PROCEEDINGS
Vaccine and Drug Ontology in the Study of Mechanism and Effect (VDOSME-2012)
VDOSME Chairs

- Yongqun “Oliver” He, University of Michigan, Ann Arbor, MI, USA
- Luca Toldo, Merck KGaA, Darmstadt, Germany
- Gully Burns, Information Sciences Institute, Marina del Rey, CA, USA
- Cui Tao, Mayo Clinic, Rochester, MN, USA
- Darrell R. Abernethy, U. S. Food and Drug Administration (FDA), Bethesda, MD, USA.

VDOSME Programme Committee

- Olivier Bodenreider, National Library of Medicine
- Richard Boyce, University of Pittsburgh
- Lindsay Cowell, University of Texas Southwestern Medical Center
- Dingcheng Li, Mayo Clinic
- Asiyah Yu Lin, University of Michigan
- Bjoen Peters, La Jolla Institute for Allergy and Immunology
- Sira Sarntivijai, University of Michigan/US FDA
- Nigam Shah, Stanford University
- Sunghwan Sohn, Mayo Clinic
- Larisa Soldatova, Aberystwyth University
- Stephen Wu, Mayo Clinic
- Yuji Zhang, Mayo Clinic
- Jim Zheng, Medical University of South Carolina
- Qian Zhu, Mayo Clinic
Programme

Saturday, 21 July 2012

8:30 – 9:00 am Registration

9:00 – 9:10 am Welcome Introduction
• Self-introductions of all attendees
• Oliver He: briefly introduce the whole day schedule with any possible updates

9:10 - 9:50 am
Session 1: Ontology modeling and analysis
Chair: Gully Burns
• Yu Lin and Yongqun He.
  Ontology representation and analysis of vaccine formulation and administration
  and their effects on vaccine immune responses (25 min)
• Drashti Dave and Gully Burns.
  A lightweight Ontology Design Pattern to curate and represent experimental
  variables from vaccine protection studies (15 min)

9:50 - 10:40 am
Session 2: Ontology-based knowledge extraction from big data case reports or
literature
Chair: Luca Toldo
• Harsha Gurulingappa, Abdul-Mateen Rajput, Luca Toldo.
  Extraction of Adverse Drug Effects from Medical Case Reports (25 min)
• Junguk Hur, Arzucan Ö zgür, Zuoshuang Xiang and Yongqun He.
  Identification of fever and vaccine-associated gene interaction networks using
  ontology-based literature mining (25 min)

10:40 – 11:10 am Break

11:10 am – 12:00 noon
Session 3: Ontology-based semantic web applications
Chair: Oliver He
• Qian Zhu, Guoqian Jiang and Christopher Chute.
  Profiling structured product labeling using NDF-RT and RxNorm (25 min)
• Charalampos Doula-verakis, George Nikolaidis, Athanasios Kleontas and
  Yiannis Kompatsiaris.
  GalenOWL: Ontology based drug recommendations discovery (25 min)
12:00 noon - 1:30 pm **Lunch break**

1:30 – 2:05 pm **Session 4: General Discussion #1**  
Moderator: Luca Toldo  
Topic: Relation between drug and vaccine ontology informatics (35 min)  
- Similarities and differences between drug and vaccine informatics  
- Collaborations between drug and vaccine ontology informatics fields  
- Collaborations between academia and industry

2:05 – 2:30 pm **Session 5: Teleconference Presentation by a presenter from USA**  
Chair: Oliver He  
- Cui Tao, Yongqun He, Gregory Poland and Christopher Chute.  
  Time modeling of vaccine adverse events in VAERS for temporal analysis (25 min)

2:30 – 3:30 pm **Session 6: General Discussion #2**  
Moderators: Melanie Courtot and Oliver He  
Topic: Updates on development of adverse event-related ontologies (60 min)  
- Melanie: Updates on the development of the Adverse Event Reporting Ontology (AERO) and related ontologies (short talk: 10-15 min)  
- Oliver: Updates on the development of the Ontology of Adverse Events (OAE) and its applications (short talk: 10-15 min)  
- Further discussion

3:30 - 4:00 pm **Break**

4:00 – 4:45 pm **Session 7: General Discussion #3**  
Moderator: Gully Burns  
Topic: Challenges and opportunities (45 min)  
- Current achievements and challenges  
- International community collaborations  
- Have another workshop next year?  
- Other open issues for discussion

4:45 pm **Workshop ends**
GalenOWL: Ontology based drug recommendations discovery
Charalampos Doulaverakis 1; George Nikolaidis 2; Athanasios Kleontas MD 2,3 and Ioannis Kompatsiaris 1

1Centre for Research and Technology Hellas, Information Technologies Institute, Thessaloniki, Greece
2Ergobyte S.A., Thessaloniki, Greece
3Theagenio Cancer Hospital, Thessaloniki, Greece

ABSTRACT
The paper presents a semantic enabled online service, named GalenOWL, capable of offering real time drug-drug and drug-diseases interaction discovery. For enabling this kind of service, medical information and terminology had to be translated to ontological terms and be appropriately coupled with medical knowledge of the field. International standards such as the ICD-10 classification or the UNII registration, provide the backbone of the common representation of medical data while the medical knowledge of drug interactions is represented by a rule base which makes use of the aforementioned standards. Details of the system architecture are presented while also giving an outline of the difficulties that had to be overcome. A comparison of the developed ontology-based system with a similar system developed using a traditional business logic rule engine is performed giving insights on the advantages and drawbacks of both implementations.

1 INTRODUCTION
One of the health sectors where intelligent information management and information sharing compose valuable preconditions for the delivery of top quality services is personalized drug prescription. This is more evident in cases where more than one drug is required to be prescribed, a situation which is not uncommon, as drug interactions may appear. The problem is magnified by the wide range of available drug substances in combination with the various concoctions in which the former are present. Another factor that makes drug prescription a complex task is the complexity that characterizes the definition of possible interactions or contraindications due to the large number of parameters that are implicated.

Indicatively it is mentioned that according to statistics men over 55 years old consume daily four different medicines on average and the reactions that can occur due to combined prescription are difficult to predict. As an example the substance Donepezil (ATC code: N06DA02) which is prescribed for the treatment of Alzheimer’s disease interacts with 9 other substances and 3 other diseases. If it is taken into account that there exist more than 18,000 pharmaceutica substances including their excipients, then it is clear that the continuous update of health care professionals is remarkably hard. Over this, the extensive literature makes discovery of relevant information a time consuming and difficult process while the different terminologies that appear between sources add more burden on the efforts of medical professionals to study available information.

Semantic Web technologies can play an important role in the structural organization of the available medical information in a manner which will enable efficient discovery and access. Semantic Web has already infiltrated in the public health sector (Wolestencroft et al., 2005) as a mean for representation of available knowledge or though the utilization of reasoning methodologies for automating procedures such as diagnosis, data classification, medical record consolidation, etc.

More specifically, with the use of ontology languages such as OWL, a rather large amount of biomedical ontologies have been developed among them ontologies of large size such as the Biological Pathways Exchange (BioPax) 1 (Ruttenberg et al., 2005), the GALEN 2 ontology (Rector and Rogers, 2006), the Foundational Model of Anatomy (FMA) 3 as well as the Gene Ontology 4 and SNOMED CT 5.

The use of OWL for the expression and representation of the aforementioned ontologies, apart from the benefits regarding knowledge reuse and sharing that come from the use of a standardized language, revealed the benefits of semantic reasoning. The validation of the ontologies using OWL reasoning engines revealed important modelling failures but also a large number of subsumption relations that were missing from the initial requirements and not locating them would mean the loss of important information in patient management systems.

Research projects funded for enabling Semantic Web technologies in the diagnosis and therapeutic procedures exist such as TUMOR 6, REMINE (Ceusters et al., 2008) and PSIP (Beuscrt et al., 2009), with the latter aiming at reducing drug prescription adverse effects through data mining and semantic interpretation of a patient’s medical record. Other projects like NeOn (Suarez-Figueroa and Gomez-Perez, 2008) and Active Semantic Documents (Sheth, 2005) employ ontologies in daily medical practice. Despite the research activity, there have been few proposals for a systematic development of a semantic knowledge base which will aid physicians when prescribing drugs. (Stephens et al., 2006) describes a framework for information integration for drug safety determination using ontologies and in (Adnan et al., 2010) authors suggest an approach to semantically annotate Electronic Discharge Summaries in order to provide decision support to physicians.

The paper presents GalenOWL, a semantic-enabled system for discovering drug recommendations and interactions. GalenOWL makes use of established and standardized medical terminologies together with a rich knowledge base of drug-drug and drug-diseases

1 BioPax, http://www.biopax.org/
5 http://www.ihtsdo.org/snomed-ct/
6 TUMOR, http://www.tumor-project.eu/
interactions expressed as rules and OWL axioms. GalenOWL is implemented as an online service having in mind, both completeness of results and responsiveness in query answering.

The paper is organized as follows: Section 2 gives development details, presents the architecture in terms of the different ontologies and rules that were integrated in the system and elaborates on the design decisions that were made. Section 3 describes the user interface and presents the series of queries for obtaining the drug recommendations. Section 4 gives metrics regarding system’s performance and compares GalenOWL with a similar system developed using traditional business logic data structures and a production rule engine. The paper concludes with Section 5 which gives an overview of the paper and discusses future work.

2 DEVELOPMENT

The stimulus for developing GalenOWL was given by an already available market product. The GALINOS drug guide, available at http://www.galinos.gr in Greek, is an online service where a user can query the drug database and get information on available drugs that are found in the market, e.g. indications, recommended dosage, excipients, interactions, adverse effects, etc, where all the latter are related to the drugs active substances. All the above were mined after extensive research in the literature and of available documents such as Summary of Product Characteristics (SPC) and Patient information leaflets (PIL). For enabling this kind of functionality GALINOS employs international medical standards which allow a unique identification of diseases and substances. It was evident that the knowledge integrated in the service could be used in order to develop an intelligent system for offering drug recommendations.

Fig. 1. GalenOWL usage

GalenOWL architecture can be seen in Fig. 1. The user issues queries to the system in order to find drug indications and contraindications that match patient data. These data populate the knowledge base and rule-based reasoning is performed. The reasoning engine makes use of the medical ontologies and the rule base for drug recommendations and a list of the drug recommendations (indications and contraindications) is returned by the engine. GalenOWL is novel in its field as, to the authors knowledge, the are no commercial systems that offer drug-diseases interactions. Systems that offer drug-drug interactions are available such as the one offered by Drugs.com

2.1 Development details

For the OWL/XML serialization, the Jena Semantic Web Framework\(^7\) was used. The OWL reasoner which provided the drug recommendations is OWLIM-Lite\(^9\) together with Sesame\(^10\) for providing the REST interface, the RDF data access and management platform and the SPARQL query interpretation layer. OWLIM was chosen as it has been found as one of the most efficient OWL reasoners (Bock et al., 2008; Bishop et al., 2011).

