# Information Extraction of Clinical Trial Eligibility Criteria

Yitong Tseo
Facebook Inc.
Menlo Park, USA
yitong@fb.com

M. I. Salkola
Facebook Inc.
Menlo Park, USA
salkola@fb.com

Ahmed Mohamed
Facebook Inc.
Menlo Park, USA
ahmedkm@fb.com

Anuj Kumar
Facebook Inc.
Menlo Park, USA
anujk@fb.com

Freddy Abnousi, MD
Facebook Inc.
Menlo Park, USA
abnousi@fb.com

## ABSTRACT

Clinical trials predicate subject eligibility on a diversity of criteria ranging from patient demographics to food allergies. Trials post their requirements as semantically complex, unstructured free-text. Formalizing trial criteria to a computer-interpretable syntax would facilitate eligibility determination. In this paper, we investigate an information extraction (IE) approach for grounding criteria from trials in ClinicalTrials.gov to a shared knowledge base. We frame the problem as a novel knowledge base population task, and implement a solution combining machine learning and context free grammar (CFG). To our knowledge, this work is the first criteria extraction system to apply attention-based conditional random field architecture for named entity recognition (NER), and word2vec embedding clustering for named entity linking (NEL). We release the resources and core components of our system on GitHub.[1] Finally, we report our per module and end to end performances; we conclude that our system is competitive with Criteria2Query, which we view as the current state-of-the-art in criteria extraction [21].

## CCS CONCEPTS

• **Information systems → Information extraction**; • **Applied computing** → *Health care information systems.*

**ACM Reference Format:**
Yitong Tseo, M. I. Salkola, Ahmed Mohamed, Anuj Kumar, and Freddy Abnousi, MD. 2020. Information Extraction of Clinical Trial Eligibility Criteria. In *Proceedings of KDD '20: Workshop on Applied Data Science for Healthcare (KDD '20).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

Clinical trials are vital for understanding diseases and testing new treatments. However trials in the United States today face significant challenges recruiting enough participants [4][8] and establishing representative diversity in their study populations [10], which regularly leads to difficulty completing trials and generalizing outcomes across populations.

ClinicalTrials.gov is a centralized public database of 330,000+ clinical studies maintained by the National Library of Medicine (NLM) [22]. In addition to hosting every American and many international trials, ClinicalTrials.gov provides filters for a handful of important eligibility criteria such as patient age, gender, trial location, and study condition. Researchers have the option to specify additional, more specific, eligibility criteria such as treatment history and pre-existing conditions. These criteria are written in free-text descriptions, the majority of which include semantically complex language and can require expert domain knowledge to understand [16]. With 32,000+ new trials added annually to ClinicalTrials.gov, automated criteria extraction is a necessary requisite for sophisticated trial discovery and cohort identification platforms.

Previous work on automated criteria extraction take many approaches [1][18]. Systems such as EliXR [19], EliXR-TIME [2], and ERGO [17] build on pattern matching and rules. Other researchers such as Butler et al. [3] and Luo et al. [13] create text mining algorithms to identify common criteria across trials. There has also been significant research focusing on information extraction including Bruijn et al.'s work [6], EliIE [9], and Criteria2Query [21].

In this work, we develop an information extraction approach for eligibility criteria extraction which combines machine learning and context free grammar. Our work makes the following contributions:

- We formulate eligibility criteria extraction as a novel knowledge base population task. Working from this theoretical framework, we achieve 0.753 end to end accuracy.
- To our knowledge, we implement the first attention-based NER for criteria extraction. Our NER detects 10 fine-grained entity classes with precision 0.911 and recall 0.716.
- To our knowledge, we implement the first NEL to leverage embedding clustering for criteria grounding. Our NEL achieves an accuracy of 0.485.
- We open sourced a portion of the end to end implementation described in this paper, and the largest dataset of clinical trial entities & attributes which we are aware of.[1]

## 2 DATASET DESCRIPTION

We present a new dataset of 121,221 clinical entities, attributes, and limits taken from 3,314 trials randomly sampled across all disease and treatment areas in ClinicalTrials.gov. Labels were double-annotated by independent layman reviewers with disagreements settled by a senior adjudicator. We define entities as non-parametric patient properties, attributes as numerical/ordinal properties, and limits as constraints on attributes. We believe this dataset represents the largest of its kind; its distribution is shown in Table 1.

