BMC Medical Informatics and Decision Making

TECHNICAL ADVANCE Open Access

Improving reference prioritisation with PICO recognition

Austin J. Brockmeier[1,2], Meizhi Ju[1], Piotr Przybyła[1,3] and Sophia Ananiadou[1,4*]

Abstract

Background: Machine learning can assist with multiple tasks during systematic reviews to facilitate the rapid retrieval of relevant references during screening and to identify and extract information relevant to the study characteristics, which include the PICO elements of patient/population, intervention, comparator, and outcomes. The latter requires techniques for identifying and categorising fragments of text, known as named entity recognition.

Methods: A publicly available corpus of PICO annotations on biomedical abstracts is used to train a named entity recognition model, which is implemented as a recurrent neural network. This model is then applied to a separate collection of abstracts for references from systematic reviews within biomedical and health domains. The occurrences of words tagged in the context of specific PICO contexts are used as additional features for a relevancy classification model. Simulations of the machine learning-assisted screening are used to evaluate the work saved by the relevancy model with and without the PICO features. Chi-squared and statistical significance of positive predicted values are used to identify words that are more indicative of relevancy within PICO contexts.

Results: Inclusion of PICO features improves the performance metric on 15 of the 20 collections, with substantial gains on certain systematic reviews. Examples of words whose PICO context are more precise can explain this increase.

Conclusions: Words within PICO tagged segments in abstracts are predictive features for determining inclusion. Combining PICO annotation model into the relevancy classification pipeline is a promising approach. The annotations may be useful on their own to aid users in pinpointing necessary information for data extraction, or to facilitate semantic search.

Keywords: Active learning, Evidence-based medicine, Logistic regression, Machine learning, Text mining, Systematic review

Background

Evidence-based research seeks to answer a well-posed, falsifiable question using existing results and a systematic and transparent methodology. The evidence—for example, results of clinical trials—should be collected and evaluated without bias using consistent criteria for inclusion [1]. For certain cases [2], a research question can be decomposed into its PICO elements: patient/population, the intervention, comparator, and outcomes [3, 4]. Along with other aspects, such as study design, PICO elements are useful for formulating search queries for literature database searches [5] and mentions of PICO elements are key to screening the search results for relevance.

A standard approach for systematic reviews (and other review types such as rapid reviews [6] and scoping reviews [7]) is to perform screening initially using only the title and abstracts of a reference collection before obtaining and analysing a subset of full-text articles [1]. While faster and more cost effective than full-text screening, manually screening all reference abstracts is a protracted process for large collections [8], especially those with low specificity [9].

Technology-assisted reviewing seeks to foreshorten this process by only screening the subset of the collection most likely to be relevant [10–13]. This subset is automatically selected using information from a manual screen-

*Correspondence: sophia.ananiadou@manchester.ac.uk
[1]National Centre of Text Mining, School of Computer Science, University of Manchester, Princess Street, M1 7DN Manchester, UK
[4]The Alan Turing Institute, 96 Euston Road, NW1 2DB, London, UK
Full list of author information is available at the end of the article

ing decisions either on another, ideally smaller, subset of the collection [14] or through multiple rounds of iterative feedback between a machine learning (ML) model and the human reviewer [15]. In effect, the machine 'reads' the title and abstract and scores the relevancy of the reference based on a model trained on relevant and irrelevant examples from the human reviewer. While previous studies [7, 16, 17] have shown the potential for time-savings, the underlying models treat each word equally and do not explicitly distinguish PICO elements within an abstract. As PICO elements are crucial for a human reviewer to making inclusion decisions or design screening filters [18], we hypothesise that a ML model with information on each reference's PICO would outperform a similar model lacking this information.

Towards this aim, we propose a PICO recognition model that is able to automatically identify text describing PICO elements within titles and abstracts. The text fragments (contiguous sequences of words) are automatically identified using a named entity recognition model [19] trained on a manually annotated corpus of clinical randomised trial abstracts [20]. Underlying the success of the network is a vector representation of words that is pre-trained on a corpus of PubMed abstracts and articles [21]. The recognition model is based on a neural network architecture [22] that is enhanced to allow the extraction of nested spans, allowing text for one element to be contained within another element. For example, consider the sentence, Steroids (*intervention*) in paediatric (*population*) kidney transplant (*intervention*) recipients (*population*) resulted in reduced acute rejection (*outcome*). The model's predictions are illustrated in Fig. 1. The words in each of the PICO spans are correspondingly marked and treated as additional binary features (in a bag-of-words representation) for a ML model based on a previously validated model [17]. Figure 2 summarizes the whole process as a flowchart.

The performance of the abstract-level screening is evaluated on a standard data set collection of drug effectiveness systematic reviews [14, 24] (DERP I) by the Pacific Northwest Evidence-based Practice Center [25]. The results indicate consistent improvement using PICO information. Furthermore, we perform statistical analysis to identify words that when marked as belonging to a particular PICO element are significant predictors of relevancy and are more precise (higher positive predictive value) than the same words not constrained to the context of PICO mentions. This illustrates how automatically extracting information, obtained by a model trained on expert PICO annotations, can enrich the information available to the machine assisted reference screening.

Related work

Previous work has shown that there are multiple avenues for automation within systematic reviews [26–28]. Examples include retrieval of high-quality articles [29–32], risk-of-bias assessment [33–36], and identification of randomised control trials [37, 38]. Matching the focus of the work, we review previous work on data extraction [39] to automatically isolate PICO and other study characteristics, can be methods for aiding abstract-level screening. The two are clearly related, since inclusion and exclusion criteria can be decomposed into requirements for PICO and study characteristics to facilitate search [40].

Extracting PICO elements (or information in broader schema [41]) at the phrase level [42–44] is a difficult problem due to the disagreement between human experts on the exact words constituting a PICO mention [45, 46]. Thus, many approaches [39] firstly determine the sentences relevant to the different PICO elements, using either rules (formulated as regular expressions) or ML models [42, 46–52]. Finer-grained data extraction can then be applied to the identified sentences to extract the words or phrases for demographic information (age, sex, ethnicity, etc.) [42, 48, 52–54], specific intervention arms [55], or the number of trial participants [56]. Instead of classifying each sentence independently, the structured form of abstracts can be exploited by identifying PICO sentences simultaneously with rhetorical types (aim, method, results, and conclusions) in the abstract [57–60]. More broadly, PICO and other information can be extracted directly from full text articles [61–65].

Rather than extract specific text, Singh et al. predict which medical concepts in the unified medical language system (UMLS) [66] are described in the full-text for each PICO element [67]. They use a neural network model that exploits embeddings of UMLS concepts in addition to word embeddings. The predicted concepts could be used as alternative features rather than just the extracted text. This would supplement manually added metadata such as Medical Subject Headings (MeSH) curated by the U.S. National Library of Medicine [68], which are not always available or have the necessary categorisations.

Our proposed approach differs from existing by both operating at the subsentence level (words and phrases) and using a neural network model for processing text [69] without hand-engineered features. In particular, the proposed approach uses an existing model architecture [19] originally designed for named entity recognition [70] to identify mentions of biomedical concepts such as diseases, drugs, anatomical parts [71, 72]. The model builds from previous neural architectures [22, 73, 74]. The model is jointly trained to predict population, intervention, and outcomes in each sentence in the abstract, and can handle nested mentions where one element's mention (like an intervention) can be contained within another like a

Fig. 1 PICO recognition example. Visualisation of the trained model's predictions of PICO elements within a reference (title and abstract) from the Proton Pump Inhibitors review. The interventions tags correspond to drug names, participant spans cover characteristics of the population, but erroneously include details of the intervention. The latter demonstrates the model's ability to nest shorter spans within longer pans. The outcomes cover spans for qualitative and quantitative measures. Screenshot from the `brat` system [23]

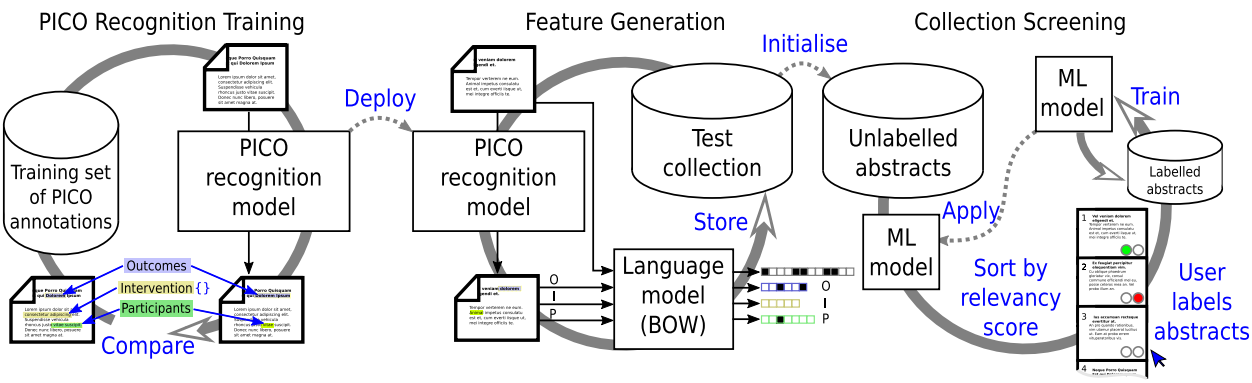

Fig. 2 PICO recognition and abstract screening process. In the first phase, the PICO recognition model is trained to predict the PICO mention spans on a human annotated corpus of abstracts. In the second phase, a collection of abstracts is processed by the PICO recognition model and the results along with the original abstract are used to create a vector representation of each abstract. In the final phase, a user labels abstracts as being included (relevant) or excluded, these decisions are used to train a machine learning (ML) model that uses the vector representation. The ML model is applied to the remaining unlabelled abstracts, which are then sorted by their predicted relevancy, the user sees the top ranked abstracts, labels them, and this process repeats

population. This capability is novel to this work, and in theory, can provide higher recall than methods that do not allow nested PICO elements.

