# PICO Entity Extraction For Preclinical Animal Literature

**Qianying Wang**
University of Edinburgh https://orcid.org/0000-0002-7779-6815
**Jing Liao**
University of Edinburgh
**Mirella Lapata**
University of Edinburgh
**Malcolm Macleod** (malcolm.macleod@ed.ac.uk)
University of Edinburgh https://orcid.org/0000-0001-9187-9839

**Research**

**Keywords:** PICO, preclinical animal study, named entity recognition, information extraction, self-training

**Posted Date:** October 28th, 2021

**DOI:** https://doi.org/10.21203/rs.3.rs-1008099/v1

# Abstract

**Background:** Natural language processing could assist multiple tasks in systematic reviews to reduce workflow, including the extraction of PICO elements such as study populations, interventions and outcomes. The PICO framework provides a basis for the retrieval and selection for inclusion of published evidence relevant to a specific systematic review question, and automatic approaches of PICO extraction have been developed particularly for reviews of clinical trial findings. Considering the difference between preclinical animal studies and clinical trials, developing separate approaches are necessary. Facilitating preclinical systematic reviews will inform the translation from preclinical to clinical research.

**Methods:** We randomly selected 400 abstracts from the PubMed Central Open Access database which described in vivo animal research and manually annotated these with PICO phrases for Species, Strain, model Induction, Intervention, Comparator and Outcome. We developed a two-stage workflow for preclinical PICO extraction. Firstly we fine-tuned BERT with different pre-trained modules for PICO sentence classification. Then, after removing text irrelevant to PICO features, we explored LSTM, CRF and BERT-based models for PICO entity recognition. We also explored a self-training approach because of the small training corpus.

**Results:** For PICO sentence classification, BERT models using all pre-trained modules achieved an F1 score over 80%, and models pre-trained on PubMed abstracts achieved the highest F1 of 85%. For PICO entity recognition, fine-tuning BERT pre-trained on PubMed abstracts achieved an overall F1 of 71%, and satisfactory F1 for Species (98%), Strain (70%), Intervention (70%) and Outcome (67%). The score of Induction and Comparator is less satisfactory, but F1 of Comparator can be improved to 50% by applying self-training.

**Conclusions:** Our study indicates that of the approaches tested, BERT pre-trained on PubMed abstracts is the best for both PICO sentence classification and PICO entity recognition in the preclinical abstracts. Self-training yields better performance for identifying comparators and strains.


# Background

Systematic review attempts to collate all relevant evidence to provide reliable summary of findings relevant to a pre-specified research question [1]. When conducting information extraction from clinical literature, the key elements of interest are Population/Problem, Intervention, Comparator and Outcome, which compose the established framework of PICO [2]. This has been used as the basis for retrieval, inclusion and classification of published evidence, and empirical studies have shown the use of the PICO framework facilitates more complex search strategies and yields more precise search results in systematic reviews [3]. As the number of publications describing experimental studies has increased, the time taken in manually extracting information has increased such that many reviews are out of date by the time they are published. The evidence-based research community has responded by advocating the

use of automated approaches to assist systematic reviews, and PICO extraction tools have been developed, particularly for clinical trials [4].

Preclinical animal studies differ from clinical trials in many aspects. The aim of animal studies is to explore new hypotheses for drug or treatment development, so they have more variations for the definition of PICO elements. For example, in animal studies, disease is not naturally present but often induced; different species can be used; and outcomes of interest can include survival, behavioural, histological and biochemical outcomes [5]. Considering the difference and the leading clinical research, the SYRCLE group developed a framework definition of preclinical PICO, where “Population” includes animal species and strain, and any method of inducing a disease model; and several outcomes can be considered [6]. Importantly, the “Comparator” for animal studies is usually simply an untreated control cohort, although the exact choice of control is sometimes a variable of interest.

Here we report the development of automatic PICO extraction approaches for preclinical animal studies which may advocate the use of preclinical PICO and facilitate the translation from preclinical to clinical research.

# Related work

To our knowledge, while automated PICO extraction in clinical reports is relatively well-explored, no method has been developed or evaluated for preclinical animal literature.

Most of the previous work for the clinical trial literature casts PICO element extraction as a sentence classification task. Byron et al use logistic regression with distant supervision to train PICO sentences derived from clinical articles [7]. More recent approaches have used recent neural networks for PICO sentence classification which requires less manual feature engineering. Such approaches include bidirectional long-short term memory network (BiLSTM) [8] with some variations [9-11]. More precise PICO phrases or snippets extraction is cast as a named-entity recognition task, and BiLSTM with conditional random field (CRF) [12] are common approaches [13-14]. Some advanced methods including graph learning [15] and BERT [16] enhance the performance.

# Methods

# Dataset

We downloaded 2,207,654 articles from PubMed Central Open Access Subset database published from 2010 to 2019 and used a citation screening filter trained to identify in vivo research from title and abstract (developed by EPPI-Centre, UCL [17]). We chose an inclusion cut-point which gave 99% precision and obtained 50,653 abstracts describing in vivo animal experiments. We randomly selected 400 abstracts for the annotation task and another 10,000 for the self-training experiments.

