NeuroTrialNER: An Annotated Corpus for Neurological Diseases and Therapies in Clinical Trial Registries

Anonymous ACL submission

Abstract

Extracting and aggregating information from clinical trial registries could provide invaluable insights into the drug development landscape and advance the treatment of neurologic diseases. However, achieving this at scale is hampered by the volume of available data and the lack of an annotated corpus to assist in the development of automation tools. Thus, we introduce NeuroTrialNER, a new and fully open corpus for named entity recognition (NER). It comprises 893 clinical trial summaries sourced from ClinicalTrials.gov, annotated for neurological diseases, interventions, and control treatments. We describe our data collection process and the corpus in detail. We demonstrate its utility for NER using large language models and achieve a close-to-human performance. By bridging the gap in data resources, we hope to foster the development of text processing applications that help researchers navigate clinical trials data more easily, efficiently, and comprehensively.

1 Introduction

Despite substantial investment, developing new treatments for human diseases is a challenging and often unsuccessful endeavour, especially for neurological conditions (Seyhan, 2019). For example, more than 99% of drugs tested in clinical trials for Alzheimer's disease fail (Cummings et al., 2014).

In this context, the synthesis of evidence from clinical trials is critical for researchers developing therapies, offering insights into the effectiveness and safety of interventions (Sutton et al., 2009). This process entails systematically evaluating data from clinical studies to form reliable conclusions about healthcare practices. Public clinical trial registries, such as ClinicalTrials.gov[1], are fundamental to this effort, fostering transparency and accessibility in clinical research (Laine et al., 2007).

However, extracting information from these resources is challenging due to the large volumes of data, incomplete and unstructured reporting, variability in medical terminology, and data quality concerns (Tse et al., 2018). Computational methods, in particular natural language processing (NLP), can help overcome some of these hurdles and ease the synthesis of clinical evidence (Marshall et al., 2017; Thomas et al., 2017). Named entity recognition (NER), a foundational step in NLP, enables text processing and standardization for downstream tasks like relation extraction and question answering (Wang et al., 2018). Yet, there is a scarcity of publicly available annotated corpora for clinical trial registries, hindering NLP's effectiveness in processing trial data.

Here we bridge this gap by introducing a new gold standard annotated dataset for clinical trial registry data in the neurological/psychiatric domain. The corpus comprises 893 clinical trial summaries from ClinicalTrials.gov, one of the largest international clinical trial registries (Zarin et al., 2019). It has been annotated by two to three annotators for key trial characteristics, i.e., condition (e.g., Alzheimer's disease), intervention (e.g., aspirin), and control (e.g., placebo).

We leverage this corpus to showcase its suitability for the NER task using models based on BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformers). Additionally, we compare the performance of these models against simple baseline methods and human experts to assess their effectiveness. The dataset, along with its documentation, guidelines, and code, is available on an anonymous GitHub repository[2]. After publication, it will be linked to Zenodo[3] and shared via the HuggingFace Dataset API, in compliance with the FAIR princi-

[1] https://clinicaltrials.gov/

[2] https://anonymous.4open.science/r/NeuroTrialNER-2FFC/

[3] https://zenodo.org/

ples (Wilkinson et al., 2016).

2 Related Work

A curated database for Aggregate Analysis of ClinicalTrials.gov[4] (AACT) has been released in 2011 to stimulate the accessibility of the clinical trial registry data (Tasneem et al., 2012). Disease and intervention information about each trial is available in two ways: 1) provided by the data contributors, and 2) in the form of Medical Subject Headings (MeSH) terms (Rogers, 1963) extracted by an algorithm developed by the National Library of Medicine (NLM) (Mork et al., 2013). In the first case, aggregation of the results is challenging due to substantial heterogeneity in terminology and maintenance quality across trials. In the second case, the rule-based NLM algorithm uses the MeSH ontology to infer terms, but this approach has several limitations. These include the risk of missing entities not in the ontology and lacking a clear strategy for grouping and analyzing trials in broader disease categories. Additionally, the annotation of MeSH terms lacks context sensitivity and specificity. This can result in the omission of clinically critical details of a disease, such as distinguishing between mild or severe COVID infections or between early- and late-stage cancer (Tasneem et al., 2012). In our work, we use AACT to sample clinical trials data and utilize the AACT disease and intervention labels as a baseline.

NER enables the automated identification and extraction of specific entities such as disease names (Wang et al., 2018). The main focus of existing work in NER for clinical trial data has been on PubMed abstracts. In Marshall et al. (2020), the authors extract PICO (Population, Intervention, Control, Outcome) elements from PubMed abstracts of clinical trial publications. Those entities are processed by a relation extraction module to infer which intervention was reported to work for which outcomes. The authors also utilized trial registry data from the World Health Organization International Clinical Trials Registry Platform (ICTRP)[5]. For both PubMed and ICTRP, the models were trained on the EBM-NLP dataset (Nye et al., 2018), an annotated corpus of PubMed abstracts describing clinical trials for cardiovascular diseases, cancer, and autism.

