Original Paper

Using Nonexperts for Annotating Pharmacokinetic Drug-Drug Interaction Mentions in Product Labeling: A Feasibility Study

Harry Hochheiser[1,2], PhD; Yifan Ning[1], MS; Andres Hernandez[1,3], MS; John R Horn[4], PharmD; Rebecca Jacobson[1,2], MS, MD; Richard D Boyce[1], PhD

[1]Department of Biomedical Informatics, School of Medicine, University of Pittsburgh, Pittsburgh, PA, United States

[2]Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, United States

[3]Center for Bioinformatics and Computational Biology, BIOS, Manizales, Colombia

[4]Department of Pharmacy, School of Pharmacy and University of Washington Medicine, Pharmacy Services, University of Washington, Seattle, WA, United States

Corresponding Author:
Harry Hochheiser, PhD
Department of Biomedical Informatics
School of Medicine
University of Pittsburgh
5607 Baum Blvd
Suite 523
Pittsburgh, PA,
United States
Phone: 1 412 648 9300
Fax: 1 412 624 5310
Email: harryh@pitt.edu

Abstract

Background: Because vital details of potential pharmacokinetic drug-drug interactions are often described in free-text structured product labels, manual curation is a necessary but expensive step in the development of electronic drug-drug interaction information resources. The use of nonexperts to annotate potential drug-drug interaction (PDDI) mentions in drug product label annotation may be a means of lessening the burden of manual curation.

Objective: Our goal was to explore the practicality of using nonexpert participants to annotate drug-drug interaction descriptions from structured product labels. By presenting annotation tasks to both pharmacy experts and relatively naïve participants, we hoped to demonstrate the feasibility of using nonexpert annotators for drug-drug information annotation. We were also interested in exploring whether and to what extent natural language processing (NLP) preannotation helped improve task completion time, accuracy, and subjective satisfaction.

Methods: Two experts and 4 nonexperts were asked to annotate 208 structured product label sections under 4 conditions completed sequentially: (1) no NLP assistance, (2) preannotation of drug mentions, (3) preannotation of drug mentions and PDDIs, and (4) a repeat of the no-annotation condition. Results were evaluated within the 2 groups and relative to an existing gold standard. Participants were asked to provide reports on the time required to complete tasks and their perceptions of task difficulty.

Results: One of the experts and 3 of the nonexperts completed all tasks. Annotation results from the nonexpert group were relatively strong in every scenario and better than the performance of the NLP pipeline. The expert and 2 of the nonexperts were able to complete most tasks in less than 3 hours. Usability perceptions were generally positive (3.67 for expert, mean of 3.33 for nonexperts).

Conclusions: The results suggest that nonexpert annotation might be a feasible option for comprehensive labeling of annotated PDDIs across a broader range of drug product labels. Preannotation of drug mentions may ease the annotation task. However, preannotation of PDDIs, as operationalized in this study, presented the participants with difficulties. Future work should test if these issues can be addressed by the use of better performing NLP and a different approach to presenting the PDDI preannotations to users during the annotation workflow.

(JMIR Res Protoc 2016;5(2):e40) doi: 10.2196/resprot.5028

KEYWORDS

crowdsourcing; natural language processing; drug interactions, drug product labeling, structured product labels

Introduction

Exposure to interacting drug combinations can lead to patient harm. Recent estimates indicate that between 5.3% and 14.3% of hospital patients in the United States experience a clinically meaningful alteration in the exposure or response of one drug occurring as a result of coadministration of another drug [1]. Fortunately, such harm can often be avoided by employing appropriate management strategies [2]. Toward that goal, US federal regulations require the mention of known, clinically relevant potential drug-drug interactions (PDDIs) in prescription drug labeling [3,4].

Structured product labels (SPLs) are mandated by the US Food and Drug Administration. The labels, produced by pharmaceutical manufacturers, are presented in a standardized format [5] and approved by regulators. As detailed descriptions subject to regulatory approval, SPLs play a vital role in disseminating drug information. However, the structure in these documents is only in the form of high-level sections such as *Description*, *Indications and Usage*, *Contraindications*, and *Warnings*. Specific PDDI details are given in plain text, tables, and figures within the *Drug Interactions* section or other locations throughout the label. Although future efforts may lead to more structured and therefore more computable labels, the regulatory importance of the SPLs and the legacy labels of more than 16,000 drugs make the labels key resources for drug-drug interaction information.

Unfortunately, product labeling is incomplete. A study of drugs that interact with the narrow therapeutic range drug warfarin found PDDI information deficiencies in 15% of relevant product labels [6]. A broader study of drugs sold in the United States, United Kingdom, and Germany found that a warning about a critical drug interaction was missing from the label of one of the interacting drugs at least 40% of the time [7]. Although publicly available PDDI information sources can serve as useful adjuncts to product label information, these collections are often far from complete. Our recent analysis of 14 collections of PDDI information found significant divergence, with overlap between pairs of sources usually less than 50% [8]. Addressing the issue of missing product label PDDI information is important to better meet the information needs of drug experts, clinicians, and patients.

We hypothesize that a computable representation of PDDIs present in product labels and other high-quality sources will enable novel methods for drug information retrieval that will in turn provide researchers and clinicians with improved capabilities for finding complete and current DDI information. Testing this hypothesis requires an efficient means of generating computable representations of PDDI mentions.

In prior work, we developed a prototype system that used simple named entity recognition (NER) and Semantic Web Linked Data [9] to link claims about PDDIs from publicly available external resources to the *Drug Interactions* section of the product label [10]. Experiments found that our system linked at least one potentially novel interaction (ie, not mentioned in the label) to the *Drug Interactions* section of product labeling for 20 antidepressants. Moreover, there were several cases where all of the PDDI mentions linked to the *Drug Interactions* section for an antidepressant were potentially novel and would complement product label information. For example, an interaction between escitalopram and tapentadol mentioned in the National Drug File-Reference Terminology [11,12] was potentially novel to all 20 escitalopram product labels.

While promising, the simple NER approach often missed potentially important links between the label and other sources. Sophisticated natural language processing (NLP) methods might prove to be more complete, accurate, and scalable than simple NER. However, there is reason to believe that even the best NLP methods would not perform well enough to guarantee automatic identification of all PDDI mentions across all drug product labels. The PDDI NLP algorithm that performed best against the 2013 SemEval Challenge text corpus had a sentence-level recall of 0.81 and a precision of 0.86 (F_1 =0.84)[13]. An algorithm we developed in prior work focusing specifically on NLP identification of pharmacokinetic PDDI mentions within product label sections had a document level recall of 0.84 and a precision of 0.88 (F_1=0.86) [14] (sentence-level performance was not evaluated).

Based on these findings, we have concluded that the involvement of human curators is necessary for the task of generating computable representations of PDDIs present in product labels and other high-quality sources. The use of semiautomatic curation is relatively common in biomedicine [15]. Unfortunately, the high cost of expert annotation is a major potential barrier to further progress. New approaches are needed to increase the scale and quality of data curation.

