OVERVIEW

Towards automating systematic reviews on immunization using an advanced natural language processing–based extraction system

David Begert[1], Justin Granek[1], Brian Irwin[1], Chris Brogly[2]*

Abstract

Evidence-informed decision making is based on the premise that the entirety of information on a topic is collected and analyzed. Systematic reviews allow for data from different studies to be rigorously assessed according to PICO principles (population, intervention, control, outcomes). However, conducting a systematic review is generally a slow process that is a significant drain on resources. The fundamental problem is that the current approach to creating a systematic review cannot scale to meet the challenges resulting from the massive body of unstructured evidence. For this reason, the Public Health Agency of Canada has been examining the automation of different stages of evidence synthesis to increase efficiencies.

In this article, we present an overview of an initial version of a novel machine learning–based system that is powered by recent advances in natural language processing (NLP), such as BioBERT, with further optimizations completed using a new immunization-specific document database. The resulting optimized NLP model at the core of this system is able to identify and extract PICO-related fields from publications on immunization with an average accuracy of 88% across five classes of text. Functionality is provided through a straightforward web interface.

Affiliations

[1] Xtract AI, Vancouver, BC

[2] Centre for Immunization and Respiratory Infectious Diseases, Public Health Agency of Canada, Ottawa, ON

***Correspondence:**
christopher.brogly@canada.ca

Suggested citation: Begert D, Granek J, Irwin B, Brogly C. Towards automating systematic reviews on immunization using an advanced natural language processing–based extraction system. Can Commun Dis Rep 2020;46(6):174–9. https://doi.org/10.14745/ccdr.v46i06a04

Keywords: automation, natural language processing, NLP, data extraction, systematic reviews, machine learning

Introduction

Evidence-based medicine relies on systematic reviews as key sources of information on a variety of topics (1) because these provide rigorous assessments and analyses of data from different studies. Although rapid publication of relevant, high-quality systematic reviews is ideal, in practice the publication process has generally been slow (1,2). This is largely due to the massive amount of unstructured information that must be filtered. Synthesizing key information from multiple articles to create a systematic review requires considerable amounts of experts' time. Publication times often exceed one year (3), and costs run into hundreds of thousands of dollars (4).

Machine learning methods have previously been identified for the automation of systematic reviews (1,5). These allow for the development of software systems capable of automatically identifying distinct types of textual information, providing there are enough examples to learn to do this from. Natural language processing (NLP) methods have also been identified for the automation of systematic reviews (1,5) as these methods analyze written text through statistical and/or knowledge-based approaches, allowing for the identification of key items and patterns.

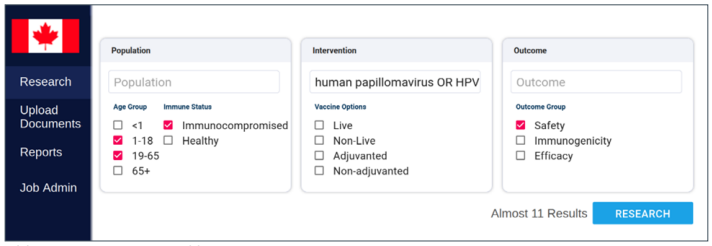
Background

The Public Health Agency of Canada has been examining the automation of aspects of evidence synthesis based on PICO principles (population, intervention, control, outcomes) to eliminate some of the barriers to obtaining systematic reviews results, namely, direct involvement of experts, time and cost. To do this, the Agency collaborated with Xtract AI (Vancouver, British Columbia) to develop a system that uses state-of-the-art machine learning and NLP to focus on extracting the PICO principles from immunization-specific articles. The functionality of our system is a result of its learning to review articles in a database composed of 249 immunization-specific articles with manually labelled PICO elements. Once the system's accuracy at extracting relevant PICO-related text from previously unseen articles is shown to be high, we can rely on it to carry out work

that would normally be completed manually. The aim was to develop the system so that far fewer articles need to be manually reviewed.

In this article, we use the terms "NLP model" and "system." Strictly speaking, the "NLP model" refers to the collection of machine learning and NLP methods used to process immunization-related documents. The "system" refers to the NLP model with a web-based user interface that allows easy use of the model (see **Figure 1**). Although the NLP model is the core of the system, performing text extraction and prediction tasks, we use the term "system" more often in this article due to the interdependencies of the NLP model and system components.

