PICO Corpus: A Publicly Available Corpus to Support Automatic Data Extraction from Biomedical Literature

Faith Wavinya Mutinda, Kongmeng Liew, Shuntaro Yada, Shoko Wakamiya, Eiji Aramaki
Nara Institute of Science and Technology
{mutinda.faith_wavinya.mz2, liew.kongmeng, s-yada, wakamiya, aramaki}@is.naist.jp

Abstract

We present a publicly available corpus with detailed annotations describing the core elements of clinical trials: Participants, Intervention, Control, and Outcomes. The corpus consists of 1011 abstracts of breast cancer randomized controlled trials extracted from the PubMed database. The corpus improves previous corpora by providing detailed annotations for outcomes to identify numeric texts that report the number of participants that experience specific outcomes. The corpus will be helpful for the development of systems for automatic extraction of data from randomized controlled trial literature to support evidence-based medicine. Additionally, we demonstrate the feasibility of the corpus by using two strong baselines for named entity recognition task. Most of the entities achieve F1 scores greater than 0.80 demonstrating the quality of the dataset.

1 Introduction

Evidence-based medicine (EBM) is an approach where doctors and health care professionals use the best available research evidence to guide them in making clinical decision about the care of patients (Sackett, 1997). Meta-analyses are one of the essential tools in EBM because they provide the highest form of medical evidence (Cook et al., 1997). A meta-analysis is a statistical technique that combines results of different research studies to determine the effectiveness of a treatment. Despite their importance, meta-analyses are labor-intensive and time-consuming as they involve manually reading hundreds of unstructured research articles and extracting data from them (Jonnalagadda et al., 2015). The number of research articles is increasing rapidly making it difficult/impossible for researchers to keep up. For instance, a recent study showed that more than 50,000 research articles related to COVID-19 have been published and more articles are being published every day (Wang and Lo, 2021).

Machine learning and natural language processing (NLP) techniques to automate data extraction from biomedical literature and speed up dissemination of biomedical evidence have been widely studied. Although automatic (or semi-automatic) approaches for extracting data from research articles have been proposed, they are still not ready for practical use (Marshall and Wallace, 2019). This is because data extraction requires high accuracy, which may be difficult for automated systems to achieve. The scarcity of publicly available corpora, which are usually expensive to create, is one barrier to the development of high-performance systems.

This paper presents a publicly available[1] corpus annotated with the core components of clinical trials, i.e., Participants, Intervention, Control, and Outcomes (PICO). We annotate in detail numeric texts especially those that identify the number of participants having certain outcomes. The annotation of the numeric texts is important for statistical analysis to determine the overall effect of an intervention. Currently, the corpus consists of 1011 research abstracts extracted from the PubMed database. The abstracts are of randomized controlled trials (RCTs) related to breast cancer, which is one of the leading causes of deaths in the world[2]. We focus on RCTs as they are considered the gold standard for clinical research methods.

2 Related work

Although there are some corpora with PICO elements annotated in abstracts and full-text articles, most of the corpora are not publicly available. Kiritchenko et al. (2010) developed a dataset containing 182 full-text articles. They annotated 21 entities including treatment dosage, frequency, funding organization, grant number, and so on. Summerscales et al. (2011) created a corpus consisting of 263 abstracts and annotated the treatment groups, out-

[1] https://github.com/sociocom/PICO-Corpus

[2] https://www.who.int/news-room/factsheets/detail/cancer

Proceedings of the 1st Workshop on Information Extraction from Scientific Publications, pages 26–31
Nov 20, 2022. ©2022 Association for Computational Linguistics

comes, group sizes, and outcome numbers. Their work is close to our study as they attempted to identify outcome numbers and group sizes for the purpose of calculating summary statistics. The annotations are however less extensive and the corpus is not publicly available.

Since constructing large corpora is expensive, Wallace et al. (2016) employed a distant supervision approach to create a large corpora consisting of full-text articles. They also manually annotated 133 articles for evaluation. Although distant supervision is a cheap way to construct large datasets, the dataset's quality might be low.