Replacing experts with nonexpert crowds (crowdsourcing) can increase the feasibility of large-scale annotation tasks for biomedical data [16-18]. Initial efforts at crowdsourcing for the annotation of medical text have found the method to be effective when the workflow is properly managed [16]. Results can be comparable in quality to those obtained via more traditional and expensive expert annotation methods [17]. Crowdsourcing is particularly attractive for obtaining results faster and at a lower cost than other participant recruitment schemes [17]. The Informatics for Integrating Biology and the Bedside (i2b2) 2010 workshop assessment found that a well-selected group of nonexperts could perform extraction of drug information from clinical reports [19]. Other biomedical efforts have applied crowdsourcing to gene-mutation mentions in the biomedical literature [20] and for clinical trial announcements [21]. A study of the feasibility of using people recruited through Amazon's Mechanical Turk to annotate medication indications found that nonexperts could achieve accuracy of greater than 95% on the binary question of whether a medication is an indication for a disease mentioned in the medication's drug label [22]. Similar approaches have been used to engage communities of experts in tackling challenges such as linking medications and problems in clinical texts from electronic medical records [23,24] and developing mappings between institutional procedure descriptions and Logical Observation Identifiers Names and Codes (LOINC) [25,26].

Our experience with NLP methods for extracting PDDI annotations suggests the possibility of using NLP annotation to provide suggestions to human annotators. Previous efforts have explored the possibility of using such preannotation. Hanauer et al [27] found that iterative alternation between human annotation and model building facilitated rapid creation of NLP models. Some comparative studies have shown that preannotation can improve annotator performance relative to unassisted annotation [28-30], but other studies have seen no difference [31].

The goal of this study was to assess the potential feasibility of using persons who are not drug experts in the task of annotating PDDIs mentioned in drug product labels. A secondary goal was to test the influence of NLP assistance on the annotation quality of both experts and nonexperts.

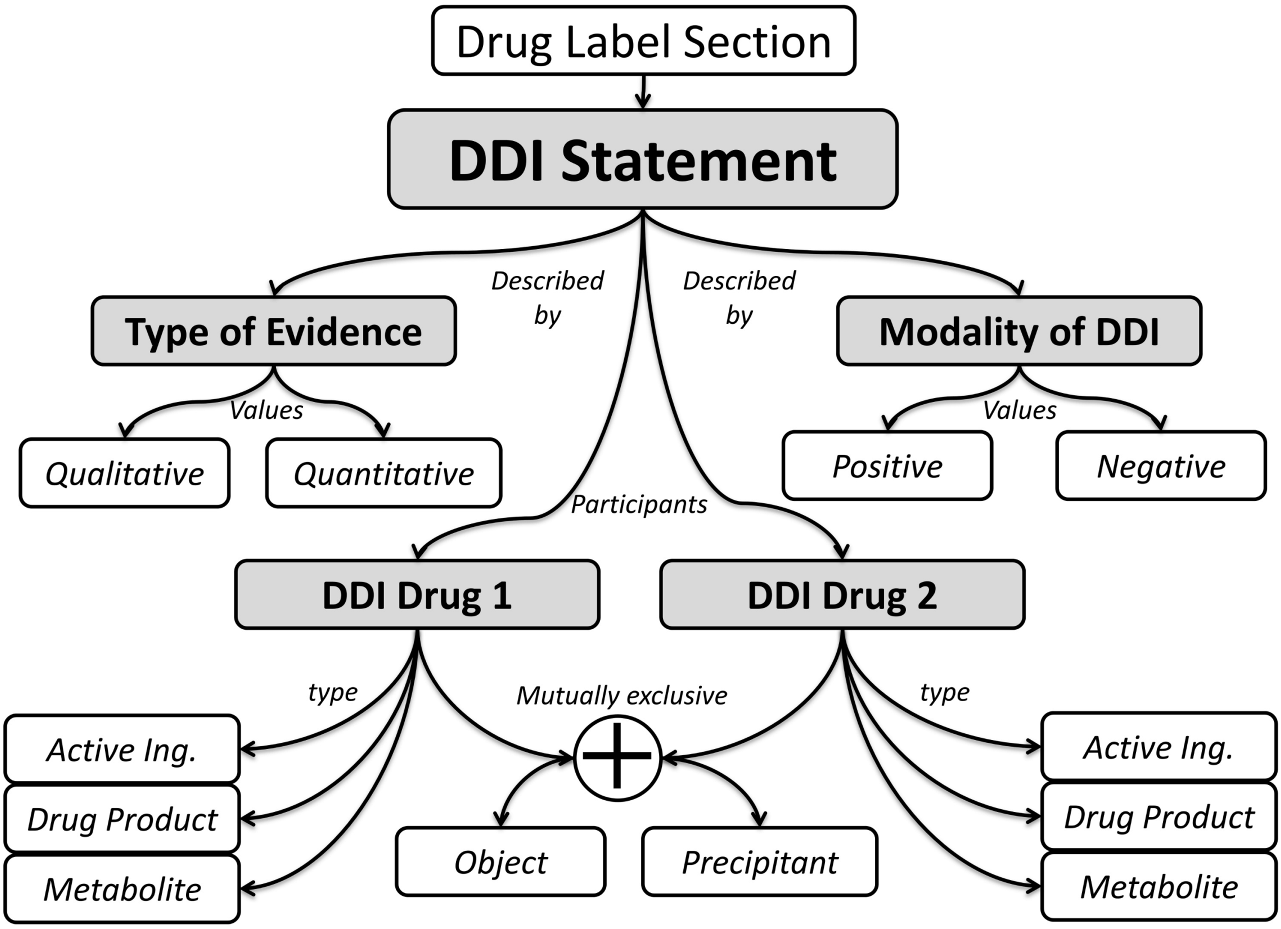
Methods

The Annotation Model

PDDI annotation requires a data model that describes the types of information that must be collected. The PDDI data model used in this study is given in Figure 1. Each PDDI mention is extracted from a span of sentences present within a product label and can include four features:

- Type of evidence (active ingredient, metabolite, or drug product): an active ingredient is a pharmacologically active chemical component used in a drug product. A metabolite is a biochemical entity produced as a result of drug metabolism. A drug product is a packaging of an active ingredient for sale or distribution, often identified by a brand name. Throughout this paper, we use the generic term “drug” to mean any of these three types.
- Role (object or precipitant): the role that each drug plays within the interaction. In pharmacokinetic PDDIs the precipitant drug affects an enzyme that regulates the absorption, distribution metabolism, or excretion of the object drug.
- Statement (quantitative or qualitative): an indication of whether the PDDI mention describes the pharmacokinetic effect of a DDI in quantitative terms (50% increase) or qualitative terms (increase or decrease) with no indication of magnitude.
- Modality (positive or negative): whether the PDDI mention is making a positive or negative claim. A positive claim is one that supports the existence of the interaction. A negative claim is one that explicitly states that no interaction exists between the drugs in question.

Figure 1. Data model used in this study for PDDIs mentioned within drug product labels.

