

PICO Entity Datasets and Methodologies in the Context of Randomized Controlled Trials: A Case Study

Julien Bezançon and Pierre Zweigenbaum and Nona Naderi

Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique

91400, Orsay, France

{julien.bezancon,pierre.zweigenbaum,nona.naderi}@lisn.fr

Abstract

Randomized Controlled Trials (RCTs) contain crucial information about the outcomes of clinical experiments. However, the sheer number of RCTs—approximately 595,000 as of September 2023—makes it difficult to efficiently navigate the available evidence. To address this challenge, medical practitioners rely on the PICO framework to identify relevant information and NLP engineers have developed automated PICO extraction pipelines. In this work, we study four PICO-related datasets created from RCTs and commonly used in the literature. We then compare two PICO extraction pipelines: *PICOX_{reproduction}* and *BIOBERT_{overlaps}*. Our findings show that (i) PICO-related datasets can be broadly categorized as either span-based or entity-based, with clear differences in how PICO information is annotated, and (ii) while *PICOX_{reproduction}* is better suited to extracting text spans containing PICO information, *BIOBERT_{overlaps}* is more suitable for fine-grained and precise PICO entity extraction. Additionally, we evaluate the performance of both models on nested and overlapping PICO entities, found in the EBM-NLP dataset. *BIOBERT_{overlaps}* achieves an entity-level micro-F1 score of 30.70 on these entities, compared with 17.58 for *PICOX_{reproduction}*.

1 Introduction

Randomized Controlled Trials (hereafter RCTs) correspond to clinical experiments during which participants are split into two or more groups: one group receives a new treatment, while the others receive either a standard treatment or no treatment at all (Braga et al., 2025). The summary of these RCTs and their outcomes are then stored in an online database, such as ClinicalTrials.gov.¹ In addition, the results of RCTs are often disseminated through publications in scientific journals. RCTs

¹<https://clinicaltrials.gov/>

42 B-participant	yrs I-participant	Canadian I-participant
tumor B-intervention	marker I-intervention	subtypes I-intervntion
5-fluorouracil B-control	plus I-control	vinorelbine I-control
risk B-outcome	of I-outcome	relapse I-outcome

Table 1: Instances of annotated PICO entities, found in the PICO-CORPUS (Mutinda et al., 2022).

rely on evidence-based medicine to report essential biomedical findings.

Recent studies have shown a substantial increase in the number of published RCTs, making it increasingly difficult to navigate and synthesize the available evidence. As of September 2023, Hu et al. (2023) report approximately 595,000 RCTs indexed in PubMed. Given the vast and continuously growing body of evidence, medical practitioners commonly rely on the PICO framework (*Participant, Intervention, Control, and Outcome*) to efficiently identify and organize the key elements of a clinical study. Originally introduced by Richardson et al. (1995), the PICO framework facilitates the rapid identification of the main components of an RCT, including the study population, the intervention administered, the control or comparator condition, and the outcomes assessed.

However, manually extracting medical information from RCTs remains a labor-intensive and time-consuming process (Borah et al., 2017). This challenge has motivated the development of automated approaches for PICO extraction, including recent generative AI-based methods, such as the approach proposed by Reason et al. (2024). Traditionally,

PICO extraction has been formulated as a Named Entity Recognition (NER) task. In this setting, each token is (i) annotated according to a NER annotation scheme, such as IO, IOB or IOB2, and (ii) assigned a label corresponding to one of the PICO components: *participant*, *intervention*, *control* or *outcome*, as illustrated in Table 1.

In the context of FAIRClinical project,² our initial objective was to use the PICOX framework (Zhang et al., 2024) to automatically extract PICO entities from publications reporting clinical trials and clinical case reports. We selected this framework for two main reasons. First, the authors reported state-of-the-art performance among BERT-based models for PICO extraction. Second, PICOX was specifically designed to handle nested and overlapping PICO entities, which are not readily captured by conventional NER approaches. However, during our experiments, we identified several inconsistencies in both (i) the PICO datasets used and (ii) the PICOX pipeline itself, which motivated a closer examination of the framework. Our findings suggest that PICOX is better suited to extracting text spans containing PICO elements than to extracting precise PICO entities. Our contributions are as follows:

PICOX_{reproduction}• We reproduce and correct the experiments described in Zhang et al. (2024). We show that PICOX was trained and evaluated on span-based PICO datasets instead of entity-based PICO datasets, resulting in higher observed performances.

BIOBERT_{overlaps}• We introduce a new BERT-based pipeline, BIOBERT_{overlaps}, which allows for precise PICO entity extraction while taking into account overlapping and nested entities.

Our observations show that the span-based PICO datasets tend to contain fewer and longer annotated entities. Accordingly, we argue that developing a PICO extraction pipeline using span-based datasets can lead to downstream performance issues, particularly when precise PICO entity boundaries are required for subsequent tasks, such as entity linking and entity normalization (Kartchner et al., 2023; Pras and Naderi, 2026).

²<https://fairclinical.github.io/>

2 PICO Extraction Workflow

2.1 Related Work

Previous approaches developed for the PICO entity extraction task consist of CRF and/or LSTM architectures (Jin and Szolovits, 2020), sometimes combined with rule-based pipelines (Chabou and Iglewski, 2015, 2018) or transformer models (Dhrangadhariya et al., 2021). More recent approaches have explored span-based methods (Liu et al., 2021; Zhang et al., 2024), Large Language Models (LLMs) (Hu et al., 2023; Ghosh et al., 2024). Additionally, Nye et al. (2018) and Mutinda et al. (2022) introduced annotation methodologies and datasets designed to support more fine-grained PICO extraction. Villarmin et al. (2024) subsequently investigated the reproducibility of the PICO-CORPUS work introduced by Mutinda et al. (2022). Given the importance of dataset construction and reproducibility in PICO extraction, we consider four publicly available datasets in our experiments, which we describe in the following section.