2.2 International standards ontologies

In order to provide such a service, coupling of Semantic Web and medical terminologies was needed. GalenOWL is built on top of OWL ontologies which express international standards of medical terminology in order to process requests for drug recommendations. The following terminologies are expressed as OWL ontologies:

- **ICD-10**: The World Health Organization classification of diseases. It is used in GalenOWL for unique identification of diseases thus uniquely identifying drug indications and contraindications related to diseases.
- **UNII**: Unique Ingredient Identifier. Used for the identification of active ingredients found in drugs. In GalenOWL it is used for uniquely identifying drug indications and contraindications related to ingredients.
- **ATC**: The Anatomical Therapeutic Chemical Classification is used for the classification of drugs. In GalenOWL it is used in similar fashion to UNII.

Each code in the above encodings is expressed as an OWL class.

2.3 Domain ontologies

Besides these international standards, two more classifications are expressed in OWL in order to make easier use of the system:

**Substance**: As the use of encodings for drug ingredients is not convenient for humans, the identification of active substances is done using its common name references in medical bibliography. These names come from international standards such as the International Nonproprietary Names (INN) and others such as USAN (United States Adopted Name) or BAN (British Approved Name). Members of this identification list are substances such as *acetazolamide* or *isradipine*. In addition, substances correspond to ATC codes and this is captured in the ontology through class equivalence such that for example *acetazolamide* $\equiv$ S01EC01.

**Condition**: As certain “groups” of substances and/or diseases are frequently present in drug interactions and these groups are not recorded explicitly in any standardized classification, it is more convenient for medical use to specify these custom groups. These often used groups are termed “conditions” in GalenOWL and are defined by medical experts. An example of such condition is *barbiturates-drugs* which is defined as

$$\text{barbituratesDrugs} = \text{a/N01AF} \mid \text{a/N01AG} \mid \text{a/N03AA} \mid \text{a/N05CA} \mid \text{a/N05CB} \mid \text{a/N05CX}$$

where “a/” stands for ATC code and “/” stands for “or”. So any member of these premises is also a member of \textit{barbituratesdrugs}. In addition a condition can appear as a premise in other condition definitions. So the condition

\texttt{hemorrhage-postoperative} \equiv \texttt{c/hemorrhage-nos} \& \texttt{c/surgical-dental-procedures}

is satisfied when two other Conditions (denoted by “c/”) are satisfied simultaneously (denoted by “\&” which stands for “and”). It is evident that conditions can be effectively expressed in OWL as defined classes. In the above examples, “a/” or “c/” would represent the names of the ontology and “/” or “&” would represent the union or intersection of classes respectively. Using DL notation the above classes are represented as

\texttt{barbituratesdrugs} \equiv \texttt{N01.AF} \sqcup \texttt{N01.AG} \sqcup \texttt{N03.AA} \sqcup \texttt{N05.CA} \sqcup \texttt{N05CB} \sqcup \texttt{N05CX}

and

\texttt{hemorrhage-postoperative} \equiv \texttt{hemorrhage-nos} \sqcap \texttt{surgical-dental-procedures}

In order to automate the definition of the Conditions ontology, a parser was developed to express the conditions from the custom format explained above to OWL/XML notation.

All the above mentioned ontologies were imported from the GalenOWL core ontology depicted in Fig. 2. Additionally, Patient is the class for patient instances. Patient instances are related with the MedicalDefinitions and with AgeGroup and SexGroup through the hasAgeGroup and hasSexGroup properties respectively.

\texttt{Fig. 2. GalenOWL core ontology}

2.4 Rule base

After the definition of the domain ontologies and the core ontology, an appropriate rule base for indications and contraindications was defined. The rules are expressed in a custom language similar to the Conditions of the previous subsection. These rules however usually have a more complex syntax. An example rule for the indication of \texttt{rimonabant} is defined as

\texttt{rimonabant} = \texttt{i/E65-E68} \& \{\texttt{i/E11} | \texttt{i/E78}\}

where “i/” stands for ICD-10 code and it reads as: \texttt{rimonabant} is indicated in cases where E65-E68 and, E11 or E78, diseases are present. In DL this is represented as \texttt{“rimonabant” \equiv E65-E68 \sqcap \{E11 \sqcup E78\}}. Due to GalenOWL being developed using OWLIM-Lite, the above expression had to be expressed in the OWLIM custom rule language. “or” could not be expressed in a rule, so two different rules were generated for \texttt{rimonabant}. To make things more complicated, drug indications also depend on the patient’s sex and age. In the above example, \texttt{rimonabant} is prescribed only for adults or elder patients so this also had to be encoded in the rules. As a result, 4 rules were built for \texttt{rimonabant} indication, i.e.

1) \texttt{Patient(?p), hasData(?p, go:E65-E68), hasData(?p, go:E11), hasAgeGroup(?p, go:adult) \rightarrow canTake(?p, go:rimonabant)}

2) \texttt{Patient(?p), hasData(?p, go:E65-E68), hasData(?p, go:E78), hasAgeGroup(?p, go:adult) \rightarrow canTake(?p, go:rimonabant)}

3) \texttt{Patient(?p), hasData(?p, go:E65-E68), hasData(?p, go:E11), hasAgeGroup(?p, go:elder) \rightarrow canTake(?p, go:rimonabant)}

4) \texttt{Patient(?p), hasData(?p, go:E65-E68), hasData(?p, go:E78), hasAgeGroup(?p, go:elder) \rightarrow canTake(?p, go:rimonabant)}

Of course indication rules have no limitation in the premises separated by “or” which can lead to a very big rule expansion. As an example, \texttt{buspirone} has 13 premises separated with “or” which leads to 13 different rules. In the current version of GalenOWL 1342 substance indications/contraindications were expressed using 9266 rules. A parser similar to the one developed for Conditions was used in order to express the indications in the OWLIM custom rule language. Although the rule base is quite large in size, OWLIM’S sophisticated indexing structure and rule engine was quite fast in evaluation of rule activation. An additional rule, elaborated as \texttt{“canTake(?p, ?s), cannotTake(?p, ?s) \rightarrow hasSubstanceConflict(?p, ?s)"}, was necessary in order to find substances that would appear both in the indications and contraindications lists.

Finally, for each rule an instance under Indications or ContraIndication class (both subclasses of SubstanceRecommendations) is created and the property hasTextualRepresentation is set to the original textual representation of the rule. This is used in order to provide tracing in rule matching so that for each rule that is activated the property activatedRule(patient, recommendation) is materialized. These relations are depicted in Fig. 2. In the GalenOWL ontologies a total of 28,867 named classes were defined.

3 INTERFACE AND QUERYING

The interface to the system is depicted in Fig. 3. As the focus on the system was on the functionalities that can be provided and on its capabilities, the design of the interface may lack in aesthetic design nevertheless it provides all the information that are returned from the system in a rather easy to use layout.

Patient data regarding diseases, allergies, population group and current medication are entered sequentially using the form. After all data are entered, the user submits all information to the system in a rather easy to use layout. In the above example, \texttt{rimonabant} is prescribed only for adults or elder patients so this also had to be encoded in the rules. As a result, 4 rules were built for \texttt{rimonabant} indication, i.e.

1) \texttt{Patient(?p), hasData(?p, go:E65-E68), hasData(?p, go:E11), hasAgeGroup(?p, go:adult) \rightarrow canTake(?p, go:rimonabant)}

2) \texttt{Patient(?p), hasData(?p, go:E65-E68), hasData(?p, go:E78), hasAgeGroup(?p, go:adult) \rightarrow canTake(?p, go:rimonabant)}

3) \texttt{Patient(?p), hasData(?p, go:E65-E68), hasData(?p, go:E11), hasAgeGroup(?p, go:elder) \rightarrow canTake(?p, go:rimonabant)}

4) \texttt{Patient(?p), hasData(?p, go:E65-E68), hasData(?p, go:E78), hasAgeGroup(?p, go:elder) \rightarrow canTake(?p, go:rimonabant)}

Of course indication rules have no limitation in the premises separated by “or” which can lead to a very big rule expansion. As an example, \texttt{buspirone} has 13 premises separated with “or” which leads to 13 different rules. In the current version of GalenOWL 1342 substance indications/contraindications were expressed using 9266 rules. A parser similar to the one developed for Conditions was used in order to express the indications in the OWLIM custom rule language. Although the rule base is quite large in size, OWLIM’S sophisticated indexing structure and rule engine was quite fast in evaluation of rule activation. An additional rule, elaborated as \texttt{“canTake(?p, ?s), cannotTake(?p, ?s) \rightarrow hasSubstanceConflict(?p, ?s)"}, was necessary in order to find substances that would appear both in the indications and contraindications lists.

Finally, for each rule an instance under Indications or ContraIndication class (both subclasses of SubstanceRecommendations) is created and the property hasTextualRepresentation is set to the original textual representation of the rule. This is used in order to provide tracing in rule matching so that for each rule that is activated the property activatedRule(patient, recommendation) is materialized. These relations are depicted in Fig. 2. In the GalenOWL ontologies a total of 28,867 named classes were defined.
In order to provide an overall view of the drug recommendations that are returned by GalenOWL, the following sequence of actions is performed: Each patient data (disease, allergy, current medication) that is in the list is inserted separately and inference is performed. This is done so that the user can have a list of recommendations that is due to each data separately. In a final step all data are entered simultaneously so that recommendations that are valid for all patient’s data are evaluated. All recommendations are separated in 4 groups, the indications list, the contraindications list, the conflicts between indications and contraindications, i.e. substances that appear both in indications and contraindications which can be expressed as \((\text{indications} \cap \text{contraindications})\), and a cleared list where only indications that do not appear in contraindications are present. This list actually represents the valid recommendations of the system for the patient’s prescription, i.e. \((\text{indications} \setminus (\text{indications} \cap \text{contraindications}))\). Results lists are separated in tabs and each tab corresponds to one of the sequential steps described above.

4 EVALUATION AND RESULTS

In order to verify GalenOWL’s functionality in terms of results completeness in drug recommendations, a series of random queries regarding patient data (diseases, current medication, population groups, etc) were submitted to the system and the results were evaluated by a medical expert. The analysis concluded that the results were as expected and all patient’s conditions were taken into account. A series of tests in order to determine initialization time, memory consumption and query response time of GalenOWL have been performed. These values are reported in the first row of Table 1 where promising results are reported especially in query response time which is kept at satisfying 16 ms average time. The initialization phase, which included compilation of the rule base and loading of the ontology in the main memory took 148 seconds which is reasonable if one takes into account the large volume of the knowledge base (ontologies plus rules) and that this is a one time task executed during initialization. What is less than ideal is the memory consumption after the initialization phase which stays constant at around 649 MB and takes up a fairly big amount of system resources which is something undesirable in a production environment.

