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

Simona E. Doneva[1], **Tilia R. Ellendorff**[1], **Beate Sick**[1,2],
Jean-Philippe Goldman[1], **Amelia E. Cannon**[1], **Gerold Schneider**[1], **Benjamin V. Ineichen**[1]
[1] University of Zurich, Zurich, Switzerland
[2] ZHAW School of Engineering, Winterthur, Switzerland
simona.doneva@uzh.ch

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 1093 clinical trial summaries sourced from ClinicalTrials.gov, annotated for neurological diseases, therapeutic 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 tools that help researchers navigate clinical trials data more easily.

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). At the same time it has been estimated that nearly 3.40 billion people, or roughly 40% of the global population, were affected by nervous system conditions in 2021 (Steinmetz et al., 2024).

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 large data volume, incomplete and unstructured reporting, variability in terminology, and data quality concerns (Tse et al., 2018). Computational methods, in particular natural language processing (NLP), can streamline information extraction with techniques for data structuring, standardization, as well as semantic analysis, ultimately facilitating the synthesis of clinical evidence (Marshall et al., 2017; Thomas et al., 2017). Named entity recognition (NER) is one such technique that identifies and categorizes key elements in text, such as drug names, and enables downstream tasks such as relationship 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 domain of neurology/psychiatry. The corpus comprises 1093 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), therapeutic intervention (e.g., aspirin), and control arms (e.g., placebo).

We demonstrate the corpus's 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. All resources are available on

[1] https://clinicaltrials.gov/

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 18868–18890
November 12-16, 2024

GitHub[2] and the corpus is being integrated into the BigBio library of biomedical NLP datasets[3] (Fries et al., 2022).

2 Related Work

The Aggregate Analysis of ClinicalTrials.gov database (AACT)[4] was released in 2011 to enhance access to clinical trial registry data (Tasneem et al., 2012). This database provides disease and intervention information in two forms: (1) directly from data contributors, and (2) through Medical Subject Headings (MeSH) terms (Rogers, 1963) extracted using a National Library of Medicine (NLM) algorithm (Mork et al., 2013). Direct contributions vary widely in terms of terminology and data quality, making the aggregation of results challenging. The NLM's rule-based algorithm applies MeSH ontology to derive terms, yet this method has limitations, such as missing non-ontological entities and lacking a coherent strategy for classifying and analyzing trials across broad disease categories. Furthermore, MeSH term annotation often fails to capture disease context and specificity, potentially overlooking critical clinical nuances—for instance, not distinguishing between mild and severe cases of COVID or between early and late stages of cancer (Tasneem et al., 2012).

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, as well as from 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. Yet, there is no evaluation provided on how this approach performed for NER from the clinical trial registry data.

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).

Existing annotated corpora of clinical trial registries are primarily focused on the eligibility criteria sections to enhance the trial recruitment process (Deleger et al., 2012; Kang et al., 2017; Kury et al., 2020). Additionally, a dataset specifically for Spanish has been released (Campillos-Llanos et al., 2021).

To the best of our knowledge, our dataset offers several unique characteristics that distinguish it from existing resources. First, we double-annotate the titles and summary sections of prospectively registered clinical trial entries rather than published abstracts of completed trials. Second, our dataset specifically targets neurological diseases, which represent a significant portion of the global disease burden, whereas existing corpora generally focus on a broader range of medical conditions. Finally, our resource includes highly detailed annotations on aspects such as disease stages and severity, as well as a variety of intervention categories. These annotations enable more granular analysis, further enhancing its value for medical research.

3 The Corpus

3.1 Data Collection

The latest available 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

[2] `https://github.com/Ineichen-Group/NeuroTrialNER`

[3] `https://github.com/bigscience-workshop/biomedical/pull/944`

[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.

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 trial summary) for the annotation step, from which we annotated 893. In a subsequent enrichment of the corpus, in order to mitigate class imbalances, we sampled another 200 trials, which were not of "DRUG" intervention type as indicated by the corresponding AACT field.

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). Examples for each entity type can be found in **Table 2**.

The annotation guidelines were iteratively refined to ensure maximum clarity and optimize interrater agreement. The final guidelines can be found in **Appendix H**.

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 PhD Program (SED). There were three rounds of annotation. A first batch of 488 annotations was performed by all three annotators. 405 additional randomly selected clinical trials, and 200 non-drug intervention 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.

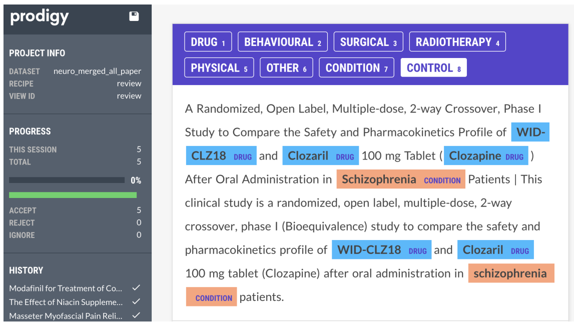

Figure 1: Annotation example shown in the annotation tool Prodigy. Blue labels indicate annotated DRUG entities and orange labels denote CONDITION entities.

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

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 (Tjong Kim 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 entity types. We also report the 95% confidence intervals (Cohen, 1960).

The overall agreement was around 0.77 across all rounds and entity types, indicating a substantial IAA. The score was highest for DRUG (range 0.83-0.87) and for CONDITION (range 0.81-0.84). The lowest agreement with most variable results was achieved for the entities BEHAVIOURAL (range 0.28-0.53) and SURGICAL (range 0.06-0.54).

[9] Calculated with sklearn.metrics.cohen_kappa_score.

Annotation Round 1 (488 annotations)									
Annotators	Overall	CONDITION	OTHER	DRUG	PHYSICAL	BEHAVIOURAL	SURGICAL	RADIOTHERAPY	CONTROL
SED;AEC	0.77 (0.76, 0.77)	0.82 (0.81, 0.83)	0.66 (0.64, 0.67)	0.85 (0.83, 0.87)	0.65 (0.61, 0.68)	0.42 (0.37, 0.48)	0.19 (0.06, 0.31)	0.91 (0.82, 1.00)	0.58 (0.53, 0.63)
AEC;BVI	0.76 (0.75, 0.77)	0.83 (0.82, 0.84)	0.63 (0.61, 0.64)	0.85 (0.83, 0.86)	0.50 (0.45, 0.54)	0.34 (0.28, 0.41)	0.46 (0.38, 0.54)	0.97 (0.91, 1.00)	0.59 (0.54, 0.64)
SED;BVI	0.76 (0.75, 0.77)	0.82 (0.81, 0.83)	0.64 (0.62, 0.65)	0.86 (0.84, 0.87)	0.60 (0.56, 0.64)	0.45 (0.39, 0.51)	0.18 (0.08, 0.28)	0.94 (0.86, 1.00)	0.68 (0.64, 0.72)
Annotation Round 2 and 3 (605 annotations)									
SED;BVI	0.77 (0.76, 0.78)	0.84 (0.84, 0.85)	0.62 (0.60, 0.63)	0.85 (0.84, 0.87)	0.64 (0.61, 0.67)	0.48 (0.44, 0.53)	0.28 (0.21, 0.35)	0.82 (0.77, 0.87)	0.68 (0.65, 0.72)

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

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 Disagreement:** Discrepancies in entity boundaries occurred, such as one annotator including punctuation marks. Additionally, there were differences in detail; for example, one annotator annotated "amnestic mild cognitive impairment" while another only annotated "mild cognitive impairment". We decided to include "amnestic" as it is important for diagnosis and treatment.
- **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. Notably, SED annotated a broader range of entities across all categories, whereas BVI more frequently classified these as "0" (no entity), suggesting a more conservative approach to annotation. Additionally, there was a notable disagreement where 30% of the instances SED categorized as BEHAVIOURAL were labeled as OTHER by BVI. Disagreements also occurred for SURGICAL and PHYSICAL, which again were annotated as OTHER by BVI, at rates of 13-15%. We further reviewed examples of discrepancies in the annotation of the SURGICAL class and observed that biological products, such as "autologous incubated macrophages" and "human placental-derived stem cells", were commonly labeled as SURGICAL by one annotator and as OTHER by the other. Since the annotation guidelines defined tissue-based therapies as part of the SURGICAL class, we determined that the correct label for these substances should be SURGICAL.

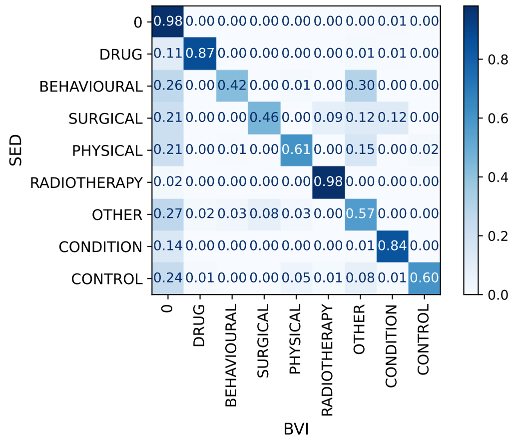

Figure 2: Confusion matrix between the labels assignments by the two independent annotators (SED and BVI). Zero (0) represents a non-entity token. For enhanced readability and comparison, the values in the matrix have been normalized by the total number of instances for each class row-wise.

