LabeledIn: Cataloging labeled indications for human drugs

Ritu Khare [a], Jiao Li [b], Zhiyong Lu [a,*]

[a] National Center for Biotechnology Information (NCBI), U.S. National Institutes of Health, 8600 Rockville Pike, Bethesda, USA
[b] Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing, China

ARTICLE INFO

ABSTRACT

Drug–disease treatment relationships, *i.e., which drug(s) are indicated to treat which disease(s)*, are among the most frequently sought information in PubMed®. Such information is useful for feeding the Google Knowledge Graph, designing computational methods to predict novel drug indications, and validating clinical information in EMRs. Given the importance and utility of this information, there have been several efforts to create repositories of drugs and their indications. However, existing resources are incomplete. Furthermore, they neither label indications in a structured way nor differentiate them by drug-specific properties such as dosage form, and thus do not support computer processing or semantic interoperability. More recently, several studies have proposed automatic methods to extract structured indications from drug descriptions; however, their performance is limited by natural language challenges in disease named entity recognition and indication selection.

In response, we report LabeledIn: a human-reviewed, machine-readable and source-linked catalog of labeled indications for human drugs. More specifically, we describe our semi-automatic approach to derive LabeledIn from drug descriptions through human annotations with aids from automatic methods. As the data source, we use the drug labels (or package inserts) submitted to the FDA by drug manufacturers and made available in DailyMed. Our machine-assisted human annotation workflow comprises: (i) a grouping method to remove redundancy and identify representative drug labels to be used for human annotation, (ii) an automatic method to recognize and normalize mentions of diseases in drug labels as candidate indications, and (iii) a two-round annotation workflow for human experts to judge the pre-computed candidates and deliver the final gold standard.

In this study, we focused on 250 highly accessed drugs in PubMed Health, a newly developed public web resource for consumers and clinicians on prevention and treatment of diseases. These 250 drugs corresponded to more than 8000 drug labels (500 unique) in DailyMed in which 2950 candidate indications were pre-tagged by an automatic tool. After being reviewed independently by two experts, 1618 indications were selected, and additional 97 (missed by computer) were manually added, with an inter-annotator agreement of 88.35% as measured by the Kappa coefficient. Our final annotation results in LabeledIn consist of 7805 drug–disease treatment relationships where drugs are represented as a triplet of ingredient, dose form, and strength.

A systematic comparison of LabeledIn with an existing computer-derived resource revealed significant discrepancies, confirming the need to involve humans in the creation of such a resource. In addition, LabeledIn is unique in that it contains detailed textual context of the selected indications in drug labels, making it suitable for the development of advanced computational methods for the automatic extraction of indications from free text. Finally, motivated by the studies on drug nomenclature and medication errors in EMRs, we adopted a fine-grained drug representation scheme, which enables the automatic identification of drugs with indications specific to certain dose forms or strengths. Future work includes expanding our coverage to more drugs and integration with other resources.

The LabeledIn dataset and the annotation guidelines are available at http://ftp.ncbi.nlm.nih.gov/pub/lu/LabeledIn/.

1. Introduction

Drug–disease treatment relationships are among the top searched topics in PubMed® [1,2]. Such relationships are

* Corresponding author.
   E-mail addresses: ritu.khare@nih.gov (R. Khare), li.jiao@imicams.ac.cn (J. Li), zhiyong.lu@nih.gov (Z. Lu).

established during the drug discovery and development process, which establishes the therapeutic intent of a given drug based on its properties and target patient characteristics. The primary application of such information is to inform healthcare professionals and patients for questions like "what drugs may be prescribed for hypertension" or "what are the indications of Fluoxetine" [3]. These relationships are also used for feeding the Google Knowledge Graph, developing computational methods for predicting and validating results of novel drug indications [4–6] and drug side effects [7], assisting PubMed Health® (http://www.ncbi.nlm.nih.gov/pubmedhealth/) editors to cross-link drug and disease monographs [8]. More recently, such information is found to be critical in validating patient notes and medication-problem links in electronic medical records (EMRs) [9–11]. Given this variety of applications, it is important to have a comprehensive gold standard of drug–disease treatment relationships, that is (i) *accurate* and derived from a credible source, and (ii) *structured* to support computer processing, and (iii) *normalized* to precise concepts in standard vocabularies, such as UMLS [12] and RxNorm [13], to facilitate semantic understanding and interoperability. The third desired property deserves further explanation. To precisely represent the treatment relationship, it is necessary that diseases and drugs are normalized to the most appropriate abstraction levels:

- Disease Normalization: The diseases should be normalized to the most *specific* concepts. For instance, if a drug is used for treating "respiratory tract infections," mapping to the generic concept "infections" would not only be imprecise but also inaccurate since the drug may not treat all kinds of infections.
- Drug Normalization: A drug can be represented at several levels of granularity based on its properties. While the therapeutic intent of a drug is largely determined by its active ingredient (IN), there is evidence showing that it may also be dictated by its dose form (DF) and strength (ST) [10,14,15]. For example, the indications of Ketorolac oral tablet are different from those of the ophthalmic solution (see Table 1).

Several existing knowledge bases such as DrugBank [16] and MedicineNet [17] already contain drug–disease relationships. However, they are unstructured (i.e. described in free text), and thus, do not allow automatic computer analysis. Google's Freebase [18] is a structured resource, but the drugs are coarsely represented as ingredients, and the diseases are not normalized. The NDF-RT [19] provides structured and normalized information. However, it is found incomplete with respect to the list of drug indications [20,21], and the drug–disease relationships are not separately labeled according to different dose forms or strengths. For instance, the Ketorolac drug is manufactured in multiple dose forms, each serving a different purpose, e.g. injectable solution is used for pain, and ophthalmic solution for conjunctivitis. Despite this, the NDF-RT links all the different forms of this drug to the same set of diseases: inflammation, allergic conjunctivitis, photophobia, and pain.

