Annotated and Normalized ADR Causal Relation Extraction Corpus for Improving Health Informatics

Samridhi Dev
Jawaharlal Nehru University
New Delhi
Samridhi9927@gmail.com

Aditi Sharan
Jawaharlal Nehru University
New Delhi
aditisharan@gmail.com

Abstract

In the ever-expanding landscape of biomedical research, development of new cancer drugs has increased the likelihood of adverse drug reactions (ADRs). However, information about these ADRs is often buried in unstructured data, requiring the conversion of this data into a structured and labeled dataset to identify potential ADRs and associations between them, making the extraction of entities and the analysis of causal relations a pivotal task. Machine learning methods have been used to identify ADRs, but current literature has several gaps in coverage, superficial manual annotation, and a lack of a labeled ADR corpus specific to cancer and normalized entities. Current datasets are generated manually on the abstracts, limiting their scope. To address these limitations, the paper presents an algorithm that automatically constructs, annotates, normalizes entities specific to cancer and identifies causal relationships among entities using linguistics and grammatical properties, MetaMap and UMLS tools enabling efficient information retrieval. A further knowledge graph was created for a case report to visualize the causal relationships.

1 Introduction

In the dynamic landscape of healthcare, Adverse Drug Reactions continue to be a challenge especially given the increasing complexity of therapeutics, terminologies, and the growth of unstructured medical literature. Significant amount of medical literature encoded in natural language, is available in the public domain through PubMed, including biomedical literature, discharge summaries, clinical trials, health reviews, electronic health records, and others. However, the main obstacle is the lack of tools and the difficulty in parsing and extracting relevant and useful information related to ADR from the available data due to its unstructured and ambiguous nature which impedes the creation of a tools and approaches for creating a structured, unambiguous, annotated dataset for machine learning algorithms for entity and causal relation extraction. Some of the available datasets for disease are NCBI(Doğan et al., 2014) and RareDis(Martínez-deMiguel et al., 2022), while MIMIC-II(Lee et al., 2011), NYPH(Duan et al., 2012), and GE HER(Harpaz et al., 2012) datasets consist of clinical observations and discharge summaries. The Stockholm EPR(Dalianis et al., 2012) Corpus and OMOP(Hripcsak et al., 2015) provide drug exposures and patient data. To the best of our knowledge, no automatically annotated dataset contains entities specific to cancer drug and their side effects. The proposed dataset has been developed and annotated automatically and contains four entities: disease, dysn, side effects, and drug along with entity CUI, normalized entity, and entity type. In succession, causal relationships were elicited among these entities. This step is essential for constructing structured knowledge repositories, and enabling efficient information retrieval. (Figure 1) shows the workflow of proposed method. Section 2 elucidates the proposed methodology and algorithm. Section 3 expounds the results and Section 4 evinces the conclusion and future scope.

