DISTANT-CTO: A Zero Cost, Distantly Supervised Approach to Improve Low-Resource Entity Extraction Using Clinical Trials Literature

Anjani Dhrangadhariya and **Henning Müller**
University of Geneva (UNIGE), Geneva, Switzerland
University of Applied Sciences Western Switzerland (HES-SO), Sierre, Switzerland
{anjani.dhrangadhariya,henning.mueller}@hevs.ch

Abstract

PICO recognition is an information extraction task for identifying participant, intervention, comparator, and outcome information from clinical literature. Manually identifying PICO information is the most time-consuming step for conducting systematic reviews (SR), which is already labor-intensive. A lack of diversified and large, annotated corpora restricts innovation and adoption of automated PICO recognition systems. The largest-available PICO entity/span corpus is manually annotated which is too expensive for a majority of the scientific community. To break through the bottleneck, we propose DISTANT-CTO, a novel distantly supervised PICO entity extraction approach using the clinical trials literature, to generate a massive weakly-labeled dataset with more than a million "Intervention" and "Comparator" entity annotations. We train distant NER (named-entity recognition) models using this weakly-labeled dataset and demonstrate that it outperforms even the sophisticated models trained on the manually annotated dataset with a 2% F1 improvement over the Intervention entity of the PICO benchmark and more than 5% improvement when combined with the manually annotated dataset. We investigate the generalizability of our approach and gain an impressive F1 score on another domain-specific PICO benchmark. The approach is not only zero-cost but is also scalable for a constant stream of PICO entity annotations.

1 Introduction

Primary care physicians rely on systematic reviews (SRs) for informed decision-making. SRs are conducted to objectively answer clinical questions and require going through a rigorous process of manually screening tens of thousands of clinical studies to identify terms describing PICO. PICO information identification is crucial to appraise the relevance of a clinical study for answering the clinical question at hand. A study is only included for writing SRs if it mentions relevant PICO information. Manual PICO information screening for a single SR consumes more than 12 months of two medical experts' time. The process can be automated using information extraction (IE) by directly pointing the human reviewers to the correct PICO descriptions. Automation will accelerate the overall process of writing SRs while reducing the burden on health professionals who are required to manually screen for PICO entities.

Automating PICO entity detection has garnered lower interest than other biomedical NER tasks because of the lack of publicly available entity annotated corpora. The largest publicly-available PICO entity/span dataset (EBM-PICO) contains only 5000 annotated abstracts, some of which were annotated through crowd-sourcing and others by hired medical experts (Nye et al., 2018). Crowd-sourcing involves hiring non-expert workers that require intensive training that is not commonly affordable. Hiring medical experts for annotation is equally often too expensive. IN GENERAL, extracting PICO entities/spans is somewhat tricky because of high disagreement between human annotators on the exact spans constituting the mentions. This leads to human errors in hand-labeled corpora. Hand-labeled datasets are static and prohibit quick manual re-labeling in case of human errors or when a downstream task requires new entities. For example, PICO entities extend to PICOS, where S denotes the "study type" of included evidence.

Distant supervision (DS) is a data-centric approach that allows generating massive weakly annotated datasets without human annotators and has previously been used to create large relation extraction corpora for the general and biomedical domains. To address the challenges above and democratize PICO entity recognition, we propose DISTANT-CTO, a distantly supervised and scalable approach to obtaining clinical trials annotations. We take an integrative approach combining methods of semi-supervised learning (SSL)

Proceedings of the BioNLP 2022 workshop, Dublin, Ireland, pages 345–358
May 26, 2022. ©2022 Association for Computational Linguistics

and gestalt pattern matching (GPM) to develop a continuously extensible dataset. We successfully demonstrate this approach for the "Intervention" and "Comparator" entity annotations as proof of concept (POC).

We summarize our contributions as follows:

- We develop a zero-cost, data-centric approach using DS to obtain "Intervention" and "Comparator" entity annotations.

- We develop and make publicly available a large weakly-labeled dataset from more than 300,000 clinical trials. The dataset offers about a million sentences with more than 977,682 annotations across 11 semantic types.

- We improve the state-of-the-art by 2% macro-F1 on the previously most poor-performing "Intervention" entity extraction on the EBM-PICO benchmark corpus without using costly manually labeled data and by 5% when combined with manually labeled data.

2 Related Work

A decade of automatic PICO information extraction was limited to sentence-level due to the unavailability of entity-annotated corpora (Boudin et al., 2010; Huang et al., 2011, 2013; Wallace et al., 2016; Jin and Szolovits, 2018). The release of the EBM-PICO corpus paved the way for the community to improve upon the PICO entity/span extraction task. (Nye et al., 2018). The corpus is biased towards pharma intervention classes overshadowing non-pharma ones leading to a substandard performance on it in the previous SOTA fully-supervised PICO entity/span recognition models (Beltagy et al., 2019; Brockmeier et al., 2019; Zhang et al., 2020) and weakly supervised model (Liu et al., 2021). Small-scale annotation projects cannot capture the range and variation of the PICO descriptions spanning the entirety of clinical trials literature. At some point, applications of such static corpora will confront the problem of insufficient and irrelevant annotations. Manual annotation projects are neither affordable nor scalable for every lab, limiting innovation.

A plethora of DS methods have been previously explored for large-scale relation extraction but not for (named) entity extraction (Etzioni et al., 2008; Smirnova and Cudré-Mauroux, 2018; Adelani et al., 2020). Entity extraction in high-impact clinical and biomedical domains largely relies on small expert annotated datasets. Commonly, obtaining weak annotations using DS rely on aligning terms (a word or phrase) from ontologies onto the unstructured text (Giannakopoulos et al., 2017; Yang et al., 2018; Peng et al., 2019; Hedderich et al., 2021). Ontologies are structured, standardized data sources that do not capture various writing variations from clinical literature. Weak annotations obtained using custom-built rules like regular expressions are restricted by either task or worse even by entity type (Ratner et al., 2017; Safranchik et al., 2020; Fries et al., 2021). Bootstrapping approaches like label propagation (LP) still require an expert annotated dataset to obtain pseudo annotations for previously unlabeled data samples (Bing et al., 2017). It is hence not zero-cost.

Our work focuses on overcoming the discussed bottlenecks using a data-centric DS approach to generate a large clinical entity annotated corpus and train a downstream NER model to assess if it yields adequate results. Unlike the reviewed DS approaches, our approach does not use ontologies or rules or LP but rather uses GPM for flexibly aligning structured text in a clinical trials database to the free-text fields in the same database using an adaptable internal scoring scheme.

