Published in final edited form as:
J Biomed Inform. 2014 February ; 47: 1–10. doi:10.1016/j.jbi.2013.12.006.

NCBI Disease Corpus: A Resource for Disease Name Recognition and Concept Normalization

Rezarta Islamaj Doğan[1], **Robert Leaman**[1,2], and **Zhiyong Lu**[1,*]

[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, 20894

[2]Department of Computer Science and Engineering, Arizona State University, USA

Abstract

Information encoded in natural language in biomedical literature publications is only useful if efficient and reliable ways of accessing and analyzing that information are available. Natural language processing and text mining tools are therefore essential for extracting valuable information, however, the development of powerful, highly effective tools to automatically detect central biomedical concepts such as diseases is conditional on the availability of annotated corpora.

This paper presents the disease name and concept annotations of the NCBI disease corpus, a collection of 793 PubMed abstracts fully annotated at the mention and concept level to serve as a research resource for the biomedical natural language processing community. Each PubMed abstract was manually annotated by two annotators with disease mentions and their corresponding concepts in Medical Subject Headings (MeSH®) or Online Mendelian Inheritance in Man (OMIM®). Manual curation was performed using PubTator, which allowed the use of pre-annotations as a pre-step to manual annotations. Fourteen annotators were randomly paired and differing annotations were discussed for reaching a consensus in two annotation phases. In this setting, a high inter-annotator agreement was observed. Finally, all results were checked against annotations of the rest of the corpus to assure corpus-wide consistency.

The public release of the NCBI disease corpus contains 6,892 disease mentions, which are mapped to 790 unique disease concepts. Of these, 88% link to a MeSH identifier, while the rest contain an OMIM identifier. We were able to link 91% of the mentions to a single disease concept, while the rest are described as a combination of concepts. In order to help researchers use the corpus to design and test disease identification methods, we have prepared the corpus as training, testing and development sets. To demonstrate its utility, we conducted a benchmarking experiment where we compared three different knowledge-based disease normalization methods with a best performance in F-measure of 63.7%. These results show that the NCBI disease corpus has the

*Corresponding author: zhiyong.lu@nih.gov.

Authors' contributions

Conceived and designed the experiments: RID ZL. Performed the experiments: RID, RL. Analyzed the data: RID. Wrote the paper: RID, RL, ZL. All authors read and approved the final manuscript.

Availability of supporting data

NCBI disease corpus is publically available at: http://www.ncbi.nlm.nih.gov/CBBresearch/Dogan/DISEASE/. The data file, including all annotations of 793 PubMed abstracts will be available in XML, plain text and in the PubTator-compatible format. The disease corpus website also includes detailed description of the annotation process and the annotation guidelines.

potential to significantly improve the state-of-the-art in disease name recognition and normalization research, by providing a high-quality gold standard thus enabling the development of machine-learning based approaches for such tasks.

Keywords

Disease name recognition; Named entity recognition; Disease name normalization; Corpus annotation; Disease name corpus

1. Background

1.1. The importance of disease name recognition for biomedical research

Disease is one of the fundamental biological entities in biomedical research, one of main research goals at the National Institutes of Health, and as a result, it is frequently searched for in literature [1] and on the Internet [2]. Hence, like other named entity recognition tasks and similar to gene name identification [3, 4], disease name recognition is an important task in biomedical text mining and plays a critical role in accelerating scientific discovery and contributing to improved information access [5].

Automatic disease recognition from free text is a challenging task due to both variation and ambiguity in disease names, as well as its definition (e.g., is *fever* a disease?) [5, 6]. From a taxonomic system point of view -- there is significant ambiguity in what is a disease nam*e: adenomatous polyposis coli* and *Friedrich ataxia* can be both gene and disease names. Abbreviated disease names are common in biomedical literature, and the same abbreviation may stand for many different diseases. For example, *AS* may stand for *Angelman syndrome, ankylosing spondylitis, aortic stenosis, Asperger syndrome, autism spectrum*, etc. Lastly, while it is extremely important to allow doctors and health practitioners the freedom to effectively describe a disease by use of descriptive expressions, this poses a higher level of difficulty for applying automatic identification methods to medical text [7, 8].

Current disease name entity recognition methods generally contain two separate steps: disease mention recognition, followed by disease concept recognition, i.e., the grounding of the pre-identified disease mentions to corresponding standard database identifiers [6, 9]. The second step, concept recognition (also known as normalization), is crucial for specifying the diseases mentioned in text. A PubMed article may reference the same disease concept multiple times, possibly using variations of one or more of its accepted full terms (long form) and also abbreviations (short forms). After normalization, unique disease concepts can be further used for a number of downstream analyses: Providing users with the ability to find a more relevant subset of resources based on user queries; allowing for detecting unique associations between diseases and genes in the literature; etc.

Previous studies on automatic disease mention recognition show that machine-learning based approaches [10, 11] compare favorably to dictionary lookup or rule-based methods. However, to our best knowledge, most approaches on disease normalization are knowledge-based as opposed to learning-based [12, 13], which is related to the lack of sufficient gold-standard training data. We also note that fewer attempts to recognize disease names have been reported compared to the large body of work on gene normalization studies, such as [14–16].

1.2. Constructing high-quality corpora for disease name recognition

Manually annotated high-quality corpora are of utmost importance for the development of sophisticated NLP systems, both as training data and for evaluative purposes. Manually

annotated biomedical corpora have consistently made possible the achievement of improved results in biomedical NLP research, with impressive results being observed in tokenization tasks [17], part-of-speech identification tasks [18, 19], clinical concept identification tasks [20], etc.

To assist the development and evaluation of the disease name recognition task, we have developed a disease corpus, namely the NCBI disease corpus. The construction of the NCBI disease corpus went through two major phases: mention-level annotation and concept-level annotation. As a result, the NCBI disease corpus was manually annotated for every disease mention and its corresponding concept for a total of 793 PubMed abstracts. The mention-level annotation work has been described in [21]. In a nutshell, disease mentions were annotated based on their relevance for biomedical information retrieval tasks that involve diseases. Annotation guidelines allowed flexible matching to UMLS concepts and categorized four annotation categories: Composite mentions, Modifiers, Disease Class mentions and Specific Diseases. The corpus was reviewed several times by several annotators and when used as gold-standard data for a state-of-the-art machine learning system, it was able to significantly improve its performance for disease name recognition [21]. The second phase was the concept-level annotation which completes the NCBI disease corpus as a disease name recognition resource, and releases it to the community in its complete form. The completed NCBI disease corpus has already enabled the creation and evaluation of the first machine learning method for disease normalization, DNorm, which relies on the joint mention-level and concept-level annotations of the corpus to learn term variations directly from the training data [22].

The objective of our work is not only to facilitate information retrieval tasks that involve diseases, but also to facilitate future applications of complex information retrieval tasks connecting diseases to treatments, causes or other types of information, specifically for biomedical literature. Hence, the next important step towards this goal is the entity normalization task that involves mapping mentions to standard database/ontology identifiers. In the present paper, we present our work on developing the NCBI disease corpus to map the individual disease mentions to standard disease controlled vocabularies, namely MeSH (http://www.nlm.nih.gov/mesh/) and OMIM (http://www.ncbi.nlm.nih.gov/omim) identifiers.

Finding a vocabulary for capturing disease data is a very important decision. We wanted a disease vocabulary that was reliable, publicly available, relatively stable, regularly maintained, and suitable for indexing biomedical literature. We were looking for a vocabulary source that was previously used and would naturally be used in the future as an annotation source for biomedical text in order to facilitate interoperability. We focused on MeSH ‘Diseases’ branch, and added the specificity of OMIM identifiers for genetic diseases, as practically combined in MEDIC, the MErged DIsease voCabulary [23]. In essence, MEDIC is a combination of the MeSH ‘disease’ branch and OMIM such that it is both deep and broad (we downloaded MEDIC on April 17, 2012, which contained 9,661 disease concepts, and about 67,000 terms). It also contains significantly fewer concepts/terms compared to the UMLS Metathesaurus [24], which makes it more convenient for real-life manual curation.