Figure 1: Screen capture of the homepage of the web interface of the NLP-based extraction system showing an example of a search query



Abbreviation: NLP, natural language processing

The system was designed to automatically extract 27 PICO-relevant classes of text from previously unseen immunization-specific articles. To test this initial version of the system we measured performance based on the system's ability to identify text about the main vaccine, the study type, the population health status and the outcome as well as the outcome's descriptive text from 40 previously unseen immunization-specific articles. We considered these five classes to be appropriate means of measuring the initial performance of the extraction of PICO-related text because they cover a wide range of texts and we suspected they were more varied than many of the other text classes the system can extract. **Table 1** shows the five main classes of text used to measure the performance of the system.

At the time of writing, the average accuracy of the system across these five key text classes was 88%. A summary of the accuracy results is shown in Table 1. To achieve this degree of accuracy, the system learned from 209 of the 249 examples in our document database. The automation task was tested using the remaining 40 examples in the document database. Two versions of these 40 test documents existed; the first version was labelled by an expert who scanned the documents for instances of the 27 text classes, and the other was unlabelled. To test the PICO-related text extraction capability of the system, the system processed the unlabelled versions of the test documents. We then compared the system's automated extractions to the text in the expert's labels. If the system extracted text that was comparable in quality to one of the expert's labelled texts, this was counted as a success that contributed to the 88% average accuracy score.

Table 1: Key text extraction classes with examples and accuracy scores

Text class	Description	Extracted example	Accuracy score (%)
Main vaccine	What vaccine is this article about?	Quadrivalent human papillomavirus vaccine, or heptavalent pneumococcal vaccine	90
Study type	What kind of study is this?	Randomized, placebo-controlled trial	92.5
Population health status	Population health status	HIV-positive	85
Outcome – adverse event	Any adverse outcome described in the article	"The most common adverse event was pain; other common events were neurological, gastrointestinal and skin related."	85
Outcome – description sentence	Any important sentence related to outcomes	Safety/ immunogenicity outcome, e.g. "VE was 93.0% (85.1–97.3) in the TVC-E (Table S1)."	87.5

As development work continues and more expert-labelled immunization-specific articles are added to the document database, we expect to be able to demonstrate similarly high accuracy scores when testing the PICO-related text extraction task on many more documents.

In this article we describe the technical approach to the development of the NLP model and the process by which the NLP model learned to perform this task. We then provide a more detailed analysis of accuracy using several performance measures.

Technical approach

The NLP model was designed as a multi-class sequence extraction model. A multi-class sequence extraction model works by processing the full text of a previously unseen document and then extracting sequences of text that correspond to each text class it learned to extract. In this case, based on the expert-labelled domains in 209 of our 249 immunization-specific documents, the system extracts up to 27 classes of text. Duplicates for each class may be included.

BioBERT, the biomedical language variant of Bidirectional Encoder Representations from Transformers (BERT), was used as a basis for the NLP model in order to increase performance (6,7).

BERT, a recent development in NLP (7), is essentially a model that has processed and learned from a massive corpus of text. BERT and variants like BioBERT are being increasingly used as the basis for new machine learning and NLP software systems. Their use has resulted in a considerable increase in the accuracy of these systems. BioBERT, as the biomedical language variant of the original BERT model, was considered more appropriate for use in this work.

Dataset creation

The initial learning data used for the system's NLP model was the evidence-based medicine NLP (EBM-NLP) corpus for PICO extraction. The EBM-NLP corpus contains 5,000 annotated abstracts of medical articles describing clinical randomized controlled trials (8). EBM-NLP corpus annotations labelled key parts of these abstracts, such as description of the participants (e.g. age range, condition), interventions (e.g. pharmacological) and outcomes (e.g. pain, adverse effects or mortality). These fields/classes of text were determined by the EBM-NLP corpus developers. A complete description of the annotation methodology used for the EBM-NLP corpus can be found at https://ebm-nlp.herokuapp.com/annotations.

We used the EBM-NLP corpus to "teach" our system to work with these fields because we needed to be able to extract the same fields/classes of text. However, because the EBM-NLP corpus annotates a small number of text fields that are not immunization-specific, we also generated the 249-document immunization-specific database by annotating data in-house. While the specific amount of time for including and labelling each article in the document database was not noted, adding and labelling a document would be equivalent to the time typically taken to manually review an article that might potentially be included in a systematic review. However, the effort at this stage will ultimately mean less human involvement in the overall systematic review process, as the system's accuracy should show that it is capable of reliably performing the task on its own.

