Integrating the Framing of Clinical Questions via PICO into the Retrieval of Medical Literature for Systematic Reviews

Harrisen Scells
Queensland University of Technology
Brisbane, Australia
harrisen.scells@hdr.qut.edu.au

Guido Zuccon
Queensland University of Technology
Brisbane, Australia
g.zucoon@qut.edu.au

Bevan Koopman
CSIRO
Brisbane, Australia
bevan.koopman@csiro.au

Anthony Deacon
Queensland University of Technology
Brisbane, Australia
aj.deacon@qut.edu.au

Leif Azzopardi
Strathclyde University
Glasgow, Scotland
leif.azzopardi@strath.ac.uk

Shlomo Geva
Queensland University of Technology
Brisbane, Australia
s.geva@qut.edu.au

ABSTRACT

The PICO process is a technique used in evidence based practice to frame and answer clinical questions. It involves structuring the question around four types of clinical information: **P**opulation, **I**ntervention, **C**ontrol or comparison and **O**utcome. The PICO framework is used extensively in the compilation of systematic reviews as the means of framing research questions. However, when a search strategy (comprising of a large Boolean query) is formulated to retrieve studies for inclusion in the review, PICO is often ignored. This paper evaluates how PICO annotations can be applied and integrated into retrieval to improve the screening of studies for inclusion in systematic reviews. The task is to increase precision while maintaining the high level of recall essential to ensure systematic reviews are representative and unbiased. Our results show that restricting the search strategies to match studies using PICO annotations improves precision, however recall is slightly reduced, when compared to the non-PICO baseline. This can lead to both time and cost savings when compiling systematic reviews.

1 INTRODUCTION

A systematic review is a type of literature review that critically analyses multiple research studies in order to answer a specific research question. They play a key role in evidence based medicine, informing practice and policy. Systematic reviews are also increasingly important for informing policy and practice outside of academia.

To answer a given research question, e.g., should beta blockers be given to heart attack survivors?, a systematic review of the literature is performed by composing a search strategy. A search strategy is first formulated to retrieve studies that satisfy inclusion and exclusion criteria set within the review's protocol. Systematic reviews intended for a biomedical audience use Boolean queries as search strategies. These often include restrictions to studies that match specific metadata like the type of studies or dates of publication.

CIKM'17, November 6–10, 2017, Singapore.
 ISBN 978-1-4503-4918-5/17/11.
DOI: 10.1145/3132847.3133080

The retrieved studies are then manually screened: assessed for inclusion in the systematic review (akin to relevance assessment). This screening process is often costly and time consuming due to numerous not-relevant studies retrieved and the effort of manually screening each study. In extreme cases, like for *scoping* systematic reviews, millions of studies may be retrieved and require screening, with only few thousands meeting the inclusion criteria; for example, a systematic review that required the screening of 1.8 million studies, of which approximately only 4,000 were found to be potentially eligible (precision=0.0022) [12].

A key question is whether some part of the retrieval and screening process could be automated [8] while ensuring that quality if not compromised? The aim is to reduce the total amount of effort, without missing any study that are eligible for inclusion in the review. In information retrieval terms, this translates in increasing the precision of the search results, while maintaining recall.

In this paper we evaluated the use of PICO annotations at both search strategy (query) and study (document) level as a method to restrict the number of retrieved studies for a given systematic review, while attempting to maintain the same recall achieved by the original search strategy (without PICO annotations). PICO is a popular framework used in medicine and clinical practice to formulate clinical questions. The framework promotes framing clinical questions according the four types of clinical information: i) *population* the question refers to (e.g., males aged 20-50); ii) *intervention* the population is administered with (e.g., weight loss drug); iii) the criteria for *comparison or control* (e.g., controlled exercise regime); and iv) *outcome* to measure (e.g., weight loss). Systematic review guidelines recommend the use of PICO in the development of search strategies (queries), and most modern systematic reviews in medicine adhere to this guideline. While PICO is commonly used to formulate the search strategy, it is not used by the underlying retrieval system executing that strategy. This paper aims to highlight whether it should be.

2 RELATED WORK

There has been increasing attention on computational methods to automate the processes of compiling systematic reviews [2, 5–7]. Automated classification methods have been studied to filter out non relevant studies once retrieved by the Boolean search strategy defined in a review.