1.3. NCBI disease corpus for disease name recognition

The work to develop the NCBI disease corpus is closely related to several other corpus construction efforts in the biomedical domain [25–34]. These studies specifically agree on: 1. the need for multiple experienced annotators; 2. the need for detailed annotation guidelines; and 3. the need for large scale high-quality annotation corpora. The NCBI

disease corpus fulfills all these needs and constitutes therefore a significant resource for disease name entity recognition research in biomedical text.

The NCBI disease corpus also is not the first resource for disease entity recognition in biomedical research. This corpus significantly extends two existing disease corpora [11, 13]. Major differences between the new corpus and the two previous corpora include three aspects (shown in Table 1). First, all sentences in an abstract are selected for annotation in the NCBI disease corpus, which is important to enable higher level text mining tasks that explore relationships between diseases and other types of entities such as genes or drugs in the same abstract. As a result, the NCBI disease corpus is also several times larger than the existing corpora. Second, a novel aspect of our concept annotation lies in the use of the recently developed MEDIC vocabulary for assigning disease concepts [23]. Finally, unlike the annotations in the two previous corpora, each annotation in our corpus was completed by at least two individuals. We believe having multiple annotations helps ensure higher quality results: both more objective and fewer missing ones, which have been recently noted as an issue by researchers developing a disease normalization system [35].

The contributions of this article are: 1. A comprehensive description of the annotation process for constructing the NCBI disease corpus, 2. A thorough exploration of the NCBI disease corpus and its characteristics, 3. An introduction of the NCBI disease corpus as a resource setup to build and evaluate new disease name recognition methods, 4. An initial evaluation of three different disease recognition models and their description, 5. An analysis of the results and future directions for improvement, and finally and most importantly 6. The public release of the NCBI disease corpus to the biomedical community for further use and research. The rest of the article is organized as follows: In the Methods section we describe the corpus and tools used to perform the annotation task, the annotation guidelines, the annotation process, and the measurement tests and the procedure to ensure the quality of resulting annotations. Furthermore, we follow by a description of the three disease name recognition methods that were used to identify disease concepts in the NCBI disease corpus. The Results section lists the inter-annotator agreement results, details and summarizes the NCBI disease corpus characteristics, and then explores and compares the results of three different approaches for disease name recognition on the NCBI disease corpus. In the next section, we discuss the choice of the terminology resource, advantages, disadvantages and limitations of preferring one terminology resource as opposed to others, the challenges of disease name normalization as opposed to other related tasks such as gene name normalization, specifically how these all relate to the NCBI disease corpus, and lastly, we give our concluding remarks in Section 5.

2. Methods

2.1. Data Sources and Preparation

The annotation work on the NCBI disease corpus was performed in two stages, where annotators worked both individually and in pairs to rectify differences in annotation and ensure the construction of a high-quality product. At the beginning, annotators were also provided help in the form of preannotations and an easy-to-use web-based annotation tool PubTator [35, 36], whose versatility in biomedical concept annotation has been demonstrated in the recent BioCreative tasks. The first stage took place in summer 2011 when a group of 12 annotators each annotated an average of 125 PubMed documents, so that for each document we had at least two annotators. This work resulted in a completely annotated disease mentions corpus [21]. Disease mentions were categorized in four categories: Specific Disease or Disease Class, (i.e., <Specific Disease> *Diastrophic dysplasia* </> is an <Disease Class> *autosomal recessive disease* </> characterized by … .) or Composite Mention (i.e., <CompositeMention> *Combined genetic deficiency of C6 and*

C7 </> in man.) or Modifier (i.e., The Israeli <Modifier> *C7-deficient* </> cases all share a C7 haplotype ….). The second stage took place in summer 2012 when a group of 14 annotators (10 overlapped with the first group) annotated 120 PubMed documents each on average, so that for each citation we had at least two annotators. This work improved consistency of the mention level annotations, and added the normalization layer of annotations, so that each disease mention in text is linked to a standardized medical vocabulary entry, namely a MeSH descriptor or OMIM identifier, as listed in the MEDIC vocabulary.

2.1.1. Annotators and the pre-annotation process—The annotator group consisted of 14 people with backgrounds in biomedical informatics research and experience in biomedical text corpus annotation. The 793 PubMed citations were divided into sets of 30 PubMed citations each. Every annotator worked on 4 or 5 sets of 30 PubMed abstracts. The sets were divided randomly among annotators such that each set was shared by at least two people. To avoid annotator bias, pairs of annotators were chosen randomly for each set, so that each pair of annotators overlapped for at most two sets.

According to our own experience [27, 36] and others [37], automatic pre-annotation can help accelerate the manual annotation process. First, each PMID document was pre-annotated using the Inference Method developed for disease name normalization [9], which properly handles abbreviation recognition, robust string matching, etc. As such, human annotators were given the pre-annotated documents as a starting point and allowed to see each pre-annotation with a computed confidence. Annotators were also told about the limitations of the automatic method, such as incomplete results for composite disease mentions.

2.1.2. Annotation guidelines—For disease concept annotation, annotators were encouraged to use their domain knowledge, as well as any other public resources such as UMLS and Wikipedia. Initially, a set of 20 randomly chosen PubMed abstracts was used as a practice set for the development of annotation guidelines. After each annotator worked individually on the set, the results were shared and discussed among all annotators. The final annotation guidelines are summarized below and also made available at the corpus download website.

To explain our annotation rules, we use the disease mentions in the following text (bolded) as a running example: “The **Schwartz-Jampel syndrome** (**SJS**) is a **recessively inherited condition** defined by **myotonia**, **short stature**, and **bone dysplasia.**”

1. Annotate the concept that matches the preferred name.

   There are several matches for “**myotonia**” in MEDIC: Myotonia [MESH:D009222], Myotonia Congenita [MESH:D009224], Myotonic Disorders [MESH:D020967], Myotonic Dystrophy [MESH:D009223], etc. For the mention “**myotonia**”, the entry MESH:D009222 is selected because it provides a perfect match to this concept’s preferred name. Moreover, the descriptions of the other possible matches do not provide a better fit when taking into account the context of the article.

2. Annotate the concept that matches the synonym name, unless there is another concept that matches the preferred name.

   Several MEDIC entries may match the mention “**bone dysplasia**”: Bone Diseases, Developmental [MESH:D001848], Gracile bone dysplasia [MESH:C537291], Fibrous Dysplasia of Bone [MESH:D005357], etc. For the mention “**bone**

dysplasia", the entry MESH:D001848 is selected because it lists a synonym which is a perfect match to this mention.

3. Annotate the most specific concept that correctly describes the disease mention. Certain disease concepts in MeSH are disease classes, which are described with a MeSH entry term and several synonymous terms corresponding to the specific disease names within that class. In such cases, MEDIC lists the correct MeSH ID for the concept as well as a secondary list of OMIM identifiers for the specific diseases within that class. For example, the MEDIC entry for "**Schwartz-Jampel syndrome**" is Osteochondrodysplasias [MeSH: D010009], with a secondary list of specific identifiers: OMIM: 239100, 255800, and 309350. The disease "**Schwartz-Jampel syndrome**" is a synonym for the MeSH entry D010009. A closer consideration of the list of OMIM identifiers reveals that OMIM:239100 corresponds to Van Buchem Disease, OMIM:255800 to Schwartz-Jampel syndrome, and OMIM:30930 to Melnick-Needles Syndrome, which are all synonyms of the MeSH entry term Osteochondrodysplasias. Hence, for the mention "**Schwartz-Jampel syndrome",** the entry OMIM:255800 is selected as the normalized concept because it is the disease concept that correctly describes the mention. In the cases where MEDIC lists one MeSH descriptor and one OMIM identifier, these two identifiers are considered equivalent; therefore either one can be selected.

4. Annotate the closest hypernym concept that logically describes the disease mention.

   Certain disease mentions in text are hard to map to a disease concept in the dictionary. For example, a search in MEDIC for "**recessively inherited condition**" produces no results and other sophisticated string matching approaches report unrelated results. The mention "**recessively inherited condition**" though, is annotated as a disease name in our corpus, and the entry Genetic Diseases, Inborn [MeSH:D030342] is chosen as the correct normalization because it provides the closest hypernym logical concept that correctly describes the disease mention. In this case however, MeSH is not specific enough and does not distinguish inherited diseases as dominant versus recessive, so a less specific mapping is produced. We discuss these cases further in the Discussion session.

