Construction and Annotation of a UMLS/SNOMED-based Drug Ontology for Observational Pharmacovigilance

Presented at IDAMAP (Intelligent Data Analysis for bioMedicine and Pharmacology), Washington, DC, 2008

Gary H. Merrill, Patrick B. Ryan, Jeffery L. Painter
GlaxoSmithKline, Research Triangle Park, North Carolina

Abstract

The primary goal of the SafetyWorks project has been the development of an integrated set of methodologies enabling the use of large observational data sources in monitoring and assessing drug safety concerns. To support its analytical and exploratory capabilities, SafetyWorks makes use of two large hierarchically structured ontologies – one for medical conditions, and one for drugs. In this paper we focus on the drug ontology employed in SafetyWorks and on its construction and annotation based on the SNOMED CT and RxNorm subsets of the Unified Medical Language System Metathesaurus. The result is a case study illustrating the value of SNOMED and its integration with UMLS and RxNorm in a critical and innovative drug safety application. We expose sufficient details of our methods to enable others to make use of these methods and to encourage the expanded use of both SNOMED and the UMLS in data exploration and analysis applications, particularly in the area of improving approaches to drug safety. [1]

1 Introduction

FDA "Guidance for Industry Good Pharmacovigilance Practices and Pharmacoepidemiologic Assessment" [FDA, 2005] describes pharmacovigilance as "all scientific and data gathering activities relating to the detection, assessment, and understanding of adverse events." While a drug is in development, one of the primary sources of safety information is clinical trials, but most trials suffer from insufficient sample size and lack of external validity to reliably estimate the risk of any potential safety concerns for the target population. Once a medicine has been approved, spontaneous adverse event reporting becomes an increasingly important tool for safety evaluation. Case review remains a key component of the ongoing surveillance of medicines, and the application of disproportionality analysis tools on spontaneous adverse event databases has greatly enhanced the signal detection process. Unfortunately, these spontaneous reporting systems have several limitations that make causal assessments difficult ([Almenoff *et al.*, 2005; Hauben *et al.*, 2005]): voluntary reporting suffers from chronic underreporting and maturation bias, and the unknown nature of underlying populations make true reporting rates difficult to obtain and use for comparisons. Several recent safety issues have received significant public attention ([Furberg *et al.*, 2006]), resulting in heightened awareness of the challenges of the current safety review process and increased demand for improved methods for understanding the effects of medicines and ensuring patient safety.

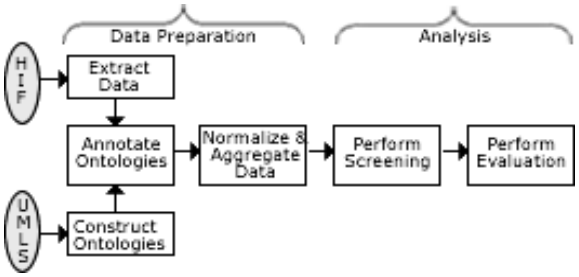

Figure 1: The SafetyWorks Process

SafetyWorks is an integrated system for leveraging observational data in support of the identification and evaluation of potential safety concerns of medicines. This system encompasses a data processing procedure that transforms disparate data sources into a common framework that enables normalized analyses across sources and the integration of automated methods for observational screening and observational evaluation. Figure 1 illustrates how raw data is extracted from the GlaxoSmithKline Healthcare Information Factory (a repository of large databases), normalized and aggregated with the help of annotated medical condition and drug ontologies constructed from the data

[1] All references to the Unified Medical Language System, the UMLS Metathesaurus, RxNorm, and the UMLS Lexical Tools are accessible through [NLM, 2008]. The SafetyWorks project began in the spring of 2005 and most of the ontology work was developed on the basis of the 2005-2006 releases of the UMLS and its documentation. However, we have continually updated our ontology as new releases have appeared.

An extended argument for the use of multiple observational databases in pharmacoepidemiology and how the methods described here may play a central role in this can be found in [Ryan, 2008]. Some additional details and related work may be found in [Painter *et al.*, 2006], [Ryan *et al.*, 2008], [Ryan and Powell, 2008], [Merrill *et al.*, 2008], and [Painter, 2008].

and UMLS, and then used in observational screening and observational evaluation to assess drug safety. Combined, this "observational pharmacovigilance" approach provides a systematic solution to supplement – rather than replace – current practices, enabling more proactive monitoring and better informed decision making.