Automatically identified PICO information can improve other automation tasks such as clinical question answering [51] and predicting clinical trial eligibility [75, 76]. Likewise, inclusion and exclusion criteria can be decomposed into requirements for PICO and study characteristics to facilitate search [40]. Recently, Tsafnat et al. have shown the screening ability of automatic PICO extraction [18] for systematic reviews. They use manually designed filters (using dictionaries and rules) [77, 78] for key inclusion criterion, mentions of specific outcomes, population characteristics, and interventions (exposures) to filter collections with impressive gains. Our goal is to replace the manually designed filters with ML modelling that leverages the automatically extracted PICO text to determine an efficient filter. A variety of ML models (different classifiers, algorithms, and feature sets) have been proposed for screening references for systematic reviews [14, 15, 79–95]. Yet, to our knowledge none of relevancy classifiers have used as input the output of PICO recognition.

Methods

The machine learning methodology consists of two main blocks: PICO recognition and relevancy classification. The two steps share some common text pre-processing. To pre-process the text in titles and abstracts, sentence boundaries are determined using the GENIA sentence splitter[1] [96], which was trained on the GENIA corpus [97, 98][2]. Within each sentence, GENIA tagger[3] is used to determine the boundaries between words and other tokens and also the lemmata (base form) of each word [99]. Capitalisation is ignored and lowercase is used for words and lemmata. Additionally, for the PICO recognition each digit is mapped to a zero [69].

PICO recognition model

The PICO annotations have the hierarchical categorisation given in Table 1 where the top-level categories consist of population, intervention/comparator, and outcomes—the comparators are merged into interventions [20]. The annotation is performed in two passes: firstly, top-level spans are identified, and secondly, spans within these are further annotated with the fine-grained types. In this manner, spans corresponding to the fine-grained types are nested within typically longer spans with top-level PICO types.

Following this annotation, the recognition model is trained to firstly extract fine-grained entities, which are under the top-level PICO. Then it extracts the spans corresponding to the top-level PICO elements. To achieve this, the training data consists of an ordered list of IOB tagging [100] sequences for each sentence that mark the beginning (B) and inside (I) of each span, as well as tokens outside (O) of these spans. The lists begin with fine-grained shorter spans and move to top-level longer spans.

Table 1 The top-level and fine-grained PICO elements in the training set for the PICO recognition model

Top-level	Patient-population-problem	Intervention/Comparator	Outcome
Fine-grained	Age	Control	Adverse effect
	Condition	Educational	Mental
	Sample size	Pharmacological	Mortality
	Sex	Physical	Pain
		Psychological	Physical
		Surgical	Other
		Other	

As described in detail [22], the network architecture for the recognition model consists of three main layers: an embedding layer, a sequence processing layer, and a output layer. Firstly, the embedding layer takes as input the sequence of tokens and the character sequence within each token and outputs a vector representation. Each token is represented using the concatenation of word embeddings [101] and representations based on processing character embeddings [102] with a bidirectional long short-term memory network (biLSTM) [103] that employ a forward and reverse LSTM [104] and concatenate the output. Words that are not found in the pre-trained word embeddings are mapped to a common vector, which is further trained by randomly dropping words (50% chance) that occur only once in the training corpus. The second layer processes the sequence of representations using another biLSTM. The third layer is an affine projection of this representation to produce the unitary potential for each of the possible tags in a conditional random field (CRF) model [105], which also models the transition probabilities between tags. Due to the IOB tagging scheme, there are $2 \times (3 + 17) + 1 = 41$ tags corresponding to beginning or inside of one of the 20 possible PICO categories (3 top-level and the 17 fine-grained) and the outside tag. The Viterbi algorithm [106] is used to efficiently infer the most likely sequence of tags marking the spans.

To make predictions of nested spans, the second layer and third layers are iteratively applied to the output of the second layer from the previous iteration until there are no more predicted spans. Specific dimensions of network

[1] http://www.nactem.ac.uk/y-matsu/geniass/
[2] The boundaries are based on punctuation and are unable to correctly split abstracts with lists of unterminated sentences.
[3] http://www.nactem.ac.uk/GENIA/tagger/

architecture are detailed in Table 2. Other choices were not explored.

The network parameters are adjusted to maximise the log likelihood of training sentences for the CRF [69]. Stochastic first-order optimisation is performed using batches of sentences, gradient clipping, and Adam [107]. Dropout [108], weight decay (L_2-regularisation), and early stopping are employed to prevent overfitting. Hyper-parameters are selected using Bayesian optimisation [109], using the design described in [19], on a development portion of the training set with the F1-score of the span-level predictions as the metric.

Relevancy classification model

The relevancy classifier is trained on screening decisions (represented as binary variables indicating inclusion or exclusion). The predictions of the classifier on the unseen references are used to prioritize them, presenting those that are most likely to be relevant. The text processing and feature set follows the description of RobotAnalyst [17], a web-based system that uses ML to prioritise relevant references. The feature set consists of a bag-of-words (BOW) representation of the title, another BOW for the title and abstract combined, and the topic distribution of the title and abstract text.

Topic distributions for title and abstract text are inferred from an LDA topic model [110] with $k = 300$ topics using MALLET [111]. The text is filtered to words consisting of alphabetic characters with initial or internal punctuation that are not on the stop word list. Topic model hyperparameters are initialized as $\alpha = 1/k$ and $\beta = 1/100$ with optimisation every 50 iterations. The topic proportions for each reference are normalised using the L_2 norm.

For the baseline model, the two contexts are title or combined title and abstract. The BOWs are formed from lemmata (base forms) of the occurring words. Included lemmata consist of more than one character, have at least one letter or number, and are not found in a list of stop words[4]. The BOW is a sparse binary vector representing whether or not a word occurred in the given context. Each BOW is normalised to have a Euclidean (L_2) norm of 1 for each reference, except when the bag is empty.

An additional feature set from the PICO recognition consists of a BOW for each of the three course-grained element types patient, intervention, and outcome (comparator is considered an intervention) recognised within the title or abstract. Although finer-grained spans are also annotated and recognised by the model, they were mapped back to the basic PICO types after recognition. In summary, the proposed model uses 5 BOWs. Note that these representations are not disjoint, as a word occurring within a PICO span would both be counted in the general BOW and in the corresponding PICO category BOW.

The classifier is a linear model implemented in `LIBLINEAR` [112]. While RobotAnalyst uses a support vector classifier, we adopt a logistic regression model with L_2-regularisation.[5] The amount of regularisation is controlled by the constraint violation cost parameter C, which is fixed at $C = 1$.

Identifying words with PICO-specific relevancy

We perform two statistical tests to identify words that are both predictive of relevancy for a particular PICO context, and are more predictive than occurrences of the word when it is not restricted to be within the context of a PICO mention. Firstly, for each context category, we compute each word's correlation with relevancy labels using Pearson's χ^2 test statistic for independence. Secondly, for each context-word pair, we compute the positive predictive value (the ratio of the number of included documents containing the word to the total number of documents containing the word) and use Leisenring et al.'s generalised score statistic for equality of positive predictive value [113, 114] to see if the PICO-specific occurrence is significantly more predictive than the word's unrestricted occurrence. The set of PICO-predictive words are those with a significant χ^2 statistic and a positive predictive value both higher and significantly different than the unrestricted context, using a significance level of 0.01 for both tests.

Datasets and simulation

A corpus of annotated references [20, 115] is used for training and evaluation the PICO recognition model. The corpus consists of 4,993 references, a subset of 4,512 are used for training and development (4,061/451). The remainder contains 191 for testing the coarse-grained spans. The remainder also contains 96 that were not used for training since they lacked at least one of the PICO

Table 2 Details of the 3-layer network architecture for the PICO recognition model

Layer		Size	Source
1a	Word embedding	200	[21], not updated
1b	Character embedding	28	trained from random initialisation
1c	Character-based word representation	2×28	biLSTM applied to 1b
1d	Combined embedding	256	concatenation of 1a and 1c
2	Recurrent layer	2×128	biLSTM over 1d
3	Linear layer	41	affine projection of 2
	CRF output	1	most likely sequence of tags

[4] http://members.unine.ch/jacques.savoy/clef/

[5] Preliminary experiments showed logistic regression consistently improved the relevancy prioritisation.

elements, and 194 references which are part of a set of 200 assigned for testing fine-grained labelling. After sentence splitting, there are 43,295 and 4,819 sentences in the training and development sets, respectively.

