Toward Creating a Gold Standard of Drug Indications from FDA Drug Labels

Ritu Khare
National Center for Biotechnology Information
U.S. National Library of Medicine
Bethesda, MD 20894, USA
ritu.khare@nih.gov

Jiao Li
Institute of Medical Information
Chinese Academy of Medical Sciences
Beijing 100020, China
li.jiao@imicams.ac.cn

Zhiyong Lu
National Center for Biotechnology Information
U.S. National Library of Medicine
Bethesda, MD 20894, USA
zhiyong.lu@nih.gov

Abstract—Having quick access to trustworthy drug-disease relationships (which drug(s) are approved for treating or preventing which disease(s)) is one of the top information needs of health providers, consumers, and researchers. This paper presents a semi-automatic approach that can lead to the creation of a gold standard of drugs and their indications. As our system input, we use the DailyMed, which houses the most current drug labels submitted to FDA by pharmaceutical companies. Extraction of specific indications from FDA labels is a challenging problem that requires distinguishing indications from other disease mentions. In response, we first identify the candidate indications from drug labels using UMLS resources and BioNLP tools, and then rely on expert judgments to validate those pre-computed indications through an interactive Web interface. For preliminary analysis, we recruited two experts to manually annotate 100 labels of frequently sought human prescription drugs at PubMed Health. We find that the resultant expert-curated gold standard on drugs and their indications is high-quality (precision=97%, recall=94%), and differs from existing resources in that it is factual, structured, and dose-form specific. The study findings suggest the feasibility of the proposed method toward building a comprehensive resource of drug indications.

Keywords—annotation; drug indication; expert curation; gold standard

I. INTRODUCTION

Having quick access to drug indications, i.e., which drug(s) are approved for treating or preventing which disease(s), is one of the top information needs of health providers, researchers, and consumers. Established drug-disease relationships are the most frequently sought information among clinicians [1], and are among the 10 most frequent multi-concept queries among PubMed users [2]. In addition to answering questions like “what are the indicated uses of the Ketorolac injection,” such relationships are useful in training biomedical systems for predicting novel drug indications [3, 4], assisting PubMed Health® (http://www.ncbi.nlm.nih.gov/pubmedhealth/) editors and developers in creating cross-links between drug and disease monographs [5], and controlling errors in electronic medical records [6]. Considering the significance and the variety of applications, it is important to build a comprehensive gold standard of drugs and their indicated uses. Such a gold standard should be:

• factual, in that it is trustworthy, accurate, and up-to-date
• structured, in that it facilitates cross-referencing between drugs and diseases, and is normalized to standard concepts facilitating future integration
• specific, in that the relationship is qualified by a specific dosage form (oral tablet, nasal solution, etc.) since the dosage form primarily dictates the disease a drug may help manage, e.g., indicated uses of Ketorolac injection are different from those of Ketorolac ophthalmic solution.

Indeed, there are several existing resources (e.g., DrugBank [7], MedicineNet [8]) with factual information on drug-disease relationships, but these are unstructured in nature, and thus, do not support interoperability. There also have been attempts [9, 10] to extract drug-disease relationships from unstructured text, such as biomedical literature. However, such relationships represent hypotheses or research findings as opposed to facts, and hence would not qualify as the gold standard. While the NDF-RT [11] provides factual as well as structured information, it is incomplete [12], and not specific in terms of the drug dosage form. For instance, the Ketorolac drug is manufactured in at least 4 different forms, each serving a different purpose (e.g., injection, capsule liquid filled).

<table>
<thead>
<tr>
<th>Drug</th>
<th>Active Ingredient (Dose Form)</th>
<th>Indication Excerpts in DailyMed (disease mentions are highlighted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>d1</td>
<td>Dutasteride (Capsule liquid filled)</td>
<td>…are indicated for the treatment of benign prostatic hyperplasia, is not approved for the prevention of prostate cancer.</td>
</tr>
<tr>
<td>d2</td>
<td>Ranitidine (Injection)</td>
<td>…is indicated in the treatment of GERD. Concomitant antacids should be given for pain relief to patients with GERD.</td>
</tr>
<tr>
<td>d3</td>
<td>Levetiracetam (Injection)</td>
<td>Levetiracetam is indicated as adjunctive therapy in the treatment of partial onset seizures in children 1 month of age and older with epilepsy, and in treatment of juvenile myoclonic epilepsy.</td>
</tr>
<tr>
<td>d4</td>
<td>Dimenhydrinate (Injection)</td>
<td>… is indicated for the prevention and treatment of nausea, vomiting, or vertigo of motion sickness.</td>
</tr>
</tbody>
</table>

TABLE I. DRUGS AND INDICATIONS

Manufactured in at least 4 different forms, each serving a different purpose (e.g., injection, capsule liquid filled).
different purpose, e.g., injectable solution is used for pain, and ophthalmic solution is used for conjunctivitis. Despite this, the NDF-RT links all the forms of this drug to the same set of diseases.