[1] https://github.com/facebookresearch/Clinical-Trial-Parser

*KDD '20, August 23–27, 2020, San Diego, CA*
2020. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

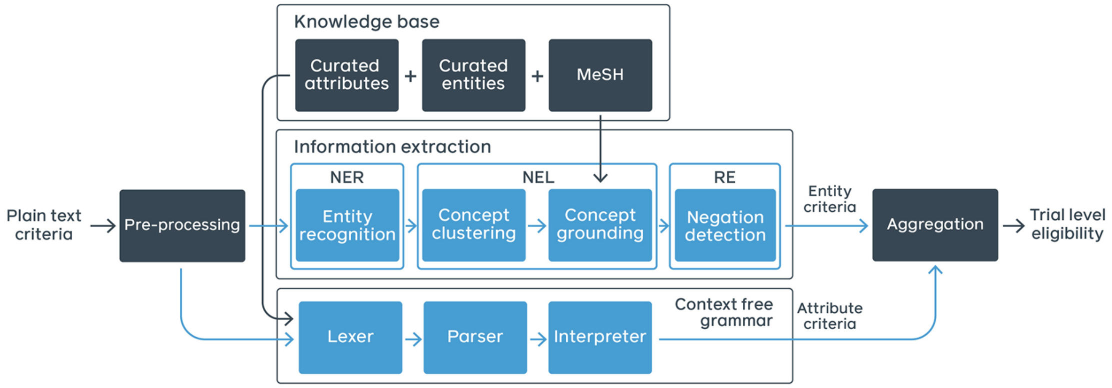

**Figure 1: System architecture.**

| | Class | Count | Examples |
|---|---|---|---|
| Entity | Treatment | 31K | surgery, remdesivir |
| | Chronic disease | 26K | kidney failure, AD |
| | Cancer | 9.3K | leukemia |
| | Gender | 3.7K | — |
| | Pregnancy | 2.8K | — |
| | Allergy | 1.9K | allergy to aspirin |
| | Contraception consent | 1.6K | — |
| | Language literacy | 482 | — |
| | Technology access | 132 | email, cellphone |
| | Ethnicity | 82 | — |
| Attribute | Clinical variable | 13K | ECOG, Hgb count |
| | Age | 2.6K | — |
| | Body mass index | 289 | — |
| Limit | Upper bound | 14K | $< 25\ kg/m^2$ |
| | Lower bound | 14K | $\geq 18\ years$ |

**Table 1: Distribution of entity, attribute, and limit classes.**

## 3 METHOD

We divide criteria extraction into two modules: (1) a classic IE pipeline to extract entity criteria and (2) a CFG engine to extract attribute criteria; see Figure 1.

*3.0.1 Criteria Definition.* Given a trial $T$ in ClinicalTrials.org, its criteria text can be split into inclusion criteria blocks ($IC_T$) and exclusion criteria blocks ($EC_T$). $IC_T$ consists of inclusion criteria ($i \in IC_T$) all of which a participant must satisfy in order to be eligible for $T$. $EC_T$ consists of exclusion criteria ($e \in EC_T$) all of which a participant must **not** satisfy in order to be eligible for $T$. For the purpose of this paper it is assumed all criteria ($c \in IC_T \cup EC_T$) are logically simple. Here a logically simple criterion is defined as a constraint on a single entity or attribute. We formalize the trial-level eligibility for $T$ ($A_T$) as:

$$A_T = IC_T \wedge !EC_T = \left(\bigcap_{x=1}^{M} i_x\right) \wedge ! \left(\bigcup_{x=M+1}^{N} e_x\right) = \bigcap_{x=1}^{N} f(x)$$
$$f(x) = \begin{cases} i_x, & 1 \leq x \leq M \\ !e_x, & M < x \leq N \end{cases} \qquad (1)$$

