Automated Suggestion of Tests for Identifying Likelihood of Adverse Drug Events

Ritwik Banerjee*, Yejin Choi*, Gaurav Piyush*, Ameya Naik* and I. V. Ramakrishnan*
Department of Computer Science
Stony Brook University
Stony Brook, NY 11794-4400

Abstract—Adverse drug events (ADE) caused by use, misuse or sudden discontinuation of medications trigger hospital emergency room visits. Information about a wide range of drugs and associated ADEs is provided in online drug databases in the form of narrative texts. Even though some ADEs can be detected by observable symptoms, several others can only be confirmed by laboratory tests. In this paper, we present a system that provides automated suggestion of tests to identify the likelihood of ADEs. Given a patient’s medications and an optional list of signs and symptoms, our system automatically produces the necessary laboratory tests needed to confirm possible ADEs associated with these drugs. The basis of our application is to map clinical symptoms to medical problems and laboratory tests. Towards that, we use template-based extraction and shallow parsing techniques from natural language processing to extract information from narrative texts. We employ relevance ranking measures to establish correspondence between the tests and ADEs. Our evaluation based on a sample set of 40 drugs shows that this system achieves relatively high sensitivity.

I. INTRODUCTION

Adverse Drug Events (ADEs) are unintended and undesired reactions experienced by an individual due to the use, misuse, or discontinuation of medication. Several studies have reported that among the adult population, over 12% of emergency room (ER) visits are caused by ADEs [1]–[3]. Safety and quality of patient healthcare are strengthened when a medical problem caused by a drug is promptly and correctly identified. Evidence of an ADE based on a patient’s clinical symptoms thus provides an important data point for clinical decision making.

As there exist way too many drugs, physicians cannot be expected to have memorized all possible ADEs associated with them. Often this is not a problem due to the availability of electronic pharmaceutical databases (e.g. [4], [5]). To assess the likelihood of ADEs, physicians manually look up pharmaceutical databases that hold extensive information on drugs including narratives of their adverse effects. This manual lookup and review is a time-consuming process, prone to lapsed vigilance, and often brings about a failure to order appropriate diagnostic tests [6].

Some ADEs can be detected by observable symptoms, but several others can only be confirmed by laboratory tests. And even though laboratory testing is the single highest-volume medical activity driving clinical decision making, the process of ordering diagnostic tests and acting upon them remains vulnerable to errors [7], [8].

In this paper, we propose a clinical decision support (CDS) system which automatically pushes laboratory test suggestions to confirm (or invalidate) potential adverse effects of a patient’s drug regimen. Our application exploits natural language information provided in online pharmaceutical databases. To this end, we use natural language processing (NLP) and template-based techniques to extract three types of information:

$\mathcal{I}_1$ Potential adverse effects of a drug.
$\mathcal{I}_2$ Observable symptoms associates with medical problems.
$\mathcal{I}_3$ Medical problems identified by abnormal values in laboratory test results.

Based on this extracted knowledge, on one hand we map symptoms to medical problems, and the problems to laboratory test results, while on the other hand we map drugs to their adverse effects. Fig. 1 shows a use-case scenario where a patient has been prescribed the drug Metformin, and is experiencing the following symptoms: (a) an upset stomach, (b) difficulty in breathing and (c) overall tiredness.

For information extraction, we use shallow parsing and pattern matching techniques to extract relevant text from available natural language databases. Further, we adapt MetaMap [9], a tool for recognizing medical concepts, to extract medical terms. The three types of information $\mathcal{I}_1$, $\mathcal{I}_2$ and $\mathcal{I}_3$ are obtained by employing these steps on three separate databases, viz. Drugs.com [10], the MedlinePlus encyclopedia [11] and Rush University Medical Center’s health encyclopedia [12], respectively.
Prior research tackling identification problem has focused on
identification of ADEs in patients in narrative form in existing pharmaceutical databases. However, the identification of known ADEs that are described text is expensive. We remark that our problem is not the dis-

sections V and VI explain the SRM and the ranking measures before the experiment results are presented in VII.

II. RELATED WORK

The topic of pharmacovigilance, defined as the science and activities relating to the detection, assessment, understanding and prevention of adverse effects or any other drug-related problem [14], has received substantial research attention in the medical literature (e.g. [15], [16]). Computer-aided approaches in this area can be divided into two orthogonal categories: (a) extraction of new ADEs from population data, and (b) detection or identification of already known ADEs in patients.

Extraction of ADEs based on population studies.

This body of work aims to discover previously unknown ADEs from population-based historical clinical data – primarily a pharmacoepidemiological approach. Much of this work is based on mining biomedical research literature [17–20]. The main drawback is that learning requires large amounts of annotated training data [18], [21], [22], or else suffers low accuracy [19], [23], [24]. Getting large amounts of annotated text is expensive. We remark that our problem is not the dis-

Identification of ADEs in patients.

