Title: Applying Natural Language Processing to Extract and Codify Adverse Drug Reaction Data in Medication Labels

Authors: Jeffrey Friedlin, D.O [1,2], Jon Duke, M.D.[1,2]

Institutions:

[1] Indiana University School of Medicine, Indianapolis, Indiana 46202
[2] Regenstrief Institute Indianapolis, Indiana 46202

Communication: Jeff Friedlin, D.O.
Regenstrief Institute
Suite 2000
410 West 10th Street
Indianapolis, Indiana 46202
Phone: (317) 423-5539
Fax: (317) 423-5695
Email: jfriedlin@regenstrief.org

Abstract

Objective To develop an automated method of identifying, extracting and codifying adverse drug reaction (ADR) data contained in the Structured Product Labels (SPLs) for drugs to be studied as part of the Observational Medical Outcomes Partnership (OMOP) project.

Background Natural Language Processing (NLP) technology has been used in the medical domain to identify, extract and codify data in biomedical literature and narrative clinical reports. It has also been applied to detect ADRs from narrative clinical reports. We will use NLP to extract ADR data from free text SPLs

Method We modified and enhanced an existing NLP program –the Regenstrief EXtraction Tool (REX) for this project to create the Structured Product Label Information Coder and ExtractoR (SPLICER). SPLICER consists of threes main modules: an SPL Parser an ADR extractor, and a MedDRA mapper. It is programmed to first identify 'raw' ADR terms contained in the SPL, and then map them to their corresponding MedDRA terms. We also performed an evaluation of SPLICER accuracy using a small sample of SPLs.

Results SPLICER processed a total of 80 OMOP drugs that represented 433 drug labels. SPLICER found a total of 40,433 unique ADRs, and achieved a recall, precision and F-measure of 96.3, 97.2 and .97 respectively.

Conclusion SPLICER accurately identified, extracted, and codified the majority of ADR data contained in SPLs. Its output – a structured ADR database – has value for several potential applications.

I. Introduction

The Observational Medical Outcomes Partnership (OMOP)[1] is a public-private partnership designed to help improve the monitoring of drugs for safety. The partnership is conducting a two-year initiative to research methods that are feasible and useful to analyze existing healthcare databases to identify and evaluate safety and benefit issues of drugs already on the market. OMOP draws on the expertise and resources of the pharmaceutical industry, academic institutions, non-profit organizations, the Food and Drug Administration (FDA), and other federal agencies. It is funded and managed through the Foundation for the National Institutes of Health. America's drug-approval process is world-renowned for its rigorous standards on safety and effectiveness, but even with pre-market clinical trials involving thousands of people, it cannot possibly uncover everything about a drug's performance that may occur once it is in use by a much larger and diverse population.

The FDA's Adverse Event Reporting System (AERS) relies primarily on voluntary, spontaneous reporting of suspected drug safety issues by health professionals, patients, and consumers. OMOP is one of a number of activities that is laying the groundwork for a supplemental approach that is systematic, proactive, and cost-effective. Utilizing databases of patient medical records and health insurance claims, researchers are developing and testing various analytical methods for their ability to detect and evaluate drug safety issues over time.

The series of studies will include assessing different types of data from across the United States, developing tools and methods to analyze the databases, and evaluating how

analyses can contribute to decision-making. Together, these studies should provide the objective evidence needed to inform best practices for using such data.

1. ACE Inhibitors
2. Amphotericin B
3. Antibiotics
4. Antiepileptics
5. Benzodiazapines
6. Beta blockers
7. Bisphosphonates
8. Tricyclic antidepressants
9. Typical antipsychotics
10. Warfarin

Figure 1. Drugs/classes in OMOP project.

1. Angioedema
2. Aplastic Anemia
3. Acute Liver Injury
4. Bleeding
5. GI Ulcer Hospitalization
6. Hip Fracture
7. Hospitalization
8. Myocardial Infarction
9. Mortality after MI
10. Renal Failure

Figure 2. Health Outcomes of Interest associated with specific drugs/classes.

One the primary goals of the OMOP project is to ascertain and evaluate methods for identifying drug-condition associations in an observational database. The objective is to implement these methods in such a way so that, when combined with a common data model, they can identify drug-condition associations without the need for extensive customization and tailoring of the methods to specific databases. There are 10 main classes of drugs that will be studied during this phase of the OMOP project as shown in Figure 1.

For each drug or class, there are two groups of outcomes OMOP is focusing on and which the methods developed during the project will be tasked to identify: 1.) Specific Health Outcomes of Interest (HOI) and 2.) Non-specified conditions. OMOP has identified 10 HOIs as shown in Figure 2. These HOIs are outcomes that may be specifically expected based on what is known about the drug or class and the outcomes associated with them. Non-specified conditions are conditions which appear in the manufacture-produced formal drug product labels for each of the drugs/classes.

Drug Product labels are a primary source of drug safety information for physicians. Studies have shown that labeled adverse event warnings can significantly influence prescribing behavior and reduce the incidence of adverse drug events/reactions (ADRs).

