Natural Language Processing Methods to Automatically Parse Eligibility Criteria in Dietary Supplements Clinical Trials

A Thesis

SUBMITTED TO THE FACULTY OF THE

UNIVERSITY OF MINNESOTA

BY

Anusha Bompelli

IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

Advisor: Dr. Rui Zhang

August 2020

Acknowledgements

I would like to acknowledge everyone who played a role in my academic accomplishments. My thesis work would not have been possible without the support of my advisor, committee members, research colleagues, professional colleagues, friends, and family.

I would especially like to thank my advisor, Dr. Rui Zhang, for providing me an opportunity to work on this project and for his continuous support, guidance, and mentorship. His guidance is instrumental and critical in the completion of my research project and publications. I would like to thank my thesis committee members, Dr. Terrence Adam and Dr. Steve Johnson for their continuous support and insights regarding the research.

I'm indebted to my research collaborators whose collaborative efforts made this project possible. I thank Dr. Zhe He from the Florida State University and Professor Serguei Pakhomov for initial insights about the research and their contribution to the manuscript. I would like to thank, Greg Silverman, Raymond Finzel, Jake Vasilakes and Benjamin Knoll for helping me with huge efforts that went into working with NLP-ADAPT. I especially thank Greg Silverman for patiently answering my questions and his work to the manuscript resulted in a much stronger paper. I'm grateful to JianFu Li, Yiqi Xu and Nan Wang for their contribution to the research work.

I'm grateful to our Director of Graduate Studies Dr. David Pieczkiewicz for being an excellent teacher and mentor who never failed to give sound advice. I thank the Institute for Health Informatics staff, and graduate program coordinators Jessica Tetzlaff and Jenna Lohnes who were always ready to help.

I thank Becky Lowery and Anna McDonagh, Contracts Managers, Office for Technology Commercialization for supporting me throughout my master's journey.

I would certainly be remiss if I don't acknowledge my classmates Surbhi Nagpal, Changye Li, Vidhya Ramesh, Xinpeng Shen, Rachel Zhang, Qui Mo, Samad Qureshi, Prithul Bom and Mohammed Mansi for amazing friendship. I'm so grateful to Surbhi Nagpal for making my graduate school experience fun and memorable.

Finally, my family and friends, who supported me with love and understanding. Without you, I could never have reached this current level of success.

Dedication

I wholeheartedly dedicate this work to my family and friends. A special feeling of gratitude to my loving parents who have been a source of inspiration and for supporting me throughout the process. My sisters and friends have never left my side and I thank them for being there for me.

Abstract

Dietary supplements (DSs) have been widely used in the U.S. and evaluated in clinical trials as potential interventions for various diseases. However, many clinical trials face challenges in recruiting enough eligible patients in a timely fashion, causing delays or even early termination. Using electronic health records to find eligible patients who meet clinical trial eligibility criteria has been shown as a promising way to assess recruitment feasibility and accelerate the recruitment process. Natural Language Processing (NLP) techniques have been used extensively to extract concepts from the clinical trial eligibility criteria. However, a significant obstacle is identifying an efficient Named Entity Recognition (NER) system to parse the clinical trial eligibility criteria.

The study comprises of two parts. In the first part of the study, the objective was to (1) understand data elements associated with DS trials' eligibility criteria and assess if they can be mapped to OMOP Common Data Model (CDM); (2) develop and evaluate NLP methods, especially deep learning-based models, for extracting eligibility criteria data elements. We analyzed the eligibility criteria of 100 randomly selected DS clinical trials and identified both computable and non-computable criteria. We mapped annotated entities to OMOP Common Data Model (CDM) with novel entities (e.g., DS). We also evaluated a deep learning model (Bi-LSTM-CRF) for extracting these entities on CLAMP platform, with an average F1 measure of 0.601. This study shows the feasibility of automatic parsing of the eligibility criteria following OMOP CDM for future cohort identification.

In the second part of the study, the objective was to examine the performance of standard open-source clinical NLP systems for the task of Named Entity Recognition (NER) for a corpus outside of the domain for which these systems were developed. we used NLP-

ADAPT (Artifact Discovery and Preparation Toolkit) to compare existing biomedical NLP systems (BiomedICUS, CLAMP, cTAKES and MetaMap) and their Boolean ensemble to identify entities of the eligibility criteria of 150 randomly selected Dietary Supplement (DS) clinical trials. We created a custom mapping of the gold standard annotated entities to UMLS semantic types to align with annotations from each system. All systems in NLP-ADAPT used their default pipelines to extract entities based on our custom mappings. The systems performed reasonably well in extracting UMLS concepts belonging to the semantic types Disorders and Chemicals and Drugs. Among all systems, cTAKES was the highest performing system for Chemicals and Drugs and Disorders semantic groups and BioMedICUS was the highest performing system for Procedures, Living Beings, Concepts and Ideas, and Devices. Whereas, the Boolean ensemble outperformed individual systems. This study sets a baseline that can be potentially improved with modifications to the NLP-ADAPT pipeline.

Table of Contents

List of Tables

List of Figures

1. Introduction

Clinical trials are one of the most valuable resources for healthcare practitioners to practice evidence-based medicine as clinical trials are usually accepted as the most unbiased measures of efficacy and safety for new interventions.[1,2] Patient recruitment is an essential part of the clinical trial with eligibility criteria (study specific patient characteristics) determine whether a patient should be included or excluded from the study.[3] However, patient recruitment is a challenging and pressing issue for researchers as it has several barriers, including the lack of patient awareness of clinical trials and access to trials, age limitations, complex study designs, fewer eligible patients than expected due to restrictive eligibility criteria and several other reasons.[4,5,6] An analysis of registered trials showed that approximately 85% of trials were not able to complete required recruitment in the pre-defined time and around 20% of the trials were closed or terminated early due to inadequate patient recruitment,[7] limiting the statistical power of the evidence related to the new interventions.[8] Moreover, more than 70% of clinical trial generalizability assessment studies reported low generalizability of completed trials, partly due to low enrollment.[9]

The rapid growth of the electronic health records (EHR) provides an unprecedented opportunity to harness its data to full potential for secondary use.[10] Moreover, the last few years have also witnessed an increasing number of clinical research networks focused on building large collections of data from EHRs and claims to provide cohort discovery services. Two notable examples are the National Patient-Centered Clinical Research Network (PCORnet), funded by Patient-Centered Outcomes Research Institute (PCORI), and the CTSA Accrual to Clinical Trials (CTSA ACT) initiative.[11,12] In addition, a number of national efforts are building tools, algorithms, and data models to identify the eligible

patients and to reduce the recruitment delays due to the aforementioned challenges. For example, (1) i2b2 has a widely used cohort discovery tool[13]; (2) the Electronic Medical Records and Genomics (eMERGE) Network is building computable phenotypes for cohort discovery[14]; (3) the stakeholders of the Observational Health Data Sciences and Informatics (OHDSI) consortium are developing open source analytical tools based on the OMOP (Observational Medical Outcome Partnership) Common Data Model (CDM).[15] The majority of the approaches develop computable representations of the clinical trial eligibility criteria and apply it to EHR data to find eligible patient cohorts. However, as the eligibility criteria is majorly in primarily in free-text format, it is essential to understand the schema of the criteria, the elements (entities and attributes) and potential to parse the data to extract the elements to provide decision support for clinical trial cohort identification.[16]

The NER approaches such as EXACT[17], EliXR[18], EliIE[19], ULTRA[20], etc., which were developed to represent eligibility criteria in a structured format and were confined to certain drugs and medical conditions. Recently, Si et al. developed a natural language processing (NLP) system to extract medical terms in eligibility criteria of Alzheimer's disease clinical trials and represent them using the OMOP CDM.[21] Criteria2Query was developed to systematically transform eligibility criteria text into SQL queries over OMOP CDM databases.[22]

Many NLP tools including MedLEE[23], the Clinical Language Annotation, Modeling, and Processing Toolkit (CLAMP)[24]; the Clinical Text Analysis and Knowledge Extraction System (cTAKES) [25]; etc. have been developed to extract information related to anatomical location, signs and symptoms, diseases, procedures, laboratory tests and medications[26, 27].

In the clinical domain, extraction of clinical information or concepts is not adequate since the concepts are significantly affected by attributes such as negation modifier, temporal information and qualifiers which describe condition status or severity[28, 29]. NLP tools are generally trained using a specific dataset and are suitable to extract certain concepts. For example, CLAMP's pipeline was trained on the 2010 VA challenge i2b2 corpus to recognize problems, drugs, treatments and lab tests[30]. Thus, it is challenging to find an NLP tool capable of extracting diverse concepts, modifiers and attributes.

To the best of our knowledge, understanding the eligibility criteria of DS trials and developing computable representations have not been investigated. From our previous published study, we identified that the dietary supplement (DS) trials eligibility criteria are different from drug clinical trials in the aspects like trial objectives and criteria related to trial objectives, demographics (such as age, gender, race), and disease or lab parameters. For certain diseases, drug clinical trials are more therapeutic oriented whereas dietary supplements are either preventive and therapeutic.[31] Thus, making the DS clinical trial eligibility criteria unique.