In addition, there have been multiple attempts to use automated methods for extracting computable (i.e. structured and normalized) indication information from existing textual resources (e.g. the DailyMed website [22]) using knowledge-based approaches. SIDER 2 [23] is a public resource focused on identifying adverse drug reactions and indications from the FDA drug labels and public documents. The method used to extract indications is based on a UMLS-based lexicon lookup technique followed by side effects filtering. The October 2012 version of SIDER 2 contains indications for 10319 drugs labels. Neveol and Lu [24] used text mining techniques to extract indications from FDA drug labels, and automatically extract 2200 relationships between 1263 ingredients and 581 diseases using SemRep [25] with precision of 73%. Wei and colleagues [26] created an ensemble indication resource called MEDI by integrating information from four resources: SIDER 2, NDF-RT, MedlinePlus, and Wikipedia. A subset of MEDI was sampled and reviewed by two physicians in two rounds to further determine the automatic inclusion strategy for a high precision dataset. The final computer-generated dataset (MEDI-HPS) contains 13304 ingredient-indication pairs corresponding to 2136 ingredients with an estimated 0.92 precision and 0.30 recall. Fung et al. [21] designed a DailyMed-based indication extraction system (SPL-X) for decision support in electronic medical records. SPL-X uses MetaMap [27], negation removal, and semantic reclassification techniques [28,29] for disease concept identification. SPL-X was applied to 2105 unique drug labels, a subset of which was evaluated by seven physicians showing 0.77 in precision and 0.95 in recall.

As found in the abovementioned studies [21,24,26], automatic methods alone are not yet sufficient to deliver a gold standard due to the challenges in natural language processing (NLP), including: (a) the difficulties with automatic disease recognition and normalization [30–32], and (b) the presence of disease mentions other than indications in drug labels. To illustrate these, Table 1 contains the indication fields of two sample drug labels (dl_1 and dl_2: same ingredient but different dose forms) in DailyMed [22], which houses the most up-to-date drug labels submitted to the FDA by drug manufacturers. Table 1 also shows the disease names found by a state-of-the-art NLP tool (column 3, in *italics*) and the final indications after human revision (column 4, in *italics*). As can be seen, recognition of disease mentions is not trivial ("*ocular itching*" vs. "*itching*" in dl_1; "*severe pain*" vs. "*moderate to severe pain*" in dl_2). In addition, drug labels could contain negative and irrelevant ("*cataract*" mention in dl_1; "*analgesia*" mention in dl_2) disease mentions.

Unlike previous studies [22,24,25], in this work we resort to human annotation to create a gold standard of drug indications with the aids from automatic text-mining tools, as they have been shown to be useful for assisting manual curation [33,34]. This

Table 1
Drug with multiple dose forms: Computer Pre-annotations and expert judgments.

Drug label	Drug concept	Indications (identified automatically)	Indications (improved using expert judgments)
dl_1	Ketorolac Ophthalmic Solution	**ACULAR** ophthalmic solution is indicated for the temporary relief of ocular *itching* due to *seasonal allergic conjunctivitis*	**ACULAR** ophthalmic solution is indicated for the temporary relief of *ocular itching* due to *seasonal allergic conjunctivitis*
	(RxNorm CUI: 377446)	**ACULAR** ophthalmic solution is also indicated for the treatment of postoperative *inflammation* in patients who have undergone *cataract* extraction	**ACULAR** ophthalmic solution is also indicated for the treatment of postoperative *inflammation* in patients who have undergone cataract extraction
dl_2	Ketorolac Oral Tablet (RxNorm CUI: 372547)	Ketorolac tromethamine tablets are indicated for short term (5 days) management of moderate to *severe pain* that requires *analgesia* at the opioid level	Ketorolac tromethamine tablets are indicated for short term (5 days) management of *moderate to severe pain* that requires analgesia at the opioid level

study is built on a pilot study [35] in which we evaluated the feasibility of our semi-automated annotation framework on a small set of 100 drug labels with two human annotators. Based on the error analysis of the previous study, we have significantly improved the annotation framework, revised the annotation guidelines, and produced a resource covering 8151 DailyMed drug labels in the current study. Another unique aspect of the current work is that each indication is linked to a specific textual location in the source drug label(s). Not only does this provide evidence and context of selected indications, such as linked textual information, but also can serve as training data for the development of supervised machine-learning methods for automatic indication extraction. Finally, this study uses a finer-grained scheme where each drug is represented as a 3-tuple <Active Ingredient (IN), Dose Form (DF), Drug Strength (ST)> when linking to its labeled indications.

2. Materials and methods

2.1. Overall workflow for annotating drug indications

Fig. 1 shows an outline of our semi-automated approach consisting of three distinct steps: (i) drug label selection; (ii) automatic disease recognition; and (iii) manual indication annotation. Each step is detailed in Sections 2.3–2.5.

2.2. DailyMed: The source of FDA drug labels

DailyMed is a drug database maintained by the National Library of Medicine (NLM) [22]. DailyMed is considered to be the largest resource on marketed drugs containing high-quality information about human and animal drugs including both over-the-counter and prescription drugs. All drug labels are available in HTML and XML formats. Fig. 2 shows the Web version for a drug label. Each label is organized into multiple sections; the "INDICATIONS AND USAGE" section provides information on drug indications in a narrative format. The NLM editors assign normalized drug concepts to the drug labels (see "RxNorm Names" box in Fig. 2), and hence the DailyMed is structured and normalized in terms of drug information. To create the gold standard of drug–disease treatment relationships, the key is to identify the most specific drug indications mentioned in the textual description, normalize them to corresponding UMLS concepts, and link to the associated drug concepts.