Most of these previous datasets are not publicly available. Nye et al. (2018) developed the EBM-NLP corpus, which is one of the largest publicly available corpora. Their annotation was done by crowd-sourcing through Amazon Mechanical Turk and a small part (200 abstracts) was done by medical professionals. The corpus consists of about 5000 abstracts of RCTs mostly related to cardiovascular diseases, cancer, and autism. They however do not annotate numeric texts that identify the number of participants who had certain outcomes.

3 Corpus annotation

3.1 Dataset collection

The corpus in this study consists of abstracts extracted from PubMed[3]. PubMed is a free search engine that provides access to the MEDLINE database[4] that indexes abstracts for biomedical and life sciences articles. We extracted research abstracts related to breast cancer whose study type is RCT, and are not meta-analysis or systematic-reviews. This was achieved by using keywords such as "breast cancer," "randomized controlled," "randomised controlled," "meta-analysis," and "systematic review."

3.2 Annotation process

The research abstracts were manually annotated. The annotator was asked to read and label text spans that identify the PICO elements, i.e., Participants (P), Interventions (I), Control (C), and Outcomes (O). For each PICO category, we developed sub-categories to capture detailed information within each category. The PICO label hierarchy is shown in Figure 2. In total we annotated 26 sub-categories (entities) which are described below.

- **Participants**: we annotate text snippets that describe the characteristics of the participants in a study. We annotate eight entities that include the total number of participants in the study, the number of participants in the intervention group, the number of participants in the control group, condition, eligibility, age, ethnicity, and location. Although breast cancer is the main condition, some studies focus on treating conditions associated with breast cancer such as hair loss, bone loss, depression, and pain.

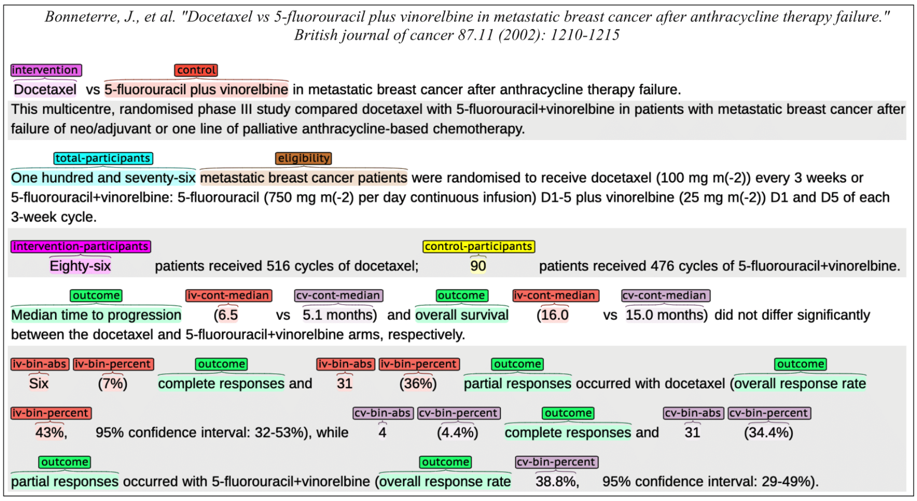
- **Intervention and Control**: we annotate text snippets that mention the specific intervention

Sub-category	Tag count	Number of abstracts
Participants (P)		
total-participants	1094	847
intervention-participants	887	674
control-participants	784	647
age	231	210
eligibility	925	864
ethinicity	101	83
condition	327	321
location	186	168
Intervention & Control (IC)		
intervention	1067	1011
control	979	949
Outcomes (O)		
outcome	5053	978
outcome-measure	1081	413
iv-bin-abs	556	288
cv-bin-abs	465	258
iv-bin-percent	1376	561
cv-bin-percent	1148	520
iv-cont-mean	366	154
cv-cont-mean	327	154
iv-cont-median	270	140
cv-cont-median	247	133
iv-cont-sd	129	69
cv-cont-sd	124	67
iv-cont-q1	4	3
cv-cont-q1	4	3
iv-cont-q3	4	3
cv-cont-q3	4	3

Table 1: Corpus statistics: The frequency of each entity (sub-category) and the number of abstracts in which each entity is found.