The FDA's DailyMed website[2] contains electronic versions of medication product labels, called Structured Product Labels (SPLs). Introduced in 2006, the SPL is a document markup standard approved by Health Level 7 (HL7) and adopted by the FDA. The 'Structured' in 'structured product label' refers not to the *content* of the SPL, but rather to the electronic format that the FDA can process, review and archive. The *content* of SPLs is free text and unstructured. Since its introduction, over 5,500 SPLs in XML format have been submitted to the FDA and made available for online review from the DailyMed website. Because the data in the SPL is in a free text, narrative format, specific data elements contained in them, such as an ADR associated with a drug, cannot be accessed or used by computerized applications. For example, there is no consistency in ADR terminology, which leads to the same ADR being referenced different ways by different SPLs. To illustrate, Stevens-Johnson Syndrome, a life-threatening condition affecting the skin which can be the result of an ADR, is referred to as SJS in some SPLs. 'Liver function tests increased' is variably referred to in SPLs as 'liver function tests elevated', 'LFTs increased', and 'increased liver enzymes'. Such inconsistency in naming ADRs makes it impossible for computerized applications to identify ADRs in SPLs through, for example, Google-like keyword searches. In order to be used by querying systems or other computerized applications, the data in the SPLs must first be identified, extracted, and then converted into a standardized format that can be understood by the

computer. Specifically relating to the OMOP project, all of the ADR data contained in the SPLs of the 80 OMOP-targeted medications needed to be identified, extracted and converted to a standard format *before* any evaluations of methods designed to extract this data from observation databases could take place. In other words, OMOP first needed to know the number and type of ADRs that were associated with a drug before it could test whether a method could accurately identify drug-condition associations. In addition to identifying the what ADRs are associated with each of the OMOP drugs/classes, OMOP also desired to know *where* in the label the information was contained because the section where an ADR is found may have implications as to the serious and/or prevalence of the ADR. A drug label contains several sections (in addition to of course the adverse events section) that may contain ADR information including the black box warnings, precautions, and post-marketing experience sections. All of these data –the ADRs associated with a drug as found in a drug label in a standardized format and the section where the ADR was found- did not exist at the time of the OMOP project and therefore needed to be created.

OMOP had two main options for identifying, extracting and standardizing or codifying the ADR data contained the in 433 SPLs of the OMOP targeted drugs. One option was to use human reviewers to gather this data. Given the volume of information contained in a typical SPL, this manual review would have been extremely time-consuming and costly. The second option was to use natural language processing (NLP) software to automatically collect this data.

NLP technology has been used in the medical domain to identify, extract and codify data in biomedical literature and narrative clinical reports[3-7]. NLP technology has

been applied to detect ADRs from narrative clinical reports[8-10]. To our knowledge, NLP technology has not been used to extract and codify free text ADR contained in SPLs. We developed computer software – Structured Product Label Information Coder and ExtractoR (SPLICER) to extract and codify ADR contained in SPLs. Here we describe the software and our experience.

II Methods

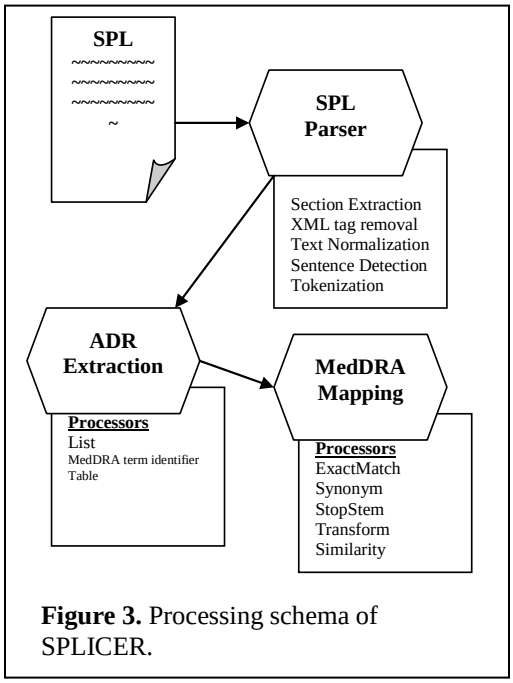

Figure 3. Processing schema of SPLICER.

To process the SPLs, we used a modified version of the Regenstrief EXtraction (REX) tool, which is an NLP software system designed to identify and extract concepts from free-text clinical narratives. Designed in 2005 at the Regenstrief Institute[11], REX has successfully extracted patient data and concepts from radiology reports[12], admission notes[13], microbiology culture results[14] and pathology reports. REX is a rule-based NLP system written in Java that uses regular expressions and algorithms to identify both the concepts as well as the context of the concepts. REX has a modular design which enables straightforward modifications and enhancements to the software. We performed several modifications to the base model of REX for this project to create the SPLICER.

SPLICER is a rule based NLP system, and uses algorithms and technologies similar to those used by other successful medical NLP programs[15]. To develop SPLICER,

we downloaded all SPLs from the DailyMed website, and performed a systematic review of these SPLs to learn the general structure and content of the XML making up the SPL. We identified consistent textual patterns that could be exploited to identify locations of ADRs within the SPL. Will also performed statistical analysis of the SPLs using proprietary text mining software to assist us in identifying text patterns that signal ADRs.