*3.0.2 Task Definition.* Extraction of criteria is framed as a classic knowledge base population task. Given unstructured text, the goal is to extract Resource Description Framework (RDF) facts representing individual criterion ($c \in IC_T \cup EC_T$) in the form:

$$(concept, constraint, trial)$$

*concept* is some entity (e.g. 'leukemia') or attribute (e.g. 'BMI') within the knowledge base. *trial* is the unique NCT identifier of a trial $T$ in ClinicalTrials.gov (e.g. 'NCT00097734'). *constraint* is the eligibility requirement $T$ places upon *concept*. Example RDF triplets: ('NCT00097734', 'excludes participants with', 'leukemia'), ('NCT03051984', 'requires participants have $\leq 38kg/m^2$', 'BMI').

### 3.1 Knowledge Base

We leverage Medical Subject Headings (MeSH) as our primary knowledge source. MeSH is an NLM controlled vocabulary originally designed as a taxonomy for biomedical research literature. It was chosen for its simplicity and versatility relative to other common knowledge bases such as the International Classification of Diseases (ICD-9/10) and Systematized Nomenclature of Medicine – Clinical Terms (SNOMED-CT). According to Yao et al.'s analysis, though MeSH (27,000 concepts) is much smaller than ICD-9 (70,000 concepts) and SNOMED-CT (350,000 concepts), it captures the most important disease concepts significantly better relative to both ICD-9 and SNOMED-CT [20].

We rely on MeSH concepts for treatment, disease, cancer, and allergy entities. We supplement 66 curated concepts for the remaining entity classes (pregnancy, contraception consent, etc), and 71 curated attributes (age, BMI, ECOG, platelet count, etc). Our curated entities and attributes are available on GitHub.[1] Entities and attributes are organized in a hierarchical structure from general to specific, and seamlessly combine to create one cohesive knowledge base.

### 3.2 Pre-Processing

Trial eligibility is segmented into inclusion and exclusion blocks using rules operating on headings. The text is delexicalized to mask digits and punctuation, normalized for case, and tokenized. Every line of text is then individually considered by the IE and CFG modules.

| Att-BiLSTM-CRF | |
|---|---|
| **Hyper − param.** | **Value** |
| batch size | 64 |
| clipping | $\tau = 1$ |
| dropout | [0.2, 0.2] |
| char_embed dim | 100 |
| BiLSTM layer | 1 |
| LSTM dim | 128 |
| attn dim | 64 |
| mlp decoder dim | 256 |

| Word2vec | |
|---|---|
| **Hyper − param.** | **Value** |
| model | cbow |
| loss | ns |
| dim | 100 |
| window size | 5 |
| epsilon | $\epsilon = 1.0^{-6}$ |
| learning rate | 0.05 |

**Table 2: Hyper-parameter configurations.**

## 3.3 Entity Criteria Extraction

*3.3.1 Named Entity Recognition (NER).* The goal of NER is to extract all entity mentions from unstructured text and categorize mentions by entity class. We trained an attention-based bidirectional Long Short-Term Memory model with a conditional random field layer (Att-BiLSTM-CRF) in PyText [15] for entity recognition of all 10 entity classes. Our model hyper-parameters can be seen in Table 2. Att-BiLSTM-CRF architecture has been shown as state-of-the-art for the task of chemical and disease NER by Luo et al. [12] and Zhai et al. [23].

*3.3.2 Named Entity Linking (NEL).* The goal of NEL is to link entity mentions with concepts in our knowledge base. We split the task into clustering and grounding. A word2vec model is trained with FastText [14] on the trial descriptions and eligibility criteria of all 300K+ trials present in ClinicalTrials.gov as of May 2019. Our model hyper-parameters can be seen in Table 2. All extracted entity mentions are projected into the embedding space and clustered with DBSCAN. Clusters are then grounded to entities and their synonyms in the knowledge base according to Sørensen-Dice similarity [7].

