PICO to PICOS: Weak Supervision to Extend Datasets with New Labels

Anjani DHRANGADHARIYA [a,b,c,1], Gaetano MANZO [d] and Henning MÜLLER [a,b,e]
[a] *Informatics Institute, HES-SO Valais-Wallis, Sierre, Switzerland*
[b] *University of Geneva (UNIGE), Geneva, Switzerland*
[c] *School of Health Sciences, HES-SO Valais-Wallis, Leukerbad, Switzerland*
[d] *Computational Health Research Branch, NLM, Bethesda, Maryland, USA.*
[e] *The Sense research and innovation center, Lausanne and Sion, Switzerland*
ORCiD ID: Anjani Dhrangadhariya https://orcid.org/0000-0003-1691-1338

Abstract. Hand-labelling clinical corpora can be costly and inflexible, requiring re-annotation every time new classes need to be extracted. PICO (Participant, Intervention, Comparator, Outcome) information extraction can expedite conducting systematic reviews to answer clinical questions. However, PICO frequently extends to other entities such as Study type and design, trial context, and timeframe, requiring manual re-annotation of existing corpora. In this paper, we adapt Snorkel's weak supervision methodology to extend clinical corpora to new entities without extensive hand labelling. Specifically, we enrich the EBM-PICO corpus with new entities through an example of "Study type and design" extraction. Using weak supervision, we obtain programmatic labels on 4,081 EBM-PICO documents, achieving an F1-score of 85.02% on the test set.

Keywords. weak supervision, information extraction, clinical NLP

1. Introduction

Despite the existence of large annotated corpora like EBM-PICO, manual PICO analysis remains essential. This analysis, which focuses on Participant, Intervention, Comparator, and Outcome elements, often requires additional information such as study type, design, context, and trial duration. These details are crucial for comprehensive evidence synthesis, but static datasets like EBM-PICO lack labels for them [1]. Clinical corpora, labelled by domain experts, serve specific purposes but are static and pose challenges in adaptability to new tasks as when PICO extends to additional analysis. Weakly supervised (WS) information extraction (IE) techniques offer promise by programmatically labeling datasets using publicly-available sources like UMLS and NCBO BioPortal. Fries *et al.* 2021 and Dhrangadhariya *et al.* 2023 used weak supervision for biomedical and clinical (PICO) entity extraction [2,3]. They, however, did not tackle the challenge of extending a corpus like EBM-PICO to new, relevant entities. Our work pioneers weak supervision for enhancing hand-labelled datasets like EBM-PICO with new clinical entities [4].

[1] Corresponding Author: Anjani Dhrangadhariya, Informatics Institute, HES-SO Valais-Wallis, Technopole 3, 3960 Sierre, Switzerland; E-mail: anjani.dhrangadhariya@hevs.ch.

Leveraging WS framework, we successfully label "Study type and design" across 4,081 EBM-PICO documents. Evaluation using a 191 manually labeled documents confirms the efficacy of our approach, offering a pragmatic solution without relying on domain experts. Our contribution expands the application of WS clinical IE, aiding faster SRs.

2. Methods

Figure 1 schematically represents our below-described approach.

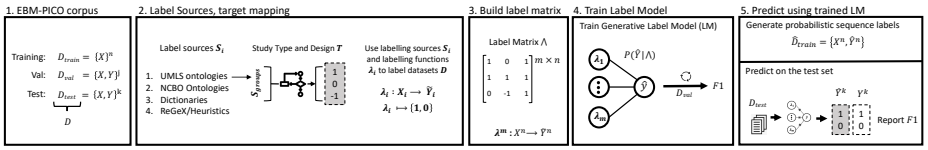

Figure 1. WS approach: 1. Define the training, validation, and test sets. 2. Define labelling sources S_i. UMLS vocabularies are reused as labelling sources and mapped to the"Study type and design" class labels. 3. LFs λ_i map the training set to class labels using S_i resulting in an $m \times n$ label matrix Λ. 4-5) The Λ is used to train a generative LM that could be used to label unlabelled training sets with probabilistic labels.

