Evaluation of SURUS: a named entity recognition NLP system to extract knowledge from interventional study records

Peeters, C.; Vijverberg, K.; Pouwer, M.; Westerman, B.; Boot, M.; Verberne, S.

BMC Medical Research Methodology

RESEARCH

Open Access

Evaluation of SURUS: a named entity recognition NLP system to extract knowledge from interventional study records

Casper Peeters[1*], Koen Vijverberg[1], Marianne Pouwer[1], Bart Westerman[2], Maikel Boot[1] and Suzan Verberne[3]

Abstract

Background Medical decision-making commonly is guided by evidence-based analyses from systematic literature reviews (SLRs). These require large amounts of time and subject matter expertise to perform. Automated extraction of key datapoints from clinical publications could speed up the process of systematic literature review assembly. To this end, we built SURUS, a named entity recognition (NER) system comprised of a Bidirectional Encoder Representations from Transformers (BERT) model trained on a fine-grained dataset. The aim of this study was to assess the quality of SURUS classifications of PICO (patient, intervention, comparator and outcome) and study design elements of clinical study abstracts.

Methods The PubMedBERT-based model was trained and evaluated using a dataset of 39,531 labels amongst 400 clinical abstracts, with an inter-annotator agreement of 0.81 (Cohen's κ) and 0.88 (F1). The labels were manually annotated using a strict annotation guide. We evaluated quality of the dataset and tested the utility of the model in the practise of systematic literature screening, by comparing SURUS predictions to expert PICO and design classifications. Additionally, we tested out-of-domain quality of the model across 7 other therapeutic areas and another study design.

Results The SURUS NER system achieved an overall F1 score of 0.95, with minor deviation between labels. In addition, SURUS achieved a NER F1 of 0.90 and 0.84 for out-of-domain therapeutic area and observational study abstracts, respectively. Finally, F1 of PICO and study design classifications was 0.89 with a recall of 0.96 compared to expert classifications.

Conclusion The system reaches an F1 score of 0.95 across 25 contextually different medical named entities. This high-quality in-domain medical entity prediction of a fine-tuned BERT-based model was the result of a strict annotation guideline and high inter-annotator agreement. This prediction accuracy was largely preserved during extensive out-of-domain evaluation, indicating its utility across other indication areas and study types. Current approaches in the field lack in the fine-grained training data and versatility demonstrated here. We think that this approach sets a new standard in medical literature analysis and paves the way for creating fine-grained datasets of labelled entities that can be used for downstream analysis outside of traditional SLRs.

*Correspondence:
Casper Peeters
casper.peeters@medstone.nl

Full list of author information is available at the end of the article

Keywords Language model, Evidence-based medicine, PICO, Systematic literature review, Natural language rrocessing, Bi-directional encoder representations from transformers, Named entity recognition

Background

Interventional trials are an important source of scientific data for medical decision-making. Unstructured data from trials are carefully evaluated in systematic literature reviews (SLRs), which are typically accompanied by meta-analyses. These efforts result in essential medical documents that help drive decision making in the medical field. The key purpose of SLRs is to provide scientific validity through the inclusion of the complete body of evidence to answer a specific research question through an evidence-based approach. As such, the generation of SLRs is an intricate process, during which a broad selection of literature in the field of study is manually screened for eligibility and evaluated for the quality of evidence. It is paramount that an SLR represents an exhaustive evaluation of an area within a scientific field, and it is of high importance that the task of eligibility assessment is scrutinized and performed with complete recall to avoid incorrect exclusion of evidence. This assessment is particularly important in the context of the growing body of scientific publications available; from 1960 onwards, the number of PubMed records has grown exponentially [1, 2]. Nowadays, the initial screening process for SLRs in active medical fields often includes more than 3000 scientific abstracts. This means that assuming one would be able to process abstracts at a pace of 1 per minute, this task alone would take a human at least 50 h of reading time. Given that the Cochrane Institute's guidelines for an SLR involve independent screening by two trained experts, a modern, manual SLR process places a disproportional workload on expensive medical experts, often resulting in months of full-time work and costs easily exceeding $100,000 per SLR [3, 4].

Clinical questions for evidence-based practice are typically structured according to elements of an established framework called PICO. For a clinical SLR project, information on the Patient, Intervention, Comparator, and Outcome (PICO) are defined to determine the scope of the work and the trials eligible for the project [5]. For example, the PICO0F0F0F[1] framework of a trial could comprise "acute coronary syndrome" (Patient), "rivaroxaban" (Intervention), "placebo" (Comparison) and "systolic blood pressure" (Outcome). Along with elements listed in the PICO framework, study design characteristics ("randomized") provide additional valuable insights for the selection of eligible studies [6, 7]. Hereafter, the combination of PICO and study design characteristics will be referred to as the PICOS framework or PICOS in short.

One of the challenges in the identification of elements of PICOS is their dependence on textual context. For example, "stroke" may refer to a criterium of study participants for their inclusion into the study (i.e. part of "Patient") or to an endpoint that is measured during the study (i.e. part of "Outcome"). It may also refer to related research, in which case its identification is of no use to the reviewer.

Over the past decades, the increasing popularity of machine learning (ML) models has given rise to the development of methods to speed up the SLR screening process. Some ML approaches rank scientific publications according to their eligibility to a research question, thus providing the reviewer with the option of a priority cut-off for screening [8–10]. Alternatively, ML methods can provide the reviewer with information on scientific publications, which can be used to include or exclude studies in further analyses. Specifically, ML-based natural language processing (NLP) methods may extract elements of PICOS from unstructured medical text or predict the eligibility of a study based on a set of eligible studies initially selected by the reviewer. Ultimately, accurate and complete extraction of study characteristics by ML models could enable reviewers to base their eligibility decisions on the model outputs during the screening process of an SLR. A subset of biomedical NLP methods currently focuses on named entity recognition (NER) classification techniques. Using NER, unstructured text is processed and words, expressions or sentences are labeled with pre-defined classes (e.g. diseases, drugs, etc.) [2].

Several approaches have been proposed for the extraction of elements of PICOS from clinical publication texts [11]. Despite their apparent advantages, these NLP tools currently have a few limitations: (1) valuable study design features are often not extracted (e.g. study duration and study size); (2) PICO-focused ML-solutions typically focus on prediction of relatively large text sequences, resulting in coarse-grained extraction of limited use to the reviewer; (3) the quality of current NER systems are insufficient to approximate expert reviewer eligibility assessment performance [12]; and (4) there are only few datasets available which are designed specifically for PICO extraction [13–16], but they are limited in terms of size and granularity, and models trained on these datasets lack performance required.