2.3. Drug label selection

We accessed DailyMed on September 1 2012, and downloaded its August 24 2012 version, which contained 18324 human prescription drug labels. For each label, we extracted its indication field from the XML file, and the assigned RxNorm concepts from the respective web page. We then applied a set of filters to ensure that a drug label is linked to RxNorm CUIs and its associated indication field is not empty. In this study, we focused on the 250 human prescription ingredients frequently found in access logs of PubMed Health. These 250 drugs correspond to 8151 drug labels. From here, our goal was to minimize annotation efforts without any loss of information downstream. We observed that a drug ingredient can have multiple drug labels in DailyMed (submitted by different manufacturers) with same or different textual descriptions in the indication field. To minimize annotation efforts, we only selected unique drug labels (in terms of indication texts) for annotation study. We grouped similar drug labels and chose a representative drug label from each group. To minimize loss of information, we only grouped highly similar (with text that is almost identical) drug labels together.

First, all 8151 drug labels were grouped based on the linked drug ingredients resulting into 250 groups. Then, each group was further sub-grouped such that all the drug labels in a sub-group are highly similar (i.e. identical) to each other in terms of their indications. We used the Dice co-efficient [36] to measure similarity between drug labels. In particular, we considered two drug labels to be identical if their Dice co-efficient lied above the threshold of 0.87. This threshold was empirically determined using our analysis of 100 drug labels. In this way, we derived 542 sub-groups wherein each sub-group contained drug labels linked to the same ingredient and having highly similar indication descriptions. As shown in Fig. 3, the size of sub-groups (i.e. the number of similar drug labels in sub-groups) ranges from 1 through 131.

We further observed, especially in case of drug labels with shorter lengths, that certain drug labels were identical to each other and were still assigned to different sub-groups. We manually merged these sub-groups resulting into 500 indication based sub-groups. For each sub-group, we randomly chose a drug label to be annotated and used as a representative of the group. In this way, we minimized the annotation effort by 93%, i.e. from 8151 drug labels to 500 drug labels. These 500 drug labels represent 250 INs, 611 < IN, DF > pairs, and 1531 < IN, DF, ST > triplets. The indication descriptions are varied in length ranging from 10 to more than 1000 words, with average of 130 words (±149 words).

2.4. Automatic disease recognition

The goal of this module was to identify all disease mentions as indication candidates from the textual descriptions of a given drug label. For this study, we prepared a disease lexicon using two seed ontologies, MeSH and SNOMED-CT, respectively useful for annotating scientific articles [30,32,37] and clinical documents [31,38,39].
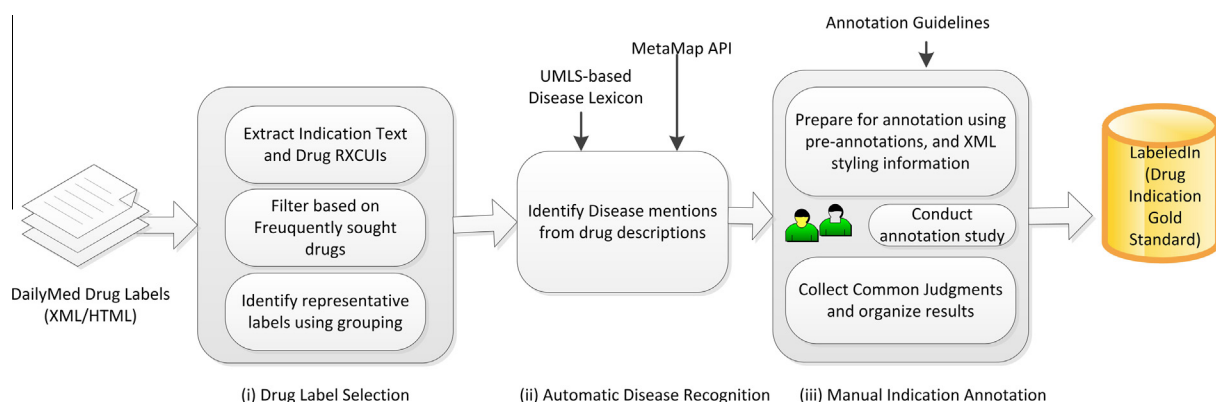

Fig. 1. The overall framework of our study for annotating drug labels.