5. Annotate all concepts in a *composite disease mention* using the "|" separator.

   Composite mentions such as: "**colorectal, endometrial, and ovarian cancers**" can be normalized to the collection of the individual constituents MeSH: D010051|D016889|D015179. The "|" character is used to compose this mapping, and this signals the presence of a composite mapping.

6. Annotate a disease mention using multiple concepts to logically describe the disease mention, using the "+" concatenator.

   In rare cases, representing a disease mention may require multiple concepts. For example, a mapping can be produced between the mention "**inherited neuromuscular disease**" and the concept **Neuromuscular Disease** [MeSH:D009468]. However, the MeSH entry description of D009468 does not imply that these diseases are inherited. Hence, a combination of identifiers: **Neuromuscular Disease** [MeSH:D009468] + **Genetic Diseases, Inborn** [MeSH:D030342] is a better concept normalization for that mention. The "+" character is used to compose this mapping, and this signals the presence of a multiple concept mapping, not to be confused with a composite mapping.

7. Annotate a disease mention even if that mention is used interchangeably as a gene name.

   If the same mention is used as both a disease name and a gene name, therefore introducing a gene name – disease name ambiguity, then the mention is considered a disease name mention. This particular guideline was motivated by the fact that we wanted to be thorough in our annotation process, and annotate every textual mention that referred to a disease. This meant that sometimes the disease mention was not a noun phrase but a modifier (disease annotation tag: Modifier). A typical example would be: "… the <Modifier> VHL </Modifier> gene…". The ambiguous mention is annotated and normalized to the VHL disease name concept; because VHL is a recognized synonym for the disease term Von Hippel-Lindau, and the phrase "VHL gene" is interpreted as "the gene causing the disease VHL".

8. When necessary, use specific concepts in OMIM not included in MEDIC.

   It is possible that a disease mention may not be covered in MEDIC. This was experienced by our annotators a handful of times. For these cases, we adopted the following solution: if OMIM contained a concept that described that disease mention, we used this identifier accordingly. For example: Leptin deficiency [OMIM:164160] and Complement Component 9 deficiency [OMIM:613825] are annotated in our corpus although they were not included in the Spring 2012 version of MEDIC that was used for this work.

2.1.3. Annotation process and the annotator software—Similar to the steps in disease mention annotation [21] the annotation of disease concepts was also organized in consecutive phases and used PubTator [36, 38] as the web-based annotation tool. Phase I consisted of each pre-annotated abstract in the corpus being read and reviewed by two annotators working independently. Annotators could agree with the pre-annotation, remove it, or modify it. Annotators could also add new mappings. The annotation software automatically updated duplicated mentions within the same document, thus ensuring consistency of annotations within each document. After this round of independent annotation, a summary document was created highlighting the agreement and differences between annotators. Accordingly, inter-annotator agreement was calculated at this step and it is reported in the Results section.

In Phase II, each annotator examined and edited his or her own annotations by reviewing the differences reported in the Phase I summary. The annotation software allowed that each annotator to compare their results for each set with those of other annotators that worked on the same set. Visual cues were implemented to signal agreement and disagreement between annotators at two granularity levels: PMID document level and PMID annotation set level.

At the PMID document level, as illustrated in Figure 1, the annotator sees the title and abstract of the document, with the highlighted disease mentions. Below the text box, a table of all mentions is produced contrasting the annotations between annotators. Each annotator may see the others' selections, but can only edit his/her own work. To facilitate the review process, this view also allows for one-click access to the MEDIC records of the normalized concepts. At the PMID annotation set level, not shown, each annotator sees the list of PMID titles of the shared annotation set. For each PubMed citation, if all annotators agree on all annotations, then that citation is highlighted in the summary view, to signal agreement.

After this round, an additional summary similar to that of Phase I was generated to highlight the remaining differences after Phase II. Then, each pair of annotators organized meetings

where they discussed and resolved their differences. After these meetings, a consensus set of annotations was produced for each PubMed abstract.

The final Phase of the annotation process consisted of the first author going over all annotated documents and ensuring that annotations were consistent across different abstracts and different annotation sets. Lists of problematic annotations were produced and discussed in a final annotators meeting, where all discrepancies were resolved.

2.1.4. Annotation Consistency Evaluation Metrics—We measured the annotators' agreement after Phase I of the annotation process. One way to measure the agreement between two annotators is to measure their observed agreement on the sample of annotated items. We measured agreement by computing the F-measure between each pair of annotators that worked on the same set of documents, as also noted in [26, 39].

Agreement statistics are measured for each annotator pair. First an F-measure agreement average is computed for all annotated mentions in each PMID document in the shared annotation set. Then the average over all PubMed documents of the shared set is computed. Next, for each annotator pair, the average agreement statistic is computed over all annotation sets that they shared.

2.2. Disease normalization methods

As an attempt to demonstrate its utility, we used our corpus as a gold standard for benchmarking three different disease normalization methods. The methods are used at an ab-initio setting, where both disease mention recognition and disease concept normalization is performed. The methods performances are compared after the normalization step. The NCBI disease corpus has also been used for the development of a successful machine learning method for disease name normalization, DNorm.

2.2.1. Dictionary look-up method—The traditional dictionary look-up performs exact matching using terms in the vocabulary. This method matches the disease name as it appears in the terminology and is therefore not robust against term variability that was not foreseen during creation of the lexicon. In addition, precision may be affected by ambiguous or nested terms. To address those issues, we used Norm, from the SPECIALIST lexical tools (http://lexsrv3.nlm.nih.gov/SPECIALIST/index.html) to preprocess all disease names in the MEDIC lexicon including their synonyms. The normalized names and synonyms were then applied to the strings and substrings of the PMID documents in the NCBI disease corpus. When a textual string in the PubMed abstracts in the NCBI testing set was mapped by Norm to a disease mention in the MEDIC lexicon, that disease mention is grounded to the corresponding MEDIC concept. The results of this string matching method are reported as Norm in the Results section.

2.2.2. MetaMap Processing—MetaMap [12] is the state-of-the-art natural language processing tool for identifying UMLS Metathesaurus concepts in biomedical text. MetaMap splits the given input text into sentences, and the set of sentences into phrases. For each phrase, MetaMap identifies possible mappings based on lexical look-up and on variants associating a score to each one of them. MetaMap identifies several possible mappings in each phrase and several candidates for each one.

In this work, the text of PubMed citations in the NCBI disease corpus, title and abstract, was run through the MetaMap Web tool to identify all UMLS concepts in that text. Next, for each PMID, the list of UMLS concept identifiers (CUIs) found by MetaMap was mapped to their corresponding MeSH descriptors and OMIM identifiers. The resulting MeSH and OMIM identifiers were further filtered in two ways:

a. Only concepts associated with the disorder Semantic Group were kept (i.e., concepts linked to these semantic types: Acquired Abnormality, Anatomical Abnormality, Congenital Abnormality, Cell or Molecular Dysfunction, Disease or Syndrome, Experimental Model of Disease, Injury or Poisoning, Mental or Behavioral Dysfunction, Neoplastic Process, Pathologic Function, Sign or Symptom).

b. Only concepts in MEDIC were kept.

The filtered set of MeSH and OMIM ids was evaluated against the gold-standard annotations. Results are reported as MetaMap results.