Another, more recent innovation that could add value in SLR practice is the emergence of large language models (LLMs), which can be leveraged to interpret and summarize large volumes of scientific texts. However, as high

[1] https://www.cochranelibrary.com/about/pico-search

recall is of paramount importance in the field of systematic literature screening, the utility of LLMs in this scope is limited due to the risk of hallucination and their mediocre performance at high-complexity annotation tasks [17–19].

In this paper, we provide an elaborate evaluation of SURUS, a BERT-based classification model fine-tuned on a fine-grained, manually annotated dataset of medical annotations. SURUS was designed for the extraction of PICOS elements from clinical texts. SURUS (which is not an acronym) was trained to classify 25 different annotation labels in the abstracts of interventional studies. In addition to this, the SURUS NER method design is intended to facilitate the extraction of the results of clinical endpoints using relation extraction. Currently, SURUS is being integrated into software for systematic literature selection and analysis by medical professionals and scientists. The purpose of the software is to identify relevant literature through the recognition of medical named entities and abstract sections.

Our primary aims are two-fold: first, to rigorously validate the detailed annotation method underlying our dataset, exploring how fine-grained annotations might influence model performance; and second, to assess the extent to which our system can recognize, interpret, and classify a diverse range of clinically significant entities, including elements of PICOS. As a secondary objective, we compare SURUS performance with classification accuracy of an LLM, instruction-tuned with the SURUS annotation manual.

We ask: can a carefully tuned BERT model capture the subtle contextual shifts of a wide range of medical entities in a way that is sufficiently reliable for practical application? To our knowledge, this represents the first deep learning-based system capable of extracting such a broad spectrum of clinically relevant information from text with accuracy suitable for practical clinical use.

Related work

Previously, data classification and categorization of scientific study records were experimented with using Support Vector Machines), Conditional Random Fields, Long Short-Term Memory or, more recently, Bidirectional Encoder Representations from Transformers (BERT) models. Examples of tools employing one or more of these techniques were recently reviewed and evaluated in a systematic review [11].

Whilst all of the models listed above have their advantages and drawbacks, the consensus is that transformer-based methods such as BERT combine high potential with relatively small (annotation) effort compared to alternatives [11]. BERT is a transformer encoder model, pretrained on a vast dataset of books and Wikipedia [20]. BERT models have shown superiority compared to BiLSTM models on several tasks including NER [21, 22]. Later on, BERT was expanded upon by adding biomedical scientific texts to its pretraining, including the specialized BERT-derivatives BioBERT [23], SciBERT [24] and PubMedBERT [25].

The most recent innovation in the field of NLP are generative large language models (LLMs). LLMs, such as GPT-4 and LLaMA, excel in summarization, contextualization and extrapolation of information from a wide range of scientific fields. In the medical field, contributions include summarization of medical texts, chat-bot mediated diagnosis and medical education [17, 26]. However, the generative nature of these models make them prone to hallucination and classification inaccuracy, which is undesirable in a task demanding extensive classification recall [17, 18]. In the context of classification, BERT-like models have still shown superiority over LLM models [19] and tuning of LLMs for a task with complex instructions has proven challenging [27]. For this reason, we decided to use a variant of BERT as the classification model of choice for validation of the quality of the dataset presented here.

BERT-based models can be fine-tuned to perform well in specialized supervised learning tasks. Manual, task-specific labeling for fine-tuning a model is work-intensive and requires expert knowledge of the task and domain. In addition, currently available datasets for NER are often of limited quality and consistency [28].

To our knowledge, there are currently 3 datasets publicly available for recognition of PICO specifically. First, Kim et al. created the NICTA-PIBOSO dataset, which consists of 1000 abstracts with manually labeled sentence annotations amongst 5 label classes [29]. Second, Jin et al. presented the PubMed-PICO dataset,[2] consisting of almost 25,000 abstracts of which relevant sentences were automatically assigned to 1 of 7 labels using a rule-based algorithm [30]. Third, Nye et al. [31] reported the EBM-NLP corpus,[3] which consists of 5190 abstracts of scientific publications, 190 of which are annotated by experts and 5000 by laymen, using Population, Intervention and Outcome labels. The EBM-NLP corpus was used to train PICO-extracting systems on a sentence [32, 33] and span level [34, 35].

For SURUS to be able to accurately and concisely predict elements of PICOS, we created a dataset that offers the following advantages: (1) the annotation approach is suitable for word-level extraction; (2) we distinguish 25 different labels allowing for fine-grained extraction of PICOS characteristics; (3) the dataset presented here consists exclusively of expert-annotated labels; (4) our

[2] https://paperswithcode.com/dataset/pubmed-pico-element-detection-dataset

[3] https://github.com/bepnye/EBM-NLP

dataset is designed in a way that would facilitate extraction of detailed study results through relation extraction in a later phase of SURUS development.

Methods

Dataset

For our dataset, we used a set of scientific articles abstracts, publicly available in the PubMed database4F.[4] PubMed is the most widely used source of clinical evidence and consists of the Medline and PMC databases. Our dataset consisted of abstracts of interventional study reports. Interventional studies are characterized by investigation of a medical intervention and group distinction is typically based on differences in therapeutic regimen [36]. Though of similar study type, the style of reporting may vary greatly between therapeutic areas and interventional study subtypes. For this reason, abstracts included in the SURUS dataset were of various interventional study subtypes and therapeutic areas.

To ascertain high versatility, 4 of the most important therapeutic areas as reported in WHO ICD-11[5] were selected to be represented in the dataset: cardiovascular diseases, endocrine diseases, neoplasms and respiratory diseases. In total, 400 article abstracts of interventional studies (100 for each therapeutic area) were randomly selected from the PubMed database for in-domain evaluation of the NER system. In addition, a set of 123 other article abstracts was randomly selected for out-of-domain therapeutic area (90) and study type (33) evaluation. During randomization, the aim was to achieve a fitting representation of the real-world diversity of interventional publication abstracts in our dataset.