3 Data

ClinicalTrials.gov (CTO hereafter) documents more than 350,000 human clinical studies conducted around the globe. The trial's principal investigator enters and updates information about each study stored in CTO. It includes the title and description of the clinical trial, participant's eligibility criteria, participant disease and demographics, interventions evaluated, outcomes, *etc.* CTO allows programmatic access to this vast amount of information in the JSON (JavaScript Object Notation) format. The information is stored as a combination of structured tabular and unstructured free-text (see Figure 1). The 'OfficialTitle' and 'BriefTitle' tags in the JSON respectively store the official and shorter version of the study title in an unstructured free-text format. The 'BriefSummary' and 'DetailedDescription' tags store study summaries. Interventions used in the study are stored under the 'InterventionName' tag and their synonyms under 'InterventionOtherName' tag each of which could be linked to their broad semantic type (drug, device, behavioral, procedural, biological, dietary supplement, diagnostic test, radiation, genetic, combination product, other) mentioned under the 'InterventionType' tag. As each intervention name is linked to its semantic type, this becomes a structured information store. The 'InterventionDescription' tag describes intervention administration procedures often in a detailed passage.

4 Approach

The approach is schematically illustrated in Figure 2 and is described below.

4.1 Distant Supervision

Distantly supervised (DS) information extraction (IE) is an efficient SSL method (Etzioni et al., 2008; Wen et al., 2019). It is used when the task at hand has 1) some strongly-labeled data, 2) abundant unlabeled data, and 3) a weak-labeling function that could sample from this unlabeled data and label them using a heuristic function. This labeling function is a heuristic algorithm that uses a heuristic to label the unlabeled data (Pinto et al., 2003; Greaves, 2014). It results in a weakly-labeled dataset with potential label noise. DS-IE models can then collectively use this strongly-labeled and weakly-labeled training data to give the final output.

4.2 Gestalt Pattern Matching

In entity extraction, the most common form of DS is to heuristically align terms from a structured information source onto the unstructured text (Wen et al., 2019). When flexible, this heuristic boils down to a substring matching problem. The weak-labeling function matches the longest common substring (LCS) between the structured term and unstructured text. Gestalt Pattern Matching (GPM), also known as Ratcliff/Obershelp similarity algorithm, is a string-matching algorithm for determining the similarity of two strings. The similarity between two strings S_1 and S_2 is measured by the formula, calculating twice the number of matching characters K_m divided by the total length $|S_1| + |S_2|$ of both strings. Matching characters are identified by the LCS algorithm followed by recursively finding matching characters in the non-matching regions on either side from both strings (Ratcliff and Metzener, 1988). $Similarity$ ranges between 0, which means no match, and 1, which means a complete match of the two strings.

$$Similarity(S) = \frac{2K_m}{|S_1| + |S_2|} \; ; \; 0 \leq S \leq 1 \quad (1)$$

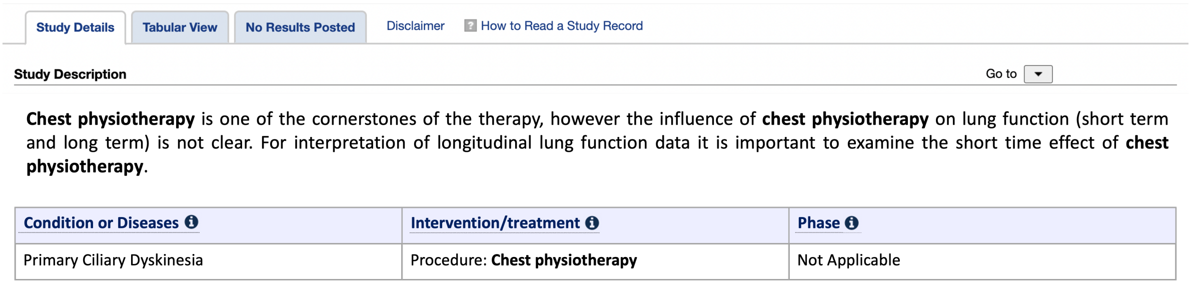
Difflib: It is a python module providing a `sequencematcher` function that extends the GPM algorithm for comparing pairs of strings. `sequencematcher` finds the longest contiguous subsequence between the sequence pair without the "junk" elements such as blank lines or white spaces. The same idea is then applied recursively to the flanks of the sequences to the left and the right of the matching subsequence. This yields matching sequences that appear normal to the human eye.

4.3 Candidate Generation

We define candidate generation as the process of automatically generating entity-annotated sentences.

Assumption and Problem formulation: As "Intervention" and "Comparator" entities represent interventions in two different roles in clinical trials and semantically the same classes, they are clubbed into a single "Intervention" entity class. Let each CTO record JSON file be $r_i \in \boldsymbol{R}, i = \{1, 2, ..., I\}$. Let the intervention terms in 'InterventionName' tags and 'InterventionOtherName' tags be the intervention source $S = \{s_1, s_2, ..., s_m\}$ used in the study r_i. Each intervention term $s_i \in S$ is linked to intervention class from 'InterventionType' tag converting it into a tuple of $\langle s_{class}, s_{name} \rangle$, s_{name} = intervention term and s_{class} = intervention category. s_{name} is a sequence of words $\{y_1, y_2, ..., y_n\}, n = \{1, 2, ..., N\}$. Let each sentence $t_i = \{x_1, x_2, ..., x_m\}, m = \{1, 2, ..., M\}$ in the 'BriefSummary', 'DetailedDescription', 'BriefTitle', 'OfficialTitle' and 'InterventionDescription' be a part of the intervention target set T. We assume that for each s_{name} in r_i there could exist a mapping to t_i meaning s_{name} is possibly either completely or partially mentioned in the t_i (see Figure 1). Our goal is to build a scalable and adaptable candidate generation pipeline that maps each s_{name} from the structured intervention source S to the target sentences $t_i \in T$ (if a loose mapping exists). In this prototypical work, we focus on *almost* direct matches between the s_{name} and t_i and keep the order-free matches for future work.

Approach For each individual CTO record r_i, we extract all $s_{name} \in S$ and $t_i \in T$ from the locally stored CTO dump. Both S and T are preprocessed by lower-casing, replacing hyphens and multiple trailing spaces with a single space and removal of Unicode characters. Given a s_{name} and t_i, our aim is to identify and score (if identified) the mapping between both



Chest physiotherapy is one of the cornerstones of the therapy, however the influence of **chest physiotherapy** on lung function (short term and long term) is not clear. For interpretation of longitudinal lung function data it is important to examine the short time effect of **chest physiotherapy**.