OMOP selected the Medical Dictionary of Regulatory Activities (MedDRA)[16] as the standard terminology to codify ADRs in SPLs. MedDRA is the standard dictionary used by the FDA for adverse event reporting, and over 2,000 organizations use MedDRA to report adverse event data from clinical trials, post-marketing reports and pharmacovigilance. It currently contains over 65,000 adverse event terms and contains five levels of terminology arranged in a hierarchical structure. SPLICER will map ADRs from SPLs to MedDRA LLT terms- the lowest level of the hierarchy that represent a single medical concept.

Based upon this preliminary analysis of SPLs, we separated the task of identifying, extracting and codifying ADR data in SPLs into three main modules. The modules and processing schema of SPLICER are shown in Figure 3. We will describe the development of each of these modules.

SPL Parser Module

SPLICERs first module is called the **SPL Parser**, which identifies sentences within the SPL, normalize the text within the SPL, and generally ‘readies’ the SPL for processing by the ADR extraction module. The FDA requires that all pharmaceutical manufacturers submit their medication information in the Extensible Markup Language

(XML) format. The SPL specification is a Health Level 7 (HL7)[17] standard for medicinal product knowledge in both human readable and computer-interpretable format. It is one of many applications of the HL7 RIM. SPL is the first comprehensive standard of medicinal products and is implemented by the U.S. FDA and all U.S. pharmaceutical industry. The SPL Parser removes XML tags, HTML formatting characters and other control characters that could interfere with NLP algorithms. The SPL parser also identifies and extracts specific SPL sections, such as "Adverse Reactions' and 'Precautions', as well as subsections such as the post-marketing portion of the adverse reactions section. The SPL Parser also identifies and extracts specific data elements from the SPL which are well structured such as pregnancy category, generic and trade medication names, and manufacture date. It also identifies and isolates tables within the adverse reaction section that typically display ADR data from clinical trials. Our initial review of SPLs revealed that the algorithms required to identify and extract ADR data contained in tables needed to be significantly different from the algorithms used to process ADR data in text. Tables in the ADR section of the SPL typically contain several data elements we wished to extract including the including the ADR, the number of patients taking the drug, the number of patients taking placebo and the frequencies of each ADR. We also observed that the structure and format of these tables are inconsistent and highly variable across SPLs. Because of these factors, SPLICER identifies tables and processes them differently than pure text data.

ADR Extraction Module

Once parsing of the SPL is complete, the normalized text from each SPL section will be passed to the **ADR Extraction** module. The goal of this module will be to extract

rarely seen: headache, nausea, rashes and GI upset

adverse events include: headache (3%) , nausea (6%) , rashes (4%) and GI upset (5%)

Figure 4 List-like textual patterns that signal mentions of ADRs in adverse reaction section.

all of the raw mentions of ADRs from the SPL and place them into a raw ADR (RA) table for later processing. The ADR Extraction module creates the RA table by using several pattern recognition algorithms to identify sentence and textual patterns that signal the presence of ADRs. We describe these processors below. Our initial review of a small sample of SPLs revealed a textual pattern which we call a 'list' that signal mentions of ADRs within the adverse reaction section with high consistency.

List Processor. Our initial review of a small sample of SPLs revealed a textual pattern which we call a 'list' that signal mentions of ADRs within the adverse reaction section with high consistency. Drug manufacturers frequently display ADRs in a list pattern with individual ADRs separated by either commas or semi-colons. Figure 4 shows examples of these list patterns. SPLICER exploits this pattern as a means of detecting terms likely to represent ADRs. It recognizes lists by syntactic analysis of sentences, such as examining the ratio of commas or semi-colons to the number of words in the sentence. This syntactic analysis determines if the sentence is a list and thereby triggering the list processor. There are three kinds of list algorithms used depending on the contents of the sentence:

L1 algorithm. This algorithm processes sentences that contain a list of ADRs with no frequency information such as "the following adverse events were seen: headache, nausea, rash and GI upset." For such a list, the L1 algorithm simply extracts the words or phrases between the list delimiters into an ADR table for later processing.

L2 Algorithm. The L2 algorithm processes sentences that contain not only lists of ADRs but frequency information as well, such as "headache (3%), nausea (5%), GI upset (2%)". L2 extracts the ADRs as well as the frequency data associated with them.

L3 Algorithm. The L3 algorithm is similar to L2 in that it processes sentences that contain both ADR and numeric frequency information, but it processes only lists that contain one ADR such as "neurologic: headache (4%)." The processing for these "lists of 1" is different than processing lists containing multiple members since SPLICER cannot exploit the presence of delimiters such as commas or semicolons.

As with all benzodiazepines, paradoxical reactions such as insomnia and hallucination have been reported rarely.

Figure 5 MedDRA terms identified in a sentence by the M1 algorithm.

If by syntactic analysis a sentence is determined not to be a list, the sentence is passed to the M1 algorithm. An example of this type of sentence is shown in Figure 5 with the MedDRA ADR terms highlighted in red.

M1 algorithm. This algorithm simply compares the sentence against the MedDRA database and identifies all MedDRA terms that are exact matches to words in the sentence. This process also looks for synonyms of MedDRA terms by consulting a table containing known, common synonyms to MedDRA terms. The synonym table was created using several different methods. We used proprietary statistical text mining software together with the Unified Medical Language System (UMLS)[18]. The UMLS contains mappings from the MedDRA vocabulary to other vocabularies such as MEdical Subject Headings (MESH)[19] and Systematized Nomenclature of Medicine--Clinical Terms (SNOMED CT)[20]. These mappings can be used to automatically identify synonyms and alternative but semantically equivalent phrasing of MedDRA terms.