2.2 PICO Datasets

We use the four benchmark datasets employed by Zhang et al. (2024): EBM-NLP (Nye et al., 2018), PICO-CORPUS (Mutinda et al., 2022), AD and COVID-19 (Hu et al., 2023). These datasets were constructed by annotating different subsets of RCTs and differ in their annotation schemes and the granularity of the PICO information they provide.

EBM-NLP. This dataset comprises approximately 5,000 abstracts describing RCTs covering cardiovascular diseases, cancer, and autism. The corpus provides two levels of PICO annotation: *starting_spans*, which identify text spans corresponding to PICO elements, and *hierarchical_labels*, which provide more fine-grained annotations within these spans. Table 2 illustrates a sequence annotated according to both annotation levels. When using the *starting_span* annotations for extraction, the entire sequence is extracted as a single span. In contrast, using the *hierarchical_labels* annotations results in four distinct entities: "atenolol", "labetalol", "metoprolol" and "pindolol". Thus, *hierarchical_labels* enables the precise, fine-grained extraction of PICO entities, whereas the *starting_spans* is better suited to identifying text segments containing PICO information.

token	<i>starting_spans</i>	<i>hierarchical_labels</i>
atenolol	B	B
,	I	O
labetalol	I	B
,	I	O
metoprolol	I	B
and	I	O
pindolol	I	B
.	I	O

Table 2: Example extracted from the EBM-NLP dataset (document ID: 43164). This example contains both *starting_spans* and *hierarchical_labels* annotations for the *outcome* label. The entities targeted for extraction are highlighted in red.

PICO-CORPUS. This dataset contains 1,011 PubMed abstracts of breast cancer RCTs. It was manually annotated using detailed guidelines, with 26 fine-grained entity sub-categories defined across the PICO components (Mutinda et al., 2022).

AD and COVID-19. Hu et al. (2023) introduced datasets of RCT abstracts focusing on Alzheimer’s disease and COVID-19. Zhang et al. (2024) subsequently used 150 RCT abstracts from each collection for PICOX and annotated them using a process grounded in the original EBM-NLP annotation guidelines, supplemented with additional rules and examples.

Table 3 reports the number of entities in each dataset. In addition to the four main PICO categories—*participant*, *intervention*, *control* and *outcome*)—EBM-NLP and PICO-CORPUS also provide finer-grained subcategories, such as *age* and *ethnicity*. To enable a consistent evaluation across all datasets, we unify their annotation scheme. Specifically, we (i) retain only the four main PICO categories and (ii) merge the *intervention* and *control* categories, as the entities annotated under these categories are highly similar in nature.

2.3 Entity-based and Span-based Datasets

We observe some differences in the annotations across datasets. Table 3 shows that *participant* entities in both the AD and COVID-19 datasets are, on average, 10 tokens longer than those in PICO-CORPUS, whereas *outcome* entities are approximately 1 token longer. Moreover, a statistical comparison of the two EBM-NLP annotation types reveals a substantial difference in annotation granularity. The *starting_spans* annotations contain 189,350 more tokens) than the *hierarchical_labels*

annotations—almost twice as many—while containing 2,586 fewer entities (2.92 %), as shown in Table 4).

As a result, we distinguish between two types of PICO datasets: entity-based PICO datasets and span-based PICO datasets. The *starting_spans* version of EBM-NLP, AD and COVID-19 were all annotated following roughly the same guidelines, designed for text span identification. Therefore, they correspond to span-based PICO datasets. In contrast, the *hierarchical_labels* version of EBM-NLP and PICO-CORPUS are based on fine-grained entity annotation guidelines. They both correspond to entity-based PICO datasets. Span-based datasets contain fewer and longer annotated spans than entity-based datasets, which may affect the performance of PICO entity extraction systems (see Section 5).

2.4 Nested and Overlapping Entities

To the best of our knowledge, EBM-NLP is the only dataset containing nested PICO entities. Among the entities within this dataset, we also identify cases where two or more entities share at least one common token, while neither entity is completely contained within the boundaries of the other. These instances correspond to overlapping entities, as defined by Dai (2018). Table 5 illustrates two examples of overlapping entities identified in the *hierarchical_labels* annotations of EBM-NLP dataset. In total, we count 2,609 nested entities and 120 overlapping entities in the EBM-NLP dataset, representing 3.08 % of the total number of entities in this dataset.

In the following Sections, we present two experiments: first, we reproduce the PICOX experiment (Section 3); second, we introduce our own pipeline, BIOBERT_{overlaps} (Section 4). Both of these approaches are designed to handle nested and overlapping entities.

3 PICOX Reproduction

In this Section, we aim to reproduce the experiments made by Zhang et al. (2024). The scripts of our reproduction are available online on GitHub.³

3.1 Model Description

PICOX relies on two models to extract PICO entities from RCTs. First, a span boundary detection

³<https://github.com/JulienBez/PICOXreproduction>

Dataset	Category	Train		Valid		Test	
		Total	Mean	Total	Mean	Total	Mean
Entity-based							
PICO-CORPUS	Participant	3,614	3.63	445	3.67	475	3.63
	Intervention	1,640	3.03	201	2.93	205	2.92
	Outcome	8,972	3.35	1,076	3.31	1,091	3.29
EBM-NLP hierarchical_labels	Participant	16,998	2.94	880	3.00	864	3.99
	Intervention	31,166	2.75	1,723	2.69	1,517	1.89
	Outcome	32,369	3.51	1,735	3.53	1,439	3.21
Span-based							
EBM-NLP starting_spans	Participant	16,104	9.75	827	9.20	632	7.66
	Intervention	30,716	3.91	1,597	3.70	1,724	2.19
	Outcome	31,171	4.75	1,541	5.00	1,793	4.51
AD	Participant	860	11.94	114	11.02	101	12.99
	Intervention	2,280	2.58	277	2.52	293	2.63
	Outcome	2,504	4.93	355	4.45	271	5.56
COVID-19	Participant	1,048	14.82	136	13.85	126	15.87
	Intervention	3,112	2.51	417	2.45	361	2.59
	Outcome	2,404	4.29	280	4.30	321	4.28

Table 3: Number of entities and mean token length of entities for each dataset, for each partition and for each entity category. We distinguish datasets focusing on entities only from datasets focusing on text spans containing entities.