Fig. 2. A snapshot of DailyMed drug label by Allergan Inc. The indication information is provided by manufacturers, and normalized (RxNorm) drug concepts are assigned by NLM curators and editors.
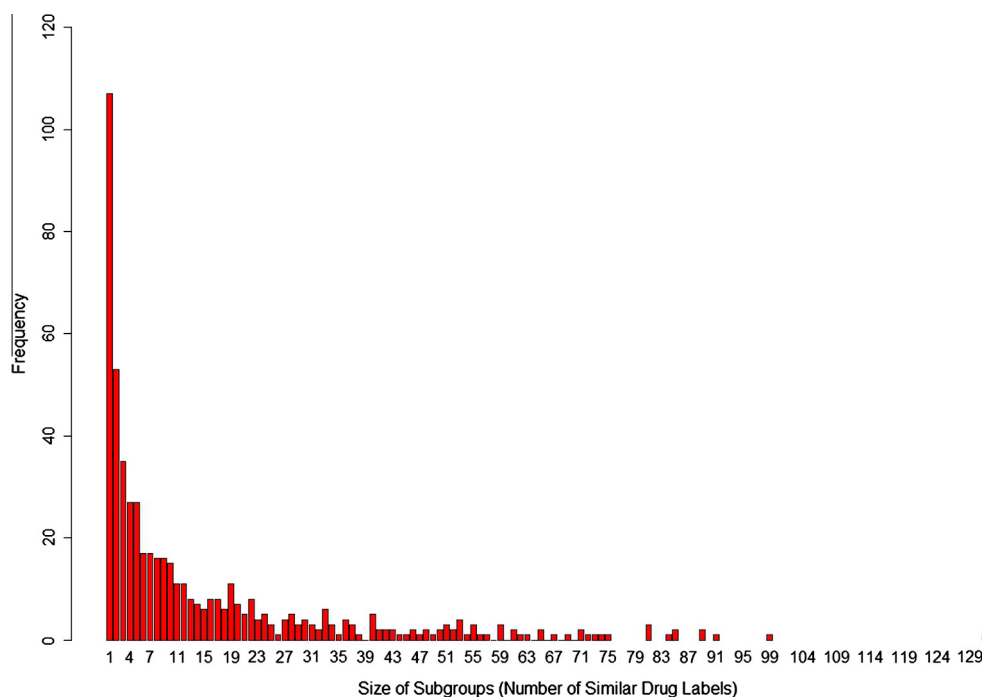

Fig. 3. An insight into the indication based sub-grouping for 8151 drug labels.

The lexicon consists of 77464 concepts taken from: (i) the disease branch in MeSH, and (ii) the 11 disorder semantic types (UMLS disorder semantic types excluding 'Finding') in SNOMED-CT as recommended in a recent shared task [30].

As for the automatic tool, we applied MetaMap [27], a highly configurable program used for mapping biomedical texts to the UMLS identifying the mentions, offsets, and associated CUIs. We used the 2012 MetaMap Java API release that uses the 2012AB version of the UMLS Metathesaurus. We experimented with multiple settings of MetaMap, and the optimal setting method for this study is illustrated in Fig. 4.

The drug descriptions may contain overlapping disease mentions, e.g. the phrase "skin and soft tissue infections" denotes two specific diseases, "skin infections" and "soft tissue infections." While the final results by MetaMap do not return such overlapping mentions, these are captured in the intermediate results of MetaMap, known as the Metathesaurus candidates. Hence, we utilized these candidate concepts, as opposed to the final results, in our disease recognition method. MetaMap provides two types of candidates, contiguous and dis-contiguous, e.g. in the phrase "skin and soft tissue infections", "soft tissue infections" is a contiguous candidate, and "skin + infections" is a dis-contiguous candidate. We found that MetaMap returns different sets of dis-contiguous candidates with and without the *term processing* feature. Hence, we conducted two runs of MetaMap for comprehensive results. Also, the *word sense disambiguation* feature was turned on to disambiguate mentions that may map to multiple CUIs, e.g. "depression."

In order to restrict the returned candidates to specific semantic types from two vocabularies as mentioned above, we used a lookup against our custom disease lexicon as opposed to running multiple rounds of MetaMap for the two vocabularies. Finally, candidates with overlapping spans were resolved in the following manner: (i) when both candidates were contiguous, the longer candidate was selected, (ii) when one candidate was dis-contiguous - (a) if the merged span contained conjunctions (e.g. "or," "and") or prepositions (e.g. "to"), then the merged span was
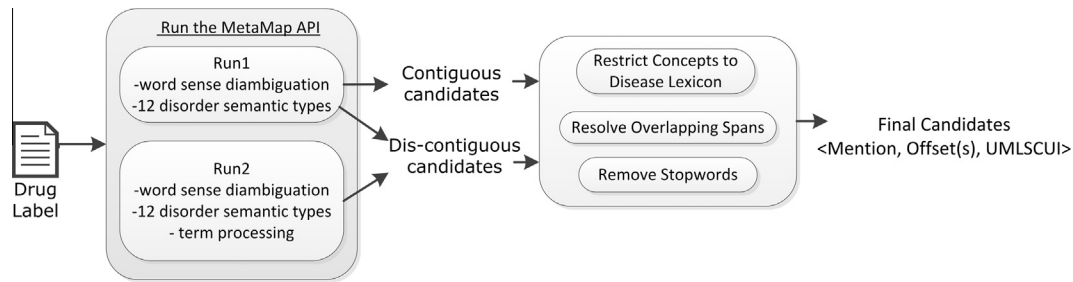
Fig. 4. Disease concept recognition method.

pre-annotated and both CUIs were retained, e.g. the elliptical coordination in "skin and soft tissue infections," (b) if the two mentions were related by a parent–child UMLS relationship (e.g., the phrase "acute bacterial otitis media" maps to hierarchically related concepts "acute + otitis media" and "otitis"), then the longer mention was retained, else, the shorter mention was retained (e.g. the phrase "drug hypersensitivity reactions" maps to non-hierarchically related concepts "drug + reactions" and "hypersensitivity reactions").

2.5. Manual indication annotation

The annotation study was conducted with two professional biomedical annotators with more than five years of experience working in an academic medical research institution. The annotators have an educational background in pharmacology and medicine, and have been trained in biomedical literature indexing. The indication section of the drug labels was presented to the two human annotators by highlighting the disease mentions (i.e., computer pre-annotations) identified in the previous automatic step. All mentions (including dis-contiguous) were presented as contiguous

by expanding the spans. At the backend, we maintained a mapping between the mentions and the associated CUIs. We used a crowd-sourcing platform (www.crowdflower.com) to build the annotation interface [35]. The system presented the pre-annotated drug labels on the annotation interface illustrated in Fig. 5. To facilitate quick and correct annotation, we leveraged the styling information available from the XML file and presented the drug label exactly as it would appear on the Web. The drug labels were presented one at a time. The human annotation was conducted in two rounds. During round-1, the annotators were asked to independently annotate the drug labels (Fig. 5a). During round-2, the annotators were asked to independently update their previous annotations based on previous disagreements (Fig. 5b).