Currently the model processes all types of documents, without pause, but returns nil results for those that do not contain any recognizable text similar to learned examples. We expect future versions of the system to be able to identify documents not related to immunization based on a lack of results. Using the BRAT annotation tool (9), we annotated the title, abstract, methods and results sections of the 249 articles in the new database. The articles were sourced from PubMed Central using a keyword search. We annotated the fields not included in the EBM-NLP corpus, that is, "study type," "main vaccine," "outcomes – adverse event," "outcomes – description sentence" and "population health status." The NLP model understands these text classes because they were manually labelled in the new immunization-specific document database. Examples of these text classes are listed in Table 1.

System learning process/accuracy testing

The BioBERT-based NLP model was initialized from the original BioBERT model, with subsequent learning using the EBM-NLP corpus. The model then completed more learning using our immunization-specific document database. While 209 samples from the document database were used during this final learning stage, 40 labelled articles were excluded in order to test the system's performance.

Although we focused on testing five key text classes in this article, the system only needs to extract and output a small amount of simple text to test several of the 27 extractable text classes (see **Table 2** for examples). For the remaining classes, the system does not output the extracted text on finding it but rather flags the article as "true" or "false" depending on the content of the research article. The EBM-NLP corpus used in the initial training of the NLP model contained examples for some of these classes. These examples were expected to allow the NLP model to identify anything included in the EBM-NLP corpus accurately.

Table 2: Additional text extraction classes with examples[a]

Text Class[b]	Extracted example
Safety	True
Efficacy	True
Pharmacological[b]	Quadrivalent human papillomavirus (types 6, 11, 16 and 18) recombinant vaccine 0.5 mL intramuscularly
Condition	HIV-infected
Country	Mali
Age	Adults aged 27 years or older
Sample size	535
Sex	Women

[a] The system also extracts "vaccine_pathogen_target_main," "adjuvanted," "immunogenicity," "immunocompromised," "healthy," "non-live," "non-adjuvanted," "live," "sex," "pregnancy," "doi," "score," "abstract," "methods," "results," "ai_version" and "keywords," but these are not listed here for reasons of brevity

[b] For some text classes, e.g. pharmacological, the system was shown many thousand learning examples via the evidence-based medicine natural language processing corpus

Evaluation

To evaluate the prediction performance for the five key text classes, we computed common measures of performance for machine learning and NLP-based systems, namely, precision, recall and F1 score (another measure of accuracy based on precision and recall). We further computed the number of successes, errors and general accuracy percentage. All performance measures are defined in **Table 3**. The measures listed apply to entire documents. The number of true positives, true negatives, false positives and false negatives are used to compute the successes and errors. Those numbers are the result of the degree of correctness of text extractions made by the

system. The general accuracy percentages shown in Table 1 are calculated based on the successes and errors.

Table 3: Definitions of model performance measures

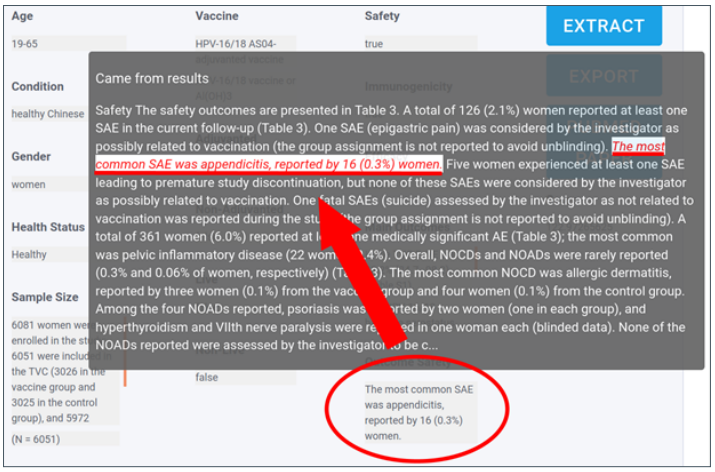
Measure of accuracy	Means or formula of calculation	Meaning of results
TP	Number of documents	Correctly identified documents Higher scores are better
TN	Number of documents	Correctly identified documents Higher scores are better
FP	Number of documents	Documents incorrectly identified Lower scores are better
FN	Number of documents	Documents incorrectly identified Lower scores are better
Successes	$TP + TN$	The sum of correctly identified documents Higher scores are better
Errors	$FP + FN$	The sum of incorrectly identified documents Lower scores are better
General accuracy percentage	$\frac{(TP + TN)}{(TP + TN + FP + FN)} * 100$	Overall accuracy from 0% to 100%
Precision	$\frac{TP}{(TP + FP)}$	Measure of document retrieval Scores from 0.0 to 1.0, the higher the better
Recall	$\frac{TP}{(FN + TP)}$	Measure of document retrieval Scores from 0.0 to 1.0, the higher the better
F1 score	2*(P*R) / (P+R)	Scores from 0.0 to 1.0, the higher the better