There have been attempts in modifying the mechanisms used for retrieval of research studies for directly reducing the number of retrieved studies, rather than devising methods that filter studies post-retrieval. Karimi et al. applied query expansion to improve the quality of retrieved studies [4]. This approach improved recall, but hampered precision; thus, its application to actual systematic review screening is often unsustainable for very large systematic reviews. A randomised control trial study by Schardt et al. compared use of PICO against no standardised framework to guide the formulation of search strategies. The study found the use of PICO tended to retrieve results for screening with greater precision than those search strategies formulated without the use of PICO [10]: this further motivates investigating the use of PICO as operators in queries to constraint retrieval. Closer to our work, Demner-Fushman and Lin [3], and Boudin et al. [1] have used PICO information to aid the retrieval of studies for systematic review screening. The former study used PICO, along with other information, to re-rank studies (thus is not applicable to our work as we examine retrieval, not ranking). The latter, instead, devised two approaches for using PICO within the retrieval process. The first approach involved creating separate language models for each of the PICO elements, using text that has been automatically annotated with respect to each PICO element. Then queries and documents were matched using a mixture model that combined the separate language models, where specific importance weights were given to matches for specific PICO elements (e.g., weighting higher matches for the population element). Empirical results only showed limited improvements and no principled way to set the weights of each language model. A second approach was proposed that exploited both PICO information and structured information of the research studies. Empirical results showed improvements in terms of precision at 10 and average precision; however, the impact of parameter tuning was unclear. Our work is different from previous approaches because we annotate both queries (search strategies) and documents (research studies) with respect to PICO, and restrict matches in documents only to corresponding PICO elements expressed in the queries.

3 USE OF PICO TO SEARCH

An increasing amount of PICO annotations over biomedical research studies has become available. The Cochrane association (a global network of health professionals that provide access to high quality medical information) is undergoing a large manual annotation effort to create a large repository of PICO annotated MEDLINE articles[1], with the intention of making the content and data in systematic reviews more discoverable. At the same time, numerous automated methods have been developed that can annotate biomedical sentences with PICO categories with high accuracy. In this work we use one such tool, called RobotReviewer [13]. RobotReviewer is a machine learning system that uses supervised distance supervision to train models that automate the extraction of PICO elements from systematic reviews. In previous evaluation of RobotReviewer [13], it was found that the system extracted PICO annotations with a precision of 0.9 (top 3 annotations), outperforming other existing methods.

Search strategies aimed at finding studies to be included in systematic reviews are often formulated using the PICO framework. However, the information of which keywords and Boolean clauses in the search strategy refer to eachPICO element is generally not included in the strategies themselves.

We next consider how to constrain the keywords (or Boolean clauses) to specific PICO elements. This could be achieved, for example, by appending operators to keywords that indicate which PICO element was elicited to decide the inclusion of the keyword itself in the search strategy. If this was the case, and if PICO annotations extracted from research studies where indexed as fields alongside the text and metadata of the studies themselves, then the matching between search strategies and research studies could be limited to matching keywords with respect to the correspondent PICO annotations. For example, suppose that a search study identified *weight loss* as the intervention. If the query had no PICO annotations (as in current systems), then studies in which *weight loss* was the measurement outcome would be retrieved. Conversely, if the query included PICO annotations as a condition of the match (as in the method we investigate here), then studies in which *weight loss* was the measurement outcome would *not* be retrieved, and only studies for which the keywords *weight loss* have been annotated as intervention would be retrieved. Note that keywords could have been annotated with more than a PICO element; similarly, some keywords may not be related to any of the PICO elements.

While the method proposed here may improve precision because the match is restricted to keywords used in the same (PICO) context in both the search strategies and the research studies, there is the concrete possibility that it also harms recall. This is because of possible noise or errors introduced in the annotation of text with respect to PICO (whether manually or automatically). In addition, note that the construction of search strategies is an iterative process, and information specialists and researchers use initial searches to formulate the final search strategy. Thus, the conversion of existing search strategies that have no PICO annotations to queries with PICO constraints may not retrieve relevant studies that were identified during search strategy formulation.

In the remainder of the paper, we aim to empirically compare the effectiveness of using PICO elements in search strategies and research studies for the compilation of a list of results to be screened for inclusion in systematic reviews.

4 EMPIRICAL EVALUATION

To evaluate the effectiveness of the retrieval methods we use the test collection compiled by Scells et. al [9]. The collection contains 26 million MEDLINE studies, 94 search strategies from Cochrane systematic reviews and the corresponding relevance assessments. To extract PICO annotations from the research studies, we use RobotReviewer, version 3. Note that this version of RobotReviewer does not extract Control elements. To obtain PICO annotations for search strategies, we employed two medical experts (a clinician and a final year biomedical science student) and compared their annotations among each other. An adjudication process was used to finalise the PICO annotations for search strategies by asking the most experienced annotator to review disagreements[2].