Prior research tackling identification problem has focused on creating ADE alerts in patient data. In one line of work, alerts

are formulated as rules that get triggered whenever signs and symptoms in patient data satisfy the conditions attached to these rules. The conditions are built with medical terms in the patient’s symptoms and ADE descriptions. These terms are drawn from discharge summaries [25], [26], ADE reports [27], [28] and ambulatory care notes [29]. NLP tools (e.g. MedLEE [30]) have also been used to extract them.

Rules vary in complexity. Simple rules look for keywords or simple linguistic patterns [29], [31], [32]. As noted by Murff et al. [33], such triggers do not achieve a desirable level of accuracy. Additionally, the rules are learned from a set of documents obtained from a small number of hospitals, often one. In other words the data set from which the rules are derived is far from being comprehensive. Not surprisingly, such systems report low accuracy even when employing more sophisticated NLP tools to extract relevant medical terms [25], [26]. More complex decision rules have been used which combine keywords and laboratory findings [27], [28], [34], [35]. Such rules are, however, manually curated. Notable exceptions include the PSIP project [36], which performs more complex semantic mining. Their approach does not, however, extend to external natural language databases.

Identification using external knowledge sources.

Some recent work has explored the utility of available semi-structured knowledge sources (e.g. RxNorm [37]) to detect and prevent ADEs [38]–[41]. They continue to offer shallow coverage, though, because of a focus on very specific drugs (e.g. [38], [41]) or because they continue to build rules manually (e.g. [39]). More importantly, this body of work does not exploit laboratory test data.

In summary, despite substantial research in ADE detection methods, gaps remain. Our approach attempts to fill this gap with three salient characteristics:

1. Extensive utilization of available pharmaceutical databases.
2. Use of automated measures that provide coverage over a large set of drugs.
3. Combines diagnostic test information with ADE information to aid CDS.

III. METHOD OVERVIEW

We use three separate medical knowledge repositories to automatically suggest diagnostic tests that identify potential ADEs. The first, a drug description dataset, is used to extract

In many cases, the output consists of more than one suggestion. This is due to multiple disorders partially or wholly matching the patient’s symptoms, as well as multiple diagnostic tests having the ability to identify a medical problem. Providing a clinician with all such suggestions can lead to alert fatigue [13]. We thus employ a similarity resolution module (SRM) to filter out spurious suggestions, and then compute the relevance of each suggestion to provide a ranked list.

The rest of this paper consists of a discussion of related work in II, followed by an overview of our method in III. The datasets along with the information extraction steps are described in IV. Sections V and VI explain the SRM and the ranking measures before the experiment results are presented in VII.
the adverse effects of drugs. The second, the laboratory test dataset, is used to obtain the medical problems indicated by abnormal test results. The third dataset contains descriptions of various medical problems. We use it to map diseases and disorders to their observable symptoms. The overview of our approach is presented in Fig. 2.

Information extraction from these datasets involves using the medical entity extraction tool MetaMap. It extracts medical terms from natural language data and assigns them semantic types defined by the Unified Medical Language System (UMLS) [42]. For the purposes of this work, we combine several semantic types into two categories, medical problems and drugs, as shown in Table I. The distinction between our semantic categorization and that followed by MetaMap is further explained in Sec. IV-A.

Given a patient’s list of medications and an optional list of exhibited symptoms, our application identifies a list of medical problems \( \{p_1, \ldots, p_k\} \) fitting two criteria:

- Each \( p_i \) is a potential adverse effect of at least one of the drugs being taken by the patient.
- The symptoms of each \( p_i \) match at least some of these exhibited symptoms.

For the second criterion, i.e. matching symptoms, synonyms are resolved so that equivalent symptoms are identified. For example, if the disease description dataset provides “fatigue” as a symptom of an ADE while the input symptoms contain “tiredness”, the synonym resolution module enables identifying them as equivalent. Sec. V provides further details.

Finally, we perform a relevance ranking of laboratory tests based on the similarity between (a) medical problems identified by test result, and (b) the list of problems \( \{p_1, \ldots, p_k\} \) obtained above.

IV. INFORMATION EXTRACTION

In the scope of this work, we extract information to map drugs to their adverse effects, medical problems to their observable symptoms and laboratory tests to the medical problems they identify. These mappings are obtained by extracting information from the drug description dataset, the disease description dataset, and the laboratory test dataset, respectively.

A. The Drug Description Dataset

We extract information about drugs from Drugs.com, which consists of 5,856 drugs. Like the previous dataset, this too is semi-structured, i.e. different types of information are pre-labeled (e.g. sections labeled “adverse effects”, “dosage”, etc). This corpus provides medical knowledge in complex descriptive paragraphs. We adapt MetaMap to extract and categorize relevant information.