We conducted the study with 50 drug labels at a time comprising 10 sets for 500 representative labels. The first two sets were annotated in our pilot study where a ground truth was curated and annotators' judgments were assessed against it [35]. The evolved guidelines from that study were used for conducting the annotation study for the remaining sets. After round-1, the annotators disagreed on average 20 (±4) labels, indicating the size of workload in round-2. The average human effort spent in double

(a) Round-1 Interface showing Pre-annotations

(b) Round-2 Interface showing the Exclusive Judgments

Fig. 5. Annotation interface.

Table 2
Description of the final annotation results (LabeledIn).

Drug specificity (size in LabeledIn)	Total annotated drug–disease pairs	Example	
		Drug	Indication(s)
Ingredient (249)	1318	Diclofenac	Osteoarthritis Rheumatoid arthritis Ankylosing Spondylitis Osteoarthritis of the knee(s) Acute pain Strains Sprains Contusions Pain Actinic keratoses Migraine Aura Inflammation
Ingredient + dose form (611)	2997	Diclofenac Topical Gel	Osteoarthritis Pain Actinic keratosis
Ingredient + dose form + strength (1513)	7805	Diclofenac 0.03MG/MG Topical Gel	Actinic keratosis

annotating a drug label in round-1 and round-2 was 3.47 min and 3.98 min, respectively. After each set annotation, we studied the comments provided by annotators and improved the guidelines accordingly. Furthermore, we consulted with a domain expert with a doctoral degree in pharmacy to validate and refine the annotation guidelines. The final version of guidelines is described thoroughly in the supplemental materials. The consensus judged by both annotators was used for deriving the gold standard.

2.6. Evaluation

We evaluated our annotated corpus in several ways. We first computed the size of the gold standard, i.e., the total number of drugs and drug–disease relationships at different levels of granularity. Then we empirically studied the effect of dose form and strength on indications in our annotation results. Next, we computed the Jaccard agreement ($P(A) = \frac{|A_1 \cap A_2|}{|A_1 \cup A_2|}$, A_i = judgments by ith annotator) and kappa agreement ($\frac{P(A) - P(e)}{1 - P(e)}$, $P(e)$ = probability of chance agreement) to assess the consistency of our annotation results. Then, we measured the precision and recall of the automatic concept recognition method with respect to the gold standard generated by the annotators. Finally, we compared our results with a similar resource by computing the Jaccard inter-source agreement $\left(\frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}, S_i = \text{indications in } i\text{th source} \right)$.

3. Results

3.1. Description of final annotation results

We name our final annotation results as "LabeledIn" since it was created from the labeled indications in the DailyMed. LabeledIn contains the relationships between drugs and indications, and is organized at three levels of granularity as shown in Table 2. It should be noted that out of the 250 ingredients, one ingredient (Varenicline) was not included in the final results because this drug is used for smoking cessation and its corresponding labels did not

have any mentions that mapped to a disease concept. On average, each drug label contained 3.43 indications.

We noticed that 136 ingredients in our results were associated with multiple drug labels in DailyMed. Using automatic analysis of <IN, DF, ST> and indication combinations, we found that 68% of these ingredients had indications that were DF-specific (same IN but different DF had different indications). For instance, the Fluticasone topical cream is indicated for "Atopic Dermatitis" whereas the Fluticasone nasal inhaler is indicated for "Seasonal and Allergic Rhinitis." About 11% of the ingredients had indications that were ST-specific (same <IN, DF> but different ST had different indications), e.g. the Finasteride 5 mg oral tablet is indicated for "Benign Prostatic Hypertrophy" whereas its 1 mg counterpart is indicated for "Androgenetic Alopecia" (a.k.a. male pattern baldness).

3.2. Inter-annotator agreement and comparison of computer pre-annotations vs. human annotation results

Our annotated text corpus contains 500 drug labels double-annotated in two rounds. The average Jaccard agreements between annotators for round-1 and round-2 were 88.77% and 94.18%, respectively. The average Kappa agreements for round-1 and round-2 were 77.48% and 88.35%, respectively. After round-2, the main cause for remaining differences is that in addition to the main indication, one of the two annotators also selected its generic or related form. Some examples include:

(i) "*Primary Prevention of Cardiovascular Disease. CRESTOR is indicated to reduce the risk of stroke and myocardial infarction*": one annotator selected "Cardiovascular Disease" in addition to "stroke" and "myocardial infarction"

(ii) "*PROVIGIL is indicated to improve wakefulness in adult patients with excessive sleepiness associated with Narcolepsy, obstructive sleep apnea, and shift work disorder. In all cases, careful attention to the diagnosis and treatment of the underlying sleep disorder(s) is of utmost importance*": one annotator selected "sleep disorder(s)" in addition to "excessive sleepiness"