	<i>starting_spans</i>	<i>hierarchical_labels</i>
Tokens	462,977	273,627
Entities	86,105	88,691
Mean	5.37	3.08

Table 4: Number of tokens belonging to an entity, entities and mean number of token per entity in EBM-NLP *starting_spans* and *hierarchical_labels* versions.

model is trained on a list of sentences containing PICO entities. Given an input sentence, the model predicts the start and end positions of potential PICO spans. The predicted boundaries are then passed to a second model, which performs span classification. For each candidate span, the classifier assigns a PICO label (either *participant*, *intervention* and/or *outcome*).

3.2 Data Processing

The datasets used for this reproduction are EBM-NLP, PICO-CORPUS, AD and COVID-19. We use these datasets to train and evaluate our models, as in Zhang et al. (2024). During the reproduction, we identified several inconsistencies in the use of these datasets that affected the reproducibility of the reported results:

EBM-NLP. We observe that PICOX was trained and evaluated on the *starting_spans* version of EBM-NLP.⁴ This choice can lead to down-

stream performance issues, as the *starting_spans* version contains PICO text spans rather than precise entities. For this reproduction, we use the *hierarchical_labels* version of EBM-NLP instead. Additionally, following Zhang et al. (2024), we randomly select 5 % of the training partition of the dataset to create a validation set.

PICO-CORPUS. Zhang et al. (2024) used 50 % of the dataset as the test set, but do not provide details on how this split was performed. Moreover, the original splits available for this dataset are not the same. Mutinda et al. (2022) use 20 % of the dataset to create the test set. We choose to use the original partitioning for this dataset to facilitate reproducibility.

AD and COVID-19. Multiple data splits are available for both the AD and COVID-19 datasets; however, Zhang et al. (2024) do not specify which split was used in their experiments. We therefore use the first split provided for each dataset in their official GitHub repository.⁵

Zhang et al. (2024) segment their datasets into sentences to train and evaluate their models. However, there was no indication regarding the sentence segmentation algorithm they used. We use the sentence-splitting functionality provided by

⁴as seen in the scripts they used: <https://github.com/WengLab-InformaticsResearch/PICOX/>

[blob/main/step_0_preprocess_bioc_ebm_nlp.ipynb](blob/main/step_0_preprocess_bioc_ebm_nlp.ipynb)

⁵https://github.com/BIDS-Xu-Lab/section_specific_annotation_of_PICO

control	of	bleeding	from	esophageal	varices
B-participant	I-participant	I-participant			
		B-outcome	I-outcome	I-outcome	I-outcome

Table 5: Two overlapping entities found in the train partition of the EBM-NLP (*hierarchical_labels*) dataset (document id: 2788816).

MEDSPACY⁶ to obtain the list of sentences needed to run the models. Consequently, sentence segmentation may differ from the one used in the original PICOX experiment.

3.3 Experimental Setup

For each dataset, we fine-tuned the models following the experimental setup of Zhang et al. (2024). In total, we conducted four runs using the following combination of training datasets: (i) EBM-NLP, (ii) EBM-NLP + PICO-corpus, (iii) EBM-NLP + AD and (iv) EBM-NLP + COVID-19. We used the same scripts, methodology, and hyperparameters as the authors. The hardware specifications used in the experiments are reported in Appendix A.

3.4 Sanity Check

As a sanity check, we reproduce the original experiment of Zhang et al. (2024) on the EBM-NLP dataset. Both the span boundary detector and span classification models are fine-tuned exclusively on the EBM-NLP training and validation partitions. For this sanity check, we use the *starting_spans* version of EBM-NLP, consistent with the original experimental setup. Table 6 shows the results we obtain on EBM-NLP using our reproduction of PICOX, alongside those reported for the original PICOX model.

We observe an average difference of 1.29 points in micro F1-score. The F1-scores for *intervention* and *outcome* are very similar, with differences of less than 1 point. In contrast, the *participant* label shows a larger gap, with a difference of 5.01 points in F1-score. The implementation of a different strategy for sentence segmentation, as well as potential differences in data partitioning, are likely contributors to these discrepancies in the reported metrics.

4 BioBERT_{overlaps} Development

In this section, we introduce BioBERT_{overlaps} for PICO entity extraction.⁷ This model was designed

to work with overlapping and nested entities. We hereafter use it for PICO entity extraction, but we designed it to be easily adaptable to any set of named entities.

4.1 Model Description

BioBERT_{overlaps} is based on BioBERT (Lee et al., 2020) and is trained using a binary cross-entropy with logits loss function. For multi-label token classification, the model produces a vector logits for each token. The loss for the n -th token and c -th label is defined as

$$\ell_{n,c} = -w_c [y_{n,c} \log \sigma(x_{n,c}) + (1 - y_{n,c}) \log (1 - \sigma(x_{n,c}))]$$

where $x_{n,c}$ denotes the logit produced by the model for the c -th label of token n , $y_{n,c} \in \{0, 1\}$ denotes the corresponding target, and w_c is the weight assigned to class c . The sigmoid function $\sigma(\cdot)$ is incorporated internally by BCEWithLogitsLoss.

