Enhancing Data Curation for Clinical Trial Registries: Application of Language Models for Drug and Disease Recognition and Normalization

Anonymous ACL submission

Abstract

Clinical trial registry reviews can reveal crucial insights into medical research quality and scope. The current process for generating reports from these registries relies heavily on manual data curation, which includes categorizing trials by disease type and classifying drugs. These tasks are time-consuming and prone to human error. In the present work, we explore the use of automated techniques for extracting drug and disease information, as well as their linking to a medical ontology. By improving the data capture and curation, our aim is to contribute to the development of new systems for reviewing and monitoring clinical trial registries. All resources are available on GitHub[1].

1 Introduction

Public clinical trial registries, such as ClinicalTrials.gov, are essential resources that enable stakeholders—including researchers, patients, healthcare providers, and policymakers—to navigate the landscape of drug development. These registries allow for the monitoring of emerging therapeutic targets and substances, and ensuring that new treatments meet safety and efficacy standards. Additionally, they can facilitate the tracking of adverse drug reactions and support the evaluation of clinical trial design quality (Saberwal, 2021).

However, extracting information from these resources is challenging due to large data volume, incomplete and unstructured reporting and variability in terminology (Tse et al., 2018; Pillamarapu et al., 2019; Shi and Du, 2024). While the Aggregate Analysis of ClinicalTrials.gov database (AACT)[2] has been released in 2011 to enhance access to the data, it provides little automated validation and harmonization of data elements (Tasneem et al., 2012). For example, a recent study found that the interventions section of ClinicalTrials.gov included non-drug-related terms, hindering comprehensive drug trend analysis (Namiot et al., 2023). Therefore, the current process of evidence synthesis from trial registries relies heavily on manual data curation, including the tasks of categorizing trials by disease type and classifying drugs (Hirsch et al., 2013; Liu et al., 2018). This approach is time-consuming and prone to human error, which might result in inconsistencies and missed information.

Computational methods, especially natural language processing (NLP), can support clinical evidence synthesis by structuring, standardizing, and semantically analyzing data (Marshall et al., 2017; Thomas et al., 2017). Techniques like Named Entity Recognition (NER) identify and categorize text elements such as drug and disease names (Wang et al., 2018). Complementary, Entity Linking (EL) matches these identified elements to unique identifiers in knowledge bases and enables entity normalization, i.e., their uniform representation (Shen et al., 2015; Shi et al., 2023).

Thus, we used NER to explore ways of enhancing the existing condition and intervention fields in ClinicalTrials.gov. We compared neural NER outputs with the existing AACT manual annotations and evaluated a state-of-the-art method for linking entities to the Systematized Nomenclature of Medicine Clinical Terminology (SNOMED CT) (Cornet and de Keizer, 2008).

2 Methods

2.1 Reference Corpus

We worked with a dataset of annotated trials from ClinicalTrials.gov (NeuroTrialNER)[3] . This dataset includes entity-level annotations in trial titles and summaries, identifying entities like disease names (called "conditions") and drugs. We analyzed the

[1] https://anonymous.4open.science/r/NeuroTrialDataCuration-3F46/

[2] https://aact.ctti-clinicaltrials.org/

[3] Developed within our group, the work is currently under anonymized review. See Anonymous GitHub.

test set of 153 trials, focusing on those with condition annotations (144 trials, 345 annotations) and drug annotations (50 trials, 100 annotations).

2.2 Named Entity Recognition

2.2.1 Model

We used BioLinkBERT as the model reported to achieve the best results on the test set for condition (F1 0.85) and drug name recognition (F1 0.90) (Yasunaga et al., 2022). In previous work the model was fine-tuned on the NeuroTrialNER train set, and we ran inference on a local CPU.

2.2.2 Evaluation

We were interested to understand what are the differences between the entities extracted from the text using BioLinkBERT, and the existing values provided in the ClinicalTrials.gov (AACT) records.