3.4 Corpus Overview

Our final annotated corpus contains 1093 trial titles/trial summaries in total (referred to as abstracts, and with a unique NCTID). It comprises of 147,377 words (12,829 unique) with an average number of 135 (min: 17, max: 829) words per trial. The most frequent entities were CONDITION (disease) which is annotated 4936 times, followed by OTHER and DRUG with a count of 1806 and 1636, respectively. On the other hand, the least frequent entity class was RADIOTHERAPY, which has a count of 77, with 30 unique instances across 19 NCTIDs (see **Table 2**).

The entity classes also vary in their average character lengths. The entity class with the longest average character number is SURGICAL, averaging 26.96 characters (range: 7.83 to 46.09). In contrast, the entity class with the shortest average character number is DRUG, with an average of 11.78 characters (range: 3.20 to 20.36). **Appendix A** provides an overview of the most frequently annotated entities in each entity type across the entire corpus.

Entity Type	Count	Unique	NCTIDs	Avg. Character Number	Annotation Examples
CONDITION	4936	1612	1032	19.23 (7.11, 31.35)	"chronic inflammation", "stroke"
OTHER	1806	1047	456	25.32 (9.27, 41.37)	"air stacking", "homeopathic remedies"
DRUG	1636	601	385	11.78 (3.20, 20.36)	"empagliflozin", "guanidinoacetic acid"
PHYSICAL	594	332	144	25.29 (10.84, 39.74)	"passive exoskeleton", "resistance exercise training"
BEHAVIOURAL	317	214	86	25.47 (9.65, 41.29)	"mindfulness", "habit reversal training"
SURGICAL	173	121	45	26.96 (7.83, 46.09)	"car t cells", "nerve transfer"
RADIOTHERAPY	77	30	19	18.13 (7.29, 28.97)	"gamma knife radiosurgery", "far infrared radiation"
CONTROL	554	218	321	19.62 (7.94, 31.30)	"un-enhanced control", "conventional medical care"
Total Counts	10,093	4175	-	-	-

Table 2: Summary of entity types with total mention counts, unique instances counts, number of unique trials containing annotations for the entity type (NCTIDs), average character number, and annotation examples.

4 Experiments

4.1 Named Entity Recognition Methods

We considered two simple baselines. First, a dictionary lookup/ regex 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]. Following the approach in Wood (2023), we annotated individual words or pairs of consecutive words that matched the lists. This approach was applicable only to the DRUG and CONDITION entities. Our second baseline consisted of the *condition* and *intervention* entries associated with each clinical trial from the AACT database. To address the absence of certain intervention entity types in the database, we mapped some of the existing labels to our target labels.

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 each 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 (Tjong Kim 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). The fine-tuning setup is described in detail in **Appendix C**. GPT models are highly effective at generating contextually relevant text for various tasks (Brown et al., 2020). We used the OpenAI API to employ these models in a zero-shot setting, without any fine-tuning. For each clinical trial and entity type, we queried the model by sending the text along with a prompt requesting a list of entities. More details about the setup are available in **Appendix G**.

4.2 Evaluation Setup

Our goal was to align the evaluation with a target application for the dataset, i.e., enabling descriptive statistics for unique diseases and drug names across the entire clinical trials corpus. To achieve this, we prioritized evaluating the model's performance at the full-text level, focusing on whether it could identify relevant entities at least once, rather than evaluating its accuracy on each individual mention. For completeness, token-level results are provided in **Appendix F**.

Furthermore, we wanted to take into account semantic equivalence. While the model was trained to recognize abbreviations of named entities, such as "MS" for "multiple sclerosis", we wanted to treat those representations as the same entity. Similarly, we aimed to consolidate "Alzheimers" and "Alzheimers Disease" into a single entity. To address the first point, we replaced all abbreviations in the test dataset 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. Details on the effectiveness of this

[10]https://go.drugbank.com/
[11]https://pypi.org/project/drug-named-entity-recognition/
[12]https://platform.openai.com/docs/models/overview
[13]https://microsoft.github.io/BLURB/leaderboard.html
[14]https://github.com/philgooch/abbreviation-extraction

Entity Type	BioLinkBERT-base	BioBERT-v1.1	BERT-base-uncased	GPT-4	GPT-3.5-turbo	AACT	RegEx-Dict
CONDITION	**0.85 (0.82, 0.89)**	0.85 (0.81, 0.88)	0.71 (0.68, 0.75)	0.76 (0.72, 0.80)	0.66 (0.62, 0.70)	0.54 (0.50, 0.58)	0.50 (0.45, 0.55)
OTHER	0.62 (0.56, 0.67)	0.73 (0.67, 0.80)	0.55 (0.50, 0.60)	0.40 (0.34, 0.45)	0.33 (0.27, 0.40)	0.36 (0.29, 0.44)	n.a.
DRUG	**0.90 (0.85, 0.95)**	0.86 (0.81, 0.92)	0.74 (0.67, 0.80)	0.77 (0.71, 0.84)	0.66 (0.58, 0.74)	0.63 (0.55, 0.71)	0.34 (0.27, 0.41)
PHYSICAL	0.71 (0.64, 0.79)	**0.74 (0.66, 0.82)**	0.72 (0.65, 0.79)	0.38 (0.31, 0.45)	0.39 (0.32, 0.46)	0.10 (0.00, 0.20)	n.a.
BEHAVIOURAL	0.68 (0.60, 0.77)	**0.77 (0.69, 0.85)**	0.46 (0.34, 0.57)	0.38 (0.30, 0.46)	0.32 (0.24, 0.41)	0.27 (0.17, 0.36)	n.a.
SURGICAL	0.29 (0.12, 0.46)	**0.69 (0.57, 0.81)**	0.41 (0.25, 0.57)	0.52 (0.39, 0.65)	0.24 (0.14, 0.33)	0.00 (0.00, 0.00)	n.a.
RADIOTHERAPY	0.00 (0.00, 0.00)	**0.88 (0.70, 1.05)**	0.00 (0.00, 0.00)	0.67 (0.43, 0.90)	0.07 (0.00, 0.16)	0.35 (0.06, 0.65)	n.a.
CONTROL	**0.85 (0.78, 0.92)**	0.84 (0.77, 0.91)	0.68 (0.58, 0.77)	0.64 (0.55, 0.72)	0.49 (0.41, 0.57)	0.42 (0.30, 0.54)	n.a.
Micro F1	0.77 (0.75, 0.79)	**0.81 (0.79, 0.83)**	0.67 (0.65, 0.69)	0.56 (0.54, 0.58)	0.48 (0.46, 0.50)	0.56 (0.54, 0.58)	0.32 (0.29, 0.36)

Table 3: Partial match abstract-level F1 score (95% confidence interval lower bound, upper bound) for the NER task across all entity types. Values below zero are set to zero.

mapping can be found in **Appendix D**.

4.2.1 Evaluation Metrics

We employed precision, recall, and F1 score calculated on the test set. 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 requires the correct entity type and a significant character overlap between the predicted and target entities, assessed through a similarity ratio. This similarity assessment is calculated considering both the number of matching characters and their positions within the strings to determine the closeness of the match[15]. For instance, if the target annotation is "hemiplegic cerebral palsy", and the prediction is "cerebral palsy", this qualifies as a partial match. We also report the micro F1 score, which aggregates the contributions of entities from all classes to compute the average (treating all entities equally) (Manning et al., 2008). For all metrics we include their confidence intervals (Gildenblat, 2023).

4.2.2 Data Split

Based on the distribution of NCTIDs across our target labels, we observed limited data availability for certain classes: RADIOTHERAPY (19 trials), SURGICAL (45), BEHAVIOURAL (86), and PHYSICAL (144). To mitigate potential skewing of performance metrics due to sparse data, we implemented a two-phase custom data splitting strategy. Initially, trials containing the minority classes were allocated into training, validation, and test sets in a 50-25-25 ratio. For instance, of the 19 RADIOTHERAPY trials, 9 were randomly assigned to train, and 5 each to validation and test sets. Subsequently, the remaining trials were distributed in an 80-10-10 split. This method ensured that each label class was represented across the datasets, particularly in the test set, to provide a more accurate assessment of model performance. At the end of this process, our dataset comprised 787 trials in the training set and 153 trials each in the validation and test sets. Overview of resulting entities distribution, as well as information about unique and overlapping entities is provided in **Appendix B**.

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

4.3 Results

4.3.1 Abstract-level Partial Match Results

Table 3 and **Figure 3** show the partial match F1 scores and their 95% confidence intervals. We preferred using partial matching because it frequently accounted for minor variations and errors that do not significantly alter the meaning of the extracted entities. The exact match results and a comparison of both metrics is provided in the **Appendix E.**

BioBERT had the highest overall performance with a micro average score of 0.81 (CI: 0.79-0.83), excelling in RADIOTHERAPY 0.88 (CI: 0.70-1.05). BioLinkBERT followed with a micro average of 0.77 (CI: 0.75-0.79), performing especially well in DRUG 0.90 (CI: 0.85-0.95). When comparing the two models, it stands out that BioLinkBERT substantially under-performed for RADIOTHERAPY, SURGICAL and OTHER. For the remaining entities BioLinkBERT's performance was similar to BioBERT's, with overlapping confidence intervals. Furthermore, we calculated the IAA on token-level between BioBERT and our target manual annotations. We reached an overall kappa score of 0.82 (0.81, 0.83), which shows that the model achieves a close to human performance.

