A Study on Agreement in PICO Span Annotations

Grace E. Lee and Aixin Sun
School of Computer Science and Engineering, Nanyang Technological University, Singapore
leee0020@e.ntu.edu.sg;axsun@ntu.edu.sg

ABSTRACT

In evidence-based medicine, relevance of medical literature is determined by predefined relevance conditions. The conditions are defined based on PICO elements, namely, **P**atient, **I**ntervention, **C**omparator, and **O**utcome. Hence, PICO annotations in medical literature are essential for automatic relevant document filtering. However, defining boundaries of text spans for PICO elements is not straightforward. In this paper, we study the agreement of PICO annotations made by multiple human annotators, including both experts and non-experts. Agreements are estimated by a standard span agreement (*i.e.,* matching both labels and boundaries of text spans), and two types of relaxed span agreement (*i.e.,* matching labels without guaranteeing matching boundaries of spans). Based on the analysis, we report two observations: (i) Boundaries of PICO span annotations by individual annotators are very diverse. (ii) Despite the disagreement in span boundaries, general areas of the span annotations are *broadly agreed* by annotators. Our results suggest that applying a standard agreement alone may undermine the agreement of PICO spans, and adopting both a standard and a relaxed agreements is more suitable for PICO span evaluation.

ACM Reference Format:
Grace E. Lee and Aixin Sun. 2019. A Study on Agreement in PICO Span Annotations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3331184.3331352

1 INTRODUCTION

In evidence-based medicine, it is crucial for medical professionals to effectively find relevant literature since medical decisions are made based on primary evidence. Relevance of literature depends on relevance conditions and the conditions are often defined using PICO framework: **P**atient (problem, population), **I**ntervention, **C**omparator, and **O**utcome. PICO elements identified in medical literature, therefore, are critical for effective retrieval.

Existing datasets used for automatic identification of PICO elements are sentence-level annotations [1, 3]. Recently, toward more accurate and detailed identification, a large-scale PICO *span* annotation dataset (EBM-PICO) is released [4]. EBM-PICO consists of 5,000 abstracts of medical literature with PICO spans annotated by medical experts and non-experts. Annotated spans range from a single word to a long phrase.

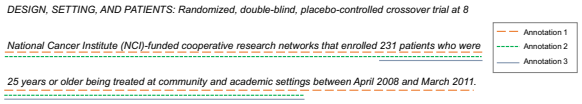


Figure 1: Example of span annotations in EBM-PICO dataset. This sentence (from PMID: 23549581) is annotated by 12 annotators (4 experts and 8 non-experts) and 3 different span annotations are made by them. The 3 spans have the same label P, but different boundaries.

PICO elements are presented in a *descriptive* manner in medical literature. Verbose descriptions of PICO elements make it nontrivial to decide their spans. Figure 1 shows an example sentence and its annotations in EBM-PICO dataset. The sentence contains information about P element. The three different spans are all annotated with label P. Even though each span has distinct boundaries, all indicate an acceptable information for P. Among the three, considering one span annotation correct and the rest incorrect may not reasonable.

In this paper, we study agreement in PICO span annotations made by different human annotators in EBM-PICO dataset. Specifically, we evaluate the annotation agreement using two types of measures: **exact span agreement** and **relaxed span agreement**. Exact span agreement is a standard evaluation approach for text span annotations. It evaluates both boundaries and labels of two spans. In relaxed span agreement, we analyze annotations in terms of two variants: *one-side boundary* (**OB**) agreement and *token overlap* (**TO**) agreement. OB and TO agreements evaluate whether annotations are with a same label but allowing the start and end boundaries between two spans to mismatch.

As the annotations in EBM-PICO dataset are made by medical experts and non-experts, we estimate the agreement within each of expert and non-expert groups, and across expert and non-expert groups. Our study shows an extremely low level of exact span agreement, but a very high level of relaxed span agreements in both within and across the groups. The large improvement in relaxed span agreement indicates the general area of annotations is mostly agreed despite unmatched boundaries. Our results suggest that applying exact agreement alone may underestimate the agreement of PICO span annotations. Therefore, adopting both exact and relaxed agreements is more suitable for PICO span evaluation such as in a PICO span recognition task. Lastly, we present how the two agreements are leveraged in a PICO span recognition task.