In this paper, we present a framework for creating a factual, structured, and specific gold standard of drug indications from the FDA drug labels. We used the DailyMed [13], which houses the drug labels submitted to FDA by drug manufacturers. Each drug label is specific to a dosage form, and provides comprehensive information on the indicated usages in a textual narrative format. Table I shows 4 example drugs \{d1, d2, d3, d4\} and their indication description in DailyMed. Automatic identification of indications is challenging due to the presence of negative (e.g., “prostate cancer” mention in the drug d1 description) and irrelevant (e.g., “pain” mention in the drug d2 description) disease mentions. Even manual identification is challenging because the textual descriptions are long (avg. 187 words), disease per se has a loose definition [14], and indications are authored in a variety of formats, e.g., specific and generic indications (drug d3), primary and associated indications (drug d4), etc.

Recently, a few studies have been conducted on mining the FDA drug labels. Deleger et al. [15] conducted a study to annotate disease mentions from 52 FDA drug labels, and achieved an agreement of 88% using 2 annotators. Our study differs in that we identify indications, as opposed to disease mentions, and demonstrate that distinguishing indications from other disease mentions is a non-trivial problem. Neveol and Lu [16] used text mining techniques on FDA drug labels, and automatically extracted 2,200 relationships between 1,263 ingredients and 581 diseases, with 73% accuracy. We differ in that we intend to qualify each relationship with a dosage form, and to involve expert intervention to achieve a highly accurate gold standard.

We developed a framework to create the drug-disease gold standard in the following manner. First, we designed an automated method to identify the candidate indications from textual descriptions of drug labels using UMLS resources and BioNLP tools. Next, we developed an interactive computer interface to project a drug label for annotation purposes. Lastly, human expert judgments were included to validate the computed indications. For preliminary evaluation, we recruited two experts to annotate 100 labels (50 labels at a time) corresponding to the most frequently-sought human prescription drugs at PubMed Health. Using the proposed framework, the experts spent only 3 min/label to identify indications from labels with 97% recall and 94% precision. We found that the judgments made by both annotators were good for building the desired gold standard. This study serves as the first evidence of annotating indications from FDA drug labels, and highlights several key challenges and guidelines in annotation.

II. MATERIALS AND METHODS

A. DailyMed: The Drug Indication Data Source

To build the gold standard of drug indications, we used the DailyMed, which provides high quality information about marketed drugs. DailyMed is a drug database of the National Library of Medicine (NLM), and houses the most recent drug labels submitted to the FDA by various pharmaceutical companies. All drug labels are available as Web pages for quick reference, and are freely downloadable in the XML format for programmatic usage. Figure 1a shows a condensed Web version for a label submitted by Allergan Inc. Each label is organized into multiple sections; the “INDICATIONS AND USAGE” section provides information on drug indications in a narrative format. In addition, the NLM editors assign normalized drug concepts to the drug labels, e.g., the drug in the figure is linked to the UMLS RxNorm concept, “Acular 0.5% Ophthalmic Solution” with concept identifier “860109.” The target gold standard should create a link between the drug and the specific diseases that the drug may treat or prevent. Essentially, the problem is to extract the indication mentions from the textual description, normalize the mentions to standard UMLS concepts, and link to the associated RxNorm concepts.

Fig. 1 (a) A condensed DailyMed drug label: ingredient = ketorolac tromethamine, dose form = ophthalmic solution, brand name = Acular, manufacturer = Allergan Inc. The indication information is provided by manufacturers, and normalized (RxNorm) drug concepts are assigned by NLM curators and editors. (b) Steps for extracting drug-disease relationships from a drug label.
We accessed the DailyMed on September 1 2012, and downloaded the August 24 2012 version, which contained 18,353 human prescription drug labels. DailyMed contains multiple labels that correspond to the same drug but are manufactured by different companies. We clustered the labels based on their associated RxNorm identifiers, and determined a representative label for each cluster based on the (min.) length of the “INDICATIONS AND USAGE” section. We thus reduced the dataset to 2,497 labels corresponding to different drugs. We further reduced this dataset and created a frequent dataset of 504 labels that were most accessed at PubMed Health. For this study, we randomly selected 100 labels from the frequent dataset.