We used the online platform tagtog for PICO phrases annotation. In addition to Intervention, Comparator and Outcome, we divided the Population category into three components: the Species, the Strain, and the method of Induction of the disease model. After the initial annotation process and discussion with a senior clinician, we proposed some general rules for the annotation task:

- Only PICO spans describing in vivo experiments are annotated, i.e. interventions or treatments should be conducted within an entire, living organism. Interventions applied to tissues derived from an animal or in cell culture (ex vivo or in vitro experiments) should not be annotated;
- Texts describing introduction, conclusion or objectives should not be annotated in most cases because these might relate to work other than that described in the publication. They should be annotated only when the remaining text lacks a clear description of the method, or where the text gives the meaning of abbreviations;
- The first occurrence of abbreviation should be annotated together with the parent text. For example, "vascular endothelial growth factor (VEGF)" should be tagged as one entity for its first occurrence; in the remainder of the text, "VEGF" or "vascular endothelial growth factor" could be annotated separately if they are not mentioned together;
- Any extra punctuations between phrases (such as commas) should not be annotated. However, if the entity appears only one time in the text, punctuations can be included in a long span of text which consists of several phrases;
- Entity spans cannot be overlapped. Annotations in tagtog are output in EntitiesTsv format which resembles the tsv output in the Stanford NER tool [18], and this does not support overlapping entities.

Figure 1 shows an example of annotated abstract using tagtog. After excluding the title, introduction sentence, first part of the objective sentence and the conclusion sentence which do not explicitly describe experimental elements, PICO entities are extracted from the remaining sentences: 1) Species: mice; 2) Strain: C57BL/6; 3) Induction: fed normal chow (NC), fed high-fat diet (HFD); 4) Intervention: aerobic exercise training, exercise, treadmill running; 5) Comparator: sedentary; and 6) Outcome: protein spots.

In total, 6837 entities were annotated across 400 abstracts, and the distribution of PICO entities is imbalanced (Table 1). Less than 50% of sentences in each abstract contains PICO phrases and using the entire abstracts to train an entity recognition model is not efficient. Therefore, we split the PICO phrases extraction task into two independent subtasks: 1) PICO sentence classification, and 2) PICO entity recognition.

Table 1
Statistics of 400 annotated PICO dataset

| | **Average number in each abstract** |
|---|---|
| PICO sentences | 5 |
| Sentences | 11 |
| Entities | 17.5 |
| | **Distribution of PICO entity** |
| Intervention | 24.1% |
| Comparator | 1.8% |
| Outcome | 40.6% |
| Induction | 10.6% |
| Species | 19.6% |
| Strain | 3.3% |
| Total | 100% |

# PICO sentence classification

Text from 400 abstracts are split into 4,247 sentences by scispaCy [19] and sentences containing at least one PICO entity are labelled as 'True' for PICO sentence. Individual sentences were randomly allocated to training, validation and test sets (80%/10%/10%). For the sentence-level classification task, we use BERT, a contextualized representation model where a deep bidirectional encoder is trained on a large text corpus. The encoder structure is derived from the powerful transformer based on multi-head self-attention, which dispenses with issues arising from recurrence and convolutions [20]. The pre-trained BERT can be fine-tuned with a simple additional output layer for downstream tasks and achieves state-of-the-art performance on many natural language processing tasks [16]. We explore the effects of using different text corpuses and methods for pre-training including 1) BERT-base, the original BERT trained on the combination of BookCorpus, and English Wikipedia [16]; 2) BioBERT, which trains BERT on the combination of BookCorpus, English Wikipedia, PubMed abstracts and PubMed Central full-text articles [21]; 3) PubMedBERT-abs, which trains BERT on PubMed abstracts only, and 4) PubMedBERT-full on a combination of PubMed abstracts and PubMed Central full-text articles [22].

The approach to training seeks to minimise cross-entropy loss using the AdamW algorithm [23]. We use a slanted triangular learning rate scheduler [24] with a maximum learning rate 5e-5 for 10 epochs of training. We apply gradient clipping [25] with a threshold norm of 0.1 to rescale gradients and gradient accumulation every 16 steps (mini-batches) to reduce memory consumption.

# PICO entity recognition

Identifying specific PICO phrases is cast as a named-entity recognition (NER) task. We convert all entity annotations to the standard BIO format [26], i.e. each word/token is labelled as 'B-XX' if it is the beginning word of the 'XX' entity, 'I-XX' if it belongs to other words inside the entity but not the beginning word, or 'O' if it is outside of any PICO entity. Hence, there are 13 unique tags for 6 PICO entities (two tags for each entity, plus tag 'O'), and a NER model is trained to assign the 13 unique tags to each token in the PICO text.

One classic NER model is the bidirectional long-short term memory (BiLSTM) with a CRF layer on top (BiLSTM-CRF) [27]. LSTM belongs to the family of recurrent neural networks which can process word embeddings sequentially. In the hidden layer, by combining the weighted hidden representations from the adjacent word through a Tanh operation, a basic recurrent neural structure can retain information from neighbouring text. However, when the document is long, retraining information from very early or late words is difficult because of the exploding or vanishing gradient problem, which stops the network learning efficiently [28]. LSTM is designed to solve this long-term dependencies problem, which uses a cell state and three gates (forget gate, input gate and output gate) for each word embedding to control information we need to flow straight, to forget, or to store and update to the next step [8]. BiLSTM contains information from words in both directions, by processing hidden vectors from previous words to the current word, and hidden vectors from future words back to the current words.

CRF is a type of discriminative probabilistic model which is often added on top of LSTMs to model dependencies and learn the transition constraints among predicted tags from LSTM output. For example, if the tag of a word in the sequence is 'I-Outcome', the tag of the previous word can only be 'B-Outcome' or 'I-Outcome', and impossible to be 'I-Intervention' or 'O' in a real sample. Models without CRF layer may lose these constraints and cause unnecessary transition errors. We explore BiLSTM models with or without CRF layers. For text representations in these models, tokens are mapped into 200-dimension vectors by word2vec [29] induced on a combination of PubMed, PMC texts and English Wikipedia [30].