The data processing procedure involves extracting key elements from each data source into a common relational model. While the representations (codes or strings) of individual elements (medical conditions and their treatments) in each source may be different, both contain common concepts of persons with drug utilization and condition incidence. We construct *drug eras* to represent periods of time where the data suggest a person may be persistently taking a medicine based on prescriptions written, prescriptions filled, or medication history provided to the physician. Similarly, we construct *condition eras* to represent common episodes of care for the same medical condition, aggregating related diagnostic codes that occur within a persistence window. Analytical methods are then applied to these eras to discover drug/condition associations and to evaluate the strength of such associations. Biomedical ontologies play an instrumental role in facilitating the normalization and aggregation of similar drug and condition concepts, and their application is the focus of this paper.

Observational screening applies an unmatched cohort design to facilitate comparisons of incidence rates of all outcomes across two populations of interest, both pre- and post-exposure. It provides an exploratory context (including information concerning patient demographics, comorbidities and concomitant medications) that can be used to understand and compare drugs, their uses, and their effects. Observational screening analyses should be considered exploratory and hypothesis-generating in nature, and should facilitate the identification and prioritization of drug-condition pairs that warrant further evaluation.

Observational evaluation (or "risk estimation") is a targeted analysis aimed at providing a robust estimate of the strength of a drug/condition association within the population of interest by systematically assessing the temporal association between a specific drug and a specific condition within the observational data sources. It models the specific exposure-outcome relationship using multivariate Poisson regression within a propensity score matched cohort design, adjusting for important covariates related to both exposure and outcome. Observational evaluation can be one mechanism to assess the hypotheses generated within observational screening by constructing cohorts that are comparable (adjusted for confounding) and representative of the population of interest.

Two classes of observational data that hold promise in this domain are administrative claims databases and electronic health records. Each type of data has its own advantages and limitations, and specific data sources may have unique features that need to be well understood and carefully considered when conducting observational analyses and interpreting results. SafetyWorks currently makes use of one instance of each type of observational data.

The administrative claims database contains health information for over 74 million persons with an average 24 months of coverage. Drug utilization is extracted from over 833 million pharmacy claims of prescriptions filled. Conditions are captured from diagnosis codes on inpatient and outpatient medical billing claims; 5.6 billion distinct diagnoses were aggregated into 1 billion condition eras. Insurance claims data has the advantage of very large sample size, and generally comprehensive summaries of health-related activities during enrollment. However, claims are also susceptible to misclassification bias, and may not adequately capture symptoms or other important aspects of the patients' medical histories. The database represents an employed, privately insured population which may not be generalizable to other populations of interest.

The electronic health record (EHR) database provides health information for 5.8 million patients. Drug utilization is extracted from prescriptions written by the provider and medication history lists to create 58 million drug eras, averaging 101 days of exposure. Condition eras are constructed from a problem list of diagnoses, symptoms, and other components of medical history, resulting in 32 million condition eras.

One key opportunity in observational pharmacovigilance lies in enabling the systematic use of disparate observational databases for a more comprehensive review of the utilization and effects of medicines in populations. By establishing a common conceptual framework to structure observational data and to normalize references to drugs and conditions, analyses can be conducted consistently across sources, thereby enabling direct comparison of otherwise disparate results. Formal biomedical ontologies provide us with the mechanism for achieving this goal.

2 Methods

In choosing a drug ontology for SafetyWorks, we were guided by several criteria. The ontology must provide a correct and uniform classification of drugs and drug categories. It must be comprehensive relative to the relevant data – which is to say that it must exhibit a sufficiently high granularity of categories to which drug references in our data (and anticipated future data) could be annotated. It must contain categories for branded drugs as well as generics. It must exhibit a hierarchical structure in terms of individual drugs, their generic forms, and various levels of drug classes; and this hierarchical structure must adequately represent the relations of drug products to multiple ingredients that they contain.