Condition or Diseases	Intervention/treatment	Phase
Primary Ciliary Dyskinesia	Procedure: **Chest physiotherapy**	Not Applicable

Figure 1: An example CTO record (ID - NCT01929356) to demonstrate the information storage format which is a combination of structured table and unstructured text.

sequences. To map and score alignment from the s_{name} to t_i, we use a distant supervision labeling function LF_{ds} which is a combination of the `sequencematcher` function and an internal scoring function to fetch almost direct annotations. The `sequencematcher` function takes as input s_{name} and t_i and outputs several matching blocks $d_{block} \in D_{blocks}$ between both strings. These matching blocks between the two strings are calculated using a modified gestalt pattern matching algorithm as elaborated in 4.2. Each $d_{block} = \langle MatchPos_t, MatchPos_s, MatchLen \rangle$. $MatchPos_t$ is the start of the match in t_i, $MatchPos_s$ is the start of the match in s_{name} and $MatchLen$ is number of characters matching between the both. `sequencematcher` provides an internal scoring function called as `ratio` that returns a similarity score between the two sequences being matched. We do not use `ratio` because it returns an overall matching score between the two full sequences s_{name} and t_i rather than a match score for s_{name} and d_{block}. Instead, to identify the matching blocks that correspond to an exact match between an entire s_{name} and a part of t_i, we calculate a match score d_s for each matching block output by `sequencematcher` using equation 2 which is dividing the number of matching characters in the match block d_{block} by number of characters in s_{name}.

$$d_s = \frac{MatchLen}{|s_{name}|} \; ; \; 0 \leq d_s \leq 1 \qquad (2)$$

Any d_{block} with the d_s score of 1.0 is considered as complete match and then the s_{name} corresponding to the d_{block} is mapped onto sentence t_i to generate a positive annotation sentence $a_+ \in A_+$. Using the d_{block} with only the match score 1.0 leads to missing out on several entities leading to an incomplete noisy weakly annotated dataset. Taking this into consideration, we retrieve the d_{block} matching with d_s score of 0.9 as fairly-accurate partial matches. We used a validation set to relax the choice of similarity match score d_s to 0.9. We relax the labeling function LF_{ds} to match bigrams in source terms to the targets. In the real-world data, not all sentences in clinical trial literature mention the intervention name and therefore in addition to the positive annotation sentences we require negative annotation sentences. We take t_i and s_{name} where no parts of d_{block} scored d_s more than 0.2 to generate the negative annotation sentences $a_- \in A_-$. We call all these sequences comprised of the positive and the negative entity annotated sentences A_{+-} our weakly annotated dataset. Next, for all A_{+-} instances we fetch part-of-the-speech (POS) tags using `POS-tagger` from NLTK (Natural Language Toolkit) resulting into A_{+-POS}. We call the resulting dataset DISTANT-CTO set. POS tags are added as additional features as they have shown to help model generalization (Augenstein et al., 2017). `difflib` in combination with the internal scoring function are previously unexplored for automatic entity annotation generation. It has to be noted that the method depends on availability of short source texts with the possibility that they will be mentioned in longer target texts.

4.4 Model Training

We train an end-to-end distant NER model on A_{+-POS} using the architecture explained below.

1. Feature Extraction: To capture the domain-specific information, we used SciBERT, which was continually pretrained and domain adapted on the scientific literature from semantic scholar (Gururangan et al., 2020). The models used SciBERT to tokenize the text input A_{+-} into encoded tokens x_t and extract dense, contextual vectors e_t from x_t at each time-step t (Beltagy et al., 2019). POS-inputs A_{+-POS} were one-hot encoded into p_t vectors.

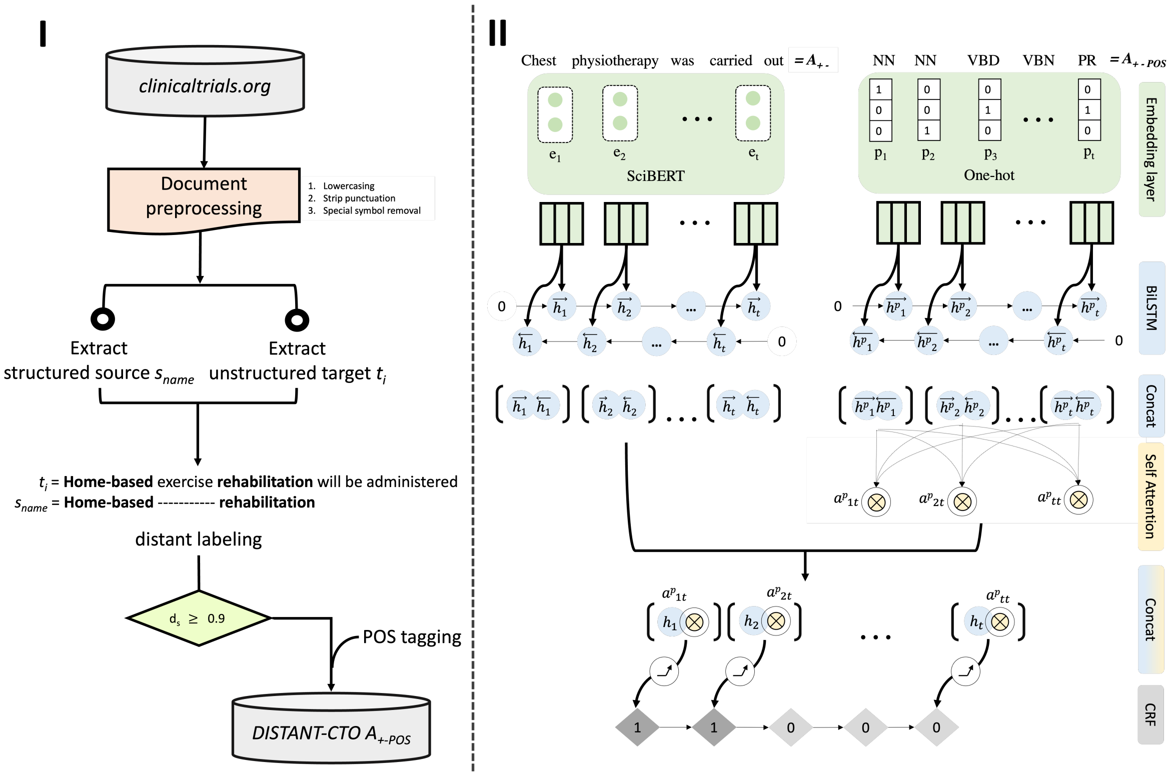

I
clinicaltrials.org
Document preprocessing
1. Lowercasing
2. Strip punctuation
3. Special symbol removal
Extract structured source s_{name}
Extract unstructured target t_i
t_i = **Home-based** exercise **rehabilitation** will be administered
s_{name} = **Home-based** ----------- **rehabilitation**
distant labeling
$d_s \geq 0.9$
POS tagging
DISTANT-CTO A_{+-POS}
II
Chest physiotherapy was carried out $=A_{+-}$
NN NN VBD VBN PR $=A_{+-POS}$
e_1 e_2 e_t
SciBERT
p_1 p_2 p_3 p_t
One-hot
Embedding layer
BiLSTM
Concat
Self Attention
a^p_{1t} a^p_{2t} a^p_{tt}
Concat
CRF
1 1 0 0 0

Figure 2: DISTANT-CTO approach - I) Distantly-supervised candidate generation approach, and II) Distantly-supervised NER model architecture.

2. Feature transformation: To further fine-tune to the training corpus, the model stacked a bidirectional LSTM (BiLSTM) on top of the SciBERT (Hochreiter and Schmidhuber, 1997). A BiLSTM layer encodes the text into a $(\overrightarrow{h})$ and $(\overleftarrow{h})$ vector using the current token embedding input e_t and the previous hidden state h_{t-1} in both the directions. $\overrightarrow{h}$ and $\overleftarrow{h}$ were shallow concatenated $([\overrightarrow{h}; \overleftarrow{h}])$ into h_t and used as the input for the next layer. Similarly, the one-hot encoded POS-vectors p_t underwent feature transformation and were concatenated $([\overrightarrow{h}_{POS}; \overleftarrow{h}_{POS}])$ into POS-features h^p_t.