Dataset: EBM-PICO was used to demonstrate the effectiveness of our approach. It comes pre-divided into training (n=4,933) and test (n=191) sets [4]. The training set was further segmented into a validation set comprising 721 documents. Hand-labelled validation and test sets are necessary for hyperparamter tuning and evaluation, respectively. To hand label these datasets, annotation guidelines for the "Study type and design" class were developed. First, the test set was doubly-annotated to calculate pairwise F1 measure as measure of inter-annotator agreement (IAA) [5]. The IAA was 78.33% and deeming it as sufficient, the validation set was singly annotated. The training set was labelled with "Study type and design" class using the weak supervision based programmatic labelling.

2.1. Weak Supervision

Weak supervision based programmatic labelling involves designing m labelling functions (LF) λ_m, each of which is a function that takes input text sequence X and a labelling source s and produces an integer label sequence $\widetilde{Y} = (\widetilde{y_1}, \widetilde{y_2}, \ldots, \widetilde{y_n})$; $\widetilde{y_i} \subset \{1, 0, -1\}$. We used programmatic labelling to label EBM-PICO training set with the "Study type and Design" class with the target labels $\widetilde{y_i}$. The label 1 represents "Study type and design" or positive class label, 0 represents a negative class label, and -1 are abstains. The ground truth Y is latent and estimated by aggregating outputs from multiple LFs, resulting in $\hat{Y}$, which serves as probabilistic token labels for X. We used programmatic labelling using the below-described methods to weakly label the EBM-PICO training set with "Study type and Design" class.

Labelling Sources: A labelling source s can be a set of terms, expert-designed ReGeX, heuristics, or a combination of these sources that encode some domain-specific knowledge. We used the 2021AB-full release of the UMLS Metathesaurus English subset after excluding zoonotic and non-English vocabularies, resulting in a pool of 112 vocabularies [6]. Labelling the "Study type and design" class entails using terms or concepts

to represent this class by aligning the UMLS concepts onto the raw text EBM-PICO training set. UMLS concepts are organized under 127 internally-defined semantic groups $S_{groups} = (s_{group_1}, s_{group_2}, ..., s_{group_n})$; $n = 127$ like "disease", "age group", "geographical location" denoting whether a concept represents a disease name or a location. Semantic groupings impart meaning to the concepts and allow the repurposing of UMLS for programmatic labelling of related entities. Our task was to map these S_{groups} to the "Study type and design" class as per their representational value, ultimately mapping the concepts to the class labels.

Non-UMLS ontologies like Clinical Trial Ontology[2], Randomized Controlled Trials Ontology[3], Ontology of Clinical Research[4], and Clinical Trials Ontology[5] were used to represent $(+1)$ class labels. Handcrafted dictionaries were designed using key-phrases from MeSH containing the generic term "trial". The terms (e.g., "quasi-experimental trial", and "crossover trial") in this dictionary were used to label positive class labels.

We examined the most common keyword patterns in "Study type and design" class in the validation set. These class-specific keyword patterns were used as ReGeX hooks along with the observed POS patterns to emit the positive class label. For e.g., the trial design information "double-blind, non-inferiority" preceded the hook pattern "randomized controlled trial". To identify such domain-specific patterns, a ReGeX was developed to identify the hook pattern "randomized controlled trial" and was combined with position and POS tags to identify preceding trial design information.