Similar to the PICO sentence classification, we also fine-tuned BERT with different pre-trained weights for the entity recognition task, using BertForTokenClassification module from Hugging Face Transformers library [31]. We also explored the effect of adding CRF and LSTM layers on top of BERT.

For more efficient training and to achieve best results for the entity recognition task, we removed sentences without any PICO annotation from each abstract and trained NER models on each remaining text, which consisted of PICO sentences only; for prediction in the future application, sentences in an individual abstract are classified by the best PICO sentence classifier from the first task, and the non-PICO sentences are then removed automatically. The workflow is illustrated in Figure 2.

For LSTM/CRF models, we tuned hidden dimension from 32 to 512, and compare Adam and AdamW optimizers with the constant or slanted triangular learning rate scheduler. We froze word embeddings because we found it achieves better performance on the validation set. Models were trained for 20

epochs and the learning rate depended on the specific model (1e-3 for BiLSTM and 5e-3 for BiLSTM-CRF). For BERT models, we fine-tuned BERT for 20 epochs with learning rate of 1e-3, BERT-CRF for 30 epochs and BERT-LSTM-CRF for 60 epochs, both with a learning rate 1e-4; other settings are similar to that of PICO sentence classification task. These settings were determined by checking overfitting or convergence issues from their learning curves. For evaluation, we used entity-level metrics [32] for each PICO text (truncated abstract):

$$\mathrm{Precision_i} = \frac{\mathrm{numberofpredictedcorrectentities}}{\mathrm{numberofpredictedentities}}$$

$$\mathrm{Recall_i} = \frac{\mathrm{numberofpredictedcorrectentities}}{\mathrm{numberoftrueentities}}$$

$$\mathrm{F1_i} = \frac{2{*}\mathrm{Precision_i} * \mathrm{Recall_i}}{\mathrm{Precision_i} + \mathrm{Recall_i}}$$

These individual metrics were then averaged across all validation/test samples to obtain the overall metrics.

# Self-training

One limitation of the previous method is the small amount of training data so we also explored a semi-supervised learning strategy, self-training, which uesd the unlabelled dataset to generate pseudo labels for training [33]. We use 400 annotated abstracts as “gold” data, and 10,000 unlabelled abstracts from 50,653 in vivo animal records as “silver” data. Non-PICO sentences were removed from the unlabelled text by the best PICO sentence classification model, and these truncated texts were used for self-training. As Figure 3 shows, we first used the fine-tuned PICO entity recognizer from the gold set (80% of 400 labelled records for training, 10% for validation) to predict entities of each token in the silver set. For each abstract in the silver set, we calculated the average prediction probabilities of all tokens within that abstract. Silver records with average probabilities larger than a threshold (0.95 or 0.99) were then combined with the original gold training/validation set, and the enlarged new dataset was used to fine-tune a newly initialised PICO entity recognizer. Then we repeat the prediction, pseudo data generation, data selection and supervised fine-tuning procedures, until no more unlabelled records with average prediction probabilities larger than the threshold are identified. Note in every data enlarging step, newly included silver records are split into training set (80%) and validation set (20%), then combined with existing gold training records (initially 320 records) and gold validation records (initially 40 records) respectively. This guarantees that the initial gold validation set is only ever used for validation. The original gold test set is used for final evaluation. All experiments were conducted using an Ubuntu machine with 16-core CPU.

[1] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist

[2] https://www.tagtog.net

# Results

The results of PICO sentence classification models on the test set (425 sentences) are shown in Table 2 (see validation performance in Appendix Table 1). All BERT models achieve an F1 score greater than 80% regardless of the pre-training corpus used, and PubMedBERT trained on PubMed abstracts achieves the highest F1 score of 85.4%. Biomedical-domain BERT improves F1 score by 4% compared with general-domain BERT, and BERT with pure biomedical-domain pre-training (two PuBMedBERT) can identify more PICO sentences than BERT with general pre-training (BERT-base) or mixed-domain pretraining (BioBERT), as recall increased by 7%. Therefore, we selected BERT trained on PubMed abstracts as the best PICO sentence classifier for self-training experiments and prediction.

Table 2
Performance of PICO sentence classification by BERT with different pre-trained weights on the test set.

| | F1 | Recall | Precision |
|---|---|---|---|
| BERT-base | 80.6 | 81.4 | 82.1 |
| BioBERT | 84.3 | 81.0 | 90.0 |
| PubMedBERT-abs | 85.4 | 88.4 | 85.0 |
| PubMedBERT-full | 84.2 | 87.1 | 83.8 |

For PICO entity recognition, for each model we used settings which achieved the best performance on the validation set, and then evaluated these them on the test set (40 truncated abstracts). As Table 3 shows, BERT models (BERT, BERT-CRF, BERT-BiLSTM-CRF) outperformed LSTM models (BiLSTM, BiLSTM-CRF), with F1 scores improved by between 3% and 27%. The use of a CRF layer improves F1 score in BiLSTM by 14%, but does not enhance performance in BERT models. Compared with the benefit from the large-scale pre-trained domain knowledge, the advantage of CRF layer might therefore be trivial. Within BERT models, biomedical BERT models improve F1 by at least 4% compared to the general BERT, and the difference among three biomedical pre-trained weights is not obvious. We selected PuBMedBERT pre-trained on PubMed abstracts and full texts as the best PICO entity recognizer based on the validation results (see Appendix Table 2), and the test performance by each PICO entity is reported in Table 4 (‘original scores’). The F1 score for identifying Species is 98%. This entity has a limited number of potential responses, so their identification is not complicated. For Intervention and Outcome, the performance is satisfactory, with F1 around 70%. F1 scores of Strain and Induction are 63% and 49% respectively, and so there remains room for improvement. The F1 score for identifying the Comparator is only 16%, which may be due to the relative lack of Comparator instances in the training corpus, and unclear boundaries in the definition of comparator and interventions in some complicated manuscripts. For instance, a manuscript may describe two experiments, and what is an intervention in the first may become a comparator in the second.