3. Self-attention: Next, the model stacked a single-head self-attention layer that calculated for each POS-tag feature at time t in the sequence a weighted average of the feature representation of all other POS-tag features in the sequence a^p_t (Vaswani et al., 2017). This improves the signal-to-noise ratio by out-weighting important POS features. Attention-weighted POS features and h_t were shallow concatenated into $([a^p_t; h_t])$ vector.

4. Decoder: The attention-weighted representation $([a^p_t; h_t])$ was fed to a linear layer to predict the tag emission sequence $\hat{y}_t$ followed by a CRF layer that takes as input the $\hat{y}_t$ sequence along with the true tag y_t sequence (Huang et al., 2015).

5 Experiments

The experiments were designed to evaluate the performance of the distant NER models trained with the DISTANT-CTO set alone *vs.* DISTANT-CTO set in combination with the EBM-PICO training set. The EBM-PICO training set is naturally composed of both positive and negative annotation sentences, but for the DISTANT-CTO, we artificially generated the negative sentences A_-. To evaluate the impact of these negative annotation sentences, we perform ablation experiments, training the models only with positive annotation sentences A_+. Finally, we also evaluate the performance when training using the entity annotations with match score $d_s = 1.0$ alone *vs.* entity annotations with $d_s \geq 0.9$. A simple SciBERT-CRF model trained using positive annotation sentences A_+ was used as the baseline. Transformer-based models incorporate sequence order and self-attention components, so our baseline served to check the impact of removing costly BiLSTM and self-attention modules.

5.1 Benchmark datasets

We evaluate our weakly annotated dataset and the NER model on the following PICO benchmarks.

1. **EBM-PICO gold.** The EBM-PICO dataset developed by Nye *et al.* consists of 5000 PICO entity/span annotated documents [1]. It comes pre-divided into a training set (n=4,933) annotated through crowd-sourcing and an expert annotated test set (n=191) for evaluation purposes. We use the training set for combined training experiments and the test set for evaluation.

2. **Physio set.** A test set comprising 153 PICO entity/span annotated documents from Physiotherapy and Rehabilitation RCTs (Randomized Controlled Trials) was used as an additional benchmark to evaluate the generalization power of our approach for this subdomain (Dhrangadhariya et al., 2021).

5.2 Experimental Setup

We define the following experimental setups based on the motivations described in section 5:

- **Exp 1.0 distant** A_{+-} **c[1,0.9] wPOS** The setup is composed of SciBERT BiLSTM CRF trained on the surface form (text) and attention-weighted POS inputs using DISTANT-CTO set comprising entity-annotated sentences A_{+-} with $d_s \geq 0.9$.

- **Exp 1.1 distant** A_{+-} **c[1] wPOS** The setup is composed of SciBERT BiLSTM CRF trained on the surface form and attention-weighted POS inputs using the DISTANT-CTO set comprising only the entity-annotated sentences A_{+-} with $d_s = 1.0$.

- **Exp 1.2 distant** A_+ **c[1] wPOS** The setup is composed of SciBERT BiLSTM CRF trained on the surface forms and attention-weighted POS inputs using DISTANT-CTO set comprising only the $d_s = 1.0$ annotations. The negative annotation sentences were removed in this case and the system was trained with positive annotated candidates A_+ only.

- **Exp 1.3 distant** A_+ **c[1] POS** $\neg$ **BiLSTM attention** The setup is composed of SciBERT CRF trained on the surface form inputs using DISTANT-CTO set comprising only the $d_s = 1.0$ annotations with only positive annotated candidates A_+. Attention weights were removed from the POS inputs. This setup was used as the baseline.

- **Exp 2.0 - Exp 2.3** These experiments are identical to their series 1.x counterparts except that the models are trained on a combination of the DISTANT-CTO with the EBM-PICO training set. Exp 2.3 using SciBERT-CRF architecture was used as another baseline.

5.3 Evaluation

To evaluate the quality of automatic annotation using the DISTANT-CTO approach, we performed manual annotation of the "Intervention" class over 200 randomly selected samples from the dataset and compared it to the automatic annotations.

Model evaluation was carried out by predicting the "Intervention" tokens for both benchmarks. Each experiment was conducted thrice with three random seeds (0, 1, and 42), and the average metrics (Precision, Recall, and F1) over three repetitions were reported. We evaluated the statistical significance of our best model using the paired student's t-test as described in (Dror et al., 2018). Further experimental details are in the Appendix.

6 Results

This section reports empirical results for the candidate generation process, evaluation for the annotation quality of DISTANT-CTO approach using the validation sets (see Table 2), and the average of the performance metrics and standard deviation σ over three random seeds on both benchmark datasets for the described NER experiments (see Table 4). We compare the performance of our weakly-supervised NER models with the previous SOTA fully supervised (FS) methods that train on the EBM-PICO training set and evaluate on EBM-PICO gold and also a weakly supervised approach (see Table 3). These models were separately trained for each of the PICO entities/spans and also clubbed the "Intervention" and "Comparator" together.

6.1 Candidate Generation

A total of 360,395 CTO records were downloaded as of March 2021. From all the downloaded CTO records, we extract 200,545 unique (391,286 redundant) intervention names from the aforementioned

[1] A single document consists of a title and an abstract.

intervention sources. Out of the 391,286 intervention terms retrieved, 104,433 terms were successfully mapped to one of the target sentences with the $d_s = 1.0$, and 3084 more were mapped with a score of 0.9. Adding $d_s \geq 0.9$ mappings did not increase the total number of annotated sentences, but it did increase the number of annotations obtained in each sentence. Table 1 shows the total number of intervention annotations obtained from mapping the source terms to target sentences. Metrics for the manual evaluation of DISTANT-CTO using the validation set show that adding annotations with $d_s \geq 0.9$ increases the recall by 3%, but lead to an expected drop in the precision (see Table 2).

Annotation level	$d_s = 1.0$	$1.0 < d_s \geq 0.9$
mention-level	943,284	17,199
token-level	1,515,868	43,096

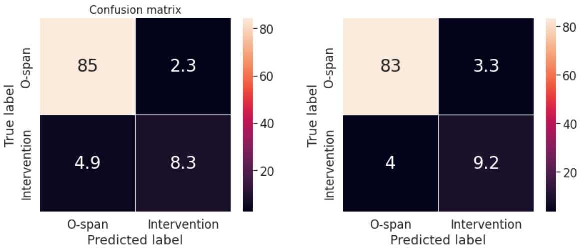
Table 1: Token-level and mention-level intervention annotations obtained in the weakly annotated DISTANT-CTO dataset grouped by their d_s scores.