Expert annotations in the NER dataset

The abstracts of these selected publications were manually annotated. During annotation, entities were labeled and assigned to one of 25 labels, amongst 7 label classes. Label classes that were not relevant to extraction of PICOS elements were designed for either extraction of additional valuable information outside of PICOS or extract entities of study results. In addition, an element of PICOS may consist of multiple labels. For example, "Population" may be composed of entities of the "Methodology Inclusion Criteria" but also "Disease Indication". We chose this structure to clearly define the contextual niche of every label class, and to add to the granularity and the utility of the predictions made. All label classes had distinct contextual dependencies and unique labels. A full overview of annotations in the dataset is visualized in Fig. 1, the mapping of labels to elements of PICOS and more detailed descriptions of the label class are available in Appendix table B. Correct labeling of text elements is dependent on the context of the element and annotations made in its vicinity. For example, when mentioned in the methods section, "overall survival" was labeled as an element of the label class 'Methodology', whereas it was labeled as an element of the 'Parameter' class in the results section. However, when "overall survival" was mentioned in the results section without any association with annotations of the 'Result' class (so without associated results), it was not labeled at all. These nuances add to the intricacy of the annotation process.

In total, the 400 scientific abstracts were labeled with 39,531 annotations, averaging 98.83 (± 29.70) annotations per abstract. The out-of-domain datasets consisted of 8,131 and 1,876 for the out-of-domain indication and study type datasets, respectively.

NER dataset annotation process

Master students with a pharmaceutical or biomedical background were tasked to annotate the scientific abstracts. To warrant the quality and consistency of the annotations made, we made four provisions: (1) a detailed annotation manual was assembled by the first author to guide the annotators; (2) all annotators followed a 2-day course, during which they were instructed about the annotation methodology and process; (3) all annotations made were reviewed by one of two expert annotators; (4) annotation consistency was manually monitored using an extensive set of restrictive rules for annotation span range and context.

The primary aims of the manual created was to facilitate complete extraction of PICOS elements and to promote consistency of annotations made between articles in different fields and of different designs. Due to the high diversity of contextual situations in medical articles, the assembly of the manual was an iterative process featuring regular 'consensus sessions', during which the judgment of one expert annotator was decisive. Figure 2 shows a fully annotated example of a study abstract. To facilitate the annotation process, a comprehensive annotation management system was developed, consisting of integrated frontend, backend, and database components. The frontend was implemented using Vue.js and Vuetify, while the backend was built in Python using FastAPI. Annotations were stored in a PostgreSQL database.

Inter-annotator agreement

To estimate the reliability of the data, we measured inter-annotator agreement (IAA) between four annotators (two expert annotators and two of the student annotators) on a randomly determined subset of the scientific abstracts. For IAA assessment, 35 scientific abstracts (5 for each therapeutic area in the annotated dataset) were randomly selected to be separately annotated by the four

[4] https://pubmed.ncbi.nlm.nih.gov/

[5] https://icd.who.int/browse11/l-m/en

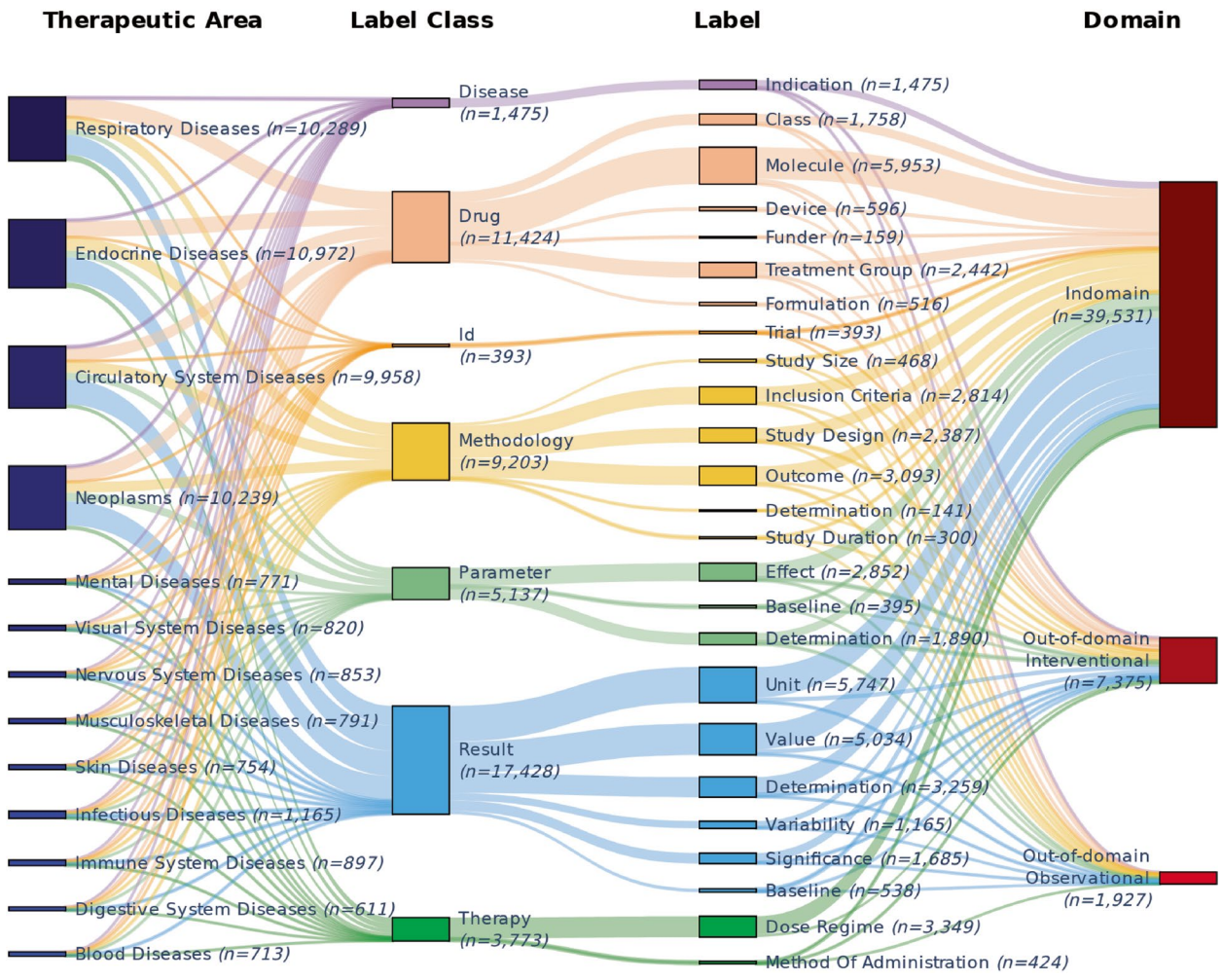

Fig. 1 Sankey diagram of all 48,833 expert annotations in the NER dataset. From left to right, the distribution of annotations is illustrated between different categories (nodes) of therapeutic area, annotation label class, annotation label and record domain. The width of the connections between nodes illustrate the number of overlapping annotations between nodes. Total number of annotations in nodes are listed between brackets behind the node name

annotators. Due to the abundance of unlabeled tokens in the dataset (introducing a positive bias), F1 on a token level was calculated in addition to Cohen's κ to approximate IAA, as unlabeled tokens may be left out of the calculation with this method [37–40]. Token-level agreement between annotators was a κ of 0.81 (± 0.05 between articles, amounting to substantial to almost perfect agreement [41]) and an F1 of 0.88 (± 0.01).