Another widely distributed dataset is the BC5CDR corpus to support the task of recognition of chemicals/diseases and mutual interactions (Li et al., 2016a). It consists of 1500 articles sampled from the CTD-Pfizer corpus, which covers a large sample of PubMed articles related to different disease classes (Davis et al., 2013).

Hence, to our knowledge, our study is the first developing a NER dataset for drugs and diseases in clinical trial registry data for neurological and psychiatric diseases. In recent years, the systems used for biomedical NER are based on deep neural networks (Song et al., 2021). Those architectures do not depend on hand-crafted rules or dictionaries and have shown a superior performance for several biomedical NLP tasks (Gu et al., 2021). Furthermore, leveraging a pre-trained biomedical language model like BioBERT requires fewer training examples as it can capture rich contextual information from existing language knowledge. We exploit such NER approaches for our dataset in section **4**.

3 The Corpus

3.1 Data Collection

A static copy of the AACT database was downloaded[6] and ingested into a local PostgreSQL database. The total number of unique clinical trials from this snapshot was 451,860.

First, we identified trials in neurological and psychiatric diseases. Since the AACT database does not provide a classification of the diseases to broader categories, we compiled a reference list of neuropsychiatric diseases. For this, we combined two sources - the International Classification of Diseases 11th Revision[7] (ICD-11) and the MeSH terms list[8]. This resulted in a list of 16,520 unique disease names (including synonyms and lexical variations) in categories such as "Mental, behavioural or neurodevelopmental disorder", and "Neurologic Manifestations". The full list with its generation code is available on our GitHub repository.

Subsequently, we used this disease list to filter the records from the AACT database, resulting in 40,842 unique trials. We further selected only the interventional trials (35,969) based on the corresponding *study type* field in the database. From this set, we randomly sampled 1,000 entries (title and

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

[5] https://www.who.int/clinical-trials-registry-platform

[6] Accessed on May 12 2023 from https://aact.ctti-clinicaltrials.org/snapshots.

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

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

trial summary) for the annotation step, from which we annotated 893.

3.2 Data Annotation

3.2.1 Annotation Guidelines

Our annotation rules were harmonized with the PICO framework (Huang et al., 2006). Within this context, the annotators were informed by the following questions:

- Disease (=Population): "Who is the group of people being studied?"
- Intervention: "What is the intervention being investigated?"
- Control: "To what is the intervention being compared?"

Furthermore, we aligned our annotation conventions for drug names with previous work (Li et al., 2016b; Krallinger et al., 2015).

We labelled the following entity types - six categories covering a broad range of common interventions (DRUG, BEHAVIOURAL, SURGICAL, RADIOTHERAPY, PHYSICAL, OTHER), one disease category (CONDITION) and one control intervention category (CONTROL).

The annotation guidelines were iteratively refined to ensure maximum clarity and optimize inter-rater agreement.

3.2.2 Annotation Process

The annotation was performed by three independent annotators - one medical doctor with > 15 years experience (BVI), one senior medical student (AEC), and a PhD candidate in the Life Sciences Graduate School (SED). There were two rounds of annotation. A first batch of 488 annotations was performed by all three annotators. 405 additional clinical trials were annotated by two annotators (BVI and SED).

The annotators used the browser-based tool Prodigy (Montani and Honnibal, 2017) to perform the manual annotation. One clinical trial example from our dataset is shown in **Figure 1**. To enhance annotation quality in case of unknown entities, the curators were encouraged to crosscheck information from reference sources such as Wikipedia, DrugBank and the ICD library.

To compile the final dataset, all conflicts were resolved by discussion. Further details about the resulting corpus can be found in section **3.4**.

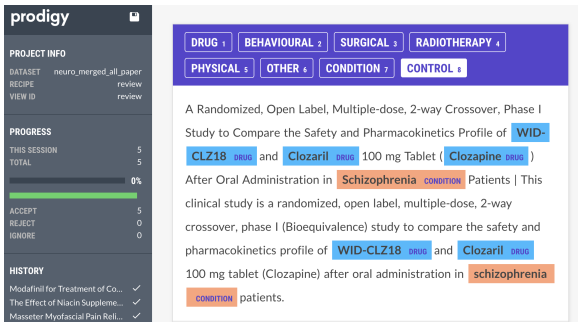

Figure 1: Annotation example shown in the annotation tool Prodigy. Blue labels indicate annotated DRUG entities and orange labels denote CONDITION entities.

	Annotation Round 1 (488 annotations)		
Annotators	Overall	DRUG	CONDITION
SED;AEC	0.77 (0.76, 0.77)	0.85 (0.83, 0.87)	0.82 (0.81, 0.83)
AEC;BVI	0.76 (0.75, 0.77)	0.85 (0.83, 0.86)	0.83 (0.82, 0.84)
SED;BVI	0.76 (0.75, 0.77)	0.86 (0.84, 0.87)	0.82 (0.81, 0.83)
	Annotation Round 2 (405 annotations)		
SED;BVI	0.79 (0.78, 0.79)	0.86 (0.84, 0.87)	0.86 (0.85, 0.87)

Table 1: Overview of inter-annotator agreement reported as the Cohen's Kappa score (95% confidence interval lower bound, upper bound).