Data element
ADR
Patient N for study group
Patient N for placebo group
Number of study patients with ADR
Number of placebo patients with ADR

Table 1 Data elements extracted from SPL tables.

Through manual review of SPLs, we discovered an additional 1074 MedDRA term synonyms not present in the UMLS, which we subsequently added to this synonym table

Once the sentence has been processed by one of the four above named algorithms, the resultant ADR table is then processed to extract only unique ADRs. For ADRs identified by the M1 algorithm, only the ADRs with the longest common string are retained. An example illustrates why this is necessary. For the sentence "Reports of chest pain occurs infrequently" the M1 algorithm finds both "chest pain" and “pain", since both are MedDRA terms. However, for this sentence, we wish to retain the more specific ADR "chest pain", and ignore the more general "pain".

The ADR Extraction Module contains an SPL table processor. The function of this processor is to identify and extract the ADR data contained in tables in the adverse reactions section that typically display information from clinical trials. Data elements extracted from tables are shown in Table 1. Our SPL review revealed wide variability in table structure and format, so we needed to create a rather complex algorithm to achieve accurate table format recognition and precise data extraction. Processing of recognized tables is performed by two separate algorithms described below.

T1 algorithm. This algorithm first determines if the table is in recognizable format, contains up to a maximum of four columns, and whether the table contains data elements (such as medication/ placebo columns, ADR frequency data etc) that can be recognized by SPLICER. If these elements are detected the T1 algorithm proceeds to

process the table by extracting, for each row of the table, the data elements listed in Table 1. The T1 algorithm links the frequency and the N numbers to each ADR extracted.

If T1 determines it cannot process the table, either because the table is not in a format SPLICER understands, is too complex, or in any of the data elements from Table 1 are missing, the table gets passed to the **T1L1 algorithm.** The T1L1 acts like the L1 algorithm in that it treats the ADR column in the table as a list. It extracts all ADRs in this column and places each one into the RA table, ignoring all frequency data in the table.

We wanted to capture ADR frequency data when it is mentioned in the text (non-table) portions of the SPL. We observed that SPLs often contain ADR frequency information that applies to whole groups of ADRs which may not be within the sentence that contains the ADR. An example would be the following:

"The following reactions occurred in <5% of patients:

Neurologic: headache.

GI: nausea, vomiting.

Skin: rash."

Initially, SPLICER processed the text portion of the SPL one sentence at a time and therefore could not reference ADR frequency data across sentences, as is needed in the above example. We programmed SPLICER with a frequency detection algorithm which assigns a frequency context value (FCV) as it processes the SPL, thus enabling the linkage of ADR frequency data across sentences. When SPLICER identifies a non-list sentence, it uses NLP technology to assign the FCV by identifying phrases likely to indicate ADR frequency, such as text phrases such as "infrequently seen", "rare", etc. as

well as numeric frequency phrases such as "<5%". These phrases were discovered through empirical review of SPLs and by using statistical text mining software. The FCV is maintained and applied to all ADRs found in subsequent “list” sentences that follow. When the next non-list sentence is encountered, the frequency detection algorithm is again triggered and the FCV is either replaced with new values or reset to null.

After completely processing the ADR section, SPLICER processes the other sections of the SPL including the overdose, precautions/warnings section, black box warning, indication, and contraindications sections. It processes each section by first normalizing text by removing punctuation, formatting, and control characters as well as removing all stop words. Stop words are common words such as "the","an", “a”,"of", etc. that provide little semantic meaning but can interfere with string comparator algorithms. The **M1 Algorithm** is applied to each normalized section and all MedDRA terms found in each section are saved to an RA table specific for that section. The indication section is subdivided into two subsections using NLP: a high value subsection and a low value subsection. The high value indication subsection contains all sentences that contain phrases specifically mentioning medication indications such as "Zoloft is indicated for depression." The low value indication subsection contains all sentences with no mention of indication. The same methods are used to subdivide the contraindication section.

MedDRA Mapping Module

After all ADR data have been extracted from the SPL and placed into the RA table, it will be passed to the **MedDRA Mapping** module. The goal of this module is to map the ‘raw’ ADR terms (hereafter referred to as SPL terms) in the RA table to their corresponding MedDRA terms. Prior to mapping, SPLICER normalizes the RA table by

removing control characters, punctuation, and extra white spaces, converting the text to lower case, etc. In additional, the RA table is filtered by several processes prior to mapping. First, each SPL term in the table is compared to ADRs extracted from the high value indication section. If a match occurs, the ADR is removed from the RA table. This is to prevent false positive ADRs from being included in the final output. For example, suppose that for medication X an ADR of depression was detected by one of SPLICER's ADR extraction algorithms. If one of the indications for drug X is depression, then it is likely that depression is not a true ADR (we grant that for some medications, an indication may also be an ADR, but our empirical analysis of SPLs revealed this is relatively rare). The next filtering process removes all erroneously captured SPL terms by comparing a list of known non-ADR terms (which we created through analysis of SPLICER output) to the RA table. These include terms such as "laboratory test", "drug interaction", "females", etc.