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

[4] https://www.nlm.nih.gov/medline/medline_overview.html



Bonneterre, J., et al. "Docetaxel vs 5-fluorouracil plus vinorelbine in metastatic breast cancer after anthracycline therapy failure." British journal of cancer 87.11 (2002): 1210-1215

Docetaxel vs 5-fluorouracil plus vinorelbine in metastatic breast cancer after anthracycline therapy failure.

This multicentre, randomised phase III study compared docetaxel with 5-fluorouracil+vinorelbine in patients with metastatic breast cancer after failure of neo/adjuvant or one line of palliative anthracycline-based chemotherapy.

One hundred and seventy-six metastatic breast cancer patients were randomised to receive docetaxel (100 mg m(-2)) every 3 weeks or 5-fluorouracil+vinorelbine: 5-fluorouracil (750 mg m(-2) per day continuous infusion) D1-5 plus vinorelbine (25 mg m(-2)) D1 and D5 of each 3-week cycle.

Eighty-six patients received 516 cycles of docetaxel; 90 patients received 476 cycles of 5-fluorouracil+vinorelbine.

Median time to progression (6.5 vs 5.1 months) and overall survival (16.0 vs 15.0 months) did not differ significantly between the docetaxel and 5-fluorouracil+vinorelbine arms, respectively.

Six (7%) complete responses and 31 (36%) partial responses occurred with docetaxel (overall response rate 43%, 95% confidence interval: 32-53%), while 4 (4.4%) complete responses and 31 (34.4%) partial responses occurred with 5-fluorouracil+vinorelbine (overall response rate 38.8%, 95% confidence interval: 29-49%).

Figure 1: An abstract with PICO elements annotated

and control used in the study. There are only two entities in this category.

- **Outcomes**: we annotate the outcome measures (primary and secondary end-points) and outcomes that were measured. We also aim to capture detailed information for the outcomes especially the numeric texts that identify the number of participants who experienced a particular outcome. In meta-analysis statistical analysis, these numeric texts are important for calculating summary statistics to ascertain the effectiveness of the intervention.

In the annotation, we mainly consider two types of outcomes, i.e, *binary outcomes* and *continuous outcomes*. Binary outcomes take two values such as the treatment was successful or failed, or survival (alive or dead). Continuous outcomes are not as straightforward as binary outcomes. Continuous outcomes such as pain are measured on a numerical scale (for instance, pain scores on a scale of 0 and 10). Continuous outcomes are usually measured at different time points (such as at baseline and at followup) and the results reported as mean, standard deviation, median, or quartiles.

We created labels to capture the various types of numeric texts in the intervention and control groups. We use "*iv*," "*cv*," "*bin*," and "*cont*" to represent intervention group, control group, binary outcome, and continuous outcome, respectively. In addition, binary outcomes numeric texts tend to be absolute values or percentage values. We use "*abs*" and "*percent*" to label absolute and percentage values respectively. Further, for the continuous outcomes, we also designed labels to capture the different types of numeric texts. We use "*mean*," "*sd*," "*median*," "*q1*," and "*q3*" to represent mean, standard deviation, median, first quartile, and third quartile respectively. In total, we have 16 entities for the outcomes. Figure 1 shows an example of an annotated abstract.

Binary outcome example:

- <*iv-bin-abs*>Four</*iv-bin-abs*> patients in the intervention group and <*cv-bin-abs*>two</*cv-bin-abs*> in the control group were <*outcome*>lost to follow-up</*outcome*>.

Continuous outcome example:

- <*outcome*>Depression scores</*outcome*> at follow-up were significantly lower in the exercise group (M = <*iv-cont-mean*>4.78</*iv-cont-mean*>, SD = <*iv-cont-sd*>3.56</*iv-cont-sd*>) compared to the control group (M= <*cv-cont-mean*>6.91</*cv-cont-mean*>, SD =<*cv-cont-sd*>5.86</*cv-cont-sd*>).