3.2.3 Annotation Data Formats

We provide the tokenized version of the trial registry texts together with the list of corresponding annotations in BIO (Beginning, Inside or Outside of an entity span) format (Sang and Buchholz, 2000). Additionally, we give the annotated entities from each trial as a tuple consisting of (start character index, end character index, entity type, entity words) like (228, 243, 'DRUG', 'botulinum toxin').

3.3 Inter-Annotator Agreement

3.3.1 Results

Table 1 shows the pairwise inter-annotator agreement (IAA) using the Cohen's kappa statistic[9] across all entities, as well as for the separate labels DRUG and CONDITION, in the two rounds of annotation. We also report the 95% confidence intervals (Cohen, 1960).

In the first round (488 clinical trial abstracts), the overall agreement was 0.77 across all entity types, indicating a substantial IAA. The score was higher for DRUG (range 0.85-0.86) and for CONDITION (range 0.82-0.83). In the second round of annotations, the overall agreement score between BVI and SED increased slightly. The small confidence intervals suggest a high level of precision in the estimated Cohen's kappa scores.

[9]Calculated with sklearn.metrics.cohen_kappa_score.

3.3.2 Examples of Annotation Disagreements

During the preparation of the final annotated dataset, conflicts were resolved by two annotators. We observed several patterns of discrepancies:

- **Span Errors:** Discrepancies in the boundaries of annotated entities. For instance, one annotator accidentally included punctuation marks within an entity annotation. Additionally, there were differences in the included level of detail, for example BVI selected the whole expression "amnestic mild cognitive impairment", while SED only annotated "mild cognitive impairment". We settled on including "amnestic" as it was important for the diagnostic and treatment of the disease.
- **Missed Entities:** In cases involving longer texts, one annotator overlooked tagging certain entities.
- **Label Disagreement:** Cases when annotators assigned different labels to the same entity. For example, one annotator classified "IGF-1" as OTHER, while another annotator labeled it as DRUG.

Figure 2 presents the confusion matrix for each entity class between two of the annotators. The most substantial disagreements occurred between "0" (no annotation) and the classes CONDITION and OTHER. It also stands out that BVI identified 194 entities as SURGICAL, which SED had classified as OTHER, while recognizing only 12 SURGICAL entities. Note that the confusion matrix represents missed entities, but also span disagreements.

3.4 Corpus Overview

Our final annotated corpus contains 893 trial summaries/titles in total. **Table 2** describes key features of the data. In total, the corpus comprises of 11,549 unique tokens with an average number of 135 tokens per trial. The most common entities among these tokens were CONDITION (disease) with 3,998 tokens, followed by DRUG with 1,477 tokens.

Figure 3 shows the top ten most frequent annotated conditions. *Stroke* was the most prevalent term, occurring 111 times, followed by *Parkinson's disease*, *schizophrenia*, *pain*, *multiple sclerosis*, and *Alzheimer's disease*, including abbreviations thereof.

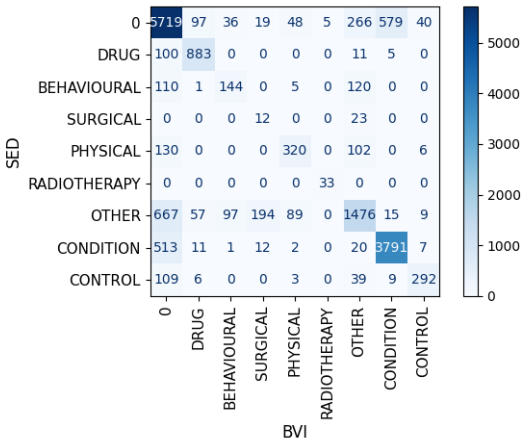

Figure 2: Confusion matrix between the labels assignments across the second annotation round of two independent annotators (SED and BVI). For readability, the number of the majority "0" class (no annotation) is shown as 10 times smaller than the actual value.

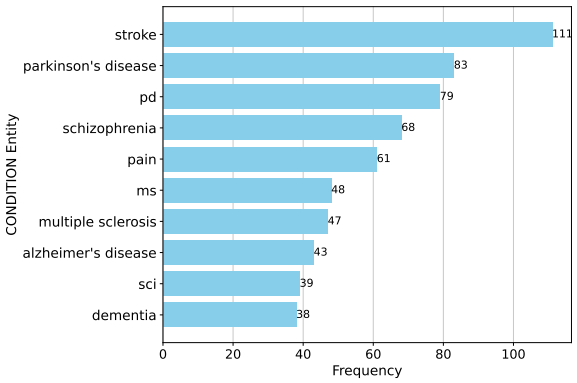

Figure 3: Top 10 most frequent CONDITION entities found in the complete dataset.

Similarly **Figure 4** presents the most frequent medication-related terms. "Aripiprazole" was the most frequently mentioned drug (n=16), followed by "dexmedetomidine" (n=14).