On average, the automatic disease recognition module identified 5.9 pre-annotations per drug label. Compared to the final human annotation results, the automatic method delivered a micro-averaged precision, recall, and F_1-measure of 0.55, 0.94, and 0.69 (and macro-averaged 0.67, 0.95, and 0.74), respectively. Fig. 6 shows the overlap between the pre-annotations and the final annotation results, the precision denotes that about 55% of the
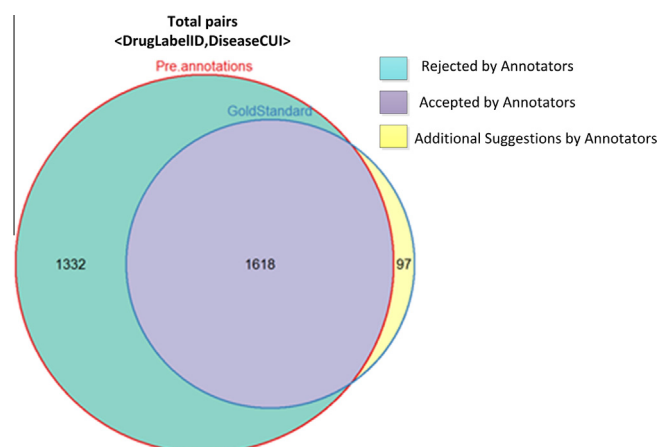

Fig. 6. Overlap between computer pre-annotations (#2950) and Gold Standard or LabeledIn (#1715).
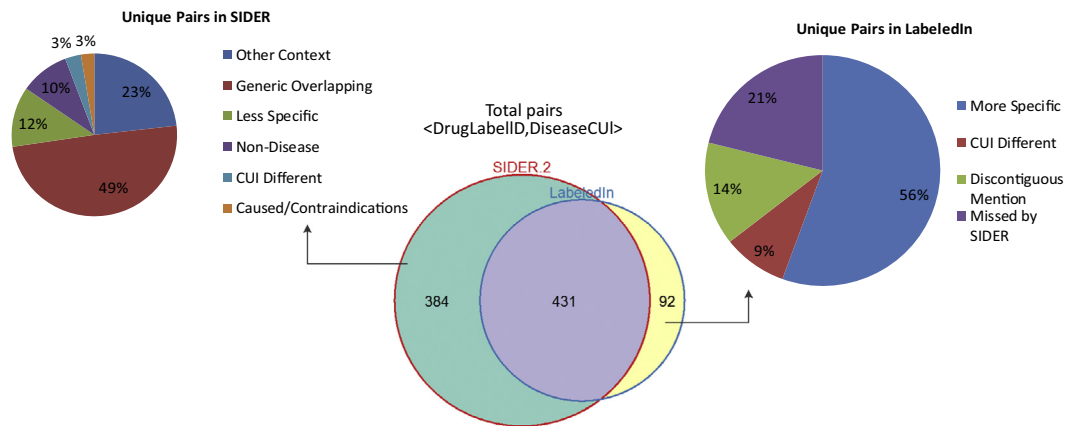
Fig. 7. Comparison of SIDER 2 indications and LabeledIn for 50 randomly selected drug labels.

computer recommended pre-annotations were accepted by the annotators (overlapping region); and the recall denotes that about 6% of the final results was created using the indications not captured by automated methods but added by the human annotators (yellow region).

3.3. Comparison results with SIDER 2

Among several similar datasets, SIDER 2 is the only one that contains indications extracted from drug labels and provides identifiers of specific drug labels. Therefore, we systematically compared LabeledIn with SIDER 2. Between LabeledIn and SIDER 2 (October 2012 version), there was an overlap of 3877 drug labels, which were reduced to 459 representative drug labels through our grouping method (Section 2.3). We compared the indications concepts for all 459 drug labels and observed micro- and macro-averaged agreements of 0.30 and 0.37, respectively.

The discrepancies in indications were found in 417 drug labels, out of which we randomly selected 50 drug labels to manually study the discrepant (unique) CUIs in both resources and identify the reasons for discrepancies as shown in Fig. 7. Table 3 illustrates the examples from different categories of discrepancies in SIDER 2.

Similarly, the unique CUIs in LabeledIn could be classified as: (i) *More Specific* and (ii) *CUI Different*, the counterparts of the *Less Specific* and *CUI Different* categories in SIDER 2, respectively, (iii) *Discontiguous Mentions*, detected due to the use of *term processing* in MetaMap, e.g. from the phrase "biliary and renal colic," LabeledIn included "biliary + colic" (a dis-contiguous mention), and (iv) *Missed by SIDER*, possibly due to their choice of lexicon, e.g. "Zollinger-Ellison syndrome," "Loeffler's syndrome," etc. The two former categories (65%) in LabeledIn could be considered as partial matches with SIDER 2 since they include cases where SIDER 2 results contain a corresponding related disease.

4. Discussion and conclusions

We have conducted a study of annotating FDA drug labels using a semi-automatic method. Deleger et al. [40] previously annotated disease mentions from FDA drug labels. Our study produced a comparable inter-annotator agreement (88%) but differs in that we yielded drug indications as opposed to all the diseases mentioned in a drug label. Distinguishing indications from other disease mentions is a non-trivial problem requiring human judgment as we observed that approximately 45% of automatically identified

Table 3
Examples of discrepant concepts in SIDER 2.