For each clinical trial, we aggregated the token-wise NER extractions into unique entities at the abstract level to enable comparison with AACT. We then determined whether each unique entity from AACT and BioLinkBERT appeared in one or both annotations. Overlaps were identified based on exact or partial token matches, with partial matches defined by significant character overlap, as described in the Appendix A.

To better understand cases where entities were present only in AACT or the BioLinkBERT extractions, we sampled 20 instances where an entity was returned by only one system. Each instance was manually reviewed and classified either as a synonym, false positive, or as a unique true positive for one system, thus a false negative for the other.

2.3 Named Entities Linking

2.3.1 Manual Annotation

NeuroTrialNER did not include annotations for linking named entities to SNOMED nomenclature. To assess performance, two annotators independently linked each manually annotated condition and drug entity from the test set to the ontology entries using the SNOMED CT web browser[4]. They identified the most accurate matches, extracting the concept name and concept IDs. The process is detailed in Appendix B. Inter-annotator agreement (IAA) was measured using Cohen's kappa statistic, and we report the 95% confidence intervals (CI) (Cohen, 1960).

[4] SNOMED CT Browser

2.3.2 Dictionary Lookup

We used a names dictionary based technique as a simple baseline for the entity linking task. We combined reference terminology from multiple knowledge bases, detailed in Appendix C. This resulted in a dictionary of 25,933 unique drug names and 18,458 unique condition names, including synonyms and lexical variations.

Following the method outlined in Wood (2023), we linked entity words that matched entries from our dictionary. This approach did not accommodate misspellings.

2.3.3 SapBERT and SNOMED

We utilized the Self-alignment Pretraining for BERT (SapBERT) model from the Huggingface library, pre-trained on PubMedBERT full texts, without further fine-tuning or change to the hyperparameters[5]. Inference with the model was performed on local CPU.

We acquired SNOMED CT data from NIH[6], isolating concepts and synonyms in the categories disorder, finding, procedure and medicinal product. SapBERT vector representations were created for each SNOMED concept and synonym.

For each named entity from the test set, we generated a SapBERT embedding and used it to match the closest SNOMED concepts based on Euclidean distance (Huang et al., 2008). Note that his setup did not take the mention's context into account (Kartchner et al., 2023). The top five closest matches and their distances (cdist) were retrieved.

2.3.4 Evaluation

The assessment of the linking quality was performed in terms of precision, recall and the F1-measure, as defined in (Shen et al., 2015) and shown in Appendix D .

2.3.5 Experiments

The "cdist" value can be interpreted as an indicator of the match's accuracy, with larger distances suggesting lower confidence in the match. We aimed to determine an optimal "cdist" threshold, above which entities should not be linked to SNOMED due to a high likelihood of being false positives. To achieve this, we explored various threshold values.

Moreover, it is possible that the manually annotated SNOMED target is not the top match returned by the system but falls within the top k closest

[5] /cambridgeltl/SapBERT-from-PubMedBERT-fulltext

[6] NIH SNOMED CT International Edition, April 1, 2024

matches (k=2,3,4,5). We therefore analyzed how performance varies when considering whether the target entity is among these closely matched entities.

3 Results

3.1 Named Entity Recognition

The UpSet plot in **Figure 1 A** shows the intersection of condition entities extracted by AACT and BioLinkBERT. Around 35% (254) of the entities were recognized by both methods. AACT and BioLinkBERT also uniquely returned 218 and 228 conditions, respectively. For drug entities, 52 drug names were overlapping (**Figure 1 B**). Additionally, BioLinkBERT uniquely identified 54 drugs, while AACT uniquely identified 32 drugs.

To understand the discrepancies between the methods, we manually reviewed a random sample of the entities recognized by only one of the systems and identified the following patterns (see also Table 1 in Appendix E):

- **Different entity surface forms**: The methods identified the same entity, but they had lexical variations. This was more frequent with drug entities (57%) than conditions (42%).
- **Unique entity by one method**: BERT detected more detailed conditions and intervention information. AACT contained entities that BERT could not extract because they were not mentioned in the trial descriptions.
- **False Positives**: AACT had only 2% false positives, while BERT had 15% for conditions and 5% for drugs.

