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

doi:10.3233/SHTI220076

Extending PICO with Observation Normalization for Evidence Computing

Ali Turfah[a], Hao Liu[a], Latoya A Stewart[b], Tian Kang[a], Chunhua Weng[a]

[a] *Department of Biomedical Informatics, Columbia University, New York, New York, U.S.A,*
[b] *Vagelos College of Physicians and Surgeons, Columbia University, New York, New York, U.S.A*

Abstract

While the PICO framework is widely used by clinicians for clinical question formulation when querying the medical literature, it does not have the expressiveness to explicitly capture medical findings based on any standard. In addition, findings extracted from the literature are represented as free-text, which is not amenable to computation. This research extends the PICO framework with Observation elements, which capture the observed effect that an Intervention has on an Outcome, forming Intervention-Observation-Outcome triplets. In addition, we present a framework to normalize Observation elements with respect to their significance and the direction of the effect, as well as a rule-based approach to perform the normalization of these attributes. Our method achieves macro-averaged F1 scores of 0.82 and 0.73 for identifying the significance and direction attributes, respectively.

Keywords

Natural Language Processing, Evidence-based Medicine, Text Mining

Introduction

Evidence-Based Medicine (EBM) is the practice of using the best evidence to make decisions about the care of patients [1]. Unfortunately the large and ever-increasing amount of information available in the medical literature, the time required to read and synthesize the findings presented, and the hectic nature of clinical schedules are barriers to implementing EBM [2; 3]. Thus, it is imperative to improve the efficiency of the EBM-driven decision-making process, especially the processes of searching for and identifying the findings in medical literature.

The PICO framework has been widely adopted to explicitly formulate clinical questions and facilitate EBM [4]. PICO is an acronym for the components of the framework, which identifies the Population, Intervention, Comparator, and Outcome of interest.

Much of the prior work using PICO has focused on incorporating PICO into document retrieval [5-7] and automatic annotation of PICO elements in medical literature [8; 9]. PICO elements are further mapped to medical concepts using standardized vocabularies such as UMLS [10] and OMOP [11]. Recent work has been made to normalize composite treatments, such as HemOnc [12] to capture chemotherapy regimen information, as well as efforts to capture more general medication information such as dosage and frequency [13; 14]. The normalization of these more nuanced medical details is necessary to enhance the expressiveness and completeness of medical literature mining, as well as to enhance the capacities of the computational systems that rely on this information. While these combined efforts greatly improve the process of identifying relevant medical evidence to a clinician's queries, they do not address the time-consuming process of reading and understanding the literature.

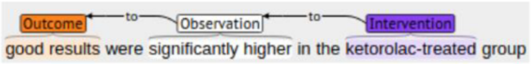
There have been both manual efforts—such as UptoDate [15] and the Cochrane Collaboration [16]—as well as automated efforts [17; 18] to capture and summarize the findings in medical literature; however they do not go beyond presenting the findings as free text. While this presentation is conducive to a human reader trying to make sense of a single article, it is not as useful for automated evidence synthesis and summarization of larger bodies of literature—as there is no standard representation of the findings these articles present.

This research seeks to leverage the widespread use of PICO in order to represent and normalize the findings in the medical literature for use in medical evidence computation and synthesis. By adding Observation elements to the PICO framework—which capture the relationship between the treatments and the Outcome measure—we enable it to represent medical findings as well as questions. As we have identified no prior work addressing this issue in this fashion, we also propose a normalization scheme to extract and normalize two attributes from the Observations: significance and the direction of the observed effect. **Our contributions** are two-fold: (1) we extend the PICO framework to include Observation elements to facilitate evidence computing tasks and (2) we implement and evaluate a framework to normalize Observation elements with respect to significance and the direction of the observed effect.

Methods

Observation Elements and Attributes

For the purposes of this study, we combine Intervention and Comparator classes, and will refer to them simply as Interventions. As described above, Observation elements represent the observed relationship between Interventions and the Outcome measure. For example, Figure 1 presents related Intervention, Observation, and Outcome elements from a snippet of text which form a medical finding. In this case, the treatment significantly increased the prevalence of "good results." We formulate "findings" as the combination of related Intervention, Observation, and Outcome elements in the text.



good results were significantly higher in the ketorolac-treated group

Figure 1– Example annotation of related Intervention, Observation, and Outcome elements [PID: 9001833]

For the Observation elements in these findings, we seek to categorize them with respect to two attributes: Significance and Clinical / Measurement Direction.