*3.3.3 Relationship Extraction (RE).* The goal of RE is to identify requirements trials place upon concepts given an unstructured text source. For amenability, we equate the RE task to a simple binary classification of negation detection: Given a trial $T$ and some extracted entity $e$ (e.g. "leukemia"), $T$ must either accept or reject subjects with $e$. This definition of the task does not distinguish temporal requirements: whether $T$ accepts subjects who previously had $e$ or subjects who currently have $e$ is not differentiated.

Medical negation detection is a known problem and solutions usually center around regular expression algorithms [5]. Our negation detection algorithm is no different. It first searches for specific negation keywords per entity class (e.g. "seronegative" for chronic diseases). Then computes negation according to string distance between keyword and entity mention.

## 3.4 Attribute Criteria Extraction

We rely on a context free grammar (CFG) engine to recognize, ground, and predict criteria for attributes in the knowledge base. A custom lexer divides and categorizes criteria into tokens (attribute, unit, comparison, number, negation, end-of-string, unknown, etc). A modified Cocke-Younger-Kasami (CYK) algorithm builds parse trees from the tokens. The interpreter analyzes the parse trees removing duplicates and sub-trees. The remaining trees are then evaluated into RDF triplets.

## 3.5 Aggregation

Following Equation 1, exclusion criteria are cast to inclusion criteria by negation. To negate RDF facts, the constraint is simply inverted (e.g. "can-have" becomes "cannot-have", ">" becomes "≤"). Inclusion criteria are left unchanged. After casting, all entity and attribute criteria are intersected to calculate trial-level eligibility of $T$ ($A_T$).

Some trials include redundant criteria for clarity; such criteria are de-duplicated during aggregation. Contradicting criteria, which are either the result of an error in our system or a semantic error in the raw criteria text, are also removed. General criteria are dropped in the presence of more specific criteria according to entity and attribute hierarchy; for example ('NCT00594516', 'excludes participants with', 'hepatitis') is dropped because ('NCT00594516', 'excludes participants with', 'hepatitis b') is also extracted. As a final step, constraints that contradict the intent of another constraint or the intent of the trial are removed. For example no trial should require participants to both be pregnant and on birth control.

# 4 RESULTS & DISCUSSION

Table 3 shows the performance of our system as evaluated on the 10 trial golden set created by Yuan et al. for Criteria2Query [21]. We release the evaluations from our internal build as ancillary files.[2]

*4.0.1 Entity Recognition.* Our NER model employs an Att-BiLSTM-CRF architecture trained with data from 3,314 randomly sampled trials. It predicts 10 fine-grained entity classes shown in Table 1 at 0.802 F1 score. Criteria2Query's NER model employs a classic CRF architecture trained with data from 230 Alzheimer's disease trials. It predicts 5 general entity classes (Condition, Drug, Measurement, Procedure, Observation) at 0.804 F1 score [21]. Our NER model extracts more fine-grained entities while maintaining a competitive F1 score. The Att-BiLSTM-CRF architecture has been successfully applied to many medical applications [12][23]. To our knowledge, this work is the first to apply Att-BiLSTM-CRF, or any other attention-based architecture, to clinical trial criteria extraction. Investigating other promising architectures such as BioBERT [11] for clinical trial NER is a direction for future work.

*4.0.2 Entity Linking.* Our NEL accuracy of 0.485 (82/169) slightly outperforms the 0.447 (51/114) accuracy of Criteria2Query's NEL module. Criteria2Query was intended as a companion tool for trial investigators. For NEL, it relies on predefined concept sets augmented by a sophisticated interface for creating new custom concept sets [21]. Our NEL took a fundamentally different approach. We trained a word2vec model from trial descriptions which we then used to project, cluster, and ground entity embeddings to our knowledge base. In this way our NEL construction was self-supervised, and can be easily configured to ground any knowledge base.