The study comprises of two parts.

1. The objective of the first part of the study was to (1) understand data elements associated with DS trials' eligibility criteria and assess if they can be mapped to OHDSI CDM; (2) develop and evaluate NLP methods, especially deep learning-based models, for extracting eligibility criteria data elements. In this, we first manually annotate free-text eligibility criteria from a sample of 100 DS clinical trials following OMOP CDM v6.0 and then train and compare both conventional

machine-learning-based versus deep-learning-based models on the CLAMP platform to automatically extract different components of eligibility criteria.

2. The objective of the second part of the study was to examine the performance of standard open-source clinical NLP systems for the task of Named Entity Recognition (NER) for a corpus outside of the domain for which these systems were developed. We examined a particular strategy for combining annotations generated from out-of-the-box clinical NLP systems into ensembles using NLP systems provided by the NLP Artifact Discovery and Preparation Toolkit (NLP-ADAPT)[32]. NLP-Ensemble-Explorer[33] integrates output from NLP-ADAPT, and through use of a custom mapping to UMLS concepts, allowed us to investigate performance of individual systems and their ensembles for the task of NER for several semantic groupings of UMLS concepts on a novel corpus. NLP-ADAPT and NLP-Ensemble-Explorer were both developed as a complete pipeline for clinical researchers to help improve the experience of the exploration phase of NLP and Information Extraction (IE) projects using individual NLP systems and their ensembles.

2. Deep Learning Approach to Parse Eligibility Criteria in Dietary Supplements Clinical Trials Following OMOP Common Data Model[1]

2.1. Background

2.1.1. Observational Medical Outcomes Partnership Common Data Model (OMOP CDM)

The OMOP CDM harmonizes the disparate observational databases with minimal loss of information and enables the interoperability among the databases.[15] It aims to facilitate research, support the conduct of EHR and manage the claims data. An integral part of CDM is OMOP standardized vocabularies which enable the exchange of patient data among different systems and allow mapping for use in research (https://ohdsi.github.io/TheBookOfOhdsi/). In this study, the definition of entities and attributes such as observation, procedure, device, condition, drug and measurement, are determined following the OMOP CDM standardized vocabularies and clinical data tables.

2.1.2. Named Entity Recognition (NER)

Clinical NER is a critical task for information extraction (IE) from text and to identify semantics (for example, entities, attributes, relations and events) to support the clinical and translational research. The widely-used clinical NER approaches include MedLEE[34], MetaMap[35], KnowledgeMap[36] and cTAKES[25]. These systems were designed for clinical notes in EHRs or text in the biomedical literature. Studies that investigate the adaptability of these systems to parse clinical trial eligibility criteria are limited. EliXR[18], EliIE[19] developed by Weng et al. and EXACT[17] developed by Yu et al. were designed specifically

[1] This chapter has been accepted for an oral presentation at the American Medical Informatics Association (AMIA) 2020 Virtual Annual Symposium

for parsing eligibility criteria of drug clinical trials. These tools are often limited to certain types of entities and attributes. Early NER methods used dictionary-based, rule-based and supervised machine learning. Later, hybrid methods became popular as combinations of different methods improved the overall performance. Most recently, state-of-the-art deep learning methods have been widely used in textual data representation. One of the most effective NER models being used is Bi-directional long short-term memory (LSTM) with a Conditional Random Field (CRF) on the top layer (Bi-LSTM-CRF).[37]

2.1.3. Clinical Language Annotation, Modeling and Processing (CLAMP)

CLAMP is a clinical language annotation, modeling and processing software designed by Xu et al.[32] It is a user-friendly tool and follows the Unstructured Information Management Architecture (UIMA) architecture. The key building blocks of CLAMP are NLP pipelines, machine learning and hybrid approaches, corpus management and annotation tool. CLAMP has multiple components such as sentence boundary detection, tokenizer, part-of-speech tagger, section header identification, abbreviation recognition and disambiguation, named entity recognizer, assertion and negation, UMLS encoder and rule engine. CLAMP's named entity recognizer contains different types of NER approaches - machine learning-based, dictionary-based, regular expression-based, and deep learning-based. The software has diverse applications in different clinical domains.[38]

2.2. Methods

2.2.1. Overview of the study

In this study as shown in **Figure 1**, we followed five steps: (1) obtaining the clinical trial eligibility criteria of DS clinical trials from ClinicalTrials.gov; (2) analyzing the eligibility

criteria and mapping to OMOP CDM v6.0; (3) developing the gold standard annotation; (4) developing named-entity recognition algorithms (i.e., CRF and Bi-LSTM-CRF) using CLAMP; and (5) evaluating the NER models using gold standard annotations.

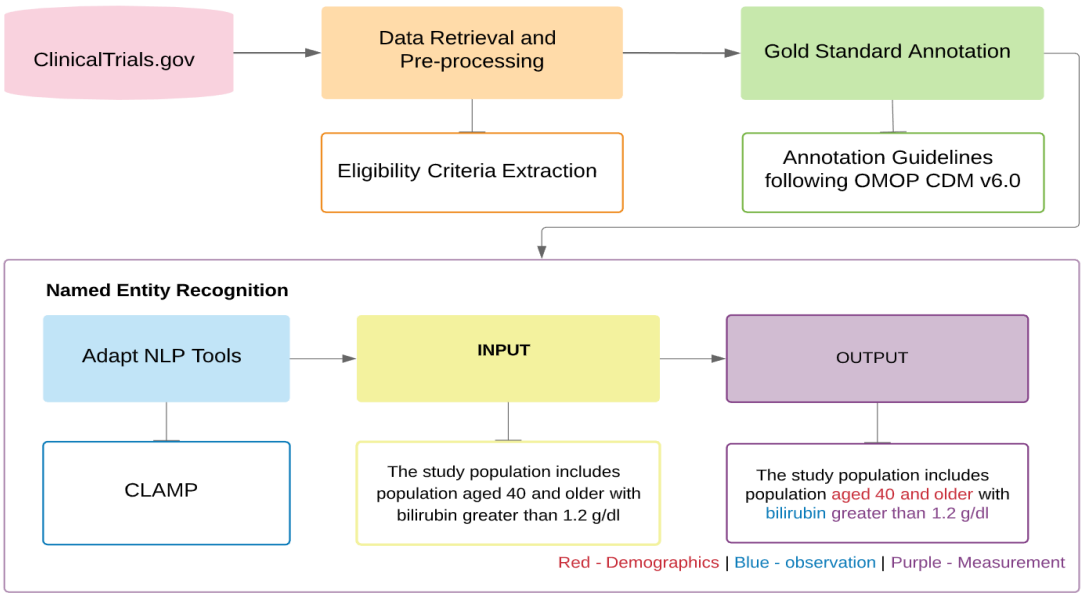

Figure 1. Overview of the Method for Extracting Clinical Trial Eligibility Criteria.

2.2.2. Data source and collection

ClinicalTrials.gov is an online repository developed by the U.S. National Library of Medicine (NLM) and the National Institutes of Health (NIH). The repository contains 332,005 research studies which are privately or publicly funded.[39] We obtained 859 clinical trials from ClinicalTrials.gov by applying the search criteria: (1) using DSs as an intervention, (2) NIH funded trials, and (3) restricting the location to the United States. The extracted trials belong to 22 disease categories. We limited the study to *Behaviors & Mental Disorders and Nervous System Diseases* consisting of 149 trials and *Nutritional and Metabolic Disorders* comprising 199 trials because these two categories of disease categories are the most prevalent domains. Clinical trial data can be obtained in various forms – XML, PDF, plain text, etc. We chose the XML format as it retains the structure of

the original document. The document contains both the structured and unstructured data clearly marked with the respective section tags as shown in **Figure 2**. The documents were parsed to extract the eligibility criteria.

```
<eligibility>
  <criteria>
    <textblock>
      Inclusion Criteria:
        - Diagnosis of AD by current McKhann et al. criteria
        - CDR global score of 0.5 or 1
        - Agreed cooperation from an appropriate study partner
        - Speaks English as primary language
        - Age 50 to 90
        - No medication changes within the past 30 days
      Exclusion Criteria:
        - Resides in a nursing home or dementia special care unit, or cannot control diet
        - A potentially confounding serious medical risk including insulin-requiring diabetes,
          cancer requiring chemotherapy or radiation within the past 5 years, or a recent
          cardiac event (i.e. heart attack, angioplasty, etc.)
        - Participating in another clinical trial or using an investigational drug or therapy
          within 30 days of the Screening Visit
        - A history of renal stones
    </textblock>
  </criteria>
</eligibility>
```

Figure 2. The eligibility criteria section of an example clinical trial (NCT03860792)

2.2.3. Analyzing the eligibility criteria and mapping to OMOP CDM

Eligibility criteria, including both the inclusion and exclusion criteria, are the list of requirements that an individual must satisfy to be enrolled in the clinical trials. Eligibility criteria can be either short or lengthy, largely free text descriptions spanning several sentences. Each trial comprises an average of 10 criteria (including inclusion and exclusion). Eligibility criteria contain the information about the individual demographics, observation and findings, condition, lifestyle and treatment, as shown in **Figure 3**. The temporal measurement, which is essential but not in the schema, is an element associated with observation, diagnosis, prognosis and treatment. As OMOP CDM standardizes data using a common information model and multiple standard terminologies bridging the interoperability among disparate observational databases, we compared the elements in the schema with OMOP CDM v6.0, domains and mapped the elements to OMOP CDM data tables. We observed that entities like condition, observation/findings and lifestyle,

procedures, demographics from the schema can be mapped to OMOP CDM data tables such as condition, observation, procedure and person, respectively. Whereas the sub-entities in the treatment such as drug and device can be mapped to drug and device in OMOP respectively. We found that information about dietary supplements is missing from the OMOP CDM data tables and this element makes this study unique.