Significance

The Significance attribute of an Observation, as the name would suggest, captures if the finding in question was statistically significant. It takes one of three possible values: TRUE, FALSE, and N/A. TRUE/FALSE values for significance are used for significant/not significant findings respectively, while N/A is used for cases where no indication of significance is made. Examples from these three classes are included below, with the relevant context to determine the class label underlined.

- TRUE: "Bone-specific alkaline phosphatase decreased ($p < 0.05$) …" [PID: 10993031]
- FALSE: "No significant difference was observed after 6 and 12 months of treatment in PEF variability" [PID: 10741095]
- N/A: "PEF variability improved in BDP and BDP + S groups" [PID: 10741095]

Clinical/Measurement Direction

The Clinical/Measurement Direction (referred to henceforth simply as Direction) captures the effect of the Intervention on the Outcome as presented in the text. It takes one of four possible values: UP, DOWN, CHANGE, UNKNOWN. Explanations of these categories as well as some examples are included below, with the relevant context to determine the class label underlined.

- UP: The UP category captures effects that are presented in terms of a positive numeric (e.g. an increase in some value) or clinical (e.g. an improvement in the patient's condition) effect
  - "good results were significantly higher in the ketorolac-treated group" [PID: 9001833]
  - "PEF variability improved in BDP and BDP + S groups" [PID: 10741095]
- DOWN: The DOWN category captures the effects that are presented in terms of a negative numeric (e.g. a decrease in some value) or clinical (e.g. a worsening of the patient's condition) effect
  - "Bone-specific alkaline phosphatase decreased ($p < 0.05$) …" [PID: 10993031]
  - "statistically significant , but clinically small , impairment of memory" [PID: 11205419]
- CHANGE: The CHANGE category captures effects that are presented as differences between the groups in question, without specifying the nature of that difference
  - "The difference in the median annual change between the two groups was significant ($P=0.013$)" [PID: 10385063]
- UNKNOWN: The UNKNOWN category captures effects where the direction of the effect is not apparent from the text
  - "in a Cox model of overall survival , but the effect of cisplatin was not significant" [PID: 8918486]

Normalization Process

With the framework above, we present and evaluate the following rule-based method to determine the values for the different attributes. There are separate processes for the categorization of an Observation with respect to these attributes, the details of which can be found below.

Significance Normalization

In order to determine the value for the Significance attribute, we first employ regular expressions to scan the Observation and its immediate surrounding context, for indicators of significance such as explicit wording or p-values. If no indication of significance is detected, then the Observation is assigned to the N/A category.

If an indication of significance is present, then the Observation is checked for negations (e.g., "no significant difference") if significance was indicated via explicit wording, and the TRUE/FALSE value is assigned accordingly. Alternatively, if significance was indicated with a p-value, then the process to determine significance is as follows. If the computed p-value is provided in the text (e.g., $p = 0.0013$), then we compare its value against the threshold of 0.05 to determine significance. If the p-value is given as greater than some value (e.g., $p > 0.05$), then the Observation is marked as non-significant and assigned the value FALSE for significance. Alternatively, if the p-value is presented as less than some value (e.g., $p < 0.01$) then the Observation is assigned the value TRUE.

Direction Normalization

We employ a string-matching approach to determine the Direction of a given Observation, using synonyms to improve coverage without having to specify every possible term. First, the Observation and its immediate context is scanned for negations (e.g. no significant difference), and if one is found the Observation is labelled UNKNOWN. Next, given a set of trigger words for the TRUE, FALSE, and CHANGE classes, we scan the Observation for matches to these terms and assign the observation to the corresponding category. For example, the UP class set of trigger words may include terms such as "increase" or "improve," whereas the DOWN class trigger words would include "decrease" or "deteriorate."

If no exact match is found, we then repeat the exact match search using synonyms of the terms in the trigger word sets. Word synonyms are determined using WordNet [19], and terms in the text are transformed to the form present in the WordNet database when possible. We use the WordNet implementation from the python NLTK package [20].