Match score	**P**	**R**	**F1**
$d_s = 1.0$	0.86	0.80	0.83
$d_s \geq 0.9$	0.84	0.83	0.84

Table 2: Macro-averaged evaluation metrics for the $d_s = 1.0$ and ≥ 0.9 entity annotations for the validation set detailed in the section 5.3

6.2 Model Training

Using the DISTANT-CTO set alone with the NER approach (Exp 1.1 Table 3 and 4) crosses the previous SOTA F1 on the EBM-PICO benchmark by 2%. The best overall F1 for both benchmarks is reached upon training the NER models with combined weakly-labeled DISTANT-CTO with the strongly-labeled EBM-PICO dataset (Exp 2.0 Table 4) crossing the previous SOTA F1 by 5% on the EBM-PICO benchmark. The improvement in F1 for the combined experiments (see Exp 2.1 and 2.0 Table 4)) is significant when compared to the their best DISTANT-CTO counterparts (see Exp 1.1 Table 4)). Using DISTANT-CTO alone has good precision across the experiment series 1.x, but combining it with the EBM-PICO further improves the recall and balances out the F1 in the experiment series 2.x. Adding the artificially generated A_{-} sentences increases the previous F1 by 5.71% and 3.77% (compare Exp 2.2 with Exp 2.1) for both the benchmarks. Note that adding these negative sentences results in an important improvement of about 9% in the F1 for the Physio dataset that is specific for the domain of physiotherapy and rehabilitation. For the combined experiment, the addition of the $d_s \geq 0.9$ annotations improves the F1 as well by a small margin for the EBM-PICO benchmark (Exp 2.0 I.) but has a marginal performance loss for the Physio benchmark (Exp 2.0 II.). While using the DISTANT-CTO alone with the $d_s \geq 0.9$ annotations boosts the precision but downgrades recall thereby reducing the F1 for both benchmarks.



(a) $d_s = 1.0$

(b) $d_s \geq 0.9$

Figure 3: Confusion matrices for the evaluation of DISTANT-CTO validation set annotations with a) $d_s = 1.0$ and b) $d_s \geq 0.9$.

Type	Method	P	R	F1
FS	Nye (2018)	**84.00**	61.00	70.00
FS	Beltagy (2019)	61.00	70.00	65.00
FS	Brockmeier (2019)	69.00	47.00	56.00
FS	Stylianou (2021)	69.04	79.24	73.29
WS	Liu (2021)	22.00	54.00	31.00
WS	Exp 1.1 (Our)	83.36	70.38	75.02
HS	Exp 2.0 (Our)	76.93	**80.17**	**78.44**

Table 3: Comparison of DISTANT-CTO NER models against the previous SOTA NER methods for “Intervention” recognition in terms of macro-averaged precision (P), recall (R), and F1 scores. Boldface represents the best score. Note: FS = Fully Supervised, WS = Weakly Supervised, HS = Hybrid Supervision.

7 Error Analysis

7.1 Candidate Generation

Confusion matrices (see Figures 3a and 3b) for manual evaluation of DISTANT-CTO validation set show that relaxing d_s from 1.0 to 0.9 does improve the true positives (TP) and reduce false negatives (FN) by 0.9% for the “Intervention” class but also reduce the precision by increasing false

Experimental setup	P	R	F1 ±σ	P	R	F1 ±σ
	I. EBM-PICO gold			II. Physio set		
Exp 1.0	88.85	65.39	71.27 ±0.007	86.13	63.70	69.14 ±0.003
Exp 1.1	83.36	70.38	75.02 ±0.013	79.45	66.28	70.63 ±0.008
Exp 1.2	74.85	68.74	71.25 ±0.005	70.52	66.37	68.14 ±0.002
Exp 1.3 (baseline 1)	85.82	64.84	70.31 ±0.002	79.97	60.79	65.14 ±0.005
Exp 2.0	76.93	80.17	**78.44*** ±0.006	75.55	79.42	77.32 ±0.010
Exp 2.1	77.10	78.83	77.89 ±0.007	76.29	80.18	**78.07*** ±0.009
Exp 2.2	67.65	85.02	72.18 ±0.009	64.80	83.69	68.75 ±0.011
Exp 2.3 (baseline 2)	70.91	77.38	73.60 ±0.025	71.50	78.40	74.38 ±0.020

Table 4: Macro-averaged performance metrics for the NER models trained on weakly annotated DISTANT-CTO alone *vs.* in combination to the strongly annotated EBM-PICO on the two described benchmarks (EBM-PICO gold and the Physio corpus). Bold is the best experiment score. Asterisk (*) denotes a significant F1-score of the experiment to its counterpart in the series 1.x. Significance tested using the paired student's t-test.

positives by 1%. Improved recall for the "Intervention" class is undoubtedly preferred, and hence it is vital to inspect the cause of false negatives. A considerable chunk of false negatives was either i) missed intervention abbreviations and the synonyms not mentioned under the sources, or ii) when only the partial intervention name was mentioned in the source, or iii) if specific intervention terms from the source were mentioned in the target but with different word order (see Table 5). This detailed post-hoc error analysis also revealed that 67% false negatives fell under non-drug type composite intervention mentions (phrase mentions of more than two words). For instance, although the term *'Home-based Rehabilitation using Interactive devices'* is expressed in the sentence *'This study investigates clinical outcomes after the rehabilitation by interactive home-based devices.'*, it will remain unmapped to it because the term does not map to the target text using our alignment heuristic. The problem lies in the lack of naming conventions for non-pharma treatment mentions that are neither clearly identified nor standardized as semantic units(Dhrangadhariya et al., 2021). There are two possible programmatic solutions to this. The first is using additional external ontologies as sources of distant supervision which improves coverage of our labeling function to detect further writing variations within the text. Another solution to matching such source and target text is using order-free string matching algorithms (Apostolico et al., 1992). Using external ontologies solves the issues of missed synonyms, and adding an external dictionary of treatment abbreviations could solve the problem of missed abbreviations (Fries et al., 2021). We noticed that the "Comparator" terms (e.g., placebo, sham, saline, etc.) were often not mentioned as structured sources. The development of a general comparator term dictionary could improve this. Improving the coverage and reducing the false negatives (thereby improving recall) using these methodologies suggests an area where future work would be valuable. Most false positives were a result of bigram matching. We will modify fuzzy bigram matching to relevant bigram matching, thereby reducing the occurrences of spurious false-positive bigrams as matches. Only frequently occurring bigrams from the source will be matched to the targets. We plan to explore the quality of DISTANT-CTO for $d_s \leq 0.9$.

Category	FN count
Missed synonym	168
Missed abbreviation	77
Partial match (incl. boundary errors)	361
Missed comparator term	43
Reorder	39
Total	688

Table 5: Distribution of the false negatives in the DISTANT-CTO evaluation corpus.

