamoxifen
Control
control
Outcome
....
Structured PICO elements
Aggregation and Visualization
Study
Experimental Events Total
Control Events Total
Risk Difference
RD
95%−CI
Veronesi et al. 3 119 11 202 −0.03 [−0.07; 0.01]
Garrone et al. 40 61 8 62 0.53 [0.38; 0.67]
Cuzick et al. 163 3579 226 3575 −0.02 [−0.03; −0.01]
Cuzick et al. 124 3295 88 3343 0.01 [0.00; 0.02]
Fixed effect model 7054 7182 0.00 [−0.01; 0.01]
Random effects model 0.03 [−0.02; 0.07]
Heterogeneity: $I^2 = 96\%$, $\tau^2 = 0.0015$, $p < 0.01$
−0.6
−0.2
0
0.2
0.4
0.6

Figure 1 – System architecture of AUTOMETA

This study aims at developing an automatic meta-analysis system based on natural language processing (NLP) by creating a corpus that has additional tags to identify detailed information for the outcomes, especially identifying the number of patients having certain outcomes. We focused on breast cancer because it is one of the leading causes of death in the world[1] and extracted 600 abstracts of RCT's related to breast cancer prevention from the PubMed database[2]. The corpus is annotated similarly to the EBM-NLP [9] corpus and with additional tags to identify detailed information for the outcomes. AUTOMETA (shown in Figure 1) extracts the Participants, Interventions, Control, and Outcomes (PICO) elements from abstracts by employing NLP techniques and turns them into structured data. Then, it parses numeric outcomes into their associated fields and aggregates and visualizes extracted outcomes for statistical analysis.

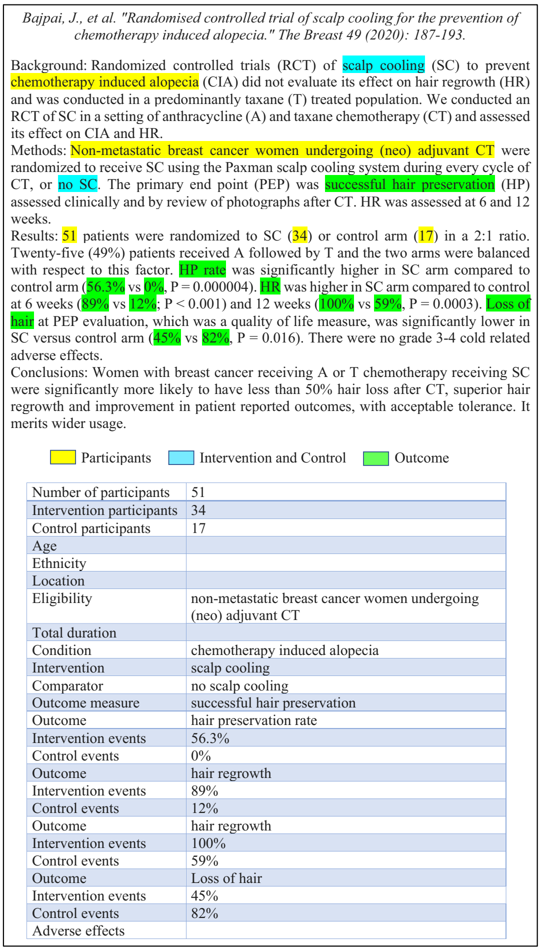
Materials and methods

Corpus

The corpus for this study consists of abstracts extracted from PubMed2. PubMed is a free search engine which provides access to the MEDLINE database3 which contains indexes, references, titles, and abstracts for biomedical and life sciences articles. We extracted articles with study type RCT, and are not meta-analysis or systematic-reviews. This was achieved by using keywords such as "randomized controlled," "randomised controlled," "meta-analysis," and "systematic review."

The annotators were asked to identify the PICO elements in each abstract as discussed below. Figure 2 shows an abstract with the PICO elements highlighted.