The target labels are represented using multi-hot encoded vectors, where each position corresponds to a label. When a label is assigned to a token, the corresponding position of this label in the vector is set to 1 (and to 0 otherwise). This formulation allows the model to represent overlapping and nested entities.

Since we also want the model to identify the exact boundaries of each entity, we use the IOB annotation scheme. Each token is represented by a multi-hot vector of size $C = 6$, corresponding to the six possible IOB-PICO labels:

$$v = [B_P, B_{Int}, B_O, I_P, I_{Int}, I_O]$$

Here, B indicates the beginning of an entity and I indicates that a token is inside of an entity. P , Int and O denote the PICO entity types *participant*, *intervention* and *outcome*, respectively. To obtain the predictions made by the model, we apply the sigmoid function:

$$Sigmoid(x) = \sigma(x) = \frac{1}{1 + \exp(-x)}$$

Then, we apply a threshold $t \in (0, 1)$ to the resulting probabilities to obtain the predicted labels.

⁶<https://github.com/medspacy/medspacy>

⁷<https://github.com/JulienBez/BioBERToverlaps>

	PICOX_{original}			PICOX_{reproduction}		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Participant	50.00	66.87	57.22	44.07	64.02	52.21
Intervention	56.96	52.61	54.70	56.05	51.91	53.90
Outcome	47.50	38.84	42.74	47.04	38.89	42.58
<i>Micro</i>	51.86	48.79	50.28	49.89	48.12	48.99
<i>Macro</i>	51.49	52.77	51.55	49.05	51.61	49.55

Table 6: Comparison between the metrics obtained with our reproduction of the PICOX model on the EBM-NLP dataset and the metrics obtained with the original one, as reported in Zhang et al. (2024). Metrics were computed at entity-level in strict mode.

By default, we use a threshold of $t = 0.3$. However, we provide an analysis of the impact of different threshold values on metrics in Section 5 by varying the threshold used.

4.2 Data Processing

To ensure comparability across the approaches, we use the same datasets and splits as those used during our reproduction of the experiments by Zhang et al. (2024). The main difference is that, unlike in our reproduction of Zhang et al. (2024), we do not apply a sentence segmentation algorithm in this setting.

4.3 Experimental Setup

BioBERT_{overlaps} was fine-tuned using the same procedure, datasets, and hardware as for our reproduction of PICOX (resulting in four runs, as described in Section 3). We used OPTUNA for hyperparameter optimization. The following hyperparameters were optimized:

- learning rate;
- weight decay;
- train batch size per device; and
- validation batch size per device.

We followed the hyperparameter search protocol established by Villarmin et al. (2024). In total, 15 trials were conducted to determine the best combination of hyperparameters. For each trial, we computed the entity-level F1-score under the strict matching criterion and used it to select the best configuration. Appendix B reports the best set of hyperparameters for each model.

5 Results

5.1 Overall Performances

Table 7 shows the performance of the models on each dataset. We observe that PICOX_{reproduction} achieves the best *micro* and *macro* F1-scores for AD and COVID-19. On EBM-NLP, BIOBERT_{overlaps} outperforms PICOX_{reproduction} by 1.23 points in *micro* F1-score. For PICO-CORPUS, BIOBERT_{overlaps} achieves a *micro* F1-score of 70.11, exceeding PICOX_{reproduction} by more than 10 points. Figure 1 shows the variation in metric scores as a function of the threshold applied to BIOBERT_{overlaps} outputs. Across all models, the best F1-scores are obtained with thresholds between 0.4 and 0.6.

The discrepancies with the original PICOX experiments—namely, the change in EBM-NLP version, differences in dataset splits and the use of a different sentence segmentation algorithm—do not allow for a direct comparison with the results of Zhang et al. (2024). Nevertheless, on the EBM-NLP dataset alone, we observe that the *micro* F1-score decreases from 48.99 (Table 6) to 44.19 (Table 7) when using the *hierarchical_labels* version of EBM-NLP instead of the *starting_spans* version. By targeting exact entities rather than spans containing these entities, the task appears to be more challenging for the PICOX architecture. Furthermore, the results on AD and COVID-19, which correspond to span-based datasets, may indicate that PICOX is better suited to span extraction than to exact entity extraction.

5.2 Evaluation on Nested Entities

To evaluate their model’s performance on nested entities, Zhang et al. (2024) divided the sentences in their test partitions into two groups: (i) sentences with nested entities and (ii) sentences without nested entities. We argue that this approach