4 Experiments

4.1 Named Entity Recognition Methods

We considered two simple baselines. First, a dictionary lookup approach based on the developed list of neurological and psychiatric diseases (see **3.1**) and a list of drug names compiled from the DrugBank [10], Wikipedia, Medline Plus, and MeSH terms [11]. We followed the approach in Wood (2023) and annotated individual words or pairs of consecu-

[10] https://go.drugbank.com/
[11] https://pypi.org/project/drug-named-entity-recognition/

Overview	Entire Corpus	
Number of trials		893
Number of tokens per abstract (min/mean/max)	17/ 134.8/ 829	
Total number of tokens/ unique	120,383/ 11,549	
Entity Class	Count	Unique
DRUG	1477	552
OTHER	1436	846
PHYSICAL	419	264
BEHAVIOURAL	222	157
SURGICAL	98	71
RADIOTHERAPY	26	15
CONDITION	3998	1349
CONTROL	462	173

Table 2: Overview of the corpus in terms of the number of manually revised trial number, number of tokens per trial (abstract), as well as vocabulary size. The number of mentions for each annotated entity class (six intervention types, one condition and one control), total and unique count, is provided in the lower half of the table.

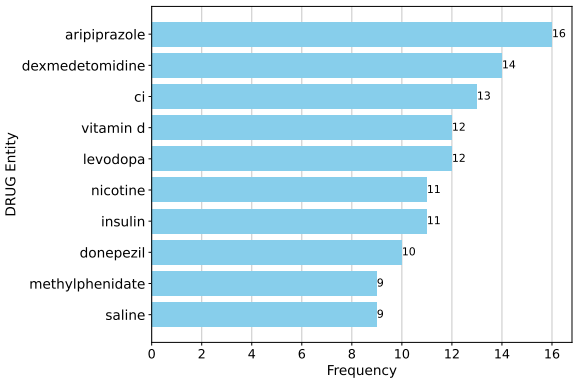

Figure 4: Top 10 most frequent DRUG entities found in the complete dataset.

tive words, that had a match in the lists. Our second baseline consisted of the *condition* and *disease* entries associated with each clinical trial from the AACT database.

For neural NER, we used three BERT-style models: BERT-base-uncased (Devlin et al., 2018), BioLinkBERT-base (Yasunaga et al., 2022), BioBER-v1.1(Lee et al., 2020), and two GPT models, gpt-3.5-turbo and gpt-4[12]. We fine-tuned BERT, BioBERT and BioLinkBERT on a single GPU in less than an hour. The latter two models have been pre-trained on biomedical domain corpora - BioBERT using PubMed abstracts and PMC full-text articles, and BioLinkBERT leveraging PubMed abstracts and citation links between PubMed articles. In contrast, BERT-base has been pre-trained on the generic BookCorpus and English Wikipedia. BioLinkBERT is notably effective in biomedical NER, ranking highly in the BLURB ranking[13]. We trained the models to classify each token as either the Beginning (B), Inside (I) or Outside (O) of an entity span (Sang and Buchholz, 2000). All BERT-based models implementations were based on the Huggingface Transformers library, using their default parameters, and Python version 3.9 (Wolf et al., 2019). We utilized the GPT models in a zero-shot setting without fine-tuning. These models excel at generating contextually relevant text for diverse tasks (Brown et al., 2020). We queried the model by sending the text of each clinical trial and asking for a list of drug and disease names. The prompt construction details are available in Appendix B.

4.2 Evaluation Setup

We evaluated the performance of the NER methods on both token and full-text level. Token-level evaluation assessed the model's performance on a per-token basis, focusing on how well it correctly labeled individual words within the text.

To evaluate the full-text level performance, we aggregated the token-level annotations to identify the unique named entities mentioned in the abstract (title and clinical trial summary). Our goal was to identify and group entities not only based on their unique textual string, but also considering semantic equivalence. For instance, we considered "MS" and "multiple sclerosis" to be equivalent. Similarly, we wanted to treat "Alzheimers" and "Alzheimers Disease" as a single entity. To address the first point, we replaced all abbreviations with their long forms using the Schwartz-Hearst algorithm (Schwartz and Hearst, 2002)[14]. To handle the cases of different spellings and synonyms, we reused the lists for diseases and drugs that we compiled for our NER baseline and mapped each synonym or spelling variation to their canonical form. By incorporating these steps, our aim was not only to enhance the evaluation process, but also to align it with a possible target application of generating descriptive statistics for unique diseases and drug names across the entire corpus.

4.2.1 Evaluation Metrics

We employed precision, recall, and F1-score calculated on the test set for the performance evaluation of the NER methods. To provide a comprehensive

[12] https://platform.openai.com/docs/models/overview

[13] https://microsoft.github.io/BLURB/leaderboard.html

[14] https://github.com/philgooch/abbreviation-extraction

NER Method	Exact	Partial
BERT-base	0.61 (0.53, 0.68)	0.63 (0.55, 0.70)
BioLinkBERT	**0.76** (0.68, 0.83)	**0.78** (0.70, 0.84)
BioBERT	0.63 (0.55, 0.70)	0.65 (0.57, 0.73)
GPT-3.5-turbo	0.26 (0.22, 0.32)	0.33 (0.27, 0.38)
GPT-4	0.45 (0.42, 0.57)	0.58 (0.50, 0.65)
AACT	0.39 (0.32, 0.47)	0.49 (0.41, 0.58)

Table 3: F1-Score (95% confidence interval lower bound, upper bound) for DRUG recognition.