All SPL terms retained in the RA table after filtering are sent to the MedDRA mapping module where a map to the corresponding MedDRA term is attempted. The mapping module contains numerous mapping algorithms progressing from simple to complex. Subsequent algorithms are triggered and utilized only after previous algorithms fail to find a match. Below we describe the mapping algorithms in the order they are implemented.

ExactMatch algorithm. The goal of this algorithm is to simply map SPL terms to MedDRA terms through exact string matching.

OR Algorithm. This algorithm processes SPL terms that contain the word "or". We discovered that due to the logic of our list extraction algorithms, occasionally SPL

terms actually contain two ADRs, such as "nausea or vomiting". If the OR algorithm detects this pattern, it splits the SPL term at the word ‘or’ and attempts to match each half of the SPL term to a MedDRA term.

Synonym algorithm. The goal of this algorithm is to map SPL terms to MedDRA terms through a synonym table lookup. This process uses the same synonym table discussed during ADR extraction using the M1 algorithm and its purpose is to identify when SPL terms are known synonyms to MedDRA terms.

StopStem algorithm. This algorithm stems and removes all stop words (described earlier during ADR extraction) from the SPL terms as well as from MedDRA terms. Stemming is the process of converting words to their base form. For example the term “rashes” gets stemmed to “rash”. Stopping and stemming SPL terms rarely changes semantic meaning and results in greater likelihood of valid string matches. For example, an extracted SPL term might be “sleep problems” and the matching MedDRA term might be “sleeping problem”. Exact string matching would fail in this instance; however, stopping and stemming the SPL term and the MedDRA term results in a common term- “sleep problem”, thereby achieving a map.

Transform algorithm. The goal of this algorithm is to map SPL terms to MedDRA terms by transforming words/phrases in an SPL term in order to make it more compatible with MedDRA terminology. This algorithm performs several kinds of transforms: **synonym transform** replaces a word/phrase within the SPL term with a ‘MedDRA compatible’ term/phrase by comparing the parts of SPL term to a synonym transform table. The synonym transform table was manually created and contains 32 members. Examples of entries in this table include “injection site”→“local” (i.e. the SPL

term ‘injection site edema’ becomes the MedDRA term ‘local edema’), and “liver enzymes”→“liver function tests”. This process is also useful in coping with inconsistencies in MedDRA terminology that can interfere with successful mapping. In reviewing MedDRA terminology, we found that some words/phrases are preferred over others. For example, there are 51 MedDRA terms that contain the words ‘central nervous system’ but only 20 MedDRA terms that contain the equivalent word ‘CNS’ (an acronym for central nervous system). Replacing ‘CNS’ with ‘central nervous system’ whenever it occurs in an SPL term increases the likelihood SPLICER will successfully map a term to MedDRA. The **switch transform** indentifies SPL terms that begin with a word/phrase, which is then deleted and a form of the term applied to the end of the SPL term. For example, all SPL terms that contain the pattern ‘decrease in X’ get transformed to the more MedDRA-like pattern ‘X decreased’ (‘decrease in potassium’ becomes ‘potassium decreased’). The switch transform table was manually created and contains 27 rows. The **append transform** identifies SPL terms that end with a word/phrase, which is then deleted and a more MedDRA-like term appended to the SPL term. For example, all SPL terms containing the pattern ‘X prolongation’ get transformed to the more MedDRA-like pattern ‘X prolonged’. This append transform table was manually created and contains 32 rows. Once the above transforms are performed, the resulting SPL term is attempted to be mapped to MedDRA through an exact string match.

Tokenizer algorithm. This algorithm attempts to find a MedDRA term that contains the exact tokens as the SPL term regardless of token order. The algorithm uses the stopped and stemmed version of the MedDRA and the SPL term. It tokenizes the SPL term and when it finds a MedDRA term with the same number of tokens, it then checks

to see if each SPL term token is present anywhere within the MedDRA term. The first MedDRA term found that contains all of the tokens of the SPL term is declared a match. The algorithm is useful for matching terms we have not transformed (described earlier). A walk-through of this process is illustrative: The SPL term “edema of the application site” gets stopped and stemmed to “edema application site”. This is not an exact match to the corresponding MedDRA term “application site edema”. However the tokenizer algorithm would achieve a match since both terms contain the same number and kind of tokens, even though they are ordered differently. We found through previous NLP research that semantic meanings of phrases are rarely affected by token order.

Tokenizer Synonym Algorithm. This algorithm performs the same order-agnostic matching as the previous algorithm, but instead of trying to match to the MedDRA term table, it tries to match to the synonym table described previously.

MedDRA-in-SPL term Algorithm. The goal of this algorithm is to map SPL terms to MedDRA terms by finding all MedDRA terms *contained in* the SPL term. We found through initial review of SPLICER output that the ADR extraction process occasionally identifies phrases that are not ADRs themselves, but may contain ADRs. One example of this kind of phrase is the SPL term ‘chest pain along with rib pain’. This phrase does not map to any single MedDRA term. However, this algorithm can identify the MedDRA terms (underlined) contained in this phrase by using processing similar to the M1 algorithm described during ADR extraction.