7.2 Model Training

Manual error analysis was carried out for both the PICO benchmarks, and the error counts for EBM-PICO gold are reported in Table 6. Each token level error was divided into either of the four classes: 1) false negative (FN) - if the entire entity that the token as part of was missed out by the NER model prediction, 2) false positive (FP) - if the entire entity that the token was part of was falsely recognized as "Intervention", 3) boundary error (BE) - if the boundary tokens were missed out but otherwise the entity was identified by the NER model prediction, and 4) overlapping error (OE) - if the NER model made an error in the non-peripheral tokens of an otherwise identified entity mention. Non-peripheral tokens are all the tokens except the first and the last token of the multi-token entity/span.

Models trained on DISTANT-CTO alone had a fewer boundary and overlapping errors, meaning they missed out on many "Intervention" entity signals leading to high precision but compromised recall. On the contrary, NER models trained on combined datasets made twice the more BE and six times more OE. While most BE and OE in the 1.x series were false negatives, they were false positives in the 2.x series leading to a higher recall. This could be because the EBM-PICO training set annotated the longest possible intervention span resulting in spans rather than pure entities in the DISTANT-CTO approach. Combined training set models also picked out names of treatments, surgeries, and enzymes not used as treatments in the RCT as intervention mentions. A huge chunk of overall FN (including the FN tokens in BE and OE) was for entities with composite intervention terms containing two or more tokens. We noticed that the NER system also missed several short intervention names and abbreviations. Overlapping errors occurred when multiple intervention names were mentioned together, separated by either comma or punctuation, or other conjunctions. The error analysis revealed some issues within EBM-PICO ground truth, which had inconsistencies with the intervention boundaries for whether intervention frequency, dose, and the way of administration should be marked as "Intervention". Several times, the ground truth marked articles preceding the entity and prepositions and punctuation succeeding the entity. Extended error analysis can be found in the Appendix.

Exp	FP	FN	BE	OE
	EBM-PICO gold			
Exp 1.0	819	1688	559	66
Exp 2.0	759	1112	1278	515
Exp 1.1	790	1152	650	55
Exp 2.1	793	1039	1327	517

Table 6: Distribution of the token-level errors made by the corresponding NER models on EBM-PICO gold.

8 Conclusions and Future Work

We exploit the freely-available clinicaltrials.org (CTO) and distant supervision for developing the largest available weakly annotated database of Intervention-Comparator entities across 11 subtypes. Using these weak annotations combined with the manual annotations, we train an "Intervention" NER model that surpasses current approaches by more than 5% in terms of F1 on the EBM-PICO gold benchmark and demonstrate strong generalizability on a domain-specific physiotherapy benchmark. When the same NER model was trained with the weakly annotated dataset alone, it surpassed other approaches by 2%. This is a prototypical work, and an automatically obtained dataset with I and C annotations are being extended for the Participant (P), Outcome (O), and Study type (S) entities. The code and data are available on Github.

References

David Ifeoluwa Adelani, Michael A Hedderich, Dawei Zhu, Esther van den Berg, and Dietrich Klakow. 2020. Distant supervision and noisy label learning for low resource named entity recognition: A study on hausa and yorùbá. *arXiv preprint arXiv:2003.08370*.

Alberto Apostolico, Maxime Crochemore, Zvi Galil, and Udi Manber. 1992. Combinatorial pattern matching third annual symposium tucson, arizona, usa, april 29–may 1, 1992 proceedings. In *Conference proceedings CPM*, page 236. Springer.

Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. 2017. Generalisation in named entity recognition: A quantitative analysis. *Computer Speech & Language*, 44:61–83.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Lidong Bing, Bhuwan Dhingra, Kathryn Mazaitis, Jong Hyuk Park, and William W Cohen. 2017. Bootstrapping distantly supervised ie using joint learning and small well-structured corpora. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Florian Boudin, Jian-Yun Nie, Joan C Bartlett, Roland Grad, Pierre Pluye, and Martin Dawes. 2010. Combining classifiers for robust PICO element detection. *BMC medical informatics and decision making*, 10(1):1–6.

Austin J Brockmeier, Meizhi Ju, Piotr Przybyła, and Sophia Ananiadou. 2019. Improving reference prioritisation with PICO recognition. *BMC medical informatics and decision making*, 19(1):1–14.

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 *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 65–77. Springer.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392.

Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S Weld. 2008. Open information extraction from the web. *Communications of the ACM*, 51(12):68–74.

Jason A Fries, Ethan Steinberg, Saelig Khattar, Scott L Fleming, Jose Posada, Alison Callahan, and Nigam H Shah. 2021. Ontology-driven weak supervision for clinical entity classification in electronic health records. *Nature communications*, 12(1):1–11.

Athanasios Giannakopoulos, Claudiu Musat, Andreea Hossmann, and Michael Baeriswyl. 2017. Unsupervised aspect term extraction with B-LSTM & CRF using automatically labelled datasets. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 180–188, Copenhagen, Denmark. Association for Computational Linguistics.

Malcolm W Greaves. 2014. *Relation Extraction using Distant Supervision, SVMs, and Probabilistic First Order Logic*. Ph.D. thesis, Carnegie Mellon University.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Michael A Hedderich, Lukas Lange, and Dietrich Klakow. 2021. Anea: Distant supervision for low-resource named entity recognition. *arXiv preprint arXiv:2102.13129*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ke-Chun Huang, I-Jen Chiang, Furen Xiao, Chun-Chih Liao, Charles Chih-Ho Liu, and Jau-Min Wong. 2013. PICO element detection in medical text without metadata: Are first sentences enough? *Journal of biomedical informatics*, 46(5):940–946.

Ke-Chun Huang, Charles Chih-Ho Liu, Shung-Shiang Yang, Furen Xiao, Jau-Min Wong, Chun-Chih Liao, and I-Jen Chiang. 2011. Classification of PICO elements by text features systematically extracted from pubmed abstracts. In *2011 IEEE International Conference on Granular Computing*, pages 279–283. IEEE.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Di Jin and Peter Szolovits. 2018. PICO element detection in medical text via long short-term memory neural networks. In *Proceedings of the BioNLP 2018 workshop*, pages 67–75. Association for Computational Linguistics.

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.

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.

Minlong Peng, Xiaoyu Xing, Qi Zhang, Jinlan Fu, and Xuanjing Huang. 2019. Distantly supervised named entity recognition using positive-unlabeled learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2409–2419, Florence, Italy. Association for Computational Linguistics.

David Pinto, Andrew McCallum, Xing Wei, and W Bruce Croft. 2003. Table extraction using conditional random fields. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 235–242. ACM.

John W Ratcliff and David E Metzener. 1988. Pattern-matching-the gestalt approach. *Dr Dobbs Journal*, 13(7):46.

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11, page 269. NIH Public Access.

Esteban Safranchik, Shiying Luo, and Stephen Bach. 2020. Weakly supervised sequence tagging from noisy rules. In *Proceedings of the AAAI Conference* on *Artificial Intelligence*, volume 34, pages 5570–5578.

Alisa Smirnova and Philippe Cudré-Mauroux. 2018. Relation extraction using distant supervision: A survey. *ACM Computing Surveys (CSUR)*, 51(5):1–35.