<table>
<thead>
<tr>
<th>Medical Entity</th>
<th>Semantic Category</th>
<th>Score (0–1000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subnormal vitamin B12 levels</td>
<td>Laboratory or Test Result</td>
<td>906</td>
</tr>
<tr>
<td>Anemia</td>
<td>Disease or Syndrome</td>
<td>746</td>
</tr>
<tr>
<td>Neuropathy</td>
<td>Disease or Syndrome</td>
<td>1000</td>
</tr>
</tbody>
</table>

B. The Laboratory Test Dataset

For extracting information about laboratory tests and procedures, we use the Rush University Medical Center’s health en-

Medical Entity Recognition (MER) is the first step of our process. A “medical entity” is a given instance of a medical concept or category (e.g. Metformin is an instance of a pharmaceutical substance). As shown in Table II, recognizing such entities consists of (a) detecting the relevant terms and (b) identifying the semantic category. In this section, we first describe how we perform MER from complex narratives, and then present the shallow parsing and template-based extraction technique we use for simpler texts.

1) The Adapted-MetaMap method for MER: In the scope of this work, we use the UMLS semantic types\(^1\) to categorize two types of medical entities: drugs and medical problems.

To extract medical entities from narratives, we use MetaMap, which splits a text into sentences and phrases and subsequently identifies medical terms. Moreover, it maps these terms to UMLS semantic categories, together with associated scores. These available categories, however, do not suffice for our proposed application. Table II presents a typical example: MetaMap identifies “subnormal vitamin B12 levels” as a test result instead of a potential adverse event. To enhance the precision of MetaMap for our application, we make the following revisions to it:

- Filter candidates based on empirically determined threshold on the score.
- Filter candidates based on additional restrictions on semantic types. E.g. the semantic type “Element, Ion, or Isotope” included by MetaMap in the category “Chemicals and Drugs” is excluded from our “Drug” category.
- Include some additional semantic types. E.g. the semantic type “Laboratory or Test Result” included by MetaMap in the category “Phenomena” is included by us in the “Medical Problem” category. The complete list of categories and their constituent UMLS semantic types are provided in Table I.

In the remainder of this paper, MetaMap-based MER revised as above is called adapted-MetaMap.

---

\(^1\)UMLS contains a semantic network which organizes a meta-thesaurus vocabulary of over two million medical concepts into 135 semantic types.
cyclopedia. It comprises of 603 tests, where information about each test is provided in semi-structured narratives. Table. III shows a typical semi-structured template.

1) Shallow Parsing and Template-based Inference: Some information is presented in relatively simple form. These include short sentences that serve as data labels, i.e. sentences from which the type of information can be inferred. Table IV illustrates this inference process with the vitamin B12 level test. Many test results, for example, have two types of abnormal results: (a) results above the reference range, and (b) results below it. Clauses of the form “causes of … deficiency include …” allows our application to relate the subsequently mentioned medical entities to the correct type of test result.

<table>
<thead>
<tr>
<th>Text snippet</th>
<th>Inference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Causes of vitamin B12 deficiency include diseases that cause malabsorption (for example, Celiac disease and Crohn’s disease), …</td>
<td>Subsequent text lists medical problems resulting from vitamin B12 deficiency.</td>
</tr>
</tbody>
</table>

As most natural language data sources (e.g. an online encyclopedia) use similar snippets across most documents, we were able to infer the type of information by using a combination of pattern matching and phrase chunking. The location of the relevant snippets were identified by studying HTML templates of our data sources.

2) Mapping Test Results to Medical Problems: Using the shallow-parsing and template-based inference, we identify the types of abnormal test results (e.g. lower/higher than normal, positive/negative). Further, we use adapted-MetaMap to extract the medical problems associated with such abnormal results. This way, we obtain a directed tripartite graph which maps each abnormal test result to a medical problem (Cf. Fig. 3).

C. The Disease Description Dataset

In order to determine the observable symptoms associated with medical problems, we make use of the National Library of Medicine’s MedlinePlus encyclopedia [11]. Among other things, MedlinePlus contains semi-structured data about the symptoms of various diseases and disorders. To extract these symptoms, the template-based inference methods described earlier suffice. A typical example is illustrated in Table. III.

V. SIMILARITY RESOLUTION

Collating information from multiple sources has its advantages, but natural language data often exhibits great variety. During experimentation, we often found instances where identical or very similar symptoms were expressed in different words or phrases. Furthermore, it was observed that ignoring such variations had a negative effect on identification of relevant ADEs.