Similarity algorithm. The goal of this algorithm is to map an SPL term to a MedDRA term when the SLP term is close to, but not an exact match to a MedDRA term. All mapping processes discussed heretofore have required the SPL term be an exact

match to a MedDRA term. This process uses a string comparator algorithm that compares two words/phrases and returns a score between 0 and 1. The closer the score is to 1, the more similar the words. Using this algorithm in previous NLP research, we found that a score of .92 or greater accurately identifies identical words that are either misspelled or spelling variants ("anaemia" vs. "anemia"). This algorithm is computationally intensive can occasionally result in false positives, which is why we placed it last in the series of matching algorithms.

If no matching MedDRA term is discovered after processing by the above algorithms SPLICER concludes the SPL term cannot be matched.

There is a post-mapping algorithm that is applied to matched MedDRA terms called '**Lab Directionality"**. A relatively common ADR found in SPLs are laboratory test abnormalities such as "decreased potassium" or "increased liver enzymes". Because SPLs are written in natural language, often the laboratory test and the effect the medication has on it are not in close proximity to each other as in: "Lasix can decrease some laboratory tests such as calcium and potassium". Because of this, we found that our ADR extraction algorithms often identify the test names (in this case, calcium and potassium), but miss the effect (decrease). We observed however that the effect is often found within the sentence where the laboratory test is mentioned. We manually created a table containing 2,392 common laboratory test names. This algorithm compares the ADRs found during MedDRA mapping to this laboratory name table. If a match occurs, the algorithm identifies the SPL sentence where the laboratory test was found and using NLP, determines the effect (or direction) of the medication on the test. Five conclusions are possible: increased, decreased, abnormal, indeterminate, and unstated. Abnormal is

concluded when the SPL sentence merely states that a lab abnormality can occur (SPLs sometimes state that a medication can cause a laboratory abnormality but provide no further details). Indeterminate means the algorithm found evidence of medication effect in the sentence but not conclusive enough to allow interpretation. Unstated means it did not find any evidence of effect in the sentence as can occur when potassium is used in reference to a medication rather than to a laboratory test. The NLP system uses regular expressions and algorithmic rules to identify mentions of lab direction in the sentence. If the lab directionality algorithm concludes increased, decreased, or abnormal the ADR is converted into a term consistent with MedDRA, i.e. ”potassium decreased” and the effect is linked to the ADR and included in the output.

After research examining the effect that medications have on laboratory tests, we identified 16 tests where medications essentially cause always the same effect. For example, serum creatinine is nearly universally increased as a result of an ADR from medication. Therefore, for these tests, we assign a default effect should NLP fail to conclusively identify an effect within the target sentence.

We programmed SPLICER to output its results to a tab delimited text file. Data elements included in its output are shown in Figure 6, and a partial view of SPLICER output for the medication Zoloft is shown in Table 2.

We performed an evaluation of SPLICER accuracy by taking a random sample of 10 OMOP drug labels and performing an independent manual review to identify ADRs in the adverse reactions section. To create the gold standard against which SPLICER output would be compared, the reviewer (who had no role in OMOP or in the development of

1. Drug ID number
2. Brand name of medication
3. Generic name of medication
4. Pregnancy category
5. Date of manufacture
6. SPL section where ADR was found
7. Mapped MedDRA term
8. Frequency of ADR
9. Original ADR extracted from SPL
10. SPL table data[1]

Figure 6 Data elements included in SPLICER output.
[1] Table data only when applicable and includes: N for study group, N for placebo group, number of study patients with ADR, number of placebo patients with ADR

SPLICER) was tasked to populate a database with all ADRs from each of the 10 SPLs. We then compared this output with SPLICER's output in order to calculate the sensitivity (recall), positive predictive value (precision) and F-measure of the SPLICER software.

The review process consisted of a third party reviewer (a specialist in the MedDRA lexicon) manually extracting ADRs from the 10 sample SPLs then comparing these results to SPLICER's output for these labels. In annotating the results, the reviewer marked each adverse reaction in one of 5 ways: 1) found by *both* SPLICER and by manual label review (B); 2) found by *label* review only (L); 3) found by *SPLICER* only (S); 4) *partial* match with SPLICER having *missed* content (PM); 5) *partial* match with SPLICER having *added* content (PA), Defining manual label review as the gold standard, we made the following interpretations for each of the above categories. Category B reflected a correctly extracted adverse event and thus a true positive. Category L reflected an ADR missed by SPLICER and thus a false negative. Category S

Drug ID	Brand Name	Pregnancy Category	SPL section	Mapped MedDRA term	Frequency	SPL term
FE9E8B7	Zoloft	C	Adv. rxn	fatigue	2	tiredness
FE9E8B7	Zoloft	C	Adv. rxn	nausea	5	nausea
FE9E8B7	Zoloft	C	Adv. rxn-post. market	headache	infreq	headaches
FE9E8B7	Zoloft	C	warnings	Potassium increased		Increased potassium

Table 2 Partial view of anticipated structure and content of SPLICER output

Section	Mean number of ADRs
Adverse Reaction	83.7
Black Box	.5
Warnings/Precautions	17.5
Post-Marketing	5.29

Table 3 Mean number of ADRs found in OMOP labels per section.