	BioBERT _{overlaps}			PICOX _{reproduction}		
	Precision	Recall	F1-score	Precision	Recall	F1-score
COVID-19						
Participant	60.33 (± 8.61)	67.50 (± 6.85)	63.67 (± 7.77)	54.29	79.17	64.41
Intervention	66.36 (± 4.09)	66.11 (± 4.87)	66.18 (± 3.84)	64.56	70.83	67.55
Outcome	48.66 (± 5.02)	36.61 (± 8.44)	41.60 (± 7.35)	56.25	45.76	50.47
Micro	59.80 (± 2.81)	55.10 (± 3.25)	57.34 (± 2.94)	59.88	62.58	61.19
Macro	58.45 (± 3.16)	56.74 (± 3.34)	57.15 (± 3.33)	58.36	65.25	60.80
AD						
Participant	26.87 (± 16.2)	28.42 (± 18.4)	27.58 (± 17.2)	18.92	36.84	25.00
Intervention	43.89 (± 5.00)	56.08 (± 1.82)	49.10 (± 2.98)	43.94	63.04	51.79
Outcome	46.37 (± 6.70)	48.61 (± 8.73)	47.36 (± 7.31)	52.94	61.63	56.96
Micro	42.83 (± 5.19)	48.32 (± 5.89)	45.41 (± 5.48)	43.09	58.69	49.69
Macro	39.04 (± 7.02)	44.37 (± 7.77)	41.34 (± 7.41)	38.60	53.83	44.58
EBM-NLP						
Participant	21.80 (± 0.97)	22.09 (± 1.42)	21.91 (± 0.75)	15.93	24.95	19.46
Intervention	58.91 (± 2.97)	54.36 (± 2.73)	56.47 (± 1.55)	53.56	59.19	56.24
Outcome	46.79 (± 1.39)	38.92 (± 1.65)	42.46 (± 0.86)	38.24	48.69	42.84
Micro	46.48 (± 0.71)	42.15 (± 1.59)	44.19 (± 0.80)	38.53	48.52	42.96
Macro	42.50 (± 0.89)	38.46 (± 1.45)	40.28 (± 0.72)	35.92	44.29	39.51
PICO-CORPUS						
Participant	69.55 (± 2.58)	72.99 (± 2.28)	71.17 (± 1.35)	51.45	75.11	61.08
Intervention	58.54 (± 1.28)	59.51 (± 5.43)	58.90 (± 2.60)	37.16	66.34	47.64
Outcome	70.67 (± 2.75)	73.41 (± 2.00)	71.95 (± 0.79)	64.41	60.83	62.57
Micro	68.77 (± 1.01)	71.54 (± 1.20)	70.11 (± 0.26)	55.37	65.16	59.87
Macro	66.25 (± 0.57)	68.64 (± 2.26)	67.35 (± 1.01)	51.01	67.43	57.08

Table 7: Performance on the PICO entity extraction task across the COVID-19, AD, EBM-NLP and PICO-CORPUS datasets. Metrics were computed at the entity-level in strict mode. For BioBERT_{overlaps}, standard deviations are computed over the five fine-tuned models. The best *micro*- and *macro*-F1 scores are highlighted in bold.

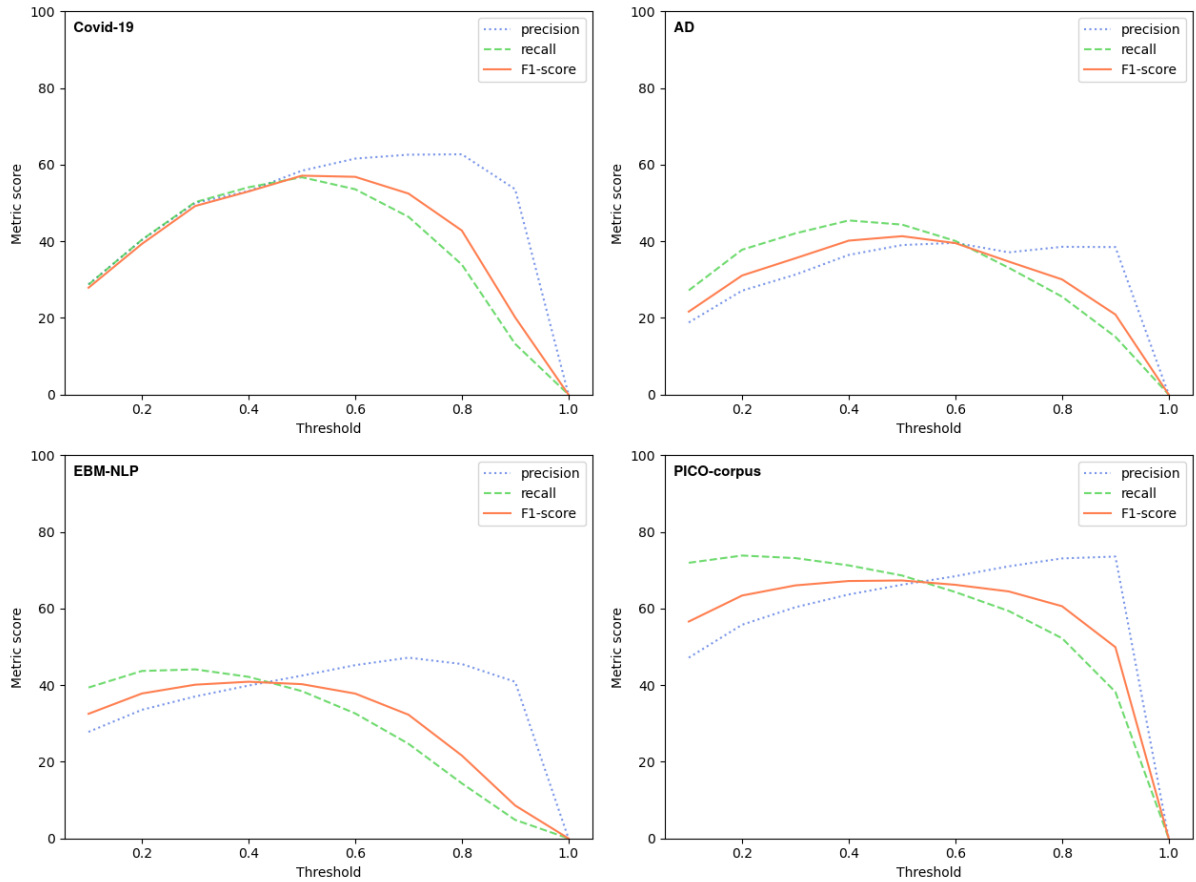


Figure 1: Impact of the threshold used on BioBERT_{overlaps} outputs on recall, precision and F-score (*macro*).

	Train	Valid	Test
Sentences	1,914	98	22
Nested entities	2,589	136	29
Non-nested entities	7,958	471	107

Table 8: Number of sentences containing at least one nested entity, along with the number of nested and non-nested entities in these sentences.