The GPT models had a weaker performance. GPT-4 scored 0.56 (CI: 0.54-0.58), doing well in CONDITION 0.76 (CI: 0.72-0.80) and DRUG 0.77 (CI: 0.71-0.84). GPT-3.5-turbo achieved an average score of 0.48 (CI: 0.46-0.50).

4.3.2 Impact of training data size

Figure 4 illustrates the impact of increasing training dataset size on the performance of the BioBERT model after fine-tuning. The reported metric is the validation micro F1 score, as computed from the

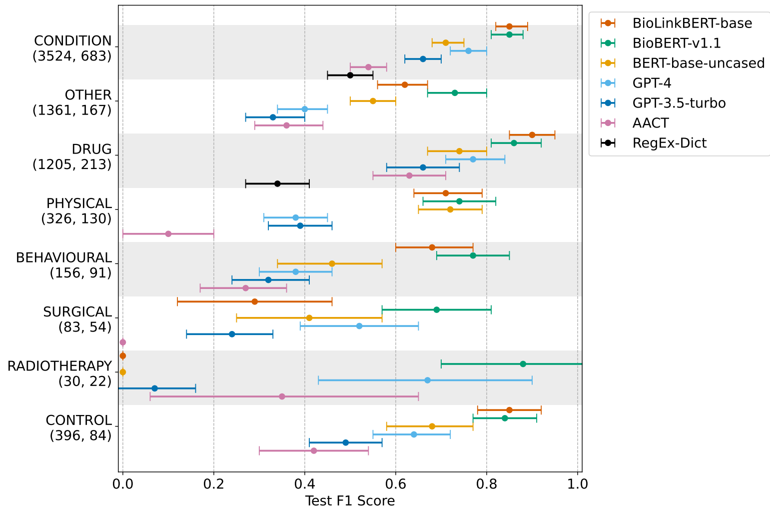

Figure 3: Partial match abstract-level F1 score (95% confidence interval lower bound, upper bound). The numbers below each entity name on the y-axis represent this entity type's frequency in the (train set, test set).

seqeval library during training (Nakayama, 2018).

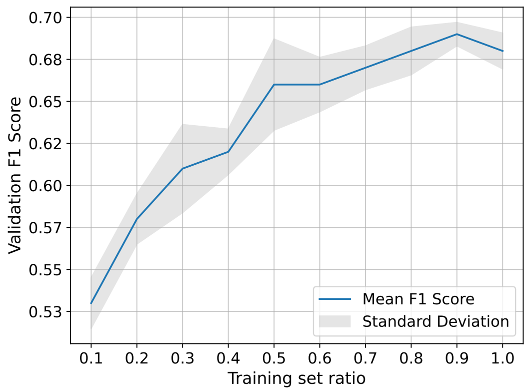

Figure 4: Micro F1 score on the validation data set versus training data size given as proportion of the full data set. The mean score (blue line) is calculated from 5 independent training runs. The shaded area shows the standard deviation.

The performance improved rapidly up to 50% utilization of the training set, after which the increase became more gradual until reaching 100% usage. A slight performance reduction at the end suggests a possible saturation point.

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 BERT-based classification:

- **Excluding relevant tokens**, e.g., "abdominal and lower limb surgeries" instead of "lower abdominal and lower limb surgeries".
- **Study outcome-related expressions**, e.g., "ear and hearing health"; "cardio-metabolic risk".
- **Non-target disease or symptom names** that were usually mentioned to give context to the study, but were not the subject of investigation or were too generic, e.g., "dyslexia"; "cerebral lesions"; "cannot walk".
- **Missed entities** include instances missed by the model, like "increased body mass index" and "immunosuppression", as well as those missed by human annotators but correctly identified by the model, such as "pain".

Furthermore, in BioBERT we noticed an issue related to the segmentation of words into sub-tokens for labelling, reported also in related work (Chen et al., 2020). 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. To address this, we used a simple strategy: taking the label of the first token of a beginning entity and merging it with subsequent sub-tokens of the same entity type. However, more sophisticated approaches recommend modifying the model architecture by replacing the last softmax layer with a BiLSTM+CRF layer (Chen et al., 2020).

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".

We also noticed instances where the model made correct annotations that the human annotators have missed. For example, BioLinkBERT annotated "agitated delirium" while the human annotator marked only "delirium".

DRUG We observed the following error patterns in the BERT-based classification:

- **Incorrect labels** annotating "soybean oil" and "fish oil" incorrectly as DRUG instead of the expected OTHER.
- **Non-target drugs**, 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.

GPT often returned non-drug interventions such as "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".

OTHER ENTITIES We observed the following error patterns in the BERT-based classification:

- **Incorrect labels**, e.g., annotating "bypass surgery" as OTHER instead of SURGERY. This error type was especially pronounced for the RADIOTHERAPY and SURGICAL entities. In many of the abstracts BioLinkBERT had correctly identified the relevant tokens, but with the incorrect label OTHER, while BioBERT had both correct.
- **Generic therapy mentions**, e.g., "therapy" instead of "meditation relaxation therapy".
- **Including irrelevant tokens**, e.g., including the word "and" or closing brackets like "cbt)".

Commonly observed error patterns from GPT models included returning the same entities for different entity types and combining interventions that should be separated. For example it extracted "onc206 in combination with radiation therapy" as a single entity for both the OTHER and RADIOTHERAPY categories. The correct annotations should have been DRUG for "onc206" and RADIOTHERAPY for "radiation therapy". Additionally, in many cases, GPT provided excessive details, such as "7 weeks of outdoor walking", instead of "outdoor walking".

4.4 Discussion

BioLinkBERT and BioBERT emerged as the top-performing models for both drug and disease recognition. An interesting observation was that BioBERT demonstrated a higher capability of learning from fewer training examples and outperformed BioLinkBERT for the minority entities SURGICAL and RADIOTHERAPY. Comparing the performance of these models with 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.

Additionally, our study highlighted the challenges in zero-shot NER with GPT models. While many results were close to our entities of interest, these models often returned unnecessary details and noise. However, we believe their output can be enhanced with more precise guidance and examples. Future work may focus on refining prompts, enriching the model context, and exploring few-shot training methods (Jimenez Gutierrez et al., 2022; Karkera et al., 2023). Furthermore, it could be beneficial to investigate the performance when all entities are returned in a single API call instead of making separate calls for each entity type.

We observed that the dictionary-lookup/ regex 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.

Finally, 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.

5 Conclusion and Outlook

We have presented NeuroTrialNER, a new, openly available corpus comprising 1093 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 their performance. Specifically, BioBERT emerged as the top-performing model with results as good as a human rater. With this, our dataset provides a fundament to enhance our understanding of disease and intervention relationships in neurological and psychiatric diseases and improve downstream tasks, such as biomedical literature summarization, ultimately improving the development of new interventions.

As future work, we plan on expanding the dataset with other disease types, including annotations for trial outcomes, and applying the NER models to other clinical trial registries or even PubMed abstracts. We are also exploring a more advanced entity normalization technique to better align the entities with a common knowledge base. Finally, we aim to conduct a comprehensive analysis of clinical trial research and envision integrating our work into the services provided by the AACT database.

Limitations

Dataset Construction. In order to select clinical trials from the neurological field, we employed a comprehensive disease terminology list, linking it to the "conditions" field of the AACT table. Despite our efforts, this method carries inherent limitations, such as potential mismatches between the terminology list and the database entries, as well as possible incomplete or inaccurate listings in the AACT "conditions" field. While we have mitigated these issues through manual validation by a medical expert, the possibility of residual inaccuracies persists. These might slightly affect the dataset's representation of certain conditions, but are unlikely to have a big impact the overall study outcomes.

The choice to utilize a random sample from the AACT database, rather than stratifying by disease, aimed to test the generalizability of our model across various conditions. Our test dataset included unique entities not seen during training, which were correctly classified, demonstrating the model's capacity to identify diseases beyond those it was explicitly trained on. This outcome suggests that a non-stratified sampling approach has the potential to highlight the robustness and adaptability of our dataset and methodology. However, it's important to note that this sampling method might not sufficiently represent less common conditions.

Finally, the random split between training and test datasets could include related trials (e.g., follow-up studies), potentially complicating the evaluation of the model's performance. However, identifying such relationships within trials is challenging due to the absence of explicit trial linkages in the database and ambiguous indicators within trial descriptions. Based on our experience with ClinicalTrials.gov, we believe that such occurrences are infrequent.

Evaluation Setup. Our custom data splitting strategy, designed to balance NCTIDs across target labels, may result in a test set that does not fully reflect the true data distribution.

A more robust evaluation method, such as cross-validation, could better capture dataset variability. However, we did not implement cross-validation due to practical constraints. Cross-validation can be computationally expensive and time-consuming. Additionally, the complexity of our custom splitting strategy and resource limitations influenced our decision to use a fixed split strategy.

It's worth noting that approximately 74% of the trials (807 out of 1093) were split using an 80-10-10 ratio. This suggests that our fixed split method may still offer a reasonable compromise between computational feasibility and model evaluation reliability.

Comparison to GPT. We acknowledge that use of GPT models in a zero-shot setting for comparison with BERT-based models, which were fine-tuned, may not constitute a fair comparison. The decision to not fine-tune the GPT models was driven by limited resources, and the limited experiments with prompting was influenced by recent research suggesting that GPT models, even with advanced prompt engineering and fine-tuning, typically underperform compared to fine-tuned BERT models in information extraction tasks such as NER (Jimenez Gutierrez et al., 2022; Ngo and Koopman, 2023; Hu et al., 2024).