NER Method	Exact	Partial
BERT-base	0.65 (0.60, 0.69)	0.69 (0.65, 0.73)
BioLinkBERT	**0.78** (0.74, 0.81)	**0.83** (0.79, 0.86)
BioBERT	0.73 (0.69, 0.77)	0.79 (0.76, 0.83)
GPT-3.5-turbo	0.40 (0.36, 0.43)	0.49 (0.45, 0.52)
GPT-4	0.49 (0.45, 0.53)	0.61 (0.57, 0.65)
AACT	0.34 (0.30, 0.39)	0.43 (0.38, 0.47)

Table 4: F1-Score (95% confidence interval lower bound, upper bound) for CONDITION recognition.

assessment, we present scores for both strict and partial matches. A "strict" match implies an exact match with the boundaries and entity type in the gold standard. A "partial" match required to have the correct entity type and a majority of words overlapping with that in the target annotations. For example, if the target is "hemiplegic cerebral palsy" and prediction "cerebral palsy", this would be a partial match since more than half of the target words are in the prediction. Confidence intervals for all evaluation metrics were calculated using the Wilson method (Wilson, 1927).

4.2.2 Data Split

To train and evaluate the methods, we randomly split the corpus into training (80%, 713 trials), validation (10%, 90 trials) and test (10%, 90 trials) sets. Overview of the number of entities in each split and their overlap is provided in Appendix A.

4.3 Results

4.3.1 Performance

Abstract Level Performance Tables **3** and **4** showcase the F1-Scores and their 95% confidence intervals for DRUG and CONDITION entity recognition tasks, respectively, comparing the different NER methods. We summarize the data for the partial match F1-Scores in Table **Figure 5**.

BioLinkBERT led in performance for the DRUG recognition task with a notable partial match F1-Score of 0.78 (CI: 0.70-0.84), outpacing BioBERT and BERT, which occupied the subsequent rankings. Despite the confidence intervals of BioBERT and BERT overlapping, BioLinkBERT's mean F1-Score surpassed BioBERT's by over 10%. GPT-3.5 trailed significantly with a partial match F1-Score of 0.33 (CI: 0.27-0.38), while GPT-4 nearly doubled its predecessor's score. Performance metrics based on AACT labels were intermediate, recording a 0.49 (CI: 0.41-0.58) for partial matches

In the CONDITION recognition task, BioLinkBERT led with an F1-Score of 0.83 (CI: 0.79-0.86), followed by BioBERT and BERT. The improvement from GPT-3.5 to GPT-4, reaching a 0.61 F1-Score, was notable but less pronounced compared to the DRUG task. Both generative models still lagged behind the fine-tuned BERT models. The AACT labels had the lowest performance for this task.

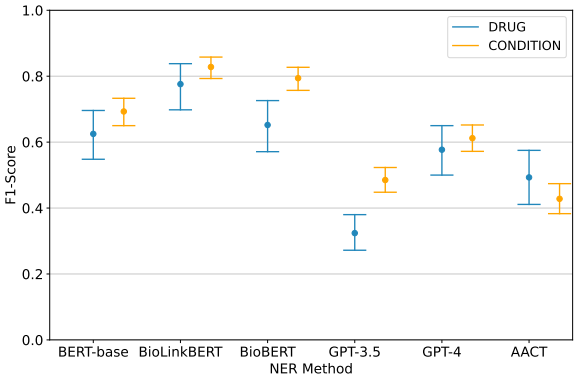

Figure 5: F1-Score (95% confidence interval lower bound, upper bound) for DRUG and CONDITION recognition on abstract level across all methods.

Entity Level Performance Table **5** provides precision (P), recall (R), and F1-Score (F1) metrics on an entity-level evaluation. The numbers are calculated using the HuggingFace seqeval implementation (Nakayama, 2018).

BioLinkBERT outperformed other methods with the highest F1-Scores for both DRUG (0.85) and CONDITION (0.83) entities, indicating a balanced precision and recall. BioBERT also demonstrated strong results, with F1-Scores close to BioLinkBERT's performance. In contrast, BERT-base's performance lagged slightly behind these domain-aware models. Dict-Lookup had the lowest performance with significantly lower F1-Scores of 0.43 for DRUG and 0.31 for CONDITION.

Furthermore, we calculated the IAA on entity level between BioLinkBERT and our target manual annotations. We reached an overall kappa score of 0.81 (0.79, 0.82), which shows that the model achieves a close to human performance.

	DRUG			CONDITION		
	P	R	F1	P	R	F1
BERT-base	0.72	0.88	0.79	0.77	0.78	0.78
BioLinkBERT	0.85	0.86	**0.85**	0.81	0.86	**0.83**
BioBERT	0.76	0.88	0.82	0.80	0.85	0.82
Dict-Lookup	0.33	0.60	0.43	0.62	0.21	0.31

Table 5: Precision (P), Recall (R) and F1-Score (F1) achieved by each method considered for entity level evaluation.