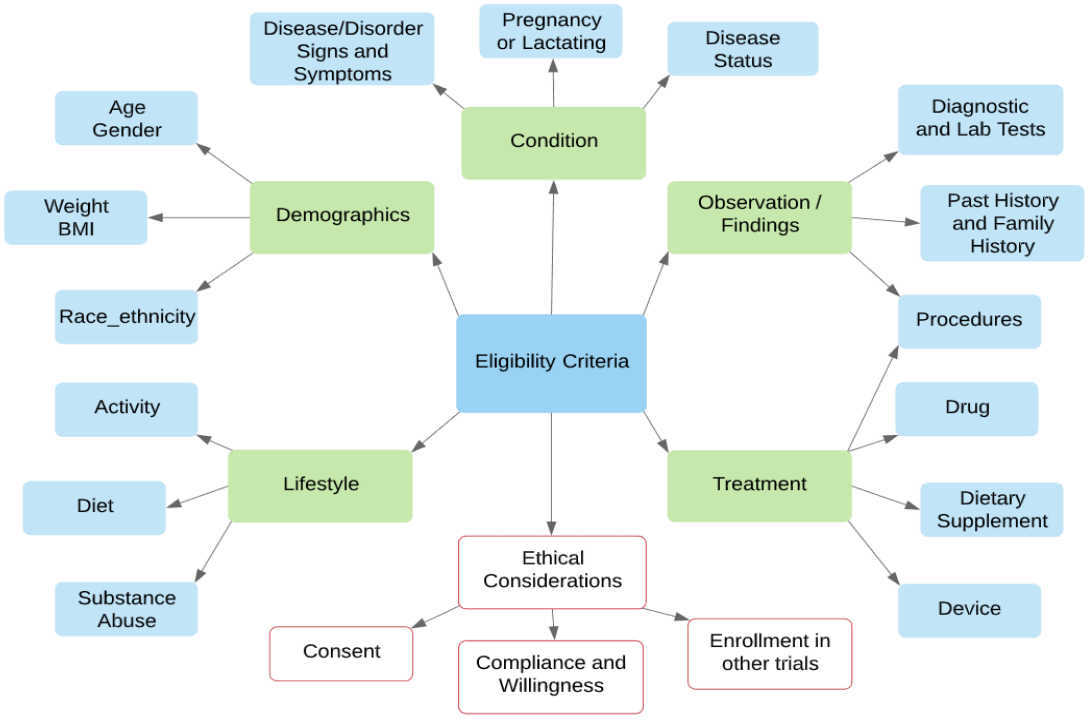

Figure 3. Schema of the elements in clinical trial eligibility criteria. In this figure, different colors are used for the main categories (green) as well as the components (light blue) in eligibility criteria. The category 'ethical considerations' and the components associated with it are marked in red as they cannot be found in EHR. The schema laid the foundation to map entities to OMOP CDM.

2.2.4. Manual Annotation

We developed the first iteration of the annotation guidelines based on the OMOP CDM v6.0. We added dietary supplements as an entity as it appears in these DS clinical trials. Three annotators (AB, YX and NW) independently annotated 5 randomly selected trials using CLAMP by understanding the first iteration of the guidelines. The team compared

the annotation results, discussed the difference of opinions and revised the annotation guidelines. While annotating, we observed that certain criteria have information about lifestyle choices and annotated the lifestyle choices as observation following OMOP CDM. Next, the team independently annotated 5 trials from each category. The team then discussed and annotated another set until a reasonable interrater agreement is reached and until no discrepancy among annotators. Inter-annotator agreement was computed over 10 trials, revealing a kappa of 0.94. After finalizing the annotation guidelines[2], totally 100 trials, 50 trials for each of the two categories, were randomly selected for final corpus annotation which comprised approximately 1843 Sentences. The entities and attributes described in the guidelines are given below in **Table 1**.

Table 1. Eligible criteria entities and attributes with selected examples.

	Semantic Class	Example Criteria (entities and attributes are underlined and marked in blue)
Entity	Demographics	Women must be > 18 to 45 years of age; BMI = 27 kg/m2;
	Observation	Bilirubin greater than 1.2 g/dl; MMSE below 24, dementia or unstable clinical depression by exam
	Procedure	History of bilateral hip replacement
	Condition	Uncontrolled hypertension (BP over 180mm HG)
	Drug	Taking metformin, propranolol and other medications
	Dietary supplement (ds)	Use of St. John's Wort or any other dietary supplement
	Device	Claustrophobia, metal implants, pacemaker or other factors affecting feasibility and / or safety of MRI scanning
Attribute	Measurement	BUN above 40 mg/dl, Cr above 1.8 mg/dl, CrCl < 60 mg/dl
	Qualifier	Signs and symptoms of increased intracranial pressure; severe hypercalcemia
	Temporal_measurement	Use of systemic corticosteroids within the last year
	Negation	Use of anti-diabetic drugs other than metformin

[2] https://z.umn.edu/annotation_guidelines

2.2.5. Developing Named Entity Recognition algorithms using CLAMP

We used CLAMP as a platform to develop NER algorithms on 1843 annotated sentences. We chose CRF as the baseline model. We also implemented a deep learning model, Bi-LSTM-CRF model, using TensorFlow framework, which has been demonstrated superior performance in other NER tasks.[37] LSTM network is better than traditional RNNs to find long range dependencies due to their updated hidden layer. As NER tasks often require contextual information (both past and future input features) from the sentence, we made use of a bidirectional LSTM network. The Bi-LSTM networks were trained using backpropagation through time technique and both forward and backward hidden states were concatenated to obtain contextual representations for the input sentence. To make full use of contextual tagged information, we then combined the Bi-LSTM networks with a CRF network to get a Bi-LSTM-CRF network. Finally, a 5-fold cross validation (80 trials for train and 20 trials for test) was applied to compare the performances of two models. The NER performances on each entity and attribute were reported using precision, recall and F1-measure.

2.3. Results

2.3.1. Distribution of dietary supplements in clinical trials

In this study, we observed that a wide range of DS have been studied as the interventions in clinical trials. Figure 4 lists the distribution of trials on each DS for two categories of diseases. As shown in Figure 4, the most studied dietary supplements in trials on Behaviors & Mental Disorders and Nervous System Diseases were fish oil, omega-3 fatty acids, vitamin D, vitamin E, DHA, EPA, soy, lipoic acid, selenium, folic acid and the those only studied once were black cohosh, boswellia serrata, chamomile extract, etc. Whereas in

those trials on Nutritional and Metabolic Diseases (**Figure 5**), vitamin D is predominated with 130 studies, followed by calcium, fish oil, omega-3 fatty acids, chromium picolinate were widely studied.