ADRs were terms extracted by SPLICER that were not actually adverse events, and thus false positives. Finally, the two partial categories PM and PA were counted as false negatives and false positives respectively.

III. Results

SPLICER processed a total of 80 OMOP-targeted drugs which represented 433 drug labels. SPLICER completely processing in just over 5 hours on a dual-core Dell Computer running the Windows Vista Operating System. SPLICER found a total of 40,433 unique ADRs in the 433 drug labels. The largest number of ADRs found in one label was 326 (clomipramine hydrochloride), while the smallest number was 2 (erythromycin eye drops). The total mean number of ADRs per entire label was 93, and the mean number of ADRs found in the different sections of the OMOP SPLs is shown in Table 3.

ADR	Number of Unique Labels
Nausea	368
Vomiting	334
Dizziness	302
Headache	301
Diarrhea	281
Hypotension	267
Thrombocytopenia	251
Death	240
Insomnia	240
Hypersensitivity	235
Fatigue	231
Constipation	227
Agranulocytosis	218
Confusion	218
Ataxia	200
Fever	199
Pruritus	199
Pregnancy	198
Alopecia	198

Table 4 Top 20 most common ADRs per OMOP drug label.

The top 20 most frequent ADRs found in the OMOP SPLs is shown in Table 4. By far the most common ADR found was nausea- found in nearly 85% of all OMOP drug labels.

Our evaluation of the software's performance in processing a random sample of 10 OMOP SPLs revealed that SPLICER correctly identified 1041 out

of 1081 labeled adverse events. It incorrectly extracted 40 terms which were not true ADRs, resulting in an overall recall of 96.3%, a precision of 97.2% and an F-measure of 0.97.

IV. Discussion

We were successful in developing an NLP system to accurately process OMOP-specific drug labels in order to identify, extract and standardize the ADR information contained in them. The database created by SPLICER is being used within OMOP to evaluate methods designed to identify non-specified drug-condition associations. Our evaluation demonstrated that SPLICER was generally accurate in identifying ADRs in SPLs, and in mapping them to their appropriate MedDRA terms.

The database created by SPLICER has many potential uses outside of the OMOP project. Standardized ADR data, obtained directly form the drug labels themselves and gathered and collated in a searchable database format, could be valuable in wide variety of applications. For example, a clinical web service could be created which would take as input a list of a patient's medications and would return a concise list of the potential ADRs - ranked by either seriousness or likelihood to occur. Similarly, such a service could also clearly indicate which of a patient's medications are known to cause an existing symptom or condition. Another example of the use of such a database is in providing an 'alternative view' of the ADR information contained in drug labels. Historically, drug labels were read by physicians in paper format, whether as package inserts or as part of the compendia such as the Physician's Desk Reference®. The recent emergence of electronic labels has allowed for review of information on a computer display and performance of basic search functions (e.g., find "thrombocytopenia"). But

studies have shown that adverse event warnings can significantly influence prescribing behavior and reduce the incidence of ADRs[21-23]. However, the effectiveness of labels in communicating drug safety information may be diminished by issues of information overload[24-28]. Research has suggested that the density of ADR data presented may result in difficulty distinguishing important adverse events from those of lesser concern[24, 25]. By using and combining the SPL with structured and standardized ADR data produced by SPLICER enables support of far more sophisticated mechanisms of review. For example, a physician could choose to see only adverse events of a certain type (e.g., cardiovascular), of a certain age group (e.g., pediatric), or associated with a particular co-morbidity (e.g., diabetes). With data inserted in a standardized form, a physician could compare the differences in adverse events among multiple medications in the same therapeutic class (e.g., weight gain in SSRI's). Furthermore, a richly encoded SPL could in fact interact with an electronic medical record system to show which of a patient's symptoms or laboratory abnormalities may be associated with their medications. In short, we believe that the solution to information overload in drug labeling is not reducing the volume of data, but leveraging technology to deliver dynamic, targeted filtering of information based on physician and patient context. Programs such as SPLICER and the data it creates can help to achieve this goal.

Despite the satisfactory performance of SPLICER processing the SPLs of OMOP-targeted drugs, based on a preliminary analysis of it's processing of SPLs of non-OMOP drugs, we see several areas where its accuracy can be improved. One is in the processing of SPL tables. Tables are especially valuable sources of ADR data in SPLs because they often contain *specific numeric frequency* information with which an ADR is encountered,

rather than less specific frequency mentions such as ‘rare’ or ‘infrequent’. Currently, SPLICER is programmed to capture the ADR and the associated numeric frequency information from tables when it correctly recognizes a table format, but only the ADR when it does not. Because of the widely varying table formats present in SPLs, we estimate that SPLICER cannot recognize a table format approximately 30% of the time. We continue to work to enhance and improve SPLICER’s table formatting recognition algorithms so that SPLICER will recognize greater numbers of table formats thereby increasing the capture of detailed ADR frequency data. The NLP technology that SPLICER uses to identify ADRs could also be improved. For example, SPLICER uses algorithms to determine when an ADR is used in a negative context and hence not a true ADR for that drug, such as in the phrase “Zoloft has **not** been shown to cause **agranulocytosis**”. However, SPLICER can miss complex phrases indicating negation of an ADR, thereby causing false positive errors. Likewise, occasionally ADRs are mentioned in SPLs in a non-positive context which can cause SPLICER to make additional false positive errors. For example, in the phrase “in patients with migraine headaches, Zoloft can sometimes cause nausea” SPLICER incorrectly identifies ‘migraine headaches’ as an ADR, even though it is not an ADR of Zoloft when mentioned in this context. We are currently developing algorithms that will improve the identification of the *context* within which an ADR is mentioned in an SPL.