Dataset split into test and train sets

After annotation, all ($n = 400$) abstracts were split randomly into 10 partitions, each consisting of 10 articles from each therapeutic area (Appendix table C), which were used for model quality evaluation through k-fold cross-validation. Randomization of the abstracts was stratified by the presence of headers in the abstract, annotation-to-word ratio and the number of study arms. The Randomice tool was used for unbiased randomized stratification of records amongst the datasets [42].

Model

Input for the NER system

All abstracts of the NER dataset were tokenized using the BERT tokenizer6F6F6F[6] and subword token embedding tensors were assigned with the BERT base uncased model.[7] It is common for clinical publication abstracts to consist of more than 512 subword tokens. To resolve this issue of exceeding the BERT input limit of 512 subword tokens, we used a sliding window approach. Scientific abstracts longer than 512 subword tokens were divided into n batches of 512 subword tokens, with a 256 subword stride. The number of batches (n) was determined according to:

$$n = \left\lceil \frac{t}{256} - 1 \right\rceil$$

[6] https://huggingface.co/transformers/model_doc/bert.html#berttokenizer

[7] https://huggingface.co/bert-base-uncased

Fig. 2 An example of a fully annotated interventional study abstract of the record with Pubmed identifier 18204830. Different types of labels are colored according to their label class. The label class of different annotations is abbreviated in the image. For example, "P EFFECT" represents an 'Effect' label of the 'Parameter' class

where t was the total number of subword tokens. Any decimal result of the formula must be rounded up to an integer, as denoted by the ceiling symbols. For example, a scientific abstract of 1200 subword tokens was divided into 4 batches.

NER model training

The NER model was trained on all train set articles of the NER dataset. 512 subword tokens at a time were fed to BERT in the sliding-window approach. For training, a learning rate of $5*10^{-5}$ (momentum 0.99) with Adam optimization was used, training for 8 epochs using a batch size of 1. The system was trained to assign a BILOU tag and one of 25 labels, based on BERT prediction. Compared to more conventional BIO tags, BILOU tags (Beginning, Intermediate, Last, Outside, Unit) allow for a more granular dataset by distinguishing between single- and multiple-token chunks [43]. In the sliding window set-up, a BILOU tag and label of a subword could be predicted up to 2 times (the label predicted may differ between predictions, due to the context difference between strides). During post-processing, the average of the probabilities for each label predicted between batches was taken as the final prediction, and the label with the highest probability was assigned to the token. Finally, adjacent tokens with the same annotation label were aggregated into a single annotation according to their BILOU classification pattern.

Quality evaluation

Evaluation of the model quality was done by calculation of the precision, recall and F1 (Eq. 1) of the model output compared to annotations in the test set. NER evaluation was done on the entity level with only complete matches as true positives. A complete match was defined by a token start, token end and label match between predicted and true labels. As such, the corresponding label class but different prediction onset or end compared to the annotation was insufficient for a complete match. For example, a span classified by NER as 'Inclusion Criteria' and annotated as 'Outcome' did not yield a full match, even though both are of the 'Methodology' label class. Similarly, comparison of a prediction of "complete remission" with an annotation of "remission", both in the 'Effect' label, yielded a false positive.

Experiments

The system quality was evaluated in two settings: in-domain and out-of-domain quality. All in-domain metrics reported were the result of tenfold cross-validation. Quality assessment was based on the F1 mean and standard deviation over the different labels resulting from the set of measurements. First, we present the experimental setup of in-domain evaluation of the NER and section prediction systems. Subsequently, we describe experiments concerned with consistency and out-of-domain quality. Finally, we describe the protocol of a utility study comparing expert PICOS annotations with the system.

Equation 1 — Equations describing calculations of precision (left), recall (middle) and F1 measure (right) using true positives (tp), false positives (fp) and false negatives (fn).

$$p = \frac{tp}{tp + fp} \; r = \frac{tp}{tp + fn} \; F1 = \frac{2p \cdot r}{p + r} \tag{1}$$

In-domain quality evaluation

For evaluation of in-domain system quality, the F1 measure of the system was evaluated on a test set of abstracts describing a similar therapeutic area. Evaluation of in-domain quality consisted of four phases: (1) the optimal BERT model for the task was selected through experiments; (2) the quality of the section prediction system was measured; (3) its added value to the F1 of the NER model was evaluated; and (4) using the optimal model, the robustness of the dataset was evaluated.

First, the optimal BERT model to be used during further experimentation was determined. NER quality of four pretrained BERT models (BERT base and domain-specific alternatives BioBERT [23], SciBERT [24] and PubMedBERT [25]) was tested through tenfold cross-validation, using the train-test splits of all 400 annotated abstracts as specified in Sect. 2.1.4. Based on the resulting F1, the best performing model was selected to be used in the remainder of the experiments. The selected optimal model was the one with the highest mean F1 score between runs.

We assessed the effect of a smaller training set on the in-domain NER prediction quality of the optimal NER system. Prediction quality was compared between systems using 2, 3, 4, 5 and 7 batches as training set (each batch consists of 10% of all dataset articles). This was done using tenfold cross-validation, where each training fold consisted of block number k as the testing set and block numbers $[k+1 \ldots k+n+1]$ as the training set where k was the fold number and n was the number of batches included in the training set.

Out-of-domain quality evaluation

We assessed the quality of the SURUS for abstracts either on another subject or of a different type than the ones included in the annotated training set. For this, we tested the performance on two out-of-domain test sets: one on out-of-domain therapeutic areas and another one on out-of-domain observational study types. For each out-of-domain NER experiment, the SURUS system was tested on abstracts manually annotated by experts as out-domain test sets, according to the annotation rules applied during the annotation of the in-domain dataset. In the out-of-domain therapeutic area test set, we randomly included 10 article abstracts from 9 ICD-11 therapeutic areas not included in the in-domain dataset. In the out-of-domain observational study type dataset, we randomly included 33 abstracts of various observational study types. Amongst the observational study types of the included articles were cohort studies, case–control studies, diagnostic accuracy studies and case studies. Abstracts included in type out-of-domain quality evaluation were of the same therapeutic area as the ones included in the SURUS dataset. A detailed overview of the composition of the out-of-domain NER datasets is provided in Appendix table C.