2.2.3. The Inference Method—The Inference Method is our prior work [9] on disease normalization. Our results (F-measure: 79%) showed that this method linked disease mentions to their corresponding medical vocabulary entry with high precision. The core of the Inference Method was built as a combination of several string matching rules that mapped the annotated strings to the names of the diseases as listed in the standard disease dictionary and/or their accompanying list of synonyms. In addition, this method successfully exploited the fact that the long form of the disease is usually defined elsewhere in the same abstract. Once the abbreviation was resolved, the knowledge of the mapping of the long form of the disease was used to infer the mapping of the abbreviated mention. Gold standard data is essential in building a successful normalization method. The Inference Method was designed and refined on a manually annotated set of PubMed abstract sentences which reflected the consensus annotation agreement of the EBI disease corpus[13] and the AZDC disease corpus [11] (the only available data at the time). Because the original annotations used UMLS as a resource, UMLS concepts were converted to MeSH descriptors. Because the textual mentions were missing from the original annotations, NCBI disease corpus annotations were consulted to build the Inference Method gold standard set. In this work, this method was used both for generating NCBI disease corpus pre-annotations, and as a disease name normalization benchmark. To evaluate its performance in an ab-initio setting, first BANNER [40] was applied on each PubMed abstract for disease mention recognition, and then, the Inference Method was applied for disease name normalization.

3. Results

3.1. Inter-annotator Agreement

Overall, we achieved a high agreement between annotators as shown in Figure 2 where each dot represents a comparison score between two annotators for one or more shared sets. The red dotted line shows the average F-score agreement over all annotator pairs.

3.2. NCBI disease corpus statistics

NCBI disease corpus contains 793 fully annotated PubMed citations. This constitutes more than 6K sentences, of which more than half contain disease names. For benchmarking purposes, the entire corpus was split into three subsets (training, development, and test sets). As shown in Table 2, there are 2,136 unique disease mentions total, mapped to 790 unique database identifiers.

3.2.1. Distribution of annotated concepts per PubMed citation—On average, the NCBI disease corpus contains 5.08 disease mentions and 3.28 disease concepts per PubMed abstract. Detailed breakdown of the number of unique mentions and concepts per abstract is shown in Figure 3. As shown, while the proportion of articles that contain only a single disease mention is merely 5%, the proportion of articles linking to a single disease concept is higher than 26%. The reason is that there are many ways in mentioning the same concept

as manifested in Figure 2. Incidentally, the concept Genetic Diseases, Inborn [D030342] is the most frequent concept, appearing in 113 different documents, while "cancer" is the most frequent mention, appearing in 44 documents. The top 10 most common disease mentions and disease concepts in the NCBI disease corpus are listed in Table 3.

3.2.2. Coverage of MeSH and OMIM concepts—One novel aspect of this work is the fact that disease mentions are linked to MeSH descriptors and/or OMIM identifiers. Out of 790 unique disease name normalizations, 698 (or 88%) are composed of MeSH identifiers, and 92 (or 12%) contain an OMIM identifier. The distribution of these across training, development and testing set of the NCBI disease corpus is given in Table 4.

3.2.3. NCBI disease corpus has low ambiguity level—Finally, 719 (or 91%) of the 790 normalizations in the NCBI disease corpus are single concepts, either MeSH descriptors or OMIM identifiers, 52 normalized concepts are composite mappings corresponding to composite mentions, and only 24 need combination of multiple concepts to describe the given disease mentions. Typical examples of composite mappings are: pancreatic, basal cell, colonic, breast, and cervical carcinomas (D010190|D002280|D015179|D001943|D002583), Prader-Willi and Angelman syndromes (D011218|D017204), syndromic and non-syndromic hearing loss (D034381|C537845), and non-familial breast and ovarian cancers (D001943| D010051), etc. Examples of multiple concept mappings are: autosomal recessive cardiodegenerative and neurodegenerative disease (Genetic Diseases, Inborn [D030342] +Neurodegenerative Diseases [D019636]), chorioretinal dystrophy (Choroid Diseases [D015862] + Retinal Dystrophies [D058499]), and ochronotic arthropathy (Joint Diseases [D007592] + Ochronosis [D009794]).

3.2.4. One-to-one and many-to-one mappings to disease concepts in the NCBI disease corpus—On average, a disease concept is associated with 2.74 disease mentions in our corpus. As pointed out above, the NCBI disease corpus is a resource with low-ambiguity level. For example, more than half of the disease concepts in the NCBI disease corpus are associated with a disease mention in a one-to-one mapping as shown in Figure 4. In our analysis, we found the following outliers: Twenty-six disease concepts, or 3.3%, are associated with more than 10 textually different disease mentions, with the maximum being 57. Of these, only four concepts correspond to general disease classes: Genetic Diseases, Inborn [D030342] ("genetic disease," "recessively inherited condition"), Neoplasms [D009369] ("cancer," "sporadic tumors," "bilateral and multifocal tumours"), Immunologic Deficiency Syndromes [D007153] and Demyelinating Diseases [D003711]. These may represent opportunities for further vocabulary refinement. The rest either correspond to specific diseases such as specific enzyme deficiencies, such as Glucosephosphate Dehydrogenase Deficiency [D005955] ("Glucose 6-phosphate dehydrogenase deficiency," "G6PD-deficient," "deficiency of G6PD"), or specific cancers, including Hereditary Breast and Ovarian Cancer Syndrome [D061325] ("hereditary human breast and ovarian cancer," "inherited breast-ovarian cancer," "breast-ovarian cancer-family syndrome"). These variabilities reflect the productivity of language; recognizing these cases would likely benefit more from the development of better machine learning methods for disease recognition.

3.3. Benchmarking results for disease normalization methods on NCBI disease corpus

In this section we present the assessment of disease name resolution from natural language text. In this task we want to link the diseases with the appropriate concept in the MEDIC dictionary. The evaluation consists in comparing the annotations produced by each one of the methods with the annotations in the gold standard (the NCBI disease corpus -- test dataset). We report the standard measures: precision, recall and F-measure at two granularity

levels (micro-average and macro-average). Micro-average evaluation takes into consideration all predictions (unique) for each of the 100 PubMed citations that constitute the NCBI test dataset, and compares them with the whole set of (unique) gold-standard annotations. Macro-average evaluation first computes the standard measures for each PubMed citation in the test dataset, and then takes a global average over all the values (see results in Tables 5 and 6).

3.4. Error analysis for the disease name normalization with the NCBI disease corpus

We performed an error analysis to help illustrate the relative strengths and weaknesses of the different methods we employed. We randomly selected 20 PubMed abstracts from the testing dataset and reviewed the output results of our methods. All methods experienced both false positives and false negatives due to term variations not present in the lexicon, such as matching mention “neuromuscular disorder” to concept Neuromuscular Diseases [D009468]. Dictionary lookup, in addition, suffered from a large number of false positives due to matches from non-entity tokens, such as “be,” “all,” and “feed,” to an abbreviation listed as a disease synonym. This type of error was successfully ignored by MetaMap due to its use of syntactic filtering to ignore tokens not part of a noun phrase, and by the Inference Method due to the specific handling of abbreviated disease terms. Very general disease terms such as disorder, dysfunction, etc., were not considered interesting per our annotation guidelines and therefore were not included in annotation. MetaMap results, although filtered accordingly to exclude such terms, still produced a few remaining artifacts such as “nonsense mutation” and “suffering”. Another problem found in the MetaMap results were mappings such as “deficiency” (from “deficiency of the ninth component of complement”) to concept Malnutrition [D044342]. The Inference Method provided an improvement by linking these mentions to concepts listed within the MeSH subtree of Immunologic Deficiency Syndromes [D007153], although mappings to the wrong component number were often observed. Another common type of error for the Inference Method was its tendency to map to the most specific concept, for example the mention “cerebellar degeneration” was mapped to concept Subacute Cerebellar Degeneration [C535352], instead of Cerebellar Degeneration, Primary [D013132]. The Inference Method was able to handle more term variation than other methods, but also experienced both false positives and false negatives due to named entity recognition errors (such as missing the “14” in “paternal uniparental disomy 14”).

Finally, as can be seen in Tables 5 and 6, the dictionary method provides strong recall compared to the other methods, but also the lowest precision. Using MEDIC to restrict the MetaMap results to diseases provides both higher precision and higher recall than restricting the results using the UMLS semantic types referring to disease. The recall increase is caused by the inclusion of some concepts in MEDIC, which are excluded from the set of UMLS semantic types such as “Findings” used for filtering to maintain reasonable precision. The increase in precision is also expected as we are restricting MetaMap results to the more focused domain by using MEDIC. The Inference Method provides the highest precision, highest f-measure and the highest macro-averaged recall so far.