Discrepancy category	Description	Example statement	SIDER 2 Discrepant Annotations
Generic – overlapping	Includes cases when SIDER 2 annotated both the generic as well as specific indication	"myocardial infarction"	"infarction" (in addition to "myocardial infarction")
Other context	Refers to cases when the disease is mentioned in some other context, e.g. risk factor, characteristics, indicated usages of other drugs, etc.	"Amantadine hydrochloride capsules are indicated in the treatment of symptomatic parkinsonism which may follow injury to the nervous system by carbon monoxide intoxication"	"intoxication"
Less specific	Refers to scenarios when SIDER 2 annotated a less specific disease as compared to the annotation in our results	"Levetiracetam is indicated as adjunctive therapy in the treatment of partial onset seizures in adults"	"seizures"(LabeledIn annotated "partial onset seizures"
Non-disease	Includes concepts that were not included in our disease lexicon such as organism names, medical procedures, etc.	"Aspirin may be continued, …use of NSAIDs including salicylates has not been fully explored (see PRECAUTIONS , Drug Interactions)"	"Drug Interactions"
CUI different	Includes mentions that had same spans but were disambiguated differently by the two resources	"Trazodone hydrochloride tablets are indicated for the treatment of depression"	"depression" (SIDER identified as C0011581 – depressive disorder, whereas LabeledIn identified as C0011570 –mental depression)
Caused/contraindications	Includes contra-indications of the drug and the diseases that are induced or caused, rather than treated or prevented, by the drug	"Phentermine hydrochloride is indicated as a short-term (a few weeks) adjunct in a regimen of weight reduction based on exercise, behavioral modification and caloric restriction in the management of exogenous obesity for patients"	"weight reduction"

disease mentions are not indications and include other concepts such as contraindications, side effects, risk factors, etc. Moreover, we observed only 30% agreement between our annotation results and an automatically generated resource, SIDER 2, due to the presence of many non-indications in SIDER 2. This further demonstrates the significance of human involvement in the creation of an accurate drug indication gold standard.

In addition to supporting computer processing, LabeledIn is linked to the source drug labels such that it captures the rich context in which a disease is judged as an indication. This enables systematic comparison with existing sources and enhancing automatically generated resources (e.g. SIDER 2 in Section 3.3). Also, because drug indications in LabeledIn are linked to the source text, they enable the development of supervised machine-learning methods for automatic drug indication extraction [24] from free text. Furthermore, an immediate application of the human-validated treatment relationships in LabeledIn is to improve accessibility in online health resources (e.g. PubMed Health) by enriching hyperlinks between drug and disease monographs.

While most existing studies represent a drug by its active ingredient, the theoretical definition of "therapeutic equivalence" suggests that the information about dose form and drug strength is also critical in ensuring effective and correct treatment [15]. Such information is also important in controlling documentation malpractices, including prescription, medication, drug nomenclature errors [9,10,14], and medication-problem linking errors [11], in the EMRs. Hence, this study regards indications as a function of all the key properties of a drug, and represents a drug as a 3-tuple <Active Ingredient (IN), Dose Form (DF), Drug Strength (ST)>. Furthermore, the automatic analysis of our annotation results helped identify the candidate drugs in DailyMed for which the indications may in practice be dictated by dose forms and/or strengths. The validation of such advanced information about the candidate drugs, however, requires further analysis by domain experts in medicine and pharmacy.

There are several limitations of the current work. First, LabeledIn currently contains information about 250 highly accessed drugs and covers nearly 50% of the human prescription FDA drug labels accordingly. For future work, we would like to expand to more ingredients and keep LabeledIn current with new releases of DailyMed. As an estimate, we studied the effective differences between the existing version of LabeledIn (August 2012) and the current version of Daily-Med (April 2014) for our 250 drugs. We found that only about 53 new drug labels need to be annotated for a period of 20 months. Second, LabeledIn only contains labeled/marketed indications. On the other hand, an existing resource MEDI [26] provides computable information regarding off-label indications from Wikipedia and MedlinePlus, in addition to labeled indications from SIDER 2 and NDF-RT. Hence, in the future we plan to investigate ways to integrate our results with existing resources such as MEDI. Lastly, certain annotated drug indications (e.g. "inflammation" in Table 1) are specific to certain procedures/conditions which are not currently captured. Given LabeledIn is linked to the source drug labels, in future work we plan to extract and organize such information in structured and computable format in order to further enrich our resource. In summary, we have produced LabeledIn, a resource containing the labeled indication information for 250 frequently accessed human drugs. We believe our human annotation results are useful in a wide variety of applications, and are complementary to existing resources.

Data Availability

The LabeledIn dataset and the annotation guidelines are publicly available at http://ftp.ncbi.nlm.nih.gov/pub/lu/LabeledIn.