Experiment Design and Evaluation

For the training set, 211 abstracts were pulled from PubMed and manually annotated for their PICO elements using the BRAT annotation tool [21]. We then generated "findings" specified by Intervention-Observation-Outcome triplets as defined above; we took a sample of 150 of such triplets and manually labeled them for their Significance and Direction according to the class definitions provided above. For the testing set, we pulled an additional 50 abstracts and followed the same process to generate the "findings" presented in them. We randomly selected 150 of these triplets and once again manually labeled them with respect to their Significance and Direction for use as a testing set. Two annotators determined the Significance and Direction; a sample of 20 "findings" were used to establish inter-annotator agreement between two annotators. After this step, one annotator determined the labels for the training set while the other labeled the testing set. A

Table 1– Attribute statistics for the Observations in the two datasets

		Significance Attribute			Direction Attribute			
Dataset	**# Observations**	**# TRUE**	**# FALSE**	**# N/A**	**# UP**	**# DOWN**	**# CHANGE**	**# UNKNOWN**
Train	150	44	15	91	58	42	10	40
Test	150	43	40	60	44	41	21	44

breakdown of the class distributions for the two attributes across the datasets can be found in Table 1. The training set was used to determine the Significance trigger phrases, as well as the candidate set of words for the Direction classes. We then evaluated the model using the testing set.

We formulate the task of determining the value for each attribute as a multi-class classification problem, and present the class-specific precision, recall, and F1 scores in Tables 2 and 3. In order to have an idea of the overall quality of the normalization performance, we also present the micro and macro-averaged values for each measure of interest, as the attributes have varying degrees of class imbalance across the datasets.

Results

Significance

The performance of the Significance attribute normalization process outlined above can be found in Table 2. Our approach achieves a macro-averaged F1 score of 0.82 on the testing set for this task. The F1 scores for the TRUE, FALSE, and N/A classes are 0.81, 0.78, and 0.85 respectively. While the performance metrics tend to be similar across the training and test sets, there is an observed degradation (-0.12) in the FALSE category's Recall as well as the N/A category's Precision (-0.11).

Clinical/Measurement Direction

The performance of the Direction normalization process outlined in the earlier section can be found in Table 3. Our approach achieves an overall macro-averaged F1 score of 0.73 on this task. The macro-averaged F1 scores for the UP, DOWN, CHANGE, and UNKNOWN Direction classes were 0.83, 0.81, 0.62, and 0.65 respectively. Between the training and testing dataset we see drops in the recall of the UP class (-0.16), as well as improvements in the CHANGE class's precision (+0.22).

Discussion

Error Analysis

Errors in determining the Significance come predominantly from limitations in the annotation of Observations rather than limitations of the logic employed, and originate from disconnected spans of text comprising the 'full' Observation. Consider the examples below (the Intervention is italicized, the Outcome is underlined, and the annotated Observation is in bold)

- Significant Doppler flow **improvement** was obtained in the *L-arginine supplemented group* [PID: 10402369]
- In the CG, *soccer training* caused an **improvement of smaller magnitude** in 10m and shooting speed ($p < 0.05$). [PID: 30431535]

In both cases, the Direction component of the Observation is separate from the Significance component, but the annotation can only capture one of them. In the first example, due to the wording of the sentence the indicator of significance is separated from the Observation by the Outcome. In the second example the significance indicator is a p-value, but it is presented at the end of the sentence. For cases like the latter one it may be possible to have an additional step to assign un-marked p-values to the closest observation, but further analysis would be necessary to determine how this affects performance.

With respect to the observed drop in performance in the proposed rule-based method's on the FALSE category, we attribute this decrease in performance to an increase in the prevalence of cases similar to the ones mentioned above in the testing dataset. A performance shift can be expected as the training set only contains 15 FALSE Observations whereas the testing set contains 40 of them, bringing the performance more in line with the TRUE category.

Errors in determining the Direction attribute arise both from the aforementioned limitation in the annotation as well as the lack of coverage in the class-specific trigger words. For example, the testing set has terms such as "alleviate" and "prolong," which are out of the scope of the synonyms generated from the training set. As mentioned in the results section, we observe an increase in the performance on the CHANGE class's precision. Our proposed method appears to have had difficulty differentiating CHANGE and UNKNOWN Observations; in the testing set 6/18 predicted CHANGE Observations were in fact UNKNOWN, and 8/21 of the CHANGE Observations were predicted to be UNKNOWN. The magnitude of the change in performance may also be due to the limited number of CHANGE examples in the original training set.