To tackle this issue at the simplest level, we resort to semantic similarity measures based on WordNet [43], which is able to directly identify several synonymous terms (e.g. “fatigue” and “tiredness”). Identifying the semantic similarity in a more general sense (i.e. going beyond synonyms), however, requires a more work. Many symptoms are expressed in words or phrases that are used interchangeably. Further, many observable symptoms are expressed in broader semantics than the actual symptom. For example, a patient suffering from a “histamine headache” may have her symptom listed only as “headache”. Such closely related expressions share a type-of relationship. If expression $e_1$ is a type of $e_2$, it is said that $e_1$ is a hyponym of $e_2$, and $e_2$ is a hypernym of $e_1$. WordNet stores hypernymy and hyponymy relations in a directed acyclic graph structure. The further away two words are from each other in this graph, the less similar they are. In other words, the path length $l$ between a term $t$ and its hyponym/hypernym $t'$ is inversely related to their semantic similarity $s(t, t')$, i.e. $l \propto 1/s(t, t')$.

For our purposes, we identify two words as synonymous if they are semantically similar enough. To determine this similarity, we collect symptom terms together with their synonyms as well as their hypernyms and hyponyms. We run several experiments with the path length varying from 0 (i.e. considering only synonyms) to 3. Fig. 4 illustrates the kind of similar terms encountered for path lengths $l = 0, 1$.

VI. RELEVANCE RANKING

The final stage of our application involves ranking several potentially useful laboratory tests according to their relevance to a given set of medical problems. In order to do this, we run adapted-MetaMap on the test descriptions, which are sections
A. Suggestions based on patient’s drug regimen only

When suggesting tests without being provided any of the patient’s symptoms, our application is unaware of the prior likelihood of any ADE. All potential adverse effects are thus treated as equally probable. Therefore, instead of ranking by relevance, we check whether a test suggested by our application has the ability to identify any potential adverse effect associated with the given drug. A total of 226 tests were suggested for the 40 drugs, of which 186 were deemed relevant by a manual verification process. Additionally, 28 laboratory tests deemed relevant by human judges were not suggested by our system.

An analysis of the results showed that in general, test suggestions to identify potential ADEs associated with psychotropic medications performed poorly. The worst results obtained for the drug Adderall, where only 2 of the 11 suggested tests were judged as relevant. For other drugs, our analysis revealed that suggestions for adverse effects involving the nervous system were comparatively less accurate. Table V presents the results of this evaluation.

B. Suggestions based on patient’s drug regimen and symptoms

Our second experiment takes into account the scenario where in addition to the list of medications, we also know a few symptoms exhibited by the patient. We test with each drug twice by providing two sets of symptoms \(s_1\) and \(s_2\) corresponding to different medical problems. For instance, Metformin medication was tested with the two sets \{nausea, vomiting, low blood pressure, abdominal pain\} and \{fatigue, dizziness, dyspnea, diarrhea\}, corresponding to the potential ADEs lactic acidosis and vitamin B12 deficiency, respectively. We thus obtain 80 test inputs for the 40 drugs. For each input, the output is a list of suggested laboratory tests ranked by the Jaccard similarity coefficient as described in Eq. 1.

We also experiment with and without the similarity resolution module (SRM). Results show that resolving similarity leads to a significant improvement in performance when hyponymy/hypernymy is restricted to short paths. Longer paths simply result in more tests being suggested, even if they are for ADEs with quite different symptoms.

Our application provides a list of laboratory tests that can confirm (or invalidate) a medical condition associated with the patient’s symptoms provided that the condition is also a known potential ADE associated with the patient’s drug regimen. The output list is ranked by the relevance of the test in identifying a potential ADE associated with the given set of drugs and symptoms. To evaluate such a list, we use the mean reciprocal rank (MRR) [44], a widely used metric in information retrieval (e.g. [45], [46]) to evaluate systems that provide ranked results instead of a single answer. It is the average of the reciprocal ranks of the output list

\[
MRR = \frac{1}{|S_i|} \sum_{k=1}^{|S_i|} \frac{1}{\text{rank}_{k}} \tag{2}
\]

where \(S_i\) is the input set of drugs and symptoms, and \(\text{rank}_{k}\) is the rank of the correct suggestion for the \(k^{th}\) test input. The preformance measurements are presented in Table VI.

VIII. Conclusion

We presented an approach for automatically suggesting diagnostic tests to identify potential ADEs. We do this by extracting information from available datasets about adverse drug effects, laboratory tests and symptoms of various medical problems. We present experiments that suggest diagnostic tests with and without relevance ranking. Further, we incorporate methods involving hyponymy/hypernymy to resolve semantically similar terms and demonstrate that resolving natural language similarities boosts performance.

Our future work in this direction intends to incorporate multiple corpora for each type of information extraction to improve coverage and to extend the current application to identify adverse effects arising from drug-drug interactions. We also intend to extend the similarity resolution work to account for medical paraphrases.

Acknowledgment

The authors would like to thank Dr. Mark C. Henry, Professor and Chair of the Department of Emergency Medicine at the Stony Brook School of Medicine for his insights and helpful comments.
REFERENCES