Abbreviations: FN, false negative; FP, false positive; P, precision; R, recall; TN, true negative; TP, true positive

It is important to note that the system extracts free-form text, where the length and content of an extracted prediction can vary greatly from the correct, labelled sequence of text on a test document (for examples, see **Figure 2**). This being the case, it is very important to clearly define what constitutes a success (true positive or true negative). For example, if the extracted prediction is "pneumococcal vaccine" while the correct answer is "heptavalent pneumococcal vaccine," the result may be classified as an error if "heptavalent" is deemed to be too important to be left out of the "main vaccine" text for this document.

If the problems associated with definitions are not addressed, the performance measures have no context. In this initial stage, the extracted predictions were manually inspected for accuracy based on defined criteria. Although not yet completed at the time of writing, we expect the application of accuracy criteria to extracted predictions to be automated in future versions of the system. These accuracy criteria were imposed upon the five key text classes for an extracted prediction to count as a predicted true positive (PTP). Since a document can have many extracted predictions, a high number of correct PTPs are needed for the document classification task to be counted as a complete success, or true positive. The easiest way to think about a PTP compared to a true positive is that a PTP is at the text level while a true or false positive is assigned to an entire document based on the number of PTPs. We define the general criteria for this in **Table 4**.

Figure 2: Screen capture of the NLP model web interface showing an extracted prediction[a]



Abbreviation: NLP, natural language processing
[a] The extracted prediction is underlined in the web interface and circled on the main results panel of the web interface

Table 4: Custom criteria for predicted true positives and the formulas used to calculate true positives and true negatives

PTP criteria 1[a]	PTP criteria 2[a]	Requirement for a success (true positive)[b,c]	Requirement for a success (true negative)[b,c]
The EP contains one or more of the labelled answers OR unlabelled but correct answer(s), including all important information	The EP cannot have too much unnecessary information	$\frac{PTP}{(PTP+ FPF + PFN)} > 0.8$ >0.5 for Outcome text classes due to subjectivity	PTP+PFP+PFN = 0

Abbreviations: EP, extracted prediction; PFN, predicted false negative; PFP, predicted false positive; PTP, predicted true positive
[a] PTPs are determined based on the extracted predictions made by the system from the document text after evaluation by an expert reviewer. There could be many extracted predictions per text class that might count as PTPs or otherwise
[b] PFPs and PFNs are equally important but the numbers of each may vary depending on the article
[c] If the requirements for a true positive or true negative are not met, a false positive may be assigned where PFP > PFN or a false negative may be assigned where PFP < PFN

The full results of the 40-document accuracy testing are shown in **Table 5**.

Table 5: All performance results for five key classes

Performance measure	Key text classes				
	Main vaccine	Study type	Outcome - adverse event[a]	Outcome - description sentence[a]	Population - health status
F1 score[b]	0.8824	0.947	0.727	0.9315	0.75
Precision[c]	1	0.964	1	0.9444	0.9
Recall[d]	0.7895	0.931	0.571	0.9189	0.643
TP[e]	15	27	4	34	9
TN[e]	21	10	33	1	25
FP[f]	0	1	0	2	1
FN[f]	4	2	3	3	5
Successes (TP or TN)[g]	36	37	37	35	34
Errors (FP or FN)[h]	4	3	3	5	6
Accuracy percentage for class, %[i]	90	92.5	92.5	87.5	85

Abbreviations: FN, false negative; FP, false positive; P, precision; R, recall; TN, true negative; TP, true positive
[a] Text classes had noticeably imbalanced positive and negative examples. Overall accuracy may be skewed in favour of the group with the greater number of examples. However, an imbalance between examples may also occur in real-world data
[b] Another measure of accuracy based on precision and recall. Scores range from 0.0 to 1.0, the higher the better
[c] Scores range from 0.0 to 1.0, the higher the better
[d] Scores range from 0.0 to 1.0, the higher the better
[e] A measure of correctly identified documents. Higher scores are better
[f] A measure of incorrectly identified documents. Lower scores are better
[g] A measure of the sum of correctly identified documents. Higher scores are better
[h] The sum of incorrectly identified documents. Lower scores are better
[i] Overall accuracy, with scores ranging from 0% to 100%

The system consistently performed well. The success rate was high and the error rate low, which demonstrates overall effectiveness at the PICO extraction task. A balance of both positive and negative test examples was not possible for every text class due to limited data, although a balance may not necessarily reflect real-world performance. For instance, there were many more true negatives for "population – health status" because the articles did not contain any text that could be extracted for this class. Regardless, one issue resulting from this imbalance is that the accuracy scores for these text classes may be skewed in favour of the group (either positive or negative) with more test examples. However, we expect that scores will remain high as this issue is addressed through the expansion of the immunization-specific documents database.