Source to Target mapping: The concepts in non-UMLS ontologies and the dictionaries were mapped to target label $+1$. To map UMLS S_{groups} to target label $+1$, we conducted a separate experiment using the validation set using the steps:

1. Label the hand-labelled validation set using all the UMLS S_{groups}.
2. Calculate recall for the target label $+1$[6].
3. Rank and sort S_{groups} based on their calculated recall.
4. Next, label the validation set using the S_{group} that ranked 1 (S_1) and calculate the initial recall r and f1-score $f1$.
5. Then loop through the ranked S_{group} starting at rank 2 and sequentially add labels to the validation set (already labelled with S_{group} rank 1) and calculate the new recall r_i and f1-score $f1_i$ with the combined labels.
6. After looping through all the S_{groups}, following heuristic was used to classify S_{group} into representing either the positive $(+1)$, negative (0) or abstain (-1) class. We consider a S_{group} representative of the "Study type and design" class (target label $+1$) if the change in the recall Δr is ≥ 1 without impacting the f1-score. Such S_{group} are marked as $+1$ and the rest as abstain or negative.

LFs: We categorize our LFs into three types depending on the labelling sources. An ontology or dictionary LF takes a set of terms (vocabularies, ontologies, *etc.*) each mapped to one of $y \subset \{0, +1, -1\}$ class labels using heuristics. A ReGeX LF used only regular expressions representative of the positive token label $+1$ and abstained from the rest. A

[2] `https://bioportal.bioontology.org/ontologies/CTO`
[3] `https://bioportal.bioontology.org/ontologies/RCTONT`
[4] `https://bioportal.bioontology.org/ontologies/OCRE/`
[5] `https://bioportal.bioontology.org/ontologies/CTONT/`
[6] The recall and the F1 score are binary metrics calculated for the "Study type and design" (positive) class.

heuristic LF took a generic ReGeX pattern, specific POS (part-of-speech) tags, and token positions to label tokens with the positive label $+1$ and abstained from the rest.

Labelling and Label aggregation: Consider $S = (s_1, s_2, \ldots, s_x)$ set of labelling sources used by m LFs ($\lambda_m; \Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_m\}$) to programmatically label the X_n EBM-PICO training tokens to the integer labels $(-1, 0, 1)$. The LFs map X_n input tokens to the integer label sequence $\widetilde{Y}_n$ leading to a label matrix $\Lambda^{m \times n}$. Majority voting (MV) and Snorkel's generative label model (LM) were tested to aggregate labels in $\Lambda^{m \times n}$ [7].

2.2. Experiments

The experiments were carried out in seven tier and aimed to evaluate the impact of sequentially adding labelling sources on the label aggregation methods. The tiers 1-4 tested the sequential addition of non-UMLS, dictionaries, and rule-based labelling sources to UMLS LFs in sequence. Tier 5 examined whether up-weighting rules could improve performance, while tiers 6 and 7 measured the effect of removing non-UMLS and dictionaries from tier 4. We evaluate performance using token-level macro F1, precision and recall over three runs of experiment tiers with three random seeds.

3. Results and Discussion

	Experiments	MV			LM		
Tiers	LF tier	P	R	F1	P	R	F1 (stdev)
1	UMLS	48.64	50.00	49.31	61.03	56.42	58.02 (4.4×10^{-5})
2	+ non UMLS	51.58	50.01	49.37	50.21	50.02	49.62 (2.2×10^{-4})
3	+ Dictionaries	48.64	49.99	49.31	64.87	62.23	63.16 (3.6×10^{-2})
4	+ Rules	48.64	50.00	49.31	86.03	78.50	81.41 (4.2×10^{-3})
5	+ Rules $\times$ 2	98.64	50.17	49.66	85.09	79.42	81.96 (7.4×10^{-3})
6	- Dictionaries	98.64	50.13	49.59	81.40	72.55	75.37 (1.5×10^{-3})
7	- non UMLS	96.22	53.31	55.56	89.96	81.41	**85.02** (1.7×10^{-2})

Table 1. Macro-averaged recall, precision and F1 % for "Study type and design" extraction models. The best F1 score is shown in bold. Standard deviation (stdev) is reported for average over three runs.