Table 3
Overall performance of PICO entity recognition models on the test set.

| Model | Weight | F1 | Recall | Precision |
|---|---|---|---|---|
| BiLSTM | -- | 43.5 | 38.1 | 50.6 |
| BiLSTM-CRF | -- | 57.9 | 54.7 | 61.6 |
| BERT | Base | 61.3 | 66.3 | 57.1 |
| | BioBERT | 65.4 | 69.8 | 61.5 |
| | PubMed-abs | 70.1 | 73.2 | 67.3 |
| | PubMed-full | 69.9 | 73.4 | 66.7 |
| BERT-CRF | Base | 62.1 | 67.2 | 57.8 |
| | BioBERT | 66.5 | 70.1 | 63.3 |
| | PubMed-abs | 68.0 | 71.5 | 64.9 |
| | PubMed-full | 67.5 | 70.9 | 64.5 |
| BERT -BiLSTM -CRF | Base | 64.6 | 69.5 | 60.3 |
| | BioBERT | 68.3 | 71.2 | 65.6 |
| | PubMed-abs | 67.2 | 70.8 | 64.0 |
| | PubMed-full | 68.5 | 72.6 | 64.8 |

In self-training experiments, we used the best PICO sentence classifier (BERT pre-trained on PubMed abstracts) to remove non-PICO sentences for unlabelled data, and the best PICO entity recognizer (BERT pre-trained on PubMed abstracts and full texts) to identify PICO phrases and calculate prediction scores across all tokens in each individual text. We explore two thresholds (0.95, 0.99) for records selection, and the results are reported in Figure 4. When the threshold is 0.99, no more silver records are included in the training set beyond the first iteration, and self-training did not improve performance. When the threshold is 0.95, the performance fluctuates and the best F1 score is improved by 5% and 1% on the gold validation set and test set respectively, achieved at the sixth iteration step. We terminated the training program after 15 iterations because the training size tends to saturate and the improvement of performance is very limited. For specific PICO entities, the main improvement using self-training was for F1 scores for Comparator and Strain, increased by 32% and 7% respectively ('self-training scores' in Table 4).

Table 4
Entity-level performance of PubMedBERT on the gold test set. Original scores refer to performance of model before self-training; self-training scores refer to performance of model at the best iteration (6th iteration) of self-training. 'R' and 'P' refer to recall and precision respectively.

| | Original scores | | | Self-training scores | | |
|---|---|---|---|---|---|---|
| | F1 | R | P | F1 | R | P |
| **Comparator** | 16.0 | 10.0 | 40.0 | 48.5 | 40.0 | 61.5 |
| **Induction** | 49.1 | 50.6 | 47.7 | 48.0 | 49.4 | 46.6 |
| **Intervention** | 70.2 | 76.1 | 65.2 | 69.8 | 74.6 | 65.6 |
| **Outcome** | 65.4 | 70.6 | 60.9 | 66.9 | 70.6 | 63.6 |
| **Species** | 98.1 | 100.0 | 96.4 | 98.1 | 100.0 | 96.4 |
| **Strain** | 63.4 | 72.2 | 56.5 | 70.0 | 77.8 | 63.6 |
| **Overall** | 69.9 | 73.4 | 66.7 | 71.0 | 74.0 | 68.2 |

We have developed an interactive application via Streamlit for potential use (see Figure 5). When the user inputs the PMID from the PubMed Open Access Subset, the app will call the PubMed Parser [34] to return its title and abstract. The background sentence model classifies and removes non-PICO sentences, and then the entity recognizer identifies the PICO phrases from those PICO sentences. This can give a quick overview of PICO elements of an experimental study.

[3]https://streamlit.io/

# Discussion

In this work we show the possibilities of automated PICO sentence classification and PICO entity recognition in abstracts describing preclinical animal studies. For sentence classification, BERT models with different pre-trained weights have generally good performance (F1 over 80%), and biomedical BERT (BioBERT or PubMedBERT) have slightly better performance than general BERT. For PICO entity recognition, all BERT models outperform BiLSTM with or without a CRF layer, with the improvement of F1 ranging from 3–27%. It is unnecessary to use a more complicated structure based on BERT, as the results of BERT, BERT-BiLSTM and BERT-BiLSTM-CRF do not differ greatly, but the latter two bring a cost in longer training time and resources. Within LSTM based models, adding a CRF layer is beneficial, where recall is increased by 16% and precision is increased by 9%. The training time of LSTM based models is much shorter than fine-tuning BERT, and this could be a quick alternative solution when computing resources are limited, at the cost of reduction of performance by 3% and 12% compared to the general BERT and PubMedBERT respectively. The self-training approach helps to identify more comparators and strains, but does not help much with the overall performance. By entity levels, F1 scores are generally

good for identifying Species (over 80%), satisfactory for Intervention, Outcome and Strain (around or over 70%), and acceptable for Induction and Outcome (around 50%).