Utility of SURUS

To determine the utility of the dataset in the workflow of a systematic literature review specialist, we compared SURUS predictions to expert-determined PICOS characteristics of interventional studies. For this evaluation, we worked with elements of PICOS from Cochrane published in a systematic literature review. 8 study records (2 for each therapeutic area included in the dataset) were randomly picked from 8 Cochrane systematic literature reviews. The Cochrane-assigned elements of PICOS were extracted from the "Characteristics of studies" section. Any element of study design or patient eligibility of the included studies mentioned in the methods section of the Cochrane review was also added to the experiment. Elements of intervention and comparison were merged as these show very limited contextual differences.

To appropriately compare Cochrane classifications to SURUS predictions, two preparatory steps preceded the comparison:

1. All Cochrane-determined elements were manually screened for presence in the study abstract. Any element not present in the abstract was excluded from the experiment. This step was included because Cochrane experts make use of the full record rather than the abstract to determine elements of PICOS.
2. SURUS predictions were mapped manually to Cochrane-assigned elements, as Cochrane-assigned elements may use different wording compared to the abstracts. The full mapping for the experiment is documented in Appendix table D.

After these steps, the precision, recall and F1 of the SURUS predictions were calculated. For these calculations, the metrics were defined as follows:

- *True positives* were unique predictions correctly mapped towards the correct constituent of PICOS.
- *False positives* were unique predictions that are either not mapped or mapped to the wrong element of PICOS.
- *False negatives* were elements of PICOS to which no prediction of SURUS was mapped or for which

elements of SURUS inadequately describe the content.
- *True negatives* were not included in the calculation of F1, which is designed to monitor the accuracy of positive predictions.

LLM performance at NER task

We tested the performance of a state-of-the-art LLM model at performing the NER task and compared it to SURUS. For this, we presented GPT-4o with a textual version of the annotation user manual, and we prompted the model for annotation of a sample of 2 test abstracts for every therapeutic area included in the in-domain dataset. In total, 8 abstracts were included in the comparison.

Availability

The full code for NER training, the full NER dataset and the detailed annotation guideline for reproduction efforts are available at our git repository.[8]

Results

We report the results of experiments regarding the quality, robustness and out-of-domain viability of SURUS. The experimental results are listed in the following order: results of the in-domain evaluation (1); results of out-of-domain evaluation (2); results of a utility case-study (3). Recall, precision and support for all classes of all evaluations are listed in Appendix table D.

PubMedBERT performs superior compared to other BERT variants when fine-tuned on SURUS dataset

To determine the optimal BERT model for SURUS, we compared the F1 using BERT, BioBERT and PubMedBERT on the full NER dataset. BioBERT and PubMedBERT showed similar prediction quality overall with an F1 of 0.95, as well as for the predictions of entities from different label classes. The results of the evaluations are listed in Table 1 and more detailed result metrics are listed in Appendix table A. Both models improved NER F1 compared to BERT for all annotation classes and compared to SciBERT for most label classes. The fine-tuned NER systems showed high prediction accuracy for Drug and Methodology, the label classes most commonly featured in PICOS. BioBERT and PubMedBERT performed superior compared to BERT and SciBERT. We expected that the performance of a PubMedBERT-fine-tuned model would extrapolate better for an out-domain task compared to a BioBERT-fine-tuned model, considering its specialization on PubMed texts. For this reason, we decided to use PubMedBERT for the remainder of the dataset validation.

Prediction quality plateaus at training on 70% of dataset items

To assess the rigidity of the annotation method, and the feasibility of further improving F1 by adding more training data, we fine-tuned the SURUS model leaving out varying percentages of the training set. High prediction quality was reached using a small selection of training data (F1 > 90% using 20% of the dataset for training, Fig. 3). For all categories, F1 mean and variability increased gradually with increasing dataset use, with the highest F1 and lowest variability eventually reached using the full train set (90% of the dataset).

Prediction quality was largely upheld testing out-of-domain abstracts

To evaluate the feasibility of using the system on other types of abstracts than the ones included in the dataset, we assessed the F1 on abstracts of out-of-domain therapeutic areas and observational study type (Table 2). The F1 of the fine-tuned model on the out-of-domain therapeutic area dataset was 0.90. Similar to the in-domain evaluation, prediction of the Parameter label class appeared to be most inconsistent in the observational dataset relative to the other label classes (the out-of-domain study type), the model scored an overall F1 of 0.84.

Table 1 F1 scores and standard deviations between folds of the 10-fold cross-validation of NER with BERT and 3 science domain-specific derivatives BioBERT, SciBERT and PubMedBERT

Label class	BERT	BioBERT	SciBERT	PubMedBERT
Disease	0.92 (±0.03)	0.95 (±0.02)	0.94 (±0.01)	0.95 (±0.02)
Drug	0.93 (±0.01)	0.95 (±0.02)	0.95 (±0.01)	0.95 (±0.01)
Identifier	0.95 (±0.02)	0.97 (±0.03)	0.97 (±0.02)	0.98 (±0.01)
Methodology	0.91 (±0.02)	0.94 (±0.01)	0.93 (±0.01)	0.94 (±0.01)
Parameter	0.81 (±0.04)	0.87 (±0.02)	0.86 (±0.03)	0.87 (±0.03)
Result	0.96 (±0.01)	0.98 (±0.00)	0.98 (±0.00)	0.98 (±0.00)
Therapy	0.90 (±0.03)	0.93 (±0.03)	0.93 (±0.03)	0.94 (±0.02)
Weighted Average	0.92 (±0.01)	0.95 (±0.01)	0.94 (±0.01)	0.95 (±0.01)