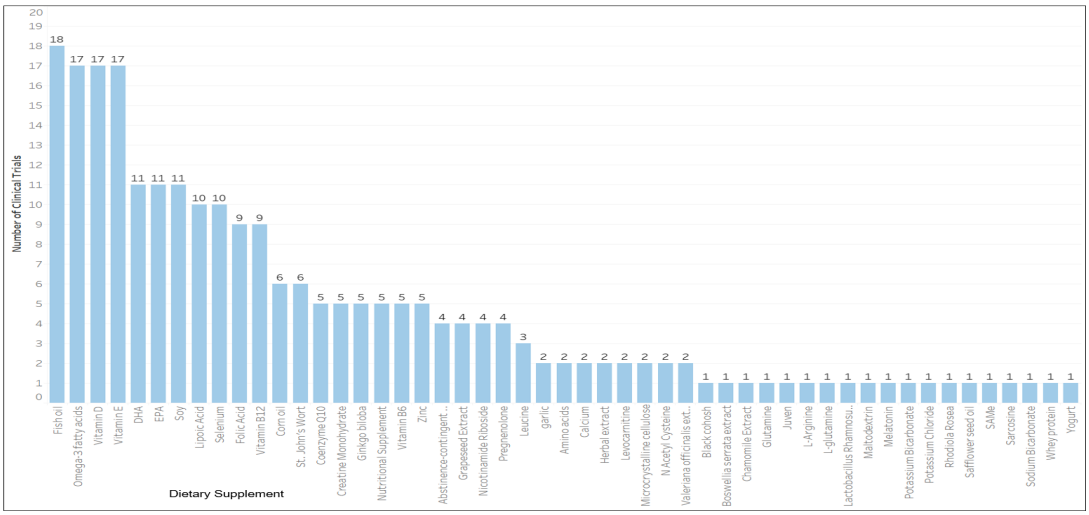

Figure 4. Distribution of DS as intervention in clinical trials on Behaviors & Mental Disorders and Nervous System Diseases

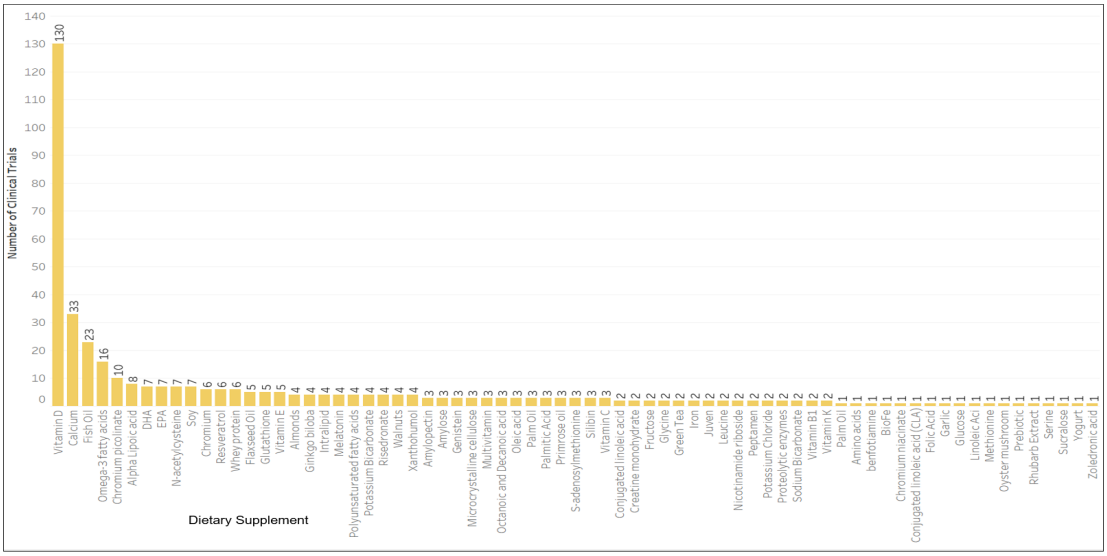

Figure 5. Distribution of DS as intervention in clinical trials on Nutritional and Metabolic Diseases

2.3.2. Statistics of eligibility criteria

Among the 1843 eligibility criteria sentences, 229 criteria were not annotated as the criteria were not computable (corresponding data not in EHR). Descriptive statistics of the

unannotated and annotated eligibility criteria corpus are provided in **Table 2** and **Table 3,** respectively. Among the unannotated 229 sentences, 23.14% of the sentences belong to the criteria referring to unwillingness or willingness of the subject, whereas 2.18% belong to partners or caregivers. Out of annotated 1614 sentences, condition entity is the largest (1401 terms), followed by drug (688 terms) and observation (671 terms) while device is the smallest (47 terms). Among the attributes, the qualifier is the largest (643 terms) followed by temporal measurement (445 terms). The average of terms of each semantic class that can be found in a trial range approximately from 2 to 5.

Table 2. Descriptive statistics of the unannotated eligibility criteria and examples

Category	**No. (%)**	**Example Criteria**
Unwillingness / Willingness	53 (23.14)	unwillingness for subject of childbearing potential to use contraception during the first year of the study
		She must be willing to practice an acceptable method of birth control
Others	39 (17.03)	The patient does not have an outside care provider for treatment of depression
		Access to a smart phone or internet and telephone
Ability / Inability	31 (13.54)	Inability to swallow oral capsules
		Ability to use a bolus calculator function with the current insulin pump with pre-defined parameters for glucose goal, carbohydrate ratio, and insulin sensitivity factor
Informed consent	28 (12.23)	Signed protocol specific informed consent prior to registration
		Subjects who refuse to sign the protocol consent document
Participation in other trials	25 (10.92)	Enrollment in any concurrent research protocols that would interfere with participant safety or research data integrity
		Previous participation in Phase 1 pharmacogenomic study
Investigator's opinion	21 (9.17)	The patient is, in the opinion of the investigator, mentally or legally incapacitated
		Any other reasons that, in the opinion of the Investigator, the candidate is determined to be unsuitable for entry into the study
Location	14 (6.11)	Lives far away from study site
		Live and work within 1 hour of the study site
Language	13 (5.67)	Fluency in English or Spanish
		Non-English speaking
Partners/care takers	5 (2.18)	Both parents/partners are required to participate in this study, not just one or the other
		Living at home with a parent or guardian

2.3.3. NER models evaluation

The detailed performance of the CRF and Bi-LSTM-CRF for the named entity recognition task is given below in **Table 3**. In almost all entities and attributes except device, Bi-LSTM-CRF outperformed CRF model. The macro-average of F1 measure for the Bi-LSTM-CRF model is 0.601. The semantic class which performed the best in the Bi-LSTM-CRF model is drug with an F1 measure of 0.769 followed by condition with 0.704. The semantic class which performed the least in Bi-LSTM-CRF is device (0.192).

Table 3. Descriptive statistics of the annotated eligibility criteria and performance of the models for entity and attribute recognition. The best F1 measures for each semantic class are in bold.

Eligibility Criteria Corpus (100 trials)			**CRF**			**BI-LSTM-CRF**		
Semantic Class		# of mentions	Precision	Recall	F1 measure	Precision	Recall	F1 measure
Entity	demographics	223	0.648	0.494	0.561	0.713	0.628	**0.668**
	observation	671	0.711	0.535	0.610	0.615	0.642	**0.623**
	procedure	122	0.750	0.214	0.333	0.712	0.679	**0.681**
	condition	1401	0.665	0.649	0.657	0.694	0.714	**0.704**
	drug	688	0.711	0.615	0.659	0.752	0.791	**0.769**
	dietary supplement (ds)	175	0.429	0.237	0.305	0.512	0.572	**0.529**
	device	47	0.750	0.200	**0.316**	0.182	0.205	0.192
Attribute	measurement	321	0.635	0.564	0.597	0.648	0.682	**0.662**
	qualifier	643	0.670	0.513	0.581	0.683	0.598	**0.637**
	temporal_measurement	445	0.714	0.542	0.616	0.653	0.666	**0.656**
	negation	215	0.817	0.519	0.635	0.867	0.618	**0.704**

2.4. Discussion

The widespread adoption and use of EHRs together with the NLP tools have led to the ability to identify and recruit patient cohort to conduct clinical trials according to eligibility criteria. However, EHRs may or may not contain all eligibility criteria data elements required for patient cohort identification. One could also incorporate clinical notes in the EHR to find patients that meet criteria that are not captured in the structured field in EHR,

e.g., MMSE of dementia patients. The eligibility criteria schema and elements should be better understood and analyzed to differentiate the criteria that are not computable.

In this pilot study, we first analyzed the eligibility criteria schema and elements (entities and attributes) for the DS clinical trials obtained from ClinicalTrials.gov for two disease categories, *Behaviors & Mental Disorders and Nervous System Diseases*, and *Nutritional and Metabolic Diseases*, the top two categories based on the frequency of DS uses. Eligibility criteria comprise multiple elements which are bound by restrictions such as qualifiers (severity of disease), negations and temporal constraints; identifying the individual entities and attributes would make EHR computability easier. As our long-term goal is to accelerate the patient recruitment using EHR data, we followed the commonly used OMOP CDM to annotate the eligibility criteria. Through our analysis on eligibility criteria of DS clinical trials, we found that dietary supplement is currently out of scope of OMOP CDM. As clinical notes will contain information about dietary supplements, we included the entity in our annotation and NER development. While annotating we observed that certain criteria can be easily computable to extract from EHR data while the rest are either hard or impossible to compute. The eligibility criteria whose corresponding data cannot be found in the EHR were not annotated even if certain terms in the criteria qualify for one of the entities or attributes as this information is not computable.

The annotators faced a few challenges while annotating the criteria. For example, the annotators had noted certain problems while annotating the criteria including (1) the criteria had both the qualifier and disease condition (e.g., multiple myeloma); (2) the criteria were difficult to differentiate as they shouldn't be annotated (e.g., aerobically trained with V02 max greater than 2 SD above age-adjusted mean); (3) ambiguity (e.g.,

certain abnormal laboratory values); (4) had interconnected entities (e.g., hepatic, renal and gastrointestinal diseases). The annotators had overcome these challenges through multiple rounds of discussions and annotations until no discrepancy among annotators.