B. Mining Indications from Drug Labels

The key task in creating the gold standard was to identify the indicated uses of a drug, i.e., step 2 in Figure 1b. In this study, we used semi-automated methods to identify the specific indications from textual descriptions. We used the UMLS Metathesaurus to automatically identify the candidate indications from drug labels, and implemented an annotation interface to display the drug labels with pre-computed candidate indications (highlighted in colors), i.e., pre-annotations [17]. To validate the pre-annotations, we recruited two expert annotators, A1 and A2, who are trained biomedical literature indexers with educational background in medical and library sciences.

1) Computing Pre-annotations:

To determine the candidate indications from a given label, we employed a UMLS-based disease lexicon, and matched the terms from the drug description with the lexicon terms.

To prepare the disease lexicon, we extracted the UMLS concepts, and the associated synonyms, belonging to the 12 disorder semantic types recommended in [14]. We first included the concepts from the MeSH source vocabulary, and later included two additional vocabularies, as described in Table II, based on manual analysis of 50 drug labels. We limited the concepts to English terms, and removed acronyms, abbreviations, fully specified names, and stop words. In addition, we extended the lexicon by using a string normalization tool [18]. Finally, the disease lexicon contained 25,823 UMLS concepts and 141,168 terms.

To prepare the pre-annotations, we tokenized a given description into consecutive terms of length 1 through 6. All tokens and their normalized versions [18] were cross-matched with the disease lexicon terms, and the matched terms were used as pre-annotations in the drug labels. To resolve overlapping pre-annotations, e.g., “arthritis” and “rheumatoid arthritis,” we chose the more specific disease, or the longest matching term. We find that comparable pre-annotation results could be achieved by using NLM’s MetaMap (http://metamap.nlm.nih.gov).

2) The Annotation Interface:

We coded the user interface using JavaScript and CSS, on a commercial annotation platform. The interface presents a drug label in a condensed format displaying the text from the “INDICATIONS AND USAGE” section along with the computed pre-annotations. A screenshot of the interface is shown in Figure 2, and the accompanied instructions are shown in Figure 3.

To avoid confusion regarding duplicate mentions, when a user validates (or deletes) a disease that has multiple mentions, all the duplicate mentions are automatically highlighted (or de-highlighted). We also created a without pre-annotation version of the interface where the user was free to highlight any portion of the text. However, based on a pilot test on 50 drug labels, this version resulted into a 12% decrease in quality of annotations confirming previous findings [17]. Hence, we continued with the pre-annotation version.
3) Annotation Workflow: We conducted the annotation study in two rounds (round-1 and round-2). During round-1, the experts were asked to independently annotate the drug labels. During round-2, the experts were asked to independently improve their previous annotations based on the disagreements observed during round-1. In round-2, we only included those labels where the two annotators disagreed, and pre-highlighted (i) the exclusive judgments, which were made by only one annotator during round-1, and (ii) the automatically computed pre-annotations that were not selected by either during round-1. The interface was re-designed for round-2, and a screenshot is shown in Figure 4.

We conducted the study with 50 drug labels at a time comprising two sets, Set-1 and Set-2. The average label lengths for both sets were 126 and 249 words, respectively. Based on error analysis, after each round, we evolved the annotation guidelines; the final guidelines are shown in Table III.

### TABLE III. ANNOTATION GUIDELINES

<table>
<thead>
<tr>
<th>What to Annotate</th>
<th>What Not to Annotate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Select all the indications. (Quite often the subtitles contain the indications.)</td>
<td>1. Do not select the contraindicated diseases.</td>
</tr>
<tr>
<td>2. Select all the indications. (Read the entire description.)</td>
<td>2. Do not select the indicated usages of another drug.</td>
</tr>
<tr>
<td>3. Select the main as well as the associated indication(s).</td>
<td>3. Do not select disease mentions that are part of an organization’s name.</td>
</tr>
<tr>
<td>4. Select the main as well as the causing indication(s).</td>
<td>4. Do not select the indications which are an explicit part of a list of symptoms.</td>
</tr>
<tr>
<td>5. Select the indication even when it is treated by a combination of drugs.</td>
<td>5. Do not select the species or organism names.</td>
</tr>
<tr>
<td></td>
<td>6. Do not select medical procedures, such as analgesia, anesthesia, or sedation.</td>
</tr>
</tbody>
</table>