Beyond these purely formal or structural constraints, we also felt it necessary to impose constraints of usability since the ontology would be employed in a graphic and interactive manner by drug safety scientists. Accordingly, it is necessary that the ontology exhibit categories and a structure of some familiarity to such users, and it must be easily navigable and searchable by them. Finally, prior research (see, for example, [Painter *et al.*, 2006]) had convinced us that the UMLS comprised a powerful resource in the areas of drug discovery, coding scheme translation, and broader areas of biomedical informatics, and we were committed to taking advantage of the richness of the relations provided in the Metathesaurus across such domains as medical conditions, diagnoses, symptoms, and drugs. As a consequence, we sought an ontology that was represented among the UMLS Metathesaurus sources.

We therefore settled on SNOMED CT as the basis of the SafetyWorks drug ontology which is constructed and annotated in a sequence of steps:

- The *Drug or medicament* sub-hierarchy of SNOMED CT is extracted from the UMLS Metathesaurus.
- RxNorm is used to extend this hierarchy by grafting leaf nodes to it for branded drugs.
- The extended ontology is annotated with drug references from the observational data sources.
- The annotated ontology is simplified by applying several transformations to its hierarchical structure.
- The resulting ontology is then emitted as a set of files suitable for importation into a relational database for use by the SafetyWorks methodologies.

The first of these steps is accomplished straightforwardly through use of the MRCONSO.RRF and MRHIER.RRF files of a Metathesaurus subset containing the SNOMED CT source. The hierarchy is represented as a set of "nodes" identified by their UMLS Atom Unique Identifiers (AUIs), or extensions of these, and is extracted simply as the *isa* hierarchy with *Drug or medicament* (AUI A6938913) as its root.

Adding Branded Drugs

As part of our data extraction process, we created unique drug product reference identifiers as product-name/strength strings (such as "Zantac 150 Mg"), and associated with these may be additional information (varying with the data source being used) in the form of codes from a variety of coding schemes. Ideally, each drug product reference would be such a string consisting only of a drug name and a strength. However, actual drug product references in the reference file and in the data sometimes contain additional information as well ("tablet", "syringe", etc.); and this makes identifying the drug and correctly annotating it to the ontology more challenging.

Occasionally it is important to distinguish between the occurrence of a drug product reference in the drug reference file (where each drug product has only a single reference) and occurrences of a drug product reference in the observational data itself (where there may be millions of references to a particular drug or drug product). In the latter case we will then refer to *instances* (in the data) of the drug reference or drug product reference.

Unfortunately, the otherwise quite satisfactory *Drug and medicament* hierarchy extracted from the UMLS lacks categories for branded drugs such as "Wellbutrin", "Zyban", etc. While for the most part the interest of drug safety scientists is focused on generic forms, we felt it necessary to achieve the granularity of branded drugs for the sake of completeness and because there are circumstances in which drug/condition associations may occur with one specific drug product and not with another. Our immediate challenge was to extend SNOMED CT with branded drug categories, and RxNorm provided us with a mechanism to meet this challenge.

The goal, then, is to take each branded drug in RxNorm (term type TTY = BN) and find the set of generic categories in the *Drug or medicament* hierarchy of SNOMED CT that represent the ingredients of that drug. The branded drug (represented by its RxNorm AUI) is then grafted to each such category as a child node in the hierarchy. In turn, this requires first finding the CUI (Concept Unique Identifier)representing the drug's "concept" and then finding the set of AUIs (Atom Unique Identifiers) *in our hierarchy* that "realize" that concept. This goal is facilitated by the RxNorm relations *tradename_of*, *ingredient_of* (and *has_ingredient*), *consists_of*, and *form_of*. These relations allow us to construct a mapping from CUIs for brand names in RxNorm to the desired sets of AUIs in our SNOMED sub-hierarchy. In fact, we restrict this mapping to only those BNs in RxNorm that have the semantic type of *Organic Chemical* since experience has shown us that this is the class of entities that most closely approximates what are intuitively regarded as the "normal" set of branded drug products.

In virtually all cases it is possible to map directly from a drug's brand name through the *tradename_of* relation to its ingredient(s). An example of this is the branded drug name (BN) "Wellbutrin" (C0085934) which in RxNorm is a tradename of the ingredient (IN) Bupropion whose CUI is C0085208) and this in turn is realized in the SNOMED CT *Drug or medicament* hierarchy as AUI A2879308. Thus we can attach the a category for the branded drug Wellbutrin as a child of the *Bupropion* category in our extended hierarchy.