From an absolute perspective, NEL is the bottleneck of our pipeline. It fails to ground 0.416 (64/154) of valid extractions. Of incorrect groundings, 0.696 (16/23) are from over generalization (e.g. "left main stem stenosis" is incorrectly grounded to "stenosis"). We believe these limitations are the consequence of our knowledge base, composed primarily of MeSH concepts. MeSH is comprised of orders of magnitude fewer and more general concepts relative to other standards such as SNOMED-CT and ICD-9/10 [20]. In the task

[2] https://arxiv.org/abs/2006.07296

| | Our System | | | Criteria2Query | | |
|---|---|---|---|---|---|---|
| | **Precision** | **Recall** | **F1** | **Precision** | **Recall** | **F1** |
| Entity Recognition | 0.911 (154/169)<br>[0.864-0.953] | 0.716 (154/215)<br>[0.656-0.772] | 0.802<br>[0.754-0.837] | 0.902 (156/173)<br>[0.844-0.936] | 0.726 (156/215)<br>[0.661-0.777] | 0.804<br>[0.760-0.841] |
| | **Accuracy** | | | **Accuracy** | | |
| Entity Linking | 0.485 (82/169) [0.408-0.556] | | | 0.447 (51/114) [0.351-0.535] | | |
| Attribute Linking | 0.750 (15/20) [0.450-0.850]$^{\star}$ | | | 0.800 (16/20) [0.500-0.900] | | |
| Relationship Extraction | 0.838 (57/68) [0.750-0.926] | | | - | | |
| End to End Performance | 0.753 (64/85) [0.661-0.844] | | | - | | |

**Table 3: Evaluation on Criteria2Query's golden set with 95% confidence interval.**
**$^{\star}$Evaluation of this metric is inferred; refer to section 4.0.3 for more detail.**

of trial criteria extraction, requirements can be exceedingly specific [16]. To accurately link entities requires a comprehensive knowledge base of equivalently specific concepts. We leave experimentation with specialized and expanded knowledge bases (e.g. curated breast cancer trial requirements, SNOMED-CT) for future work.

*4.0.3 Attribute Linking.* Attribute recognition, linking and constraint extraction are performed in tandem by our CFG engine. Evaluated on the golden set, the end to end precision of our attribute criteria extraction is 0.938 (15/16). Criteria2Query similarly relies upon rules for attribute extraction, but does not report its end to end accuracy, only its linking accuracy of 0.800 (16/20) [21]. For the purpose of comparison, we infer the attribute linking accuracy of our CFG engine to be 0.750 (15/20).

*4.0.4 Relationship Extraction.* We achieve an accuracy of 0.838 (57/68) for our rule-based negation detection module. There is no counterpart to our RE module in Criteria2Query. Criteria2Query frames its RE task to infer relations between entities and attributes, rather than concepts and trials.

The expressiveness of our framework is constrained by the simplification of RE to just binary negation detection, and our assumption that all criteria are logically simple. Collectively, 0.636 (7/11) of RE errors could be avoided by expanding the relation ontology to capture a wider array of requirements (e.g. history of disease), and handling logically complex criteria (e.g. conditional and compound constraints).

*4.0.5 End to End Performance.* Our end to end accuracy on Criteria2Query's golden set is 0.753 (64/85) as evaluated by a medical professional. Criteria2Query does not report its end to end accuracy for comparison. However, EliIE, the precursor to Criteria2Query, does report its end to end accuracy of 0.71 [9].

## 5 CONCLUSION

We present a novel formulation of clinical trial eligibility criteria extraction as a knowledge base population task. As far as we are aware, this work is the first to apply attention-based architecture to clinical trial entity extraction, and word2vec embedding clustering to clinical trial entity linking. We open source a library containing our training data, CFG, embeddings, and NER model binary. We evaluate our system against Yuan et al.'s Criteria2Query pipeline [21], which we consider the state-of-the-art, to demonstrate the competitiveness of our system.

## REFERENCES

[1] S. G. Alves, J. S. Costa, and J. Bernardino. Information extraction applications for clinical trials: A survey. *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–6, 2019.