3.2 Named Entity Linking

3.2.1 Manual Linking

The Cohen's kappa score between the two annotators for linking drug entities was 0.85 (CI: 0.78, 0.92), and for linking conditions, it was 0.79 (CI: 0.75, 0.84). The two annotators manually reviewed the disagreements and reached a consensus on the final target SNOMED entity, which was then used for model evaluation.

3.2.2 Dictionary Lookup

Of the 100 drug mentions, 52% were successfully linked using the exact string matching dictionary lookup strategy. This method also successfully linked 123 (36%) of the annotated condition entities. Linking conditions was more challenging because the annotations included extra disease characteristics such as stage and severity, which were not present in our target disease knowledge bases.

3.2.3 Optimal Entity Linking Performance

The highest F1 scores were obtained at a cdist threshold of 7.73 for conditions, achieving an F1 score of 0.76 (**Figure 2 A**). For drug entities, the highest F1 score of 0.92 was achieved at a cdist threshold of 8.18 (**Figure 3 A** in Appendix F).

As seen in **Figure 2 B**, at lower cdist thresholds, the model was more stringent, accepting only very close matches. This resulted in higher precision but lower recall, as the model missed some true matches that had a higher Euclidean distance. Conversely, at higher cdist thresholds, the model was less strict, which increased recall by including more true matches, but also decreased precision due to the inclusion of more false positives.

3.2.4 Performance at different k

Figure 2 B demonstrates the relationship between the number of included closest entities (k) and the performance of the entity linking model. The results showed that while precision and recall increase with the number of closest entities considered. This indicates that the correct entity is frequently found within the top 5 closest entities, suggesting that these entities are closely related. Similar results were obtained for for drug entities, see **Figure 3 B** in Appendix F.

4 Discussion

Analysis of entities unique to either AACT or BERT revealed that the same entity often appeared in both extractions with different surface forms, highlighting the challenge of handling extensive synonyms in the biomedical domain (Kartchner et al., 2023). The neural NER approach offered more detailed and standardized annotations. For example it included disease stages and severity grades, while excluding drug dosage information. This suggests the potential for using this technique to automatically extract and standardize entities from trial descriptions, enhancing the granularity and completeness of the data.

We also showed that a neural entity linker to a standard medical vocabulary could address the challenge of different entity surface forms. This would facilitate data aggregation across different trials and enable analysis at various hierarchical

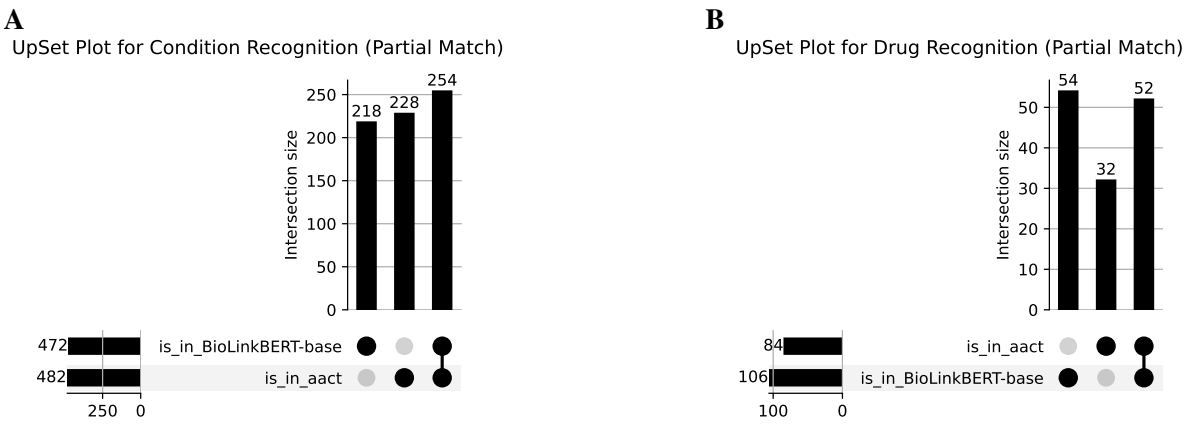

Figure 1: Named entities comparison. Horizontal bars show the total number of unique entities (sets) recognized by each method. The vertical bars indicate the size of intersections between sets. A single filled dot means the set is coming from only one of the outputs, and a connecting line indicates overlapping entities.