- Participants: identify text snippets which describe the characteristics of the participants. Here we defined 7 labels for identifying participants' characteristics which include the number of participants (total participants, participants in the intervention group, or participants in the control group), average age, ethnicity, location of the study, eligibility, total duration, and condition. Although breast cancer is the main condition, we are also interested in identifying the condition/symptom of breast cancer that is being treated (such as hair loss, bone loss, and vomiting).
- Intervention and Control: identify the specific intervention and control used in the study.
- Outcome: identify what is being measured in the study so as to identify if the treatment worked. Here we defined 5 labels which include the outcomes that were measured, number of events in the intervention group, number of events in the control group, outcome measure, and adverse effects.



Bajpai, J., et al. "Randomised controlled trial of scalp cooling for the prevention of chemotherapy induced alopecia." The Breast 49 (2020): 187-193.

Background: Randomized controlled trials (RCT) of scalp cooling (SC) to prevent chemotherapy induced alopecia (CIA) did not evaluate its effect on hair regrowth (HR) and was conducted in a predominantly taxane (T) treated population. We conducted an RCT of SC in a setting of anthracycline (A) and taxane chemotherapy (CT) and assessed its effect on CIA and HR.
Methods: Non-metastatic breast cancer women undergoing (neo) adjuvant CT were randomized to receive SC using the Paxman scalp cooling system during every cycle of CT, or no SC. The primary end point (PEP) was successful hair preservation (HP) assessed clinically and by review of photographs after CT. HR was assessed at 6 and 12 weeks.
Results: 51 patients were randomized to SC (34) or control arm (17) in a 2:1 ratio. Twenty-five (49%) patients received A followed by T and the two arms were balanced with respect to this factor. HP rate was significantly higher in SC arm compared to control arm (56.3% vs 0%, P = 0.000004). HR was higher in SC arm compared to control at 6 weeks (89% vs 12%; P < 0.001) and 12 weeks (100% vs 59%, P = 0.0003). Loss of hair at PEP evaluation, which was a quality of life measure, was significantly lower in SC versus control arm (45% vs 82%, P = 0.016). There were no grade 3-4 cold related adverse effects.
Conclusions: Women with breast cancer receiving A or T chemotherapy receiving SC were significantly more likely to have less than 50% hair loss after CT, superior hair regrowth and improvement in patient reported outcomes, with acceptable tolerance. It merits wider usage.

🟨 Participants 🟦 Intervention and Control 🟩 Outcome

Number of participants	51
Intervention participants	34
Control participants	17
Age	
Ethnicity	
Location	
Eligibility	non-metastatic breast cancer women undergoing (neo) adjuvant CT
Total duration	
Condition	chemotherapy induced alopecia
Intervention	scalp cooling
Comparator	no scalp cooling
Outcome measure	successful hair preservation
Outcome	hair preservation rate
Intervention events	56.3%
Control events	0%
Outcome	hair regrowth
Intervention events	89%
Control events	12%
Outcome	hair regrowth
Intervention events	100%
Control events	59%
Outcome	Loss of hair
Intervention events	45%
Control events	82%
Adverse effects	

Figure 2– An abstract with PICO elements highlighted. The top part shows the abstract while the bottom part shows the PICO elements transformed into a structured format. Some slots are empty if their corresponding information is not in the abstract.

Data extraction

The data extraction task can be formulated as a sequence labelling task, i.e., given a token classify it as one of predefined named entity recognition (NER) tags. Previous studies on extraction of PICO elements have proposed rule-based, Support Vector Machines (SVM), Naive Bayes (NB), and Conditional Random Fields (CRF)-based models [2]. Although these models are useful in information extraction, they heavily rely on hand-crafted features. Designing hand-crafted features is time-consuming and requires domain knowledge in determining useful features.

Deep learning-based models have gained popularity for the data extraction task since they do not require predefined hand-crafted features. Deep learning-based models have achieved state-of-the-art performance in information retrieval by using contextualized text embeddings [10,11]. Jin et al. [12] proposed a bidirectional long short-term memory (Bi-LSTM) model for extraction of PIO elements from PubMed abstracts. Mezaoui et

[1] https://www.who.int/news-room/fact-sheets/detail/cancer
[2] https://pubmed.ncbi.nlm.nih.gov/
[3] https://www.nlm.nih.gov/bsd/pmresources.html

al. [13] proposed an improvement of the Jin et al. [12] model by adding a multi-label PIO classifier based on BERT which provides state-of-the-art embeddings.