One limitation of our work is that the training corpus is at the level of the abstract, but some PICO elements in preclinical animal studies are often not described in the abstract. This limits the usefulness of our applications and we cannot transfer it to full-text identification without further evaluation. Of note, this same limitation applies to manual approaches to identifying PICO elements based on the abstract alone. In a related literature we have shown, for instance, that manual screening for inclusion based on TiAb has substantially lower sensitivity than manual screening of full texts (https://osf.io/nhjeg). Another limitation is that the amount of training, validation and test data is not adequate. Although our best models do not show very inconsistent results between validation and test set (except for “Comparator”), the conclusions may still be biased using small dataset. Previous studies show that self-training can propagate both knowledge and error from high confidence predictions on unlabelled samples [35], and that training from larger annotated corpora may reduce the error propagation and boost performance. Large datasets also provide possibilities for exploring more complicated models which are proved effective in other tasks.

In future work we will evaluate our PICO sentence classification and entity recognition models on some full-text publications, to observe any heuristic implications. We will also evaluate existing clinical PICO extraction tools on preclinical text to identify interventions and outcomes because these two categories may be more similar in preclinical and clinical studies than other PICO elements. As the training corpora for clinical PICO are relatively larger and in more standard forms, we think that training using a combined preclinical/clinical corpus may yield better performance.

# Conclusions

We demonstrate a workflow for PICO extraction in preclinical animal text using LSTM and BERT based models. Without feature engineering, BERT pre-trained on PubMed abstracts is optimal for both PICO sentence classification, and BERT pre-trained on PubMed abstracts and full texts is optimal for PICO entity recognition tasks in preclinical abstracts. PICO entities including Intervention, Outcome, Species and Strain have acceptable precision and recall (around or over 70%), while Comparator and Induction have less satisfactory scores (around 50%). We encourage the collection of a more standard PICO annotation corpus and the use of natural language processing models for PICO extraction in preclinical animal studies, which may achieve better results for publications retrieval, reduce workflow of preclinical systematic reviews, and narrow the gap between preclinical and clinical research.

# Abbreviations

BERT: Bidirectional Encoder Representations from Transformers

BiLSTM: Bidirectional Long-Short Term Memory

BIO: Beginning-Inside-Outside tagging format

CRF: Conditional Random Field

NER: Named Entity Recognition

# Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Availability of data and materials

The datasets supporting the current study are available in the Preclinical PICO extraction repository, https://osf.io/2dqcg.

### Competing interests

The authors declare that they have no competing interests.


### Funding

This work is jointly funded by China Scholarships Council and a John Climax UK Reproducibility Network PhD Studentship


### Authors' contributions

JL and QW collected and processed the in vivo abstracts. QW and MM conducted PICO phrases annotations. QW developed and implemented the classification, entity recognition and self-training models. ML was involved in the design of self-training experiments. QW analysed and evaluated results. All authors reviewed and provided comments on preliminary versions. All authors read and approved the final manuscript.


### Acknowledgements

None.


# References

1. Higgins JPT, Green S, (editors). Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. 2011.

2. Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. ACP journal club. 1995;123. doi:10.7326/acpjc-1995-123-3-a12.

3. Huang X, Lin J, Demner-Fushman D. Evaluation of PICO as a knowledge representation for clinical questions. AMIA Annu Symp Proc. 2006;2006:359–63. http://www.fpin.org/. Accessed 29 Mar 2021.

4. Marshall IJ, Wallace BC. Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. Systematic Reviews. 2019;8:163. doi:10.1186/s13643-019-1074-9.

5. Hooijmans CR, Rovers MM, De Vries RBM, Leenaars M, Ritskes-Hoitinga M, Langendam MW. SYRCLE's risk of bias tool for animal studies. BMC Med Res Methodol. 2014;14:43. doi:10.1186/1471-2288-14-43.

6. Hooijmans CR, De Vries RBM, Ritskes-Hoitinga M, Rovers MM, Leeflang MM, IntHout J, et al. Facilitating healthcare decisions by assessing the certainty in the evidence from preclinical animal studies. PLoS One. 2018;13.

7. Wallace BC, Kuiper J, Sharma A, Zhu MB, Marshall IJ. Extracting PICO Sentences from Clinical Trial Reports using Supervised Distant Supervision. J Mach Learn Res. 2016;17. http://www.ncbi.nlm.nih.gov/pubmed/27746703. Accessed 3 Mar 2019.

8. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Comput. 1997;9:1735–80. doi:10.1162/neco.1997.9.8.1735.

9. Jin D, Szolovits P. PICO Element Detection in Medical Text via Long Short-Term Memory Neural Networks. In: Proceedings of the BioNLP 2018 workshop. Stroudsburg, PA, USA: Association for Computational Linguistics; 2018. p. 67–75. doi:10.18653/v1/W18-2308.

10. Chabou S, Iglewski M. Combination of conditional random field with a rule based method in the extraction of PICO elements. BMC Med Inform Decis Mak. 2018;18:128. doi:10.1186/s12911-018-0699-2.

11. Jin D, Szolovits P. Advancing PICO Element Detection in Biomedical Text via Deep Neural Networks. Bioinformatics. 2018;36:3856–62. http://arxiv.org/abs/1810.12780. Accessed 6 Feb 2021.

12. Sutton C, McCallum A. An introduction to conditional random fields. Found Trends Mach Learn. 2011;4:267–373. doi:10.1561/2200000013.

13. Brockmeier AJ, Ju M, Przybyła P, Ananiadou S. Improving reference prioritisation with PICO recognition. BMC Med Inform Decis Mak. 2019;19:256. doi:10.1186/s12911-019-0992-8.