For having a broader view of GalenOWL’s performance, a similar system has been developed using standard business logic programming technologies. This system has been termed GalenDrools as in its core for drug recommendations lies the Drools rule engine (Drools, 2012) which is an open source and efficient framework for business logic integration.

To give a brief description of GalenDrools implementation, ICD-10, ATC and UNII encodings as well as Substance and Conditions, are stored in a database. For building the rule base the indications/contraindications rules are parsed and translated to the Drools rule language (DRL). When premises for ICD-10 or ATC classification codes are present in the rule body, the latter is automatically populated with upper level codes of of the classification, in a manner similar to the Class/SubClass relation in ontologies. One more different aspect of GalenDrools architecture is the way that Conditions are handled. While in GalenOWL Conditions are translated into OWL defined classes, here each condition that appears in a rule is recursively expanded to its primitive elements, i.e. ICD-10, ATC, UNII or Substance codes. When requesting drug recommendations, patient data are inserted as facts in the Drools truth maintenance table and rule execution is initiated. These facts actually correspond to the database IDs of the ICD-10, ATC, UNII and Substance codes which makes rule matching quite fast.

A direct comparison between GalenOWL and GalenDrools reveals that in almost all aspects the business logic implementation of the drug recommendations system outperforms the semantic-enabled implementation by an order of magnitude. Initialization of GalenOWL takes more time as the rule base has to be compiled and all inferences computed during the ontology loading. Memory consumption is high as the whole ontology and rules base have to be loaded in memory. On the contrary, in GalenDrools the initialization phase includes only the compilation of the rule base which is the only structure stored in memory thus making it more efficient both in startup time and in memory consumption. Regarding query response time, in GalenOWL when a new patient instance is inserted inference is performed which leads to increased response time compared to GalenDrools where simple rule matching is performed.

On the other hand, although inference adds burden and overhead to query response, it actually makes development of the system easier. In the business logic implementation both the subclass relations and the expansion of Conditions had to be implemented programmatically by hand and encoded in the rule body, a process which requires effort and increases the possibility to induce errors but also combine medical knowledge, e.g. for Condition definition, with drug administration rules. In contrast, the OWL-based implementation all the above were taken care by the reasoning engine, i.e. issues such as hierarchical class relationships and derived consequences such as class membership, and medical knowledge is defined in the ontology. Table 2 provides a qualitative comparison between the two approaches.

<table>
<thead>
<tr>
<th></th>
<th>Initialization</th>
<th>Memory</th>
<th>Query time</th>
</tr>
</thead>
<tbody>
<tr>
<td>GalenOWL</td>
<td>148 s</td>
<td>649 MB</td>
<td>16 ms</td>
</tr>
<tr>
<td>GalenDrools</td>
<td>41 s</td>
<td>74 MB</td>
<td>5 ms</td>
</tr>
</tbody>
</table>

Table 1. GalenOWL system performance compared to similar developed in Drools (GalenDrools)
The efficiency of production rule engines has already been utilized in semantic web literature. In (Meditskos and Bassiliades, 2008) the authors use the CLIPS rule engine as an OWL reasoner after transforming the OWL ontology to the COOL object oriented language of CLIPS. However ontology management and querying are made difficult. In OWLJessKB (OWLJessKB, 2007) the Jess rule engine is used for OWL reasoning where the RDF triples are inserted as facts and OWL entailments are materialized using production rules. This approach though suffers from memory limitations. It should be noted that business rule engines have been around for much longer time than OWL reasoners and they are aimed at much larger audience than Semantic Web technologies. This alone corresponds to a much larger community contributing to frameworks like Drools. These two facts can account for the exceptional performance that these systems present. The authors believe that as the Semantic Web community grows larger, more frameworks that will be able to compete traditional rule engines will be made available. OWLIM is an example of an efficient reasoning engine and up to now several other reasoners are claiming increased performance such as Hermes (Glimm et al., 2010) and TrOWL (Thomas et al., 2010).

5 CONCLUSION

In this paper a drug recommendation system based on semantic web technologies, termed GalenOWL, was presented. It has been shown that OWL and semantic web can provide a good match for drug recommendations as OWL is expressive enough to effectively encapsulate medical knowledge. Rule-based reasoning can model medical decision making and provide assistance to experts. A comparison of the semantic-enabled implementation to a traditional business logic implementation was presented. Although the latter has shown better performance in time and memory requirements, semantic technologies provide a better alternative for integrating knowledge in the system than simple rule engines. Future work, apart from the expansion of the semantic rule base, will include prioritization of interactions as not all interactions have the same importance. Additional work will be directed to research oriented performance optimizations, such as context extraction from medical knowledge and from queries which will lead to modular ontologies, so that not to take into account the whole ontology during query time. This will result in less memory utilization and better query response times.

ACKNOWLEDGEMENTS

This work was supported by GSRT Hellas under the “Vouchers for SMEs” funding program 2010 and by the national project GNORASI co-funded by GSRT and EU.

REFERENCES


Time Modeling of Vaccine Adverse Events in VAERS for Temporal Analysis

Cui Tao1,*, Yongqun He2, Gregory A Poland3, Christopher G Chute1
1 Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, MN
2 Unit for Laboratory Animal Medicine, Department of Microbiology and Immunology, and Center for Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI
3 Mayo Vaccine Research Group, Mayo Clinic, Rochester, MN

ABSTRACT
The U.S. FDA/CDC Vaccine Adverse Event Reporting System (VAERS) provides a valuable data source for post-vaccination adverse event analyses. The structured data in the system has been widely used, but the information in the write-up narratives is rarely included in this kind of analyses. In fact, the unstructured nature of the narratives makes the data embedded in them difficult to be used for any further studies. In this paper, we introduce an ontology-based approach for representing the data in the narratives in a “machine-understandable” way, so that it can be easily queried and further analyzed. Our focus is the time aspect in the data for time trending analysis. The Time Event Ontology (TEO), Ontology of Adverse Events (OAE), and Vaccine Ontology (VO) are leveraged for the semantic representation of both structured and unstructured data in the VAERS reports. A use case evaluation has been conducted to illustrate how the above three ontologies can be used together to model the data in from the VAERS report and how we envision to use LifFlow, an event sequence visualization tool, for time trend analysis.

1 INTRODUCTION
Effective analyses of time trends for post-vaccine adverse events (AEs) can enhance clinical research in different areas such as vaccine safety analyses, causality assessments, and retrospective studies. The FDA Vaccine Adverse Event Reporting System (VAERS) provides a valuable data set for these purposes. VAERS maintains a database for reports of AEs following vaccination. These reports contain both structured data (e.g., gender, age, vaccination date, and onset date), as well as short narratives that usually provide more detailed descriptions of the vaccination, the related events, and their time constraints.

The structured data in the VAERS database have been widely leveraged in different medical analyses for vaccine adverse events. The unstructured nature of the narratives, however, makes the data embedded in them difficult to be used in further analyses. These narratives usually contain additional valuable information (e.g., patient ages that were not reported in a structured way, vaccination doses, and durations or time stamps for multiple events following and concrete clinical analyses.

Targeting these challenges, we have designed the Temporal Information Modeling, Extraction, and Reasoning (TIMER) framework for extracting, querying, and inferring useful temporal information automatically. Figure 1 shows the TIMER system overview. One core component of TIMER is the modeling component. Our vision is to leverage ontologies to semantically model the domain and the time knowledge. TIMER relies on the ontologies as the annotation schema for its extraction component, and as the knowledge base for its reasoning component. It is essential to ensure that the ontologies are capable to represent related data faithfully in an integrated, machine-understandable way, so that computer programs can automatically process the data, infer new knowledge, sort clinically relevant events over the timeline, and facilitate data querying for clinical research analyses.

This paper focuses on the modeling component. We introduce our efforts to leverage Semantic Web mechanisms to represent time-related information for vaccine adverse events from the VAERS database. We use the Ontology of Adverse Events (OAE, previously named Ontology of Adverse Event or AEO) and the Vaccine Ontology (VO) for representing vaccine names and event names in a standard way. We use the Time Event Ontology (TEO) to represent

Figure 0: TIMER System Overview

* To whom correspondence should be addressed: tao.cui@mayo.edu
time information among the events.

The rest of the paper is organized as follows. We first introduce the three ontologies (TEO, OAE, and VO) and how they can be used to represent time in the area of vaccine adverse events. A VAERS case report is then presented as a use case for the ontological representations. The advantages of using our ontology-based semantic web representation and data analysis are emphasized. Finally, we draw conclusions and discuss further improvements.

2 ONTOLOGY REPRESENTATION FOR VACCINE ADVERSE EVENT TIME ANALYSIS

2.1 TEO and time relation reasoning

The TEO is an OWL ontology designed for representing the temporal relations among events and time stamps. It is an extended version of the Clinical Narrative Temporal Relation Ontology (CNTRO)\(^6\), which is our effort on representing basic temporal relations among events extracted from clinical narratives. We expand the CNTRO to TEO since we believe that the time modeling should be general for all the domains. The TEO aims to model time event relationships in general, not specific to the clinical or biomedical domain. Figure 2 shows a high-level overview of the TEO. All the TEO concepts were derived from the Basic Formal Ontology (BFO)\(^8\) to ensure upper-level interoperability with other ontologies. A TEO event is equivalent to the BFO term ‘processual entity’ that is defined as: “an occurrence that exists in time by occurring or happening, has temporal parts and always involves and depends on some entity”. Events could be administration of medication (e.g., a vaccine), follow-up lab testing, patient complaints, adverse events with different symptom outcomes, etc. Time stamps of when these events occurred can be represented if known via the hasTime relation. Duration information of a particular event (if it spanned minutes, hours, days) or between events can be annotated, as well as the temporal relationships between events implied throughout the narrative through use of words such as “before”, “after”, “equal”, etc. Granularity of a specific time stamp or frequency of the re-occurred time can also be modeled. In addition, the TEO also covers periodic time interval which models the time of the events that repeat themselves (e.g., “Exercise 20 minutes 3 times/day starting from July 21 for 2 weeks”). It semantically defines specific temporal regions such as holidays, weekdays, etc, as well as specific temporal intervals such as days, weeks, months, today, tomorrow, etc.

The TEO provides a representation mechanism to model temporal information stated within adverse event reports in a “machine-understandable” way. Sorting out the events on a timeline or answering time-related clinically significant questions, however, usually cannot be accomplished by querying the information explicitly stated with the reports. Many times it requires semantic inference to fully answer the time-relevant questions.

2.2 OAE and VO Usage

The OAE is an OWL ontology for representing adverse events. In addition to the principles of Ontological Realism\(^11\), the development of OAE follows the OBO Foundry principles: openness, collaboration, and use of a common shared syntax\(^12\). OAE is aligned with the top ontologies such as Basic Formal Ontology (BFO)\(^13\) and the Relation Ontology (RO)\(^13\).