In this study, we use a BERT-based model since BERT has achieved state-of-the-art performance in various NLP tasks including NER [10,11]. BERT is a general-purpose language model trained on a large dataset and uses the encoder structure of the Transformer, which is an attention mechanism that learns contextual relations between words (or subwords) in a text. BERT was first pre-trained on general English domain texts including Wikipedia and BooksCorpus. However, biomedical domain texts, such as our corpus, contain domain-specific words and general-purpose language models might perform poorly in domain-specific tasks. Domain-specific BERT models such as BioBERT [14] have been developed to address this challenge. BioBERT is initialized by BERT and further trained on biomedical domain texts including PubMed abstracts and PubMed Central full-text articles.

The data extraction step aims to extract PICO elements from the research articles and convert them into a structured form as shown in Figure 2. After extraction of PICO elements, we parse numeric texts to identify the number of participants having certain outcomes, as shown in Figure 3. This is a challenging task because research articles lack uniformity, and different articles report results differently. Some of the common patterns to indicate which patients have certain outcomes include, Z of Y , X% (n = Z), Z (X%), and X%. In cases such as X%, we require knowledge of the number of participants in the intervention and control groups so as to calculate the number of affected participants.

Acronym expansion

Acronyms are commonly used in research articles to avoid repeating long phrases and save space. Although acronyms simplify writing and reading, they pose a challenge to text understanding tasks [15]. In research articles, acronyms mostly occur in the words preceding their first use in parentheses, for example, "The primary end point (PEP) was successful hair preservation (HP) assessed clinically and by review of photographs." In this study, we adopt a rule-based acronym expansion method using regular expressions. First acronyms are identified by finding terms in parentheses if they are between two and ten characters. By using regular expressions, expansion candidates are found from the surrounding text.

System architecture

The architecture of the proposed AUTOMETA system is as shown in Figure 1. Our main goal is to provide a system for automating the meta-analysis process as much as possible so as to reduce the time taken in conducting a meta-analysis. The proposed system consists of four major components: crawling PubMed articles, NLP module, creating structured data, and aggregation and visualization. First, a user queries the PubMed database and related articles are returned. Abstracts are then extracted from the articles and passed to the NLP module for preprocessing and extraction of PICO elements. The extracted data is then converted into a structured form as shown in Figure 2. In this study, we also parse numeric texts to identify the number of patients having certain outcomes (Figure 3). Identification of the number of patients having certain outcomes is important for statistical analysis to determine the effectiveness of an intervention. The final step of the system is to aggregate the studies and present them for statistical analysis such as visualizing the data using forest plots (Figure 1) which provide a summary and the extent to which results from different studies overlap.

			Intervention group		Control group	
Study	Sentence	Predicted outcome	#events	#total	#events	#total
Veronesi et al.	Temporary discontinuation occurred in 2.5% of patients in the adjuvant studies and in 5.4% of women in the chemoprevention study …	Temporary discontinuation occurred	2.5%	119	5.4%	202
Garrone et al.	A significantly higher proportion of patients in the tamoxifen group had increased ET at 6 and 12 months from randomisation compared with the exemestane group (66.1% and 64.3% versus 12.1% and 6.8%, respectively; P < 0.0001).	significantly higher proportion of patients in the tamoxifen group had increased ET at 6 and 12 months from randomisation compared with the exemestane group	66.1%	61	12.1%	62
Cuzick et al.	The risk of developing breast cancer was similar between years 0-10 (226 [6.3%] in 3575 women in the placebo group vs 163 [4.6%] in 3579 women in the tamoxifen group; hazard ratio [HR] 0.72 [95% CI 0.59-0.88], p=0.001) and after 10 years (124 [3.8%] in 3295 women vs 88 [2.6%] in 3343, respectively; HR 0.69 [0.53-0.91], p=0.009).	risk of developing breast cancer was similar between years 0-10	163 [4.6%]	3579	(226 [6.3%]	3575
		and after 10 years	124 [3.8%]	3295	88 [2.6%]	3343

Figure 3– Sample outcomes extracted from three studies, Veronesi et al. [16], Garrone et al. [17], and Cuzick et al. [18]. The studies are clinical trials for investigating the effect of tamoxifen (intervention) in breast cancer patients. The red text shows model prediction error.