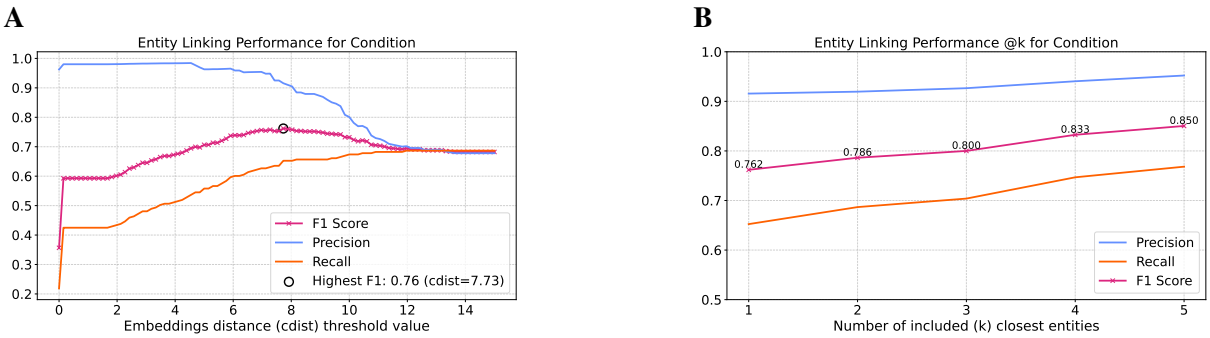

Figure 2: Entity linking experiments for condition. **A:** Impact of different Euclidean distance (cdist) threshold values. **B:** Performance change when considering a k number of closest entities.

levels. However, our experiments also uncovered two challenges. The first was the insufficient coverage of concepts in SNOMED. For example, while "Everolimus" is correctly categorized as a substance, its well-known brand name, "Afinitor", is missing, as well as the drug "Priopidine". The second challenge stemmed from the lack of contextual information when using SapBERT, which made it difficult to determine the correct target of the entities. Specifically, for disease names, it was challenging to decide whether an entity referred to a disease or a symptom/finding without the context it appeared in. For example, human annotators preferred "Depressive disorder (disorder)" for the entity "depressive", whereas SapBERT returned "Symptoms of depression (finding)".

5 Conclusion

This study evaluated the impact and limitations of NLP-based techniques for automated data extraction from clinical trial registry data. Our findings indicate that NER can retrieve entities from trial titles and summaries, potentially replacing or complementing the manually provided data in AACT. Additionally, we explored linking entities to structured representations in an ontology and standardize variations, addressing a gap in AACT.

Future work could expand upon these findings in several directions. First, we identified the need for additional ontologies or knowledge bases to address the issue of missing entities. Second, while we tested a single entity linking approach, there is a need for a more comprehensive benchmarking of different methodologies.

A promising future application would be integrating these techniques into existing trial registry platforms. This could enhance the data capturing and curation process, making it more complete, standardized and less prone to human errors, thus enhancing the usability and interoperability of the data for downstream tasks, such as monitoring and evidence synthesis.

Limitations

Our research was limited to trials in neuroscience from ClinicalTrials.gov. However, we believe that the methodologies and approaches we employed could be adapted for use with other clinical trial registry platforms and medical domains.

The estimation of optimal parameters for entity linking was exploratory and conducted on the test set of the NeuroTrialNER corpus. However, it would be more rigorous to annotate the validation corpus, optimize the parameters based on that, and then report the performance scores on the test set. Furthermore, the manual annotations for entity linking did not take into account the context in which the entity appeared. It might be necessary to refine the annotation guidelines and differentiate more clearly between target SNOMED concepts such as disorders and findings.