Our study has a number of limitations. First, due to time and funding constraints, we performed our evaluation of SPLICER using only a small sample of SPLs and using only one reviewer. In the future, we plan on performing a more extensive and thorough evaluation using a larger number of SPLs and more reviewers. Second, SPLICER is

programmed to map ADR data from SPLs to MedDRA terms only. A computerized system that requires ADR SPL data to be mapped to some other standardized vocabulary (SNOMED-CT for example) would not be able to use SPLICER output directly. And although mapping tables do exist between the MedDRA terms that SPLICER outputs and other standardized vocabularies, the potential for imprecise and/or incomplete mappings can resultant data likely will occur.

In the future, we plan to continue to enhance and improve SPLICER in order to achieve highly accurate processing and data mining of the entire set of SPLs (over 7,000) on the DailyMed website.

REFERENCES

1. OMOP Web site. Available at:http://omop.fnih.org/node/22 Accessed August 13, 2010.
2. FDA Web site. Available at: http://www.fda.gov/ForIndustry/DataStandards/StructuredProductLabeling/default.htm Accessed September 1, 2009.
3. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001:17-21.
4. Friedman C, Shagina L, Lussier Y, Hripcsak G. Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc. 2004 Sep-Oct;11(5):392-402.
5. Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. J Am Med Inform Assoc. 2008 Jan-Feb;15(1):87-98.
6. Elkin PL, Froehling D, Wahner-Roedler D, et al. NLP-based identification of pneumonia cases from free-text radiological reports. AMIA Annu Symp Proc. 2008:172-6.
7. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 2001 Oct;34(5):301-10.
8. Melton GB, Hripcsak G. Automated detection of adverse events using natural language processing of discharge summaries. J Am Med Inform Assoc. 2005 Jul-Aug;12(4):448-57.
9. Honigman B, Lee J, Rothschild J, et al. Using computerized data to identify adverse drug events in outpatients. J Am Med Inform Assoc. 2001 May-Jun;8(3):254-66.

10. Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. J Am Med Inform Assoc. 2009 May-Jun;16(3):328-37.
11. Regenstrief Institute Web site. Available at: http://www.regenstrief.org/ Accessed October 4, 2008.
12. Friedlin J, McDonald CJ. A natural language processing system to extract and code concepts relating to congestive heart failure from chest radiology reports. AMIA Annu Symp Proc. 2006:269-73.
13. Friedlin J, McDonald CJ. Using a natural language processing system to extract and code family history data from admission reports. AMIA Annu Symp Proc. 2006:925.
14. Friedlin J, Grannis S, Overhage JM. Using natural language processing to improve accuracy of automated notifiable disease reporting. AMIA Annu Symp Proc. 2008:207-11.
15. Turchin A, Kolatkar NS, Grant RW, Makhni EC, Pendergrass ML, Einbinder JS. Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes. J Am Med Inform Assoc. 2006 Nov-Dec;13(6):691-5.
16. MedDRA Web site. Available at:http://www.meddramsso.com/ Accessed March 13, 2010.
17. Health Level 7. HL7 Web site. Available at: http://www.hl7.org. Accessed Oct 1, 2009. .
18. UMLS Web site. Available at: http://www.nlm.nih.gov/research/umls/ Accessed August 1, 2010.
19. MESH Web site. Available at:http://www.nlm.nih.gov/mesh/ Accessed March 13, 2010.
20. SNOMED Web site. Available at: http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html Accessed August 13, 2010.
21. Starner CI, Schafer JA, Heaton AH, Gleason PP. Rosiglitazone and pioglitazone utilization from January 2007 through May 2008 associated with five risk-warning events. J Manag Care Pharm. 2008 Jul-Aug;14(6):523-31.
22. Olfson M, Marcus SC, Druss BG. Effects of Food and Drug Administration warnings on antidepressant use in a national sample. Arch Gen Psychiatry. 2008 Jan;65(1):94-101.
23. Jacoby JL, Fulton J, Cesta M, Heller M. After the black box warning: dramatic changes in ED use of droperidol. Am J Emerg Med. 2005 Mar;23(2):196.
24. Watson KT, Barash PG. The new Food and Drug Administration drug package insert: implications for patient safety and clinical care. Anesth Analg. 2009 Jan;108(1):211-8.
25. Avorn J, Shrank W. Highlights and a hidden hazard--the FDA's new labeling regulations. N Engl J Med. 2006 Jun 8;354(23):2409-11.
26. Hollister LE. Drug product information: sensible disclosure of clinically important facts. Drugs. 1974;7(6):414-8.
27. Herxheimer A, Lionel ND. Minimum information needed by prescribers. Br Med J. 1978 Oct 21;2(6145):1129-32.

28. Schommer JC, Doucette WR, Worley MM. Processing prescription drug information under different conditions of presentation. Patient Educ Couns. 2001 Apr;43(1):49-59.