Table 2– The proposed method's performance for Significance classification

	Train				Test			
Class	**Precision**	**Recall**	**F1**	**Supp.**	**Precision**	**Recall**	**F1**	**Supp.**
TRUE	1.00	0.70	0.83	44	0.97	0.70	0.81	43
FALSE	1.00	0.80	0.89	15	0.93	0.68	0.78	40
N/A	0.85	1.00	0.92	91	0.74	1.00	0.85	67
Macro-Avg.	0.95	0.83	0.88	--	0.88	0.79	0.82	--
Micro-Avg.	0.91	0.89	0.89	--	0.86	0.82	0.82	--

Table 3– The proposed method's performance for Direction classification

	Train				Test			
Class	**Precision**	**Recall**	**F1**	**Supp.**	**Precision**	**Recall**	**F1**	**Supp.**
UP	0.91	0.91	0.91	58	0.92	0.75	0.83	44
DOWN	0.97	0.78	0.87	42	1.00	0.68	0.81	41
CHANGE	0.45	0.50	0.48	10	0.67	0.57	0.62	21
UNKNOWN	0.64	0.75	0.69	40	0.54	0.84	0.65	44
Macro-Avg.	0.74	0.74	0.74	--	0.75	0.71	0.73	--
Micro-Avg.	0.82	0.81	0.81	--	0.76	0.73	0.74	--

Limitations and Extensions

We recognize the limitations of our simple framework. Primarily, there is no distinction made between clinical findings and numerical ones. This is particularly relevant in cases where a clinical increase corresponds to a numeric decrease (e.g., blood pressure in hypertension patients). An improvement (UP) in such a patient's blood pressure corresponds to a decrease (DOWN) in the numeric value; assigning it to any of the UP/DOWN Direction categories results in a dissonance with the other type of finding. This issue also arises in the example provided earlier which describes improvements in PEF variability; an improvement actually corresponds to a decrease in some numeric value.

There is some ambiguity in this framework when it comes to representing negative findings—findings that report the absence of an effect. Consider an intervention that is reported to operate "without increasing the rate of complications." This finding would fall into the UNKNOWN category defined above because the actual effect on the outcome, if any, is not specified. However, it may be meaningful to capture the fact that this finding is presented in terms of an "increase in complications," even if it presents the absence of an increase. In the case of a non-inferiority trial, this type of negative finding would be one of the main conclusions to draw from the text. For cases like this it may be better—for interpretation—to split up the Direction attribute into sub-components that capture the direction presented in the finding along with a negation.

We recommend further work on synthesizing propositions that operate on the same Intervention and Outcome to have a more complete understanding of the effect presented in the text. In the example below, the presented finding is a significant increase in the levels of peripheral leukocytes and lymphocytes.

- "Over a 24-month observation period the immunized group always had **higher levels** of peripheral leukocytes and peripheral lymphocytes ; this **difference was significant** for the first 21 months." [PID: 8908288]

It is clear to a human reader that the "difference" specified is in fact an increase, however with our current approach there would be two extracted findings (the observations for which are bolded): an increase of unspecified significance, and a significant change in the levels of the leukocytes and lymphocytes.

With respect to the normalization process, we recommend future work explore more robust statistical or computational approaches, such as neural networks, to determine the Direction attribute. While we do offset some of the brittleness of using a manually specified list of words by enriching it with synonyms, such an approach is not scalable to deal with the many possible ways to describe findings in medical literature.

Use Cases for Normalization

Suppose that a clinician is looking to compare the effectiveness of different interventions with respect to some outcome measure. With a PICO-based search, at best this provides them a list of potentially relevant studies that they would need to review manually. While available summaries of the relevant documents would help speed up this process, they do not address the issue of needing to sift through a potentially large volume of relevant articles. Making use of our approach to normalize findings, it would be possible to generate summary views of the literature as a whole, indicating for example that of the 15 relevant documents retrieved, 12 of them indicate a significant positive effect of the intervention on the outcome while the remaining three do not have significant findings. While the simplicity of the current framework somewhat limits its expressivity, it can assist in streamlining the evidence review portion of EBM.

This framework has uses beyond simply summarizing the literature. With a standard representation of "findings," it is possible to detect when there is a contradiction in the results of different studies. In addition, it has been noted that abstracts may report findings in a more positive light than in the main text of the article [22]; our approach can be extended to full-text articles to automatically identify such instances where the findings of a study are "spun" differently in the Abstract.

Conclusions

In this study, we propose a novel extension to the PICO framework: Observation elements which capture the effect of the Intervention on the Outcome, in order to extend the framework to more explicitly represent findings in the literature. In addition, we propose and evaluate a rule-based method to normalize these elements by capturing information about significance and the direction of the reported effect. We are able to achieve F1 scores of 0.82 and 0.73 for determining the Significance and Direction of medical findings respectively.

Acknowledgments

This work was supported by 5R01LM009886-11 (Bridging the semantic gap between research eligibility criteria and clinical data; PI: Weng). We thank Mathew Spotnitz, Konstantin Stojanovic, and Anna Ostropolets for helpful discussions.