Finally, our research was conducted exclusively using English-language data. Expanding this work to include other languages could enrich the dataset and offer more comprehensive insights into global clinical practices.

References

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Ronald Cornet and Nicolette de Keizer. 2008. Forty years of SNOMED: a literature review. *BMC Medical Informatics and Decision Making*, 8(1):S2.

Bradford R. Hirsch, Robert M. Califf, Steven K. Cheng, Asba Tasneem, John Horton, Karen Chiswell, Kevin A. Schulman, David M. Dilts, and Amy P. Abernethy. 2013. Characteristics of Oncology Clinical Trials: Insights From a Systematic Analysis of ClinicalTrials.gov. *JAMA Internal Medicine*, 173(11):972–979.

Anna Huang et al. 2008. Similarity measures for text document clustering. In *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand*, volume 4, pages 9–56.

David Kartchner, Jennifer Deng, Shubham Lohiya, Tejasri Kopparthi, Prasanth Bathala, Daniel Domingo-Fernández, and Cassie S. Mitchell. 2023. A Comprehensive Evaluation of Biomedical Entity Linking Models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, 2023:14462–14478.

Xu Liu, Yuan Zhang, Ling-Long Tang, Quynh Thu Le, Melvin L. K. Chua, Joseph T. S. Wee, Nancy Y. Lee, Brian O'Sullivan, Anne W. M. Lee, Ying Sun, and Jun Ma. 2018. Characteristics of Radiotherapy Trials Compared With Other Oncological Clinical Trials in the Past 10 Years. *JAMA Oncology*, 4(8):1073–1079.

Iain J Marshall, Joël Kuiper, Edward Banner, and Byron C Wallace. 2017. Automating biomedical evidence synthesis: RobotReviewer. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2017, page 7. NIH Public Access.

Eugenia D. Namiot, Diana Smirnovová, Aleksandr V. Sokolov, Vladimir N. Chubarev, Vadim V. Tarasov, and Helgi B. Schiöth. 2023. The international clinical trials registry platform (ICTRP): data integrity and the trends in clinical trials, diseases, and drugs. *Frontiers in Pharmacology*, 14. Publisher: Frontiers.

Mounika Pillamarapu, Abhilash Mohan, and Gayatri Saberwal. 2019. An analysis of deficiencies in the data of interventional drug trials registered with Clinical Trials Registry - India. *Trials*, 20(1):535.

Gayatri Saberwal. 2021. The Many Uses of Data in Public Clinical Trial Registries. *Current Science*, 120(11):1686.

Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

Jiyun Shi, Zhimeng Yuan, Wenxuan Guo, Chen Ma, Jiehao Chen, and Meihui Zhang. 2023. Knowledge-graph-enabled biomedical entity linking: a survey. *World Wide Web*, 26(5):2593–2622.

Xuanyu Shi and Jian Du. 2024. Constructing a finer-grained representation of clinical trial results from ClinicalTrials.gov. *Scientific Data*, 11(1):41. Publisher: Nature Publishing Group.

Asba Tasneem, Laura Aberle, Hari Ananth, Swati Chakraborty, Karen Chiswell, Brian J McCourt, and Ricardo Pietrobon. 2012. The database for aggregate analysis of ClinicalTrials. gov (AACT) and subsequent regrouping by clinical specialty. *PloS one*, 7(3):e33677.

James Thomas, Anna Noel-Storr, Iain Marshall, Byron Wallace, Steven McDonald, Chris Mavergames, Paul Glasziou, Ian Shemilt, Anneliese Synnot, Tari Turner, et al. 2017. Living systematic reviews: 2. combining human and machine effort. *Journal of clinical epidemiology*, 91:31–37.

Tony Tse, Kevin M Fain, and Deborah A Zarin. 2018. How to avoid common problems when using ClinicalTrials.gov in research: 10 issues to consider. *Bmj*, 361.