[2] M. R. Boland, S. W. Tu, S. Carini, I. Sim, and C. Weng. EliXR-TIME: A Temporal Knowledge Representation for Clinical Research Eligibility Criteria. *AMIA Jt Summits Transl Sci Proc*, 2012:71–80, 2012.

[3] A. Butler, W. Wei, C. Yuan, T. Kang, Y. Si, and C. Weng. The data gap in EHR for clinical research eligibility screening. *AMIA Jt Summits Transl Sci Proc*, 2018.

[4] B. Carlisle, J. Kimmelman, T. Ramsay, and N. MacKinnon. Unsuccessful trial accrual and human subjects protections: an empirical analysis of recently closed trials. *Clin Trials*, 12(1):77–83, Feb 2015.

[5] W. Chapman, W. Bridewell, P. Hanbury, G. F. Cooper, and B. G. Buchanan. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *Journal of Biomedical Informatics*, 34(5):301–310, Oct 2001.

[6] B. de Bruijn, S. Carini, S. Kiritchenko, J. Martin, and I. Sim. Automated information extraction of key trial design elements from clinical trial publications. *AMIA Annu Symp Proc*, pages 141–145, 2008.

[7] L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[8] S. Gubar. The Need for Clinical Trial Navigators. *The New York Times*, Jun 2019.

[9] T. Kang, S. Zhang, Y. Tang, G. Hruby, A. Rusanov, N. Elhadad, and C. Weng. EliIE: An open-source information extraction system for clinical trial eligibility criteria. *JAMIA*, 24(6):1062–1071, Nov 2017.

[10] T. Knepper and H. McLeod. When will clinical trials finally reflect diversity? *Scientific American*, (557):157–159, May 2018.

[11] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. 2019.

[12] L. Luo, Z. Yang, P. Yang, Y. Zhang, L. Wang, H. Lin, and J. Wang. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition . *Bioinformatics*, 32(8):1381–1388, Apr 2018.

[13] Z. Luo, R. Miotto, and C. Weng. A human-computer collaborative approach to identifying common data elements in clinical trial eligibility criteria. *J Biomed Inform*, 46(1):33–39, Feb 2013.

[14] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[15] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.

[16] J. Ross, S. Tu, S. Carini, and I. Sim. Analysis of eligibility criteria complexity in clinical trials. *Summit Transl Bioinform*, 2010:46–50, Mar 2010.

[17] S. W. Tu, M. Peleg, S. Carini, M. Bobak, J. Ross, D. Rubin, and I. Sim. A practical method for transforming free-text eligibility criteria into computable criteria. *J Biomed Inform*, 44(2):239–250, Apr 2011.

[18] C. Weng, S. W. Tu, I. Sim, and R. Richesson. Formal representation of eligibility criteria: a literature review. *J Biomed Inform*, 43(3):451–467, Jun 2010.

[19] C. Weng, X. Wu, Z. Luo, M. R. Boland, D. Theodoratos, and S. B. Johnson. EliXR: an approach to eligibility criteria extraction and representation. *J Am Med Inform Assoc*, 18 Suppl 1:i116–124, Dec 2011.

[20] L. Yao, A. Divoli, I. Mayzus, J. A. Evans, and A. Rzhetsky. Benchmarking ontologies: bigger or better? *PLoS Comput Biol*, 7(1), 2011.

[21] C. Yuan, P. B. Ryan, C. Ta, Y. Guo, Z. Li, J. Hardin, R. Makadia, P. Jin, N. Shang, T. Kang, and C. Weng. Criteria2Query: a natural language interface to clinical databases for cohort definition. *JAMIA*, 26(4):294–305, Apr 2019.

[22] D. A. Zarin, T. Tse, R. J. Williams, R. M. Califf, and N. C. Ide. The ClinicalTrials.gov results database–update and key issues. *N Engl J Med*, 364(9):852–860, 2011.

[23] Z. Zhai, D. Q. Nguyen, and K. Verspoor. Comparing CNN and LSTM character-level embeddings in BiLSTM-CRF models for chemical and disease named entity recognition. *arxiv*, Aug 2018.