There is some question as to whether, and to what degree, the coverage and accuracy of our annotation could be enhanced by the use of other information the data might contain – such as associations with codes from various coding schemes. This is still something of an open question, but at one stage of the project substantial effort was put into making use of NDC codes in the data and their occurrences within several sources (NCI, NDFRT, NDDF, and VANDF) in the Metathesaurus. After a careful and thoughtful implementation, it was determined that the use of this approach yielded not a single enhancement to our lexically-based heuristic approach, and so it was removed from the annotation component.

396 of the brand names in RxNorm did not map to ingredients (i.e., these were BNs that had no corresponding INs) and consequently were not added as categories to our hierarchy. A single case (Meclomen) failed to map by means of the *tradename_of* relation. However, in our experimental approach to mapping arbitrary drug names into UMLS sources, we had developed a set of sophisticated algorithms involving relations among semantic clinical drugs (SCD), semantic clinical drug components (SCDC), and semantic branded drug components (SBDC); and the Meclomen case fell to these.

The *Drug or medicament* hierarchy extracted directly from the Metathesaurus contained 6,800 categories, and adding categories for branded drugs raised this count to 15,159.

Annotating Drug Data References to the Ontology

The goal of annotation is to associate each drug reference in our data with one or more categories in the drug ontology. Our fundamental approach to annotating the drug ontology with such drug references is then to match the string rep-

resentation of the product-names to category names in the drug ontology, and we employ a number of algorithms and heuristics in this pursuit.

Matching of this sort requires a careful approach to string normalization, and initially our approach was to depend on the UMLS normalization of strings found in the MRXNS_ENG.RRF file of our SNOMED CT and RxNorm subsets, and couple this with the use of the UMLS Lexical Tools *norm* utility. However, for a variety of reasons we cannot detail here, we ultimately abandoned this approach in favor of developing our own string normalizer which is tuned more specifically to the needs of a clinical drug vocabulary.

The fundamental concept supporting our annotation of the drug ontology with drug references from our data is that of *Product Instance*. A Product Instance represents a single "drug product", distinguished by the product name, and also associates with this an "expanded" version of that name, a set of generic ingredients (if known), and potentially other information as well (such as codes from a variety of coding schemes, if these are known and might be useful). We annotate the ontology with drug references from each data source in turn, and the first step in annotation is to construct a Product Instance Table comprising Product Instances for each distinct drug product referenced in the given data source. Once the Product Instance Table is created, we then consider each Product Instance in turn and attempt to annotate it to the ontology.

The high level heuristic we follow in attempting to annotate a Product Instance to the ontology is a sequence of steps, each of which is tried if the previous ones have failed to produce a successful annotation. Table 1 illustrates what proportion of drug reference matches are captured by each method in the case of annotating our data to our SNOMED-based and enhanced drug ontology.

	Drug Coverage	
Annotation by matching normalized form of	**Claims**	**EHR**
The exact product name	51.60%	45.40%
An expanded form of product name	3.16%	1.40%
The generic name(s)	44.89%	52.13%
A variant of the product name	0.35%	1.07%
A variant of the expanded product name	0.00%	0.00%

Table 1: Drug Reference Coverage

The relatively high percentage of generic matches as compared to product name matches reflects our current conservative strategy of preferring a generic match over a match to a product name that has been modified in ways that might render a resulting match inaccurate. This is a challenging problem in the case of various forms of products (such as "extended release", "flu", "nighttime", "cold/cough" vs. "cold/allergy" variants, etc.) where the variant may contain significantly different additional ingredients than the base product; and tuning our matching heuristics is an ongoing research project.