3.3 Corpus statistics

The corpus contains 1011 manually annotated abstracts. The annotation was performed using BRAT, an open-source web annotation tool (Stenetorp et al., 2012). The abstracts were annotated by two annotators. One of the annotators was hired from an annotation company and has extensive experience annotating medical documents and the second annotator is one of the authors. The first annotator annotated all the abstracts while the second annotator annotated 45% of the abstracts. The inter-annotator agreement was calculated based on Cohen Kappa and achieved a score of 0.72. Annotator disagreements were mainly found in the *outcome* and *eligibility* entities where the annotators had challenges in determining the start and end spans. How to minimize these disagreements during the annotation process is an important future work. Annotator disagreements for the other entities were minimal since they could be identified by one or two words and these disagreements are easy to resolve.

Currently the corpus has 17,739 entities and the frequencies of the annotated entities are shown in Table 1. The most frequent entity type is *outcome*, which comprises about 28% of all the annotations. Continuous outcomes quartile values (*q1* and *q3*) are the least frequent entity types. Table 1 also shows the number of abstracts containing each of the entities. The entities found in most abstracts are *intervention*, *outcome*, and *control* which are in 100%, 97%, and 94% of the abstracts, respectively. Most abstracts do not contain continuous outcomes values (*mean*, *median*, *sd*, *q1*, *q3*) and *ethnicity*.

4 Baseline experiments

We evaluate the corpus using named entity recognition (NER) task. This task is important for automatic information extraction from RCT research articles. Since deep learning language models have gained a lot of attention in NLP tasks, we adopt Bidirectional Encoder Representations from Transformers (BERT)-based models. BERT-based models have achieved state-of-the-art results in NLP tasks including NER (Devlin et al., 2018). These models are usually pre-trained on huge amounts of unlabeled data and can be fine-tuned to specific tasks. They use the encoder structure of the transformer which is an attention mechanism that learns contextual relations between words (or subwords).

We chose two pre-trained transformer-based baseline models, BioBERT (Lee et al., 2020) and LongFormer (Beltagy et al., 2020). BioBERT is initialized with general domain corpora and further trained on biomedical domain texts (PubMed abstracts and PubMed Central articles). LongFormer is pre-trained on general domain corpora including books, wikipedia, news, stories.

The 1011 abstracts were randomly split into 80% training data and 20% test data. As baseline experiments, we followed the standard BERT practice of formulating NER task as a sequential tagging task. Since neural networks provide different results when initialized with different seeds, we trained the models with five different seeds and averaged the results.

Sub-category	**Bio-BERT**	**Long-Former**
total-participants	0.94	**0.95**
intervention-participants	**0.85**	**0.85**
control-participants	**0.88**	**0.88**
age	0.80	**0.87**
eligibility	0.74	**0.88**
ethinicity	**0.88**	0.79
condition	**0.80**	0.79
location	0.76	**0.87**
intervention	**0.84**	**0.84**
control	0.76	**0.81**
outcome	0.81	**0.85**
outcome-measure	0.84	**0.90**
iv-bin-abs	0.80	**0.82**
cv-bin-abs	**0.82**	**0.82**
iv-bin-percent	**0.87**	0.86
cv-bin-percent	**0.88**	0.85
iv-cont-mean	0.81	**0.84**
cv-cont-mean	**0.86**	**0.86**
iv-cont-median	**0.75**	0.69
cv-cont-median	**0.79**	0.73
iv-cont-sd	0.83	**0.89**
cv-cont-sd	0.82	**0.89**
iv-cont-q1	0	0
cv-cont-q1	0	0
iv-cont-q3	0	0
cv-cont-q3	0	0

Table 2: NER models results in terms of F1 score

The performance of the models was evaluated using F1 score. Table 2 shows the results of the NER models. The models achieved satisfactory performance and several sub-categories achieved high F1 scores. *Total-participants* achieved the highest F1 score of 0.95. Most of the sub-categories achieved F1 scores greater than 0.80. The models could not predict for sub-categories with the lowest frequency (F1 score=0).

We performed an error analysis and identified misclassified entities and boundary detection as the common types of errors. In the case of misclassified entities errors, the models identified the correct boundaries but assigned the wrong entities. For example, *iv-bin-abs* was misclassified as *cv-bin-abs* and vice-versa. Boundary detection errors were common in the *outcome* and *eligibility* entities, where the models identified longer or shorter entities than those marked in the gold set.