4. Discussion

Despite the efforts in the biomedical research community and the named-entity recognition challenge tasks organized in the biomedical domain (BioCreative, BioNLP, TREC, i2b2, etc.,), disease name recognition research has lacked in the development of competitive machine-learning methods due to the unavailability of suitably-sized gold-standard corpora. The NCBI disease corpus is significantly larger than the other available corpora with disease mention and/or concept annotations.

In addition, when we compare the NCBI disease corpus with the Jimeno et al., [13] and Leaman at al. [11] corpora we can identify several substantial differences:

1. The Jimeno et al. corpus does not have mention level annotation, and consists of only 856 PubMed sentences. It is important to have both mention and concept level annotation for purposes of building better quality recognition methods.
2. The Leaman at al. corpus, while providing mention-level annotation for 2,783 PubMed sentences, consists of concept level annotations to the most specific concept. This leads to several inconsistencies, for example:

   Cancer, cancers is found normalized to these concepts: “C0206663 – Neuroecodermal tumor”, “C0006826 – malignant neoplasms”, “C1306459 – primary malignant neoplasm”, “C0009402 – colorectal carcinoma”, “C1527249 – colorectal cancer”, “C0476089 – endometrial carcinoma”, etc.
3. Both previous corpora consist of selected PubMed sentences, while the NCBI disease corpus contains whole abstract annotations. We can see this difference in the annotation of abbreviated disease names, and their mapping to the correct concept. We have studied this issue in [9] and [22]. Our findings indicated that whole-document annotations improve the ability to automatically identify the correct concepts, in particular for abbreviated disease names, which occur frequently in biomedical literature.

Finally, the disease normalization methods results reported in this paper and in DNorm [22] show that the NCBI disease corpus is a suitable resource of gold-standard data to build better, more accurate models for disease name recognition.

4.1. Choosing a terminology resource

Several terminological resources are available that provide disease terms. Amongst the most used resources are the Medical Subject Headings (MeSH), the National Cancer Institute (NCI) thesaurus (http://www.cancer.gov/), SNOMED CT (http://www.ihtsdo.org/) and the UMLS National Library of Medicine UMLS resource that encloses the whole medical language terminology into a single source. Primarily, each of these has a different scope, and none of them has been designed to meet text mining needs. Hence, the need for a useful resource for building more precise tools for disease name recognition is still unmet.

Previous corpora have used UMLS as a backbone to normalize disease mentions [11, 13]. While UMLS Metathesaurus is the broadest resource of medical concepts, disease concepts correspond to only a small subset of UMLS, namely the 12 semantic types that comprise the Disorder group. The latest release of UMLS combines 136 source vocabularies, with a resulting set of approximately 3 million (distinct) concepts. If we counted only the concepts comprising the 12 semantic types of the Disorder group — creating smaller and hence more useful views of large terminology resources has also been suggested by other studies in the literature, for example in [41] — we would still count more than 540,000 distinct concepts.

SNOMED CT [42] was developed to enable a consistent way to index clinical data and structure medical records. At 66,000 concepts the SNOMED disorder branch may offer the right degree of granularity which would be very useful for mining the clinical aspects of disease. An enrichment of the disease mentions in the NCBI disease corpus with SNOMED terms could certainly make the NCBI disease corpus more valuable for clinical applications. However for the present this remains as future work.

The MEDIC disease lexicon[23] is a manually curated resource that associates a MeSH descriptor from the "Diseases" category, and/or a genetic disorder identifier from the OMIM database to disease names found in PubMed literature. This collection, at 10,000 entries, created a sufficient combination for our purposes, especially considering that neither MeSH nor OMIM have licensing restrictions. In addition, all PubMed abstracts undergo MeSH terminology indexing. Therefore, linking the disease mention in PubMed abstracts with the corresponding MeSH/OMIM identifiers would allow for easy integration with these other NLM resources. Our work, which in a sense may serve as the first real life evaluation of MEDIC, shows that MEDIC provides sufficient coverage for the disease concepts found in PubMed citations, making this the right environment for disease name normalization for such purposes.

Other possible terminology resources that could be considered for such purposes are the Disease Ontology [43] and the Human Phenotype Ontology [44]. A future work that links MeSH/OMIM ids in the NCBI disease corpus with these other terminologies would undoubtedly be very useful for related research on gene and protein functions. Another important detail to consider is the right level of annotation with the terminology terms. Recent research efforts on large-scale community-wide assessment of protein function annotation [45] illustrate the importance of annotation and the use of the appropriate ontology for both annotation and evaluation of computational methods. A computational method that only reports highly specific terminology terms is typically penalized in recall, while a method that only reports the most general ontology terms would have higher recall but its results would not be very useful in practice. Achieving the right balance is important and should be investigated both from the evaluation point of view and form the annotation point of view.

4.1.1. Mapping disease mentions to disease concepts—Our annotation results revealed some limitations of using the MEDIC disease vocabulary. First, a handful of disease concepts were discovered that were not included in MEDIC. For those, we decided to include the appropriate OMIM identifiers (as explained in the Annotation guidelines section, guideline 8).

Next, certain disease mentions were found to not be easily represented using the standard categorizations. Thus we allowed multiple concept normalization. Multiple concept normalization was used for *composite disease mentions* such as "**colorectal and endometrial cancers**" -> MeSH: D010051|D016889. Of interest is the fact that the composite mention "**hereditary breast and ovarian cancer syndrome**" has a special MeSH descriptor: D061325, and therefore there is no need for multiple concept normalization MeSH: D001943 (breast cancer) | D010051 (ovarian cancer). Multiple concept normalization was also used for providing a correct representation of a disease mention, when normalization to a single disease concept did not sufficiently convey the meaning intended by the authors, for example, "**inherited muscular disease**" is normalized to MeSH: D009468 - Neuromuscular Diseases + D030342 - Genetic Diseases, Inborn.

Our annotation work also pointed out at certain future directions for standard vocabulary development. For example, for some disease entries as shown in Figure 3, the large number of disease mentions mapped to the same concept suggests that perhaps a more refined classification is needed. We also identified disease nuances described in the literature which were lost during mapping, due to the lack of specificities and or detailed categorizations of diseases:

> Ex: "PRAD1 mRNA is abundantly expressed in seven of seven **centrocytic lymphomas** (Kiel classification), in contrast to 13 closely related but

noncentrocytic lymphomas." The current mapping is: centrocytic lymphomas -> Mantle cell lymphoma[MESH: D020522], noncentrocytic lymphomas -> Lymphoma [MESH: D008223] since there exists no disjoint classification of lymphoma in MeSH or OMIM, into centrocytic versus noncentrocytic.

"We explored the 30 Dutch kindreds well known to the Dutch **X-ALD**/**AMN** Study Group and phenotyped 77 male patients: 35 (46%) had **adrenomyeloneuropathy** (**AMN**) and 24 (31%) **childhood cerebral ALD** (**CCALD**) or **adolescent cerebral ALD** (**AdolCALD**)." The current mapping is Adrenoleukodystrophy [MeSH: D000326] since the specific categories identified in the article are not present in the standard nomenclature. Note that, this is a normal and expected observation. Disease research articles are expected to report variability, distinctions and subtypes of known diseases. With time, certain distinctions will enter the vocabulary and get listed in the terminology resources. Regardless, a successful disease recognition program is expected to recognize all the mentions listed in the article and all the concepts they correspond to.

4.2. Disease name normalization as opposed to gene name normalization

The logic employed to determine whether two entities should be considered synonymous is not the same for genes and diseases. Genes can typically be sequenced and traced to a specific genomic location, providing a significant degree of precision to the decision of whether or not two genes (or gene products) represent the same entity. Diseases, on the other hand, are always defined descriptively, through a collection of attributes, including symptoms, systems affected, references to disease processes, or to the disease etiology. Since disease definitions require some judgment to apply, there is therefore a significant amount of ambiguity present in any discussion of disease. The resulting variation is reflected in both disease names and the mentions used to refer to diseases.