The DERP collections [24, 116] are used to test whether including the PICO features will improve the prioritisation of relevant references using simulated screening. Table 3 describes the collections for the different reviews.

The simulation is modelled after the RobotAnalyst framework [17], where the classification model is updated at multiple stages during the screening process. Specifically, we run 100 Monte Carlo simulations. In each simulation, we begin with a random batch of 25 references. If this batch contains any relevant references, this forms the initial training set, otherwise batches of 25 are sampled randomly and appended to the training set until at least one relevant reference is found. Given the training set, a classifier is trained and applied to the remaining references. The references are prioritised by the classifier's score, which is proportional to the posterior probability of being relevant (using a logistic regression model). The 25 highest ranked references are then included in the training set, a classifier is retrained, and so on. This continues until all references are screened. This iterative process is readily comparable to relevance feedback methods [117].

To compare against other baselines from the literature we also use a stratified 2-fold setting, where half of the inclusions and half of the exclusions are used for training. Internal results are reported for the average of 100 Monte Carlo trials of stratified training with 50% of each class for training and 50% for testing.

To test the wider applicability of the methodology we applied it to five additional collections introduced by Howard et al. [95]. Four of the collections were produced by the National Institute of Environmental Health Sciences's National Toxicology Program's Office of Health Assessment and Translation (OHAT), and the fifth was produced by the Edinburgh CAMARADES group [118]. Table 4 describes the collections for the different reviews.

Table 3 DERP systematic review descriptive statistics

Review	Inc.	Exc.	Tot.	Prev.
ACE Inhibitors	2544	41	2503	1.61%
ADHD	851	20	831	2.35%
Antihistamines	310	16	294	5.16%
Atypical Antipsychotics	1120	146	974	13.04%
Beta Blockers	2072	42	2030	2.03%
Calcium Channel Blockers	1218	100	1118	8.21%
Estrogens	368	80	288	21.74%
NSAIDS	393	41	352	10.43%
Opioids	1915	15	1900	0.78%
Oral Hypoglycemics	503	136	367	27.04%
Proton Pump Inhibitors	1333	51	1282	3.83%
Skeletal Muscle Relaxants	1643	9	1634	0.55%
Statins	3465	85	3380	2.45%
Triptans	671	24	647	3.58%
Urinary Incontinence	327	40	287	12.23%

Abbreviated columns correspond to the number of inclusions (relevant references), exclusions, total number of references, and the prevalence (percentage of inclusions compared to total)

Evaluation

Firstly, the PICO recognition model is evaluated by its ability to identify top-level (patient, intervention, and outcome) mentions as annotated by experts. Performance is calculated in terms of the model's recall and precision at the level of individual tokens. Each token is treated as an individual test case. True positives for each category are tokens in the category's span that matches the one assigned by the model, and false positives are tokens assigned to the category by the model but not in the original span. This solves the problem of comparing two spans that have matching category, but partially overlapping spans.

The performance is also calculated at the document level in terms of the set of included words. This is a looser evaluation that tests whether the annotated PICO words would be captured when each document is represented as filtered BOW with lemmata, which using the same processing (removing single letter tokens, stop words, etc.) as the BOW for the relevancy classification model. In other words, the document-level matching tests how well individual documents could be retrieved by searching for words within specific PICO contexts. The evaluation uses a held out test set from the same collection as the recognition model training data [20].

Table 4 OHAT and COMARADES systematic review descriptive statistics

Review	Inc.	Exc.	Tot.	Prev.
PFOA/PFOS and Immunotoxicity	6331	95	6236	1.50%
Bisphenol A (BPA) and Obesity	7700	111	7589	1.44%
Transgenerational Inheritance of Health Effects	48638	765	47873	1.57%
Fluoride and Neurotoxicity in Animal Models	4479	51	4428	1.14%
Neuropathic Pain	29207	5011	24196	17.16%

Abbreviated columns correspond to the number of inclusions (relevant references), exclusions, total number of references, and the prevalence (percentage of inclusions compared to total)

Secondly, we test the hypothesis that adding automatically recognised PICO elements to the feature set improves the prioritisation of relevant references. In this setting, the main objective is to prioritise references such that relevant references are presented as early as possible. To compare against baselines from the literature we use both a two-fold relevancy prioritisation [84, 95, 119], and a relevancy feedback setting [120, 121]. In both cases, references with the highest probability of being relevant are screened first [88, 89, 91, 94, 122], like in relevance feedback [117].

As an internal baseline for BOW we consider an average of context-dependent word vectors. Word vectors are trained using algorithms, such as word2vec [123] and GloVe [124], on large corpora such that the vector-space similarity among words reflects the words' distributional similarity: words with similar vectors appear in similar contexts. In comparison, with BOW each word is assigned a vector orthogonal to the rest, such that all words are equally dissimilar. Word vectors perform well on a variety of language tasks, and even better performance is possible when the vector representation of a word depends on its surrounding context [125]. In this case, the context-dependent word vector is computed by the hidden layers of a neural network trained on language modeling tasks. As suggested by a reviewer, we use the context-dependent word vectors from the BERT language model [126], specifically the BioBert model trained on *PubMed* abstracts to better reflect the language of biomedical research papers [127] . For each PICO mention, we compute the average of the output vectors of the last layer hidden of the model for all tokens covered by the span, and then average these for a given PICO category. The BERT representation of abstracts is obtained in the same way, except we average across the vectors for all of the abstract's tokens.

Following previous work, we quantify the performance in terms of work saved over sampling at 95% recall (*WSS*@95%) [14]. This is computed as the proportion of the collection that remains after screening 95% of the relevant reference and subtracting 5% to account for the proportion expected when screening in random order. The recall after screening i references is

$$recall(i) = \frac{TP(i)}{TP(i) + FN(i)}, \tag{1}$$

where $TP(i)$ is the number of relevant references found and $FN(i)$ is the number of relevant references that have not been screened. Likewise, $FP(i)$ denotes the number of irrelevant references found, and $TP(i) + FP(i) = i$. Let i_{R95} denote the number of references screened when 95% recall is firstly achieved. Precisely,

$$i_{\mathrm{R95}} = \min_{\substack{i\in\{1,\ldots,N\} \\ recall(i)\geq 0.95}} i. \tag{2}$$

Under random ordering the expected value for i_{R95} is $95\%N$, where N denotes the total number of references. Work saved is $\frac{N-i_{\mathrm{R95}}}{N}$, and

$$\begin{aligned} WSS@95\% &= \frac{N - i_{\mathrm{R95}}}{N} - 5\% \\ &= 95\% - \frac{i_{\mathrm{R95}}}{N}, \end{aligned} \tag{3}$$

where N denotes the total number of references. The metric is intended to express how much manual screening effort would be saved by a reviewer that would stop the process after finding 95% of the relevant documents. While this metric is useful to compare algorithms, in practice a reviewer will not be able to recognise when 95% recall has been obtained and thus the work saving is a theoretical one, unless a perfect stopping criterion is available.

Results

The test set of 191 abstracts [20, 115] is used to evaluate the model's PICO annotation. The token-wise performance for the three categories is reported in Table 5. The model achieves an F-1 score (geometric mean of precision and recall) of 0.70 for both participants and outcomes, and 0.56 for interventions. The latter is caused by a much lower recall of 0.47. The performance metrics are higher for document-level matching, which uses the same processing (lemmatisation, removing single letter tokens, stop words, etc.) as the BOW for the relevancy classification model. For outcomes, a promising recall of 0.81 is achieved.

The results of relevancy feedback experiment are in Table 6 with the column labelled LR corresponding to the baseline set of features from RobotAnalyst with logistic regression, and PICO indicating the model with the additional PICO bag-of-words features. On average, the inclusion of PICO features increases the work saved metric by 3.3%, with substantial gains for the Opioids and Triptans collections.