To consider some examples:

- A reference to "Wellbutrin" succeeds as a direct match.
- "Aber-Tuss HC" fails to match (it does not occur in RxNorm). Its generic in the EHR data source appears as "PHENLYEPH-CHLORPHEN-HYDROCOD" which also fails to match. "CHLORPHEN" does match a SNOMED category, but neither of "PHENLYEPH" or "HYDROCOD" do. However, expansions of these do match, and so drug product references involving "Aber-Tuss HC" in the EHR data are annotated to each of the categories *Phenlyephrine*, *Chlorphen*, and *Hydrocodone*.
- "Dextrose in water" fails to match, but its variant "Dextrose" succeeds and so "Dextrose in water" is annotated to *Dextrose*.
- The product "Haleys M-O" fails to match, as does its generic "Mag hydroxide in mineral oil" and the expansion "Magnesium hydroxide in mineral oil". But the simplified variant of the expanded generic, "Magnesium hydroxide", succeeds and so "Haleys M-O" is annotated to the category *Magnesium hydroxide*.

Any drug reference that fails to be annotated to a drug category is annotated to an *unclassified substances* category added to the ontology for this purpose, and the consequence of this is that no purported drug reference in the data is ever lost to analysis.

2.1 Simplifying the Ontology

Initial attempts to use the annotated ontology as described in the previous section showed us that it exhibited some unfortunate features relative to our criteria and our plans for using it to support the SafetyWorks analytical methodologies.

Our last stage in the construction of the drug ontology is then to perform a series of refinements in which we

- Prune unnecessary "forms" of drugs from the hierarchy.
- Ensure that no drug reference annotates both a node and an ancestor of that node.
- Create "generic product" nodes to ensure that only hierarchy nodes at the lowest level are annotated.

Note that no hierarchy pruning or restructuring should take place until annotation is complete in order to maximize the degree and accuracy of the annotations.

Pruning forms

If we look at the ontology immediately after annotating it from the data sources, we will see a number of categories that serve no useful purpose and are something of a hazard to efficient navigation. These categories represent "forms" of a drug and are of no interest in the drug safety context within which we are working. The most common example of such categories are salts of substances such as *Fluvastatin sodium* which in RxNorm is the *tradename_of* the branded drug *Lescol*. But *Fluvastatin sodium* is a direct child of *Fluvastatin* in the extended hierarchy.

The RxNorm documentation describes the *form_of* relation as holding "between a base ingredient and a precise ingredient", and in our analytical and exploratory context

such "precise" ingredients are unnecessary. In addition, to facilitate application of some of the SafetyWorks analytical methodologies and to provide more meaningful results to our users, we adopt the principle of annotating only branded drug categories or generic drug categories. Leaving forms (such as salts) in the hierarchy yields a structure that essentially contains (non-uniformly, since only at some places) two levels of generic drug categories. As a consequence, we choose to eliminate these unnecessary intermediate categories and prune the ontology of them by making use of the RxNorm *form_of* relation. As this pruning takes place, it is necessary to move any annotations attached to these categories to the higher level (base ingredient) category that remains.

Eliminating ancestral annotations

It is possible that a data reference has been annotated to multiple ontology categories, and we have seen examples in previous sections where this makes perfect sense (as in the case where a drug has multiple components). However, it is also possible that a data reference has been annotated to a category and also to an *ancestor* of that category. How does this happen?

The answer to this question lies in what are often slight incommensurabilities between the concept structure of UMLS and the content or structure of the specific ontology we are annotating (in this case a subset of SNOMED CT). Recall that we accomplish annotation of the ontology by first finding the UMLS concept (CUI) that represents a data reference and then "projecting" that concept into SNOMED CT to find the UMLS AUIs that realize the concept in that source.

It is true that a given AUI will be associated with only one CUI, but a given CUI may be associated with (realized by) multiple AUIs *in the same source*. This is simply a feature of how CUIs, AUIs, and their relationships have been implemented in the Metathesaurus. After all, the UMLS concept structure is simply yet another ontology (though it is intended to be a very general one). So it should not be at all surprising that mappings of Metathesaurus concepts into Metathesaurus sources will not, in some cases, be structure-preserving. A simple example of this is found in the case of the concept C0028040 which is realized in SNOMED CT by both A2877800 (*Nicotine*) and A3581984 (*Nicotine Agent*), where the latter is in fact a child (in the SNOMED CT hierarchy) of the former. And as a consequence of this, our data reference of "Nicotrol NS 1-Wk 10MG/ML" ends up being annotated to both of these categories.

For our purposes of data analysis (and also from the perspective of a user attempting to navigate the ontology) such redundant higher level annotations are confusing and can be computationally problematic. We therefore modify the annotated ontology to ensure that if a drug reference has been annotated to a node and also to one or more ancestors of that node, then the annotation is detached from the ancestor nodes.