Notably, none of the normalization methods we applied achieved performance comparable to systems developed for gene normalization tasks. Gene normalization is better studied, partially due to the existence of appropriate corpora for system development and evaluation, and also because shared tasks such as BioCreative have consistently included a task involving gene normalization. In addition, recent work in gene normalization has demonstrated that machine learning techniques provide a performance advantage if the necessary training data is available [14–16]. We believe that the NCBI disease corpus, with the joint annotation of mentions and concepts over the same text, will allow exploration of similar machine learning techniques for disease names, which were previously not possible. In this regard, the NCBI disease corpus will provide a valuable resource to the text mining researchers for the development of more capable, more powerful system in the field of disease name recognition.

5. Conclusions

This work presents the NCBI disease corpus, a richly annotated corpus with disease names and their corresponding MeSH and/or OMIM identifiers. This resource contains 793 PubMed abstracts, and lists 6,892 disease mentions, which are linked to 790 unique concepts, thus providing an important foundation for improving the text-mining research on disease named entity recognition. Our experiments demonstrated the feasibility of using the corpus as the basis for training learning models in both named entity identification and concept recognition, and we expect these results to serve as benchmark for other future methods. To facilitate future benchmarking experiments, the corpus is also divided into training, development and testing sets.

The NCBI disease corpus contains annotations for all sentences in a PubMed document (title and abstract), an important aspect in facilitating development of complex information retrieval tasks that connect diseases to treatments, causes or other types of information. This corpus provides annotation of disease mentions in four major categories: Specific Disease (i.e., clear-cell renal cell carcinoma), Disease Class (i.e., cystic kidney diseases), Composite mentions (i.e., prostatic, pancreas, skin, and lung cancer), and Modifier (i.e., hereditary breast cancer families) [21]. Disease normalization annotation guidelines were designed with the goal of allowing flexible matching to MeSH (diseases branch) and OMIM concepts, while retaining the true meaning of the specific mention. The current corpus was reviewed several times by several annotators and describes a refined scale of the annotation categories. The NCBI corpus can act as a basis for the development of more accurate machine learning systems for disease name recognition and normalization, as well as boost research into other areas of biomedical knowledge discovery pertaining to diseases.

Acknowledgments

We thank the team of 14 annotators for their time and expertise during the annotation of this corpus, and Francois Lang, Lan Aronson, and Jim Mork for help with Metamap, UMLS and other library tools.

Funding: This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

References

1. Islamaj Dogan R, Murray GC, Neveol A, Lu Z. Understanding PubMed user search behavior through log analysis. Database: the journal of biological databases and curation. 2009; 2009 bap018.
2. Brownstein JS, Freifeld CC, Madoff LC. Digital disease detection--harnessing the Web for public health surveillance. N Engl J Med. 2009; 360:2153–5. 7. [PubMed: 19423867]
3. Cohen KB, Hunter L. Getting started in text mining. PLoS Comput Biol. 2008; 4:e20. [PubMed: 18225946]
4. Krallinger M, Leitner F, Valencia A. Analysis of biological processes and diseases using text mining approaches. Methods Mol Biol. 2010; 593:341–82. [PubMed: 19957157]
5. Neveol, A.; Li, J.; Lu, Z. Linking multiple disease-related resources through UMLS. Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium; 2012. p. 767-72.
6. Neveol, A.; Kim, W.; Wilbur, WJ.; Lu, Z. Exploring two biomedical text genres for disease recognition. Proccedings of the ACL 2009 Workshop on Natural Language Processing in Biomedicine (BioNLP 2009); 2009. p. 144-52.
7. Islamaj Dogan R, Neveol A, Lu Z. A context-blocks model for identifying clinical relationships in patient records. BMC bioinformatics. 2011; 12 (Suppl 3):S3. [PubMed: 21658290]
8. Mork JG, Bodenreider O, Demner-Fushman D, Dogan RI, Lang FM, Lu Z, et al. Extracting Rx information from clinical narrative. Journal of the American Medical Informatics Association: JAMIA. 2010; 17:536–9. [PubMed: 20819859]
9. Islamaj Dogan, R.; Lu, Z. AAAI Fall Symposium on Information Retrieval and Knowledge Discovery in Biomedical Text. Arlington, Virginia: AAAI; 2012. An Inference Method for Disease Name Normalization.
10. Chowdhury, FM.; Lavelli, A. Disease mention recogntion with specific features. Proceedings of the ACL 2010 Workshop on Natural Language Processing in Biomedicine (BioNLP 2010); 2010. p. 83-90.
11. Leaman, R.; Miller, C.; Gonzalez, G. Enabling recognition of diseases in biomedical text with machine learning: Corpus and benchmark. Proceedings of the 2009 Symposium on Languages in Biology and Medicine; Jeju Island, South Korea. 2009. p. 82-9.
12. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001:17–21. [PubMed: 11825149]

13. Jimeno A, Jimenez-Ruiz E, Lee V, Gaudan S, Berlanga R, Rebholz-Schuhmann D. Assessment of disease named entity recognition on a corpus of annotated sentences. BMC bioinformatics. 2008; 9 (Suppl 3):S3. [PubMed: 18426548]
14. Lu Z, Kao HY, Wei CH, Huang M, Liu J, Kuo CJ, et al. The gene normalization task in BioCreative III. BMC bioinformatics. 2011; 12 (Suppl 8):S2. [PubMed: 22151901]
15. Morgan AA, Lu Z, Wang X, Cohen AM, Fluck J, Ruch P, et al. Overview of BioCreative II gene normalization. Genome biology. 2008; 9 (Suppl 2):S3. [PubMed: 18834494]
16. Hirschman L, Colosimo M, Morgan A, Yeh A. Overview of BioCreAtIvE task 1B: normalized gene lists. BMC Bioinformatics. 2005; 6 (Suppl 1):S11. [PubMed: 15960823]
17. Tomanek K, Wermter J, Hahn U. A reappraisal of sentence and token splitting for life sciences documents. Studies in health technology and informatics. 2007; 129:524–8. [PubMed: 17911772]
18. Kulick, S.; Bies, A.; Liberman, M.; Mandel, M.; McDonald, R.; Palmer, M., et al. Integrated annotation for biomedical information extraction. Proc of the Human Language Technology Conference and the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL); 2004.
19. Lease, M.; Charniak, E. Parsing Biomedical Literature. In: Dale, R.; Wong, K-F.; Su, J.; Kwong, O., editors. Natural Language Processing – IJCNLP 2005. Springer; Berlin Heidelberg: 2005. p. 58-69.
20. Roberts, A.; Gaizauskas, RJ.; Hepple, M.; Guo, Y. Combining Terminology Resources and Statistical Methods for Entity Recognition: an Evaluation. LREC: European Language Resources Association; 2008.
21. Islamaj Dogan, R.; Lu, Z. An improved corpus of disease mentions in PubMed citations. Proceedings of the 2012 Workshop on Biomedical Natural Language Processing; 2012. p. 91-9.
22. Leaman R, Islamaj Dogan R, Lu Z. DNorm: Disease name normalization with pairwise learning to rank. Bioinformatics. 2013 In press.
23. Davis AP, Wiegers TC, Rosenstein MC, Mattingly CJ. MEDIC: a practical disease vocabulary used at the Comparative Toxicogenomics Database. Database: the journal of biological databases and curation. 2012; 2012 bar065.
24. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004; 32:D267–70. [PubMed: 14681409]
25. Tanabe L, Xie N, Thom LH, Matten W, Wilbur WJ. GENETAG: a tagged corpus for gene/protein named entity recognition. BMC bioinformatics. 2005; 6 (Suppl 1):S3. [PubMed: 15960837]
26. Thompson P, Iqbal SA, McNaught J, Ananiadou S. Construction of an annotated corpus to support biomedical information extraction. BMC bioinformatics. 2009; 10:349. [PubMed: 19852798]
27. Neveol A, Islamaj Dogan R, Lu Z. Semi-automatic semantic annotation of PubMed queries: a study on quality, efficiency, satisfaction. Journal of biomedical informatics. 2011; 44:310–8. [PubMed: 21094696]
28. Chapman WW, Savova GK, Zheng J, Tharp M, Crowley R. Anaphoric reference in clinical reports: characteristics of an annotated corpus. Journal of biomedical informatics. 2012; 45:507–21. [PubMed: 22343015]
29. Verspoor K, Cohen KB, Lanfranchi A, Warner C, Johnson HL, Roeder C, et al. A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. BMC Bioinformatics. 2012; 13:207. [PubMed: 22901054]
30. Bada M, Eckert M, Evans D, Garcia K, Shipley K, Sitnikov D, et al. Concept annotation in the CRAFT corpus. BMC Bioinformatics. 2012; 13:161. [PubMed: 22776079]
31. Kim JD, Ohta T, Tateisi Y, Tsujii J. GENIA corpus--semantically annotated corpus for bio-textmining. Bioinformatics. 2003; 19 (Suppl 1):i180–2. [PubMed: 12855455]
32. Gerner M, Nenadic G, Bergman CM. LINNAEUS: a species name identification system for biomedical literature. BMC Bioinformatics. 2010; 11:85. [PubMed: 20149233]
33. Pyysalo S, Ginter F, Heimonen J, Bjorne J, Boberg J, Jarvinen J, et al. BioInfer: a corpus for information extraction in the biomedical domain. BMC Bioinformatics. 2007; 8:50. [PubMed: 17291334]