To explore the feasibility of the NLP techniques, especially state-of-the-art deep learning models, for parsing these eligibility criteria automatically, we leveraged the annotation and CLAMP platform to compare two models. The precision and recall of the entities and attributes for the CRF model are mostly lower than the Bi-LSTM-CRF model except for entity 'device'. In the Bi-LSTM-CRF model, within the 5 folds for the entity 'device', three folds did not find the entity whereas the other two showed low precision and recall values. By analyzing the original train/dev/test dataset, we observed that the difference in values in different folds is due to a small dataset for device (totally only 47 mentions). CRF requires a list of features, while Bi-LSTM-CRF is a feature engineering free model but require a large dataset. This is one main reason why the current performance is still suboptimal. The most common semantic classes "drug", "condition" and "negation" in clinical trials reached the F1 measure over 0.70 whereas the F1 measure for other semantic classes is above 0.50 except for the semantic class "device". The low F1 measures could be due to a small dataset and lower number of mentions in the annotation corpus.

This study has a few limitations. The major limitation is a small dataset which resulted in low performance with respect to certain entities and attributes. The other notable limitation is the differences in individual perception while annotating certain concepts (e.g., condition, qualifier) in the dataset which resulted in some inconsistent annotations. Future work will focus on using a large dataset and improving the annotations consistency which would eventually improve the performance of the NER models. We will also try other deep

learning modes, such as BERT, which shows promising performance in 11 NLP common tasks.

2.5. Conclusion

In this study, we investigated the data elements associated with eligibility criteria associated with the clinical trials which use DS as an intervention. We analyzed the criteria and found both computable and non-computable criteria. We manually created eligibility criteria entities followed the OMOP CDM v6.0. We annotated these entities for 100 trials and used the annotated data to develop a Bi-LSTM-CRF model for NER task. This study demonstrates the feasibility of using CDM to represent the DS clinical trial eligibility criteria and using deep learning models for NER task in clinical trials. This study lays the foundation for future matching patients using their EHR data to DS clinical trials.

3. Comparing NLP Systems to Extract Entities of Eligibility Criteria in Dietary Supplements Clinical Trials using NLP-ADAPT[3]

3.1. Background

3.1.1. Unified Medical Language Systems (UMLS)

The Unified Medical Language Systems (UMLS) developed and maintained by the National Library of Medicine (NLM) provides a unified global biomedical terminology[40]. The UMLS semantic network organizes many concepts and groups them according to the semantic types[41]. There are 15 semantic groups, 133 semantic types and 54 semantic relationships[42]. The semantic network has been widely used in information extraction, clinical annotation, and knowledge representation[43].

3.1.2. NLP systems

Parsing clinical notes is a critical task for information extraction (IE) as it leverages information from narrative text to support clinical and translational research[27]. Clinical NLP systems such as the BioMedical Information Collection and Understanding System (BioMedICUS)[44]; the Clinical Language Annotation, Modeling, and Processing Toolkit (CLAMP)[45]; the Clinical Text Analysis and Knowledge Extraction System (cTAKES)[46]; and MetaMap[47] have been developed to perform Named Entity Recognition (NER) and IE tasks on free text clinical notes or biomedical literature. Because many of these systems were developed to extract specific types of information, adopting these systems for use

[3] This chapter has been accepted for a long presentation at the 2020 International Conference on Artificial Intelligence in Medicine (AIME)

beyond their original purpose without customization of each system's statistical models and dictionaries can potentially result in reduced performance[24].

3.2. Methods

3.2.1. Overview of the study

This study compares performances of different NLP systems and their ensembles for the task of concept extraction from unstructured dietary supplements (DSs) clinical trial eligibility criteria. The study was performed following these steps: (1) obtain the clinical trial eligibility criteria of DS clinical trials from ClinicalTrials.gov; (2) develop gold standard annotations; (3) map entities to UMLS semantic types; (4) Apply NLP-ADAPT to extract entities mapped to semantic types; and (5) use NLP-Ensemble-Explorer to create ensembles and compare the performance of individual and ensembled annotator systems against the gold standard annotations (see Fig. 6).

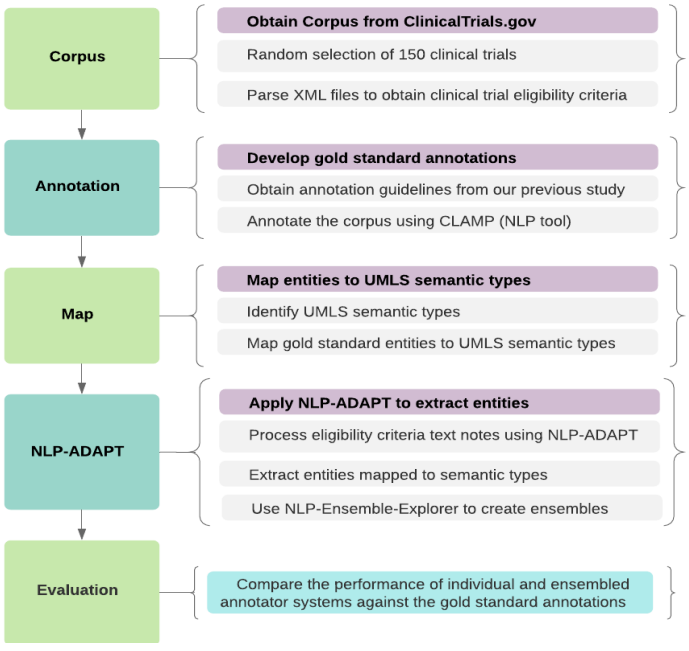

Fig. 6. Overview of the study

3.2.2. Corpus and annotation

We obtained the dietary supplements clinical trial data corpus from ClinicalTrials.gov, which is an online repository developed by the National Library of Medicine (NLM) and the National Institutes of Health (NIH). We randomly selected 150 clinical trials from the Behaviors & Mental Disorders and Nervous System Diseases categories and parsed the clinical trial XML files to obtain the eligibility criteria. We annotated the eligibility criteria by following the annotation guidelines[4] developed in our previous unpublished study. Three annotators independently annotated 5 randomly selected clinical trials by understanding the first iteration of the guidelines. The team compared the annotation results, discussed the difference of opinions and revised the annotation guidelines. The team then annotated another set until a reasonable interrater agreement is reached and until no discrepancy among annotators. Later, Inter-annotator agreement among three annotators was computed over 10 trials, revealing a kappa of 0.94. The annotated entities and attributes include: *Demographics*, *Observation*, *Condition*, *Procedure*, *Device*, *Drug*, *Dietary Supplement*, *Negation Modifier*, *Qualifier*, *Measurement*, and *Temporal Measurement*. While annotating we observed that certain criteria can be easily computable to extract from EHR data while the rest are either difficult or impossible to compute. Examples of such criteria are given below: criteria referring to willingness or unwillingness to use or decline certain medications or methods (contraception); requiring consent; ability/inability and compliance of the individual, and caregivers or study partners; criteria about participant's enrollment in any other clinical trial. The eligibility criteria whose corresponding data cannot be found in the EHR were not annotated even if certain terms

[4] https://z.umn.edu/annotation_guidelines

in the criteria qualify for one of the entities or attributes as this information is not computable.

3.2.3. Mapping to UMLS semantic groups across NLP systems

To compare the performance of individual NLP systems against our gold standard annotations, the entities and attributes present in the gold standard annotation were mapped to UMLS semantic types/groups and annotation types available in individual NLP systems. We observed that some entities can be mapped to one or more semantic groups. For example, *Condition* and *Observation* were mapped to *Disorders* (see Figure 7.) Figure 7 also illustrates that CLAMP and cTAKES don't annotate four and two semantic groups respectively.

UMLS Semantic Groups	BioMedICUS	CLAMP	cTAKES	MetaMap	Gold Standard Entities
Chemicals and Drugs	T130, T121, T195, T192, T123, T131, T103, T196, T125, T116, T120, T129, T114, T122, T104, T197, T109, T127, T200, T126	• drug	• MedicationMention	irda, phsu, antb, rcpt, bacs, hops, chem, elii, horm, aapp, chvf, imft, nnon, bodm, chvs, inch, orch, vita, clnd, enzy	Drug Dietary Supplement
Disorders	T049, T033, T047, T019, T048, T184, T191, T050, T020, T190, T037, T046	• problem • lab value	• DiseaseDisorderMention • SignSymptomMention	comd, fndg, dsyn, cgab, mobd, sosy, neop, emod, acab, anab, inpo, patf	Condition Observation
Procedures	T060, T059, T058, T063, T062, T065, T061	• test • treatment	• ProcedureMention	diap, lbpr, hlca, mbrt, resa, edac, topp	Procedures Observation
Living Beings	T100, T098	-	-	aggp, popg	Demographics
Device	T203, T074, T075	-	-	drdd, medd, resd	Device
Concepts and Ideas	T080, T081, T079	-	DateAnnotation	qlco, qnco, tmco	Temporal Measurement Measurement Qualifier
Phenomena	T034	-	MeasurementAnnotation	lbtr	Measurement Qualifier

Fig. 7. Mapping to UMLS semantic groups across NLP systems

3.2.4. NLP- ADAPT

Text notes representing eligibility criteria were processed using the version of NLP-ADAPT for Kubernetes (NLP-ADAPT-kube), which includes the following NLP systems that are compatible with the Unstructured Information Management Architecture (UIMA)[48]: BioMedICUS, CLAMP, cTAKES, and MetaMap (with UIMA adapter). All NLP systems in NLP-ADAPT utilized their default pipelines to extract entities mapped to UMLS semantic types. To minimize false positives, as determined by our prior experience with these systems, we used 800 as the threshold for MetaMap's evaluation score. MetaMap outputs all entity mapping candidates with corresponding mapping scores in a range of 0-1000 (where 1000 indicates a complete mapping) [49]. Annotations produced by NLP-ADAPT-kube were extracted using dkpro-cassis, a software library developed by the Technische Universität Darmstadt[50]. The following versions of the UMLS, by system, were used: 2019AB by MetaMap; 2016AB by cTAKES; 2016AA by BioMedICUS; 2014AB by CLAMP.