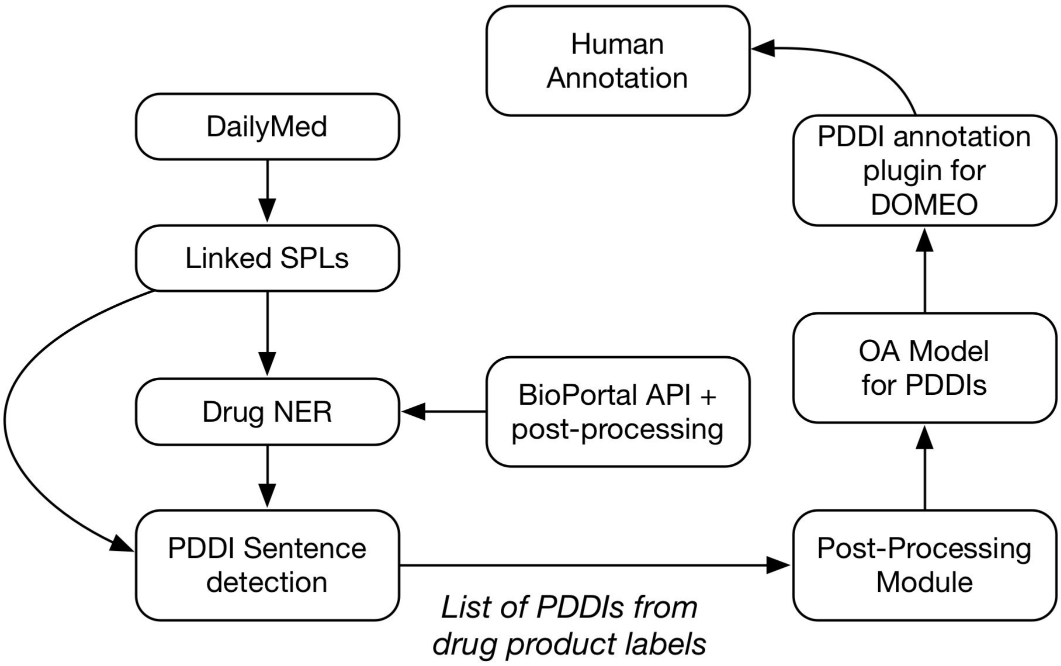


Natural Language Processing Pipeline and Postprocessing Module

In prior work, we developed algorithms for extracting drug named entities and pharmacokinetic PDDI mentions from drug product labels [14]. We integrated our NLP algorithms into a preannotation pipeline (Figure 2). The pipeline used the following steps:

1. The NLP process applies NER to each product label section [32]. The NER algorithm uses the National Center for Biomedical Ontology BioPortal Annotator to extract drug mentions and synonyms from the RxNorm and MeSH terminologies [33]. The results are postprocessed to improve recall and precision by filtering out entities that are not active ingredients, drug products, or metabolites based on entity relationships provided by RxNorm and WordNet [34].
2. Output from the NER process is then processed by an NLP algorithm for identifying pharmacokinetic PDDI mentions [14]. For each product label section, the PDDI extraction algorithm outputs a table of sentence spans labeled as to whether they include a pharmacokinetic PDDI (true or false). Spans including PDDI mentions are also labeled to indicate the modality of the mention (positive or negative). Output of the NLP algorithm is passed to a postprocessing module designed to increase the process's precision and recall (Figure 2). This module uses RxNorm relationships and exact case-insensitive matching to map drug product mentions to unique identifiers of the sole active ingredients.
3. The PDDI mentions present in the corpus are transformed into a machine-readable annotation schema using the Open Annotation data model [35], necessary for subsequent loading into the study annotation tool.
4. Finally, the resulting preannotated named entities and PDDI mentions are loaded into the study annotation tool.

Figure 2. Pipeline for extraction of pharmacokinetic PDDIs from drug labels sections.