In the current version of OAE, the term ‘adverse event’ is defined as a pathologico bodily process (OGMS:0000061) that occurs after a medical intervention and is likely induced by the medical intervention. Examples of the medical interventions include vaccination, drug administration, usage of medical devices, and surgery. An adverse event may or may not be caused by a medical intervention. In OAE, we specifically define a term ‘causal adverse event’ as a pathologico-
cal bodily process that is induced/caused by a medical intervention. Currently, OAE has 2464 representational units, annotated by means of 981 terms with specific such as OAE identifiers, and the other terms imported from existing ontologies including BFO, RO, and the Ontology of Biomedical Investigations (OBI). The importing of external ontology terms avoids regeneration of new ontology terms that are not in the scope of the adverse event domain and supports efficient ontology reuse on the condition that the feeder ontologies are based on the same principles.

The VO is a community-based ontology in the domain of vaccine and vaccination. Like TEO and OAE, VO is developed in OWL and aligned with BFO and RO. VO has classified all existing vaccines licensed for human and animal uses in the U.S.A. and Canada. For each licensed vaccine, VO also includes relevant attributes such as vaccine type, vaccine components (e.g., antigens, adjuvant and preservatives), vaccination route, manufacturer, and the disease and pathogen targeted by the vaccine. All these data are organized in ontological format and shared syntax, supporting automated reasoning and SPARQL query.

MedDRA (http://www.meddrasso.com/) is used as the default controlled vocabulary for describing adverse event terms in VAERS. To better represent adverse events in OAE, we have made a match (cross-reference) between many OAE terms and MedDRA terms. Compared to MedDRA, OAE uses a formal ontology format with machine-readable logical definitions and structures. OAE also imports vaccine-specific information from VO, making it an ideal platform for analyzing vaccine time events.

In order to make OAE, VO, and TEO work seamlessly, alignments between these three ontologies have been made. All these ontologies use BFO as the top ontology. TEO is imported to OAE as a middle-layer ontology for representation of time. All adverse events in OAE are subclasses of TEO events (i.e., BFO term ‘processual_entity’). Therefore, these adverse event classes in OAE automatically inherit TEO methods for time representation.

3 CASE EVALUATION

Let us use the sample file in Figure 3 as a running use case to illustrate how we can use an ontology-based approach to represent and infer useful temporal information, as well as facilitate time trend analysis. The database provides dates such as “Vaccinated”, “Onset”, “Submitted”, and “Entered” in a structured way. However, these dates, in many cases, are not accurate or sufficient for time trend analysis. In the above example, we can observe that the onset date in the structured report is only one date: 1990-03-29. In the write-up, we can see there were actually a series of effects that happened after the vaccination, on different days. The additional information in the write-up narrative, however, is not easy to query or be processed by computers, or even by human experts. If the information of interest from the narrative is represented and inferred, it can be queried and processed easily for further analysis.

3.1 OAE and VO modeling

OAE and VO can be used to represent adverse events and related vaccine information in an ontological format. The original case reports use MedDRA to represent symptoms. In OAE, we normalize the symptoms shown after vaccination as adverse events (AEs). For example, we represent the nausea in Figure 3 as ‘nausea AE’. The OAE representation shows that the patients had adverse events in different areas (e.g., skin, joint, and digestive system). It is noted that an ‘influenza AE’ may not be induced by influenza virus. The better term to represent the case is an OAE term ‘influenza like illness AE’ (OAE_000100). The patient in the case reported was vaccinated with the vaccine Engerix-B, which is semantically defined in VO with a VO ID of VO_0010711. VO provides the hierarchical information as well as the semantic assertions associated for different vaccines. In the future, the vaccine information can be directly imported to OAE to support efficient automated reasoning.

3.2 TEO modeling of the time events in this case:

The recognized named entities will be annotated using terms in TEO as well as domain ontologies (OAE & VO) to create a set of attribute-value pairs. Figure 4 shows some examples. We have recognized 7 events (shown in the list) as well as 6 time instants (not shown) from the short narratives. These events are also annotated with respect to domain ontologies so that the computer program “understands” the semantic meaning of the events. For example,
“FLU-LIKE SYMPTOMS” is annotated with the OAE term ‘influenza like illness AE’, which is supposed to be aligned with a MedDRA term “influenza like illness” (MedDRA ID: 10022004). This step is very important for normalizing event names for time trending analyses. For example, there could be multiple ways to describe “FLU-LIKE SYMPTOMS” such as “Influenza-like illness”, “flu-like illness”, etc. Only when these different expressions are annotated with the same ontology term, can a computer program know that they refer to the same meaning. The extracted relationships will also be represented in a computer-understandable way (in our case in the Rich Description Format or RDF, which is a standard data representation mechanism in Semantic Web), so that they can be easily queried too. Figure 4, for example, shows how we represent the relationship between “1ST ENGERIX-B DOSE” and its timestamp “3-29-90”.

Figure 4: The annotation results of the example in Figure 1 (Partial): The upper panel shows the information of interest highlighted. The bottom left panel shows classes with annotated instances. And the bottom right panel shows a recognized temporal relation between “1ST ENGERIX-B DOSE” and its time stamp. This figure is generated as a screenshot using the Semantator (http://informatics.mayo.edu/CNTRO/index.php/Semantator) semi-automatic annotation environment.

3.3 Visualization of time events:

With our ontology-based approach, we can represent both the structured data and the data extracted from the write-up narratives with respect to our ontologies to create an integrated data repository for further data analysis. Since the data is represented in a “machine-understandable” way, it is therefore possible for computer programs to query or infer temporal relationships between the events, or to sort events of interest on the timeline. We believe this is a very necessary step for improving the time trending analysis for vaccine adverse events. Figure 5, for example, shows the event sequences for three sample data files in the LifeFlow event sequence visualization environment.

3.4 Interpretation of time event analysis results:

We randomly selected three VAERS reports for patient vaccination with HEP vaccine. Although analyzing three reports may not be statistically significant, we would like to use these three samples to illustrate how we may interpret clinical significant or meaningful results from the data.

(i) Many symptoms are shared but may occur at different time points. For example, rash is shared by the first two patients, one at day 3, and the other at day 8. Influenza and arthralgia symptoms also are reported by the first two patients but at different days. The third patient had an injection site reaction on the second day instead of the first day (for the first two patients).

(ii) Frequency of a symptom may be different. Nausea occurred twice for the first patient. However, it occurred only once for the second patient and did not happen in the third patient.

(iii) Sequence of the adverse events may be different: The first patient had an injection site reaction then arthralgia. The two symptoms also showed up in the second patient in the opposite order.

Our approach provides a way to identify these differences, which can be used for further investigation.

4 CONCLUDING REMARKS AND FUTURE WORK

In this paper, we introduce an ontology-based approach for representing time-related information from the VAERS repository. We believe that an ability to representing both the structured data and the data from write-up narratives in an integrated, unified, and “machine-understandable” way can enable research in vaccine safety analyses, causality as-

---

1 Note that in Figure 2, “FLU-LIKE SYMPTOMS” in the write-up was annotated as MedDRA term “influenza” (in the symptom section). This is an inaccurate alignment done by the VAERS database as MedDRA has a more accurate term “influenza like illness” (MedDRA ID: 10022004).
Time Modeling of Vaccine Adverse Events in VAERS for Temporal Analysis

...sessments, and related retrospective studies. Based on the representation mechanisms defined in the ontology, we will implement tools for automatically extracting information such as event names, vaccine names, as well as temporal relationships from the VAERS system. We also plan to build a tool for statistical analysis on top of the integrated data integrated by representation with ontologies.

ACKNOWLEDGEMENTS

This project was supported by the National Science Foundation under Grant #0937060 and the National Center for Biomedical Ontologies (NCBO) to C.T. and the NIH–NIAID grant R01AI081062 to Y.H.

REFERENCES

Identification of fever and vaccine-associated gene interaction networks using ontology-based literature mining

Junguk Hur\(^1\), Arzucan Özgür\(^2\), Zuoshuang Xiang\(^3\), Yongqun He\(^3\)*

\(^1\) Department of Neurology, University of Michigan, Ann Arbor, MI 48109, USA
\(^2\) Department of Computer Engineering, Bogazici University, 34342 Istanbul, Turkey
\(^3\) Unit for Laboratory Animal Medicine, University of Michigan, Ann Arbor, MI 48109, USA

**ABSTRACT**

Fever is one of the most common vaccine adverse events. The detailed mechanisms of fever and vaccine-associated gene interaction networks are not fully understood. In the present study, we employed a genome-wide, centrality and ontology-based literature mining approach to analyse the genes and gene interaction networks associated with fever or vaccine-related fever responses. Over 170,000 fever-related articles from PubMed abstracts and titles were retrieved and analysed at the sentence level to identify genes and vaccines (including 186 Vaccine Ontology terms) as well as their interactions. In total, 403 genes and 577 gene interactions were identified in the generic fever network, and 29 genes and 28 gene interactions were found in the vaccine-specific fever sub-network. Vaccines (including 4 specific vaccine names) were found to directly interact with 26 genes. Making scientific discoveries and generating new hypotheses were possible by using network centrality and gene set enrichment analyses. For example, our study found that the genes in the generic fever network were more enriched in cell death and responses to wounding, and the vaccine sub-network had more gene enrichment in leukocyte activation and phosphorylation regulation. Interestingly, no Toll-like receptors (TLRs) were found in the gene-vaccine interaction network. Since multiple TLRs were found in the generic fever network, it is reasonable to hypothesize that vaccine-TLR interactions may play an important role in inducing fever response, which deserves a further investigation. This study demonstrated that ontology-based literature mining is a powerful method for analyzing gene interaction networks and generating new scientific hypotheses.

1 INTRODUCTION

Fever, or pyrexia, is a symptom of abnormal elevation of body temperature, usually as a result of a pathologic process. Normal body temperature is ranged 98-100°F (36.5-37.5°C), and any temperatures above this range are usually considered febrile. Increased body temperature usually indicates possible presence of infection or sepsis. Once an infection occurs, the body responds to control the infection, often resulting in increased temperature. The fever in response to infection is likely a cure to remove infection and create a favorable environment for immune compartments such as white blood cells (Gardner, 2012). Nevertheless, a long-lasting fever can cause devastating effects. Therefore, reducing fever either with medication or physical cooling methods remains as a common practice.

Body temperature is mainly regulated by the hypothalamus in the brain, which is responsible for coordinating complex homeostatic mechanisms (Nakamura, 2011). A pyrogen, a trigger of the fever, stimulates a release of prostaglandin E2 (PGE2), in turn stimulating the hypothalamus to generate a systemic response to increase the body temperature. Many genes involved in this fever response have been identified to date. For examples, phospholipase A2 (PLA2), cyclooxygenase-2 (COX-2), and prostaglandin E2 synthases (PTGES) are directly involved in the arachidonic acid metabolism pathway that releases PGE2 (KEGG, 2012). Major cytokines including interleukin 1 α/β (IL1A/B), interleukin 6 (IL6), and tumor necrosis factor α (TNFA), are also pyrogens (Dinarello, 2004). These major cytokines and other minor pyrogens such as interleukin-8 (IL8) and interferon-α/β/γ (INF-A/B/G) activate the arachidonic acid pathway. Many vaccines can also frequently cause fever. Vaccine-induced fever may be required for the induction of protective immunity or might be an undesired adverse effect. However, how vaccination perturbs which fever-related genes to cause the adverse event is still unclear.