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, et al. 2018. Clinical information extraction applications: a literature review. *Journal of biomedical informatics*, 77:34–49.

Thomas A Wood. 2023. Drug named entity recognition (computer software), version 1.0.1. To appear.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016, Dublin, Ireland. Association for Computational Linguistics.

A NER Overlap Calculation Details

The partial match similarity assessment was calculated considering both the number of matching characters and their positions within the strings to determine the closeness of the match[7]. For instance, if the AACT annotation is “hemiplegic cerebral palsy”, and the BioLinkBERT prediction is “cerebral palsy”, this qualifies as a partial match.

B Annotation Guideline for Entity Linking

Guidelines

1. Read the entity from the list of extracted NERs (e.g., column `unique_condition_target`).

2. If the entity is not clear, look up the clinical trial from which it was extracted and read the context in which it appears.

   2.1. If the linking would be possible only through the context, add this as a flag in the designated column of the annotations file (column `context_required`).

3. If it is clear what concept is represented, search for it in the SNOMED browser.

   3.1. If the concept is found:

      3.1.1. Preferably look for (disease) or (substance) main concept; e.g., for the entity "tic", prefer "Tic disorder (disorder)" instead of "Tic (finding)".

      3.1.2. Copy the concept and the concept ID into columns `target_snomed_concept` and `target_snomed_concept_id`.

      3.1.3. Always keep the semantic tag, even if the concept is of another semantic tag, e.g., (procedure) - this will be used to know which other semantic concept to include in the SNOMED graph.

      3.1.4. If more than one entity is extracted, add all the corresponding matches for linking, separated with a comma.

   3.2. If the concept is not found:

      3.2.1. Try to reduce the entity to its main components, e.g., if the entity was "post-operative atrial fibrillation" and this returns no hits from the database, look for "atrial fibrillation" only; also, if the entity is an adjective, try with the noun form, e.g., if the entity was "acromegalic", try looking for "acromegaly".

      3.2.2. Consider using a synonym.

      3.2.3. If a more generic concept is returned, then use this generic concept instead, e.g., for the entity "autoimmune neurological diseases", the resulting match is "Autoimmune disease (disorder)".

      3.2.4. If there is still no meaningful concept returned, write `n.a.` for `snomed_concept` and `snomed_concept_id`.

C Dictionary Sources for EL

For a comprehensive list of neurological and psychiatric diseases, we combined two primary sources: the International Classification of Diseases 11th Revision[8] (ICD-11) and the MeSH terms list[9]. This integration resulted in a list of 18,458 unique disease names, including synonyms and lexical variations, categorized under “Mental, behavioural or neurodevelopmental disorder” and “Neurologic Manifestations”. For drug names, we compiled data from DrugBank[10], Wikipedia, MedlinePlus, and MeSH terms[11].

D Linking Evaluation Measures

Following (Shen et al., 2015), we measured the following metrics to assess the entity linking per-

[7] We used the get_close_matches function with cutoff=0.6 from: https://docs.python.org/3/library/difflib.html

[8] https://icd.who.int/icdapi

[9] Version 2023 obtained as an XML file from https://www.nlm.nih.gov/databases/download/mesh.html

[10] https://go.drugbank.com/

[11] https://pypi.org/project/drug-named-entity-recognition/

formance.

Precision measures the accuracy of the entity linking system by evaluating the proportion of correctly linked entity mentions out of all mentions linked by the system.

$$\text{Precision} = \frac{|\text{correctly linked entity mentions}|}{|\text{linked mentions generated by system}|}$$

Recall assesses the completeness of the entity linking system by evaluating the proportion of correctly linked entity mentions out of all mentions that should have been linked.

$$\text{Recall} = \frac{|\text{correctly linked entity mentions}|}{|\text{entity mentions that should be linked}|}$$

The F1 score is combines precision and recall to provide a single score for evaluation:

$$F_1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

E Overview of NER Discrepancies

Table 1 presents various types of discrepancies observed in the extraction of condition and drug entities using BioLinkBERT and AACT. The evaluation was based on 20 randomly sampled entities for each type of disagreement, totaling 40 examples per entity category (drug and disease).