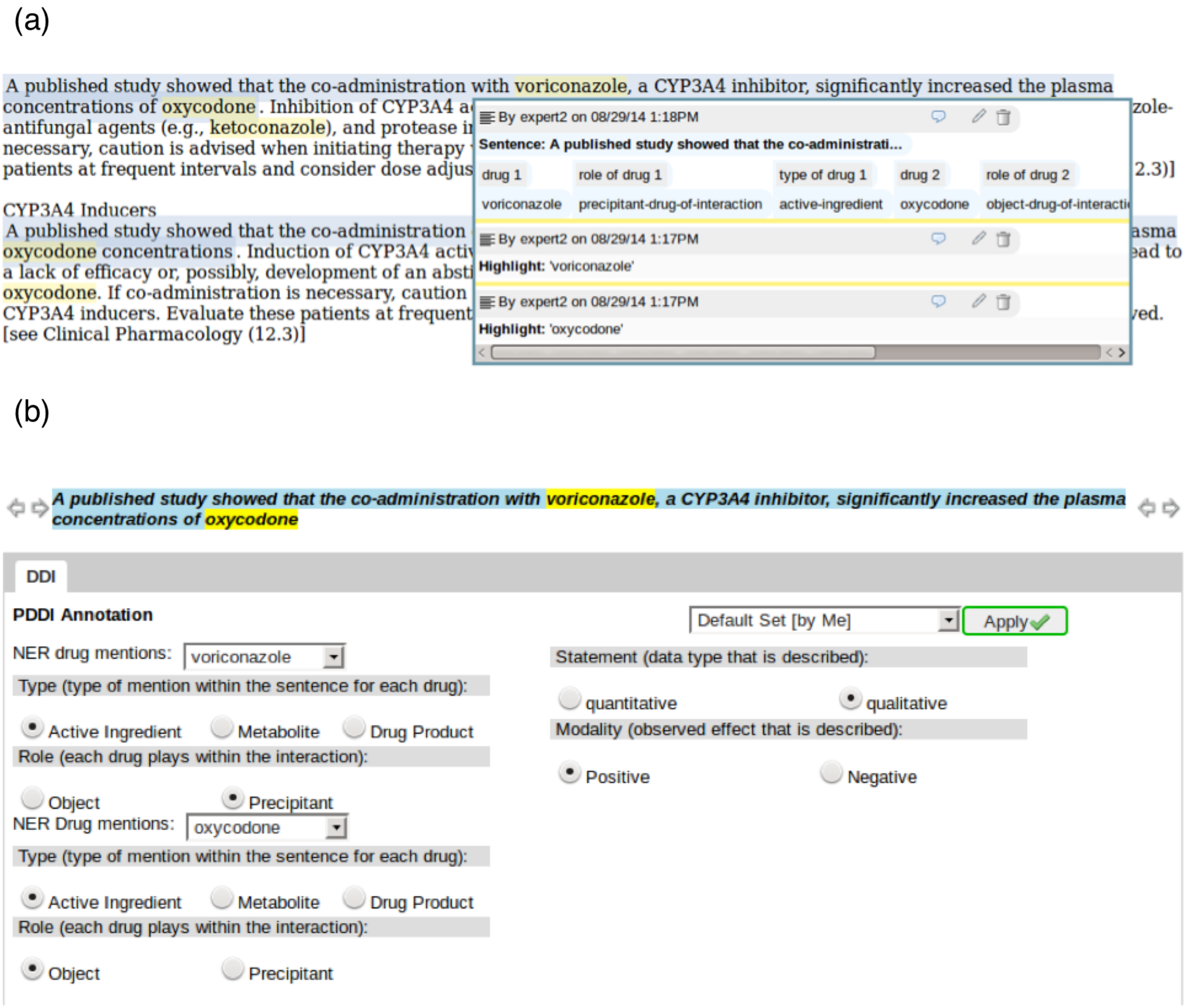
Reference Standard

In an earlier study, we developed a corpus of 208 annotated PDDI statements from SPLs. Two experts in drug information used the data model described above to annotate these sections, with subsequent discussions used to develop a consensus model. The resulting corpus contains 607 pharmacokinetic PDDI mentions along with 3351 active ingredients, 234 drug products, and 201 metabolite mentions [14]. These sections were used in the current study, with the consensus annotations acting as a gold standard.

Drug-Drug Interaction Annotation Tool

Participants used a custom-designed user interface (Figure 3) based on the DOMEO Web-based system [36] to annotate PDDI mentions. DOMEO is an extensible Web application that supports scalable Web-based annotation necessary for crowdsourcing efforts [37]. We extended DOMEO with a plugin that can be used to link text in drug label sections with details of the PDDI data model (Figure 1) [38]. To complete a PDDI annotation task, users would view a product label section and select one or more sentences from the section that discusses the PDDI. They would then use our PDDI annotation plugin to provide values in a Web form indicating the two drugs involved in the interaction, the type and role for each drug, the type of PDDI mention (quantitative or qualitative), and the modality of the mention (positive or negative).

Figure 3. Screenshots of the DOMEO PDDI annotation plugin: (a) product label excerpt with text selected by an annotator as being relevant to a PDDI and (b) form with the fields that the annotator must complete in order to describe the PDDI using the data model described in Figure 1.



A published study showed that the co-administration with voriconazole, a CYP3A4 inhibitor, significantly increased the plasma
concentrations of oxycodone . Inhibition of CYP3A4 a zole-
antifungal agents (e.g., ketoconazole), and protease i
necessary, caution is advised when initiating therapy
patients at frequent intervals and consider dose adjus 2.3)]

CYP3A4 Inducers
A published study showed that the co-administration asma
oxycodone concentrations . Induction of CYP3A4 activ ead to
a lack of efficacy or, possibly, development of an absti
oxycodone. If co-administration is necessary, caution
CYP3A4 inducers. Evaluate these patients at frequent ved.
[see Clinical Pharmacology (12.3)]

By expert2 on 08/29/14 1:18PM

Sentence: A published study showed that the co-administrati...

drug 1	role of drug 1	type of drug 1	drug 2	role of drug 2
voriconazole	precipitant-drug-of-interaction	active-ingredient	oxycodone	object-drug-of-interacti

By expert2 on 08/29/14 1:17PM

Highlight: 'voriconazole'

By expert2 on 08/29/14 1:17PM

Highlight: 'oxycodone'

A published study showed that the co-administration with voriconazole, a CYP3A4 inhibitor, significantly increased the plasma concentrations of oxycodone

DDI

PDDI Annotation

Default Set [by Me] Apply

NER drug mentions: voriconazole

Type (type of mention within the sentence for each drug):

● Active Ingredient ○ Metabolite ○ Drug Product

Role (each drug plays within the interaction):

○ Object ● Precipitant

NER Drug mentions: oxycodone

Type (type of mention within the sentence for each drug):

● Active Ingredient ○ Metabolite ○ Drug Product

Role (each drug plays within the interaction):

● Object ○ Precipitant

Statement (data type that is described):

○ quantitative ● qualitative

Modality (observed effect that is described):

● Positive ○ Negative

Annotation Scenarios

We explored four annotation scenarios aimed at assessing the impact of different approaches to NLP preannotation. The 208 product label sections from our reference standard [14] were distributed across four scenarios so that each scenario had roughly the same number of long and short sections:

- Scenario 1 (no assistance) consisted of 52 label sections with no NLP assistance for annotation. Annotators had to read and highlight all drugs and PDDI mentions within the assigned drug label sections.
- Scenario 2 (drug mentions) consisted of 52 drug label sections with preannotations for drug mentions but not PDDI mentions. Annotators had to correct preannotated drug mentions, identify any drug mentions that the NLP missed, and highlight all PDDIs mentioned in the label sentences.
- Scenario 3 (drug mention plus PDDIs) consisted of 53 label sections preannotated with both drug and PDDI mentions. The annotator had to edit and correct NLP preannotations and add any mentions missed by the NLP.
- Scenario 4 (no assistance, second time), a second completely unassisted scenario, was included with the intent of measuring any learning effects associated with the completion of the NLP-assisted tasks. This scenario consisted of 48 drug label sections.

Each participant completed all four scenarios in order. Three of the 208 sections were reserved for training purposes to familiarize participants with the annotation tool and process, leaving 205 sections to be annotated by each participant.

Participants

A drug expert was defined as a professional in pharmacy or related field with a Doctor of Pharmacy degree or equivalent and more than five years' experience in drug-drug interaction research. A drug nonexpert was defined as an undergraduate or graduate student with some basic training in chemistry. Both expert and nonexpert participants were recruited from personal contacts of the investigative team. All participants were compensated for participating in this study. The University of Pittsburgh Institutional Review Board approved the study protocol as exempt.

Annotator Guidelines and Training

Annotators were provided with guidelines describing the annotation task. Guidelines were written based on assumption of college-level formal training in chemistry (eg, general chemistry) for both groups. The complete guidelines are provided in Multimedia Appendix 1. Participants attended a half-day training session that introduced the goal of the annotation task and provided the annotation guidelines.

Annotation Tasks

Each participant completed all of the four scenarios in the order given above. For each task, the annotators were asked to read the entire content of the relevant drug label sections, identify all drug and PDDI mentions, and record information about the PDDI corresponding to the PDDI annotation model. They were also asked to self-report the amount of time it took to completely annotate each section. The results of each training task were verified to ensure that each annotator completed each scenario according to the study requirements. A short questionnaire completed at the end of each scenario included closed- and open-ended questions about the participant's perception of the usability and effectiveness of the annotation tool and NLP preannotation.

Annotation Performance Metrics

Performance metrics were calculated by comparing the PDDIs in each participant's results with the reference standard described above [14]. User’s annotations were considered true positives if they (a) matched the precipitant, object, and modality of the reference standard and (b) used sentences that either partially or exactly overlapped the sentences used in the reference standard. Metrics were computed by label and then averaged by scenario.