Results and discussion

The motivation of this paper is to present the feasibility of automating meta-analysis. This study is preliminary and the entire AUTOMETA system was not evaluated. However, we investigate the performance of the most important module, the NLP module. Our corpus consists of 600 PubMed abstracts annotated with PICO elements, and the frequency of each element is as shown in Table 1. The dataset was split into 80% training set and 20% test set. We developed a BioBERT-based model and the performance was evaluated using precision, recall, and F1 score. The results are shown in Table 2. The performance for several categories such as the number of participants, average age, and total duration is relatively high. The system achieved the highest F1 score for number of participants, which had a high frequency (1435) in the dataset. The F1 score for intervention and control was the lowest indicating that the model could not identify intervention and control effectively.

Figure 3 shows examples of studies whose intervention is tamoxifen. The model was able to capture the outcomes and their respective intervention events and control events relatively well. In the corpus, the number of participants irrespective of whether they are in the intervention group or control group are labelled as the number of participants. Therefore, to identify the number of participants in the intervention and control groups, first the system finds the extracted number of participants, and then assigns them to the intervention group or control group based on which they are closest to. In the Cuzick et al. study, the model misidentified the intervention events and control events. In most articles, intervention events tend to appear before control events. The model might have learned this pattern and hence the reason for the error.

The evaluation of how well the system identifies outcomes and their respective intervention events and control events is challenging. Although the performance of this step largely depends on the performance of the data extraction step, how to effectively evaluate is one of our important future work. Moreover, our corpus is small, and we believe that by increasing the annotated data the model performance can be significantly improved. However, considering that this study is preliminary, we believe the proposed system, AUTOMETA, is technically feasible.

Table 1– Corpus statistics

Category	Sub-category	# tags
Participants	Number of participants	1435
	Average age	168
	Ethnicity	75
	Location	130
	Eligibility	654
	Total duration	463
	Condition	454
Intervention and control	Intervention	619
	Control	606
Outcome	Outcome measure	1019
	Outcome	2584
	Intervention events	1340
	Control events	854
	Adverse effects	119

Limitations

One limitation is that our study uses abstracts only. Abstracts sometimes lack important information that may be presented in the full text document. A manual check of the abstracts in our corpus found that some do not mention the number of participants in the intervention and control groups. This will present a challenge when determining the number of the patients having certain outcomes for statistical analysis. A second limitation is that we do not account for participants who drop out of a study and this might affect the final results of the meta-analysis. Abstracts often lack information about the number of participants who drop out from a study. Therefore, for future work it will be important to consider full-text articles.

Table 2– BioBERT model results in terms of precision, recall and F1 score on the test set.

Sub-category	Precision	Recall	F1
Number of participants	0.87	0.94	0.91
Average age	0.93	0.88	0.90
Ethnicity	0.83	0.83	0.83
Location	0.71	0.92	0.80
Eligibility	0.82	0.87	0.84
Total duration	0.78	0.84	0.81
Condition	0.69	0.63	0.66
Intervention	0.65	0.61	0.63
Control	0.62	0.63	0.63
Outcome measure	0.80	0.82	0.81
Outcome	0.77	0.87	0.82
Intervention events	0.64	0.80	0.71
Control events	0.71	0.68	0.69
Adverse effects	0.91	0.59	0.71

Conclusion

In this study, we presented AUTOMETA, a system for automating meta-analysis by using NLP techniques. Our main goal is to provide a system for automating the meta-analysis process as much as possible so as to reduce the time taken in conducting a meta-analysis, increase the dissemination of medical evidence, and allow for automatic update when new evidence becomes available. The proposed AUTOMETA system extracts PICO elements from research articles, performs advanced meta-analysis by parsing numeric outcomes to identify the number of patients having certain outcomes, and presents results in a structured form for statistical analysis. We also presented a new dataset which improves previously released datasets by providing detailed annotation for the outcomes.