14. Nye B, Yang Y, Li JJ, Marshall IJ, Patel R, Nenkova A, et al. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In: ACL 2018

- 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers). 2018. p. 197–207. doi:10.18653/v1/p18-1019.

15. Perozzi B, Al-Rfou R, Skiena S. DeepWalk: Online Learning of Social Representations. Proc ACM SIGKDD Int Conf Knowl Discov Data Min. 2014;:701–10. doi:10.1145/2623330.2623732.

16. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR. 2018. https://github.com/tensorflow/tensor2tensor. Accessed 21 Oct 2019.

17. Liao J, Ananiadou S, Currie GL, Howard BE, Rice A, Sena ES, et al. Automation of citation screening in pre-clinical systematic reviews. bioRxiv. 2018;:280131. doi:10.1101/280131.

18. Finkel JR, Grenager T, Manning C. Incorporating non-local information into information extraction systems by Gibbs sampling. In: ACL-05 - 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference. Association for Computational Linguistics (ACL); 2005. p. 363–70. doi:10.3115/1219840.1219885.

19. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. Association for Computational Linguistics (ACL); 2019. p. 319–27.

20. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in Neural Information Processing Systems. 2017. p. 5999–6009. http://arxiv.org/abs/1706.03762. Accessed 26 Aug 2019.

21. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2019. doi:10.1093/bioinformatics/btz682.

22. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, et al. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. 2020. http://arxiv.org/abs/2007.15779. Accessed 18 Sep 2020.

23. Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. 7th Int Conf Learn Represent ICLR 2019. 2017. http://arxiv.org/abs/1711.05101. Accessed 1 Oct 2020.

24. Howard J, Ruder S. Universal language model fine-tuning for text classification. In: ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers). 2018. p. 328–39. doi:10.18653/v1/p18-1031.

25. Zhang J, He T, Sra S, Jadbabaie A. Why gradient clipping accelerates training: A theoretical justification for adaptivity. 2019. http://arxiv.org/abs/1905.11881. Accessed 1 Oct 2020.

26. Ramshaw LA, Marcus MP. Text Chunking using Transformation-Based Learning. 1995;:157–76. http://arxiv.org/abs/cmp-lg/9505040. Accessed 7 May 2021.

27. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. Neural Architectures for Named Entity Recognition. 2016 Conf North Am Chapter Assoc Comput Linguist Hum Lang Technol NAACL HLT 2016 - Proc Conf. 2016;:260–70. http://arxiv.org/abs/1603.01360. Accessed 19 Apr 2021.

28. Pascanu R, Mikolov T, Bengio Y. On the difficulty of training Recurrent Neural Networks. 30th Int Conf Mach Learn ICML 2013. 2012; PART 3:2347–55. http://arxiv.org/abs/1211.5063. Accessed 18 Nov 2020.

29. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. 2013. http://ronan.collobert.com/senna/. Accessed 1 Apr 2019.

30. Pyysalo S, Ginter F, Moen H, Salakoski T, Ananiadou S. Distributional Semantics Resources for Biomedical Text Processing. Proc 5th Lang Biol Med Conf (LBM 2013). 2013;:39–44.

31. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. HuggingFace's Transformers: State-of-the-art Natural Language Processing. 2019. http://arxiv.org/abs/1910.03771. Accessed 13 Feb 2021.

32. Hiroki Nakayama. seqeval: A Python framework for sequence labeling evaluation. 2018. https://github.com/chakki-works/seqeval. Accessed 7 May 2021.

33. Ruder S, Plank B. Strong Baselines for Neural Semi-supervised Learning under Domain Shift. ACL 2018 - 56th Annu Meet Assoc Comput Linguist Proc Conf (Long Pap. 2018;1:1044–54. http://arxiv.org/abs/1804.09530. Accessed 16 Apr 2021.

34. Achakulvisut T, Acuna D, Kording K. Pubmed Parser: A Python Parser for PubMed Open-Access XML Subset and MEDLINE XML Dataset XML Dataset. J Open Source Softw. 2020;5:1979. doi:10.21105/joss.01979.

35. Gao S, Kotevska O, Sorokine A, Christian JB. A pre-training and self-training approach for biomedical named entity recognition. PLoS One. 2021;16 2 February. doi:10.1371/journal.pone.0246310.