4.3.2 Impact of training data size

Figure 6 illustrates the impact of increasing training dataset size on the performance of the BioLinkBERT model after fine-tuning, measured by the validation F1-Score. The performance increased rapidly up to the utilization of 30% of the training set, after which the increase became more gradual through to 100% usage of the training set. However, even at the size of the full training set the performance has not fully reached a plateau.

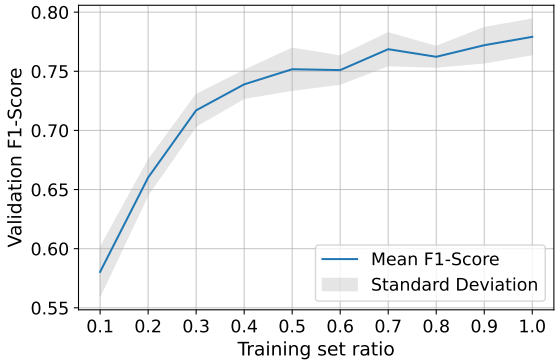

Figure 6: F1-Score on the validation data set versus training data size given as proportion of the full data set. The mean score (blue line) is calculated from 5 independent training runs. The shaded area shows the observed variation.

4.3.3 Error Analysis

Our qualitative error-analysis focused on the abstract-level errors. We consider it to be a good proxy for the errors on entity-level as it covers all unique entities found in the trial registries.

CONDITION We observed the following error patterns in BioLinkBERT's classification of CONDITION entities:

- **Over-specifying** words that we would not annotate for the disease classification, e.g., "age-related hearing loss" instead of "hearing loss"; "prolonged covid symptoms" instead of "prolonged covid".
- **Under-specifying**, e.g., "abdominal and lower limb surgeries" instead of "lower abdominal and lower limb surgeries".
- **Generic diseases symptoms** that we could consider as relevant for the study, but we did not annotate in the target as they were not specific, e.g. "loss of muscle"; "fear"; "cannot walk";
- **Study outcome-related expressions**, e.g., "ear and hearing health"; "cardio-metabolic risk".
- **Non-target disease names** that were usually mentioned to give context to the study, but were not the subject of investigation, e.g., "dyslexia"; "cerebral lesions".
- **Missed entities** as clear false negatives, e.g. "lumbosacral radiculopathy"; "immunosuppression".

There were also a few cases that should have been annotated in the target and the annotators had missed, e.g., the word "pain". BioLinkBERT had annotated those correctly, but the evaluation considered them as false positive.

BioBERT made similar qualitative errors like BioLinkBERT. Furthermore, we observed an issue related to the segmentation of words into sub-tokens for labelling. BioBERT erroneously assigned "B-LABEL" (indicating the start of a new entity) instead of "I-LABEL" (indicating continuation within an entity) to sub-tokens that should represent ongoing entities. For example in one case the word "chronic" was split into "ch" and "##ronic", and for both sub-parts the assigned labels were "B-CONDITION". This misclassification resulted in the the wrong grouping of entities, and led to more false positives.

The GPT models showed high sensitivity to the prompt formulation. Furthermore, additional post-processing was required, as the model outputs did not consistently generate the requested output list. GPT frequently extracted the trial outcome and intervention words together with the conditions, e.g. "quality of life", "functional status", "education outcomes". Also, generic terms were returned, e.g. "symptoms", "sleep".

DRUG BioLinkBERT annotated "soybean oil" and "fish oil" incorrectly as DRUG instead of the expected OTHER. Another issue was the reporting

of additional drugs in the trial summary not being tested, e.g. “Remimazolam combines the safety of midazolam and [...] of propofol.” While “remimazolam” is the target drug of the trial, the other two are only there to provide context and should not be annotated. Similar cases were observed for substances used for diagnostic purposes such as contrast agents for imaging, e.g. “gadabutrol” for MRA imaging. BioBERT missed more relevant annotations. Furthermore, again several of the errors stemmed from wrong labelling of tokenized subwords. For example the drug name “propranolol” was split and erroneously annotated as “prop” (B-DRUG), “##rano” (B-DRUG), “##lo” (I-DRUG), “##l” (I-DRUG).

GPT often returned non-drug interventions such “chamomile”, “acupuncture", and “speech therapy". There were also overall correct extractions, yet too specific according to our annotations guidelines. For example, GPT returned “diazepam nasal spray" and “diazepam rectal gel”, while we would only annotate “diazepam”.

4.4 Discussion and Limitations

BioLinkBERT and BioBERT emerged as the top-performing models for both drug and disease recognition. This was true when evaluating on entity-level, as well as the full-text aggregated target. The larger confidence intervals for drug recognition suggest that this task presents a bigger challenge for both models. However, we should note that there were also less training examples including DRUG annotations. Comparing the performance of these models with expert inter-rater agreements showed that the models achieved human like performances. The lower performance of BERT-base highlights the importance of domain-aware pre-training, as biomedical texts contain specialized terminology and complexities that generic language models might struggle to capture.