Entity Availability. Our methodology primarily focused on extracting entity names from the abstract or title of clinical trial records, effectively capturing a vast majority of relevant data. However, we also identified instances where essential information was located within the AACT database's *condition* and *intervention* fields. This highlights the need for future work to address these scenarios and potentially adapt our methodology.

Acknowledgments

We thank Emma-Lotta Säätelä from the Karolinska Institute University Library for her assistance in developing the initial MeSH term list for neurological conditions, which was used to filter for relevant clinical trials.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Franz Calvo, Bryant T Karras, Richard Phillips, Ann Marie Kimball, and Fred Wolf. 2003. Diagnoses, syndromes, and diseases: a knowledge representation problem. In *AMIA annual symposium proceedings*, volume 2003, page 802. American Medical Informatics Association.

Leonardo Campillos-Llanos, Ana Valverde-Mateos, Adrián Capllonch-Carrión, and Antonio Moreno-Sandoval. 2021. A clinical trials corpus annotated with umls entities to enhance the access to evidence-based medicine. *BMC medical informatics and decision making*, 21:1–19.

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.

Louise Deleger, Qi Li, Todd Lingren, Megan Kaiser, Katalin Molnar, Laura Stoutenborough, Michal Kouril, Keith Marsolo, Imre Solti, et al. 2012. Building gold standard corpora for medical natural language processing tasks. In *AMIA Annual Symposium Proceedings*, volume 2012, page 144. American Medical Informatics Association.

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*.

Jason Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, Sunny Kang, Rosaline Su, Wojciech Kusa, Samuel Cahyawijaya, Fabio Barth, Simon Ott, Matthias Samwald, Stephen Bach, Stella Biderman, Mario Sänger, Bo Wang, Alison Callahan, Daniel León Periñán, Théo Gigant, Patrick Haller, Jenny Chim, Jose Posada, John Giorgi, Karthik Rangasai Sivaraman, Marc Pàmies, Marianna Nezhurina, Robert Martin, Michael Cullan, Moritz Freidank, Nathan Dahlberg, Shubhanshu Mishra, Shamik Bose, Nicholas Broad, Yanis Labrak, Shlok Deshmukh, Sid Kiblawi, Ayush Singh, Minh Chien Vu, Trishala Neeraj, Jonas Golde, Albert Villanova del Moral, and Benjamin Beilharz. 2022. BigBio: A framework for data-centric biomedical natural language processing. In *Advances in Neural Information Processing Systems*, volume 35, pages 25792–25806. Curran Associates, Inc.

Jacob Gildenblat. 2023. A python library for confidence intervals. `https://github.com/jacobgil/confidenceinterval`.

Yan Hu, Qingyu Chen, Jingcheng Du, Xueqing Peng, Vipina Kuttichi Keloth, Xu Zuo, Yujia Zhou, Zehan Li, Xiaoqian Jiang, Zhiyong Lu, et al. 2024. Improving large language models for clinical named entity recognition via prompt engineering. *Journal of the American Medical Informatics Association*, page ocad259.

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.

Bernal Jimenez Gutierrez, Nikolas McNeal, Clayton Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about GPT-3 in-context learning for biomedical IE? think again. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tian Kang, Shaodian Zhang, Youlan Tang, Gregory W Hruby, Alexander Rusanov, Noémie Elhadad, and Chunhua Weng. 2017. Eliie: An open-source information extraction system for clinical trial eligibility criteria. *Journal of the American Medical Informatics Association*, 24(6):1062–1071.

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.

Michael H. Kottow. 1980. A medical definition of disease. *Medical Hypotheses*, 6(2):209–213.

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.

Fabrício Kury, Alex Butler, Chi Yuan, Li-heng Fu, Yingcheng Sun, Hao Liu, Ida Sim, Simona Carini, and Chunhua Weng. 2020. Chia, a large annotated corpus of clinical trial eligibility criteria. *Scientific data*, 7(1):281.

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.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

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.

Duy-Hoa Ngo and Bevan Koopman. 2023. From free-text drug labels to structured medication terminology with bert and gpt. In *AMIA Annual Symposium Proceedings*, volume 2023, page 540. American Medical Informatics Association.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Melbourne, Australia. Association for Computational Linguistics.

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

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.

Jaimie D Steinmetz, Katrin Maria Seeher, Nicoline Schiess, Emma Nichols, Bochen Cao, Chiara Servili, Vanessa Cavallera, Ewerton Cousin, Hailey Hagins, Madeline E Moberg, et al. 2024. Global, regional, and national burden of disorders affecting the nervous system, 1990–2021: a systematic analysis for the global burden of disease study 2021. *The Lancet Neurology*, 23(4):344–381.

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.

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the CoNLL-2000 shared task chunking. In *Fourth Conference on Computational Natural Language Learning and the Second Learning Language in Logic Workshop*.

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 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. 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.

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 Corpus Details

Figure 5 outlines the top 10 annotations across the different entity categories based on frequency. In the CONDITION category, prevalent conditions like *stroke* (196 occurrences) and Parkinson's disease (130 occurrences) are featured, shedding light on major themes within the dataset. The OTHER category encompasses various treatments and techniques, with *transcranial direct current stimulation (tdcs)* and *continuous positive airway pressure (cpap)* being the most frequent. In the DRUG category, medications and treatments such as *melatonin* (19 occurrences) and *risperidone* (18 occurrences) are listed, indicating a focus on pharmacological interventions. The PHYSICAL category outlines physical and rehabilitative therapies, with *exercise* being the most present (41 occurrences). BEHAVIOURAL shows therapeutic approaches such as *cognitive-behavioral therapy (cbt)* and *action observation therapy*, with frequencies ranging from 9 to 4. SURGICAL presents various surgical methods, with *car t cells* and *carotid endarterectomy* among the top, showcasing specialized medical interventions. RADIOTHERAPY covers radiation-based treatments, with *radiation therapy* having the highest frequency (12 occurrences). Lastly, CONTROL describes control conditions in experiments, with *placebo* (217 occurrences) leading, underscoring its common use in controlled studies.

B Data Split Details

Table 4 displays the frequency and uniqueness of the different entity types across training, validation, and testing datasets.

CONDITION, OTHER, and DRUG are the most frequently annotated entity types, with relatively moderate novelty in the test data; CONDITION features 25% (171/683) unique entities and DRUG has 36% (77/213). It also stands out that while OTHER is the second most frequently annotated entity, around 62% (103/167) of the test entities are unique for the test set. This is due to the nature of this label - it captures anything that does not fit in the other categories.

On the other hand, PHYSICAL and BEHAVIOURAL have fewer annotations but exhibit higher novelty, with 46% (60/130) and 60% (55/91) of their test entities being unique, respectively. At the lower end, SURGICAL and RADIOTHERAPY have the fewest annotations but also a substantial portion of novel entities in the test datasets, 69% (37/54) and 23% (5/22) respectively. This configuration underscores different challenges for predictive models, ranging from handling familiar entities to adapting to largely unseen ones in testing.

C Fine-Tuning Setup

The datasets used for training, validation, and testing were loaded from JSON files. The tokenization process utilized the HuggingFace `AutoTokenizer` and ensured that the tokens are aligned with their corresponding labels. Padding, truncation, and a maximum sequence length of 512 tokens was applied to ensure consistent input sizes. The labels were mapped to integer IDs using a dictionary, where each unique label in the dataset was assigned

Entity Type	Train Total	Train Unique	Valid Total	Valid Unique	Test Total	Test Unique	Train ∩ Valid	Train ∩ Test	Test ∩ Valid	Train ∩ Valid ∩ Test
CONDITION	3524	1068	729	191	683	171	123	110	63	57
OTHER	1361	749	278	164	167	103	17	18	10	7
DRUG	1205	415	218	62	213	77	25	26	8	6
PHYSICAL	326	191	138	63	130	60	13	4	5	2
BEHAVIOURAL	156	105	70	48	91	55	4	3	1	1
SURGICAL	83	58	36	24	54	37	1	1	0	0
RADIOTHERAPY	30	13	25	7	22	5	3	4	4	3
CONTROL	396	138	74	37	84	31	7	10	5	5

Table 4: "Train Total", "Valid Total", and "Test Total" represent total entity counts in the training, validation, and test datasets, respectively. "Train Unique", "Valid Unique", and "Test Unique" indicate unique entity counts in these datasets. "Train ∩ Valid", "Train ∩ Test", and "Test ∩ Valid" denote entity overlaps between training-validation, training-test, and test-validation sets, respectively. "Train ∩ Valid ∩ Test" shows entities common to all three datasets.