References

[1] Islamaj Dogan R, Murray GC, Neveol A, Lu Z. Understanding PubMed user search behavior through log analysis. Database: J Biol Databases Curation 2009. 2009:bap018. Epub 2010/02/17.
[2] Neveol A, Islamaj Dogan R, Lu Z. Semi-automatic semantic annotation of PubMed queries: a study on quality, efficiency, satisfaction. J Biomed Infor 2011;44(2):310–8. Epub 2010/11/26.
[3] Ely JW, Osheroff JA, Gorman PN, Ebell MH, Chambliss ML, Pifer EA, et al. A taxonomy of generic clinical questions: classification study. BMJ 2000;321(7258):429–32. Epub 2000/08/11.
[4] Lu Z, Agarwal P, Butte AJ. Computational Drug Repositioning – Session Introduction. Pacific Symposium on Biocomputing; 2013. p. 1–4.
[5] Li J, Lu Z. A new method for computational drug repositioning using drug pairwise similarity. IEEE Int Conf Bioinfor Biomed 2012:1–4.
[6] Li J, Lu Z. Pathway-based drug repositioning using causal inference. BMC Bioinfor 2013;14((Suppl 16):S3).
[7] Chang RL, Xie L, Bourne PE, Palsson BO. Drug off-target effects predicted using structural analysis in the context of a metabolic network model. PLoS Comput Biol 2010;6(9):e1000938. Epub 2010/10/20.
[8] Li J, Khare R, Lu Z. Improving online access to drug-related information. In: IEEE second international conference on healthcare informatics, imaging and systems biology, La Jolla, CA; 2012.
[9] Khare R, An Y, Wolf S, Nyirjesy P, Liu L, Chou E. Understanding the EMR error control practices among gynecologic physicians. In: iConference 2013 fort worth, Texas; 2013. p. 289–301.
[10] Lesar TS. Prescribing errors involving medication dosage forms. J Gen Int Med 2002;17(8):579–87. Epub 2002/09/06.
[11] McCoy AB, Wright A, Laxmisan A, Ottosen MJ, McCoy JA, Butten D, et al. Development and evaluation of a crowdsourcing methodology for knowledge base construction: identifying relationships between clinical problems and medications. J Am Med Infor Assoc: JAMIA 2012;19(5):713–8. Epub 2012/05/15.
[12] Unified Medical Language System (UMLS). U.S. National Library of Medicine, <http://www.nlm.nih.gov/research/umls/>.
[13] RxNorm. U.S. National Library of Medicine, <http://www.nlm.nih.gov/research/umls/rxnorm/>.
[14] Lesar TS, Briceland L, Stein DS. Factors related to errors in medication prescribing. JAMA: J Am Med Assoc 1997;277(4):312–7. Epub 1997/01/22.
[15] Administration USFaD. Drug Approvals and Databases > Drugs@FDA Glossary of Terms. <http://www.fda.gov/drugs/informationondrugs/ucm079436.htm>.
[16] Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, et al. DrugBank 3.0: a comprehensive resource for 'omics' research on drug. Nucl Acids Res 2011;38(Database Issue). D1035–41.
[17] MedicineNet: We Bring Doctor's Knowledge to You. <http://www.medicinenet.com/script/main/hp.asp>.
[18] Freebase: A Community-curated database of well-known people, places, and things. <http://www.freebase.com/>.
[19] 2012AA National drug file – reference terminology source information. <http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NDFRT/>.
[20] Barriere C, Gagnon M. Drugs and disorders: from specialized resources to Web Data. In: The 10th international semantic web conference; October 23–27; Bonn, Germany; 2011.
[21] Fung KW, Jao CS, Demner-Fushman D. Extracting drug indication information from structured product labels using natural language processing. J Am Med Infor Assoc: JAMIA 2013;20(3):482–8. Epub 2013/03/12.
[22] DailyMed: Current Medication Information. <http://dailymed.nlm.nih.gov>.
[23] SIDER 2 Side Effect Resource. <http://sideeffects.embl.de/>.
[24] Neveol A, Lu Z. Automatic integration of drug indications from multiple health resources. ACM International Health Informatics Symposium; Arlington, VA2010. p. 666–73.
[25] Rindflesch TC, Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic

propositions in biomedical text. J Biomed Infor 2003;36(6):462–77. Epub 2004/02/05.

[26] Wei WQ, Cronin RM, Xu H, Lasko TA, Bastarache L, Denny JC. Development and evaluation of an ensemble resource linking medications to their indications. J Am Med Infor Assoc: JAMIA 2013;20(5):954–61. Epub 2013/04/12.

[27] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proceedings/AMIA annual symposium AMIA symposium; 2001. p. 17–21. Epub 2002/02/05.

[28] Fan JW, Friedman C. Semantic reclassification of the UMLS concepts. Bioinformatics 2008;24(17):1971–3. Epub 2008/07/16.

[29] Fan JW, Friedman C. Semantic classification of biomedical concepts using distributional similarity. J Am Med Infor Assoc: JAMIA 2007;14(4):467–77. Epub 2007/04/27.

[30] Dogan RI, Lu Z. An improved corpus of disease mentions in PubMed citations. Workshop on Biomedical Natural Language Processing; 2012. p. 91–9.

[31] Leaman R, Khare R, Lu Z. NCBI at 2013 ShARe/CLEF eHealth shared task: disorder normalization in clinical notes with DNorm. In: Conference and labs of the evaluation forum 2013 Working Notes; 2013.

[32] Leaman R, Islamaj Dogan R, Lu Z. DNorm: disease name normalization with pairwise learning to rank. Bioinformatics 2013. Epub 2013/08/24.

[33] Wei CH, Kao HY, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. Nucl Acids Res 2013;41(Web Server issue). W518-22. Epub 2013/05/25.

[34] Wei CH, Harris BR, 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: J Biol Databases Curation 2012. 2012:bas041. Epub 2012/11/20.

[35] Khare R, Li J, Lu Z. Toward creating a gold standard of drug indications from FDA Drug Labels. In: IEEE international conference on health informatics; September 09–11, 2013; Philadelphia, PA; 2013.

[36] Smadja F, Hatzivassiloglou V, McKeown KR. Translating collocations for bilingual lexicons: a statistical approach. Comput Linguistics 1996.

[37] Huang M, Neveol A, Lu Z. Recommending MeSH terms for annotating biomedical articles. J Am Med Infor Assoc: JAMIA 2011;18(5):660–7. Epub 2011/05/27.

[38] Khare R, An Y, Li J, Song I-Y, Hu X. Exploiting semantic structure for mapping user-specified form terms to SNOMED CT concepts. In: ACM SIGHIT international health informatics symposium; Miami, FL; 2012. p. 285–94.

[39] An Y, Khare R, Hu X, Song I-Y. Bridging encounter forms and electronic medical record databases: Annotation, mapping, and integration. In: IEEE international conference on bioinformatics and biomedicine (BIBM 2012); October 04–07, 2012; Philadelphia, PA; 2012.

[40] Deleger L, Li Q, Lingren T, Kaiser M, Molnar KDU, Stoutenborough L, et al. Building gold standard corpora for medical natural language processing tasks. AMIA Annu Symp Proc 2012:144–53.