[8] https://github.com/surus-ai/dataset

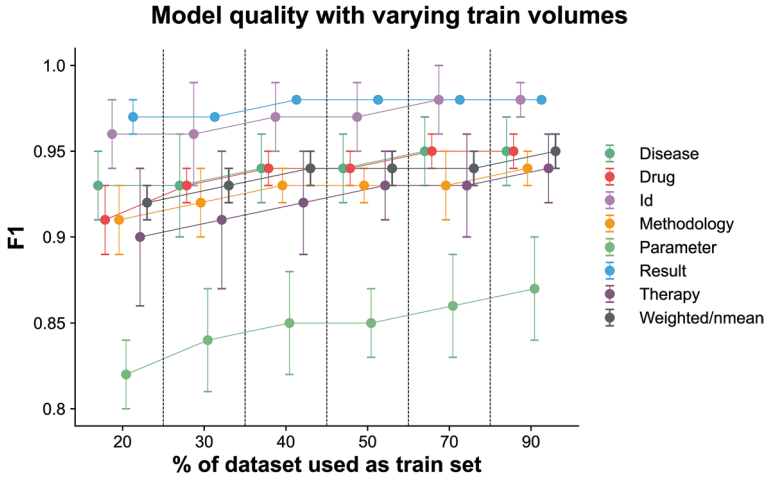

Fig. 3 Effect of limiting the volume of train data on the model quality. Weighted mean F1 does not dip below 0.9 even when 80 annotated abstracts are used for finetuning. Mean F1 steadily increases up to 0.95 with full use of train corpus (90% of the dataset). Individual label classes show a similar trend, with a relatively steep increase in context understanding for the Parameter label class, improving up to 0.05 in F1

Table 2 Out-of-domain evaluation metrics of PubMedBERT fine-tuned on the full SURUS dataset

Label class	Interventional				Observational			
	Precision	Recall	F1	Support	Precision	Recall	F1	Support
Disease	0.99	0.90	0.94	664	0.95	0.87	0.91	302
Drug	0.91	0.85	0.87	4,759	0.81	0.74	0.76	338
Id	1.00	0.98	0.99	341	1.00	1.00	1.00	15
Methodology	0.96	0.89	0.92	3,851	0.91	0.77	0.82	1,627
Parameter	0.83	0.76	0.79	3,003	0.78	0.68	0.73	1,345
Result	0.96	0.96	0.96	5,164	0.93	0.91	0.92	1,976
Therapy	0.97	0.85	0.90	1,273	0.33	0.50	0.40	2
Weighted Mean	0.93	0.88	0.90	19,055	0.88	0.80	0.84	5,605

Table 3 Utility assessment metrics, matching SURUS predictions to mapped Cochrane extracts of elements of PICOS

PICOS label	TP	FP	FN	P	R	F1
Participants	26	4	1	0.87	0.96	0.91
Interventions/Comparisons	32	3	2	0.91	0.94	0.93
Outcomes	27	9	1	0.75	0.96	0.84
Study Design	16	5	0	0.76	1	0.86
Weighted Mean	101	21	4	0.83	0.96	0.89

Abbreviations: *TP* True Positive, *FP* False Positive, *FN* False Negative, *P* Precision; R: Recall

High recall on PICOS classification task shows utility of SURUS

To assess the utility of SURUS in the practice of systematic literature screening, we compared SURUS predictions to Cochrane-assigned PICOS labels for 8 randomly chosen interventional abstracts for the relevant therapeutic area. The results of the experiment are shown in Table 3. The overall F1 of SURUS during the utility assessment was 0.89. Most false positive predictions could be attributed to prediction of entities that made no appearance in the Cochrane "Characteristics of Studies" section. The high recall reflected a minimal risk of missing relevant elements of PICOS.

Low F1 and high deviation of state-of-the-art LLM on NER task

The GPT-4o model performs worse than SURUS at the NER classification task, with a character-level F1 of 0.35 compared to 0.95 by SURUS on the subset of 8 articles. Evaluated on entity-level, the LLM performs worse with an F1 of 0.1 compared to 0.94 by SURUS. Full results of the comparison are listed in Appendix table E.

Discussion

In this paper, we evaluated a densely annotated and fine-grained medical dataset for finetuning NLP text classification models. We compared the quality of multiple BERT model variants, fine-tuned on this dataset to

identify named entities from clinical abstracts. Our measurements confirm that SURUS is capable of fine-grained classification and extraction of 25 different medically relevant categories, with a weighted mean F1 of 0.95 on interventional abstracts across 4 key therapeutic areas. The relatively high inter-annotator agreement (κ of 0.81) and the adequate out-of-domain performance of the fine-tuned underline the quality of the dataset. The high recall measured during the utility assessment demonstrate the value of SURUS to systematic literature reviewers in the screening process. The dataset and the annotation manual are available for non-commercial use and allow for expansion of the dataset for use in other domains.

To the best of our knowledge, of annotated medical NLP corpora published, the SURUS annotated dataset allows for the highest label prediction quality, for the largest diversity of clinical entity types. In addition, it shows the highest prediction quality of elements of PICOS as extracted by experts. This metric provides the key utility advantage of SURUS, granting high, time-saving opportunities to systematic literature reviewers with low risk of missing relevant elements of PICOS.

Current classification model alternatives typically focus on sentence or sentence clause classification, leaving much of the interpretation to the scientist performing the screening. In addition, mapping such text strands towards an ontology is laborious and inefficient. The fine-grained extraction of 25 labels allows SURUS to provide the reviewer with more detailed information on the PICOS element of studies in their selection. Important study features, such as information on drugs and treatments (0.95), elements of methodology (0.94) and disease (0.95) are predicted with high reliability (likely due to their contextual consistency throughout medical reporting), with limited variation between runs of the k-fold validation and in-domain therapeutic areas. Prediction quality in the current paper exceeds the current state-of-the-art prediction quality on other datasets focused on clinical studies such as EBM-PICO (0.73, PubmedBERT [25]), NICTA-PIBOSO (0.57–0.91, BioBERT [44]) and comparable to PubMedPICO (0.85–0.99, BioBERT [44]), recognizing more granular text spans and more label classes in the process. In addition, SURUS is the only PICOS classification system for which the utility is assessed compared to mapped expert extractions, rather than annotation span comparison, which typically introduces a layer of subjectivity and inconsistency.

Out-of-domain therapeutic area evaluation of the model shows a modest drop of prediction quality from 0.95 to 0.90, with most of the important label classes retaining high prediction quality. This signifies the utility of SURUS to systematic researchers specialized in any therapeutic area. The prediction quality of SURUS falls off slightly for abstracts of observational studies compared to out-domain therapeutic area prediction (F1 of 0.84 vs 0.90). The discrepancy is likely because of the methodological and stylistic differences between study types. For example, in some observational studies diseases may be key study group differentiators, whereas in interventional studies, study groups are defined based on the therapeutic regimens received. In addition, there is substantial variety in writing style between different types of observational studies, which include study types such as diagnostic accuracy studies, cohort studies and case reports. Still, important NER class categories such as Disease and Methodology can relatively reliably be extracted from observational studies (F1 of 0.91 and 0.82, respectively).