34. Lu, Z.; Bada, M.; Ogren, P.; Cohen, KB.; Hunter, L. Improving biomedical corpus annotation guidelines. Proceedings of the joint BioLink and 9th bio-ontologies meeting; 2006. p. 89-92.
35. Kang N, Singh B, Afzal Z, van Mulligen EM, Kors JA. Using rule-based natural language processing to improve disease normalization in biomedical text. J Am Med Inform Assoc. 2012
36. Wei CH, Harris RB, Li D, Berardini TZ, Huala E, Kao HY, et al. Accelerating literature curation with text mining tools: A case study of using PubTator to curate genes in PubMed abstracts. Database: the journal of biological databases and curation. 2012 bas041.
37. Lingren, T.; Deleger, L.; Zhai, H.; Meinzen-Derr, J.; Kaiser, M.; Stoutenborough, L., et al. Pre-annotating clinical notes and clinical trial announcements for gold standard corpus development: Evaluating the impact on annotation speed and potential bias. Proceedings of The Second IEEE International Conference on Healthcare Informatics, Imaging and Systems Biology (HISB 2012); San Diego, CA. 2012. p. 108
38. Wei, CH.; Kao, HY.; Lu, Z. PubTator: A PubMed-like interactive curation system for document triage and literature curation. Proceedings of the BioCreative 2012 Workshop; Washington, DC. USA. 2012. p. 145-50.
39. Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. Journal of the American Medical Informatics Association: JAMIA. 2005; 12:296–8. [PubMed: 15684123]
40. Leaman, R.; Gonzalez, G. BANNER: an executable survey of advances in biomedical named entity recognition. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing; 2008. p. 652-63.
41. Wu ST, Liu H, Li D, Tao C, Musen MA, Chute CG, et al. Unified Medical Language System term occurrences in clinical notes: a large-scale corpus analysis. Journal of the American Medical Informatics Association: JAMIA. 2012; 19:e149–56. [PubMed: 22493050]
42. Cornet R, de Keizer N. Forty years of SNOMED: a literature review. BMC medical informatics and decision making. 2008; 8 (Suppl 1):S2. [PubMed: 19007439]
43. Schriml LM, Arze C, Nadendla S, Chang YW, Mazaitis M, Felix V, et al. Disease Ontology: a backbone for disease semantic integration. Nucleic acids research. 2012; 40:D940–6. [PubMed: 22080554]
44. Robinson PN, Mundlos S. The human phenotype ontology. Clinical genetics. 2010; 77:525–34. [PubMed: 20412080]
45. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, et al. A large-scale evaluation of computational protein function prediction. Nat Methods. 2013; 10:221–7. [PubMed: 23353650]

Highlights

- NCBI disease corpus is built as a gold-standard resource for disease recognition.
- 793 PubMed abstracts are annotated with disease mentions and concepts (MeSH/OMIM).
- 14 annotators produced high consistency level and inter-annotator agreement.
- Normalization benchmark results demonstrate the utility of the corpus.
- The corpus is publicly available to the community.

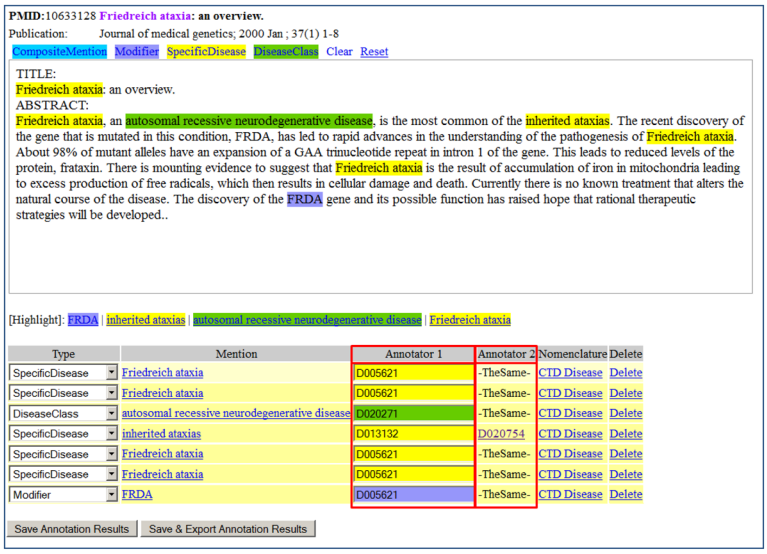

Figure 1.
Screenshot of our annotation software PubTator displaying the differences in annotation for two annotators at the end of Phase I.

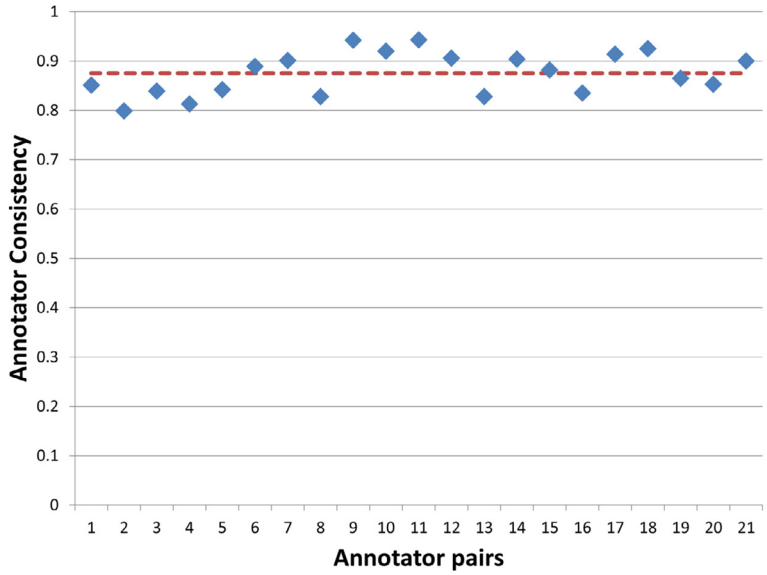

Figure 2.
Inter-annotator annotation agreement for the disease name normalization task. The red dotted line shows the average F-score among all annotator pairs.