# Figures

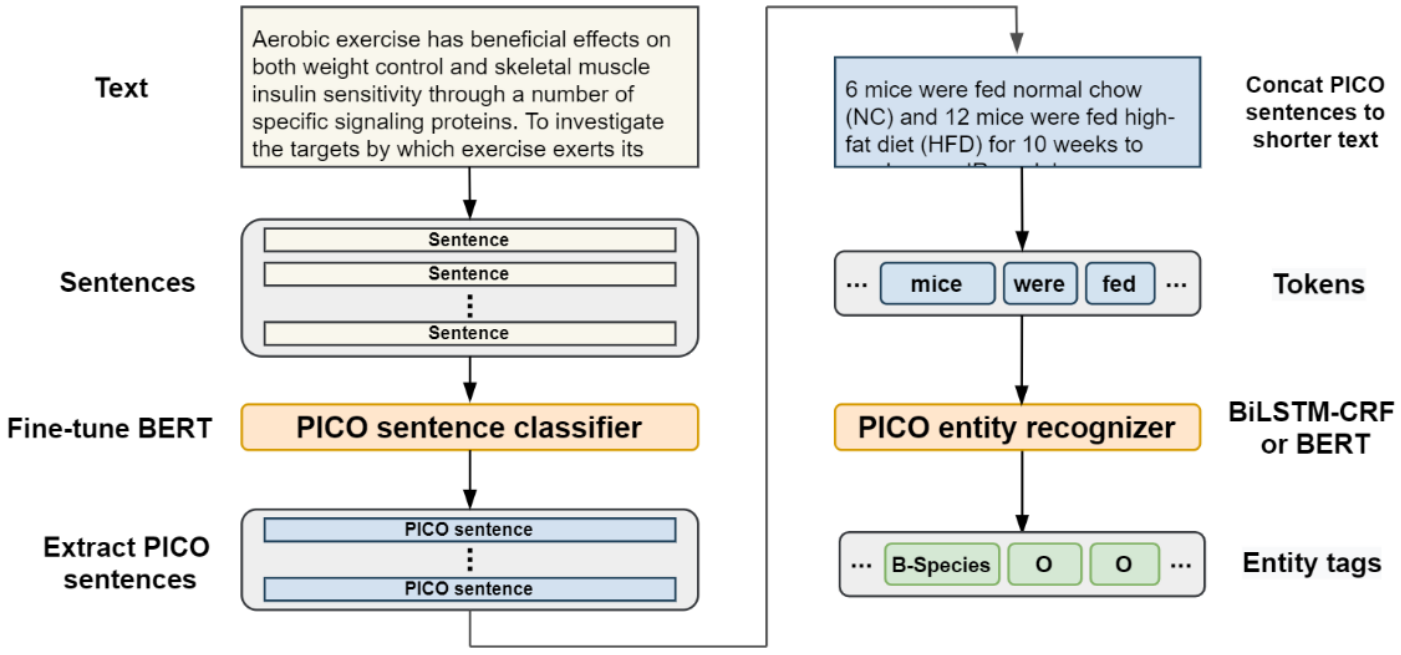
Proteomic Analysis of Skeletal Muscle in Insulin-Resistant Mice: Response to 6-Week Aerobic Exercise.

Aerobic exercise has beneficial effects on both weight control and skeletal muscle insulin sensitivity through a number of specific signaling proteins. To investigate the targets by which exercise exerts its effects on insulin resistance, an approach of proteomic screen was applied to detect the potential different protein expressions from skeletal muscle of insulin-resistant mice after prolonged aerobic exercise training and their sedentary controls. Eighteen C57BL/6 mice were divided into two groups: 6 mice were fed normal chow (NC) and 12 mice were fed high-fat diet (HFD) for 10 weeks to produce an IR model. The model group was then subdivided into HFD sedentary control (HC, n = 6) and HFD exercise groups (HE, n = 6). Mice in HE group underwent 6 weeks of treadmill running. After 6 weeks, mice were sacrificed and skeletal muscle was dissected. Total protein (n = 6, each group) was extracted and followed by citrate synthase, 2D proteome profile analysis and immunoblot. Fifteen protein spots were altered between the NC and HC groups and 23 protein spots were changed between the HC and HE groups significantly. The results provided an array of changes in protein abundance in exercise-trained skeletal muscle and also provided the basis for a new hypothesis regarding the mechanism of exercise ameliorating insulin resistance.