Limitations of the approach include the low diversity of the train dataset, focusing on interventional studies on 4 of the most common therapeutic areas. Prediction of named entity labels is less accurate outside these domains, or the domain may require additional labels which are not defined in our methodology (for example, a designated label for animals used in animal studies). Researchers may want to consider adding to the fine-tuning dataset to improve SURUS performance on any other therapeutic area of interest. Furthermore, the complexity of the annotation process may represent a considerable hurdle to producing a significant contribution to the dataset. The annotation manual may need to be adjusted when processing other study types to reach similar prediction quality levels as is shown here (for example, there is no “intervention” in observational studies). Nevertheless, the current prediction quality offers perspective for additional fine-tuning efforts to improve the prediction quality of relevant medical labels in observational studies.

In our experiment, SURUS performs better than a state-of-the-art, instruction-tuned LLM model in classification of NER labels in accordance with our annotation manual. In general, it appears that the number of different annotation labels was too high and the instructions for label span cut-off were too complex for the LLM to approximate SURUS NER prediction accuracy. As LLMs will likely continue to improve, it would be interesting to see whether LLMs will, in the future, approximate SURUS classification accuracy through instruction tuning.

Conclusion

Our findings show that the SURUS system is well-suited to classify 25 different medically relevant entity labels in interventional study abstracts with high prediction quality. Combined, its predictions can be used to extract elements of PICOS from clinical abstracts with high accuracy. Prediction quality is highest for articles on indications the system is trained on but remains considerable when applying SURUS to other indications. In addition,

SURUS shows considerable practical utility when used to extract elements of PICOS from scientific abstracts, with very limited risk of failing to identify elements of PICOS.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12874-025-02624-z.

Supplementary Material 1.

Acknowledgements
Khadija Ahmiane and Finn Bohte contributed to the assembly of the inter-annotator agreement set, Chayenne van Dongen assisted in the assembly of out-of-domain datasets. Finally, we would like to acknowledge all the master students who participated in the annotation process of the NER dataset.

Authors' contributions
C.P. contributed to experimental design and performed the experiments, wrote the manuscript and designed and imple-mented the annotation strategy. K.V. built the model and designed the infrastructure for the annotation. M.P. assisted in the experimental design and contributed to the writing/reviewing process. B.W. and M.B. contributed to reviewing and editing the manuscript. S.V. advised on experimental design, contributed to the writing process and reviewed/edited the manuscript.

Funding
This work was funded by Medstone Science B.V., Amsterdam, the Netherlands. The funding bodies played no role in the design of the study, the collection, analysis, and interpretation of data and in writing the manuscript.

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Medstone Science, Amsterdam, The Netherlands
[2]Amsterdam University Medical Center (UMC), Amsterdam, The Netherlands
[3]Leiden Institute of Advanced Computer Science (LIACS), Leiden University, Leiden, The Netherlands

Received: 27 November 2024 / Accepted: 24 June 2025
Published online: 31 July 2025

References

1. Larsen PO, Von Ins M. The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. Scientometrics. 2010;84(3):575–603. https://doi.org/10.1007/s11192-010-0202-z.
2. Fleuren WW, Alkema W. Application of text mining in the biomedical domain. Methods. 2015;74:97–106. https://doi.org/10.1016/j.ymeth.2015.01.015.
3. Lefebvre C, et al. Chapter 4: searching for and selecting studies'. In: Cochrane Handbook for Systematic Reviews of Interventions, 6.2. Cochrane; 2021. Available: https://training.cochrane.org/handbook/current/chapter-04. Accessed 30 June 2021.
4. Michelson M, Reuter K. The significant cost of systematic reviews and meta-analyses: a call for greater involvement of machine learning to assess the promise of clinical trials. Contemp Clin Trials Commun. 2019;16:100443. https://doi.org/10.1016/j.conctc.2019.100443.
5. Higgins J, et al. Chapter 4: searching for and selecting studies. In: Cochrane Handbook for Systematic Reviews of Interventions, 6.2. Cochrane; 2021. Available: https://training.cochrane.org/handbook/current/chapter-04. Accessed: 30 Jun 2021.
6. Cooke A, Smith D, Booth A. Beyond PICO: the SPIDER tool for qualitative evidence synthesis. Qual Health Res. 2012;22(10):1435–43. https://doi.org/10.1177/1049732312452938.
7. Eriksen MB, Frandsen TF. The impact of Patient, Intervention, Comparison, Outcome (PICO) as a search strategy tool on literature search quality: a systematic review. J Med Libr Assoc. 2018;106(4):420. https://doi.org/10.5195/jmla.2018.345.
8. Scells H, Zuccon G, Koopman B, Deacon A, Azzopardi L, Geva S. A Test Collection for evaluating retrieval of studies for inclusion in systematic reviews. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. Shinjuku Tokyo Japan: Association for Computing Machinery; 2017. pp. 1237–1240. https://doi.org/10.1145/3077136.3080707
9. Wang S, Scells H, Mourad A, Zuccon G. Seed-driven document ranking for systematic reviews: a reproducibility study. 2021. arXiv: arXiv:2112.04090. Available: http://arxiv.org/abs/2112.04090. Accessed 23 Jan 2024.
10. Wang S, Scells H, Koopman B, Zuccon G. Neural rankers for effective screening prioritisation in medical systematic review literature search. In: Proceedings of the 26th Australasian document computing symposium. 2022. pp. 1–10. https://doi.org/10.1145/3572960.3572980.
11. Schmidt L, Mutlu AN, Elmore R, Olorisade BK, Thomas J, Higgins JP. Data extraction methods for systematic review (semi)automation: update of a living systematic review. F1000Res. 2023;10:401. https://doi.org/10.12688/f1000research.51117.2.
12. Marshall IJ, Wallace BC. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. Syst Rev. 2019;8(1):163. https://doi.org/10.1186/s13643-019-1074-9.
13. Lee GE, Sun A. A study on agreement in PICO span annotations. In: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, in SIGIR'19. New York: Association for Computing Machinery; 2019. pp. 1149–1152. https://doi.org/10.1145/3331184.3331352.
14. Uzuner O, Solti I, Xia F, Cadag E. Community annotation experiment for ground truth generation for the i2b2 medication challenge. J Am Med Inform Assoc. 2010;17(5):519–23. https://doi.org/10.1136/jamia.2010.004200.
15. Herrero-Zazo M, Segura-Bedmar I, Martínez P, Declerck T. The DDI corpus: an annotated corpus with pharmacological substances and drug–drug interactions. J Biomed Inform. 2013;46(5):914–20. https://doi.org/10.1016/j.jbi.2013.07.011.
16. Li J, et al. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database (Oxford). 2016;2016:baw068. https://doi.org/10.1093/database/baw068.
17. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. Front Artif Intell. 2023;6:1169595. https://doi.org/10.3389/frai.2023.1169595.
18. Chen Q, et al. An extensive benchmark study on biomedical text generation and mining with ChatGPT. Bioinformatics. 2023;39(9):btad557. https://doi.org/10.1093/bioinformatics/btad557.
19. Wang S, et al. GPT-NER: named entity recognition via large language models. 2023. arXiv: arXiv:2304.10428. https://doi.org/10.48550/arXiv.2304.10428.
20. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805. 2019. Available: http://arxiv.org/abs/1810.04805. Accessed 1 Jul 2021.
21. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. arXiv:1906.05474. 2019. Available: http://arxiv.org/abs/1906.05474. Accessed 21 Jul 2021.
22. Si Y, Wang J, Xu H, Roberts K. Enhancing clinical concept extraction with contextual embeddings. J Am Med Inform Assoc. 2019;26(11):1297–304. https://doi.org/10.1093/jamia/ocz096.
23. Lee J, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234–40. https://doi.org/10.1093/bioinformatics/btz682.
24. I Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. arXiv:1903.10676. 2019. Available: http://arxiv.org/abs/1903.10676. Accessed 11 Aug 2021.