We compare these results against two baselines that use relevancy feedback rather ML. The first baseline is a relevance feedback system exploiting the lexical network induced by shared word occurrence [120]. This is a strong baseline as it uses a deterministic seed for retrieval based on custom set of terms in the research questions and

Table 5 PICO recognition performance in terms of a token-wise evaluation and a document-level filtered bag-of-words (BOW)

	Token-wise			Document-level BOW		
	Precision	Recall	F-1	Precision	Recall	F-1
Participants	0.81	0.62	0.70	0.86	0.71	0.78
Interventions	0.69	0.47	0.56	0.83	0.52	0.64
Outcomes	0.66	0.75	0.70	0.73	0.81	0.77

Table 6 Relevancy feedback performance in terms of *WSS*@95% on DERP systematic review collections

	[120]	[121]	LR	PICO	Δ
ACE Inhibitors	74.3	*82.7	74.7	74.4	-0.3
ADHD	67.9	*82.1	67.5	68.9	1.4
Antihistamines	*24.5	17.7	-1.7	-1.9	-0.1
Atypical Antipsychotics	18.0	*33.6	18.0	20.5	2.5
Beta Blockers	65.0	*68.5	54.7	55.7	1.1
Calcium Channel Blockers	17.3	12.8	*47.6	47.1	-0.5
Estrogens	22.6	28.5	36.6	*39.1	2.4
NSAIDS	*77.4	64.1	60.9	63.1	2.2
Opioids	9.0	17.4	19.5	*34.1	14.6
Oral Hypoglycemic	13.5	*15.9	6.9	9.2	2.3
Proton Pump Inhibitors	19.7	21.0	*21.2	18.3	-2.9
Skeletal Muscle Relaxants	*58.6	29.9	25.9	32.4	6.5
Statins	27.8	*43.7	42.9	43.3	0.3
Triptans	39.6	*54.1	34.3	52.4	18.1
Urinary Incontinence	20.8	41.6	44.8	*46.4	1.6
Average	37.1	40.9	36.9	40.2	3.3

[Δ] indicates the change between adding the PICO features to the baseline logistic regression classifier (LR)
[*] indicate best performance per review

the search strategy (in particular the inclusion criterion) and proceeds with relevance feedback adding one reference at a time. Ji et al. follow the same experiment and for a fair comparison we report their results for the case when parameters are fixed ($DT = 1$) across collections using SNOMED-CT and MeSH features for a semantic network [121]. The overall performance with the PICO features is comparable to the semantic network based relevance feedback [121]. This is encouraging since the latter uses a human selected seed query, versus the random initialisation for the proposed method.

Other baselines from the literature only report results in the stratified 2-fold setting. The first baseline [84] uses a naive Bayes classifier, and the reported values are the average across five 2-fold cross-validations, in each of the 10 runs the WSS value for a threshold with at least 95% recall is reported. This includes a weight engineering factor for different groups of features that is maximised on the training set. The second baseline is an SVM-based model [79, 119] with the feature set that performed the best consisting of abstract and title text, MeSH terms, and Meta-map phrases. The final baseline [95] uses cross-validation on the training sets to select the following hyperparameters: the number of topics, the regularisation parameter, and the inclusion or exclusion of additional bigram, trigram, or MeSH term features. The reported values are an average across 25 Monte Carlo trials.

The results are reported in Table 7. The inclusion of PICO features improves the work saved performance metric versus the default logistic regression model, with an average improvement of 1.6%. The results are competitive against the earlier baselines, but the cross-validation selection of hyperparameters [95] yields the best average performance. Searching for these hyperparameters using cross-validations is computational demanding, especially in the relevance feedback setting, where there is not a large initial training set, but rather a different training set at each stage.

Results on the additional OHAT and CAMARADES collections are shown in Table 8. The inclusion of PICO features improves performance on three of the five collections, with an average improvement of 0.3%.

Considering all 20 collections, the addition of PICO features yields a significant improvement in two-fold *WSS*@95% performance over the baseline logistic regression classifier as assessed by a one-sided sign-test (p-value of 0.0207) at a significance level of 0.1.

In Fig. 3, we report the two-fold performance on the DERP collections comparing BOW to BERT with and without the additional PICO features. On this internal comparison, we log and report the number of times a representation performs best across the Monte Carlo trials. BERT performs better on the most difficult collections, but on average, BOW outperforms BERT. Interestingly, the collections that have the highest gain between

Table 7 Two-fold relevancy prediction in terms of *WSS*@95% on DERP systematic review collections

	[84]	[119]	[95]	LR	PICO	Δ
ACE Inhibitors	52.3	73.3	*80.1	78.5	77.6	-0.9
ADHD	62.2	52.6	*79.3	75.5	74.5	-0.9
Antihistamines	14.9	*23.6	13.7	4.9	5.0	0.1
Atypical Antipsychotics	20.6	17.0	*25.1	19.9	20.9	1.0
Beta Blockers	36.7	46.5	42.8	*55.5	54.1	-1.4
Calcium Channel Blockers	23.4	43.0	*44.8	38.8	39.3	0.6
Estrogens	37.5	41.4	*47.1	41.0	43.7	2.7
NSAIDS	52.8	67.2	*73.0	65.3	66.5	1.2
Opioids	55.4	36.4	*82.6	53.3	57.0	3.7
Oral Hypoglycemic	8.5	*13.6	11.7	7.1	8.9	1.8
Proton Pump Inhibitors	22.9	32.8	*37.8	32.6	31.0	-1.6
Skeletal Muscle Relaxants	26.5	37.4	*55.6	40.1	45.3	5.3
Statins	31.5	*49.1	43.6	42.2	44.3	2.1
Triptans	27.4	34.6	41.2	40.6	*51.2	10.5
Urinary Incontinence	29.6	43.2	*53.0	52.4	52.4	0.0
Average	33.5	40.8	48.8	43.2	44.8	1.6

[Δ] indicates the change between adding the PICO features to the baseline logistic regression classifier (LR)
[*] indicate best performance per review

Table 8 Two-fold relevancy prediction in terms of *WSS*@95% on OHAT and CAMARADES systematic review collections

	[95]	LR	PICO	Δ
PFOA/PFOS and Immunotoxicity	80.5	84.0	*84.6	0.7
Bisphenol A (BPA) and Obesity	75.2	77.9	*78.6	0.8
Transgenerational Inheritance of Health Effects	71.4	*74.3	*74.3	0.0
Fluoride and Neurotoxicity in Animal Models	87.0	89.3	*89.4	0.1
Neuropathic Pain	*69.1	64.3	64.1	-0.1
Average	76.6	77.9	78.2	0.3

[Δ]indicates the change between adding the PICO features to the baseline logistic regression classifier (LR)
[*]indicate best performance per review

PICO(BOW) and BOW—Statins, Estrogens, Triptans, and Skeletal Muscle Relaxants—also have a large gap between BOW and BERT. This highlights the utility of the precision that BOW and PICO tagging provide. To assess whether the performance differences were statistically significance, we consider the performance rank of each representation per collection. The average ranks (where the best performing is assigned rank 1) are 2.1 for PICO(BOW), 2.4 for PICO(BERT), 2.7 for BOW, and 2.9 for BERT. The differences in average rank are not significant using a Friedman test at a significance level of 0.1.

To better illustrate the methodology, a subset of PICO features selected by the hypothesis tests for strong relevancy are shown in Tables 9 and 10. The two examples over the cases where the inclusion of PICO features lowered the performance on the Proton Pump Inhibitor review, and raised the performance on the Triptans review. In both cases, the strongly relevant features are clearly indicative of key inclusion aspects. For example, given an occurrence of the word 'complete' there is less than a 50% chance of the reference being relevant; however, within the spans marked as outcome the chance is over 70%. The lower performance in the case of the Proton Pump Inhibitor review corresponds to a lower positive predictive value on these features.

Discussion

The results indicate that the additional PICO tagging is useful for improving machine learning performance in both the two-fold and relevancy feedback scenarios with