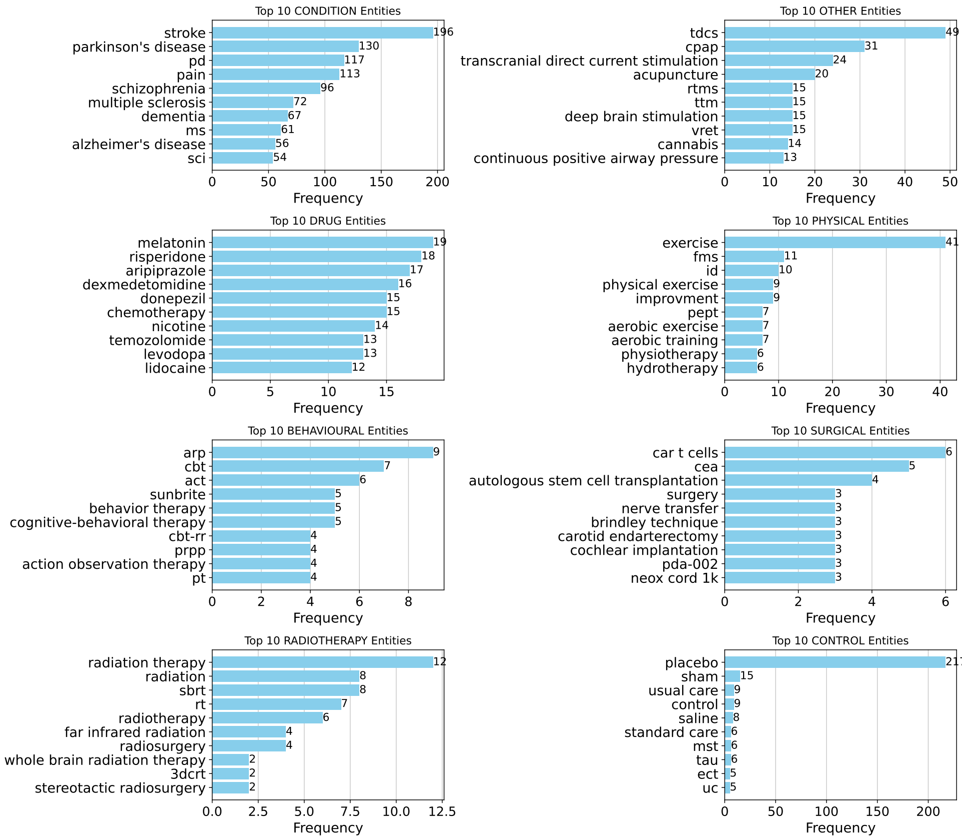

Figure 5: Top 10 most frequent annotated entities per entity type in the complete dataset.

a unique index.

The model architecture was based on the pre-trained `AutoModelForTokenClassification` from HuggingFace, initialized from the different model's checkpoint. The configuration was adjusted to match the number of labels in the dataset. Label-to-ID and ID-to-label mappings were provided during the initialization to ensure correct classification during training and evaluation.

The `TrainingArguments` class from the HuggingFace Transformers library was utilized to configure the training process. Below, we detail the key parameters used in the experiments[16]:

- **Training Epochs:** The model was trained for a total of 15 epochs.
- **Batch Size:** A batch size of 16 was used for training on each device, and a batch size of 64 was used during evaluation.
- **Warmup Steps:** A warmup ratio of 10% of the total training steps was applied to gradu-

[16]The code is available at https://github.com/Ineichen-Group/NeuroTrialNER/blob/main/models/bert/train_script.py

ally increase the learning rate at the beginning of training.

- **Weight Decay:** A weight decay of 0.01 was employed to regularize the model and prevent overfitting.

- **Logging and Evaluation:** Logging, evaluation, and model saving were configured to occur at the end of every epoch.

- **Model Selection:** The best model was selected based on the evaluation loss (`eval_loss`), with lower values indicating better performance.

- **Checkpointing:** To limit the storage space used by checkpoints, a maximum of two checkpoints were saved. The best checkpoint was always retained, and older checkpoints were deleted when new ones were created.

- **Reporting:** The training process was tracked and reported using the Weights and Biases (`wandb`) platform.

D Entity Mapping Details

As described in **Section 4.2** we used a basic mapping technique to link entities recognized by different NER models to their canonical forms in a target dictionary. Here we present a brief evaluation about how well this technique performed. Table 5 details the results of applying our mapping technique to the aggregated unique abstract-level entities, obtained from the various NER methods.

The RegEx-Dict method, employing regular expression-based dictionary matching, shows a 100% success rate in mapping both DRUGs and CONDITIONs. This perfect mapping is attributable to the source of these annotations, which are derived directly from the same dictionaries used for mapping.

The results further revealed that, generally, DRUG entities were mapped more successfully to the dictionary compared to CONDITION entities. This disparity could be due to the inclusion of additional information related to CONDITIONS, such as severity and stage, in the manual and therefore fine-tuned model extractions. These detailed attributes make CONDITION entities more complex and harder to map accurately to the dictionary. In contrast, the AACT database typically contains high-level condition descriptions that exclude such detailed attributes, resulting in higher mapping success (61.5%) as these broader terms align better with the dictionary entries.

For DRUG entities, the highest number of successful mappings was produced by entities identified using BioLinkBERT-base (49.1%), followed closely by the GPT models, with GPT-4 mapping 56 out of 120 processed entities (46.7%) and GPT-3.5-turbo mapping 44 out of 99 (44.4%). Interestingly, the AACT DRUG entities were mapped in only 35.8% of cases.

The results suggest that a more advanced neural linking approach would be better for entity linking.

E Abstract-level Exact Match Results

Table 6 presents the F1 scores calculated based on the exact match between target and predicted annotations. The comparative performance of the different models remained consistent: BioLinkBERT led in DRUG and CONDITION categories, while BioBERT outperformed in all other entity types. Notably, there was a drop in performance for the minority classes: PHYSICAL, BEHAVIOURAL, SURGICAL, and RADIOTHERAPY.

Table 7 helps interpret the differences between partial and exact matches taking BioBERT as a reference model. It provides the target and predicted named entities from three randomly sampled trials per entity type where the exact F1 score was lower than the partial F1 score. The total number of trials exhibiting this discrepancy is also reported below each entity type.

We can see that the partial match metric allowed for flexibility in the span of extracted entities, such as ignoring additional terms in “aerobic dance training practice” or minor variations like the suffix in “seizure rms”. It also disregarded unnecessary characters added by the model, exemplified by the erroneus bracket in “meditation-relaxation)”.

However, there were instances where model extractions missed parts of a word, such as extracting "pre gait training" instead of "precision gait training." This issue was particularly relevant for the CONTROL category, where the frequent entity "placebo" was often reduced to "place." Additionally there were cases where missing a part of the entity changes the semantic meaning, e.g., extracting only "cannabis" from "cannabis misuse" did not capture the actual condition. In these cases, the partial match metric was more forgiving, potentially obscuring some limitations of the model.

Source Annotations	Annotated Drug	Matched Drug	% Mapped Drug	Annotated Condition	Matched Condition	% Mapped Condition
Manual Target Annotations	100	52	52.0	345	120	34.8
BioLinkBERT-base	112	55	49.1	424	131	30.9
BioBERT-v1.1	121	50	41.3	433	127	29.3
BERT-base-uncased	123	41	33.3	549	125	22.8
GPT-3.5-turbo	99	44	44.4	488	111	22.8
GPT-4	120	56	46.7	268	128	47.8
AACT	81	29	35.8	405	249	61.5
RegEx-Dict	189	189	100.0	126	126	100.0

Table 5: Mapping of abstract level entities to a canonical in a target dictionary. Each row in the table quantifies the total number of entities identified by the different NER methods (Annotated Drug and Annotated Condition) and the number that were accurately mapped (Matched Drug and Matched Condition) along with their respective percentages.

This type of evaluation highlights the trade-offs between partial and exact matching approaches. Partial matching can be advantageous for handling variations and minor errors in entity extraction, offering a more lenient and potentially more informative measure of model performance. However, it can also mask inaccuracies and semantic differences that exact matching would capture.

F Token-level Results

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. Table **8** presents the results of token-level evaluation for micro F1 score across different entity types. Since the GPT models and the AACT database did not provide token-level annotations, we only provide the scores achieved by the BERT-based models.

BioLinkBERT-base achieved an average F1 score of 0.94. BioBERT-v1.1 showed a slightly higher performance with an average F1 score of 0.95. On the other hand, BERT-base-uncased performed slightly lower with an average F1 score of 0.93.

Notably, BioLinkBERT-base and BioBERT-v1.1 generally exhibited higher performance across most entity types compared to BERT-base-uncased. However, there were variations in performance across different entity types. For instance, BioBERT-v1.1 outperformed other models in RADIOTHERAPY (F1 score of 0.93) and SURGICAL (F1 score of 0.74) categories, while BERT-base-uncased struggled particularly in BEHAVIOURAL (F1 score of 0.36) and SURGICAL (F1 score of 0.30) categories.