We previously demonstrated that high throughput literature mining and the use of ontology can significantly enhance our understanding of vaccine research (Hur, et al., 2011; Ozgur, et al., 2010; Ozgur, et al., 2011). First, to better understand interferon gamma (IFN-γ) and vaccine-mediated gene-interaction networks, we developed a literature-based discovery (LBD) approach that integrates text mining with network centrality analysis (Ozgur, et al., 2010). Here the interaction network represents a network with various direct and indirect interactions. Gene interaction networks were generated from the biomedical literature using natural language processing (NLP) and the most important genes in these networks were identified by network centrality analyses using four types of centrality measures: degree, eigenvector, closeness, and betweenness centralities. Integrating these multiple centrality-based core gene sets in the vaccine subdomain resulted in identification of vaccine-specific sub-network of IFN-γ (Ozgur, et al., 2010).

\(^*\) To whom correspondence should be addressed: yongqunh@umich.edu
\(^\xi\) These authors equally contributed.
In an extended study, the application of the Vaccine Ontology (VO) has significantly improved the analysis of the vaccine-specific IFN-γ sub-network (Ozgur et al., 2011). The community-based VO classifies existing vaccines in licensed use, on trial, or in research (He, et al., 2009). The relations between different VO vaccine terms have been logically defined and support advanced semantic reasoning. Another study from our group demonstrated that VO-based literature mining provided a better performance in retrieving Brucella vaccine-related literature and building gene interaction network than the MeSH-based approach (Hur, et al., 2011). These studies helped to generate new candidate genes for vaccine development.

Here we report the application of ontology-based literature mining methods to retrieve gene-gene and gene-vaccine interaction networks associated with fever or vaccine-associated fever processes. Central genes and enriched biological functions are identified in these interaction networks.

2 METHODS

2.1 Literature corpus
Fever-related literature was defined by a PubMed (http://www.ncbi.nlm.nih.gov/pubmed) query, “Fever OR Hyperthermia OR Pyrexia OR Febrile OR Pyrexial”. The sentences in the titles and abstracts of this fever literature cohort were obtained from the BioNLP database in the National Center for Integrative Biomedical Informatics (NCIBI; http://ncibi.org/).

2.2 Gene and vaccine name identification
SciMiner, a dictionary- and rule-based literature mining tool (Hur, et al., 2009), was used to identify gene names in the fever-related literature. Identified genes were reported in terms of the official human genes based on the HUGO Gene Nomenclature Committee (HGNC) database (http://www.genenames.org/). To identify vaccine names, a modified version of VO-SciMiner (Hur, et al., 2011), was employed using a set of 186 VO vaccine terms. These terms are specific vaccine names at the bottom-level of the ontology hierarchy under the term “vaccine”. This list of terms does not include those names that are common and do not help in specific name identification.

2.3 Generation of gene-gene and gene-vaccine interaction networks
To include only gene-gene and gene-vaccine pairs with potentially true interactions rather than a simple co-occurrence in a sentence, we employed a machine-learning-based scoring system (Ozgur, et al., 2011). Briefly, the dependency parse trees of the sentences were obtained using the Stanford Parser (http://nlp.stanford.edu/software/lex-parser.shtml). The shortest dependency path between each pair of genes (or a gene and a vaccine) in a sentence was then extracted. The Support Vector Machine (SVM) (Joachims, 1999) with an edit distance-based kernel function among these dependency paths was used to classify if a path describes an interaction between a gene or a gene-vaccine pair. Only the pairs with positive confidence scores were used to build the fever-related interaction networks (Ozgur, et al., 2011). In the gene-vaccine network, a node represents a gene. In the gene-vaccine network, a node represents a gene or a vaccine. The nodes in the networks are connected by edges, which represent literature-derived interactions with positive SVM scores.

2.4 Centrality analysis of networks
Based on the generated interaction networks, four different types of centralities were calculated: degree, eigenvector, closeness, and betweenness. Each centrality measures a specific type of importance. In degree centrality, a node is considered important if it is connected to many other nodes in the network. In contrast to degree centrality, in eigenvector centrality each neighbor does not contribute equally to the centrality of a node. A node is considered more important if it is connected to many “central” nodes. In other words, besides the quantity of the connections of a node, their quality is also taken into account. In closeness centrality, a node is more important if its total distance to the other nodes in the network is smaller. In betweenness centrality the importance of a node is higher if it occurs on many shortest paths between other nodes. Each centrality measures a specific role of a node in a network.

2.5 Gene set enrichment analysis
A modified Fisher’s exact test is frequently used in high-throughput gene expression studies as a preferred method for identifying enriched biological functions among given gene sets. The Database for Annotation, Visualization and Integrated Discovery (DAVID) (Huang da, et al., 2009) was employed to identify over-represented biological functions and pathways between the genes in the fever- and vaccine-associated fever networks.

3 RESULTS

3.1 Study design
This study aimed to identify a generic fever gene interaction network and vaccine-related sub-networks (Fig. 1). All fever-related publications from PubMed were downloaded, and the sentences from titles and abstracts were parsed and pre-processed. Human gene names and VO terms were tagged by SciMiner and VO-SciMiner. A SVM-based method was used to identify possible interactions between gene pairs and gene-vaccine pairs. Centrality scores of each gene in the interaction networks were calculated and compared among the fever-network and vaccine-associated fever-network.
Identification of fever and vaccine-associated gene interaction networks using ontology-based literature mining

3.2 Fever-related literature-derived network

A PubMed query of fever-related literature resulted in a total of 179,156 articles (as of 12/31/2011). Two subsets of vaccine-related literature were defined with additional criteria: (1) including the terms “vaccine”, “vaccination”, and their variants (e.g., “vaccines”) (6,224 articles) or (2) including 186 specific vaccine names from VO (6,537 articles). SciMiner and VO-SciMiner were used to identify gene symbols and VO terms in the sentences of abstracts and titles of retrieved articles. Gene-gene interaction pairs at the sentence level only with a positive SVM score were used to generate three interaction networks, including the generic fever network (Fig. 2A and 2C), and two sub-networks associated with “vaccine” and its derived terms (Fig. 2B). The application of VO terms increased the numbers of retrieved papers, genes, and interactions (Fig. 2C).

3.3 Centrality analysis

All the genes in the generic fever network and vaccine/VO-specific fever sub-network were sorted based on centrality scores. The most central genes (the genes ranked among the top 10 by at least one of the centrality measures) are listed in Table 1. These genes are predicted to be associated with fever and potentially related to vaccine as well. These top genes can be grouped into three sub-sets:

1. **Genes ranked high in both networks.** Five genes (IL1B, TNF, IL6, IFNG, and CD8A) were ranked among top 10 in both networks by at least one centrality measure. These genes are well studied in both generic fever research and vaccine specific research.

2. **Genes ranked high in generic fever network but not in the vaccine/VO-specific sub-network.** This group includes nine genes (HSPA1A, NFKB1, IL8, IL2, MEFV, MAPK1, POMC, CD4, and IL10). These genes have not been well studied in the vaccine context. However, since these genes are associated with fever, it can be hypothesized that many of these genes are also important in vaccine-induced fever immune responses.

3. **Genes ranked high in vaccine/VO-specific sub-network but not in the generic fever network.** Seven genes (CSF2, IL7R, ERVWE1, APC, MC4R, IL1R1, and TLR2) were ranked among the top 10 only in the Vaccine/VO-specific sub-network. These genes might have been heavily studied in the vaccine domain but not in the general fever research field. It is reasonable to hypothesize that some of the genes are also critical to other fever-associated research domains.

### Table 1. Centrality score rankings of genes related to fever and vaccine networks

<table>
<thead>
<tr>
<th>Genes</th>
<th>Fever-network</th>
<th>Vaccine/VO-associated fever-network</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D</td>
<td>E</td>
</tr>
<tr>
<td>IL1B</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TNF</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>HSPA1A</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>IL6</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>IFNG</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>NFKB1</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>IL8</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>CD8A</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>IL2</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>MEFV</td>
<td>10</td>
<td>---</td>
</tr>
<tr>
<td>MAPK1</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>POMC</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>CD4</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>IL10</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>CSF2</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>IL7R</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>ERVWE1</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>APC</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>MC4R</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>IL1R1</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>TLR2</td>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

---

3.4 Functional enrichment analysis
DAVID identified 997 and 239 significantly over-represented functional terms (GO or KEGG) in the fever-network and VO-associated fever-network, respectively. The top 10 most significant terms are listed in Table 2. Although the majority of the functions are significant in both sets, the fever-network was more significantly enriched with processes related to cell death (apoptosis) and responses to wounding. Considering the roles of fever, such finding is reasonable. While “immune response” is commonly over-represented in both sets, leukocyte-related processes were the most significant in the Vaccine/VO-associated network, suggesting its important roles in vaccine-induced fever.

It is interesting that the most enriched processes in the fever-vaccine sub-network are associated with positive regulation of phosphorylation and phosphate metabolic processes. How the phosphorylation process is involved in the fever and vaccines domain deserves further investigation.