### C. Judgment Assessment and Gold Standard Creation

To evaluate the expert judgments, we created a ground truth for the 100-label dataset. To identify accurate indication(s) for each drug, we manually reviewed the drug label description, and determined the indicated usages and the corresponding UMLS concepts. To verify the identified indications, we also consulted other resources, such as, NDF-RT and PubMed Health. In all, 461 ground truth indications were identified. These indications are distributed over 9 of the 12 UMLS disorder semantic types [14], barring “Cell or Molecular Dysfunction,” “Anatomical Abnormality,” and “Experimental Model of Disease.”

After each annotation round, we collected the judgments, and measured the duration of annotations. We mapped the collected judgments to UMLS concepts. We wanted to investigate whether the two experts could jointly produce a high-quality gold standard. Hence, we measured their “joint performance,” i.e., the performance of those judgments wherein both experts agree with each other. We measured the recall and precision of their joint judgments by comparing the UMLS concept identifiers with the ground truth. We also calculated the inter-annotator agreement ($num_{match}/num_{match}+num_{nonmatch}$) between the two annotators, $A_1$ and $A_2$.

### III. RESULTS

**A. Pre-annotations Quality**

In all, 850 pre-annotations were computed for 100 labels. Out of these, 441 overlapped with ground truth indications, leading to a precision of 51.88%. The remaining pre-annotations corresponded to the irrelevant or negative disease mentions (e.g., drugs $d_1$ and $d_2$ in Table I). The pre-annotations recalled 95.67% of the ground truth indications. The few failure cases were due to (i) natural language challenges, e.g., identifying “tinea versicolor” from “tinea (pityriasis) versicolor,” and “skin infections” from “skin and structure infections,” or (ii) limitations of the lexicon e.g., the concepts “tick fever,” and “pylori infection,” were not included. Given these limitations, we gave extra performance
credit to the judgment that span-overlapped with the ground truth indications, e.g., annotating “fever” from “tick fever.”

B. Judgment Assessment

It should be noted that round-2 is a shorter version of round-1 where only those labels were included for which the two experts disagreed. The total number of labels annotated during round-1(Set-1), round-2(Set-2), round-1(Set-1), and round-2(Set-2) were 50, 22, 50, and 26, respectively. Each annotator spent an average 124 min and 173 min to finish both rounds of annotations for Set-1 and Set-2, respectively. Table IV summarizes the joint performance of the annotators, computed using the judgments where the two agreed. The performances for round-2 are calculated by combining the round-2 labels and the round-1 labels where the annotators had perfect agreement.

<table>
<thead>
<tr>
<th>TABLE IV. EXPERT JOINT PERFORMANCE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
</tbody>
</table>

The round-1 for Set-1 was conducted with very brief annotation guidelines. This round achieved a nearly perfect (0.99) joint precision, suggesting that the agreed upon judgments are quite accurate. Based on disagreements and error analysis, we added more details to our annotation guidelines (Table III), and conducted round-2 of guided annotations. This resulted into 11% improvement in F1-measure, 3% decrease in precision, 26% improvement in recall. Also, we noticed a 38% improvement in the inter-annotator agreement from round-1 to round-2, and the final agreement between the two annotators was 76.2%.

The round-1 for Set-2 was conducted with annotation guidelines, and achieved a high (0.97) precision performance, similar to that of Set-1. The guidelines were further revised based on disagreements and error analysis. In round-2, F1-measure, precision, and recall, improved by 4%, 1%, and 7%, respectively. Also, a 20% improvement in inter-annotator agreement was observed, and the final agreement was 93.9%.

C. Error Analysis

In the absence of a detailed guidelines during round-1 of Set-1, A1 missed several important indications (Table V row 2), leading to a lower recall (0.78). Both A1 and A2 made a few incorrect judgments, e.g., A2 selected the indications of another drug (Table V row 1). These errors led us to design “what to” #1-5 and “what not to” #1-3 guidelines (Table III). During round-2, the annotators missed only 1 ground truth indication; from the excerpt, “Alprazolam is also indicated for the treatment of panic disorder, with or without agoraphobia,” they did not select agoraphobia. However, their precision declined since they both selected several symptoms as indications, e.g., “Panic disorder is characterized by following symptoms: palpitations, pounding heart, or accelerated heart rate ...” This led us to design the “what not to” #4 guideline.