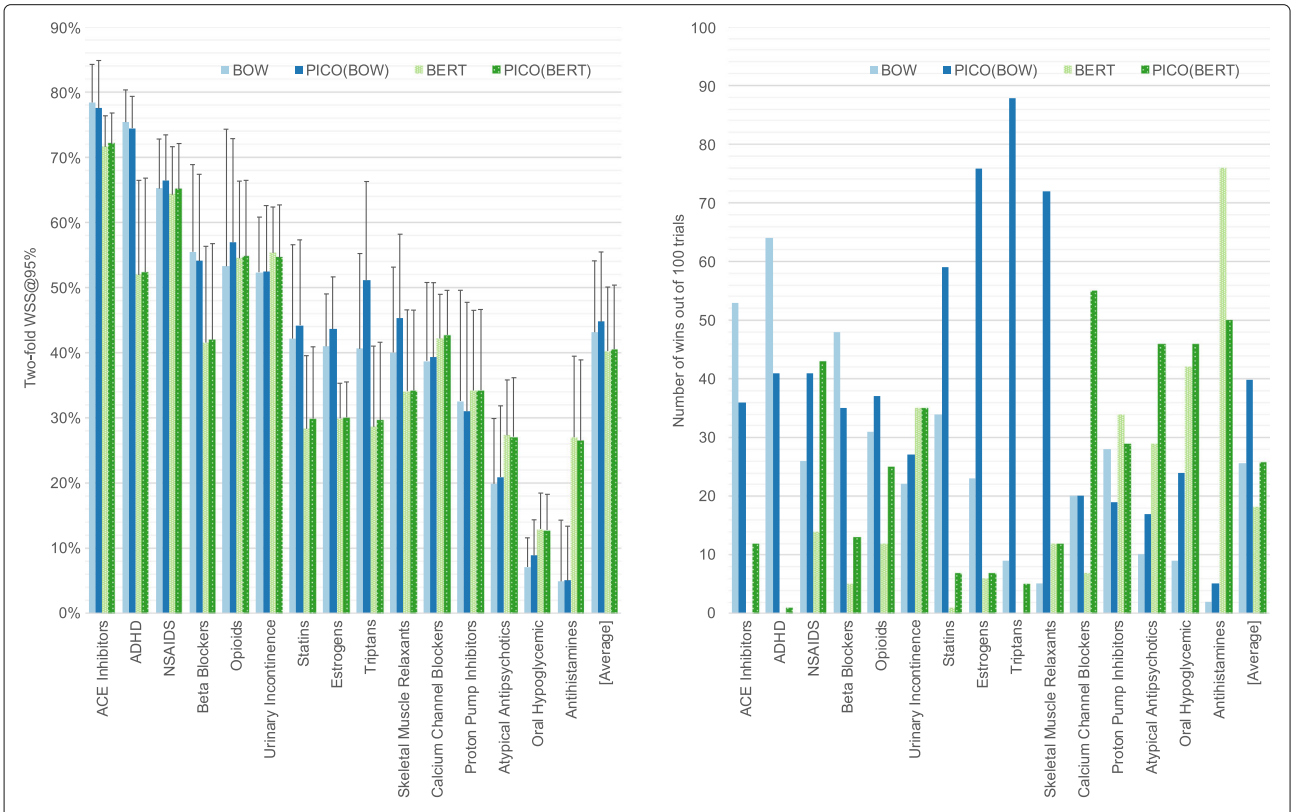

Fig. 3 Comparison of BOW and BERT word vectors as the machine learning representation. The two-fold relevancy prediction performance is reported in terms of WSS@95% across the DERP collections, sorted by BOW performance. In each group, the different colored bars correspond to BOW, BOW including PICO features, BERT, and BERT including PICO features. Bar heights are the average across 100 Monte Carlo trials. In the WSS@95% plot, the upper error bars indicate the standard deviation across the 100 Monte Carlo trials

Table 9 PICO features with strong relevancy within the Proton Pump Inhibitors systematic review

PICO		PPV		TP/FP	
Tag	Lemma	PICO	BOW	PICO	BOW
O	relief	0.21	0.17	21/78	22/111
O	healing	0.13	0.11	33/215	33/264
O	heartburn	0.15	0.11	16/94	16/125
O	pain	0.15	0.12	14/79	14/98
P	oesophagitis	0.14	0.11	13/77	13/107
O	rate	0.07	0.07	35/439	35/501
P	grade	0.15	0.08	8/44	8/90
O	safety	0.10	0.08	11/94	11/122
P	reflux	0.07	0.05	23/311	23/441

Positive predictive value (PPV) is the proportion of true positives (TP) to the total number of TP and false positives (FP). Each TP corresponds to an inclusion containing the feature; each FP corresponds to an exclusion containing the feature

a bag-of-words representation. This could only be the case if the additional features carry information about the relevancy decisions and are not redundant with the existing feature sets. These questions are answered by statistical analysis, which shows that when restricted to a specific PICO context certain words are more reliable predictors. As inclusion criteria are often stated in terms of PICO (and other study characteristics) this is not a surprising result, but nonetheless, requires a well-trained PICO recognition model to transfer the knowledge from the training set of annotations. In a way, the proposed methodology connects with previous work on generalisable classifiers that can learn from the screening decisions of other systematic reviews [128].

Table 10 PICO features with strong relevancy within the Triptans systematic review

PICO		PPV		TP/FP	
Tag	Lemma	PICO	BOW	PICO	BOW
O	relief	0.68	0.61	96/46	106/67
O	headache	0.53	0.43	130/113	161/212
P	migraine	0.50	0.41	138/138	198/281
P	treat	0.78	0.59	49/14	124/85
O	pain	0.59	0.52	90/63	96/89
O	severe	0.80	0.60	40/10	89/60
O	moderate	0.79	0.63	34/9	94/55
O	response	0.59	0.49	51/35	71/75
I	sumatriptan	0.43	0.41	141/187	145/211
O	mild	0.73	0.53	29/11	71/62
O	migraine	0.51	0.41	74/70	198/281
O	functional	0.81	0.56	21/5	25/20
O	effective	0.82	0.47	18/4	106/120
O	patient	0.67	0.43	26/13	194/253
O	complete	0.71	0.47	15/6	36/40
O	reduction	0.64	0.42	16/9	30/42
O	reduce	0.87	0.38	7/1	39/63
O	migraine-specific	0.80	0.50	8/2	10/10

Positive predictive value (PPV) is the proportion of true positives (TP) to the total number of TP and false positives (FP). Each TP corresponds to an inclusion containing the feature; each FP corresponds to an exclusion containing the feature

Furthermore, PICO tagging is an interpretable process meant to emulate human annotation and can readily be used by reviewers themselves. For instance, highlighting the mentions of outcomes may accelerate data extraction, since identifying outcome measures and data are a critical step in many systematic reviews. In the context of the ML model, the influence of a specific PICO feature in prioritising an abstract can be assessed by the corresponding coefficients of the logistic regression model. This can be used to check which of the PICO categories has contributed the most to the score assigned to a certain abstract—for example, the presence of an outcome-specific word with a relatively large coefficient. If this raises doubts, the text spans assigned to this type can be verified. The ability to interact with the model in such ways would increase its interpretability, which could aid a user in understanding and trusting the current model's predictions [129]. While this can be done for all of the words, the semantics, sparsity and higher precision of PICO features make them more meaningful.

There are a number of avenues for future work. The first is to consider PICO tagging in new systematic reviews. The simulation results remains a surrogate for actual live screening evaluation as was performed by Przybyła et al. [17]. In practice, users may benefit from more precise queries where search terms are restricted to appear in PICO recognised spans, or integrated into additional facets for semantic search [130]. That is, the semantic classes of interventions and outcomes may be useful for users to search large collections and databases. For example, if instead of searching for a phrase or word describing an outcome measure in the whole text of the references, a reviewer would be able to search just within the fragments categorised as outcomes, the results would better align with the reviewer's intention. The word 'reduce' in Table 10 is a strong example, where only 8 results with 7 being relevant are returned for ouctome-specific usage compared to 102 results with only 39 relevant in general. This demonstrates that a query-driven approach with PICO tagging has the potential to greatly reduce screening efforts needed to obtain an initial seed of relevant documents. User selected queries could be combined with RobotAnalyst's ability to prioritise the results based on relevance predictions. Essentially, this would combine

the approach proposed here with the ability for human design [18] of screening rules using PICO classes. Finally, in this work the fine-grained PICO recognition was not evaluated, but this may be useful to highlight population information (sample size, age, sex, condition).

During peer review, it was noted that the DERP collections also contain the reasons for most exclusions. Reasons for exclusions are often recorded in systematic reviews, and may be coded using PICO categories. Thus, a system with PICO-specific feature sets has the potential of incorporating the additional information into a ML model. This is an interesting area for future work.

Finally, we note that the proposed methodology is not able to beat relevancy screening baselines previously reported in the literature. This can largely be attributed to differences in evaluation. For the relevancy feedback experiments, the baseline methods [120, 121] start from deterministic queries that use expert knowledge of the inclusion criteria, versus the random initialisation for the propose method. In the case of two-fold predictions, the best performing method [95] uses cross validation to select the best from among different hyperparameters combinations, including distinct feature set choices. This would require additional computation in the online setting and it is not clear if this approach would perform well in the limited data setting (without access to half of the inclusions).

Conclusion

Screening abstracts for systematic reviews requires users to read and evaluate abstracts to determine if the study characteristics match the inclusion criterion. A significant portion of these are described by PICO elements. In this study, words within PICO tagged segments automatically identified in abstracts are shown to be predictive features for determining inclusion. Combining PICO annotation model into the relevancy classification pipeline is a promising approach to expedite the screening process. Furthermore, annotations may be useful on their own to aid users in pinpointing necessary information for data extraction, or to facilitate semantic search.

Authors' contributions

SA conceived the plan of work, which was overseen by PP. MJ developed the PICO recognition model and its implementation. AJB applied the model to the abstract screening task. PP implemented the context dependent word embeddings for abstract and PICO snippets. The machine learning models were designed, implemented and evaluated by AJB. AJB and PP drafted the manuscript, and all authors reviewed and provided input on preliminary versions. All authors read and approved the final manuscript.

Funding

This work was supported by the Medical Research Council: grant number MR/N00583X/1 "Manchester Molecular Pathology Innovation Centre" and grant number MR/L01078X/1 "Supporting Evidence-based Public Health Interventions using Text Mining", and the Biotechnology and Biological Sciences Research Council: grant number BB/P025684/1 "Japan Partnering Award. Text mining and bioinformatics platforms for metabolic pathway modelling". MJ was financially supported by the University of Manchester's 2016 President's Doctoral Scholar Award. PP was partially supported by the Polish Returns programme of the Polish National Agency for Academic Exchange (PPN/PPO/2018/1/00006). These funding sources had no role in the design of this study nor any role during its execution, analyses, interpretation of the data, or manuscript preparation.