G GPT Setup

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

```
completion =
  client.chat.completions.create(
  model=gpt_model,
  temperature=0.1,
  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

We also explored a suggested approach to prevent GPT from generating tokens that are not in the original input text (Jimenez Gutierrez et al., 2022). Specifically, by employing *logit bias*[17], we could add a fixed value to the final probability of a specified set of tokens, thereby constraining the

[17] https://platform.openai.com/docs/api-reference/completions

Entity Type	BioLinkBERT-base	BioBERT-v1.1	BERT-base-uncased	GPT-4	GPT-3.5-turbo	AACT	RegEx-Dict
CONDITION	**0.77 (0.73, 0.81)**	0.72 (0.68, 0.76)	0.61 (0.57, 0.64)	0.58 (0.53, 0.63)	0.50 (0.45, 0.55)	0.31 (0.26, 0.35)	0.35 (0.29, 0.41)
OTHER	0.39 (0.33, 0.46)	**0.47 (0.40, 0.55)**	0.28 (0.21, 0.34)	0.15 (0.09, 0.20)	0.09 (0.04, 0.14)	0.05 (0.01, 0.10)	n.a.
DRUG	**0.83 (0.77, 0.89)**	0.73 (0.66, 0.80)	0.54 (0.46, 0.61)	0.67 (0.60, 0.75)	0.58 (0.50, 0.66)	0.46 (0.37, 0.55)	0.30 (0.23, 0.37)
PHYSICAL	0.41 (0.31, 0.50)	**0.45 (0.35, 0.55)**	0.41 (0.32, 0.50)	0.14 (0.07, 0.20)	0.11 (0.05, 0.17)	0.03 (0.00, 0.08)	n.a.
BEHAVIOURAL	0.32 (0.21, 0.42)	**0.50 (0.38, 0.61)**	0.22 (0.11, 0.34)	0.07 (0.01, 0.13)	0.04 (0.00, 0.09)	0.02 (0.00, 0.05)	n.a.
SURGICAL	0.09 (0.00, 0.22)	**0.44 (0.29, 0.59)**	0.08 (0.00, 0.19)	0.09 (0.00, 0.20)	0.11 (0.03, 0.19)	0.00 (0.00, 0.00)	n.a.
RADIOTHERAPY	0.00 (0.00, 0.00)	**0.80 (0.58, 1.02)**	0.00 (0.00, 0.00)	0.13 (0.00, 0.37)	0.05 (0.00, 0.12)	0.13 (0.00, 0.37)	n.a.
CONTROL	**0.69 (0.59, 0.78)**	0.58 (0.49, 0.68)	0.05 (0.00, 0.12)	0.40 (0.30, 0.50)	0.22 (0.14, 0.30)	0.30 (0.18, 0.43)	n.a.
Micro F1	0.66 (0.64, 0.68)	**0.68 (0.66, 0.70)**	0.54 (0.52, 0.56)	0.42 (0.40, 0.44)	0.37 (0.35, 0.39)	0.45 (0.43, 0.47)	0.25 (0.21, 0.28)

Table 6: Exact-match F1 score (95% confidence interval lower bound, upper bound) for the NER task across all entity types.

tokens that GPT can generate. However, we observed a substantial amount of new noise in the outputs, and due to time constraints, we did not further investigate this approach. Instead we defined some post-processing rules based on the observed outputs as described later.

Prompting Strategy Only briefly we experimented with a simpler (v1) and more sophisticated (v2) prompt formulations for the DRUG (**Listing 2**) and CONDITION (**Listing 3**) entities. Curiously, we observed that the simpler prompt versions for both entity types resulted in better results for GPT-4. For GPT-3 the opposite was true, and the outputs produced using the more complex prompts seemed to be better. We leave a more systematic evaluation of the prompt strategies and their impact to future research.

```
interventions_prompt_v1 = "List the drug
     names mentioned in the following
    sentences separated with the |
    symbol. If none is found, return
    only the word none.: "

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 2: DRUG Extraction Prompts

```
conditions_prompt_v1 = "List the
    diseases mentioned in the following
    sentences separated with the |
    symbol. If none is found, return
    only the word none.: "

conditions_prompt_v2 = "Examine the
    clinical trial document within the
    triple quotes. Identify and list
    only the names of diseases and
    related symptoms under investigation
    . Format this list with each name or
     symptom separated by the '|' symbol
    , omitting any additional
    descriptions or text. Exclude
    diseases and symptoms that are only
    mentioned but not investigated. If
    there are no diseases or symptoms
    actively investigated, answer with '
    none'. The response should strictly
    contain the list of names and
    symptoms."
```

Listing 3: CONDITION Extraction Prompts

The prompt strategies for PHYSICAL, BEHAVIOURAL, SURGICAL, RADIOTHERAPY, CONTROL entities followed the same template as illustrated in **Listing 4**. In each case, only the relevant portion highlighted in orange was utilized from the prompt template.

```
prompt_template = "Extract the therapeutic
    physcial | therapeutic behavioural |
    surgical | radiotherap |
    comparator interventions 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 4: Different Entities Prompt

Finally, for the OTHER category, we instructed GPT to identify interventions that didn't fit into any other predefined category, see **Listing 5**.