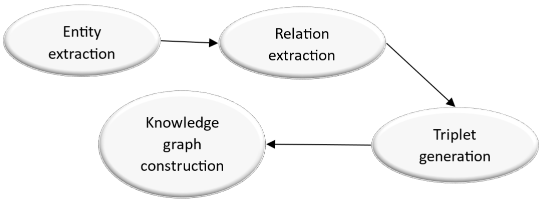


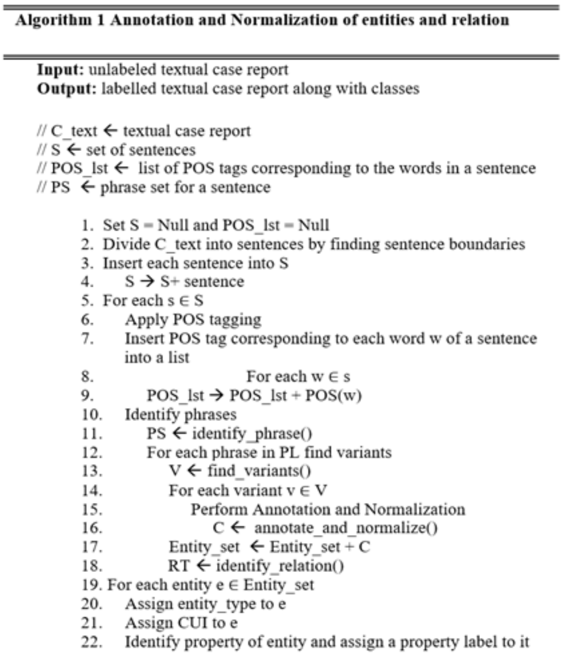
Figure 1: Proposed work flow

2 Proposed Work

This study presents the algorithm proposed for the automatic construction of a labeled semantic ADR dataset specifically to detect adverse drug reactions during treatment of cancer and its types. We observed that most of the datasets available for ADR prediction are based on the abstracts of PubMed articles, a part of which are unlabeled and unstructured and requires expert knowledge, which limits their coverage of entities and makes it difficult to extract latent information. These limitations were addressed by the proposed algorithm (Figure 2) discussed briefly in this section.

Defining entities of interest lays the foundation as it define the scope of the dataset. For our corpus, four types of entities were defined (Table 1): cancer types, drugs, dysn, and Adverse Drug reaction. There is occasionally a chance that a certain condition might be caused by another disease, by ingesting any substance, or by any medication as a side effect. In the suggested dataset, this particular condition is categorised as a Dysn.

Proposed algorithm can be divided into four parts where first part find phrases in every sentence by calling 'identify_phrase()' function. Now for each phrase, in second part, all possible variants are generated and best variant 'v' based on the threshold mechanism is selected by calling the 'find_variants()' function. In third part, annotation and normalization is performed by calling the 'annotate_and_normalize()' function. After annotation, relations are extracted in the form of triplets by calling 'identify_relation()' function followed by assigning CUI and property to the entity in fourth part of algorithm.



Algorithm 1 Annotation and Normalization of entities and relation

```
Input: unlabeled textual case report
Output: labelled textual case report along with classes

// C_text ← textual case report
// S ← set of sentences
// POS_lst ← list of POS tags corresponding to the words in a sentence
// PS ← phrase set for a sentence

Set S = Null and POS_lst = Null
Divide C_text into sentences by finding sentence boundaries
Insert each sentence into S
    S → S+ sentence
For each s ∈ S
    Apply POS tagging
    Insert POS tag corresponding to each word w of a sentence
    into a list
                    For each w ∈ s
        POS_lst → POS_lst + POS(w)
    Identify phrases
        PS ← identify_phrase()
        For each phrase in PL find variants
            V ← find_variants()
            For each variant v ∈ V
                Perform Annotation and Normalization
                    C ← annotate_and_normalize()
            Entity_set ← Entity_set + C
            RT ← identify_relation()
For each entity e ∈ Entity_set
    Assign entity_type to e
    Assign CUI to e
    Identify property of entity and assign a property label to it
```

Figure 2: Proposed algorithm

Entities of interest	Label
Cancer types	Neop
Drugs	drug
Adverse drug reaction	Adr
Dysn	Dysn

Table 1: Entities of interest.

2.1 Raw Data Collection

A good amount of primary data sources is available for collecting raw data like survey data, clinical trials, medical literature, case reports, product labeling, social media data, review data, and electronic health records commonly known as EHRs. Social media data is vulnerable to inaccuracies. Case reports were chosen as a raw data because in comparison to all, case reports have high sensitivity for detecting ADRs and high coverage of cancer types and drugs used in treatment for cancer along with its associated side effects. 500 case reports were chosen as a raw dataset.

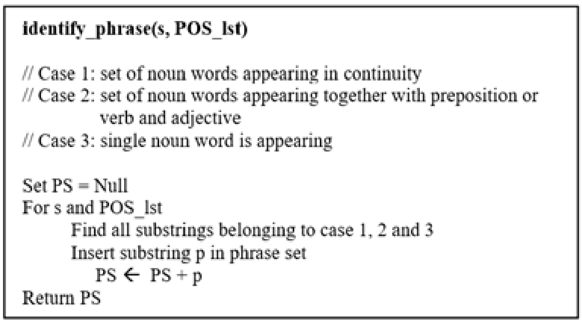
2.2 Phrase Generation

Annotation necessitates the identification of initial phrases containing the entity. Entity identification for dataset construction is based on the assumption that the word which will be considered as the entity must be a noun. For phrase identification three cases have been considered and explained below and MedPost tagger (Smith et al., 2004) has been employed to find POS for each word for phrase generation(Figure 3).