Restricting annotation to the lowest level

Up to this point, we have allowed annotations to attach to both branded drug categories and to generic drug categories. Thus, for example, "Zantac 15MG/ML" annotates the *Zantac* category while "Ranitidine 15MG/ML" annotates the *Ranitidine* category. But this means that annotations are being made to two distinct levels of the hierarchy: branded drugs and their generics.

Again, this may complicate certain computations and it can be confusing to users navigating the hierarchy or searching for annotations. For these reasons we decided to annotate drug data references to only the lowest-level categories of the hierarchy. In order to do this coherently and uniformly we introduce the concept of an "NOC" (not otherwise characterized) category. An NOC category is grafted to the hierarchy at the same level as branded drug categories (i.e., as a child of a generic), and is used to hold annotations which would otherwise annotate the parent generic node. In the context of our Zantac example, then, we introduce the *Ranitidine NOC* category, make it a child of *Ranitidine*, and move any annotations from the *Ranitidine* category to the new lowest-level *Ranitidine NOC* category (which is a sibling of *Zantac*). Another way of thinking of Ranitidine NOC is as "unbranded Ranitidine product". More generally, an NOC category is expected to be annotated with unbranded *products* (or otherwise unknown/unrecognized branded products) of its parent generic category. Thus generic categories are never directly annotated, and annotations apply only to the leaves (lowest levels) of the hierarchy graph.

At this point our drug ontology (now with 16,100 categories) is a modified extended version of the SNOMED CT *Drug or medicament* hierarchy and is complete for our purposes. As shown in Figure 1, we then make use of the annotated ontology to "normalize" all drug references in each source into ontology categories, and on the basis of these normalized drug references we then create the aggregated drug eras described in the *Overview* section. That (together with a similar process involving our medical conditions ontology and the creation of condition eras) completes the SafetyWorks *Data Preparation* process, and the normalized and aggregated data is then used to perform observational screening and observational analysis for drugs and medical conditions of interest within the drug safety monitoring domain.

3 Discussion

Our claims data drug reference set contains 73,553 distinct drug product (product-name/strength) references, from which we can identify 15,848 distinct drug references (distinct product names, independent of strength). The EHR drug reference set contains 38,723 distinct drug product references and 22,851 distinct drug references. After annotating the ontology from both sources, we achieve a drug reference coverage (drug names from our drug reference set annotated to generic or brand name categories) of 90.16% for the claims drug references and 69.16% for the EHR drug references. Two questions that immediately arise are: "What explains the discrepancy in the coverage?" and "What explains the failure rate?".

As we have hinted in earlier sections, the data we are dealing with (in the form of drug product names) is not completely "clean". In fact, many of the purported "drug references" in the data are not references to drugs at all but to medical devices (syringes, braces, ice bags, lancets,

etc.), eyecare products, bathroom products, chemicals and minerals, and herbal remedies. The most humorous example of such items occurring in the drug column of our data is "Contour fitted sheets", but others abound. This is the nature of observational data.

Given this, the coverage of over 90% for the claims drug names is quite impressive, and the much lower rate for the EHR drug names is explained by a much higher density of non-drug items that are referenced in that data. In addition, references in the EHR data to generics is of much lower quality than in the claims data. Strings in that data that are supposed to represent drugs sometimes contain a non-drug substring such as the strength, formulation, or delivery system – which makes parsing out the drug name more difficult. Another difference in the quality of data between the two sources is that the claims data represents generic components individually in separate columns while the EHR data represents generics in a single column with complex generic combinations such as "PSEUDOEPH-CHLORPHEN-HYDROCOD" which must be tokenized correctly and then matched through the application of more complicated heuristics.

The coverage of actual drug reference *instances* in the data is even better. While we have successfully classified only about 70% of the EHR drug data references, this represents a successful annotation of 95.6% of the 57.7 million *instances* of drug utilization observed in the EHR data. And of the 494 million *instances* of drug utilization observed in the claims data, 98.2% are successfully annotated.