Nikolaos Stylianou and Ioannis Vlahavas. 2021. Transformed: End-to-end transformers for evidence-based medicine and argument mining in medical literature. *Journal of Biomedical Informatics*, 117:103767.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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.

Zeyi Wen, Dong Deng, Rui Zhang, and Ramamohanarao Kotagiri. 2019. : An efficient entity extraction algorithm using two-level edit-distance. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 998–1009. IEEE.

Yaosheng Yang, Wenliang Chen, Zhenghua Li, Zhengqiu He, and Min Zhang. 2018. Distantly supervised ner with partial annotation learning and reinforcement learning. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2159–2169. Association for Computational Linguistics.

Tengteng Zhang, Yiqin Yu, Jing Mei, Zefang Tang, Xiang Zhang, and Shaochun Li. 2020. Unlocking the power of deep PICO extraction: Step-wise medical ner identification. *arXiv preprint arXiv:2005.06601*.

A Appendix

A.1 DISTANT-CTO characteristics

The total number of entity-level “Intervention” mentions in DISTANT-CTO are almost 30 times more than in the EBM-PICO dataset as shown in Table 7. For the EBM-PICO training set, 57.48% of mentions fell under the “drug” class and the rest under the six remaining classes.

Total	**DISTANT-CTO**	**EBM-PICO**
mention-level	977,682	32,890
token-level	1,558,964	125,920

Table 7: Comparing the number of “Intervention” annotations in DISTANT-CTO *vs.* EBM-PICO.

Out of all the mention-level annotations in the DISTANT-CTO dataset, 59.90% corresponded to “drug” class and 40% to the rest of 10 classes. The pie chart (upper pie in Figure 4) shows the class distribution of the semantic classes for the retrieved “Intervention” mentions s_{name} about half of which fall under the “drug” (or Pharma) class and the rest under the remaining 10 non-pharma classes. Out of the total retrieved mentions, almost two-thirds that get mapped to a target t sentences also fall under the “drug” class (lower pie in Figure 4). Table 8 and 9 shows the number of retrieved intervention mentions by their semantic class *vs.* the percentage of these intervention mentions that get mapped to some target sentence with the match score d_s of 1.0 and score 0.9 respectively. Notice that collectively the intervention mentions that fall under the non-pharma classes outnumber the pharma (“drug”) mentions.

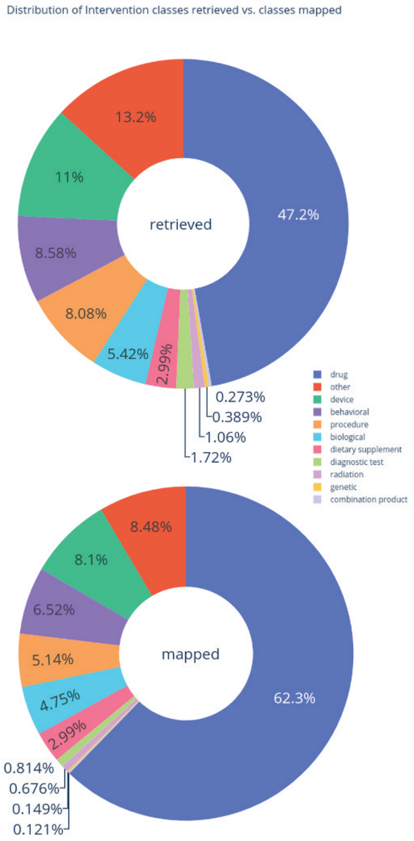


Figure 4: upper) Class distribution for the retrieved “Intervention” mentions, and lower) Class distribution for the mapped “Intervention” mention.

.

Top semantic classes for the most mapped and most unmapped intervention mentions from the total retrieved mentions are shown in the figure 6 and 5. As evident from the tables 8 and 9 “drug” class intervention mentions are the most mapped

Domain	retrieved - (mapped)
drug	184835 (35.50%)
device	43134 (20.09%)
other	51703 (16.19%)
procedure	31630 (21.38%)
behavioral	33590 (16.03%)
biological	21225 (22.86%)
dietary supplement	11699 (25.46%)
radiation	4134 (20.44%)
diagnostic test	6742 (10.13%)
combination product	1070 (14.39%)
genetic	1524 (07.94%)
all non-pharma	206,451 (18.80%)

Table 8: Number of intervention mentions retrieved *vs.* percentage mapped with $d_s = 1.0$

Domain	retrieved - mapped
drug	184835 (36.22%)
device	43134 (21.13%)
other	51703 (16.84%)
procedure	31630 (22.16%)
behavioral	33590 (16.44%)
biological	21225 (24.07%)
dietary supplement	11699 (27.44%)
radiation	4134 (21.17%)
diagnostic test	6742 (10.78%)
combination product	1070 (14.95%)
genetic	1524 (08.53%)
all non-pharma	206,451 (19.64%)

Table 9: Number of intervention mentions retrieved *vs.* percentage mapped with a d_s of 0.9

followed by “dietary supplement” and “procedure” classes which also reflects in the pie chart of most mapped lengths and common phrase lengths for each class (see Figure 5). The most frequent phrase length for these classes is one (unigram) and the second most frequent length is two (bigram).

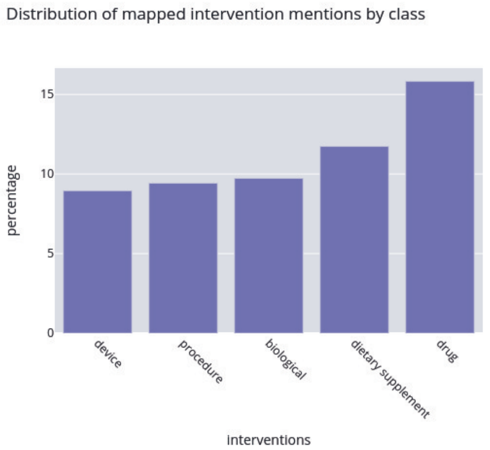


Figure 5: Top five semantic classes, source intervention mentions from which get mapped to the target.

.

Domain	Most common length
drug	1
dietary supplement	1
biological	1
procedure	2
device	1

Table 10: Lengths for the most mapped classes

The least mapped intervention mention classes are “combination product”, “diagnostic test” and “behavioral” (refer figures 6) with most intervention mentions in these classes containing either trigrams or bigrams. This very well reflects with the numbers in figures 7 which shows that trigram and bigram intervention mentions constitute almost half the right pie showing the top phrase lengths for intervention mentions that remain unmapped. One of the ways to retain some of the missed bigram and trigram intervention mentions is to explore the matches with lower match scores. The Table 12 shows some of the $d_s \geq 0.9$ source-target matches not captured by the $d_s = 1.0$ constraint because of the difference of either a single missing space or singular-plural differences. It is also interesting to note that the radiographic procedure “cystourethrography” matches the name of the test “cys-

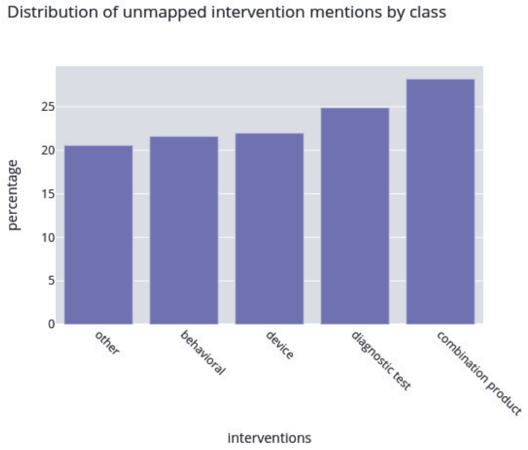


Figure 6: Top five semantic classes of the source intervention mentions that remain unmapped to the target.