Case 1: During POS tagging, when multiple nouns appear together in a continuity. It will be considered as a single phrase having a high probability of occurrence of multiple word entities.

Case 2: While implementing POS tagging, when multiple nouns appear together along with prepositions, adjectives or verbs it will also be considered as a phrase. For example, “pain in chest” is a group of two noun words with one preposition in between.

Case 3:In most of the cases during POS tagging, a single noun word is considered as a phrase for entity annotation. Consider an example sentence 'mild fever in the evening' where "fever" is a single noun word considered as a phrase.



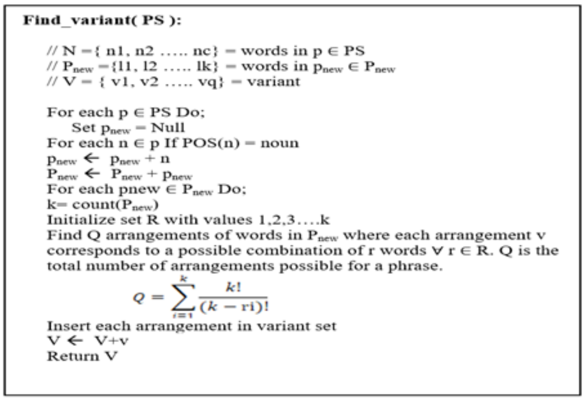
```
identify_phrase(s, POS_lst)

// Case 1: set of noun words appearing in continuity
// Case 2: set of noun words appearing together with preposition or
           verb and adjective
// Case 3: single noun word is appearing

Set PS = Null
For s and POS_lst
      Find all substrings belonging to case 1, 2 and 3
      Insert substring p in phrase set
         PS ← PS + p
Return PS
```

Figure 3: Phrase identification

2.3 Annotation and Normalization

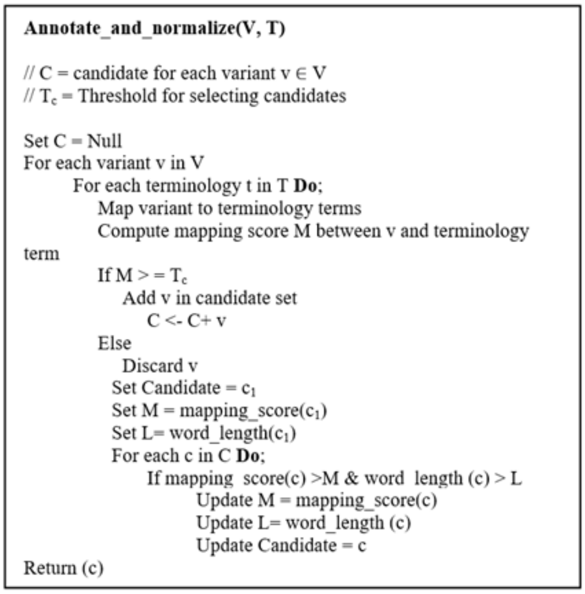
To increase the efficiency of entity annotation and normalization and to maximize the number of named entities that can be mapped to medical terminologies, variants are generated since this helps to ensure that the corpus annotations are as useful as possible (Figure 4). Initially, if the phrase contains any biomedical entity, then variants of that phrase are generated. METAMAP tool (Aronson and Lang, 2010) has been used for entity recognition. Variant generation is based on three schemes: string distance, phrase rearrangement, and stop word removal. In a string distance-based scheme, a string is manipulated on the grounds of cohesiveness and minimum edit distance. In the phrase rearrangement scheme, all possible combinations of words in phrases are taken care off, while in stop word removal, all prepositions, adjectives and verbs are removed from phrases. Giving these phrase variants as input for lexical look up from medical terminologies, mapping is performed with the help of the METAMAP tool. During lexical look up, the mapped medical terms, which are considered as a candidate set for each phrase, is generated. Each candidate is ranked or eliminated based on the candidate score. If the candidate score exceeds the defined threshold, these candidates are considered as the set of normalized names for the actual phrase mentioned in the case reports. Threshold value was set to the value on which it gave best results during multiple iterations.