<table>
<thead>
<tr>
<th>Terms</th>
<th>Fever</th>
<th>Vaccine</th>
<th>VO</th>
</tr>
</thead>
<tbody>
<tr>
<td>regulation of cell death</td>
<td>45.3</td>
<td>3.9</td>
<td>3.6</td>
</tr>
<tr>
<td>regulation of apoptosis</td>
<td>45.4</td>
<td>3.9</td>
<td>3.6</td>
</tr>
<tr>
<td>regulation of programmed cell death</td>
<td>43.1</td>
<td>3.9</td>
<td>3.6</td>
</tr>
<tr>
<td>response to organic substance</td>
<td>37.8</td>
<td>3.4</td>
<td>3.9</td>
</tr>
<tr>
<td>response to wounding</td>
<td>30.7</td>
<td>2.5</td>
<td>2.3</td>
</tr>
<tr>
<td>extracellular space</td>
<td>34.8</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>positive regulation of multicellular organismal process</td>
<td>33.7</td>
<td>3.1</td>
<td>3.9</td>
</tr>
<tr>
<td>defense response</td>
<td>33.6</td>
<td>3.0</td>
<td>2.8</td>
</tr>
<tr>
<td>regulation of cell proliferation</td>
<td>31.6</td>
<td>3.2</td>
<td>2.9</td>
</tr>
<tr>
<td>immune response</td>
<td>30.7</td>
<td>7.8</td>
<td>7.2</td>
</tr>
<tr>
<td>cell activation</td>
<td>15.2</td>
<td>9.2</td>
<td>8.7</td>
</tr>
<tr>
<td>positive regulation of phosphorylation</td>
<td>16.6</td>
<td>7.2</td>
<td>7.0</td>
</tr>
<tr>
<td>positive regulation of phosphate metabolic process</td>
<td>16.3</td>
<td>7.2</td>
<td>6.9</td>
</tr>
<tr>
<td>positive regulation of phosphorus metabolic process</td>
<td>16.3</td>
<td>7.2</td>
<td>6.9</td>
</tr>
<tr>
<td>positive regulation of protein amino acid phosphorylation</td>
<td>15.3</td>
<td>7.3</td>
<td>7.0</td>
</tr>
<tr>
<td>leukocyte activation</td>
<td>14.3</td>
<td>7.3</td>
<td>6.9</td>
</tr>
<tr>
<td>leukocyte differentiation</td>
<td>9.7</td>
<td>8.9</td>
<td>8.5</td>
</tr>
<tr>
<td>hemopoiesis</td>
<td>9.6</td>
<td>7.3</td>
<td>6.9</td>
</tr>
<tr>
<td>lymphocyte differentiation</td>
<td>5.8</td>
<td>7.2</td>
<td>6.9</td>
</tr>
</tbody>
</table>

Table 2. Top 10 most significantly enriched biological functions for each gene-gene (i.e., ‘GG’) interaction network set. Values are \(-\log_{10}(\text{Benjamini-Hochberg corrected P-values})\) from DAVID results. Highlighted values correspond to top 10 lowest p-values in each set.

3.5 Gene-vaccine interaction networks

Gene-vaccine interactions with positive SVM scores were integrated into the fever networks. The whole PubMed contained 1,716 articles containing 2,835 positively-scored interactions between genes and vaccines (including VO vaccine terms). Among these articles, 32 were also related to fever, which contained 52 sentences with 44 unique gene-vaccine interactions. Specific vaccine names included Brucella vaccine RB51, Shigella flexneri vaccine SC602, Shigella sonnei strain WRSS1, and Shigella dysenteriae 1 strain WRSd1. Fig. 3 illustrates the combination of gene-gene interactions and gene-vaccine interactions retrieved from the fever-associated literature.

Different levels of gene interaction networks were identified. Twenty-six genes were found to directly interact with the general term ‘vaccine’. Three of them (including CD4, CD8A, and VIPR1) also directly interact with the four vaccines. These 26 genes can be considered as the first layer genes that interact with vaccines. The other 376 genes in the network are considered as the genes at the second layer (genes directly interacting with the 26 genes) or beyond. Although these genes have not been found to directly interact with vaccines through literature mining studies, they may indirectly interact with vaccines at different levels. Some of them may be more important than others in regulating vaccine-induced fever immune responses. It is interesting that no Toll-like receptors (TLRs) were found to directly interact with the vaccine in the fever-associated literature. Within the fever domain, many TLR receptors including TLR2, TLR4, TLR5, and TLR7 are found in the gene-gene interaction network. However, many TLRs are related with vaccines in the research domains not associated with fever. Therefore, it is reasonable to hypothesize that many vaccines directly or indirectly interact with these TLR receptors under the scope of fever domain.

4 DISCUSSION

Vaccine-induced fever may be a positive host response to the induction of protective immunity. However, vaccine-induced fever adverse event is annoying and can even result in severe outcomes. The detailed interaction network mechanisms among genes and vaccines leading to fever are not fully understood. This study is an attempt to identify potential gene interaction mechanisms that may contribute to vaccine-induced immune responses, including fever adverse
event. In this study, an ontology-based literature mining approach was applied to discovering gene-gene and gene-vaccine interaction networks associated with fever and fever-vaccine. Centrality and functional enrichment analyses were further employed to identify the most central genes and enriched biological functions in the networks. Our study demonstrates that ontology-based literature mining is able to efficiently discover gene-gene and gene-vaccine interactions in the fever domain.

Compared to our previous studies that analyzed the IFNG and IFNG-vaccine gene interaction networks (Ozgur, et al., 2010; Ozgur, et al., 2011), the current study focuses on analysis of fever and fever-vaccine gene interaction networks. Both studies demonstrated that the application of the VO enhanced the performance of mining vaccine-specific gene interaction networks. Therefore, the genome-wide, ontology-based literature mining approach can be applied to different domains, which can be defined by a gene name(s) (e.g., IFNG) or by a research domain (e.g., fever).

It is noted that not many gene-vaccine interactions were obtained through our literature mining analysis. This could be due to a few reasons. Our current approach uses sentence-level SVM-positively-scored interactions. The number of hits (gene-gene or gene-vaccine pairs) will increase if we employ a simple sentence-level co-citation rather than SVM-based approach. Even more hits would be identified with an abstract level co-citation for detecting possible gene-vaccine interactions. However, these methods may result in lower specificity of detecting true interactions. The same argument applies to the identification of gene-gene interactions from the literature.

Many scientific findings and hypotheses have been generated in our study. For example, our finding of the phosphorylation-focused regulation enriched in the fever-vaccine sub-network suggests the key role of the phosphorylation process in the vaccine-induced fever phenomenon. The generic fever network has an enriched gene set in cell death regulation, while leukocyte-associated genes are over-expressed in the fever-vaccine sub-network. This suggests that leukocyte cell death is critical to vaccine-induced immunity. It is also interesting that TLRs have been studied frequently in the vaccine domain, but none of them are directly associated with fever in the literature (at least in the sentence level). Since the TLRs are included in the generic fever gene interaction network, it can be hypothesized that TLRs may be potential key factors in vaccine-induced fever responses, including fever adverse events. Therefore, our ontology-based literature mining is able to advance scientific discovery by generating new findings and hypotheses. Future research efforts will be focused on expanding our initial networks by including more specific vaccines. VO is dynamically updated and growing with the support of VO community. Increased number of VO terms will substantially improve the sensitivity of identifying gene-vaccine interactions. The method described in this paper is generic and can be applied to study other gene interaction networks in different domains. The Ontology of Adverse Events (OAE; http://www.oae-ontology.org) is a community-based ontology in the domain of adverse events. The future integration of OAE with VO-based literature-mining would be of great interest to better understand the interactions between various types of vaccines and adverse effects.

ACKNOWLEDGEMENTS

This work has been supported by grant R01AI081062 from the US NIH National Institute of Allergy and Infectious Diseases and by a Marie Curie Career Integration Grant within the 7th European Community Framework Programme. The authors thank Ms. Rebecca Racz for her editorial review.

REFERENCES


Gardner, J. (2012) Is fever after infection part of the illness or the cure?, Emerg Nurse, 19, 20-25; quiz 27.


Hur, J., et al. (2011) Ontology-based Brucella vaccine literature indexing and systematic analysis of gene-vaccine association network, BMC Immunology, 12, 49.


Ontology representation and analysis of vaccine formulation and administration, and their effects on vaccine immune responses

Yu Lin, Yongqun He
Department of Microbiology and Immunology, Unit for Laboratory Animal Medicine, and Center for Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI 48109, USA

ABSTRACT

Vaccine formulation and administration methods may affect vaccine efficacy and safety significantly. In this report, the detailed classifications and definitions of vaccine components and vaccine administration processes are represented using OWL within the framework of Vaccine Ontology (VO). Different use cases are then studied to show how different vaccine formulations and routes of administering vaccines affect the protection efficacy, general immune responses, and adverse events following vaccination. Our results indicate that the two factors, vaccine formulation and administration route, can independently affect host response outcomes after vaccination. These two factors also interact with each other in their influence to the vaccine immune responses. The adverse events of live attenuated and nonliving licensed vaccines approved by the US FDA were analysed and compared. The live attenuated and nonliving vaccines are usually administered in different routes and have different local and systematic adverse effect manifestations.

1 INTRODUCTION

A vaccine is a processed material with the function that when administered, it prevents or ameliorates a disorder in a target organism by inducing or modifying adaptive immune responses specific to the antigens in the vaccine. In the field of infectious diseases, a vaccine stimulates an adaptive immune response to an infectious agent so that on subsequent exposure to this agent the infection is reduced or the disease does not occur. Vaccines include, for example, suspensions of killed or attenuated microorganisms, or products or derivatives of microorganisms. The most common method of administering vaccines is by injection, but some are given by mouth or nasal spray.

The community-based Vaccine Ontology (VO) categorizes various licensed vaccines, and vaccine candidates in clinical trial or under research (He, et al., 2009). VO logically represents all types of vaccines and vaccination approaches and how they interact with host immune systems. VO provides a platform for vaccine knowledge creation, sharing, and utilization, which further supports automated reasoning.

To study the efficacy and safety of a vaccine, knowledge of the formulation and administration route of a vaccine is important for investigating immune responses and adverse events following vaccination. Yet there is no published paper to address these topics in the perspective of ontology. In this report, we first represent various vaccine formulations and routes of vaccine administration in VO, and then discuss how different vaccine formulations and vaccination routes influence the outcomes of vaccination, including protective immune responses and adverse events.

2 METHODS

2.1 Vaccine Ontology (VO) development

The VO development follows the OBO Foundry principles, including openness, collaboration, and use of a common shared syntax (Smith, et al., 2007). VO is aligned with the Basic Formal Ontology (BFO) (http://www.ifomis.org/bfo) and the Relation Ontology (RO) (Smith, et al., 2005). The core VO terms include vaccine, vaccine component, route of administration, vaccination, immunization, vaccine protection assay, and vaccine-induced host response (Fig. 1).

Fig. 1. VO core terms asserted under the BFO framework.

The format of VO ontology is W3C standard Web Ontology Language (OWL2) (http://www.w3.org/TR/owl-guide/). For this study, many new terms and logical definitions were added into existing VO using the Protégé 4.1 OWL ontology editor (http://protege.stanford.edu/). VO project website is: http://www.violinet.org/vaccineontology.

2.2 Data sources for ontology development

Reference books are the main resource for classification and definition of vaccine formulation and administration (CDC, 2011; Plotkin, et al., 2008). Definitions of terms related in this paper were generated and reviewed by experts from the VO development team. A mail listing of VO developers was used to discuss and achieve agreement on logical definitions. The USA CDC pink book serves as a source for populating information related with U.S. licensed human vaccines. The vaccine ingredients used in U.S. licensed
Vaccines were manually extracted from an appendix table in the CDC pink book. The VIOLIN vaccine database, a web-based central resource for vaccine-related research data across various human pathogens (Xiang, et al., 2008), is another source for populating VO. Data used for use case analyses were manually extracted from: (i) PubMed papers and (ii) package inserts of licensed vaccines listed in US Food and Drug Administration (FDA) website (http://www.fda.gov/BiologicsBloodVaccines/Vaccines/ApprovedProducts/).