An interesting observation was the inability of the BioBERT model to recognize contiguous phrases, a limitation observed in other work as well (Chen et al., 2020). A proposed approach in (Chen et al., 2020) to mitigate this involves model architecture modification by replacing the last softmax layer with a BiLSTM+CRF layer. We did not explore this alternative extensively, as our primary focus was on the direct application of existing models to the task.

We observed that the "Dictionary-Lookup" approach fell short, particularly in recall, suggesting a propensity to miss relevant entities. This underlines the importance of leveraging more sophisticated models for the proposed entity recognition tasks.

Additionally, our study underscores the importance of prompt design in GPT models and the difficulties in eliciting specific information without comprehensive annotation guidelines. Future work may focus on improving prompts, enriching model context, and investigating few-shot training methods (Karkera et al., 2023).

We also showed that the training data size has a large impact on the model’s performance and we expect to see small improvements with more annotations.

Finally, it is important to acknowledge the assumption made in our methodology, namely, that drug and disease names are mentioned in the abstract or title of the clincal trial. Although this assumption holds in many cases, we did encounter instances where relevant information was only available in the trial’s *condition* and *intervention* AACT fields. This highlights the need for future work to address these scenarios and potentially adapt our methodology.

5 Conclusion and Outlook

We have presented NeuroTrialNER, a new, openly available corpus comprising 893 clinical trial registry abstracts annotated for diseases, interventions, and controls. We further demonstrated that the dataset was effective in training neural NER models and analyzed the performance of DRUG and CONDITION recognition. Specifically, BioLinkBERT emerged as the top-performing model with results approaching the level of a human rater. With this, our dataset has the potential to enhance our understanding of disease and drug relationships in neurological and psychiatric diseases and improve downstream tasks, such as biomedical literature summarization, ultimately improving the development of drugs to treat neurological and psychiatric diseases.

As future work, we plan on expanding the dataset with more annotated trials, other disease types, including trial outcomes, and applying the NER models to other clinical trial registries. We aim to conduct a comprehensive analysis of neurology/psychiatry clinical trial research and envision integrating our work into the services provided by the AACT database.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Miao Chen, Fang Du, Ganhui Lan, and Victor S Lobanov. 2020. Using pre-trained transformer deep learning models to identify named entities and syntactic relations for clinical protocol analysis. In *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering (1)*, pages 1–8.

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

Jeffrey L Cummings, Travis Morstorf, and Kate Zhong. 2014. Alzheimer's disease drug-development pipeline: few candidates, frequent failures. *Alzheimer's research & therapy*, 6(4):1–7.

Allan Peter Davis, Thomas C Wiegers, Phoebe M Roberts, Benjamin L King, Jean M Lay, Kelley Lennon-Hopkins, Daniela Sciaky, Robin Johnson, Heather Keating, Nigel Greene, et al. 2013. A CTD–Pfizer collaboration: manual curation of 88 000 scientific articles text mined for drug–disease and drug–phenotype interactions. *Database*, 2013:bat080.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Xiaoli Huang, Jimmy Lin, and Dina Demner-Fushman. 2006. Evaluation of PICO as a knowledge representation for clinical questions. In *AMIA annual symposium proceedings*, volume 2006, page 359. American Medical Informatics Association.

Nikitha Karkera, Sathwik Acharya, and Sucheendra K Palaniappan. 2023. Leveraging pre-trained language models for mining microbiome-disease relationships. *BMC bioinformatics*, 24(1):1–19.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M. Lowe, and et al. 2015. The CHEMDNER corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, 7:1–17.

Christine Laine, Richard Horton, Catherine D DeAngelis, Jeffrey M Drazen, Frank A Frizelle, Fiona Godlee, Charlotte Haug, Paul C Hébert, Sheldon Kotzin, Ana Marusic, et al. 2007. Clinical trial registration: looking back and moving ahead. *The Lancet*, 369(9577):1909–1911.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016a. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. 2016b. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database: The Journal of Biological Databases and Curation*, 2016:68.

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.

Iain J Marshall, Benjamin Nye, Joël Kuiper, Anna Noel-Storr, Rachel Marshall, Rory Maclean, Frank Soboczenski, Ani Nenkova, James Thomas, and Byron C Wallace. 2020. Trialstreamer: A living, automatically updated database of clinical trial reports. *Journal of the American Medical Informatics Association*, 27(12):1903–1912.

Ines Montani and Matthew Honnibal. 2017. Prodigy: A modern and scriptable annotation tool for creating training data for machine learning models.

James G Mork, Antonio Jimeno-Yepes, Alan R Aronson, et al. 2013. The NLM Medical Text Indexer System for Indexing Biomedical Literature. *BioASQ@ CLEF*, 1.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

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.

Frank B Rogers. 1963. Medical subject headings. *Bulletin of the Medical Library Association*, 51:114–116.

Erik F Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 shared task: Chunking. *arXiv preprint cs/0009008*.