```
prompt_other = "Extract any other
    therapeutic interventions from the
    following clinical trial, which are
    not behavioural, surgical,
    radiotherapy or physical. Return
    them in a list separated with the |
    symbol. If none is found, return
    only the word none."
```

Listing 5: Different Entities Prompt

Post-processing Our post-processing rules were developed based on observation of the model's outputs. These rules guided the following steps:

Entity Type (Diff Cases)	Target Entities	Predicted Entities	Exact F1	Partial F1
CONDITION (40)	emergent seizure, seizure	emergent seizure rm, seizurel seizure rm, seizure rms	0.33	1.00
	drug abuse, spm	drug abuse, drug use, dual disordered, spmi, substance abuse	0.57	0.75
	cannabis misuse, misuse cannabis, schizophrenia	cannabis, schizophrenia	0.40	1.00
OTHER (21)	electromagnetic tracking, electromagnetic tracking system	electromagnetic tracking tracking	0.00	1.00
	imaginal exposure sessions, imaginal exposure therapy, online format of ie	imaginal exposure, imaginal exposure therapy, online format of	0.33	1.00
	environmental enrichment online spatial navigation	online spatial navigation intervention remotely delivered environmental enrichment intervention	0.00	1.00
DRUG (13)	pasireotide, somatostatin analogues	pasireotide, pasireotide lar, somatostatin analogue	0.40	1.00
	lanreotide, octreotide	lanreotide autogel, lanreotidegel, octreotide	0.40	1.00
	lithium, lurasidone, lurasidone hcl	lithium, lurasidone	0.80	1.00
PHYSICAL (16)	inspiratory muscle strengthening exercise, inspiratory muscle training	inspiratory muscle strengthening exercise, inspiratory muscle training care	0.50	1.00
	aerobic dance training, aerobic dance training with home practice	aerobic dance training, aerobic dance training practice, physical exercise	0.40	0.86
	precision gait retraining	pre gait retraining	0.00	1.00
BEHAVIOURAL (9)	brief talking therapy	brief intervention, talking therapy	0.00	0.80
	meditation relaxation therapy, meditation-relaxation, mr therapy	meditation-relaxation (, meditation relaxation therapy, mr therapy	0.67	1.00
	prevention prompts tailored to familial risk, tools for health promotion and disease prevention	familial risk assessment and prevention prompts tailored to familial risk	0.00	0.80
SURGICAL (4)	femoral derotation osteotomies, femoral derotation osteotomy	femoral derotation osteotomy, transversal plane femoral derotation osteotomies tracking	0.50	1.00
	biostar septal repair implant, patent foramen ovale closure, pfo closure	biostar septal repair implant, biostar septal repair implant system, patent foramen ovale closure, pfo closure	0.85	1.00
	(autologous) stem cells, stem cell transplant, syngeneic or autologous hematopoietic cell transplantation	stem cell transplant, stem cell transplant (autologous) stem cells, syngeneic or autologous hematopoietic cell transplantation	0.67	1.00
RADIOTHERAPY (1)	3d conformal palliative rt, 3d conformal radiotherapy, 3d crt, radiotherapy, stereotactic body radiotherapy	3d conformal palliative rt, 3d conformal radiotherapy, 3d crt, stereotactic body radiotherapy	0.88	1.00
CONTROL (14)	placebo	place	0.00	1.00
	standard of care	standard of care method, standard of care techniques	0.00	1.00
	the usual post-transplant care, usual care	usual post-liver transplant care, usual post-transplant care	0.00	0.85

Table 7: Examples of cases for BioBERT where where the exact F1 score was lower than the partial score. Below each entity type the number of trials where this was true is presented. The "Target Entities" column contains the unique manual annotations, while the "Predicted Entities" are the annotations obtained from the model.

1. Replacement with 'none': Certain phrases like "not mentioned," "interventions: none," or variations were replaced with "none" to indicate absence of information.

2. Removal between specific phrases: Remove

Entity Type	BioLinkBERT-base	BioBERT-v1.1	BERT-base-uncased
CONDITION	**0.89 (0.88, 0.9)**	0.88 (0.87, 0.89)	0.85 (0.83, 0.86)
OTHER	0.59 (0.56, 0.62)	**0.66 (0.62, 0.69)**	0.52 (0.49, 0.56)
DRUG	**0.90 (0.88, 0.93)**	0.85 (0.82, 0.88)	0.85 (0.81, 0.88)
PHYSICAL	0.70 (0.66, 0.73)	**0.77 (0.74, 0.8)**	0.69 (0.65, 0.72)
BEHAVIOURAL	0.64 (0.59, 0.69)	**0.72 (0.67, 0.76)**	0.36 (0.30, 0.43)
SURGICAL	0.31 (0.24, 0.39)	**0.74 (0.69, 0.79)**	0.30 (0.22, 0.37)
RADIOTHERAPY	0.00 (0.00, 0.00)	**0.93 (0.87, 0.99)**	0.00 (0.00, 0.00)
CONTROL	**0.79 (0.75, 0.84)**	0.75 (0.71, 0.8)	0.33 (0.25, 0.41)
Micro F1	0.94 (0.94, 0.95)	**0.95 (0.95, 0.95)**	0.93 (0.92, 0.93)

Table 8: Token-level evaluation F1 score (95% confidence interval lower bound, upper bound) for all entity types.

text between specific phrases, such as between "The" and "are," "The" and "are as follows:", "Therefore" and "is:", "The therapeutic intervention" and "is:", and "not" and "is:".

3. Cleaning text: Various cleaning operations were applied, such as removing newlines, hyphens, redundant spaces, periods, and quotes.

These steps collectively aimed to enhance the coherence of the GPT-generated text.

H Annotation Guidelines

H.1 General Guidelines

1. The curators are encouraged to crosscheck information from reference sources such as Wikipedia, and chemical databases (ChEBI, DrugBank, etc.) to facilitate the annotation process and ensure compliance with the guidelines.

2. Do not tag unclear cases. If the annotator is not sure about a given mention, even after consulting some external sources, the corresponding mention should remain unlabelled.

3. Mentions should be annotated considering the context in which they are used and only if fulfill the definitions for Condition and Intervention described in later chapters. E.g. While the word *Immunotherapy* is a valid Intervention in some cases, it is not to be annotated in the sentence "The Efficacy and Safety of the United Allergy Service (UAS) Immunotherapy Protocol", as it has a different semantics in this context. If the text mentions the same intervention/condition in another context, e.g. existing research such as animal studies, it should be annotated. Example of the latter is the text: "Different Efficacy Between Rehabilitation Therapy and Umbilical Cord Derived Mesenchymal Stem Cells Transplantation in Patients With Chronic Spinal Cord Injury in China | [...] However, it can not repair the damaged nerve function. Studies show that mesenchymal stem cell transplantation can remarkably improve the neurological function of SCI in animals without any severe side effect." Here the tokens "mesenchymal stem cell transplantation" and "SCI" should be labeled in the last sentence.

4. Conditions are more reliably maintained in AACT than Interventions. Therefore we have a more broad inclusion criteria for Interventions than Conditions, which need to be more specific to be annotated. If there is an overlap in the phrase, we prefer annotating for the intervention rather than the condition, e.g. in "Clinical Assessment of Perfusion Techniques During Surgical Repair of Coarctation of Aorta With Aortic Arch Hypoplasia in Infants" the phrase "Surgical Repair of Coarctation of Aorta With Aortic Arch Hypoplasia" should be annotated as INTERVENTION.

5. Conditions and Interventions should be annotated only if they appear in relation to the target study population or intervention. E.g. in "Pain is a common symptom of Multiple Sclerosis. In the present study we assess whether aspirin relieves headache." the words "Pain" and "Multiple Sclerosis" should not be annotated, while "aspirin" (DRUG) and "headache" (CONDITION) should be annotated.

6. Interventions or Conditions mentioned within the context of the study name, should not be annotated. E.g. "Nova Scotia Chronic Pain

Collaborative Care Network: A Pilot Study" should result in no annotations.

7. If there are multiple CONDITION or INTERVENTION mentioned which are separated with "versus", "vs", "and", "or", "/" or similar, annotate preferably as separate entities. A positive example is "Rehabilitation program by rhythmic auditory cueing" - here "Rehabilitation program" and "rhythmic auditory cueing" should be annotated separately. However, if the words can't stand by themselves, the whole phrase should be annotated as one entity. E.g. "Moderate and Severe Dementia", "early versus standard AR therapy" should be annotated together. In "Multimodal Opiate-sparing Analgesia Versus Traditional Opiate Based Analgesia", the two INTERVENTIONs can be clearly separated in two entities: "Multimodal Opiate-sparing Analgesia" and "Traditional Opiate Based Analgesia".

8. If possible, the labeled word string should not be a combination of terms with and without brakets. E.g. "oral appliance (OA) device" should result in two labeled words "oral appliance" and "OA".

9. Typing errors or formatting errors should be labelled, unless they have impact on the tokenization provided by Prodigy and would result in wrong entity span.

H.2 Condition Mention Annotation

Our working definition for a **Condition** is any "state labeled as diseases by virtue of consensus on prevalent sociocultural and medical values". It has to have "clearly identifiable diagnostic features and disease progression, and response to specific treatment." (Calvo et al., 2003) In contrast, we do not label the symptomatic manifestation of a disease, that is the "self-conscious sensation of dysfunction and/or distress that is felt to be limitless, menacing and aid-requiring." (Kottow, 1980)

Whenever possible we will follow closely the annotations presented in (Li et al., 2016b).

What to annotate?

1. As a general guideline, annotated should be conditions that have an ICD-11 code [18].

[18] https://icd.who.int/browse11/l-m/en

2. We annotate conditions even in the absences of an intervention or if a diagnostic/explorative method was investigated in the trial.

3. Further defining characteristics should be included: Acute/Chronic; Active/Inactive; Mild/Moderate/Severe; End Stage/Early Stage; Drug-resistant; Total/Partial; Intermittent/Relapsing and others. Similarly, "Post-stroke" should be annotated instead of only "stroke" because it refers to the phase after the acute stroke. This includes genotypes further specifying diseases, e.g. "GBA-associated Parkinson's Disease."

4. Annotate deficiencies of one or more essential vitamins, e.g. "Vitamin B deficiency", "Zinc deficiency".

5. Annotate words like "pain" and "cognitive dysfunction", only if is a clear target for the intervention. It should not be annotated if its role is an OUTCOME, e.g. In the case of "Test if [...] offer a better pain relief.", the word "pain" should not be annotated.

6. Compound strings like "PwMS" (Person with Multiple Sclerosis) should not be annotated.

7. Symptoms should be annotated only if they are a clear target of the Intervention, e.g. in "depressive symptoms after stroke" both "depressive symptoms" and "stroke" should be annotated separately.

8. Annotate the most specific disease mentions. For instance, the complete phrase "partial seizures" should be preferred over "seizures" as it is more specific.

9. Annotate minimum necessary text spans for a disease. For example, select "hypertension" instead of "sustained hypertension."

10. Annotate all mentions of a disease entity in an abstract. All occurrences of the same disease mention should be marked, including duplicates within the same sentence.

11. Annotate abbreviations.Abbreviations should be annotated separately. For instance,"Huntington disease (HD)"should be separated into two annotations: "Huntington disease" and "HD".