Availability of data and materials

The datasets supporting the conclusions of this article are available in the Drug Effectiveness Review Project (DERP) repository [24], the EBM-NLP corpus [115], and as additional files [95].

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The author declares that he/she has no competing interests.

Author details

[1]National Centre of Text Mining, School of Computer Science, University of Manchester, Princess Street, M1 7DN Manchester, UK. [2]University of Delaware, 139 The Green, 19716 Newark, Delaware, USA. [3]Linguistic Engineering Group, Institute of Computer Science, Polish Academy of Sciences, Jana Kazimierza 5, 01-248 Warszawa, Poland. [4]The Alan Turing Institute, 96 Euston Road, NW1 2DB, London, UK.

Received: 5 April 2019 Accepted: 22 November 2019
Published online: 05 December 2019

References

1. Higgins JP, Deeks JJ. Selecting studies and collecting data. In: Higgins JP, Green S, editors. Cochrane Handbook for Systematic Reviews of Interventions, Chap. 7. Version 5.1.0. Chichester: The Cochrane Collaboration. John Wiley & Sons; 2011. updated March 2011.
2. Huang X, Lin J, Demner-Fushman D. Evaluation of PICO as a knowledge representation for clinical questions, vol. 2006. In: AMIA Annual Symposium Proceedings. American Medical Informatics Association; 2006. p. 359.
3. Oxman AD, Sackett DL, Guyatt GH, Browman G, Cook D, Gerstein H, Haynes B, Hayward R, Levine M, Nishikawa J, et al. Users' guides to the medical literature: I. how to get started. JAMA. 1993;270(17):2093–5.
4. Richardson WS, Wilson MC, Nishikawa J, Hayward RSA. The well-built clinical question: a key to evidence-based decisions. ACP J Club. 1995;123(3):12.
5. Schardt C, Adams MB, Owens T, Keitz S, Fontelo P. Utilization of the PICO framework to improve searching PubMed for clinical questions. BMC Med Inf Dec Making. 2007;7(1):16.
6. Wagner G, Nussbaumer-Streit B, Greimel J, Ciapponi A, Gartlehner G. Trading certainty for speed - how much uncertainty are decisionmakers and guideline developers willing to accept when using rapid reviews: an international survey. BMC Med Res Methodol. 2017;17(1):121. https://doi.org/10.1186/s12874-017-0406-5.
7. Shemilt I, Simon A, Hollands GJ, Marteau TM, Ogilvie D, O'Mara-Eves A, Kelly MP, Thomas J. Pinpointing needles in giant haystacks: use of text mining to reduce impractical screening workload in extremely large scoping reviews. Res Synth Methods. 2014;5(1):31–49. https://doi.org/10.1002/jrsm.1093.
8. Allen IE, Olkin I. Estimating time to conduct a meta-analysis from number of citations retrieved. Jama. 1999;282(7):634–5.
9. Lefebvre C, Glanville J, Wieland LS, Coles B, Weightman AL. Methodological developments in searching for studies for systematic reviews: past, present and future? Syst Rev. 2013;2(78):. https://doi.org/10.1186/2046-4053-2-78.
10. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Syst Rev. 2015;4(5):. https://doi.org/10.1186/2046-4053-4-5.

11. Shemilt I, Khan N, Park S, Thomas J. Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews. Syst Rev. 2016;5(140):. https://doi.org/10.1186/s13643-016-0315-4.
12. Kanoulas E, Li D, Azzopardi L, Spijker R. CLEF 2017 technologically assisted reviews in empirical medicine overview. In: CEUR Workshop Proceedings, vol. 1866; 2017. p. 1–29. http://ceur-ws.org/Vol-1866/invited_paper_12.pdf. Accessed 27 Sept 2018.
13. Kanoulas E, Li D, Azzopardi L, Spijker R. Clef 2018 technologically assisted reviews in empirical medicine overview. In: CEUR Workshop Proceedings, vol. 2125; 2018. p. 1–34. http://ceur-ws.org/Vol-2125/invited_paper_6.pdf. Accessed 27 Sept 2018.
14. Cohen AM, Hersh WR, Peterson K, Yen PY. Reducing workload in systematic review preparation using automated citation classification. J Am Med Inform Assoc. 2006;13(2):206–19. https://doi.org/10.1197/jamia.M1929.
15. Wallace BC, Trikalinos TA, Lau J, Brodley C, Schmid CH. Semi-automated screening of biomedical citations for systematic reviews. BMC Bioinf. 2010;11(1):55.
16. Rathbone J, Hoffmann T, Glasziou P. Faster title and abstract screening? Evaluating Abstrackr, a semi-automated online screening program for systematic reviewers. Syst Rev. 2015;4(80):. https://doi.org/10.1186/s13643-015-0067-6.
17. Przybyła P, Brockmeier AJ, Kontonatsios G, Le Pogam M-A, McNaught J, von Elm E, Nolan K, Ananiadou S. Prioritising references for systematic reviews with robotanalyst: A user study. Res Synth Meth. 9(3):470–88. https://doi.org/10.1002/jrsm.1311.
18. Tsafnat G, Glasziou P, Karystianis G, Coiera E. Automated screening of research studies for systematic reviews using study characteristics. Syst Rev. 2018;7(1):64.
19. Ju M, Miwa M, Ananiadou S. A neural layered model for nested named entity recognition. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans: Association for Computational Linguistics; 2018. p. 1446–1459. http://aclweb.org/anthology/N18-1131.
20. Nye B, Li JJ, Patel R, Yang Y, Marshall IJ, Nenkova A, Wallace BC. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. 2018197–207. https://doi.org/10.18653/v1/p18-1019.
21. Chiu B, Crichton G, Korhonen A, Pyysalo S. How to train good word embeddings for biomedical nlp. In: Proceedings of the 15th Workshop on Biomedical Natural Language Processing; 2016. p. 166–74. https://doi.org/10.18653/v1/w16-2922.
22. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural architectures for named entity recognition. In: Proceedings of NAACL-HLT; 2016. p. 260–70. https://doi.org/10.18653/v1/n16-1030.
23. Stenetorp P, Pyysalo S, Topić G, Ohta T, Ananiadou S, Tsujii J. 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. Stroudsburg: Association for Computational Linguistics; 2012. p. 102–7.
24. Pacific Northwest Evidence-based Practice. OHSU Center for Evidence-Based Policy: Drug Effectiveness Review Project (DERP) Systematic Drug Class Review Gold Standard Data. https://dmice.ohsu.edu/cohenaa/systematic-drug-class-review-data.html. Accessed 16 Jan 2018.
25. Pacific Northwest Evidence-based Practice Center. Drug Effectiveness Review Project (DERP). https://www.ohsu.edu/xd/research/centers-institutes/evidence-based-practice-center/drug-effectiveness-review-project/current-past-reports.cfm. Accessed 16 Jan 2018.
26. Thomas J, McNaught J, Ananiadou S. Applications of text mining within systematic reviews. Res Synth Meth. 2011;2(1):1–14.
27. Tsafnat G, Glasziou P, Choong MK, Dunn A, Galgani F, Coiera E. Systematic review automation technologies. Syst Rev. 2014;3(1):74.
28. Beller E, Clark J, Tsafnat G, Adams C, Diehl H, Lund H, Ouzzani M, Thayer K, Thomas J, Turner T, et al. Making progress with the automation of systematic reviews: principles of the international collaboration for the automation of systematic reviews (ICASR). Syst Rev. 2018;7(1):77.
29. Aphinyanaphongs Y, Aliferis CF. Text categorization models for retrieval of high quality articles in internal medicine. In: AMIA Annual Symposium Proceedings, vol. 2003. Bethesda: American Medical Informatics Association; 2003. p. 31–5. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1480096/.
30. Aphinyanaphongs Y, Tsamardinos I, Statnikov A, Hardin D, Aliferis CF. Text categorization models for high-quality article retrieval in internal medicine. J Am Med Inform Assoc. 2005;12(2):207–216.
31. Choi S, Ryu B, Yoo S, Choi J. Combining relevancy and methodological quality into a single ranking for evidence-based medicine. Inf Sci. 2012;214:76–90.
32. Del Fiol G, Michelson M, Iorio A, Cotoi C, Haynes RB. A deep learning method to automatically identify reports of scientifically rigorous clinical research from the biomedical literature: Comparative analytic study. J Med Int Res. 2018;20(6):. https://doi.org/10.2196/preprints.10281.
33. Marshall IJ, Kuiper J, Wallace BC. Automating risk of bias assessment for clinical trials. IEEE J Biomed Health Inf. 2015;19(4):1406–12.
34. Marshall IJ, Kuiper J, Wallace BC. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. J Am Med Inform Assoc. 2015;23(1):193–201.
35. Millard LA, Flach PA, Higgins JP. Machine learning to assist risk-of-bias assessments in systematic reviews. Int J Epidemiol. 2015;45(1):266–77.
36. Zhang Y, Marshall I, Wallace BC. Rationale-augmented convolutional neural networks for text classification. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, vol. 2016. NIH Public Access; 2016. p. 795. https://doi.org/10.18653/v1/d16-1076.
37. Cohen AM, Smalheiser NR, McDonagh MS, Yu C, Adams CE, Davis JM, Yu PS. Automated confidence ranked classification of randomized controlled trial articles: an aid to evidence-based medicine. J Am Med Inform Assoc. 2015;22(3):707–17.
38. Marshall IJ, Noel-Storr A, Kuiper J, Thomas J, Wallace BC. Machine learning for identifying randomized controlled trials: an evaluation and practitioner's guide. Res Synth Methods. 2018. https://doi.org/10.1002/jrsm.1287.
39. Jonnalagadda SR, Goyal P, Huffman MD. Automating data extraction in systematic reviews: a systematic review. Syst Rev. 2015;4(1):78.
40. Sim I, Tu SW, Carini S, Lehmann HP, Pollock BH, Peleg M, Wittkowski KM. The ontology of clinical research (OCRe): an informatics foundation for the science of clinical research. J Biomed Inf. 2014;52:78–91.
41. Dawes M, Pluye P, Shea L, Grad R, Greenberg A, Nie J-Y. The identification of clinically important elements within medical journal abstracts: Patient_population_problem, exposure_intervention, comparison, outcome, duration and results (PECODR). J Innov Health Inf. 2007;15(1):9–16.
42. Hara K, Matsumoto Y. Extracting clinical trial design information from medline abstracts. N Gener Comput. 2007;25(3):263–75.
43. Summerscales R, Argamon S, Hupert J, Schwartz A. Identifying treatments, groups, and outcomes in medical abstracts. In: The Sixth Midwest Computational Linguistics Colloquium (MCLC 2009). Bloomington: Indiana University; 2009.
44. Summerscales RL, Argamon S, Bai S, Hupert J, Schwartz A. Automatic summarization of results from clinical trials. In: 2011 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2011. p. 372–377. https://doi.org/10.1109/bibm.2011.72.
45. Niu Y, Hirst G. Analysis of semantic classes in medical text for question answering. In: Proceedings of the Conference on Question Answering in Restricted Domains; 2004.
46. Demner-Fushman D, Lin J. Answering clinical questions with knowledge-based and statistical techniques. Comput Linguist. 2007;33(1):63–103.
47. Demner-Fushman D, Lin J. Knowledge extraction for clinical question answering: Preliminary results. In: Proceedings of the AAAI-05 Workshop on Question Answering in Restricted Domains. Pittsburgh: AAAI Press (American Association for Artificial Intelligence); 2005. p. 9–13.
48. Xu R, Garten Y, Supekar KS, Das AK, Altman RB, Garber AM. Extracting subject demographic information from abstracts of randomized clinical trial reports. In: Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems, Stud Health Technol Inform. Amsterdam: IOS Press; 2007. p. 550–54.
49. Kim SN, Martinez D, Cavedon L, Yencken L. Automatic classification of sentences to support evidence based medicine. BMC Bioinformatics. 2011;12:5. BioMed Central.
50. Boudin F, Nie J-Y, Bartlett JC, Grad R, Pluye P, Dawes M. Combining classifiers for robust PICO element detection. BMC Med Inf Dec Making. 2010;10(1):29.