Ariel S Schwartz and Marti A Hearst. 2002. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Biocomputing 2003*, pages 451–462. World Scientific.

Attila A Seyhan. 2019. Lost in translation: the valley of death across preclinical and clinical divide–identification of problems and overcoming obstacles. *Translational Medicine Communications*, 4(1):1–19.

Bosheng Song, Fen Li, Yuansheng Liu, and Xiangxiang Zeng. 2021. Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison. *Briefings in Bioinformatics*, 22(6):bbab282.

Alexander J Sutton, Nicola J Cooper, and David R Jones. 2009. Evidence synthesis as the key to more coherent and efficient research. *BMC medical research methodology*, 9(1):1–9.

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.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

Edwin B Wilson. 1927. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

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. *arXiv preprint arXiv:2203.15827*.

Deborah A Zarin, Kevin M Fain, Heather D Dobbins, Tony Tse, and Rebecca J Williams. 2019. Ten-year update on ClinicalTrials. gov Results Database. *The New England journal of medicine*, 381(20):1966.

A Data Split Details

To train and evaluate the methods, we randomly split the corpus into training (80%, 713 trials), development (10%, 90 trials) and test (10%, 90 trials) sets. **Figure 7** illustrates the intersection of unique DRUG mentions (n= 552 tokens) across the three datasets. The numbers within each set signify the count of unique DRUG mentions found only in the corresponding dataset. Additionally, we show the number of overlapping entities: 18 DRUG mentions are shared between Train and Validation, 16 between Train and Test, and 2 between Validation and Test. Seven DRUG mentions were found within all three datasets (1%). **Figure 8** presents overlap of unique diseases (n=1349 tokens). It shows that diseases show a higher overlap between different datasets with 31 mentions in all three datasets (2%).

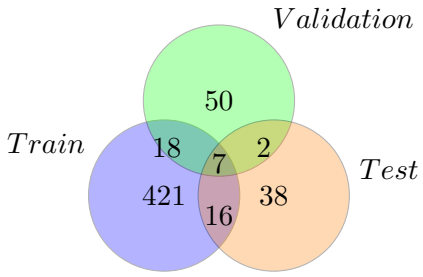

Figure 7: Overlap of unique DRUG entity mentions across datasets.

B GPT Prompting

Here we briefly describe the prompting implementation used for querying GPT. The code in Listing 1 shows the API call we used for each clinical trial. The *gpt_model* variable was replaced with the name of the GPT model, i.e., either gpt-3.5-turbo or gpt-4. The *input_raw_text* variable serves as a placeholder

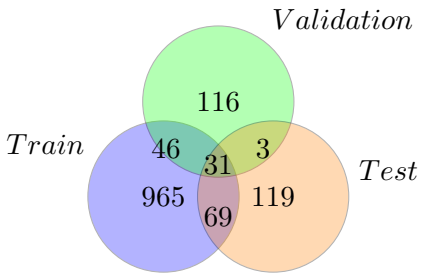

Figure 8: Overlap of unique CONDITION entity mentions across datasets.

for the actual content of the clinical trial, including both its title and detailed description. This is the text from which the GPT model is tasked with extracting relevant information based on the given prompt. The nature of the prompt varies depending on the information extraction task at hand.

```
completion =
  openai.ChatCompletion.create(
  model=gpt_model,
  temperature=0.6,
  max_tokens=2000,
  messages=[
    {"role": "system", "content":
    "You are an expert
       information
    extraction assistant from
    clinical trials."},
    {"role": "user", "content":
    prompt + "'''" +
    input_raw_text + "'''"}
  ]
)
```

Listing 1: GPT Chat Completion API Call

For the drug name extraction task, we utilize a prompt specifically designed to solicit a concise list of drug names mentioned within the clinical trial text. This is exemplified by the *interventions_prompt* variable, which reads:

```
interventions_prompt= "Extract the drug
    names from the following clinical
    trial and return them in a list
    separated with the | symbol. If none
     is found, return only the word none
    : "
```

Listing 2: DRUG Extraction Prompt

Similarly, for the disease name and symptoms extraction task, the *conditions_prompt* is tailored to extract both the diseases being investigated and any related symptoms, as demonstrated below:

```
conditions_prompt = "Extract the
    investigated disease names and
    related symptoms from the following
    clinical trial. Return them in a
    single list separated with the |
    symbol. If none is found, return
    only the word none: "
```

Listing 3: CONDITION Extraction Prompt

We also investigated a variation of the prompts that we show in Listing 4 for DRUG. However, this did not result in consistently better performance.

```
interventions_prompt_v2 = "Review the
    clinical trial document enclosed
    within triple quotes. Extract only
    the names of drugs that are actively
     being investigated in the trial.
    List these names separated by the
    '|' symbol without any additional
    text or explanation. Exclude drugs
    merely mentioned and not under
    investigation. If there are no drugs
     actively investigated, simply
    respond with 'none'. Focus solely on
     the drug names for clarity and
    precision."
```

Listing 4: DRUG Extraction Prompt v2

The final reported results were from queries executed on 29/01/2024.