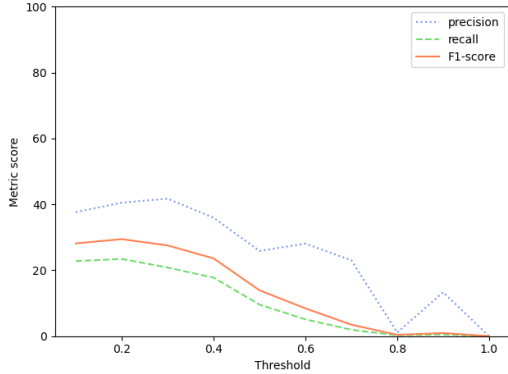


Figure 2: Impact of the threshold on nested and overlapping entities.

is flawed. Table 8 shows the number of sentences belonging to group (i) along with the numbers of nested and non-nested entities they contain. The number of non-nested entities is approximately three times higher than the number of nested entities in each partition. As a result, their reported performance on nested entities is heavily influenced by the model’s performance on non-nested entities.

To address this limitation, we evaluate the model’s performance on nested entities using two metrics: a token-level metric and an entity-level metric. Hereafter, P refers to the model’s predictions and R to the references (the true labels).

For the token-level metric, we go through each token one-by-one. If a token has two or more labels in either R or P , we keep it in both lists and we filter it out otherwise. This ensures that our token-level evaluation is carried *only* on overlapping tokens. Finally, we ignore IOB labels for this metric. For the entity-level metric, we begin by extracting entity spans from R and P . We then apply the same selection criterion to the resulting entity spans.

Table 9 shows the results obtained with both approaches for nested and overlapping entities. BIOBERT_{overlaps} outperforms PICOX_{reproduction} by 12.94 points in *micro* F1-score according to the token-level metric and by 13.20 points according to the entity-level metric. Figure 2

shows the variation in metric scores for nested and overlapping entities as a function of the threshold applied to BIOBERT_{overlaps} outputs. The best F1-score is achieved at a threshold of 0.2.

6 Conclusion

In this work, we show that PICO datasets created from RCTs can be broadly categorized as either span-based or entity-based. In span-based datasets, PICO annotations consist of fewer and longer spans of text containing relevant PICO information, whereas entity-based datasets focus on fine-grained and precise annotations of PICO entities in RCTs.

We then conducted a comparative study of two pipelines: PICOX_{reproduction} and BIOBERT_{overlaps}. Our findings show that PICOX_{reproduction} is better suited to extracting text spans containing PICO information, whereas BIOBERT_{overlaps} is more suited to extracting precise PICO entities. Moreover, we observe that BIOBERT_{overlaps} performs better on nested and overlapping PICO entities, reaching an entity-level *micro* F1-score of 30.70, compared with 17.58 for PICOX_{reproduction}. Future work will focus on applying these models to a large-scale RCT database. We plan to annotate a sample of these documents to determine which approach is better suited to the needs of medical practitioners.

Limitations

This study focuses exclusively on BERT-based models. It would be interesting to compare them with other approaches, such as traditional machine learning systems or generative LLMs. However, BERT-based models have shown to achieve strong performance on biomedical NLP tasks (Naguib et al., 2024), which motivates our focus on this model family.

Regarding the sanity check in Table 6, we used the same EBM-NLP version as Zhang et al. (2024). However, since the authors do not provide clear instructions regarding data segmentation or data split, we indeed couldn’t reproduce the same methodology, and therefore, results. This lack of reproducibility impacted our experiments in several ways. We offer a reproducible version of PICOX in this paper.

Additionally, we noticed several inconsistencies within the EBM-NLP dataset. Some entities are incorrectly annotated. While the test

	BioBERT _{overlaps}			PICOX _{reproduction}		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Token-level (IO)						
Participant	100.0 (+0.00)	58.06 (+19.4)	71.82 (+16.7)	78.57	35.48	48.89
Intervention	100.0 (+0.00)	65.07 (+8.35)	78.58 (+6.51)	97.72	64.18	77.48
Outcome	100.0 (+0.00)	48.18 (+4.93)	64.91 (+4.58)	61.90	29.54	40.00
<i>Micro</i>	100.0 (+0.00)	58.30 (+5.11)	73.56 (+4.06)	84.81	47.18	60.62
<i>Macro</i>	100.0 (+0.00)	57.11 (+6.83)	71.77 (+5.96)	79.40	43.07	55.46
Entity-level (strict)						
Participant	20.57 (+15.1)	10.91 (+9.95)	14.17 (+12.0)	00.00	00.00	00.00
Intervention	48.57 (+5.63)	27.41 (+2.03)	34.90 (+1.82)	46.67	25.92	33.33
Outcome	56.13 (+11.7)	24.21 (+7.06)	33.58 (+8.44)	09.09	05.26	06.67
<i>Micro</i>	45.66 (+6.41)	23.16 (+2.88)	30.70 (+3.83)	23.53	14.04	17.58
<i>Macro</i>	41.76 (+8.85)	20.84 (+4.10)	27.55 (+5.42)	18.59	10.40	13.33

Table 9: Performance on the PICO extraction task for nested and overlapping entities in the EBM-NLP dataset. Metrics are token-level IO and entity-level strict. The best *micro*- and *macro*-F1 scores are highlighted in bold.

partition of this dataset was manually annotated, its training partition was automatically annotated. Other studies have addressed this issue and proposed corrected versions of EBM-NLP, such as EBM-NLP_{COMET} (Abaho et al., 2022) or EBM-NLP_{rev} (Hu et al., 2023). However, these corrections cover only a fraction of the entries in EBM-NLP, resulting in smaller training datasets and potentially limiting their usefulness for model training.

Finally, the threshold used to obtain predictions from BIOBERT_{overlaps} outputs can be considered a limitation. Although we found that the best threshold across the studied datasets was between 0.5 and 0.6, this range should not be considered universal. The optimal threshold may differ for other datasets.