12. Annotate mentions with morphological variations such as adjectives. Only when the adjective describes a specific disease. For instance, “hypertensive” should be annotated as it comes from “hypertension.”

13. Annotate all words from a composite disease mention should be annotated. For example in "ovarian and peritoneal cancer", "ovarian and peritoneal cancer" should be annotated as one entity.

What not to annotate?

1. Do NOT annotate words that define *how* a disease is expressed, e.g. plaque in "plaque psoriasis".

2. Do NOT annotate patient demographics, e.g. "elderly people".

3. Do NOT annotate the word "patient", e.g. "knee surgery patients".

4. Do NOT include species names as part of a disease. Organism names such as “human” are generally excluded from the preferred mention unless they are critical part of a disease name. Viruses, bacteria, and other organism names are not annotated unless it is clear from the context that the disease is caused by these organisms. e.g. “HIV-1-infected” means the disease caused by the organism “HIV”.Thus, “HIV” should be included.

5. Do NOT annotate symptoms, e.g. stomach ache, headache, arm weakness. Unless it’s a clear target of the Intervention, e.g. in "depressive symptoms after stroke" both "depressive" and "stroke" should be annotated separately.

6. Do NOT annotate general terms that occur individually and are not specific, such as: disease, syndrome, deficiency, complications, etc.

7. Do NOT annotate references to biological processes such as “tumorigenesis” or “cancerogenesis”.

8. Do not annotate the condition if it is within another linguistic expression. For example, in "Total Tic Severity Index", "Tic" should not be annotated.

H.3 Intervention Mention Annotation

Our working definition of **Intervention** includes any "treatment, procedure, or other action taken to prevent or treat disease, or improve health in other ways."[19].

For the annotation on Drug/Chemical-based therapies, we follow closely the guidelines of constructing CHEMDNER corpus for annotating chemical mentions (Krallinger et al., 2015), as well as (Li et al., 2016b). The basic rule for chemical entity annotation is that the chemical should have a specific structure.

General guidelines:

1. Annotate both the tested intervention and its control intervention, e.g. "home visits (OTHER) vs out-patient visits (CONTROL)" results in two annotations. A special label for CONTROL is provided.

2. In the case of a non-drug intervention, annotate all further specifying terms. E.g. in the sentence "[...] a single injection Transmuscular Quadratus Lumborum (TQL) block, when compared to [...]", the whole phrase "single injection Transmuscular Quadratus Lumborum (TQL) block" should be annotated. Words in parenthesis that give further details about the intervention should not be annotated, e.g. in "remote visit (via phone or videochat)" only "remote visit" is to be annotated. An exception are abbreviations or a clear synonym of the intervention. E.g. in "Brindley technique (anterior sacral root stimulation with posterior rhizotomy) is the only technique" both "Brindley technique" and the defintion in the brackets should be annotated.

3. Prophylaxis and prevention related Interventions should be annotated as "OTHER". E.g. in "safe and efficacious ischemic stroke prophylaxis for [...]." the phrase "ischemic stroke prophylaxis" is to be annotated. This holds only if there is no other more specific intervention stated. E.g in "Migrane prevention using Short Pulswave Therapy", "migrane" should be annotated as CONDITION while the INTERVENTION is "Short Pulswave Therapy".

4. Monitoring and diagnostic procedures should not be annotated as interventions, e.g. in

[19]https://www.cancer.gov/publications/dictionaries/cancer-terms/def/intervention

"The aim of this study is to evaluate nocturnal hypertension with 24-hour ambulatory blood pressure [...]" the phrase "24-hour ambulatory blood pressure" is not an intervention.

5. We annotate any interventions that aim at improving the health quality outcomes, even if the population/condition is not of immediate relevance. E.g. in "Evaluation of Computer-based Training to Educate Japanese Physicians in the Methods of Interpreting PET Scans." the terms "Computer-based Training" should be labeled.

6. Words that can not stand alone as a specific intervention outside of the study context should not be annotated, e.g. "stimulation", "rehabilitation" alone should not be included. At the same time "rehabilitation treatment" should be annotated. An exception should be made if the generic word is the only mention of the tested intervention in the text.

7. Both umbrella terms, and more specific annotations (if eligible) should be annotated, e.g. If those two terms appear in different positions of the sentence, "rehabilitation treatment [...] yoga exercise", both need to be annotated. Equally valid in "Mitoxantrone (MITO, Novantronae), a synthetic anthracenedione approved for [...]", both "Mitoxantrone" and "anthracenedione" should be annotated.

8. If the intervention is part of an accepted therapeutic regiment, e.g. "radio-chemotherapy", all involved interventions need to be annotated as such. E.g. In "study will evaluate whether the dosage of 1500 mg/m2 of capecitabine is tolerable after radiation" both "capecitabine" (DRUG) and "radiation" (RADIOTHERAPY) should be annotated.

What to annotate?

I. DRUG

1. Below are general guidelines for Chemical annotation that should help identify entities for annotation. Chemicals' sub-types are represented in Fig. 6. They are to be annotated with the single label **DRUG**. :

   (a) Chemical Nouns convertible to:
   -A single chemical structure diagram: single atoms, ions, isotopes, pure elements and molecules such as: Calcium(Ca), Iron(Fe), Lithium (Li),Potassium(K), Oxygen(O2),
   -A general Markush diagram with R groups such as: Amino acids

   (b) General class names where the definition of the class includes information on some structural or elemental composition such as: steroids, sugars, fatty acids, saturated fatty acids

   (c) Small Biochemicals
   - Monosaccharides, disaccharides and trisaccharides: Glucose, Sucrose...
   - Peptides and proteins with less than 15 aminoacids: Angiotensin II...
   - Monomers, dimmers, trimmers of nucleotides: e.g. ATP, cAMP...
   - Fatty acids and their derivatives excluding polymeric structures. e.g. Cholesterol, glycerol, prostaglandin E1

   (d) Synthetic Polymers such as: Polyethylene glycol

   (e) Special chemicals having well-defined chemical compositions. E.g. "ethanolic extract of Daucus carota seeds (DCE)"; "grape seed proanthocyanidin extract"

   (f) Other substances, that cannot be associated to a clear molecular structure, such as Olive Oil, Herbal Extracts, Cannabis, Tea, are to be annotated as **OTHER**.

2. For combined drugs, mark them separately, e.g. "levodopa/carbidopa" should be two entities "levodopa" and "carbidopa".

3. Chemicals that are compared in a study and separated with a "vs" should be annotated separately, e.g. "GLP-1 analogues vs DPP4 inhibitors for the treatment of type 2 diabetes mellitus".

4. Annotate all mentions of a chemical entity in an abstract.

5. Annotate the word "Vaccine" together with the immunogenic component.

6. Annotate abbreviations. Some abbreviations are ambiguous by convention. Take "Nitric Oxide (NO)" as an example, "NO" could also be interpreted as a negative response. Ambiguity should be avoided using context, i.e. in this case "NO" should not be annotated.

7. If a DRUG mention is present that is already part of the patient treatment (but is not the primary target of investigation), it should still be label as DRUG, as it is part of the overall treatment.

II. Other interventions
The below mentions represent individual labels.

1. **BEHAVIOURAL**, e.g. meditation, cognitive behavioural therapy, or other education related interventions.
2. **SURGICAL** (incl. tissue-based therapy), e.g. organ transplantation, stem cell transplantation. Injections and transfusions do not fall into this category and should be annotated as "OTHER" instead.
3. **RADIOTHERAPY**, e.g. proton beam therapy, radioactive iodine.
4. **PHYSICAL**, interventions requiring active participation from the study population e.g. cardiovascular strengthening. In case the intervention does not clearly state that active participation is required, but it could involve it based on the intervention description, the label PHYSICAL should be used, e.g. "Kinesiology".
5. **OTHER**, other types of interventions that should be annotated in a more inclusive/broad way e.g. gluten-free diet, clear liquid diets, gene therapy, Virtual Reality, medical massage. An example for a broad inclusion is "Ultrasound-guided Erector Spinae Plane Block".
6. **CONTROL**, The most specific mention of the control interventions should be annotated, e.g. in "sham product (vitamins)" the word "vitamins" should be annotated. However if there is no specific mention, general words such as "placebo", "sham product" should be labeled. Drugs should be annotated as drugs even if they are a control intervention. If in doubt about whether something is a control intervention, annotate as "Other" (or the respective intervention class). e.g., "Test catheters compared to SL catheters".

What not to annotate?

1. Do NOT annotate words that describe *how* an intervention is delivered, unless it is an essential part of the intervention. For example *Household Water Treatment Device* in "Trial of a Household Water Treatment Device as a Delivery System for Zinc in Zinc Defficient children." should NOT be annotated, while *computer-guided interpositional sandwich osteotomy* should be annotated in "The aim was to assess the efficiency of the computer-guided interpositional sandwich osteotomy [...]." Other examples include "Vitamin B (DRUG) supplement (not annotated)", "THC (DRUG) infusion (not annotated)"
2. Do NOT annotate other terms different from chemical nouns. Adjective forms of chemical names are also excluded. For instance, muscarinic, adrenergic and purinergic.
3. Do NOT annotate chemical nouns named for a role or similar, that is, nonstructural concepts (e.g. anti-HIV agents, anticonvulsants, anticholinesterase drug, antipsychotic, anticoagulant, etc).
4. Do NOT annotate very nonspecific structural concepts.e.g. Atom, Ion, Molecular, Lipid, Protein. Exception is when some of these workds are part of a longer specific chemical name, e.g. "chloride ion", "thiol dimers".
5. Do NOT annotate words that are not chemicals in context, even if they are co-incidentally the same set of characters (synonyms and metaphors). For instance,"Gold" should not be annotated if it appears in "gold standard." This applies also to general drug names, e.g. cellulose, glucocorticoid.
6. Do NOT annotate general vague compositions. For instance, according to Wikipedia, the term opiate describes any of the narcotic opioid alkaloids found as natural products in the opium poppy plant, Papaver somniferum, and thus should be excluded.
7. Do NOT annotate special words not to be labeled by convention (e.g. Water, saline, juice, etc).
8. Do NOT tag acronyms that are of 1 letter in length.
9. Do NOT include trademark symbols, e.g. Mesupron®should result in the annotation "Mesupron".

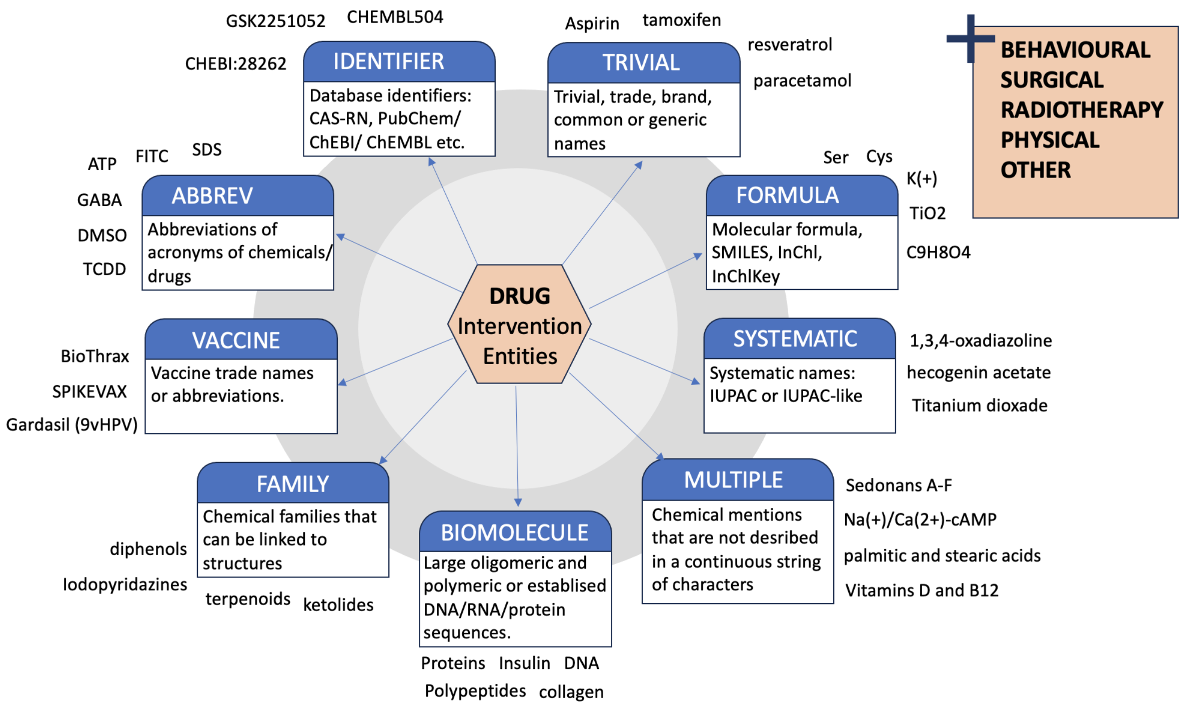

Figure 6: Overview of chemical-based interventions, adapted from (Krallinger et al., 2015) and other types of interventions of interest.