5 Conclusion

We presented a publicly available corpus with detailed annotation of the PICO elements. The corpus contains 1011 abstracts related to breast cancer RCTs. The corpus provides detailed annotation for outcomes especially numeric texts to identify the number of participants having certain outcomes. This is important for statistical analysis to determine the effectiveness of a treatment. The corpus will facilitate NLP research on automatic information extraction from biomedical literature and contribute towards evidence-based medicine. Since the corpus consists of breast cancer related abstracts, one of the future works is to extend it to include other diseases. The corpus is publicly available at https://github.com/sociocom/PICO-Corpus.

Acknowledgment

This work was supported by JST CREST Grant Number: JPMJCR22N1, Japan.

References

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Deborah J Cook, Cynthia D Mulrow, and R Brian Haynes. 1997. Systematic reviews: synthesis of best evidence for clinical decisions. *Annals of internal medicine*, 126(5):376–380.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Siddhartha R Jonnalagadda, Pawan Goyal, and Mark D Huffman. 2015. Automating data extraction in systematic reviews: a systematic review. *Systematic reviews*, 4(1):1–16.

Svetlana Kiritchenko, Berry De Bruijn, Simona Carini, Joel Martin, and Ida Sim. 2010. Exact: automatic extraction of clinical trial characteristics from journal publications. *BMC medical informatics and decision making*, 10(1):1–17.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Iain J Marshall and Byron C Wallace. 2019. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic reviews*, 8(1):1–10.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain J Marshall, Ani Nenkova, and Byron C Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, page 197. NIH Public Access.

David L Sackett. 1997. Evidence-based medicine. In *Seminars in perinatology*, volume 21, pages 3–5. Elsevier.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Rodney L Summerscales, Shlomo Argamon, Shangda Bai, Jordan Hupert, and Alan Schwartz. 2011. Automatic summarization of results from clinical trials. In *2011 IEEE International Conference on Bioinformatics and Biomedicine*, pages 372–377. IEEE.

Byron C Wallace, Joël Kuiper, Aakash Sharma, Mingxi Zhu, and Iain J Marshall. 2016. Extracting pico sentences from clinical trial reports using supervised distant supervision. *The Journal of Machine Learning Research*, 17(1):4572–4596.

Lucy Lu Wang and Kyle Lo. 2021. Text mining approaches for dealing with the rapidly expanding literature on covid-19. *Briefings in Bioinformatics*, 22(2):781–799.

A Appendix

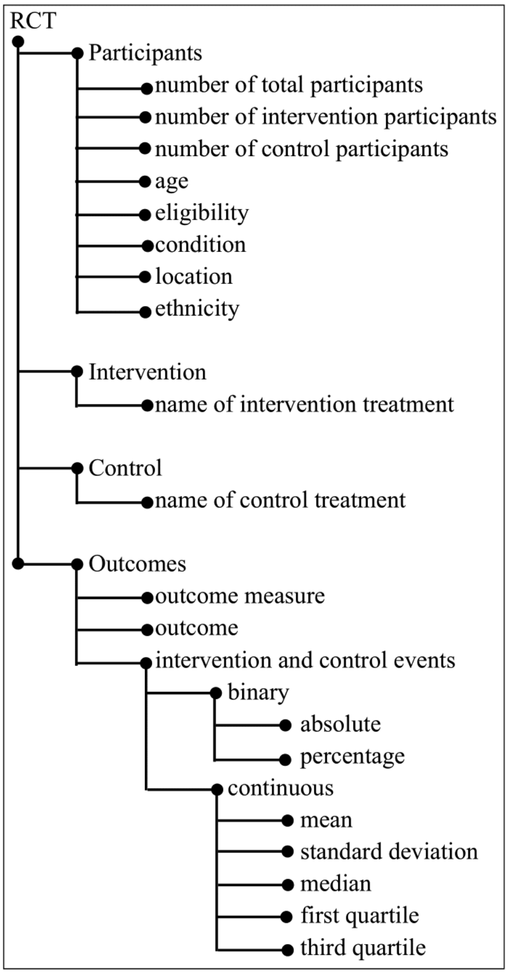


Figure 2: PICO label hierachy