We supplemented the standard metrics of precision, recall, and F_1 with additional metrics to gain more insight into the effect of NLP preannotation on the PDDI annotation task. Specifically, for Scenario 3 (ie, the full NER plus NLP preannotation), we evaluated how often participants decided to change NLP annotations and whether those NLP annotations agreed or disagreed with the reference standard.

Results

Two experts and 4 nonexperts were recruited into the study. One expert left the study after experiencing too many difficulties with the PDDI annotation user interface. One nonexpert left the study because of not having time to complete annotations due to work and school commitments. The remaining participants completed the annotation task for all scenarios.

Annotation performances measured in F_1 score relative to the reference standard [14] indicate relatively strong performance (F_1>0.7) for all participants for the first two scenarios, with a drop in performance for the last two scenarios. (Figure 4 and Table 1; full recall and precision results in Multimedia Appendix 2). The performance of the entire NLP pipeline is included for each scenario for comparison even though participants were only provided PDDI preannotations in Scenario 3.

Tables 2 and 3 summarize self-reported task completion times and subjective feedback across the first three scenarios. As Scenario 4 was conducted solely to assess learning effects, task completion time and subjective responses were not collected. Participants differed in their reports of time required, ranging from Nonexpert 1 reporting times comparable to those of the expert to Nonexpert 3 reporting more than 5 hours spent completing Scenario 3 (Table 2).

Results from the subjective question assessing ease of use are given in Table 3. Users agreed that the PDDI annotation interface was moderately difficult when full preannotation assistance was enabled and also agreed that the PDDI annotation plugin without NLP assistance or using a lower level of assistance is relatively easy to use. Full questionnaires and results are given in multimedia appendices 3-7.

Tables 4 and 5 illustrate the agreement between the participants, NLP, and reference standard in the scenario with NLP and preannotation assistance (Scenario 3). Table 4 addresses performance on the 151 PDDI annotations found in the reference standard, while Table 5 summarizes false positives—mentions extracted in the NLP or by users that were not found in the reference standard.

Though exploratory because of the very small sample size, success in detecting true-positive PDDI mentions missed by the NLP was similar between the expert and nonexperts (Table 4, column 2). The expert also had slightly more false negatives than the nonexpert participants irrespective of whether spans were found by NLP (Table 4, columns 1 and 3). False-positive rates for the expert were comparable to those of the nonexperts (Table 5, column 1). Nonexperts also seemed to be slightly more likely to agree with false-positive mentions extracted by the NLP (Table 5, columns 2 and 3).

Table 1. F_1 measures for all participants and NLP system across all scenarios and overall.

Annotator	Scenario 1[a]	Scenario 2[b]	Scenario 3[c]	Scenario 4[d]	Overall
Expert	0.80	0.79	0.54	0.66	0.68
Nonexpert 1	0.79	0.83	0.59	0.53	0.66
Nonexpert 2	0.76	0.68	0.57	0.70	0.67
Nonexpert 3	0.74	0.62	0.53	0.62	0.61
NLP	0.58	0.40	0.41	0.46	0.46

[a] No assistance.

[b] Preannotation of drug mentions.

[c] Preannotation of drug mentions and PDDIs.

[d] No assistance.

Table 2. Participant self-reported task completion times.

Participant	Scenario	<1 hour	1-3 hours	3-5 hours	>5 hours
Expert					
	1	X			
	2	X			
	3	X			
Nonexpert 1					
	1	X			
	2	X			
	3	X			
Nonexpert 2					
	1		X		
	2	X			
	3	X			
Nonexpert 3					
	1			X	
	2			X	
	3				X

Table 3. Usability questionnaire results. All results reported on a 5-point scale (1=*very difficult* to 5=*very easy*).

Participant	Scenario 1	Scenario 2	Scenario 3	Mean
Expert	2	4	2	2.67
Nonexpert 1	4	5	2	3.67
Nonexpert 2	4	5	2	3.67
Nonexpert 3	3	3	2	2.67
Mean	3.25	4.25	2	—

Table 4. Comparison of agreement between the participants, NLP preannotation, and PDDI annotations (N=151) in the reference standard during the scenario with NER and NLP preannotation assistance (Scenario 3).

NLP Result	No mention found		Mention	
Participant	No mention	Mention[a]	No mention[b]	Mention
	NLP FN[f] User FN n (%)	NLP FN User TP n (%)	NLP TP[c] User FN n (%)	NLP TP User TP n (%)
Expert	59 (39.1)	50 (33.1)	23 (15.2)	19 (12.6)
Nonexpert 1	46 (30.5)	63 (41.7)	11 (7.3)	31 (20.5)
Nonexpert 2	43 (28.5)	66 (43.7)	11 (7.3)	31 (20.5)
Nonexpert 3	49 (32.5)	60 (39.7)	13 (8.6)	29 (19.2)

[a]Indicates case where the user corrected an NLP error.

[b]Indicates cases where the NLP was correct and the user was incorrect.

[c]FN: false negative

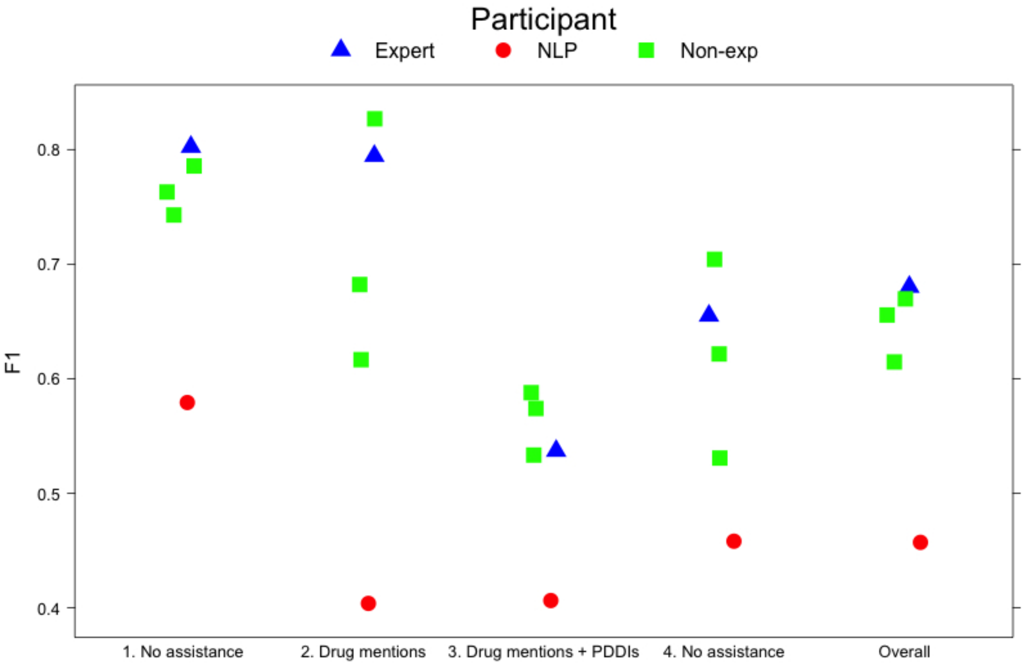
[d]TP: true positive

Table 5. Analysis of user and NLP false positives relative to the reference standard for Scenario 3.