For round-1 of Set-2, A2 missed several indications from the long labels (600-800 words), generating a relatively lower (0.83) joint recall. Also, A1 made some incorrect judgments, such as, selecting species name as indications, e.g., selecting pneumococcus from “Respiratory tract infections caused by Streptococcus pneumoniae” (pneumococcus was included as a pre-annotation since it is a normalized term for the disease “pneumonia”). This led us to design the “what not to” #5-6 guidelines. In round-2, the few incorrect judgments included symptoms (e.g., obsession, compulsion, sensitivity, etc.), and conditions (e.g., sedation) caused by drugs. Certain important indications were still missed in round-2 because of the presence of long labels in Set-2.

<table>
<thead>
<tr>
<th>TABLE V. SOME ANNOTATIONS' WITH DISAGREEMENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotation by A1</td>
</tr>
<tr>
<td>Cimetidine hydrochloride injection is indicated in short-term treatment of active duodenal ulcer. Concomitant antacids should be given as needed for relief of pain.</td>
</tr>
<tr>
<td>For the prevention and treatment of nausea, vomiting, or dizziness associated with motion sickness.</td>
</tr>
</tbody>
</table>

A annotations in orange, pre-annotations in light-orange

IV. DISCUSSION

A. Study Implications

Identifying specific indications from FDA drug labels is a challenging problem. We find that about half of the disease mentions in a drug label cannot be classified as indications. We used the judgment abilities of two experts to validate the automatically generated pre-annotations for drug labels. Even without detailed annotation guidelines in round-1 of Set-1, the experts jointly delivered a promising performance. The challenges faced by the annotators helped us in developing specific indication annotation guidelines. The performances of the annotators, with guidelines, further validated our hypothesis of using joint performances as gold, and confirmed the usefulness of the proposed framework.

While most failure cases could be handled by improving the annotation guidelines, certain scenarios such as annotating long descriptions, and differentiating symptoms from indications, need further research. We also measured the inter-annotator agreement, which intuitively improved with each round. However, a higher agreement does not necessarily lead to a higher quality, e.g., the performance of annotators on Set-1 is better but their agreement is lower as compared to Set-2. In this study, rather than accomplishing a higher agreement, our goal was to devise a mechanism to assist the experts in annotation, and identify the judgments that could be regarded as gold with high confidence.

B. Contributions

We have proposed a semi-automatic method for building factual, structured, and specific gold standard of drug indications by leveraging the information extracted from the
FDA labels. The automatically generated pre-annotations covered over 95% of the ground truth indications. The annotators spent almost no time in learning to use the annotation interface for validating the pre-annotations. Our preliminary results on 100 labels confirmed that using the proposed framework, even a limited number of annotators can jointly and quickly (3 min/label) produce a gold standard with an average 95% F1-measure performance. In addition, this is the first study involving annotation of drug indications wherein we highlighted the key challenges, and provided specific guidelines for annotators.

C. Limitations and Future Work

One limitation of this study is that the gold standard is built using the judgments of only two annotators. However, the system and the annotation workflow helped produce a high-quality gold standard by combining the common judgments. We plan to confirm these results by involving other experts in the future. The corpus used for experimentation represented only about 4% of the entire prescription drug dataset in DailyMed. We plan to investigate whether lessons learnt through these representative labels could apply on a larger dataset. In addition, the algorithm for preparing pre-annotations needs sophisticated text mining techniques to automatically eliminate some negative pre-annotations, and further assist the annotators. In the future, we plan to evaluate the training ability of the expert-curated gold standard to classify indications from other DailyMed labels. We also intend to systematically compare the gold standard with other resources on drug indications, and to mine finer level of information on indications, such as the type of indications, e.g., acute treatment, prevention, symptomatic treatment, etc.

ACKNOWLEDGMENT

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine, the National Key Technology R&D Program of China (Grant No. 2013BAI06B01), and the Peking Union Medical College Youth Fund. The authors would like to thank the two human annotators, Zanmei Li and Yujing Ji.

REFERENCES