Acknowledgments

This work has been supported by CHIST-ERA grant CHIST-ERA-22-ORD-02 and by the Agence Nationale de la Recherche (ANR) under project numbers ANR-23-CHRO-0008-01 and ANR-22-CPJ1-0087-01.

References

- Micheal Abaho, Danushka Bollegala, Paula Williamson, and Susanna Dodd. 2022. *Assessment of contextualised representations in detecting outcome phrases in clinical trials*.
- Rohit Borah, Andrew W. Brown, Patrice L. Capers, and Kathryn A. Kaiser. 2017. *Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry*. *BMJ open*, 7(2):e012545.
- Luis H. Braga, Forough Farrokhyar, M. İrfan Dönmez, Caleb P. Nelson, Bernhard Haid, Kathy Herbst, Massimo Garriboli, Salvatore Cascio, Anka Nieuwhof-Leppink, Martin Kaefer, Darius J. Bāgli, Nicholas Kalfa, Christina Ching, Magdalena Fossum, and Luke Harper. 2025. *Randomized controlled trials – The what, when, how and why*. *Journal of Pediatric Urology*, 21(2):397–404.
- S. Chabou and M. Iglewski. 2015. *PICO extraction by combining the robustness of machine-learning methods with the rule-based methods*. In *2015 World Congress on Information Technology and Computer Applications (WCITCA)*, pages 1–4.
- Samir Chabou and Michal Iglewski. 2018. *Combination of conditional random field with a rule based method in the extraction of PICO elements*. *BMC Medical Informatics and Decision Making*, 18(1):128.
- Xiang Dai. 2018. *Recognizing Complex Entity Mentions: A Review and Future Directions*. In *Proceedings of ACL 2018, Student Research Workshop*, pages 37–44, Melbourne, Australia. Association for Computational Linguistics.
- Anjani Dhrangadhariya, Gustavo Aguilar, Thamar Solorio, Roger Hilfiker, and Henning Müller. 2021. *End-to-End Fine-Grained Neural Entity Recognition of Patients, Interventions, Outcomes*. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 65–77, Cham. Springer International Publishing.
- Madhusudan Ghosh, Shrimon Mukherjee, Asmit Ganguly, Partha Basuchowdhuri, Sudip Kumar Naskar, and Debasis Ganguly. 2024. *AlpaPICO: Extraction of PICO Frames from Clinical Trial Documents Using LLMs*. *Methods*, 226:78–88. ArXiv:2409.09704 [cs].
- Yan Hu, Vipina K Keloth, Kalpana Raja, Yong Chen, and Hua Xu. 2023. *Towards precise PICO extraction from abstracts of randomized controlled trials using a section-specific learning approach*. *Bioinformatics*, 39(9):btad542.
- Di Jin and Peter Szolovits. 2020. *Advancing PICO element detection in biomedical text via deep neural networks*. *Bioinformatics*, 36(12):3856–3862.

- David Kartchner, Jennifer Deng, Shubham Lohiya, Tejasri Kopparthi, Prasanth Bathala, Daniel Domingo-Fernández, and Cassie Mitchell. 2023. [A comprehensive evaluation of biomedical entity linking models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14462–14478, Singapore. Association for Computational Linguistics.
- 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.
- Shifeng Liu, Yifang Sun, Bing Li, Wei Wang, Florence T. Bourgeois, and Adam G. Dunn. 2021. [Sent2Span: Span Detection for PICO Extraction in the Biomedical Text without Span Annotations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1705–1715, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Faith Mutinda, Kongmeng Liew, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. 2022. [PICO Corpus: A Publicly Available Corpus to Support Automatic Data Extraction from Biomedical Literature](#). In *Proceedings of the first Workshop on Information Extraction from Scientific Publications*, pages 26–31, Online. Association for Computational Linguistics.
- Marco Naguib, Xavier Tannier, and Aurélie Névoul. 2024. [Few-shot clinical entity recognition in English, French and Spanish: masked language models outperform generative model prompting](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6829–6852, Miami, Florida, USA. Association for Computational Linguistics.
- Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. 2018. [A Corpus with Multi-Level Annotations of Patients, Interventions and Outcomes to Support Language Processing for Medical Literature](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Melbourne, Australia. Association for Computational Linguistics.
- Baptiste Pras and Nona Naderi. 2026. [Fine-grained mention-level analysis of biomedical entity linking models](#). *Studies in Health Technology and Informatics*, 336:999–1003.
- Tim Reason, Julia Langham, and Andy Gimblett. 2024. [Automated Mass Extraction of Over 680,000 PICOs from Clinical Study Abstracts Using Generative AI: A Proof-of-Concept Study](#). *Pharmaceutical Medicine*, 38(5):365–372.
- W. S. Richardson, M. C. Wilson, J. Nishikawa, and R. S. Hayward. 1995. The well-built clinical question: a key to evidence-based decisions. *ACP journal club*, 123(3):A12–13.
- Carlos Cuevas Villarmin, Sarah Cohen-Boulakia, and Nona Naderi. 2024. [Reproducibility in Named Entity Recognition: A Case Study Analysis](#). In *2024 IEEE 20th International Conference on e-Science (e-Science)*, Osaka, Japan. IEEE.
- Gongbo Zhang, Yiliang Zhou, Yan Hu, Hua Xu, Chunhua Weng, and Yifan Peng. 2024. [A span-based model for extracting overlapping PICO entities from randomized controlled trial publications](#). *Journal of the American Medical Informatics Association: JAMIA*, 31(5):1163–1171.

A Hardware Specifications

Table 10 shows the pieces of hardware we used to run our experiments. Both experiments were run on the same setup.

Processor:	2 x Intel Xeon Gold 6148 20 cores / 40 threads @ 2.4 GHz (Skylake)
GPU:	4 x NVIDIA Tesla V100 with 32 GiB of RAM (NVLink)
RAM:	384 GiB

Table 10: Hardware specifications of our experiments.