NLP result	No mention	Mention (n=93)	
Participant	Mention[a]	No mention[b]	Mention
	NLP TN User FP n	NLP FP User TN n (%)	NLP FP User FP n (%)
Expert	25	93 (100)	0 (0)
Nonexpert 1	16	88 (94.6)	5 (5.4)
Nonexpert 2	37	86 (92.5)	7 (7.5)
Nonexpert 3	24	88 (94.6)	5 (5.4)

[a]Indicates cases where the user identified spans that were not identified by the NLP. [b]Indicates cases where the NLP identified spans that the participant did not annotate.

Figure 4. Annotator and NLP performance (F1 scores) for each of the four scenarios and overall performance across all four scenarios.



Discussion

Overview

Our long-term goal is to develop tools that will deliver computable representations of reliable, accurate PDDI information to clinicians, facilitating decision support and hopefully reducing adverse events. The number of drugs that might need to be addressed (more than 16,000) and the complexity of the content in the SPLs make this a daunting task. Our experience in building NLP tools for the extraction of PDDI information [10,14] illustrated some of the difficulty and led us to the conclusion that some amount of manual involvement in the process was required.

Our goal in this study was to address two key questions in the development of human-assisted processes for curating PDDI information. Specifically, who should conduct the annotation and what sort of assistance should they receive? Although pharmacists and other domain experts familiar with drug information presumably have the background and training necessary to interpret SPLs, annotation by experts is often prohibitively difficult. Thus, we set out to gain some preliminary insight into the practicality of asking participants not specifically trained in drug information to annotate this data. Second, we were interested in understanding what level of assistance might be helpful for users. If our NLP tools were found to speed completion of annotation tasks without reducing accuracy, this would decrease the cost of PDDI annotation even for nonexperts.

Is It Possible for Nonexperts to Produce Reliable PDDI Annotations From Drug Labels?

Annotation results from the nonexpert group were relatively strong in every scenario and better than the performance of the NLP pipeline (Figure 4). These findings suggest that nonexperts might be able to produce reliable PDDI annotations from drug labels with accuracy levels similar to those of experts and that crowdsourcing might be a feasible option for annotating PDDIs across a broader range of drug product labels. Our results are consistent with earlier demonstrations of the feasibility of applying crowdsourcing to related problems in annotation of biomedical texts [19-22].

Ensuring the success of nonexpert annotations of PDDI mentions will likely require greater attention to two keys issues: the usability of the annotation tools and the selection of the annotators.

Although self-reported task completion times (Table 2) indicated that two of the nonexperts were able to complete all tasks in times comparable to those of the expert, one nonexpert (Nonexpert 3) needed substantially more time. Differences in F_1 scores (Table 1) suggest that annotations provide by Nonexpert 3 were of slightly lower quality than those of the other two nonexperts. The combination of increased task-completion time and lower F_1 scores suggest that Nonexpert 3 may have struggled more than the other participants with the annotation task. In addition, difficulties with the annotation interface prevented one expert user from completing the annotation tasks.

Despite these difficulties, responses to the usability questions (Table 3) were generally positive, suggesting that usability concerns should not be insurmountable. The small sample size and self-reported time results limit our ability to develop a nuanced understanding of specific issues that might have led to increased task completion times or dissatisfaction with the user interface. Observational user studies, including think-aloud feedback from participants, would likely provide insight into usability problems, potential opportunities for redesign [39], and any difficulties associated with the longer task completion times and lower performance of Nonexpert 3.

Results from the repeated no assistance scenario (Scenario 4) do not appear to show any learning effect based on the previous three scenarios. Exposure to the preannotations in scenarios 2 and 3 may have confused participants, pointing out complexities in interpretation of the labels that might have negatively impacted performance.

Identification of individuals who are likely to produce high-quality results will be a key challenge for successful nonexpert annotation results. Although our nonexpert participants all had relevant educational backgrounds and computer experience, variations in the task completion times and F_1 scores suggest that some participants might find PDDI annotations more approachable than others. Future nonexpert PDDI annotation recruitment might draw on experience from prior efforts in crowdsourcing which have found that appropriate screening and training of participants can help improve outcomes [40,41].

What Is the Influence of NLP Assistance on Annotation Quality?

Most participants perceived PDDI annotation to be easier when NER preannotation was provided. However, the full NLP assistance (Scenario 3: drug mention plus PDDI preannotations) was associated with lower levels of perceived usability for both expert and nonexpert participants (Table 2). Complaints about deleting false positives were commonly expressed in the questionnaire data. These results suggest that the performance of the NLP algorithm and the presentation of NLP preannotated PDDIs might have adversely impacted participant performance. Participants suggested several possible improvements, including preannotating with NER and then presenting NLP preannotations only after a section is annotated. The purpose then would be to highlight possibly missed interactions. We think this approach would depend on an NLP algorithm with much better sentence-level performance than the algorithm used in this study.

Although exploratory, the comparison of the agreement between users, NLP preannotations, and the reference standard (tables 4 and 5) suggests several questions for future study. The expert was slightly less likely than the nonexperts to correct NLP false negatives (Table 4, column 1), possibly because the expert might have inappropriately used knowledge of the domain or applied an overly strict interpretation of the PDDI identification guidelines. The expert user was also more likely to reject a correct NLP interpretation (Table 4, column 3) and more likely to reject an incorrect NLP assertion (Table 5, column 2) suggesting that the expert user's thought processes were somehow different than those of the nonexperts. It is also possible that the nonexpert agreement with NLP false positives might be associated with greater trust in NLP on the part of the nonexpert participants. Of course, given the small size of this study, it is entirely possible that these participant-level observations are not statistically significant. Subsequent studies involving more participants and including investigation of user thought processes—perhaps via think-aloud protocols or retrospective interviews—would be needed to understand these phenomena.

Limitations

The generalizability of this study is limited by the small sample size; a larger study would be needed to more accurately characterize the differences between nonexperts, experts, and the NLP annotation. Another potential limitation of our study is that we could not evaluate the characteristics of label sections that might be more difficult to read and annotate by nonexperts. The experimental design attempted to address this concern by balancing the number of sections across each scenario to minimize the effect of differences in difficulty level. The study results might have been influenced by the accuracy of the NLP algorithm and the reliability and usability of the annotation user interface. Interface revisions based on usability might lead to improved performance for experts and nonexperts. Finally, as we did not conduct any debriefing interviews or otherwise assess participant mental states, we are only able to speculate as to factors that might contribute to differences in task performance.

Conclusions

Our goal was to explore of use of nonexperts to annotate PDDI mentions in drug product labels. Our results suggest that nonexperts could produce reliable PDDI annotations from drug labels with efficiency comparable to that of an expert annotator with training in pharmacy or pharmaceutics, indicating that the task of extracting PDDIs from drug product labeling might be suitable for crowdsourcing. Although NER preannotation was found useful to both experts and nonexperts, NLP preannotation as implemented in this study seemed to present an obstacle to all participants. A high performance NLP algorithm might still be helpful if NLP preannotations are shown to annotators after a section is annotated, if only to highlight possibly missed interactions. Improvements in the usability of the annotation tool and screening of potential annotators might further increase performance.