[1] http://community.cochrane.org/tools/data-management-tools/pico-annotation-project

[2] The collection which has also been augmented with PICO annotations is made available at https://github.com/ielab/SIGIR2017-PICO-Collection.

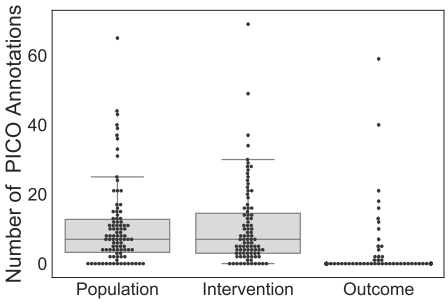

Figure 1: Distribution of the number of PICO annotations per query. 3 outliers have been omitted for Intervention (number of Intervention fields were 96, 113, and 380).

The collection was indexed using Elasticsearch version 5.1.1, including indexing studies with respect to the PICO elements present in the text as separate fields. For each study, we indexed title, abstract and metadata (including publication date and medical subject heading terms) separately because search strategies may contain conditions that restrict the matching of keywords or Boolean clauses to only some of these elements.

We compared four approaches: (B) a baseline approach that use the original Boolean queries, (P) an approach that uses queries for which matches of specific keywords or Boolean clauses may be restricted to some PICO elements, and finally, (cP & fP) two approaches that both use the same PICO restrictions, but attempt to select the optimal PICO elements.

Our evaluation criteria consists of four levels of relevance since not all studies may have been retrieved by the queries we acquired. The studies are classified as such: excluded and not retrieved (*l1*), included and not retrieved (*l2*), excluded and retrieved (*l3*), and included and retrieved (*l4*) [9]. The evaluation described in this section is performed using *l3* and *l4*.

In the first approach (cP), we performed a 'coarse grain selection' of the PICO elements, in which, for each query, we found a combination of the PICO fields by iterating through each possible combination of fields and selecting the combination that lead to the lowest recall loss compared to the baseline. This shows the effect that maintaining only the PICO fields that least reduce recall has on precision. In the second approach (fP), we performed a 'fine grain selection' of the PICO elements, in which we use an oracle heuristic to remove the individual PICO elements (and not the entire fields) that caused a reduction in recall. This last approach is thus explicitly aimed at maintaining the same recall as the Boolean baseline, but improving precision with PICO, where possible.

During preliminary experiments we found that 13 queries matching on only PICO fields (P method) retrieve no studies: this may be due to the reasons outlined in Section 3. When the PICO-based strategy (P) did not retrieve any relevant studies but the baseline did, we removed the PICO annotations from the query (fallback).

As a retrieval task, we considered the task of retrieving studies for screening. The task aims to retrieve all relevant studies, while at the same time minimising non relevant studies; this thereby minimises the time researchers need to spend in reviewing the full text of studies. The balance between ensuring high recall and decreasing the number of not relevant studies to be screened is key for this task. As such, for this task the collection identified as appropriate evaluation measures the F_β-measure and the work saved over sampling (WSS) measure [2][3]:

$$WSS = \frac{N - |retrieved|}{N} - (1 - \frac{|relevant\ retrieved|}{|retrieved|}) \quad (1)$$

WSS measures the work saved (with respect to the number of studies required to be screened) by comparing the number of not relevant studies that have not been retrieved (true negatives), those that have been retrieved and recall. In the F_β-measure, β controls the preference towards recall over precision; studies in systematic reviews automation used $\beta = 1$, $\beta = 3$ (recall three times more important than precision) and $\beta = 0.5$ [8]. In addition, we also report precision and recall.

5 RESULTS AND DISCUSSION

5.1 Analysis of PICO Annotations

Of the 26 million studies in the collection, we found that RobotReviewer annotated with PICO 63% of the studies. All studies annotated with PICO contained a Population field; all studies except for one contained an Intervention field; and all studies except for four contained an Outcomes field. On average, we found that the Population field is 15.8 words in length, the Intervention field is 16.46 words in length, and the Outcomes field is 16.43 words in length. We found that many of the studies that had not been annotated with PICO comprised of title only (no abstract).