NLP-Ensemble-Explorer[33] was used to create ensembles and evaluate individual systems and their ensembles on the task of Named Entity Recognition (NER) of text spans representing UMLS concepts across the DS clinical trial corpus. Individual systems and their ensembles were evaluated using standard performance measures of precision, recall and F1-score. NLP-Ensemble-Explorer takes comprehensive lists of all permutations for NLP systems as input and transforms these into an exhaustive set of Boolean combinations using the logical ∨ operator - to represent a UNION set operation (or ∪); and the logical ∧ operator - to represent an INTERSECTION set operation (or ∩). NLP-Ensemble-Explorer then evaluates Boolean combinations by creating a merged set of system annotations to

assess performance against gold standard annotations. Once a Boolean expression is generated it is stored and evaluated as a binary tree using the parse tree algorithms provided by Miller and Ranum[51]. NLP-Ensemble-Explorer uses character-level binary i-o classification on the positive label (labeled as 1 and 0, respectively) to determine whether there is overlap between an annotated span in the system, merged span set and gold standard span set for each document[52]. We used a character-level partial matching scheme to adjust the weight based on the length of the match, in order to appropriately weight matches to the number of characters in overlap.

3.3. Results

3.3.1. Entities and attributes in DS clinical trials

The distribution of entities and attributes in the DS clinical trials is shown in Fig. 8. Out of the annotated 150 trials, *Condition* entity (mapped to *Disorders* in Fig. 7.) was the largest (1,832 terms) followed by *Qualifier* (1,137 terms), *Drug* (890 terms) and *Observation* (868 terms) while *Device* was the smallest (37 terms).

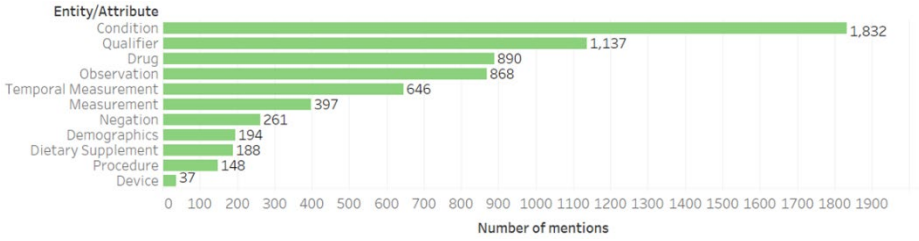

Fig. 8. Distribution of entities and attributes in DS clinical trials

3.3.2. Performances of individual NLP systems and boolean ensemble

As individual NLP systems were trained on different datasets and had distinct strengths and weaknesses, the top performers for one semantic group often struggled in other areas. Among the individual NLP systems, cTAKES was the highest performing system for

Chemicals and Drugs and *Disorders* semantic groups and BioMedICUS was the highest performing system for *Procedures*, *Living Beings, Concepts and Ideas,* and *Devices*.

Table 4. NLP-ADAPT individual system and Boolean combinations performance for NER; measures are **p**=precision, **r**=recall, **F1**=F1-score; systems are A: BioMedICUS, B: CLAMP, C: cTAKES and D: MetaMap. The first part of the table contains individual NLP system performance and the second part of the table contains only the highest performing Boolean combination in each semantic group.

Individual NLP system performance for NER												
Corpus	**BioMedICUS**			**CLAMP**			**cTAKES**			**MetaMap**		
Semantic Group	**p**	**r**	**F1**	**p**	**r**	**F1**	**p**	**r**	**F1**	**p**	**r**	**F1**
Chemicals and Drugs	0.542	0.528	0.535	0.638	0.476	0.545	0.518	0.591	**0.552**	0.464	0.345	0.396
Disorders	0.641	0.532	0.582	0.617	0.603	0.610	0.687	0.565	**0.620**	0.619	0.429	0.507
Procedures	0.292	0.267	**0.279**	0.237	0.328	0.275	0.293	0.184	0.226	0.209	0.172	0.188
Living Beings	0.033	0.051	**0.040**	-	-	-	-	-	-	0.021	0.040	0.027
Concepts and Ideas	0.605	0.229	**0.332**	-	-	-	0.277	0.001	0.001	0.306	0.186	0.231
Devices	0.278	0.483	**0.353**	-	-	-	-	-	-	0.247	0.386	0.301
Phenomena	0.109	0.006	0.011	-	-	-	0.377	0.019	0.036	0.100	0.007	**0.014**
Micro-average	0.528	0.354	0.424	0.508	0.520	0.514	0.578	0.315	0.408	0.426	0.270	0.331

Boolean combination performance for NER												
Corpus	**Highest F1-score**				**Highest precision**				**Highest recall**			
Semantic Group	**combination**	**p**	**r**	**F1**	**combination**	**p**	**r**	**F1**	**combination**	**p**	**r**	**F1**
Chemicals and Drugs	((A∧C)∨B)	0.583	0.576	0.579	(((A∧B)∧C)∧D)	0.704	0.186	0.294	(((A∨B)∨C)∨D)	0.457	0.662	0.540
Disorders	(((A∨B)∧D)∨C)	0.669	0.608	0.637	(((A∨D)∧B)∧C)	0.841	0.249	0.384	(((A∨B)∨C)∨D)	0.523	0.746	0.615
Procedures	(((C∧D)∨A)∨B)	0.228	0.399	0.290	(((C∧D)∨A)∧B)	0.370	0.135	0.198	(((A∨B)∨C)∨D)	0.209	0.422	0.279
Living Beings	(A∧D)	0.041	0.040	0.041	(A∧D)	0.041	0.040	0.041	(A∨D)	0.021	0.051	0.029
Concepts and Ideas	((A∨C)∨D)	0.404	0.355	0.378	((C∧D)∨A)	0.605	0.229	0.332	((A∨C)∨D)	0.404	0.355	0.378
Devices	(A∨D)	0.238	0.524	0.328	(A∧D)	0.294	0.307	0.301	(A∨D)	0.238	0.524	0.328
Phenomena	((A∨C)∨D)	0.188	0.029	0.050	((A∧D)∨C)	0.377	0.019	0.037	((A∨C)∨D)	0.188	0.029	0.050

We evaluated Boolean combinations of all 4 systems on the DS corpus for all semantic groups in our mapping with the exceptions. For *Concepts and Ideas*, and *Phenomena*, only 3 systems were evaluated, because CLAMP does not annotate for these semantic groups, whereas for *Living Beings*, and *Devices*, only 2 systems were evaluated, because CLAMP and cTAKES does not annotate for these groups. The total number of potential Boolean combinations when combining 4 systems was 238; while for 3 systems it was 28; and for 2 systems it was 4^{53}. The highest performing Boolean combination in each semantic group were shown in Table. 4. For example, Boolean combinations like ((BioMedICUS ∧ cTAKES) ∨ CLAMP) and (((BioMedICUS ∨ CLAMP) ∧ MetaMap) ∨ cTAKES) have

higher F1-scores for the *Chemicals and Drugs* and *Disorders* semantic groups, respectively than any single system.

3.4. Discussion

The four NLP annotator systems used in this study were developed with different datasets. Both BioMedICUS and cTAKES utilize pipelines that were developed on the MiPACQ corpus, which consists of fully anonymized outpatient clinical narratives[54,55]. CLAMP's pipeline has elements that were trained on the 2010 VA challenge i2b2 corpus, which consists of hospital discharge summaries and progress notes taken from multiple independent institutions[30]. MetaMap was designed for extracting text from biomedical literature[56].

Because of this, each individual system has its own internal strengths and weaknesses but may improve performance for particular tasks when ensembled with systems that have complementary strengths and weaknesses, as discussed by Derczynski[57]. Thus, ensemble performance of systems would provide increased performance over any one individual system. The DS clinical trial eligibility criteria corpus used in this study is significantly different from any corpora used to develop these annotator systems. The corpus covers criteria related to patient characteristics, disease characteristics, laboratory tests, lifestyle and concurrent therapies which can further be classified according to several dimensions that characterize the content of corresponding eligibility criteria, including but not limited to temporal status, time independent status, constraint types and subject.