2.3 Use case studies

Use case studies were designed to address the following questions: how the differences of vaccine ingredients and/or administration routes influence vaccine efficacy or adverse event reactions? To simplify the story, the effects of other factors, such as patient health condition, vaccine dose, and vaccination frequency, are not analyzed in this study. VO is used for classification of vaccine types in the use cases.

3 RESULTS

3.1 Representation of vaccine components in VO

The final form of a vaccine is a mixture of different ingredients. The vaccine formulation refers to such a mixture of all vaccine components that forms the final vaccine to be administrated. A ‘vaccine component’ represents entities that exist in the formulation of a vaccine. The logical definition of ‘vaccine component’ in VO is:

- ‘processed material’ and (participates_in some ‘vaccine preparation’) and (part_of some ‘vaccine’)

Different ingredients of vaccines are classified into subclasses under the class ‘vaccine component’ (Fig. 2).

Fig. 2. Different vaccine components represented in VO.

Vaccine components related to this study are described below:

‘Pathogen organism vaccine component’: a vaccine component that is either a pathogen organism or derived from a pathogen organism. This term is logically defined as:

- ‘vaccine component’ and (organism or (‘is part of some organism)).

The pathogen organism used for vaccine development is an output of vaccine preparation process, which includes propagating the pathogen in a culture medium, isolating and purifying the organism. Live attenuated or killed virulent pathogens can both be used as a key component of a vaccine. A vaccine that utilizes a live attenuated pathogen as a component is defined as a ‘live attenuated vaccine’ in VO. Similarly, a vaccine that includes a killed or inactivated pathogen as a component is defined as a ‘killed or inactivated vaccine’. A vaccine that uses a protein, peptide, lipopolysaccharide, or polysaccharide of a pathogen as a component is defined as a ‘subunit vaccine’ in VO. A vaccine that uses a DNA or RNA sequence of a pathogen organism as a component is a DNA or RNA vaccine.

‘Vaccine antigen’: an immunogenic antigen in a vaccine. A vaccine antigen is typically a modified or partial form of the virus, bacteria or the toxin that causes a disease. Its logical definition is: ‘vaccine component’ and has role some ‘antigen role’. An example of a vaccine antigen is a protein from a pathogen organism as described above.

‘Vaccine conjugate protein’: is a protein that is chemically covalent to a vaccine antigen and is used to enhance vaccine-induced immune response. A vaccine that uses a conjugate protein called a conjugate vaccine. The definition for ‘vaccine conjugate protein’ in VO is ‘vaccine component’ and (is part_of some ‘conjugate vaccine’) and (bearer_of some ‘vaccine conjugate role’).

‘Vaccine adjuvant’: a vaccine additive incorporated into a vaccine to enhance the immunogenicity of vaccine antigens. In general, the adjuvant effect can be divided into two principal components: (1) antigen depot and delivery and (2) immune potentiation by targeting antigens to antigen presenting cells (APC) (Sayers, et al., 2012; Scholl, et al., 2005). The logical definition of this term in VO is: ‘vaccine additive’ and (‘has role’ some ‘vaccine adjuvant role’).

Other vaccine components are described similarly. Different vaccine components may share similar functions. For example, vaccine adjuvant and vaccine conjugate protein both help to induce stronger immune responses. A vaccine conjugate protein is chemically covalently attached to a vaccine antigen, whereas, a vaccine adjuvant is just mixed with vaccine antigen without any chemical conjugation. A vaccine conjugate protein, but usually not a vaccine adjuvant, may induce a strong immune response to itself. Vaccine vector and vaccine adjuvant can both carry vaccine antigen. A vaccine vector carries the genetic material that encodes for a protein antigen. However, a vaccine adjuvant delivery system carries the vaccine antigen per se.

The important differentiations among different vaccine components are given in the definition of roles. For example, the definition of ‘vaccine adjuvant role’ in VO is “an adjuvant role that inheres in a material entity which is added as a part of a vaccine and induces enhanced adaptive immune response to the vaccine antigen”. Some other roles including ‘vaccine additive role’, ‘vaccine vector role’,

2
‘vaccine stabilizer role’, ‘vaccine preservative role’, and ‘vaccine conjugate role’ are generated in VO. Bearers of these roles are corresponding vaccine components. Through the definition of roles, the usages and purposes of different vaccine components become clearly defined in VO.

3.2 Vaccine administration

The route of a vaccine administered into a vaccinee is another key factor in evaluating vaccine efficacy and safety. For example, if given through nose, a vaccine’s efficacy may be much weaker than that through injection. The term ‘vaccination’ was introduced in He’s previous work as a child term of ‘administering substance in vivo’ (Brinkman, et al., 2010). There are two ways to differentiate vaccination types: 1) based on frequency; and 2) based on entry location. Here we introduce detailed information about subtypes of vaccination based on the second categorization. Specifically, a vaccination can be classified as intramuscular, intraperitoneal, intragastric, intradermal, subcutaneous, percutaneous, intravenous, intravesical, oral, and intranasal vaccination (Fig. 3A). The ontological definition for this subtype of vaccination utilizes the relation ‘unfolds in’ and a route of administration. For example, an ‘intravenous vaccination’ is defined as ‘vaccination and unfolds in some intravenous route’. For this purpose, another term ‘route of administration’ has been defined in VO. We assert that ‘route of administration’ is a subclass of BFO:site (Fig. 3A). The term ‘route of administration’ is asserted with constrain: ‘is located in’ some ‘gross anatomical part’. Following this treatment, an oral route is logically defined as: ‘route of administration and is located in some mouth’.

![Fig. 3. VO classification of (A) vaccination types and (B) routes of administration.](image)

During a vaccine administration process, the site of vaccination often needs to be recorded (Fig. 2B). The term ‘vaccine injection site’ refers to a site on the skin where a needle pierced through, and it is part of an administration route. By this treatment, oral route and intranasal route have no vaccine injection site as a part.

3.3 Effects of formulation and administration to vaccine-induced immune responses

The immune responses induced by a vaccine may be affected by many factors, including the type of vaccine, vaccine dose, route of administration, and the use of an adjuvant. Host factors such as age, nutritional factors, genetics, and coexisting disease may also affect the response (Atkinson, et al., 2011). Here we show how different vaccine formulations and routes of administration will affect the immune responses in the host after vaccination.

3.3.1. Effects of formulation to vaccine-induced immune responses

Many examples can be used to demonstrate the important role of formulation to the vaccine-induced immune responses. For instance, killed or live attenuated vaccines turn to induce different immune responses. Compared to killed vaccines, live attenuated vaccines usually induce superior protection against subsequent challenge (Ficht, et al., 2009). However, they often present safety risks. Live and killed vaccines induce different immune responses in the host. Live attenuated vaccines tend to induce strong cell-mediated immunity. For example, Brucella abortus strain RB51 is a live attenuated vaccine against cattle brucellosis. It is licensed in the US Department of Agriculture (USDA) and many other countries. Live RB51 can induce both cell-mediated and humoral immunity against heterologous challenges. However, killed whole pathogen of RB51 fails to induce protective cell-mediated immunity (Jimenez de Bagues, et al., 1994). Another example is influenza vaccines. Currently there are both trivalent live attenuated or inactivated influenza vaccines available in the market. Although both live and inactivated influenza vaccines induce similar humoral responses, only live vaccines induce diverse T-cell responses in young children (Hoft, et al., 2011). Killed vaccines typically induce a Th2 immune response. To improve the performance of a killed vaccine, an effective vaccine adjuvant is often used.

The usage of conjugate vaccines is another example of the effect of formulation to vaccine-induced immune response. Polysaccharide antigens of many pathogenic bacteria induce only weak and short immune responses, especially in young children population, due to their immunological naivety and a degree of immunoincomptence condition (Finn, 2004). The response to a capsular polysaccharide is T-cell-independent, meaning that B lymphocytes proliferate and produce antibody without the help of T cells (Dintzis, 1992). Conjugate vaccines are formed by linking capsular polysaccharides to a protein carrier (a conjugate protein), such as tetanus toxoid, diphtheria toxoid, or Corynebacterium diphtheriae cross-reactive material (CRM197). Conjugate vaccines may induce T-cell-dependent immunologic responses that confer immune memory. In the USA, three meningococcal vaccines were proved by FDA: conjugate vaccine Menactra...
by Sanofi Pasteur, conjugate vaccine Menveo by Novartis Vaccines and Diagnostics, and polysaccharide vaccine Menomune -A/C/Y/W-135 by Sanofi Pasteur. All these vaccines use meningococcal polysaccharides from serogroups A,C,W-135 and Y as vaccine antigens. The conjugate vaccines Menactra and Menveo, but not polysaccharide vaccine Menomune, are able to induce high immunogenicity and immunologic memory especially in young children (Plotkin, et al., 2008).

3.3.2. Effect of vaccine administration route to protective immunity

The route of administration also plays a role in the efficacy of a vaccine. For example, RB51 is a well-studied live attenuated Brucella vaccine. Many studies investigated the relationships between RB51 protection efficacy and different factors including vaccine dose, administration route, and vaccination frequency. Although the intraperitoneal (IP) RB51 vaccination in mice induces protection against IP challenge of S2308, RB51 is unable to protect against intranasal (IN) challenge of virulent strain 2308 in the lung (Olsen, et al., 2007). On the other hand, the IN administration of RB51 is able to provide the protection against an intranasal challenge in the lung (Surendran, et al., 2012). The data extracted from these two experiments are modeled using individuals in VO (Fig. 4).

Using the information from the package inserts of 53 licensed US vaccines listed by the US FDA, an analysis of the effects of vaccine formulation and vaccination routes on the advance event occurrences was conducted. In total, 53 licensed vaccines are used in the US market. These 53 vaccines are classified into two groups: ‘live attenuated vaccine’ and ‘nonliving vaccine’. Applying the VO’s definition of ‘live vaccine’ and ‘live attenuated vaccine’, all the vaccines that are not live vaccine are considered nonliving vaccines, which includes inactivated vaccines, subunit vaccines, and conjugate vaccines. Local reaction is defined as any adverse event that occurs at the vaccination site after the vaccination. Systematic reactions are the adverse event unfolds in the whole body or any anatomical system in human body.

In total, 13 (24.5%) vaccine products are live attenuated vaccines, and their routes of administration are: subcutaneous injection (7, 53.8%), oral (3, 23%), intranasal spray (1, 7.7%), intranasal injection (2, 15.4%) and intravesical injection (1, 7.7%). Among 40 vaccines in the nonliving vaccine group, 38 (95%) are administrated by intramuscular injection, and 3 (7.5%) administrated by subcutaneous injection. The vaccine IPOL (Poliovirus Vaccine Inactivated manufactured by Sanofi Pasteur) is administrated either intramuscularly or subcutaneously. No local reactions were reported after oral administration. The most common local adverse events following vaccine injection are pain, induration and tenderness at the injection site. Intranasal administration may result in nose symptoms. Systematic adverse events were reported in the package inserts of all vaccines.