51. Boudin F, Shi L, Nie J-Y. Improving medical information retrieval with pico element detection. In: European Conference on Information Retrieval. Springer; 2010. p. 50–61. https://doi.org/10.1007/978-3-642-12275-0_8.
52. Zhao J, Kan M-Y, Procter PM, Zubaidah S, Yip WK, Li GM. Improving search for evidence-based practice using information extraction. In: AMIA Annual Symposium Proceedings, vol. 2010. Bethesda: American Medical Informatics Association; 2010. p. 937.
53. Zhao J, Bysani P, Kan M-Y. Exploiting classification correlations for the extraction of evidence-based practice information. In: AMIA Annual Symposium Proceedings, vol. 2012. Bethesda: American Medical Informatics Association; 2012. p. 1070.
54. Kelly C, Yang H. A system for extracting study design parameters from nutritional genomics abstracts. J Integr Bioinform. 2013;10(2):82–93.
55. Chung GY-C. Towards identifying intervention arms in randomized controlled trials: extracting coordinating constructions. J Biomed Inform. 2009;42(5):790–800.
56. Hansen MJ, Rasmussen N. Ø., Chung G. A method of extracting the number of trial participants from abstracts describing randomized controlled trials. J Telemed Telecare. 2008;14(7):354–8.
57. Chung GY, Coiera E. A study of structured clinical abstracts and the semantic classification of sentences. In: Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing. Association for Computational Linguistics; 2007. p. 121–128. https://doi.org/10.3115/1572392.1572415.
58. Chung GY. Sentence retrieval for abstracts of randomized controlled trials. BMC Med Inf Dec Making. 2009;9(1):10.
59. Dernoncourt F, Lee JY, Szolovits P. Neural networks for joint sentence classification in medical paper abstracts. arXiv preprint arXiv:1612.05251. 2016.
60. Jin D, Szolovits P. Pico element detection in medical text via long short-term memory neural networks. In: Proceedings of the BioNLP 2018 Workshop; 2018. p. 67–75. https://doi.org/10.18653/v1/w18-2308.
61. De Bruijn B, Carini S, Kiritchenko S, Martin J, Sim I. Automated information extraction of key trial design elements from clinical trial publications. In: AMIA Ann Symp Proc, vol. 2008. Bethesda: American Medical Informatics Association; 2008. p. 141.
62. Kiritchenko S, de Bruijn B, Carini S, Martin J, Sim I. ExaCT: automatic extraction of clinical trial characteristics from journal publications. BMC Med Inf Dec Making. 2010;10(1):56.
63. Hsu W, Speier W, Taira RK. Automated extraction of reported statistical analyses: towards a logical representation of clinical trial literature. In: AMIA Annual Symposium Proceedings, vol. 2012. Bethesda: American Medical Informatics Association; 2012. p. 350.
64. Bui DDA, Del Fiol G, Hurdle JF, Jonnalagadda S. Extractive text summarization system to aid data extraction from full text in systematic review development. J Biomed Inf. 2016;64:265–72.
65. Wallace BC, Marshall IJ. Extracting PICO sentences from clinical trial reports using supervised distant supervision. J Mach Learn Res. 2016;17:1–25.
66. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32(suppl_1):267–70.
67. Singh G, Marshall IJ, Thomas J, Shawe-Taylor J, Wallace BC. A neural candidate-selector architecture for automatic structured clinical text annotation. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. ACM; 2017. p. 1519–28. https://doi.org/10.1145/3132847.3132989.
68. Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM indexing initiative's medical text indexer. 2004;107:268–72. https://doi.org/10.1037/e615572012-009.
69. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. J Mach Learn Res. 2011;12(Aug):2493–537.
70. Nadeau D, Sekine S. A survey of named entity recognition and classification. Lingvisticae Investigationes. 2007;30(1):3–26.
71. Leaman R, Gonzalez G. BANNER: an executable survey of advances in biomedical named entity recognition. In: Pacific Symposium on Biocomputing, vol. 13. Hawaii: World Scientific; 2008. p. 652–663.
72. Chowdhury M, Faisal M, et al. Disease mention recognition with specific features. In: Proceedings of the 2010 Workshop on Biomedical Natural Language Processing. Stroudsburg: Association for Computational Linguistics; 2010. p. 83–90.
73. Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991. 2015.
74. Ma X, Hovy E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1; 2016. p. 1064–74. https://doi.org/10.18653/v1/p16-1101.
75. Restificar A, Ananiadou S. Inferring appropriate eligibility criteria in clinical trial protocols without labeled data. In: Proceedings of the ACM Sixth International Workshop on Data and Text Mining in Biomedical Informatics, DTMBIO '12. New York: ACM; 2012. p. 21–8. https://doi.org/10.1145/2390068.2390074.
76. Restificar A, Korkontzelos I, Ananiadou S. A method for discovering and inferring appropriate eligibility criteria in clinical trial protocols without labeled data. In: BMC Medical Informatics and Decision Making, vol. 13; 2013. p. 6. https://doi.org/10.1186/1472-6947-13-s1-s6.
77. Karystianis G, Buchan I, Nenadic G. Mining characteristics of epidemiological studies from medline: a case study in obesity. J Biomed Semant. 2014;5(1):22.
78. Karystianis G, Thayer K, Wolfe M, Tsafnat G. Evaluation of a rule-based method for epidemiological document classification towards the automation of systematic reviews. J Biomed Inf. 2017;70:27–34.
79. Cohen AM. Optimizing feature representation for automated systematic review work prioritization. In: AMIA Annual Symposium Proceedings, vol. 2008. Bethesda: American Medical Informatics Association; 2008. p. 121–125.
80. Cohen AM, Ambert K, McDonagh M. A prospective evaluation of an automated classification system to support evidence-based medicine and systematic review. In: AMIA Annual Symposium Proceedings, vol. 2010. Bethesda: American Medical Informatics Association; 2010. p. 121.
81. Bekhuis T, Demner-Fushman D. Towards automating the initial screening phase of a systematic review. In: Safran C, Reti S, Marin H, editors. World Congress on Medical Informatics (MEDINFO), Stud Health Technol Inform, vol. 160. Amsterdam: IOS Press; 2010. p. 146–50.
82. Bekhuis T, Demner-Fushman D. Screening nonrandomized studies for medical systematic reviews: a comparative study of classifiers. Artif Intell Med. 2012;55(3):197–207.
83. Bekhuis T, Tseytlin E, Mitchell KJ, Demner-Fushman D. Feature engineering and a proposed decision-support system for systematic reviewers of medical evidence. PLoS ONE. 2014;9(1):86277.
84. Matwin S, Kouznetsov A, Inkpen D, Frunza O, O'Blenis P. A new algorithm for reducing the workload of experts in performing systematic reviews. J Am Med Inform Assoc. 2010;17(4):446–53.
85. Frunza O, Inkpen D, Matwin S. Building systematic reviews using automatic text classification techniques. In: International Conference on Computational Linguistics (COLING). Stroudsburg: Association for Computational Linguistics; 2010. p. 303–11.
86. Frunza O, Inkpen D, Matwin S, Klement W, O'Blenis P. Exploiting the systematic review protocol for classification of medical abstracts. Artif Intell Med. 2011;51(1):17–25.
87. Small K, Wallace B, Trikalinos T, Brodley CE. The constrained weight space SVM: learning with ranked features. In: International Conference on Machine Learning (ICML-11). Norristown: Omnipress; 2011. p. 865–72.
88. Wallace BC, Small K, Brodley CE, Lau J, Trikalinos TA. Deploying an interactive machine learning system in an evidence-based practice center: abstrackr. ACM SIGHIT Symp Int Health Inf. 2012;819:. https://doi.org/10.1145/2110363.2110464.
89. Jonnalagadda S, Petitti D. A new iterative method to reduce workload in systematic review process. Int J Comput Biol Drug Des. 2013;6(1-2):5–17.
90. Dalal SR, Shekelle PG, Hempel S, Newberry SJ, Motala A, Shetty KD. A pilot study using machine learning and domain knowledge to facilitate comparative effectiveness review updating. Med Decis Making. 2013;33(3):343–55.
91. Miwa M, Thomas J, O'Mara-Eves A, Ananiadou S. Reducing systematic review workload through certainty-based screening. J Biomed Inform. 2014;51:242–53. https://doi.org/10.1016/j.jbi.2014.06.005.
92. Timsina P, Liu J, El-Gayar O. Advanced analytics for the automation of medical systematic reviews. Inf Syst Front. 2016;18(2):237–52.
93. Khabsa M, Elmagarmid A, Ilyas I, Hammady H, Ouzzani M. Learning to identify relevant studies for systematic reviews using random forest and external information. Mach Learn. 2016;102(3):465–82.