References

[1] D.L. Sackett, W.M. Rosenberg, J.M. Gray, R.B. Haynes, and W.S. Richardson, Evidence based medicine: what it is and what it isn't, in, British Medical Journal Publishing Group, 1996.

[2] 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** (2010), e1000326.

[3] J.W. Ely, J.A. Osheroff, M.H. Ebell, G.R. Bergus, B.T. Levy, M.L. Chambliss, and E.R. Evans, Analysis of questions asked by family doctors regarding patient care, *Bmj* **319** (1999), 358-361.

[4] S.A. Miller and J.L. Forrest, Enhancing your practice through evidence-based decision making: PICO, learning how to ask good questions, *Journal of Evidence Based Dental Practice* **1** (2001), 136-141.

[5] D. Demner-Fushman and J. Lin, Answering clinical questions with knowledge-based and statistical techniques, *Computational Linguistics* **33** (2007), 63-103.

[6] E. Znaidi, L. Tamine, and C. Latiri, Answering PICO clinical questions: A semantic graph-based approach, in: *Conference on Artificial Intelligence in Medicine in Europe*, Springer, 2015, pp. 232-237.

[7] F. Boudin, J.-Y. Nie, and M. Dawes, Clinical information retrieval using document and PICO structure, in: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2010, pp. 822-830.

[8] B.C. Wallace, J. Kuiper, A. Sharma, M. Zhu, and I.J. Marshall, Extracting PICO sentences from clinical trial reports using supervised distant supervision, *The Journal of Machine Learning Research* **17** (2016), 4572-4596.

[9] T. Kang, S. Zou, and C. Weng, Pretraining to recognize PICO elements from randomized controlled trial literature, *Studies in health technology and informatics* **264** (2019), 188.

[10] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, *Nucleic acids research* **32** (2004), D267-D270.

[11] J.M. Overhage, P.B. Ryan, C.G. Reich, A.G. Hartzema, and P.E. Stang, Validation of a common data model for active safety surveillance research, *Journal of the American Medical Informatics Association* **19** (2012), 54-60.

[12] J.L. Warner, D. Dymshyts, C.G. Reich, M.J. Gurley, H. Hochheiser, Z.H. Moldwin, R. Belenkaya, A.E. Williams, and P.C. Yang, HemOnc: A new standard vocabulary for chemotherapy regimen representation in the OMOP common data model, *Journal of biomedical informatics* **96** (2019), 103239.

[13] H. Xu, S.P. Stenner, S. Doan, K.B. Johnson, L.R. Waitman, and J.C. Denny, MedEx: a medication information extraction system for clinical narratives, *Journal of the American Medical Informatics Association* **17** (2010), 19-24.

[14] S. Sohn, C. Clark, S.R. Halgrim, S.P. Murphy, C.G. Chute, and H. Liu, MedXN: an open source medication extraction and normalization tool for clinical text, *Journal of the American Medical Informatics Association* **21** (2014), 858-865.

[15] T.W. Post, UpToDate: Evidence-based Clinical Decision Support, in, 2021.

[16] I. Chalmers, The Cochrane collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care, *Annals of the New York Academy of Sciences* **703** (1993), 156-165.

[17] X. Zhang, P. Geng, T. Zhang, Q. Lu, P. Gao, and J. Mei, Aceso: PICO-guided Evidence Summarization on Medical Literature, *IEEE journal of biomedical and health informatics* **24** (2020), 2663-2670.

[18] M. Fiszman, D. Demner-Fushman, H. Kilicoglu, and T.C. Rindflesch, Automatic summarization of MEDLINE citations for evidence-based medical treatment: a topic-oriented evaluation, *Journal of biomedical informatics* **42** (2009), 801-813.

[19] Princeton University, About WordNet, in, Princeton University, 2010.

[20] S. Bird, E. Loper, and E. Klein, Natural Language Processing with Python, *O'Reilly Media Inc* (2009).

[21] P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J.i. Tsujii, BRAT: a web-based tool for NLP-assisted text annotation, in: *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 2012, pp. 102-107.

[22] I. Boutron and P. Ravaud, Misrepresentation and distortion of research in biomedical literature, *Proceedings of the National Academy of Sciences* **115** (2018), 2613-2619.

Address for correspondence

Chunhua Weng, chunhua@columbia.edu. Department of Biomedical Informatics, Columbia University. PH-20, 622 W 168 ST, PH-20, New York, NY 10032, USA