As shown in Figure 1, PICO-related extraction results are accessible through a user-friendly web interface. **Figure 3** shows an example of a completed search displaying results for many of the text classes.

System limitations

As previously stated, a balance of positive and negative test example groups for all text classes was not possible due to limited data. This may skew the accuracy scores in favour of the group with the higher number of test examples. However, it is important to note that there may be an imbalance between these positive and negative examples on unseen documents in real-world situations.

Figure 3: Example extraction results for HPV after submitting search terms

Developing the new immunization-specific document database required some involvement by experts, that is, it was not automated. There was also some manual effort in reviewing extracted predictions from the document text for correctness. This early manual effort is ultimately required to enable automation later.

Next steps

At the time of writing, the system was still being developed. Future work will include increasing the number of labelled documents in the new immunization-specific document database to improve system learning. The web interface will also continue to be refined. Ideally, the system will identify documents that are not related to immunization and stop processing them immediately to prevent even the brief delay that is currently needed to scan a text. A related system, designed to encompass all biomedical literature (based on the same technology in this article), is also being developed.

Finally, the effectiveness of a more complete system will need to be tested in consultation with public health decision makers.

Conclusion

We described a system based on machine learning and NLP methods for automating the repetitive manual work of analyzing documents that is part of the systematic review process. This system focuses on immunization-specific documents only. The promising performance results in this initial work demonstrate that there is potential to move away from the manual and laborious approaches of systematic reviews and move towards

automated systems, in an effort to eventually eliminate (or significantly reduce) expert involvement in the repetitious tasks of the process.

The system's overall design presents a promising way for public health decision makers to utilize unstructured data more quickly and economically when making policy decisions and applying the principles of evidence-based medicine. Our unique contribution to this area is the system's ease of use via the straightforward web interface combined with the performance resulting from the application of state-of-the-art machine learning and NLP methods on our new immunization-specific document database.

Authors' statement

DB — System design/implementation, testing, writing, review, editing
JG — System design/implementation, testing, writing, review, editing
BI — System design/implementation, testing, writing, review, editing
CB — Testing requirements, writing, review, editing

Conflicts of interest

None.

Acknowledgements

The authors would like to acknowledge the contribution of Xtract AI and the following people from the Public Health Agency of Canada: O Baclic, M Tunis, M Laplante, K Young, H Swerdfeger, C Doan.

Funding

This work was supported by the Public Health Agency of Canada.

References

1. Tsafnat G, Glasziou P, Choong MK, Dunn A, Galgani F, Coiera E. Systematic review automation technologies. Syst Rev 2014;3(74):1-15. DOI PubMed

2. Jonnalagadda SR, Goyal P, Huffman MD. Automating data extraction in systematic reviews: a systematic review. Syst Rev 2015;4(78):1-16. DOI PubMed

3. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. BMJ Open 2017;7(2):e012545. DOI PubMed

4. Lau J. Editorial: systematic review automation thematic series. Syst Rev 2019;8(70):1-2. DOI PubMed

5. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. BMC Syst Rev 2019;8(163):1-10 DOI PubMed

6. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb;36(4):1234–40. DOI PubMed

7. Devlin J, Chang MW, Lee K, Toutanova K. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers); BERT: pre-training of deep bidirectional transformers for language understanding; 2019-06. NAACL-HLT. Minneapolis, Minnesota (US): Association for Computational linguistics: 2019;4171–86. DOI

8. Nye B, Li JJ, Patel R, Yang Y, Marshall I, Nenkova A, Wallace B. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics; A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. Melbourne (AUS): Association for Computational linguistics; July 15-20, 2018. DOI

9. Stenetorp P, Pyysalo S, Topíc G, Ohta T, Ananiadou S, Tsujii J. BRAT: A web-based tool for NLP-assisted text annotation. Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. 2012 April. Avignon (FR): EACL.