Of the 94 queries, all were annotated with PICO elements. On average, there was a similar number of Boolean clauses annotated with Population and Intervention fields (approximately 10.6 clauses per queries). However, there was a significant lower number of clauses annotated as Outcome (2.3 on average). Figure 1 shows the distribution of the number of PICO annotations per query, across the three types of annotations. We found that the majority of queries contained a similar number of Population and Intervention annotations. Most queries have an Intervention annotation (80) and Population annotation (78), while only 18 queries contain an Outcome annotation.

5.2 Analysis of Retrieval Effectiveness

Table 1 reports the evaluation of the retrieval results obtained by the Boolean baseline (B) and the methods that exploits PICO annotations (P, cP, fP). For all measures we computed statistical significance using a paired one-tail t-test. Overall, all PICO-based methods significantly improved precision over the baseline, except for cP which exhibited no significant change in precision. The increase in precision for PICO-based methods translated to a saving of studies retrieved for screening, thus likely allowing potential time savings and cost reduction[4]. Figure 2 (left) illustrates these savings in total number of studies retrieved by method P for the 20 queries with highest number of relevant studies. Overall, method P retrieved 46.4% less studies to be screened than the baseline (B). Figure 2 (right) shows the amount of reduction P achieved for all queries in the collection. (A reduction of 0 indicates the use of *fallback* because the PICO query retrieved no studies). The savings

[3] As such, our evaluation is concerned only with the (set-based) retrieval of studies, not their ranking.

[4] For an estimation of cost reduction at the expense of increased bias due to the lower recall, refer to [11].

	Recall	Precision	F3	F1	F0.5	WSS
B	$\mathbf{0.7553}^{p}$	0.0137^{pf}	0.0901^{pf}	0.0255^{pf}	0.0168^{pf}	0.0120^{pf}
P	0.6509^{bcf}	0.0215^{bc}	0.1128^{bc}	0.0375^{bc}	0.0258^{bc}	0.0206^{bc}
cP	0.7002^{pf}	0.0139^{pf}	0.0886^{pf}	0.0257^{pf}	0.0170^{pf}	0.0124^{pf}
fP	$\mathbf{0.7553}^{pc}$	$\mathbf{0.0223}^{bc}$	$\mathbf{0.1263}^{bc}$	$\mathbf{0.0400}^{bc}$	$\mathbf{0.0271}^{bc}$	$\mathbf{0.0214}^{bc}$

Table 1: Comparison of the effectiveness of the Boolean baseline (B), PICO based retrieval (P), coarsely selected PICO retrieval (cP), and finely selected PICO retrieval (fP). Statistical significance ($p < 0.01$) is denoted as ^{b} (wrt. Boolean) ^{p} (wrt. PICO), ^{c} (wrt. coarsely selected PICO) and ^{f} (wrt. finely selected PICO).

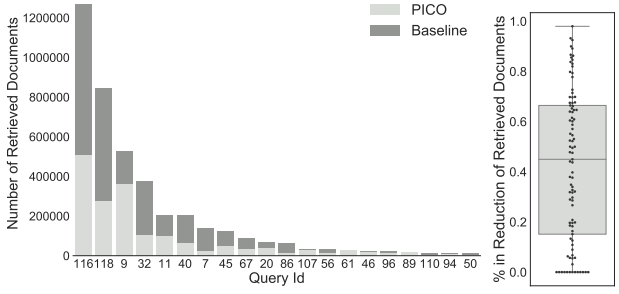

Figure 2: (Left) Number of studies retrieved using PICO search strategies overlaid over the number of studies retrieved using the baseline Boolean search strategies. The first 20 queries are shown; (Right) Percentage reduction of retrieved studies for Boolean search strategies versus PICO search strategies for every query.

in number of studies to be screened is achieved in a simpler and more controllable way[5] (e.g., without resorting to parameter tuning) compared to previous studies that attempted to exploit PICO elements within the retrieval process, e.g. [1].

While the P method reduced recall compared to the baseline, a trade-off between a reduction in studies to be screened and the relevant studies to be retrieved can be obtained using fP. The results in Table 1 in fact show that fP maintains the same recall as the baseline (by definition), while significantly increasing precision (42.58% reduction in number of studies retrieved for screening).

By analysing the other evaluation measures reported in Table 1 we can observe that method fP provides the highest gains over the baseline for all measures, apart from recall which is unchanged.