Using the described labelling sources and functions, we developed a total of 144 LFs: 112 UMLS LFs, 10 non-UMLS LFs, 2 dictionary LFs and 20 ReGeX LFs. The results of the experiments are listed in Table 1. LF aggregation via MV fails to detect any meaningful signals and performs at a level close to or even worse than random. For tier 7, however, removing non-UMLS LFs boosts the recall and therefore the F1 for MV. The performance of UMLS alone for the LM tier 1 is poor. Incorporating non-UMLS sources into the model results in a significant drop in F1 score by as much as 8.4% again pointing towards the low representational value of this labelling source. If a labelling source like non-UMLS sources, does not contain many of the terms that are representative of the entity in question, this could cause the F1 score to decrease upon their addition. Our results prove this claim by conducting the ablation experiments. The F1 score for the "Study type and design" entity deprecated on adding non-UMLS LF, suggesting that these func-

tions were not typical for the "Study type and design" entity. When non-UMLS labelling sources were removed from tier 7, a shoot-up in F1 score by 3.61% from tier 4, where all the labelling sources were used. The inclusion of dictionaries boosted the F1 by 13.52%, yet it plateaued at 63.16% As expected, adding generic rules in tier 4 boosted the recall by 18.25% from tier 1. Up-weighting rule-based LFs in tier 5 led to a nominal F1 increase by 0.55%. In tier 6, removing handcrafted dictionaries decreases the previous best recall by 6.87%, demonstrating performance contribution. In tier 7, removing the non-UMLS labelling sources improves the overall F1 by 3.06%.The utility and representational value of dictionaries are evidenced by a decrease of 6.05% in the F1 upon their removal in tier 6. While ReGeX and heuristics designed for the "Study type and design" class may not be directly transferable to other entities, the methodology of developing ReGeX using hook patterns and a small labeled validation set can be effectively extended to other entity classes.

4. Conclusion

We adopted a weak supervision approach to enhance existing EBM-PICO dataset by incorporating additional categories, like "Study type and design" without relying on manual annotation. This is achieved through the application of weak supervision techniques using Snorkel. Our approach achieved exceptional performance, with an F1 score of 85.05% on the hand-labelled EBM-PICO test set, highlighting the potential of this method for rapidly generating large amounts of annotated data compared to traditional supervised approaches. The resources to reproduce this work are available on GitHub.

References

[1] A. M. Methley, S. Campbell, C. Chew-Graham, R. McNally, and S. Cheraghi-Sohi, "PICO, PICOS and SPIDER: a comparison study of specificity and sensitivity in three search tools for qualitative systematic reviews," *BMC Health Serv Res* **14**(1), pp. 1–10, 2014.

[2] J. A. Fries, E. Steinberg, S. Khattar, S. L. Fleming, J. Posada, A. Callahan, and N. H. Shah, "Ontology-driven weak supervision for clinical entity classification in electronic health records," *Nat. Commun* **12**(1), pp. 1–11, 2021.

[3] A. Dhrangadhariya and H. Müller, "Not so weak PICO: leveraging weak supervision for participants, interventions, and outcomes recognition for systematic review automation," *JAMIA open* **6**(1), p. ooac107, 2023.

[4] B. Nye, J. J. Li, R. Patel, Y. Yang, I. J. Marshall, A. Nenkova, and B. C. Wallace, "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*, **2018**, p. 197, NIH Public Access, 2018.

[5] L. Deleger, Q. Li, T. Lingren, M. Kaiser, K. Molnar, L. Stoutenborough, M. Kouril, K. Marsolo, I. Solti, *et al.*, "Building gold standard corpora for medical natural language processing tasks," in *AMIA Annual Symposium Proceedings*, **2012**, p. 144, American Medical Informatics Association, 2012.

[6] B. L. Humphreys, D. A. Lindberg, H. M. Schoolman, and G. O. Barnett, "The unified medical language system: an informatics research collaboration," *J Am Med Inform Assoc* **5**(1), pp. 1–11, 1998.

[7] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré, "Snorkel: Rapid training data creation with weak supervision," in *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, **11**, p. 269, NIH Public Access, 2017.