Once the phrases are mapped to their medical terminology terms and their entity type is matched to the entity type of our interest, the phrase is tagged as the entity with its entity type, and the mapped terms are assigned as the normalized terms for the entity. For every entity, a unique concept identifier (CUI) is assigned. We considered SNOMED-CT (Chang and Mostafa, 2021), RXNorm (Liu et al., 2005), NCI (Mailman et al., 2007) and MeSH (Dhammi and Kumar, 2014) as potential normalization resources for cancer types, dysn, drugs related to cancer and their corresponding side effects. In this manner, entities of interest were tagged in case reports (Figure 5).



```
Find_variant( PS ):

   // N ={ n1, n2 ..... nc} = words in p ∈ PS
   // $P_{new}$ ={l1, l2 ..... lk} = words in $p_{new}$ ∈ $P_{new}$
   // V = { v1, v2 ..... vq} = variant

   For each p ∈ PS Do;
      Set $p_{new}$ = Null
   For each n ∈ p If POS(n) = noun
   $p_{new}$ ← $p_{new}$ + n
   $P_{new}$ ← $P_{new}$ + $p_{new}$
   For each pnew ∈ $P_{new}$ Do;
   k= count($P_{new}$)
   Initialize set R with values 1,2,3....k
   Find Q arrangements of words in $P_{new}$ where each arrangement v
   corresponds to a possible combination of r words ∀ r ∈ R. Q is the
   total number of arrangements possible for a phrase.
            $Q = \sum_{i=1}^{k} \frac{k!}{(k - ri)!}$
   Insert each arrangement in variant set
   V ← V+v
   Return V
```

Figure 4: Variant generation

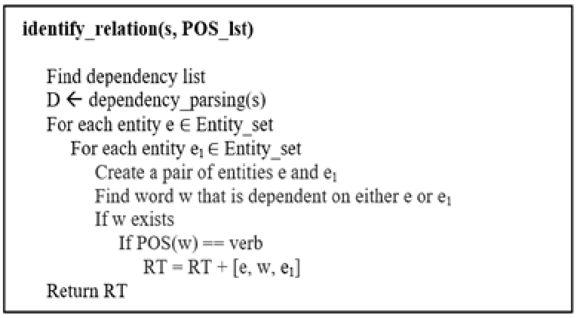


```
Annotate_and_normalize(V, T)

// C = candidate for each variant v ∈ V
// $T_c$ = Threshold for selecting candidates

Set C = Null
For each variant v in V
      For each terminology t in T Do;
         Map variant to terminology terms
         Compute mapping score M between v and terminology
term
         If M > = $T_c$
            Add v in candidate set
               C <- C+ v
         Else
            Discard v
           Set Candidate = $c_1$
           Set M = mapping_score($c_1$)
           Set L= word_length($c_1$)
           For each c in C Do;
              If mapping_score(c) >M & word_length (c) > L
                    Update M = mapping_score(c)
                    Update L= word_length (c)
                    Update Candidate = c
Return (c)
```

Figure 5: Annotation and normalization

2.4 Causal Relation Extraction

In order to effectively identify high-quality relationships from biomedical resources, there exists a process called biomedical causal relation extraction. A causal (cause-effect) relation is an association between two occurrences where the first event must happen before the second in order for the relationship to exist. This paper proposes the usage of linguistics rules and grammatical dependencies for relation extraction (Figure 6). The proposed approach to extract relations employs the stanza dependency parser for extracting dependencies among the words in a sentence. Word dependency is a crucial part as it provides syntactical information about sentence through which semantics of a sentence can be extracted. Relations were extracted in the form of triplets encompassing three elements defined as subject, object and predicate, where subject and object are the entities and predicate indicates the relationship between them. Causal relations are typically more complex and explicitly concerned with cause-and-effect connections. This relationship was taken into consideration while developing the triplet generation rule. The rule specifies that if there is a linking word that is a verb between two entities in a text, then the combination of these two elements in a phrase should be taken into consideration. Then this path should be regarded as a triplet.



```
identify_relation(s, POS_lst)

  Find dependency list
  D ← dependency_parsing(s)
  For each entity e ∈ Entity_set
      For each entity e₁ ∈ Entity_set
          Create a pair of entities e and e₁
          Find word w that is dependent on either e or e₁
          If w exists
              If POS(w) == verb
                  RT = RT + [e, w, e₁]
  Return RT
```

Figure 6: Causal relation extraction

2.5 Entity property assignment

Most of the previous studies on the construction of biomedical corpus are limited to the tagging of mentioned single word entities in a text, due to which useful information remains hidden. In this regard, the proposed dataset construction algorithm facilitates the extraction of hidden information by identifying entities mentioned in different forms. These forms of entity are described by their properties. Entity properties give a clear and precise description of the linguistic formation of entities. These properties require careful handling because they are complex to annotate and are defined in table 2.