94. Hashimoto K, Kontonatsios G, Miwa M, Ananiadou S. Topic detection using paragraph vectors to support active learning in systematic reviews. J Biomed Inform. 2016;62:59–65.
95. Howard BE, Phillips J, Miller K, Tandon A, Mav D, Shah MR, Holmgren S, Pelch KE, Walker V, Rooney AA, Macleod M, Shah RR, Thayer K. SWIFT-Review: a text-mining workbench for systematic review. Syst Rev. 2016;5(87):. https://doi.org/10.1186/s13643-016-0263-z.
96. Sætre R, Yoshida K, Yakushiji A, Miyao Y, Matsubayashi Y, Ohta T. AKANE system: protein-protein interaction pairs in BioCreAtIvE2 challenge, PPI-IPS subtask. In: Proceedings of the Second Biocreative Challenge Workshop, vol. 209. Madrid: CNIO Centro Nacional de Investigaciones Oncológicas; 2007. p. 212.
97. Ohta T, Tateisi Y, Kim J-D. The genia corpus: An annotated research abstract corpus in molecular biology domain. In: Proceedings of the Second International Conference on Human Language Technology Research. Morgan Kaufmann Publishers Inc.; 2002. p. 82–86. https://doi.org/10.3115/1289189.1289260.
98. Kim J-D, Ohta T, Tateisi Y, Tsujii J. Genia corpus—a semantically annotated corpus for bio-textmining. Bioinformatics. 2003;19(suppl_1):180–2.
99. Tsuruoka Y, Tsujii J. Bidirectional inference with the easiest-first strategy for tagging sequence data. In: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2005. p. 467–74. https://doi.org/10.3115/1220575.1220634.
100. Sang EF, Veenstra J. Representing text chunks. In: Conference of the European Chapter of the Association for Computational Linguistics (EACL). Association for Computational Linguistics; 1999. p. 173–9. https://doi.org/10.3115/977035.977059.
101. Bengio Y, Ducharme R, Vincent P, Jauvin C. A neural probabilistic language model. J Mach Learn Res. 2003;3(Feb):1137–55.
102. Santos CD, Zadrozny B. Learning character-level representations for part-of-speech tagging. In: Proceedings of the 31st International Conference on Machine Learning (ICML-14). PMLR; 2014. p. 1818–1826. http://proceedings.mlr.press/v32/.
103. Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Netw. 2005;18(5-6):602–10.
104. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80.
105. Lafferty JD, McCallum A, Pereira FCN. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc.; 2001. p. 282–9.
106. Viterbi A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Trans Inf Theory. 1967;13(2):260–9.
107. Kinga D, Adam JB. A method for stochastic optimization. In: Proceedings of the 3rd International Conference on Learning Representations (ICLR). San Diego; 2015.
108. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. J Mach Learn Res. 2014;15:1929–58.
109. Snoek J, Larochelle H, Adams RP. Practical bayesian optimization of machine learning algorithms. In: Advances in Neural Information Processing Systems. Red Hook: Curran Associates, Inc.; 2012. p. 2951–9.
110. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. J Mach Learn Res. 2003;3:993–1022.
111. McCallum AK. MALLET: A machine learning for language toolkit. 2002. http://mallet.cs.umass.edu. Accessed 16 Jan 2018.
112. Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J. LIBLINEAR: A library for large linear classification. J Mach Learn Res. 2008;9:1871–4. https://doi.org/10.1038/oby.2011.351.
113. Leisenring W, Alono T, Pepe MS. Comparisons of predictive values of binary medical diagnostic tests for paired designs. Biometrics. 2000;56(2):345–51.
114. Kosinski AS. A weighted generalized score statistic for comparison of predictive values of diagnostic tests. Stat Med. 2013;32(6):964–77.
115. Nye B, Li JJ, Patel R, Yang Y, Marshall IJ, Nenkova A, Wallace BC. EBM-NLP. https://ebm-nlp.herokuapp.com. Accessed 9 June 2018.
116. Cohen AM. An effective general purpose approach for automated biomedical document classification. In: AMIA Annual Symposium Proceedings, vol. 2006. Bethesda: American Medical Informatics Association; 2006. p. 161.
117. Salton G, Buckley C. Improving retrieval performance by relevance feedback. J Assoc Inf Sci Technol. 1990;41(4):288–97.
118. Collaborative Approach to Meta-Analysis and Review of Animal Data from Experimental Studies. CAMARADES. www.camarades.info. http://www.dcn.ed.ac.uk/camarades/default.htm.
119. Cohen AM. Performance of support-vector-machine-based classification on 15 systematic review topics evaluated with the WSS@ 95 measure. J Am Med Inf Assoc: JAMIA. 2011;18(1):104.
120. Ji X, Yen P-Y. Using medline elemental similarity to assist in the article screening process for systematic reviews. JMIR Med Inf. 2015;3(3):. https://doi.org/10.2196/medinform.3982.
121. Ji X, Ritter A, Yen P-Y. Using ontology-based semantic similarity to facilitate the article screening process for systematic reviews. J Biomed Inf. 2017;69:33–42.
122. Kontonatsios G, Brockmeier AJ, Przybyła P, McNaught J, Mu T, Goulermas JY, Ananiadou S. A semi-supervised approach using label propagation to support citation screening. J Biomed Inform. 2017;72:67–76.
123. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Adv Neural Inf Process Syst. Red Hook: Curran Associates, Inc.; 2013. p. 3111–9.
124. Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP); 2014. p. 1532–43. http://www.aclweb.org/anthology/D14-1162. https://doi.org/10.3115/v1/d14-1162.
125. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L. Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans: Association for Computational Linguistics; 2018. p. 2227–37. https://doi.org/10.18653/v1/N18-1202.
126. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics; 2018. p. 4171–86. 1810.04805.
127. 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. arXiv:1901.08746. 2019.
128. Cohen AM, Ambert K, McDonagh M. Cross-topic learning for work prioritization in systematic review creation and update. J Am Med Inform Assoc. 2009;16(5):690–704.
129. Lipton ZC. The Mythos of Model Interpretability. In: Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016). New York; 2016. https://doi.org/10.1145/3233231.
130. Soto AJ, Przybyła P, Ananiadou S. Thalia: semantic search engine for biomedical abstracts. Bioinformatics. 2018;35(10):1799–801.

Publisher's Note