25. Gu Y, et al. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthcare. 2020;3(1):1–23. https://doi.org/10.1145/3458754.
26. Wójcik S, Rulkiewicz A, Pruszczyk P, Lisik W, Poboży M, Domienik-Karłowicz J. Beyond ChatGPT: what does GPT-4 add to healthcare? The dawn of a new era. Cardiol J. 2023;30(6):1018–25. https://doi.org/10.5603/cj.97515.
27. Keloth VK, et al. Advancing entity recognition in biomedicine via instruction tuning of large language models. Bioinformatics. 2024;40(4):btae16. https://doi.org/10.1093/bioinformatics/btae163.
28. Li J, Sun A, Han J, Li C. A survey on deep learning for named entity recognition. arXiv:1812.09449. 2020. Available: http://arxiv.org/abs/1812.09449. Accessed 22 Nov 2021.
29. Kim SN, Martinez D, Cavedon L, Yencken L. Automatic classification of sentences to support evidence based medicine. BMC Bioinformatics. 2011;12(2):S5. https://doi.org/10.1186/1471-2105-12-S2-S5.
30. Jin D, Szolovits P. PICO element detection in medical text via long short-term memory neural networks. In: Proceedings of the BioNLP 2018 workshop. Melbourne: Association for Computational Linguistics; 2018. pp. 67–75. https://doi.org/10.18653/v1/W18-2308.
31. Nye B, et al. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. arXiv:1806.04185. 2018. Available: http://arxiv.org/abs/1806.04185. Accessed 22 Jul 2021.
32. L Schmidt L, Weeds J, Higgins J. Data mining in clinical trial text: transformers for classification and question answering tasks. arXiv:2001.11268. 2020. Available: http://arxiv.org/abs/2001.11268. Accessed 11 Aug 2021.
33. Nye BE, Nenkova A, Marshall IJ, Wallace BC. Trialstreamer: mapping and browsing medical evidence in real-time. Proc Conf. 2020: 63–69. https://doi.org/10.18653/v1/2020.acl-demos.9.
34. Yang Y, Agarwal O, Tar C, Wallace BC, Nenkova A. Predicting annotation difficulty to improve task routing and model performance for biomedical information extraction. arXiv:1905.07791. 2019. Available: http://arxiv.org/abs/1905.07791. Accessed 16 Aug 2021.
35. Kang T, Zou S, Weng C. Pretraining to recognize PICO elements from randomized controlled trial literature. Stud Health Technol Inform. 2019;264:188–92. https://doi.org/10.3233/SHTI190209.
36. Thiese MS. Observational and interventional study design types; an overview. Biochem Med Zagreb. 2014;24(2):199–210. https://doi.org/10.11613/BM.2014.022.
37. Brandsen A, Verberne S, Wansleeben M, Lambers K. Creating a dataset for named entity recognition in the archaeology domain'. In: Proceedings of the 12th Language Resources and Evaluation Conference. Marseille: European Language Resources Association. 2020. pp. 4573–4577. Available: https://www.aclweb.org/anthology/2020.lrec-1.562. Accessed 30 Jun 2021.
38. Deleger L, et al. Building gold standard corpora for medical natural language processing tasks. AMIA Annu Symp Proc. 2012;2012:144–53.
39. Grouin C, Rosset S, Zweigenbaum P, Fort K, Galibert O, Quintard L. Proposal for an extension of traditional named entities: from guidelines to evaluation, an overview'. In: Proceedings of the 5th Linguistic Annotation Workshop. Portland: Association for Computational Linguistics. 2011. pp. 92–100. Available: https://aclanthology.org/W11-0411. Accessed 5 Apr 2022.
40. Hripcsak G, Rothschild AS. Agreement, the F-measure, and reliability in information retrieval. J Am Med Inform Assoc. 2005;12(3):296–8. https://doi.org/10.1197/jamia.M1733.
41. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–74. https://doi.org/10.2307/2529310.
42. van Eenige R, Verhave PS, Koemans PJ, Tiebosch IA, Rensen PC, Kooijman S. RandoMice, a novel, user-friendly randomization tool in animal research. PLoS ONE. 2020;15(8):e0237096. https://doi.org/10.1371/journal.pone.0237096.
43. Ratinov L, Roth D. Design challenges and misconceptions in named entity recognition. In: Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009). Boulder: Association for Computational Linguistics. 2009. pp. 147–155. Available: https://aclanthology.org/W09-1119. Accessed 30 Mar 2022.
44. Jin D, Szolovits P. Advancing PICO element detection in biomedical text via deep neural networks. Bioinformatics. 2020;36(12):3856–62. https://doi.org/10.1093/bioinformatics/btaa256.

Publisher's Note