.

Domain	Most common length
device	3
other	2
behavioral	3
diagnostic test	2
combination product	3

Table 11: Lengths for the most unmapped classes

tourethrogram”.

A.2 Experimental Details

For the candidate generation process, we did not define any junk elements for using the `sequencematcher` function. All the NER experiments in this article were conducted in PyTorch and the models were trained for 10 epochs with a mini-batch size of 10 for training and 6 for evaluation. We used the IO (Inside, Outside) also called raw labeling for all the NER tasks to make the experiments compared with the previous studies. The maximum sequence length was set to 100 because the average length of each input text sequence was about 68 words. For both experiments types, either using the DISTANT-CTO alone or with the EBM-PICO training set, 80% of the data was used for training and 20% for development. The [CLS] embeddings from the SciBERT layer were used as features of the input text. SciBERT was fine-tuned by not freezing weights during the experiments. The hidden size for LSTM/BiLSTM was set to 512/1024 for the text input embeddings and 20/40 for the POS one-hot embeddings. ReLU was used as the activation function before feeding emission outputs to the CRF layer. Model training was optimized using AdamW using a learning rate of 5e-5. The gradients were clipped to 1.0 to mitigate the problem of exploding gradients. Due to very specific RAM and GPU requirements for each experiment and the institute’s capacity for sharing the GPUs amongst the group members, experiments were carried out on the following GPUs. Each experiment was carried out on a single GPU without any data and model parallelization.

B Extended Error Analysis

Manual error analysis results for Physio corpus are reported in the Table 14. FP error count was always lower than the FN error count in the EBM-PICO gold but for the Physio set, the combined NER experiments (series 2.x) lead to a higher FP compared the FN. The ratio of BE in Exp series 2.x is on an average 1.2 times that of series 1.x. However, a large chunk of BE in series 1.x are false negatives in contrast to the BE in series 2.x which are false positives. Upon closer inspection of false-negative BE in series 1.x, we found that they were either missed intervention synonyms inside brackets, missed information accompanying intervention terms like dose, type, medium of intervention, administrator of intervention, or location of administration. This is due to the fact that distantly supervised annotation does not take into account labeling the additional intervention information except the name. The addition of the manually annotated EBM-PICO in the combined training experiments reduces the number of false-negative BE. This is due to the fact that EBM-PICO guidelines required the annotators to mark the longest possible phrase describing intervention including the additional information like dose, mode, medium, and location of administration.

For both the evaluation corpora, the combined NER experiments lead to more TP for the “Intervention” class which is vital to PICO entity/span recognition. This could be the case because the combination of weakly and strongly annotations reduce the percentage of unseen surface forms (words) from both test sets. 27.70% of the intervention entity surface forms in the EBM-PICO gold benchmark remain unseen in the EBM-PICO training set while for the DISTANT-CTO training set it drops to 21.38%. 27.29% of the intervention entity surface forms in the Physio benchmark remain unseen in the EBM-PICO training set while for

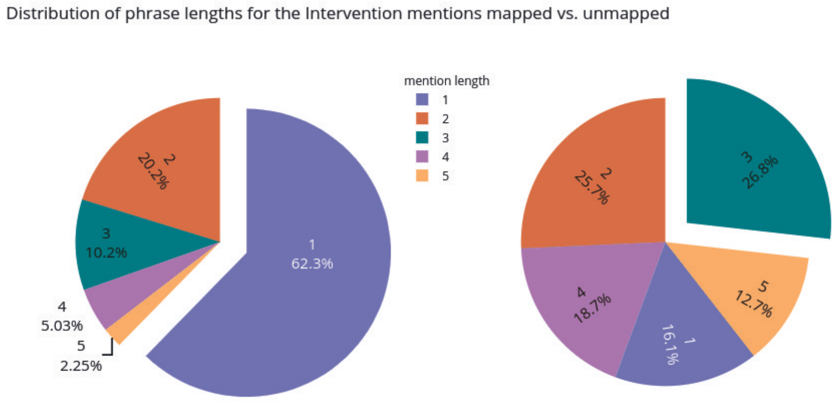


Figure 7: Left) Phrase length distribution of mapped intervention mentions, Right) Phrase length distribution of unmapped intervention mentions.

.

Characteristic	Source	Target	d_s
Single missing space	"l carnitine"	"lcarnitine"	0.923
Missing negations	"no pumice prophylaxis"	"pumice prophylaxis"	0.900
Plurals	"punch skin biopsies"	"punch skin biopsy"	0.941
Abbreviations	"rfsh alone"	"recombinant fsh alone"	0.926
Specific treatment name to generic treatment name	"biphasic insulin aspart 50"	"biphasic insulin aspart"	0.923
Procedure matches the instrument	"cystourethrography"	"cystourethrogram"	0.900

Table 12: Example "Intervention" mentions from CTO that get mapped to target sentences t with a d_s of 0.9

GPU	RAM	Experiment
Tesla V100-PCIE-16GB	1TB	1.1, 2.1, 1.3
TeslaK80 GPU	126GB	1.2, 2.2
Tesla V100-PCIE-32GB	1TB	2.0, 1.0, 2.3

Table 13: Experiments and the details of GPUs they were carried out on.

Exp	FP	FN	BE	OE
	Physio set			
Exp 1.0	963	1586	654	20
Exp 2.0	1168	897	867	347
Exp 1.1	990	1420	723	19
Exp 2.1	1116	904	1025	228

Table 14: Distribution of the token-level errors made by the corresponding NER models on Physio set.

the DISTANT-CTO training set it drops to 22.97%. Combining both training sets leads to a reduction in unseen surface forms to 16.29% and 15.13% for the EBM-PICO gold and Physio benchmarks respectively. (Augenstein et al., 2017) has shown that recall on unseen surface forms is significantly lower than on seen surface forms for NER tasks.

C Ethical Statement

This paper studies clinical NER with a small strongly labeled and a large weakly labeled dataset. Our investigation neither introduces any social or ethical bias to the model nor amplifies any bias in the data. We do not foresee any direct social consequences or ethical issues.

D License Information

DISTANT-CTO uses all of clinicaltrials.gov (CTO) data that allows downloading and using it given that any publication/distribution states and describes any modifications made to the content of the data. It is public data that anyone can download and reproduce the outcomes with the code made available on Github.