BioMedICUS uses a tiered scoring technique for matching UMLS concepts to phrases by first performing direct dictionary phrase matches, second by lower-cased dictionary phrase

matches, and lastly using a discontinuous bag of SPECIALIST normalized terms matches. cTAKES matches UMLS concepts to phrases, by each phrase's lexical and non-lexical permutations and variations against concepts in a dictionary and a list of maintained terms[25]. CLAMP matches UMLS concepts to phrases using the BM25 algorithm for UMLS lookup to find candidate concepts from the UMLS and then applies RankSVM to rank those candidates, from which the top ranked concept is selected. MetaMap uses a shallow parser to generate candidate phrases then, for each candidate phrase, many lexical variations are generated; finally, each phrase is then assigned a score based on its distance to concepts in the UMLS[58]. For this study, we did not use word sense disambiguation functionality from these systems.

We tested the above-mentioned NLP systems and ensemble methods on clinical trial corpus to examine the differences between systems. We saw improved performance when systems were combined. The performance improvement from particular Boolean ensembles confirms the complementary nature of the individual NLP systems, which we suspect exists due to the previously mentioned differences in development.

In order to evaluate the system annotations, we compared the FPs and FNs for each system against the gold standard annotations. Out of 556 FNs, 28 belong to *Chemicals and Drugs*, *Disorders* and *Procedures* semantic groups whereas 515 belong to *Living Beings*, *Concepts* and *Ideas*, *Devices*, and *Phenomena*. Among the 8622 FPs, 3984, 941, 485 and 3222 belong to BioMedICUS, CLAMP, cTAKES and MetaMap, respectively. We investigated the reasons and found that the systems annotated the text from the sentences which were not manually annotated resulting in high FPs and affecting NER performance. According to our annotation guidelines, we omitted annotating the sentences which in our view cannot

be converted into computable queries. For example, any sentence which is focused on the willingness of the patient, informed consent or in the investigator's opinion.

This study has some limitations. As this was our first attempt at defining the mapping to UMLS semantic types across NLP systems, our choice of mappings might not be optimal. Additionally, since the NLP systems used in this study are by default each configured with different versions of UMLS, it is possible different versions can supplement or hinder the system performance. However, in our study, we observed that the older version of UMLS outperformed the systems which used newer versions in some tasks (e.g., BioMedICUS). We believe inclusion of different versions of the UMLS may be beneficial, since, for example, developers of SemRep continue to use 2006AA, as opposed to newer versions of the UMLS, because there are fewer concepts/synonyms which decreases ambiguity[59]. Furthermore, since systems within NLP-ADAPT are configured with different versions of the UMLS, we believe this exploits complementarity between systems (as discussed by Derczynski), with potential for increased ensemble performance. Thus, complementarity due to UMLS version differences warrants further research. Lastly, we only examined Boolean combination ensembles in this study. NLP-Ensemble=Explorer has an option for generating all combinations of majority-vote ensembles, and will be explored in future work.

All systems used in NLP-ADAPT were based on the UIMA architecture, and as such this could be a potential source of bias in our results. Thus, use of non-UIMA based NLP/IE systems, such as QuickUMLS and SpaCy for use in extraction of UMLS concepts based on our mapping would be worth exploring. Also, customization of the systems in NLP-

ADAPT to use customized dictionaries such as iDISK[60] would be worth exploring, which demonstrated better performance to identify supplement entities compared to UMLS[61].

Our results indicate that currently publicly available traditional biomedical NLP systems do not seem to generalize well beyond the tasks for which they were originally designed. These findings are consistent with other previously published results of applying standard NLP tools and their combinations in the domain of pre-hospital trauma notes which also showed limited generalizability[62, 63]. It is possible that the new generation of biomedical NLP tools based on neural network models may help overcome some of these issues; however, given the results obtained so far we believe that heavy domain adaptation is necessary in order to realize the full potential of general-purpose biomedical NLP tools.

3.5. Conclusion

We used NLP-ADAPT which is configured with NLP systems and ensemble methods to extract data elements from the unstructured DS clinical trial eligibility criteria. Results indicated that the ensemble of NLP systems can improve NER performance of each individual system, thus setting a baseline that can be potentially improved with modifications to the NLP-ADAPT pipeline.

4. Bibliography

1. Guyatt GH, Haynes RB, Jaeschke RZ, Cook DJ, Green L, Naylor CD, et al. Users' Guides to the Medical LiteratureXXV. Evidence-Based Medicine: Principles for Applying the Users' Guides to Patient Care. JAMA. 2000 Sep 13;284(10):1290–6.
2. Van Spall HGC, Toren A, Kiss A, Fowler RA. Eligibility Criteria of Randomized Controlled Trials Published in High-Impact General Medical JournalsA Systematic Sampling Review. JAMA. 2007 Mar 21;297(11):1233–40.
3. Wang AY, Lancaster WJ, Wyatt MC, Rasmussen LV, Fort DG, Cimino JJ. Classifying Clinical Trial Eligibility Criteria to Facilitate Phased Cohort Identification Using Clinical Data Repositories. AMIA Annu Symp Proc. 2018 Apr 16;2017:1754–63.
4. Frank G. Current challenges in clinical trial patient recruitment and enrollment. SoCRA SOURCE. 2003 Nov 30;
5. Kadam RA, Borde SU, Madas SA, Salvi SS, Limaye SS. Challenges in recruitment and retention of clinical trial subjects. Perspect Clin Res. 2016;7(3):137–43.
6. McDonald AM, Knight RC, Campbell MK, Entwistle VA, Grant AM, Cook JA, et al. What influences recruitment to randomised controlled trials? A review of trials funded by two UK funding agencies. Trials. 2006 Apr 7;7(1):9.
7. Huang GD, Bull J, Johnston McKee K, Mahon E, Harper B, Roberts JN. Clinical trials recruitment planning: A proposed framework from the Clinical Trials Transformation Initiative. Contemporary Clinical Trials. 2018 Mar 1;66:74–9.
8. Treweek S, Lockhart P, Pitkethly M, Cook JA, Kjeldstrøm M, Johansen M, et al. Methods to improve recruitment to randomised controlled trials: Cochrane systematic review and meta-analysis. BMJ Open. 2013 Jan 1;3(2):e002360.
9. He Z, Tang X, Yang X, Guo Y, George TJ, Charness N, Quan Hem KB, Hogan W, Bian J. Clinical Trial Generalizability Assessment in the Big Data Era: A Review. Clinical and Translational Science. 2020. In press.
10. Embi PJ, Jain A, Clark J, Harris CM. Development of an electronic health record-based Clinical Trial Alert system to enhance recruitment at the point of care. AMIA Annu Symp Proc. 2005;2005:231–5.

11. PCORnet | The National Patient-Centered Clinical Research Network [Internet]. The National Patient-Centered Clinical Research Network. https://pcornet.org/. Accessed March 6, 2020.
12. CTSA Consortium Tackling Clinical Trial Recruitment Roadblocks | National Center for Advancing Translational Sciences. National Center for Advancing Translational Sciences. https://ncats.nih.gov/pubs/features/ctsa-act. Accessed March 6, 2020.
13. i2b2: Informatics for Integrating Biology & the Bedside. I2b2.org. https://www.i2b2.org/. Accessed March 6, 2020.
14. Electronic Medical Records and Genomics (eMERGE) Network. Genome.gov. https://www.genome.gov/Funded-Programs-Projects/Electronic-Medical-Records-and-Genomics-Network-eMERGE. Accessed March 6, 2020.
15. OHDSI – Observational Health Data Sciences and Informatics. Ohdsi.org. https://ohdsi.org/. Accessed March 6, 2020.
16. Luo Z, Miotto R, Weng C. A human–computer collaborative approach to identifying common data elements in clinical trial eligibility criteria. Journal of Biomedical Informatics. 2013 Feb 1;46(1):33–9.
17. Pradhan R, Hoaglin DC, Cornell M, Liu W, Wang V, Yu H. Automatic extraction of quantitative data from ClinicalTrials.gov to conduct meta-analyses. J Clin Epidemiol. 2018/09/23. 2019 Jan;105:92–100.
18. Weng C, Wu X, Luo Z, Boland MR, Theodoratos D, Johnson SB. EliXR: an approach to eligibility criteria extraction and representation. Journal of the American Medical Informatics Association. 2011 Jul 31;18(Supplement_1):i116–24.
19. Kang T, Zhang S, Tang Y, Hruby GW, Rusanov A, Elhadad N, et al. EliIE: An open-source information extraction system for clinical trial eligibility criteria. J Am Med Inform Assoc. 2017 Nov 1;24(6):1062–71.
20. He Z, Chen Z, Bian J. Analysis of Temporal Constraints in Qualitative Eligibility Criteria of Cancer Clinical Studies. Proceedings (IEEE Int Conf Bioinformatics Biomed). 2017/01/19. 2016 Dec;2016:717–22.