Intervention Comparator Outcome Induction Species Strain

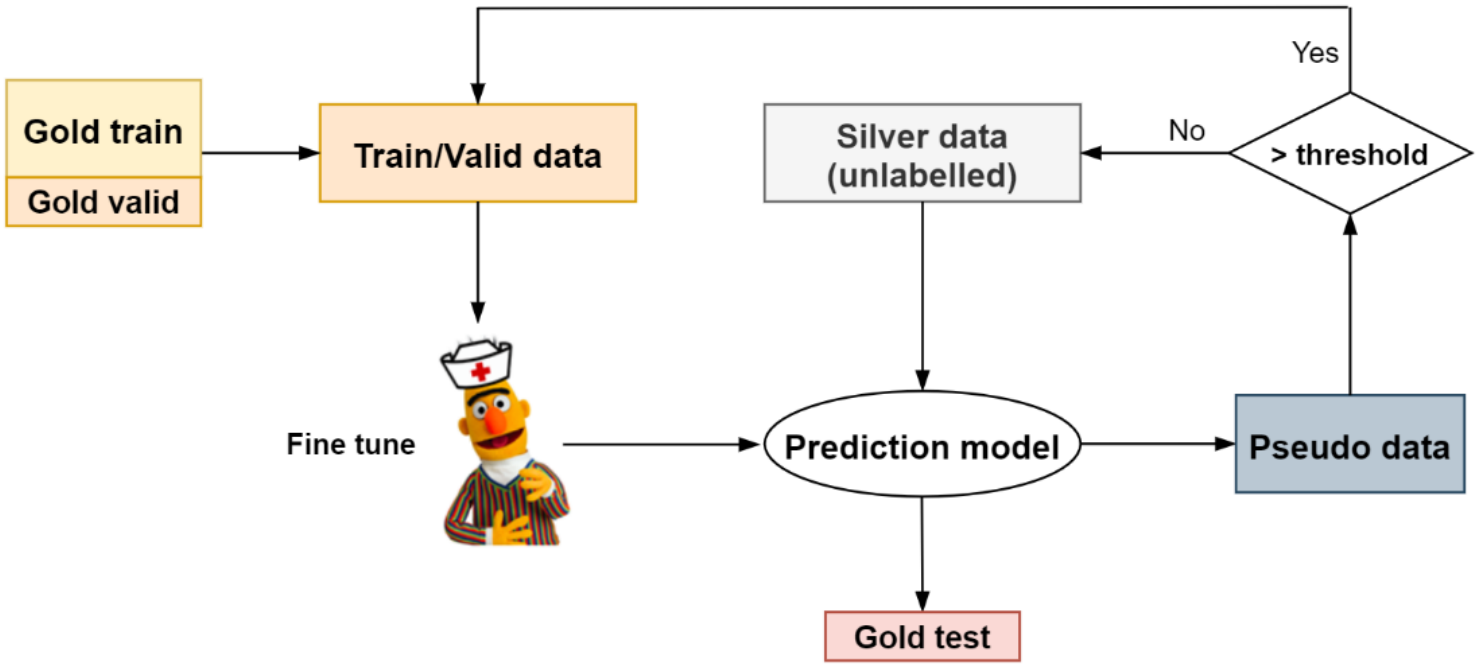
## Figure 1

Preclinical PICO annotation example. Screenshot from tagtog.



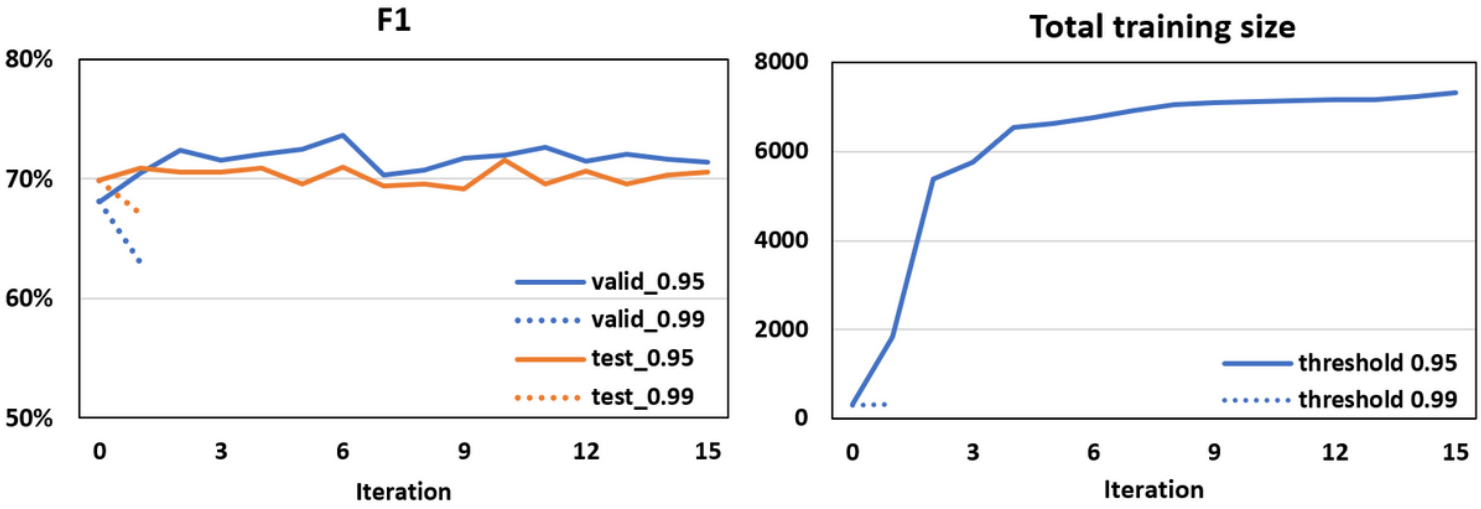

## Figure 2

The workflow of PICO extraction.




## Figure 3

The workflow of self-training in our experiments.




## Figure 4

Performance of PubMedBERT for PICO entity recognition using self-training.

# PICO extraction for in vivo abstract

Input one PMID:

27231887

## Extracted PICO text

In this study, we investigated the potential of the ethanolic extract of Acanthopanax koreanum Nakai (AK) to protect against experimental alcoholic liver disease in a mouse model that couples diet and daily ethanol bolus gavage. Fifty-six C57BL/6J mice were randomly divided into seven groups: normal control (NC), alcohol control (AC), alcohol/HFD control (AH), low-dose (1%) AK in alcohol group (ACL), high-dose (3%) AK in alcohol group (ACH), low-dose AK in alcohol/HFD group (AHL), and high-dose AK in alcohol/HFD group (AHH). The AH group showed more severe damage than the AC group in terms of biochemical and molecular data that were observed in this study. The administration of AK exerted remarkable effects in: plasma ALT ($p < 0.0001$), total lipid ($p = 0.014$), TG ($p = 0.0037$) levels; CPT-1α ($p = 0.0197$), TLR4 ($p < 0.0001$), CD14 ($p = 0.0002$), IL-6 ($p = 0.0264$) and MCP-1 ($p = 0.0045$) gene expressions; and ALDH ($p < 0.0001$) and CAT ($p = 0.0076$) activities.

## Extracted PICO phrases

Species: `{'mouse', 'mice'}`

Strain: `{'C57BL/6J'}`

Induction: `{'alcohol/HFD', 'alcohol'}`

Intervention: `{'Acanthopanax koreanum Nakai (AK)', 'AK'}`

Outcome: `{'TLR4', 'CD14', 'IL-6', 'severe damage', 'plasma ALT', 'CPT-1α', 'CAT (p=0.0076) activities', 'ALDH', 'TG (p=0.0037)', 'total lipid', 'MCP-1 (p=0.0045) gene expressions'}`

**Figure 5**

The visualization Streamlit app.

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix.docx