Acknowledgments

Research reported in this paper was partially supported by the Fogarty International Center of Global Health of the National Institutes of Health under grant number 1D43TW008443-01, National Institute on Aging grant K01AG044433, and National Library of Medicine grant 1R01LM011838-01. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We also thank Paolo Ciccarese and Tim Clark for their advice on the development of the PDDI annotation plugin for DOMEO.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Annotation guidelines provided to study participants.

[PDF File (Adobe PDF File), 2MB-Multimedia Appendix 1]

Multimedia Appendix 2

Full precision, recall and F1 results for all participants and for the NLP pipeline in each of the four scenarios.

[PDF File (Adobe PDF File), 64KB-Multimedia Appendix 2]

Multimedia Appendix 3

Subjective questionnaires.

[PDF File (Adobe PDF File), 44KB-Multimedia Appendix 3]

Multimedia Appendix 4

Subjective responses, Scenario 1.

[XLSX File (Microsoft Excel File), 26KB-Multimedia Appendix 4]

Multimedia Appendix 5

Subjective responses, Scenario 2.

[XLSX File (Microsoft Excel File), 29KB-Multimedia Appendix 5]

Multimedia Appendix 6

Subjective responses, Scenario 3.

[XLSX File (Microsoft Excel File), 35KB-Multimedia Appendix 6]

Multimedia Appendix 7

Subjective responses, Scenario 4.

[XLSX File (Microsoft Excel File), 25KB-Multimedia Appendix 7]

References

1. Magro L, Moretti U, Leone R. Epidemiology and characteristics of adverse drug reactions caused by drug-drug interactions. Expert Opin Drug Saf 2012 Jan;11(1):83-94. [doi: 10.1517/14740338.2012.631910] [Medline: 22022824]
2. Hansten PD. Drug interaction management. Pharm World Sci 2003 Jun;25(3):94-97. [Medline: 12840961]
3. Dal-Ré R, Pedromingo A, García-Losa M, Lahuerta J, Ortega R. Are results from pharmaceutical-company-sponsored studies available to the public? Eur J Clin Pharmacol 2010 Nov;66(11):1081-1089. [doi: 10.1007/s00228-010-0898-y] [Medline: 20844869]
4. Code of Federal Regulations Title 21. Washington DC: US Food and Drug Administration URL: http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?fr=201.57 [accessed 2015-06-10] [WebCite Cache ID 6ZBdCKzvJ]
5. Guidance for industry: indexing structured product labeling. Washington DC: US Food and Drug Admininstration URL: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm072317.pdf [accessed 2015-11-25] [WebCite Cache ID 6dJGbbaTV]
6. Hines LE, Ceron-Cabrera D, Romero K, Anthony M, Woosley RL, Armstrong EP, et al. Evaluation of warfarin drug interaction listings in US product information for warfarin and interacting drugs. Clin Ther 2011 Jan;33(1):36-45. [doi: 10.1016/j.clinthera.2011.01.021] [Medline: 21397772]
7. Pfistermeister B, Saß A, Criegee-Rieck M, Bürkle T, Fromm MF, Maas R. Inconsistencies and misleading information in officially approved prescribing information from three major drug markets. Clin Pharmacol Ther 2014 Nov;96(5):616-624. [doi: 10.1038/clpt.2014.156] [Medline: 25062063]
8. Ayvaz S, Horn J, Hassanzadeh O, Zhu Q, Stan J, Tatonetti NP, et al. Toward a complete dataset of drug-drug interaction information from publicly available sources. J Biomed Inform 2015 Jun;55:206-217 [FREE Full text] [doi: 10.1016/j.jbi.2015.04.006] [Medline: 25917055]