21. Si Y, Weng C. An OMOP CDM-Based Relational Database of Clinical Research Eligibility Criteria. Stud Health Technol Inform. 2017;245:950–4.
22. Yuan C, Ryan PB, Ta C, Guo Y, Li Z, Hardin J, et al. Criteria2Query: a natural language interface to clinical databases for cohort definition. J Am Med Inform Assoc. 2019 Apr 1;26(4):294–305.
23. Friedman, C.: Towards a comprehensive medical language processing system: methods and issues. Proc AMIA Annu Fall Symp. 595–599 (1997).
24. Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, et al. CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines. Journal of the American Medical Informatics Association. 2017 Nov 24;25(3):331–6.
25. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010;17(5):507–13.
26. Conway, M., Keyhani, S., Christensen, L., South, B.R., Vali, M., Walter, L.C., Mowery, D.L., Abdelrahman, S., Chapman, W.W.: Moonstone: a novel natural language processing system for inferring social risk from clinical narratives. J Biomed Semantics. 10, (2019). https://doi.org/10.1186/s13326-019-0198-0.
27. Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., Liu, H.: Clinical information extraction applications: A literature review. Journal of Biomedical Informatics. 77, 34–49 (2018). https://doi.org/10.1016/j.jbi.2017.11.011.
28. Friedman, C., Shagina, L., Lussier, Y., Hripcsak, G.: Automated Encoding of Clinical Documents Based on Natural Language Processing. J Am Med Inform Assoc. 11, 392–402 (2004). https://doi.org/10.1197/jamia.M1552.
29. Teije, A. ten, Völker, J., Handschuh, S., Stuckenschmidt, H., d'Acquin, M., Nikolov, A., Aussenac-Gilles, N., Hernandez, N.: Knowledge Engineering and Knowledge Management: 18th International Conference, EKAW 2012, Galway City, Ireland, October 8-12, 2012, Proceedings. Springer (2012).

30. Uzuner, Ö., South, B.R., Shen, S., DuVall, S.L.: 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc. 18, 552–556 (2011). https://doi.org/10.1136/amiajnl-2011-000203.
31. He Z, Rizvi RF, Yang F, Adam TJ, Zhang R. Comparing the study populations in dietary supplement and drug clinical trials for metabolic syndrome and related disorders. AMIA Jt Summits Transl Sci Proc. 2019;2019:799–808.
32. University of Minnesota, NLP/IE. nlp-adapt-kube. 2019, https://github.com/nlpie/nlp-adapt-kube, last accessed 2020/01/06.
33. University of Minnesota, NLP/IE, nlp-ensemble-explorer, UMN NLPIE, 2020, https://github.com/nlpie/ensemble-explorer, last accessed 2020/01/06.
34. Friedman C. A broad-coverage natural language processing system. Proc AMIA Symp. 2000;270–4.
35. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010;17(3):229–36.
36. Denny JC, Irani PR, Wehbe FH, Smithers JD, Spickard A 3rd. The KnowledgeMap project: development of a concept-based medical school curriculum database. AMIA Annu Symp Proc. 2003;2003:195–9.
37. Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991
38. CLAMP | Natural Language Processing (NLP) Software. Clamp.uth.edu. https://clamp.uth.edu/. Accessed March 6, 2020.
39. National Institutes of Health. Clinicaltrials.gov. https://clinicaltrials.gov/. Accessed March 6, 2020.
40. Azam, S.S., Raju, M., Pagidimarri, V., Kasivajjala, V.: Q-Map: Clinical Concept Mining from Clinical Documents. arXiv:1804.11149 [cs]. (2018).
41. McCray, A.T., Burgun, A., Bodenreider, O.: Aggregating UMLS semantic types for reducing conceptual complexity. Stud Health Technol Inform. 84, 216–220 (2001).
42. Semantic Types and Groups, https://metamap.nlm.nih.gov/SemanticTypesAndGroups.shtml, last accessed 2020/05/01.

43. He, Z., Perl, Y., Elhanan, G., Chen, Y., Geller, J., Bian, J.: Auditing the Assignments of Top-Level Semantic Types in the UMLS Semantic Network to UMLS Concepts. Proceedings (IEEE Int Conf Bioinformatics Biomed). 2017, 1262–1269 (2017). https://doi.org/10.1109/BIBM.2017.8217840.
44. University of Minnesota N, biomedicus, 2019, https://github.com/nlpie/biomedicus, last accessed 2020/01/06.
45. University of Texas, UT Health, CLAMP, 2020, https://clamp.uth.edu, last accessed 2020/01/06.
46. Apache Software Foundation, cTAKES, https://ctakes.apache.org, last accessed 2020/01/06.
47. The National Institutes of Health, MetaMap, 2019, https://metamap.nlm.nih.gov, last accessed 2020/01/06.
48. Apache Foundation. UIMA Project, UIMA Proj. 2013, https://uima.apache.org, last accessed 2020/02/08.
49. Aronson, A.R.: MetaMap Evaluation. 2001, https://ii.nlm.nih.gov/Publications/Papers/mm.evaluation.pdf
50. Technische Universität Darmstadt, Ubiquitous Knowledge Processing Lab, dkpro-cassis, 2019, https://github.com/dkpro/dkpro-cassis, last accessed 2020/01/06.
51. Miller BN, Ranum DL. Parse Tree. In: Problem Solving with Algorithms and Data Structures using Python.Section 7.6.https://runestone.academy/runestone/books/published/pythonds/Trees/ParseTree.html, last accessed 2020/01/06.
52. Sang, E.F.T.K., Veenstra, J.: Representing text chunks. In: Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics. pp. 173–179. Association for Computational Linguistics, Bergen, Norway (1999). https://doi.org/10.3115/977035.977059.
53. University of Minnesota, NLP/IE. expected_number_boolean_combinations_n_eq_5.py. expected_number_boolean_combinations_n_eq_5.py. 2020, https://gist.github.com/GregSilverman/3e09cb6b7c7bf664b4df14d309192bb3, last accessed 2020/02/07.

54. Knoll, B.C., Melton, G.B., Liu, H., Xu, H., Pakhomov, S.V.S.: Using synthetic clinical data to train an HMM-based POS tagger. In: 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI). pp. 252–255 (2016). https://doi.org/10.1109/BHI.2016.7455882.
55. Albright, D., Lanfranchi, A., Fredriksen, A., Styler, W.F., Warner, C., Hwang, J.D., Choi, J.D., Dligach, D., Nielsen, R.D., Martin, J., Ward, W., Palmer, M., Savova, G.K.: Towards comprehensive syntactic and semantic annotations of the clinical narrative. J Am Med Inform Assoc. 20, 922–930 (2013). https://doi.org/10.1136/amiajnl-2012-001317.
56. Aronson, A.R.: Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 17–21 (2001).
57. Derczynski, L.: Complementarity, F-score, and NLP Evaluation. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). pp. 261–266. European Language Resources Association (ELRA), Portorož, Slovenia (2016).
58. Aronson, A.R., Lang, F.-M.: An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 17, 229–236 (2010). https://doi.org/10.1136/jamia.2009.002733.
59. Kilicoglu, H., Rosemblat, G., Fiszman, M., Shin, D.: Broad-coverage biomedical relation extraction with SemRep. BMC Bioinformatics. 21, (2020). https://doi.org/10.1186/s12859-020-3517-7.
60. Rizvi, R.F., Vasilakes, J., Adam, T.J., Melton, G.B., Bishop, J.R., Bian, J., Tao, C., Zhang, R.: iDISK: the integrated DIetary Supplements Knowledge base. Journal of the American Medical Informatics Association. 27, 539–548 (2020). https://doi.org/10.1093/jamia/ocz216.
61. Vasilakes, J., Bompelli, A., Bishop, J., Adam, T., Bodenreider, O., Zhang, R.: Assessing the Enrichment of Dietary Supplement Coverage in the UMLS. Journal of American Medical Informatics Association (accepted). (2020).
62. Silverman, G.M., Lindemann, E.A., Rajamani, G., Finzel, R.L., McEwan, R., Knoll, B.C., Pakhomov, S., Melton, G.B., Tignanelli, C.J.: Named Entity

Recognition in Prehospital Trauma Care. Stud Health Technol Inform. 264, 1586–1587 (2019). https://doi.org/10.3233/SHTI190547.

63. Tignanelli, C.J., Silverman, G.M., Lindemann, E.A., Trembley, A.L., Gipson, J.C., Beilman, G., Lyng, J.W., Finzel, R., McEwan, R., Knoll, B.C., Pakhomov, S., Melton, G.B.: Natural language processing of prehospital emergency medical services trauma records allows for automated characterization of treatment appropriateness. J Trauma Acute Care Surg. 88, 607–614 (2020). https://doi.org/10.1097/TA.0000000000002598.