While overall fP provides promising results, the automatic selection of the optimal PICO fields that guaranteed no loss in recall compared to the Boolean baseline is still an open challenge. Future work will investigate the use of statistical predictors, e.g., query performance prediction, applied to the individual PICO elements in the Boolean queries clauses.

6 CONCLUSION

In this paper we investigated the effectiveness of exploiting PICO annotations in both search strategies (queries) and studies (documents) to narrow down the matches between keywords or Boolean clauses for retrieving studies for screening in systematic reviews.

[5]The ability to carefully control and replicate the retrieval of studies in a systematic review has a key importance within protocols for systematic reviews compilation.

The use of PICO was compared to a Boolean retrieval baseline that represents current search technology employed when performing retrieval for systematic review screening.

Our empirical evaluation on a collection containing 26 million studies (documents) and 94 systematic reviews (queries) showed the effectiveness of the PICO-based methods in reducing the number of false positives retrieved for screening (increase in precision). Despite this, the use of PICO showed a consequent decreased in recall compared to the Boolean baseline, unless the PICO query was modified to explicitly remove the PICO constrains that led to recall losses. However, the analysis of F_{β}-measure and WSS values suggested that, in general, recall losses were more than balanced by savings in terms of number of non relevant studies to be examined in the screening process. This outcome has the potential to drastically and immediately decrease the screening time of research studies for systematic reviews. The use of a large test collection (both in documents and queries) and a tool to automatically annotate abstracts with PICO allowed us to show the large scale effects that searching using PICO elements matching in both queries and documents has for systematic reviews screening, whereas previous studies have used significantly smaller data sets.

These results have considerable implications for the future of biomedical systematic reviews. Firstly, we recommend that systematic review creation guidelines include annotating elements of the search strategy with PICO. Secondly, we highlight the need for an automatic tool able to extract PICO elements from queries in existing systematic reviews: this would benefit the common process of updating existing reviews.

REFERENCES

[1] F Boudin, J Nie, and M Dawes. 2010. Clinical information retrieval using document and PICO structure. In *HLT*. 822–830.

[2] A M Cohen, W R Hersh, K Peterson, and P Yen. 2006. Reducing workload in systematic review preparation using automated citation classification. *JAMIA* 13, 2 (2006), 206–219.

[3] D Demner-Fushman and J Lin. 2007. Answering clinical questions with knowledge-based and statistical techniques. *CL* 33, 1 (2007), 63–103.

[4] S Karimi, S Pohl, F Scholer, L Cavedon, and J Zobel. 2010. Boolean versus ranked querying for biomedical systematic reviews. *BMC MIDM* 10, 1 (2010), 1.

[5] M Khabsa, A Elmagarmid, I Ilyas, H Hammady, and M Ouzzani. 2016. Learning to identify relevant studies for systematic reviews using random forest and external information. *ML* 102, 3 (2016), 465–482.

[6] D Martinez, S Karimi, L Cavedon, and T Baldwin. 2008. Facilitating biomedical systematic reviews using ranked text retrieval and classification. In *ADCS*. 53–60.

[7] M Miwa, J Thomas, A O'Mara-Eves, and S Ananiadou. 2014. Reducing systematic review workload through certainty-based screening. *JBI* 51 (2014), 242–253.

[8] A O'Mara-Eves, J Thomas, J McNaught, Makoto Miwa, and S A. 2015. Using text mining for study identification in systematic reviews: a systematic review of current approaches. *SR* 4, 1 (2015), 5.

[9] Harrisen Scells, Guido Zuccon, Bevan Koopman, Anthony Deacon, Shlomo Geva, and Leif Azzopardi. 2017. A Test Collection for Evaluating Retrieval of Studies for Inclusion in Systematic Reviews. In *ACM SIGIR*.

[10] C Schardt, M B Adams, Thomas Owens, Sheri K, and F P. 2007. Utilization of the PICO framework to improve searching PubMed for clinical questions. *BMC MIDM* 7, 1 (2007), 16.

[11] I Shemilt, N Khan, S Park, and J Thomas. 2016. Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews. *SR* 5, 1 (2016), 140.

[12] I Shemilt, A Simon, G J Hollands, T M Marteau, D Ogilvie, A O'Mara-Eves, M P Kelly, and J Thomas. 2014. Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. *RSM* 5, 1 (2014), 31–49.

[13] B C Wallace, J Kuiper, A Sharma, M B Zhu, and I J Marshall. 2016. Extracting PICO sentences from clinical trial reports using supervised distant supervision. *JMLR* 17, 132 (2016), 1–25.