2.6 Knowledge Graph Construction

After generating triplets and relation extraction, final step is to create knowledge graph using the generated triplets (Figure 7).

3 Result and Analysis

This paper contributes to medical society by meeting the challenges and research gaps identified in the literature with the aid of natural language processing and providing the algorithm for the automatic construction of semantic, labeled, and normalized dataset by extracting entities and their relations in the form of triplets. Further, as a final point, a knowledge graph was created. Proposed approach for entity extraction was compared with the three state of art approaches: Hunflair, scispacy and Pubtator. Table 3 shows the comparison based on the listed parameters which are different entities having different grammatical properties and associations. State of art models were unable to extract some complex entities, our approach is efficiently able to extract those entities. A knowledge graph of a case report having PMID 22508979 and title "Hand foot syndrome related to chemotherapy" (Qiao and Fang, 2012) was created using the proposed approach and further compared it with the existing knowledge graph creation method scispacy. knowledge graph constructed by the proposed approach encompasses more semantics than the state of art method. Figure 7 and 8 show a knowledge graph created by our approach and scispacy. As there is no benchmark dataset available for validation. Extracted entities or the nodes of the knowledge graph were manually validated by a domain expert.

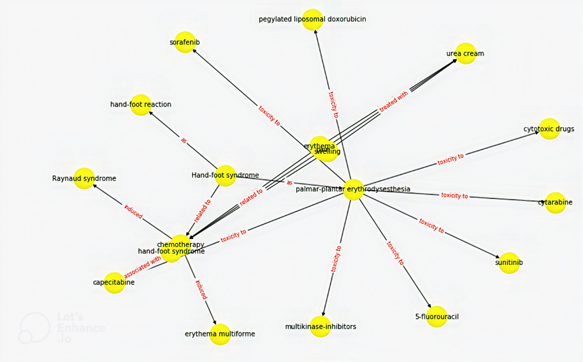


Figure 7: Proposed Causal relation extraction graph

Property	Example	Explanation
Multiple word entity	First-line treatment of myeloid leukaemia	‘myeloid leukaemia’ is a two word entity
Abbreviations and Acronyms	TEN is also known as Lyell’s syndrome and VCZ is a new-generation triazol antifungal agent	‘TEN’ is an abbreviated form of ‘Toxic epidermal necrolysis’ and ‘VCZ’ is an acronym for ‘Voriconazole’
Composite entity mentions	Bleeding from oral, conjunctival and genital mucous membranes	It is a presence of a single entity that leverages multiple entities that are related to each other. Composite entity is converted into single entities: ‘oral bleeding’, ‘conjunctival bleeding’ and ‘genital mucous membrane bleeding’
Presence of modifier before entity	Lady was admitted with a dry cough	Normal entity may change the meaning if it comes before a biomedical entity. Word ‘dry’ modifies the meaning of word ‘cough’
Combined entity	Resolved after switching to lenalidomide-dexamethasone regimen	Extracted entity ‘lenalidomide-dexamethasone’ is a combination of two drugs ‘lenalidomide’ and ‘dexamethasone’
Discontinous enitity	Pain started in hands while walking	Biomedical multi-word entity can be discontinuous. Pain and hand should be perceived as a single entity yet are not continuous.

Table 2: Property definition

Entity|Approach	Proposed entity	Hunflair	Pubtator	Scispacy
Pain in hands	Yes	No	Yes	No
Metastic renal cell carcinoma	Yes	No	Yes	Yes
TMZ	Yes	Yes	Yes	Yes
Lenalidomide - dexamethasone	Yes	No	No	No