2 PICO SPAN ANNOTATION DATASET

We use EBM-PICO dataset [4] which provides PICO *span* annotations on 5,000 medical literature abstracts. Examples of annotations

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331352

are shown in Figure 1. We study this dataset because of its fine-grained annotations and the large scale. Other existing PICO annotation datasets are sentence-level annotations [1, 3] and/or contain only a few hundreds of documents [2]. Note that, in this dataset Intervention (I) and Comparator (C) are combined as a single element I, so that each document can have three types of annotations: **P**, **I**, and **O**. Each annotation (*i.e.,* a text span as in Figure 1) has one of the three labels.

EBM-PICO provides annotations by individual annotators (used in Section 3), and also aggregated annotations which combine individuals' annotations (used in Section 4). We focus on the individuals' annotations to study agreements of PICO span annotations. Individual annotators include two groups: Amazon Mechanical Turk (MTurk) workers as non-experts and medical experts. Specifically, all the 5,000 documents have annotations of PIO elements by MTurk workers. For each document, annotation process is conducted by at least three MTurk workers. Among the 5,000 documents, 200 documents have annotations by medical experts.[1] Each of the 200 documents has at least two experts' annotations. Hence, these 200 documents have annotations by both MTurk workers and medical experts.

3 PICO SPAN AGREEMENT

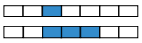
We measure agreement of PICO span annotations by computing $F1$. Given a pair of annotators, A and B, for a document, annotations made by A are first considered as *gold standard*, and annotations by B are considered as *predicted annotations*. We also switch A and B to calculate $F1$. Similar evaluation scheme was used in [5]. As a document has more than two annotators in EBM-PICO dataset, we average these values for all pairs of annotators, depending on the evaluation scenario (*e.g.,* within or cross non-expert and expert group evaluations). $F1$ is estimated by two kinds of agreement definitions.

Exact and Relaxed Span Agreements. In exact span agreement, two annotations agree with each other, if both have the same label and the same boundaries of text spans. By comparison, we also report two types of relaxed span agreement: one-side boundary (OB) agreement and token overlap (TO) agreement. Both OB agreement and TO agreement estimate the agreement in terms of whether two annotations are labeled as a same label, but allowing non-exact matching boundaries. For example, Table 1 shows 5 predicted annotations and their evaluations under the different span agreement definitions. The example sentence has 7 tokens and each cell represents a token. The cells in blue indicate a text span annotation. It is assumed that all examples in Table 1 have a same label.

In one-side boundary (**OB**) agreement, if either side of boundaries in predicted annotation is matched with a corresponding boundary of gold standard annotation, and there is at least one overlapped token between the two span annotations, then the predicted annotation is considered a correct prediction. In other words, a predicted span annotation can be larger or smaller than a gold standard span annotation, but the two spans share at least the same start or end boundaries. For instance, in Table 1, No. 1, 2, and 3 predicted annotations are correct annotations in OB agreement.

[1] We note that a few documents in the dataset have no annotations.

Table 1: Five examples of predicted annotations (Predicted) and their evaluations by different agreements (*i.e.,* Exact, OB, TO). The example sentence consists of 7 tokens, and tokens in blue denote annotated text span (the same label is assumed in all annotations). o/x indicates a correct/wrong prediction against gold annotation (Gold).

No.	Predicted	Gold	Exact	OB	TO
1			o	o	o
2			x	o	o
3			x	o	o
4			x	x	o
5			x	x	o

Token overlap (**TO**) agreement is defined with a more relaxed setup than one-side boundary (OB) agreement. Without considering boundaries of span annotations, if a predicted span annotation has at least one overlapped token with a gold standard span annotation, TO agreement counts the predicted annotation as a correct prediction. In Table 1, all five predicted annotations are correct under the TO agreement because of overlapped tokens.