References

[1] D.J. Cook, C.D. Mulrow and R.B. Haynes, Systematic reviews: synthesis of best evidence for clinical decisions, Annals of internal medicine 126(5) (1997), 376–380.

[2] S.R. Jonnalagadda, P. Goyal and M.D. Huffman, Automating data extraction in systematic reviews: a systematic review, Systematic reviews 4(1) (2015), 78.

[3] I.J. Marshall and B.C. Wallace, Toward systematic review automation: a practical guide to using machine learning tools in research synthesis, Systematic reviews 8(1) (2019), 163.

[4] H. Bastian, P. Glasziou and I. Chalmers, Seventy-five trials and eleven systematic reviews a day: how will we ever keep up?, PLoS med 7(9) (2010), e1000326.

[5] S.B. Kambhampati, R. Vaishya and A. Vaish, Unprecedented surge in publications related to COVID-19 in the first three months of pandemic: A bibliometric analytic report, Journal of clinical orthopaedics and trauma 11(Suppl 3) (2020), S304.

[6] K.G. Shojania, M. Sampson, M.T. Ansari, J. Ji, S. Doucette and D. Moher, How quickly do systematic reviews go out of date? A survival analysis, Annals of internal medicine 147(4) (2007), 224–233.

[7] R. Borah, A.W. Brown, P.L. Capers and K.A. Kaiser, Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry, BMJ open 7(2) (2017), e012545.

[8] M. Michelson, Automating meta-analyses of randomized clinical trials: a first look., in: AAAI Fall Symposia, Citeseer, (2014).

[9] B. Nye, J.J. Li, R. Patel, Y. Yang, I.J. Marshall, A. Nenkova and B.C. Wallace, A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature, in: Proceedings of the conference. Association for Computational Linguistics. Meeting, Vol. 2018, NIH Public Access, (2018), 197.

[10] A. Radford, K. Narasimhan, T. Salimans and I. Sutskever, Improving language understanding by generative pre-training, Preprint (2018).

[11] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proc. of NAACL'19, Vol. 1, (2019), 4171– 4186.

[12] D. Jin and P. Szolovits, Pico element detection in medical text via long short-term memory neural networks, in: Proceedings of the BioNLP 2018 workshop, (2018), 67–75.

[13] H. Mezaoui, A. Gontcharov and I. Gunasekara, Enhancing PIO element detection in medical text using contextualized embedding, arXiv preprint arXiv:1906.11085 (2019).

[14] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C.H. So and J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36(4) (2020), 1234–1240.

[15] A. Pouran Ben Veyseh, F. Dernoncourt, T.H. Nguyen, W. Chang and L.A. Celi, Acronym identification and disambiguation shared tasks for scientific document understanding, arXiv e-prints (2020), arXiv–2012.

[16] A. Veronesi, M.A. Pizzichetta, M.A. Ferlante, M. Zottar, M.D. Magri, D. Crivellari and S. Foladore, Tamoxifen as adjuvant after surgery for breast cancer and tamoxifen or placebo as chemoprevention in healthy women: different compliance with treatment, Tumori Journal 84(3) (1998), 372–375.

[17] O. Garrone, G. Bertelli, E. Principe, P.D. Lewis, M. Occelli, E. Miraglio and M.C. Merlano, A prospective randomised study of transvaginal ultrasound effects of tamoxifen and exemestane in postmenopausal women with early breast cancer, Tumori Journal 100(6) (2014), 620–624.

[18] J. Cuzick, I. Sestak, S. Cawthorn, H. Hamed, K. Holli, A. Howell, J.F. Forbes, I.-I. Investigators et al., Tamoxifen for prevention of breast cancer: extended long-term follow-up of the IBIS-I breast cancer prevention trial, The lancet oncology 16(1) (2015), 67–75.

Address for correspondence

Eiji Aramaki: aramaki@is.naist.jp