B BIOBERT_{overlaps} Hyperparameters

Table 11 shows the best hyperparameters for each fine-tuned BIOBERT_{overlaps} instance.

	Learning rate	Weight decay	Train size	Valid size
COVID-19	0.00004	0.00005	8	16
AD	0.00004	0.00004	8	8
EBM-NLP	0.00003	0.00010	16	32
PICO-CORPUS	0.00004	0.00347	8	16

Table 11: Best sets of hyperparameters for BIOBERT_{overlaps}.

C Additional Results

Table 12 shows the performance of the models on each dataset and Table 13 show the results obtained with both approaches on nested and overlapping entities. We did not perform a hyperparameter search for BIOBERT_{overlaps} on these runs and instead used the default hyperparameters.

	BioBERT _{overlaps}			PICOX _{reproduction}		
	Precision	Recall	F1-score	Precision	Recall	F1-score
COVID-19						
Participant	14.02 (± 8.82)	18.33 (± 10.4)	15.87 (± 9.61)	54.29	79.17	64.41
Intervention	54.61 (± 1.83)	57.49 (± 2.52)	56.01 (± 2.12)	64.56	70.83	67.55
Outcome	40.08 (± 3.26)	20.68 (± 2.21)	27.27 (± 2.65)	56.25	45.76	50.47
<i>Micro</i>	41.77 (± 3.42)	37.41 (± 2.62)	39.47 (± 2.97)	59.88	62.58	61.19
<i>Macro</i>	36.24 (± 3.20)	32.17 (± 4.01)	33.05 (± 3.66)	58.36	65.25	60.80
AD						
Participant	06.55 (± 4.01)	12.62 (± 6.00)	08.59 (± 4.83)	18.92	36.84	25.00
Intervention	36.88 (± 1.04)	60.42 (± 1.82)	45.80 (± 1.23)	43.94	63.04	51.79
Outcome	19.75 (± 4.05)	16.66 (± 4.05)	18.06 (± 4.05)	52.94	61.63	56.96
<i>Micro</i>	24.12 (± 2.37)	30.80 (± 2.98)	27.05 (± 2.63)	43.09	58.69	49.69
<i>Macro</i>	21.06 (± 2.46)	29.90 (± 3.36)	24.15 (± 2.84)	38.60	53.83	44.58
EBM-NLP						
Participant	18.97 (± 0.73)	27.87 (± 0.77)	22.57 (± 0.57)	15.93	24.95	19.46
Intervention	51.97 (± 1.61)	58.96 (± 1.68)	55.22 (± 0.89)	53.56	59.19	56.24
Outcome	38.48 (± 0.86)	44.26 (± 2.85)	41.14 (± 1.40)	38.24	48.69	42.84
<i>Micro</i>	39.17 (± 0.36)	47.27 (± 1.70)	42.83 (± 0.86)	38.53	48.52	42.96
<i>Macro</i>	36.48 (± 0.22)	43.70 (± 1.35)	39.64 (± 0.61)	35.92	44.29	39.51
PICO-CORPUS						
Participant	53.43 (± 2.97)	68.07 (± 1.25)	59.83 (± 2.04)	51.45	75.11	61.08
Intervention	39.30 (± 5.97)	61.37 (± 2.65)	47.65 (± 4.01)	37.16	66.34	47.64
Outcome	57.50 (± 1.78)	72.33 (± 0.71)	64.05 (± 1.20)	64.41	60.83	62.57
<i>Micro</i>	53.50 (± 1.65)	69.81 (± 0.36)	60.57 (± 1.01)	55.37	65.16	59.87
<i>Macro</i>	50.08 (± 1.81)	67.25 (± 0.96)	57.18 (± 1.17)	51.01	67.43	57.08

Table 12: Performance on the PICO entity extraction task across the COVID-19, AD, EBM-NLP and PICO-corpus datasets. Metrics are computed at the entity-level in strict mode. For BioBERT_{overlaps}, standard deviations are computed over five fine-tuned models. The best *micro*- and *macro*-F1 scores are highlighted in bold.

	BioBERT _{overlaps}			PICOX _{reproduction}		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Token-level (IO)						
Participant	57.98 (± 5.05)	82.93 (± 12.9)	68.05 (± 7.51)	78.57	35.48	48.89
Intervention	70.49 (± 3.81)	87.94 (± 2.08)	78.19 (± 2.13)	97.72	64.18	77.48
Outcome	76.95 (± 3.76)	93.02 (± 1.41)	84.16 (± 1.75)	61.90	29.54	40.00
<i>Micro</i>	71.72 (± 3.35)	90.08 (± 2.62)	79.77 (± 1.44)	84.81	47.18	60.62
<i>Macro</i>	68.47 (± 2.56)	87.96 (± 5.0)	76.80 (± 2.13)	79.40	43.07	55.46
Entity-level (strict)						
Participant	11.33 (± 10.4)	05.45 (± 4.97)	07.35 (± 6.72)	00.00	00.00	00.00
Intervention	51.39 (± 6.66)	29.63 (± 5.24)	37.45 (± 5.71)	46.67	25.92	33.33
Outcome	58.45 (± 14.2)	26.32 (± 6.45)	36.18 (± 8.55)	09.09	05.26	06.67
<i>Micro</i>	45.98 (± 6.78)	23.86 (± 3.64)	31.40 (± 4.68)	23.53	14.04	17.58
<i>Macro</i>	40.39 (± 7.95)	20.47 (± 3.39)	26.99 (± 4.74)	18.59	10.40	13.33

Table 13: Performance on the PICO extraction task for nested and overlapping entities in the EBM-NLP dataset. Metrics are token-level IO and entity-level strict. The best *micro*- and *macro*-F1 scores are highlighted in bold.