Note that, the original EBM-PICO dataset paper reports token-level annotation agreement within expert annotators [4]. In their evaluation, the agreement is estimated on individual tokens (token-level) and it does not consider boundaries of annotations. In our evaluation, both exact span agreement and the two relaxed span agreements are at span-level.[2]

Within Group Agreement. There are 2 groups of annotators (*i.e.,* MTurk workers and medical experts). In this section, we study pairwise annotation agreement *within each annotator group.*

As discussed in Section 2, there are 5000 documents annotated by MTurk workers and among them 200 documents are additionally annotated by experts. We report agreement within MTurk workers on the 5000 documents (MTurk-5000), and within medical experts on the 200 documents (Expert-200). Since these 200 documents also contain MTurk annotations, we also compute the agreement within MTurk workers on the 200 documents (MTurk-200), for a direct comparison with Expert-200.

Observation 1. *The overall exact span agreement, within MTurk worker group and also within medical expert group, is very low.*

Table 2 reports averaged $F1$ values (standard deviations) of pairwise agreement within MTurk workers, and within experts, by different agreement evaluations. The overall exact span agreement within each group is very low. The exact span agreement among MTurk workers is extremely low. For labels I and O, $F1$ values are even lower than 0.1. The similar low agreement is also observed among the expert annotators. The agreement for label I is slightly higher than 0.5 but for labels P and O, $F1$ values are lower than 0.4. These values show that more than a half of annotations are not agreed with other annotators, even among domain experts. We believe that the low exact span agreement is caused by the high verbosity of PICO elements. According to the exact agreement, none of the annotation pairs in Figure 1 are agreed.

[2] Token-level agreement has more evaluation instances than span-level agreement since a span consists of several tokens.

Table 2: Average (standard deviation) $F1$ scores estimated within MTurk workers on 5000 documents (MTurk-5000). For the 200 documents containing annotations by both MTurk workers and medical experts, the within group agreement of MTurk workers (MTurk-200) and medical experts (Expert-200) are reported. The $F1$ scores are computed based on different agreements.

	Exact span agreement			Relaxed One-side Boundary (OB)			Relaxed Token Overlap (TO)		
	MTurk-5000	MTurk-200	Expert-200	MTurk-5000	MTurk-200	Expert-200	MTurk-5000	MTurk-200	Expert-200
P	0.187 (0.136)	0.202 (0.125)	0.395 (0.201)	0.361 (0.179)	0.385 (0.157)	0.680 (0.200)	0.421 (0.190)	0.441 (0.161)	0.737 (0.189)
I	0.093 (0.085)	0.137 (0.095)	0.576 (0.301)	0.187 (0.112)	0.235 (0.120)	0.732 (0.271)	0.241 (0.123)	0.282 (0.128)	0.758 (0.269)
O	0.064 (0.053)	0.078 (0.054)	0.357 (0.167)	0.139 (0.089)	0.175 (0.093)	0.654 (0.178)	0.215 (0.121)	0.256 (0.115)	0.713 (0.169)

Table 3: Precision, Recall, and $F1$ values of cross-group annotation agreement between MTurk workers (predicted annotation) and medical experts (gold standard). The average values (standard deviation in parenthesis) of 200 documents are reported.

	Exact span agreement			Relaxed One-side Boundary (OB)			Relaxed Token Overlap (TO)		
	Pre	Rec	$F1$	Pre	Rec	$F1$	Pre	Rec	$F1$
P	0.266 (0.126)	0.338 (0.144)	0.275 (0.125)	0.483 (0.144)	0.624 (0.172)	0.496 (0.144)	0.537 (0.146)	0.707 (0.178)	0.553 (0.143)
I	0.243 (0.131)	0.332 (0.186)	0.257 (0.141)	0.360 (0.149)	0.570 (0.239)	0.391 (0.159)	0.408 (0.154)	0.769 (0.364)	0.457 (0.173)
O	0.147 (0.078)	0.220 (0.108)	0.159 (0.082)	0.295 (0.113)	0.450 (0.141)	0.316 (0.112)	0.364 (0.123)	0.789 (0.410)	0.412 (0.132)

Observation 2. *Annotators agree with the general areas where PIO elements appear, even though they made different choices in annotation boundaries.*

On relaxed span agreement, both MTurk workers and medical experts show largely increased $F1$ than that of exact span agreement. Specifically, on one-side boundary (OB) agreement, for MTurk worker group, $F1$ values are greater than twice of the $F1$ values estimated on exact span agreement. On token overlap (TO) agreement, for expert group, $F1$ values for all PIO elements are greater than 0.7, which is a clear indication of high level of agreement. The improved agreement in relaxed span agreement demonstrates annotators made annotations at similar areas but not necessarily with same boundaries. For instance, in Figure 1 some pairs of annotations are agreed depending on the OB or TO agreements, and it is different from the zero pair agreed on the exact span agreement.