F Linking Performance for DRUG

Figure 3 A illustrates the performance of the entity linking model for drug entities at various embeddings distance (cdist) thresholds. As the cdist threshold increases, the model becomes less stringent, which impacts these metrics. The highest F1 score of 0.92 is achieved at a cdist threshold of 8.18, indicating an optimal balance between precision and recall at this point.

Figure 3 B shows the entity linking performance for drug entities as a function of the number of included closest entities (k). As the number of closest entities increases from 1 to 5, both Recall and F1 Score improve, reaching their peak at k=4 and k=5 with an F1 Score of 0.964. Precision remains high and stable throughout, indicating that including more closest entities improves recall without significantly compromising precision.

	Conditions		Drug		
Difference Type	Frequency	Example (BERT vs AACT)	Frequency	Example (BERT vs AACT)	**Comment**
Both extractions represent the same entity	17 (42%)	spine cancer vs spinal bone metastases	23 (57%)	dextrose vs dextrose 5% in water	Often the BERT extractions contained less noise and could be more easily aggregated.
Correct entity available only in AACT	7 (18%)	no annotation vs ovarian cancer	7 (18%)	migraine medications vs verapamil + paroxetine	In all cases, those entities were not available in the title of trial brief description.
Correct entity available only in BERT	8 (20%)	respiratory muscle dysfunction vs muscle weakness	6 (15%)	clozapine vs nmdae plus aifa	BERT's extractions contained more fine-grained details for conditions. Also, interventions tested together with a new intervention are annotated.
False positives AACT	1 (2%)	ketamine	1 (2%)	blood sampling	Observed entities that do not belong to the class.
False positives BERT	6 (15%)	post-, lack	2 (5%)	5 (from dextrose 5% in water)	Observed extraction errors from BERT, e.g., partial extractions of an entity (2 cases).

Table 1: Types of discrepancies for condition and drug entities extraction using BioLinkBERT and AACT. The evaluation of the results for each entity type was based on 20 randomly sampled entities for each disagreement type, i.e., 20 examples where entity extracted only by BERT and not by AACT, and 20 examples where only by AACT but not by BERT.

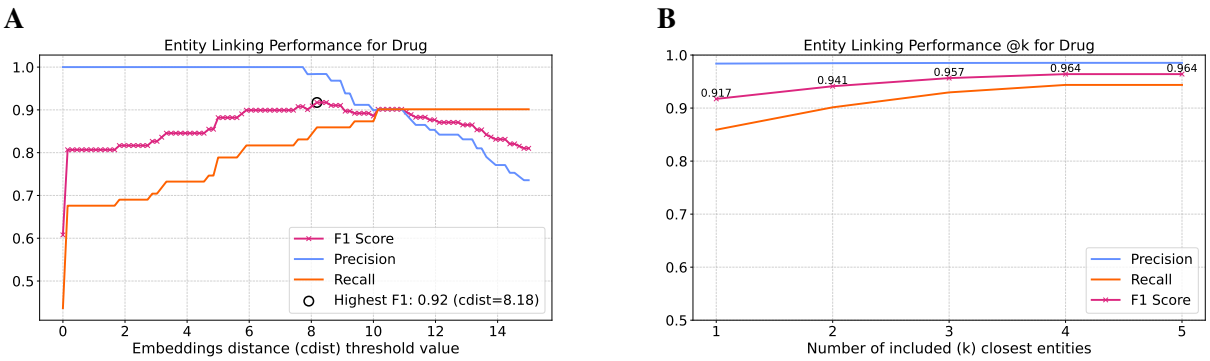

Figure 3: Entity linking experiments for drug. **A:** Impact of different Euclidean distance (cdist) threshold values. **B:** Performance change when considering a k number of closest entities.