Although we have not done a formal expert-based analysis of the accuracy of our annotations, any misclassifications appear to be extremely infrequent and for the most part these occur when a fairly specific (e.g., brand) reference is annotated to its generic. An example of this is where "Zyban" is annotated to *Bupropion NOC*, and the reason in this case is that while "Zyban SR" occurs in RxNorm, "Zyban" itself does not. This raises, once again, difficulties in correctly classifying different forms of drug products, and we plan to more fully address these issues in future work. In general, the approach we take to drug name matching can be expected to be highly accurate since the heuristics it employs make use of partial matching of normalized strings but do not make use of any form of probabilistic matching (which we have found to substantially increase the chance of misclassification in both drug and medical condition ontologies). However, our work in the future will focus on improving these methods and assessing their validity.

The drug ontology described here forms an integral part of the SafetyWorks methodologies and has been used in test cases of those methodologies in the assessment of known drug/condition associations. We continue to improve and test these methodologies as SafetyWorks moves from a prototype application to a production-level application that can be used with confidence by drug safety scientists and epidemiologists. And we hope to have exposed enough details of our approach to make it usable by others.

4 Acknowledgements

The authors would like to thank their colleagues Robertino Mera, Kathleen Beach, Hoa Le, and Gregory Powell for contributing medical and drug safety knowledge in evaluating the SafetyWorks drug ontology and its application. Thanks are due Henry Krzywy for his work in understanding the data sources and in developing the technology for extracting the raw data for use in SafetyWorks.

References

[Almenoff *et al.*, 2005] June Almenoff, Joseph Tonning, A. Lawrence, Ana Szarfman, et al. Perspectives on the use of data mining in pharmacovigilance. *Drug Safety*, 28(11):981–1007, 2005.

[FDA, 2005] FDA. *Guidance for industry: good pharmacovigilance practices and pharmacoepidemiologic assessment*. U.S. Food and Drug Administration, March 2005. Available at `http://www.fda.gov/cder/guidance/6359OCC.htm`.

[Furberg *et al.*, 2006] D. Furberg, A. Levin, P. Gross, R. Shapiro R, and B. Strom. The fda and drug safety: A proposal for sweeping changes. *Arch Intern Med.*, 166:1938–1942, 2006.

[Hauben *et al.*, 2005] M. Hauben, D. Madigan, C. Gerrits, et al. The role of data mining in pharmacovigilance. *Expert Opin Drug Saf*, 4(5):929–948, 2005.

[Merrill *et al.*, 2008] Gary H. Merrill, Patrick B. Ryan., and Jeffery L. Painter. *Using SNOMED to Normalize and Aggregate Drug References in the SafetyWorks Observational Pharmacovigilance Project*. KR-MED, Phoenix, AZ, USA, 2008. (Poster session.).

[NLM, 2008] NLM. *The Unified Medical Language System*. U.S. National Library of Medicine, 2008. Available at `http://www.nlm.nih.gov/research/umls/`.

[Painter *et al.*, 2006] Jeffery L. Painter, Kristoph Kleiner, and Gary H. Merrill. Inter-translation of biomedical coding schemes using umls. Technical Report FS-06-06, American Association for Artificial Intelligence, Washington, DC, 2006. Fall Symposium on Semantic Web For Collaborative Knowledge Acquisition.

[Painter, 2008] Jeffery L. Painter. *A Mapping Between SNOMED-CT and the OXMIS Coding Scheme*. KR-MED, Phoenix, AZ, USA, 2008. (Poster session.).

[Ryan and Powell, 2008] Patrick B. Ryan and Gregory E. Powell. *Exploring Candidate Differences Between Drug Cohorts Prior To Exposure: A Systematic Approach Using Multiple Observational Databases*. International Society of Pharmacoeconomics and Outcomes Research, Toronto, 2008.

[Ryan *et al.*, 2008] Patrick B. Ryan, Gary H. Merrill, and Jeffery L. Painter. *Defining medical conditions by mapping ICD-9 to MedDRA: A systematic approach to integrating disparate observational data sources for enabling enhanced pharmacovigilance analyses*. Drug Information Association, Boston, 2008. (Poster session.).

[Ryan, 2008] Patrick B. Ryan. A call to action: Emerging opportunities for pharmacoepidemiology to advance the understanding of the effects of medicines. *The PharmacoEpi Newsletter*, 2008. Forthcoming.