Table 3: Comparison table

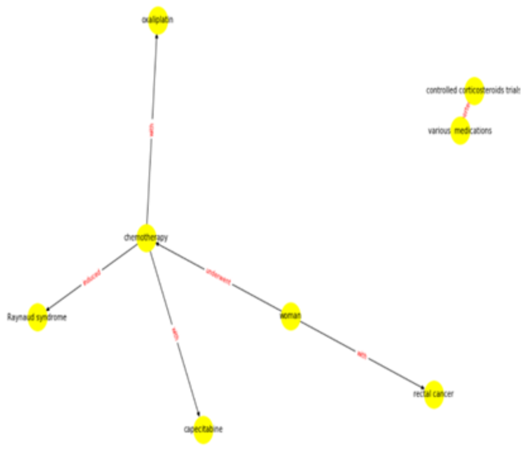


Figure 8: Proposed Causal relation extraction graph

4 Conclusion and future Scope

This study delves into state-of-the-art named entity and relation extraction methods to advance our understanding of the complex biomedical landscape. The proposed dataset can be utilized for model training for ADR prediction specific to cancer. Further knowledge graph was constructed which can assist healthcare professionals in making informed decision and identifying relevant treatments. In future, this corpus can be enlarged by adding more case reports. More entity types can be extracted and corpus can be used for training predictive models.

References

Alan R Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *J. Am. Med. Inform. Assoc.*, 17(3):229–236.

Eunsuk Chang and Javed Mostafa. 2021. The use of SNOMED CT, 2013-2020: a literature review. *J. Am. Med. Inform. Assoc.*, 28(9):2017–2026.

Hercules Dalianis, Martin Hassel, Aron Henriksson, and Maria Skeppstedt. 2012. Stockholm epr corpus: A clinical database used to improve health care. In *Swedish Language Technology Conference*, pages 17–18.

Ish Kumar Dhammi and Sudhir Kumar. 2014. Medical subject headings (MeSH) terms. *Indian J. Orthop.*, 48(5):443–444.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Lian Duan, Mohammad Khoshneshin, W Nick Street, and Mei Liu. 2012. Adverse drug effect detection. *IEEE journal of biomedical and health informatics*, 17(2):305–311.

Rave Harpaz, William DuMouchel, Nigam H Shah, David Madigan, Patrick Ryan, and Carol Friedman. 2012. Novel data-mining methodologies for adverse drug event discovery and analysis. *Clinical Pharmacology & Therapeutics*, 91(6):1010–1021.

George Hripcsak, Jon D Duke, Nigam H Shah, Christian G Reich, Vojtech Huser, Martijn J Schuemie, Marc A Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R Rijnbeek, Johan van der Lei, Nicole Pratt, G Niklas Norén, Yu-Chuan Li, Paul E Stang, David Madigan, and Patrick B Ryan. 2015. Observational health data sciences and informatics (OHDSI): Opportunities for observational researchers. *Stud. Health Technol. Inform.*, 216:574–578.

Joon Lee, Daniel J Scott, Mauricio Villarroel, Gari D Clifford, Mohammed Saeed, and Roger G Mark. 2011. Open-access mimic-ii database for intensive care research. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 8315–8318. IEEE.

Simon Liu, Wei Ma, Robin Moore, Vikraman Ganesan, and Stuart Nelson. 2005. Rxnorm: Prescription for electronic drug information exchange. *IT Professional*, 7(5):17–23.

Matthew D Mailman, Michael Feolo, Yumi Jin, Masato Kimura, Kimberly Tryka, Rinat Bagoutdinov, Luning Hao, Anne Kiang, Justin Paschall, Lon Phan, Natalia Popova, Stephanie Pretel, Lora Ziyabari, Moira Lee, Yu Shao, Zhen Y Wang, Karl Sirotkin, Minghong Ward, Michael Kholodov, Kerry Zbicz, Jeffrey Beck, Michael Kimelman, Sergey Shevelev, Don Preuss, Eugene Yaschenko, Alan Graeff, James Ostell, and Stephen T Sherry. 2007. The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.*, 39(10):1181–1186.

Claudia Martínez-deMiguel, Isabel Segura-Bedmar, Esteban Chacón-Solano, and Sara Guerrero-Aspizua. 2022. The raredis corpus: A corpus annotated with rare diseases, their signs and symptoms. *Journal of Biomedical Informatics*, 125:103961.

Jianjun Qiao and Hong Fang. 2012. Hand-foot syndrome related to chemotherapy. *Canadian Medical Association Journal*, pages cmaj–111309.

Larry Smith, Thomas Rindflesch, and W John Wilbur. 2004. Medpost: a part-of-speech tagger for biomedical text. *Bioinformatics*, 20(14):2320–2321.