To summarize, due to the verbosity of PICO elements in medical literature, annotations made by annotators have diverse boundaries. As the exact span agreement requires matching boundaries as well as labels, the low level of agreement is estimated for PICO span annotations. However, the relaxed span agreements reflect the characteristics of PICO elements and show greater agreement by allowing unmatched boundaries.

Cross Group Agreement. The finding, the low agreement in exact span agreement and the high agreement in relaxed span agreement, is observed between annotators who share similar understanding about domain knowledge (*i.e.,* MTurk-MTurk or Expert-Expert pairs). In this section, we study the annotation agreement when annotators have different levels of domain knowledge, by measuring annotation agreement between a MTurk worker and a medical expert (*i.e.,* MTurk-Expert). Furthermore, we study differences in annotations between by MTurk workers and by medical experts in terms of lengths of span annotations.

On the 200 documents having annotations by both MTurk workers and medical experts, we measure the cross-group agreement. Specifically, we consider annotations by medical experts as gold standard annotations, and then estimate Precision, Recall, and $F1$ values for annotations by MTurk workers as predicted annotations. As each document has annotations from multiple MTurk and expert annotators, values derived by all possible MTurk-Expert pairs are averaged for a document.

Table 3 presents Precision/Recall/$F1$ values of cross-group agreement averaged on the 200 documents, with the three agreement types. In Table 3, the cross-group agreements between MTurk workers and medical experts present the similar trend shown in within-group agreement (Table 2). The exact agreement is low and the relaxed agreements are much higher. Based on the results of cross-group agreement as well as within group agreement, we make the third observation as follows.

Observation 3. *Our finding, the high agreement on the general areas of PICO annotations with unmatched boundaries, is consistent regardless of domain knowledge that annotators have.*

Besides, an interesting finding is that when the agreement changes from exact to relaxed, Recall increases to greater than 0.7 in the TO agreement while the improvement in precision is relatively small. This shows that the annotations by MTurk workers are a 'superset' of the annotations by medical experts in relaxed span agreements (high recall and low precision). Another finding is that on each agreement definition, the agreement in MTurk-Expert pair is always higher than the agreement in MTurk-MTurk pair, and lower than the agreement in Expert-Expert pair, comparing $F1$ values reported in both Tables 2 and 3. It shows that some MTurk workers are capable of annotating PICO elements similar to medical experts.

Next, we study differences in annotations made by the two groups in terms of the number of tokens in each span annotation. Table 4 shows the average number of tokens and its standard deviation in each label of spans for the 5000 documents by MTurk workers, and for the 200 documents by MTurk workers and by medical experts. Observe that the number of tokens in annotations

Tokens	Efficacy	and	safety	of	selamectin	against	fleas	and	heartworms	in	dogs	and	cats	presented	as	veterinary	patients	in	North	America	.
Gold standard	O	N	O	N	I	N	N	N	N	N	P	P	P	P	P	P	P	P	P	P	N
Predicted label	O	O	O	N	I	N	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P

Figure 2: Examples of correct PICO spans and predicted PICO spans in a PICO span recognition task. The example sentence is from a medical document PMID: 10940525 in EBM-PICO dataset. N indicates a token without PICO label.

Table 4: Comparison of the length of span annotations. The average (standard deviation in parenthesis) number of tokens are presented within MTurk workers on 5000 documents (MTurk-5000), and on 200 documents (MTurk-200) and medical experts on 200 documents (Expert-200).

Label	MTurk-5000	MTurk-200	Expert-200
P	9.355 (10.406)	8.268 (9.113)	6.356 (6.500)
I	4.356 (7.999)	3.322 (5.743)	1.903 (2.075)
O	6.792 (16.257)	6.134 (14.822)	4.379 (5.347)

by medical experts is smaller than the number of tokens in annotations by MTurk workers. We believe the difference in the length of spans is attributed to medical domain knowledge. Medical experts make more specific and short annotation spans for being able to identify essential information. Moreover, the annotations by MTurk workers show higher standard deviations than by experts.