9. Marshall MS, Boyce R, Deus HF, Zhao J, Willighagen EL, Samwald M, et al. Emerging practices for mapping and linking life sciences data using RDF: a case series. Web Semantics: Science, Services and Agents on the World Wide Web 2012 Jul;14(4):2-13. [doi: 10.1016/j.websem.2012.02.003]
10. Boyce RD, Horn JR, Hassanzadeh O, Waard AD, Schneider J, Luciano JS, et al. Dynamic enhancement of drug product labels to support drug safety, efficacy, and effectiveness. J Biomed Semantics 2013;4(1):5 [FREE Full text] [doi: 10.1186/2041-1480-4-5] [Medline: 23351881]
11. Brown SH, Elkin PL, Rosenbloom ST, Husser C, Bauer BA, Lincoln MJ, et al. VA National Drug File Reference Terminology: a cross-institutional content coverage study. Stud Health Technol Inform 2004;107(Pt 1):477-481. [Medline: 15360858]
12. Olvey EL, Clauschee S, Malone DC. Comparison of critical drug-drug interaction listings: the Department of Veterans Affairs medical system and standard reference compendia. Clin Pharmacol Ther 2010 Jan;87(1):48-51. [doi: 10.1038/clpt.2009.198] [Medline: 19890252]
13. Bui Q, Sloot PM, van Mulligen EM, Kors JA. A novel feature-based approach to extract drug-drug interactions from biomedical text. Bioinformatics 2014 Dec 1;30(23):3365-3371 [FREE Full text] [doi: 10.1093/bioinformatics/btu557] [Medline: 25143286]
14. Boyce R, Gardner G, Harkema H. Using natural language processing to identify pharmacokinetic drug-drug interactions described in drug package inserts. 2012 May 07 Presented at: Proceedings of the Workshop on Biomedical Natural Language Processing; 2012; Montreal, Quebec, Canada p. 206-213 URL: https://dl.acm.org/citation.cfm?id=2391151
15. Rzhetsky A, Shatkay H, Wilbur WJ. How to get the most out of your curation effort. PLoS Comput Biol 2009 May;5(5):e1000391 [FREE Full text] [doi: 10.1371/journal.pcbi.1000391] [Medline: 19461884]
16. Good BM, Su AI. Crowdsourcing for bioinformatics. Bioinformatics 2013 Aug 15;29(16):1925-1933 [FREE Full text] [doi: 10.1093/bioinformatics/btt333] [Medline: 23782614]
17. MacLean DL, Heer J. Identifying medical terms in patient-authored text: a crowdsourcing-based approach. J Am Med Inform Assoc 2013;20(6):1120-1127 [FREE Full text] [doi: 10.1136/amiajnl-2012-001110] [Medline: 23645553]
18. Khare R, Good BM, Leaman R, Su AI, Lu Z. Crowdsourcing in biomedicine: challenges and opportunities. Brief Bioinform 2016 Jan;17(1):23-32. [doi: 10.1093/bib/bbv021] [Medline: 25888696]
19. Uzuner O, Solti I, Xia F, Cadag E. Community annotation experiment for ground truth generation for the i2b2 medication challenge. J Am Med Inform Assoc 2010;17(5):519-523 [FREE Full text] [doi: 10.1136/jamia.2010.004200] [Medline: 20819855]
20. Burger JD, Doughty E, Khare R, Wei C, Mishra R, Aberdeen J, et al. Hybrid curation of gene-mutation relations combining automated extraction and crowdsourcing. Database (Oxford) 2014 [FREE Full text] [doi: 10.1093/database/bau094] [Medline: 25246425]
21. Zhai H, Lingren T, Deleger L, Li Q, Kaiser M, Stoutenborough L, et al. Web 2.0-based crowdsourcing for high-quality gold standard development in clinical natural language processing. J Med Internet Res 2013;15(4):e73 [FREE Full text] [doi: 10.2196/jmir.2426] [Medline: 23548263]
22. Khare R, Burger JD, Aberdeen JS, Tresner-Kirsch DW, Corrales TJ, Hirchman L, et al. Scaling drug indication curation through crowdsourcing. Database (Oxford) 2015 [FREE Full text] [doi: 10.1093/database/bav016] [Medline: 25797061]
23. McCoy AB, Wright A, Krousel-Wood M, Thomas EJ, McCoy JA, Sittig DF. Validation of a Crowdsourcing Methodology for Developing a Knowledge Base of Related Problem-Medication Pairs. Appl Clin Inform 2015;6(2):334-344. [doi: 10.4338/ACI-2015-01-RA-0010] [Medline: 26171079]
24. 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 Inform Assoc 2012;19(5):713-718 [FREE Full text] [doi: 10.1136/amiajnl-2012-000852] [Medline: 22582202]
25. Vreeman DJ, Hook J, Dixon BE. Learning from the crowd while mapping to LOINC. J Am Med Inform Assoc 2015 Nov;22(6):1205-1211. [doi: 10.1093/jamia/ocv098] [Medline: 26224334]
26. Dixon BE, Hook J, Vreeman DJ. Learning from the crowd in terminology mapping: the LOINC experience. Lab Med 2015;46(2):168-174. [doi: 10.1309/LMWJ730SVKTUBAOJ] [Medline: 25918199]
27. Hanauer D, Aberdeen J, Bayer S, Wellner B, Clark C, Zheng K, et al. Bootstrapping a de-identification system for narrative patient records: cost-performance tradeoffs. Int J Med Inform 2013 Sep;82(9):821-831. [doi: 10.1016/j.ijmedinf.2013.03.005] [Medline: 23643147]
28. Gobbel GT, Garvin J, Reeves R, Cronin RM, Heavirland J, Williams J, et al. Assisted annotation of medical free text using RapTAT. J Am Med Inform Assoc 2014;21(5):833-841 [FREE Full text] [doi: 10.1136/amiajnl-2013-002255] [Medline: 24431336]
29. Lingren T, Deleger L, Molnar K, Zhai H, Meinzen-Derr J, Kaiser M, et al. Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements. J Am Med Inform Assoc 2014;21(3):406-413 [FREE Full text] [doi: 10.1136/amiajnl-2013-001837] [Medline: 24001514]

30. Fort K, Sagot B. Influence of pre-annotation on POS-tagged corpus development. 2010 Presented at: Proceedings of the Fourth Linguistic Annotation Workshop; 7/15/2010; Uppsala, Sweden p. 56-63 URL: http://www.aclweb.org/anthology/W10-1807
31. South BR, Mowery D, Suo Y, Leng J, Ferrández Ã, Meystre SM, et al. Evaluating the effects of machine pre-annotation and an interactive annotation interface on manual de-identification of clinical text. J Biomed Inform 2014 Aug;50:162-172 [FREE Full text] [doi: 10.1016/j.jbi.2014.05.002] [Medline: 24859155]
32. Boyce R, Guzman R, Gardner G. SPL Drug Name Entity Recognizer. URL: https://github.com/dbmi-pitt/u-of-pitt-SPL-drug-NER [accessed 2015-06-09] [WebCite Cache ID 6ZAD5y4tT]
33. Shah NH, Bhatia N, Jonquet C, Rubin D, Chiang AP, Musen MA. Comparison of concept recognizers for building the Open Biomedical Annotator. BMC Bioinformatics 2009;10 Suppl 9:S14 [FREE Full text] [doi: 10.1186/1471-2105-10-S9-S14] [Medline: 19761568]
34. Miller GA. WordNet: a lexical database for English. Commun. ACM 1995;38(11):39-41. [doi: 10.1145/219717.219748]
35. World Wide Web Consortium (W3C). W3C Open Annotation Data Model URL: http://www.openannotation.org/spec/core/ [accessed 2015-06-11] [WebCite Cache ID 6ZD4c1fEe]
36. Ciccarese P, Shotton D, Peroni S, Clark T. CiTO + SWAN: The web semantics of bibliographic records, citations, evidence and discourse relationships. Semantic Web 2012:1-28 [FREE Full text]
37. Ciccarese P, Ocana M, Garcia Castro LJ, Das S, Clark T. An open annotation ontology for science on web 3.0. J Biomed Semantics 2011;2 Suppl 2:S4 [FREE Full text] [doi: 10.1186/2041-1480-2-S2-S4] [Medline: 21624159]
38. Ning Y, Hernandez A, Hochheiser H, Ciccarese P, Boyce R. DOMEO Client Drug-drug Interaction Plugin. URL: https://github.com/rkboyce/DomeoClient [accessed 2015-06-10] [WebCite Cache ID 6ZBq8SVE0]
39. Lazar J, Feng J, Hochheiser H. Research Methods in Human-Computer Interaction. London: Wiley; 2009.
40. Downs J, Holbrook M, Sheng S, Cranor L. Are your participants gaming the system? Screening Mechanical Turk workers. 2010 Presented at: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; April 10-15, 2010; Atlanta, GA p. 2399-2402. [doi: 10.1145/1753326.1753688]
41. Mitra T, Hutto C, Gilbert E. Comparing person- and process-centric strategies for obtaining quality data on Amazon Mechanical Turk. 2015 Presented at: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems; 2015; Seoul, Republic of Korea p. 1345-1354. [doi: 10.1145/2702123.2702553]

Abbreviations

NER: named entity recognizer
NLP: natural language processing
SPL: structured product label
PDDI: potential drug-drug interactions

Edited by G Eysenbach; submitted 17.08.15; peer-reviewed by K Doing-Harris, P Schulz; comments to author 12.11.15; revised version received 25.11.15; accepted 19.12.15; published 11.04.16

Please cite as:
Hochheiser H, Ning Y, Hernandez A, Horn JR, Jacobson R, Boyce RD
Using Nonexperts for Annotating Pharmacokinetic Drug-Drug Interaction Mentions in Product Labeling: A Feasibility Study
JMIR Res Protoc 2016;5(2):e40
URL: http://www.researchprotocols.org/2016/2/e40/
doi: 10.2196/resprot.5028
PMID: 27066806