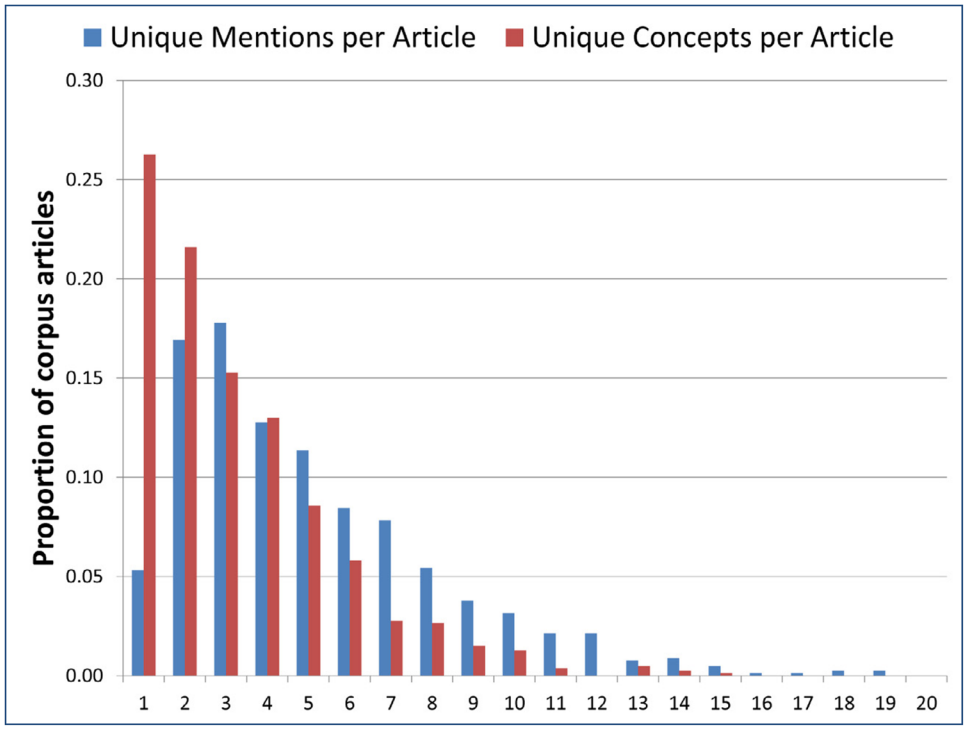

Figure 3.
Distribution of mentions and concepts in the corpus documents.

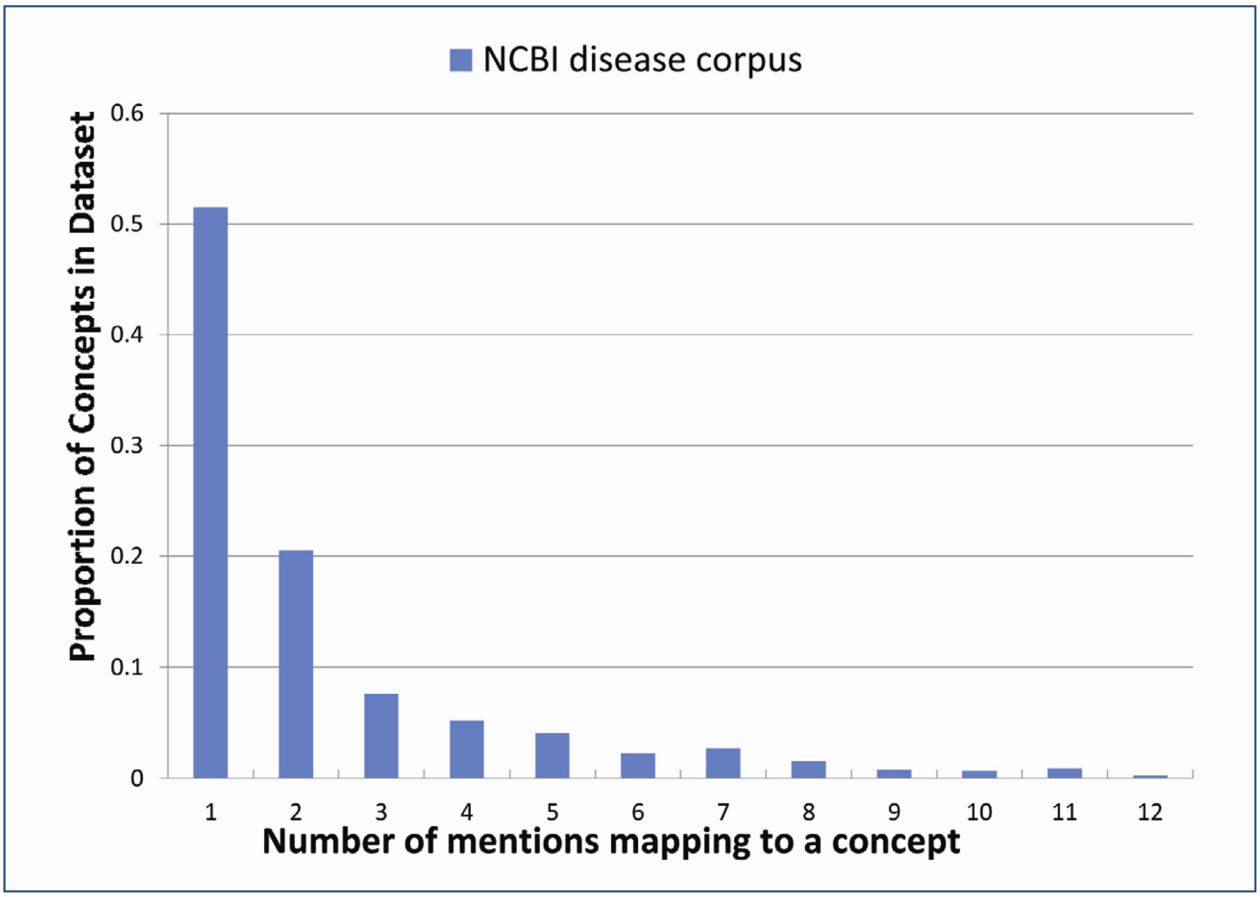

Figure 4.
Distribution of disease name concepts with respect to the number of disease mention strings that map to them.

Table 1

Characteristics of disease corpora

Disease Corpus	Corpus Size	Disease Mention	Disease Concept Ontology	Number of Annotators
Jimeno et al., 2008	856 selected sentences in 642 abstracts	No	UMLS	2
Leaman et al., 2009	2,783 selected sentences in 793 abstracts	Yes	UMLS	1
NCBI disease corpus	All 6,881 sentences in 793 abstracts	Yes	MEDIC	14

Table 2

NCBI disease corpus as training, development and testing sets for disease name recognition

Corpus Characteristics	Training set	Development set	Test set	Whole Corpus
PubMed Citations	593	100	100	793
Total disease mentions	5,145	787	960	6,892
Unique disease mentions	1,710	368	427	2,136
Unique Concept ID	670	176	203	790

Table 3

The most common disease mentions and disease concepts in the NCBI disease corpus, with the corresponding number of abstracts they appear in.

Disease mentions (number of abstracts)	Disease concepts (number of abstracts)
Cancer (44)	D030342 - Genetic Diseases, Inborn (113)
Tumor (43)	D009369 - Neoplasms (112)
Breast cancer (41)	D061325 - Hereditary Breast and Ovarian Cancer Syndrome (52)
DM (39)	D001943 - Breast Neoplasms (46)
Myotonic dystrophy (35)	D009223 - Myotonic Dystrophy (42)
G6PD deficiency (33)	D005955 - Glucosephosphate Dehydrogenase Deficiency (36)
DMD (33)	D020388 - Muscular Dystrophy, Duchenne (33)
Ataxia-telangiectasia (30)	D011125 - Adenomatous Polyposis Coli (33)
APC (29)	D001260 - Ataxia Telangiectasia (31)
Duchene muscular dystrophy (27)	D010051 - Ovarian Neoplasms (27)

Table 4

Distribution of MeSH versus OMIM concepts

Set	MeSH (unique)	OMIM (total)	Annotated Concepts (unique)
Training	1512 (599)	198 (71)	1710 (670)
Development	316 (153)	52 (23)	368 (176)
Testing	378 (182)	49 (21)	427 (203)

Table 5

Benchmarking results of three disease normalization methods (micro-average)

Method	Precision	Recall	F-measure
Dictionary look-up	0.218	0.685	0.331
MetaMap (semantic type filtering)	0.475	0.644	0.547
MetaMap (MEDIC filtering)	0.502	0.665	0.572
Inference Method	0.533	0.662	0.591

Table 6

Benchmarking results of three disease normalization methods (macro-average)

Method	Precision	Recall	F-measure
Dictionary look-up	0.213	0.718	0.316
MetaMap (semantic type filtering)	0.495	0.679	0.541
MetaMap (MEDIC filtering)	0.510	0.702	0.559
Inference Method	0.597	0.731	0.637