4 PICO SPAN RECOGNITION

Our evaluation shows that the exact span agreement is low and the relaxed span agreements are much higher, among human annotators, regardless within experts, within non-experts, or cross-group. In this section, we demonstrate how the exact and relaxed span agreements can be used for the evaluation of PICO span recognition task. We also show how differently the exact and relaxed span agreements present the quality of performance.

The original EBM-PICO dataset paper also conducted a PICO recognition task. The performance is reported by token-level evaluation [4]. That is, boundaries are not evaluated since they are predetermined as a token separation, and also the evaluation instances in token-level evaluation are more than that of span evaluation. In this work, we conduct the same task as [4], but evaluate its performance on *span* evaluation, specifically, the exact, OB, and TO span agreements.

For PICO span recognition, Bi-directional LSTM-CRF (BiLSTM-CRF) model is used. We follow the same experiment settings and train/validation/test data splits of the aggregated annotations (see section 2) used in [4].[3] Table 5 presents performance evaluated by the exact span agreement, and OB and TO relaxed span agreements. Low performance on the exact span agreement is observed, as it is challenging even for human annotators. Performance evaluated by the two relaxed span agreements shows significant improvement. The values are even comparable to the results reported in Tables 2 and 3.

[3]From data exploration, it is found that about 10 percent of tokens have multiple labels in the aggregated annotations. Before training a model, we resolve the multiple labels of tokens into a single label with the priority order of I, P, and O. We believe I element is the most important element among the three elements, followed by P and O.

Table 5: Performance of BiLSTM-CRF in PICO span recognition evaluated by the exact and relaxed agreements. Precision/Recall/*F1* values are estimated for each element label (P, I, O) and Micro-averaged value (Micro) for all labels.

Eval	Exact span			One-side Boundary			Token Overlap		
Label	Pre	Rec	*F*1	Pre	Rec	*F*1	Pre	Rec	*F*1
P	0.227	0.205	0.216	0.766	0.692	0.727	0.840	0.758	0.797
I	0.465	0.283	0.352	0.792	0.481	0.599	0.835	0.508	0.632
O	0.406	0.276	0.329	0.790	0.538	0.640	0.838	0.571	0.679
Micro	0.387	0.267	0.316	0.785	0.541	0.640	0.837	0.577	0.683

Figure 2 shows examples of correct spans and predicted spans in the recognition task. There are three predicted spans (*i.e.,* O, I and P in the order). Indeed, the model correctly makes predictions on the general areas of correct spans. However, boundaries of the predicted spans are incorrect except I element prediction. Hence, by the exact span agreement, the only I predicted span is a correct prediction and the other two are incorrect predictions. On the other hand, in the OB and TO agreements, 2 and 3 out of the predicted spans are considered correct predictions, respectively.

5 CONCLUSION

We report observations made from agreements in PICO span annotations. The exact span agreement presents very low-level of agreement but the two relaxed span agreements show high-level of agreements in human annotations. The result shows that even though boundaries of PICO annotations are unmatched, the annotations are in the similar areas in general. Based on our observations, we argue that the evaluation of PICO span-related tasks shall consider not only the exact span agreement but also the relaxed span agreements, because even human annotators do not agree on exact spans due to the high verbosity of PICO elements.

ACKNOWLEDGMENTS

This work was supported by Data Science & Artificial Intelligence Research Centre, NTU Singapore.

REFERENCES

[1] Grace Y Chung. 2009. Sentence retrieval for abstracts of randomized controlled trials. *BMC medical informatics and decision making* (2009).
[2] Dina Demner-Fushman and Jimmy Lin. 2007. Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics* (2007).
[3] Su Nam Kim, David Martinez, Lawrence Cavedon, and Lars Yencken. 2011. Automatic classification of sentences to support evidence based medicine. In *BMC bioinformatics*.
[4] 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 *ACL*.
[5] Jie Yang, Yue Zhang, Linwei Li, and Xingxuan Li. 2018. YEDDA: A Lightweight Collaborative Text Span Annotation Tool. In *ACL, System Demonstrations*.