This result indicates that formulation of vaccine and route of administration interactively affect immune response. The vaccine formulation may decide the route of administration. Since live vaccines have the ability to invade into cell and replicate, they can be administrated using subcutaneous, oral and intranasal. Most of nonliving vaccines are delivered by intramuscular injection. Most nonliving vaccines need adjuvant or effective carrier component in their formulation. Since oral administration does not relate with local reaction, the local adverse reactions after vaccination are often linked to the use of needle for vaccine injection.

4 DISCUSSION

The vaccine formulations and routes of vaccination have been represented using OWL in VO and described in this paper. The presented use cases support our hypothesis that there are combinatorial effects of formulation and administration route on the outcomes of vaccination, such as protective immune responses and/or vaccine adverse events. This study semantically models the knowledge of this study. Such modeling and analysis can be used as a starting point to systematically analyze various vaccine data and identify new knowledge.

Fig. 4. VO modeling of successful protection against IP challenge of virulent S2308 by RB51 IN vaccination but not by RB51 IP vaccination. In short, RB51 induced immune response with/without protection against S2308 in lung is preceded by the RB51 vaccination process, which unfolds in the intraperitoneal route or intranasal route. RB51 is the specific input of the RB51 vaccination and RB51 induced immune response against S2308, in which S2308 is a participant. Square boxes represent class level terms. Boxes with round corners indicate instances collected from related papers.

It was found in another study that oral vaccination of RB51 was able to protect against IP challenge of S2308; however the protection is lower and less effective than that obtained from RB51 IP vaccination route (Stevens, et al., 1996). This result further demonstrates the effect of vaccination route to the vaccine protection efficacy.

3.3.3. Combinatorial effects of vaccine formulation and administration routes to vaccine adverse event profiles

The vaccine formulations and routes of vaccination have been represented using OWL in VO and described in this paper. The presented use cases support our hypothesis that there are combinatorial effects of formulation and administration route on the outcomes of vaccination, such as protective immune responses and/or vaccine adverse events. This study semantically models the knowledge of this study. Such modeling and analysis can be used as a starting point to systematically analyze various vaccine data and identify new knowledge.
Live attenuated vaccines can induce stronger protection, but it is considered risky. Killed vaccines or vaccines using part of pathogen organism are considered safer but need more additives, such as vaccine adjuvant. In the formulation of vaccines, less additives and residuals are more beneficial for the vaccines and strongly recommended. For example, using mutant non-toxic CRM197 (as in Menveo) instead of toxic diphtheria toxoid (as in Menactra) as conjugate protein will reduce the use of formaldehyde or other detoxifiers (Cooper, et al., 2011).

How risky is it to use a live attenuated vaccine instead of a nonliving vaccine? The case study in section 3.3.3 shows that an orally administrated live attenuated vaccine induces stronger immune response and less local adverse events. However, the vaccine formulation is not the only determinant for vaccine-induced immune responses. As demonstrated in our study of the 53 licensed vaccines in the US, the live and nonliving vaccines are administrated via different routes. Therefore, no example can be used to compare vaccine-induced adverse event or immune response if the difference of only one variable - the vaccine formulation is considered. One exception is live vs. nonliving Japanese encephalitis vaccines. Both vaccines contain the same vaccine antigen and administrated in the same route. However, they are licensed and used in different countries, thus cannot be easily analyzed. To better identify the differential immune responses by live vs. nonliving licensed Japanese encephalitis vaccines, a systematic analysis of protective immune responses and adverse events on a global scale is needed.

It is relatively straightforward to link local vaccine-induced adverse events (e.g., redness and local pain) with the vaccine and the administration route. However, it is often difficult to assert that systematic adverse events are induced by the administration of a vaccine. Asserting the causal relations between systematic adverse events and vaccine administrations is rather another topic, which is out of the scope of this paper and shall be addressed by the Ontology of Adverse Event (OAE) (He, et al., 2011). Finally, a system bioinformatics approach supported by integrating ontologies analyzing gene level data, literature mining data and adverse event case reporting data will more likely provide answers for these important questions.

ACKNOWLEDGEMENTS

This project was supported by a NIH-NIAID grant (R01AI081062).

REFERENCES


Smith, B., et al. (2005) Relations in biomedical ontologies, Genome Biology, 6, R46.


A lightweight Ontology Design Pattern to curate and represent experimental variables from vaccine protection studies.

Gully A.P.C. Burns¹* and Drashti Dave¹

¹ Biomedical Knowledge Engineering Group, Intelligent Systems Division, USC Information Sciences Institute.

ABSTRACT

The ‘Ontology of Experimental Variables and Values’ (OoEVV) is based on the principle of ontology design patterns. The main purpose behind OoEVV is to facilitate the interlinking of multiple variables that measure the same underlying quality but are expressed using different measurement scales. It is designed to enable biologists to generate preformatted Excel worksheets, to populate them with structured definitions of variables and measurement scales and then to generate Ontology Web Language (OWL) files directly from these worksheets. Users may also include identifiers for terms from existing ontologies as part of their definitions so that the OoEVV system will use online lookup services from the National Center of Biomedical Ontology (NCBO) to incorporate imported terminology within the OoEVV representation. We curated terminology definitions for experimental variables and measurement scales from eight published vaccine protection studies that were originally curated into the to Los Alamos National Laboratories HIV Vaccine Trials Database to yield 72 variable and 14 measurement scale definitions, which we then added to NCBO’s public ontology repository.

INTRODUCTION

As processes, both experimental protocols and computational workflows involve a form of information propagation as parameters are set, material entities are manipulated, data is gathered and then analyzed. Tracing which constants and parameters influence which measurements in experimental protocols provides a focused line of scientific inquiry (Russ et al 2011) and showcases the need for developing a standardized framework that represents scientific variables.

Ontologies such as the Experimental Factors Ontology (EFO, Malone et al. 2010) and Ontology of Biomedical Investigations (OBI, Brinkman et al. 2010) are in the process of developing community-driven formal representations for data elements. The National Institute of Neurological Diseases and Stroke (NINDS) have developed an extensive catalogue of ‘Common Data Elements’ in a top-down effort to develop standard elements for data (Saver et al. 2012). These efforts represent steps towards vocabulary standardization, but they do not address the issues of representing variables on a variety of scales with sufficient generality. Neither OBI, EFO or the NINDS CDE projects offer the ability to represent multiple variables using different scales to measure the same underlying quality (in order to support inference about equivalence of quantities across scales). Thus, we are developing an efficient, practical method of consolidating of equivalent data between measurement scales as a long-term goal. In this paper we describe the design of an ontology design pattern (‘ODP’, Gangemi & Presutti 2009) called ‘Ontology of Experimental Variables’ (OoEVV), meant to provide a lightweight methodology for capturing definitions of variables in terms of a variety of different measurement scales. We present OoEVV as complementary technology to both OBI and EFO to address a representational question that is still under discussion within the community. Hence, we have (a) created OoEVV as a relatively simple representation that may be easily adapted in multiple contexts, (b) provided practical curation tools that may be used by non-specialists and (c) curated a carefully-defined sample set directly from primary research literature as a preliminary demonstration of our approach. An intrinsic feature of our implementation is the ability to link concepts in OoEVV to other terms accessed from the National Center for Biomedical Ontology’s (NCBO) Bioprotal (http://bioportal.bioontology.org/ontologies/3006).

A priority of our approach is to provide a flexible methodology for modeling experiments that is usable by biologists. The Los Alamos National Laboratories (LANL) HIV Vaccine Trials Database presents a comprehensive index of research outcomes derived from vaccine studies in nonhuman primates. Each entry in the database is a summary based on the main parameters of a study. Our objective is to represent the research summaries as logical representations of dependencies between the variables of each study (see Russ et al. 2011 for an overview of this approach). The variability between experimental designs in vaccine protection studies is a challenge when designing a general-purpose relational database for this domain. Here, we use our curation framework to define variables pertaining to vaccine protection studies by surveying a small set of papers from the HIV Vaccine Trials Database.

MATERIALS AND METHODS

2.1 - Basic Curation Process

We selected eight studies from the HIV vaccine trials database and made detailed representations each study using our model editing tool (see http://www.bioscholar.org/). This system allows a scientist to draw a diagram of an experimental protocol based on a base set of elements for entities and processes: (a) entities (material and information), (b) processes (including material processing, assays, data transformation), (c) control points (for branching and fork-
ing). This approach follows OBI’s representation of protocol elements. In addition, the tool allows users to link variables to processes and entities within the model. We built complete, fully populated models of the experimental protocols of Belyakov et al. (2001), Cafaro et al. (2001), Muthumani et al. (2003), Buge et al. (2003), Rosati et al. (2005), Gomez-Roman et al. (2006), Lun et al. (2004) and Fuller et al. (2002). We extracted representations of parameters and variables from the editor to generate OoEVV-formatted spreadsheets, which we then edited manually.

2.2 - The OoEVV System

Figure 1: High-level meta-model of the OoEVV system.

2.2.1 Metamodel design - The basic UML design of the OoEVV meta-model is shown in Figure 1. Each ExperimentalVariableSet designates sets of ExperimentalVariable elements. Each ExperimentalVariable is linked by a measures attribute to a Term (denoting the underlying quality being measured) and a MeasurementScale (denoting the mathematical properties of its values). Term objects are used as a design pattern for ontology terms accessible via NCBO’s BioPortal interface. MeasurementValue are the values of ExperimentalVariable objects. Both MeasurementScale and MeasurementValue classes have several subtypes: Binary (true/false values); Numeric (quantitative numerical values), Hierarchical (imported values from other ontologies), Nominal (named entities), Ordinal (ranked values) and Relative (anonymous values defined by relations to named entities). This technical design is documented at http://www.isi.edu/projects/ooevv/overview.

Figure 2: OoEVV ontology creation process.

2.2.2 Data processing – Figure 2 shows the data processing pipeline used to generate the OoEVV ontology. Each ExperimentalVariableSet is curated into a formatted Excel spreadsheet that may then be uploaded into a local OWL file. The conversion process consists of mapping UML classes that include a term attribute linking to a Term class into OWL classes. We reproduce inheritance hierarchies between classes and map UML roles to OWL domain/range relationships. Spreadsheet data are instantiated as OWL individuals.

2.3 - Populating OoEVV From Expert Curated Spreadsheets

We have built a framework for enabling the conversion of expert curated terms into instances in an ontology design pattern like OoEVV. The framework, known as View-Primitive Data Modeling framework (VPDMf, Burns et. al 2003), relies on the meta-model specification as input expressed in the Unified Modeling Language (UML). We extended this UML-based system to support the generation of OWL-based models from the underlying data stored in our VPDMf-based system. This UML-to-OWL conversion leverages industry-standard development tools but has the does not directly support OWL-based description-logic reasoning (which is not needed for lightweight ODP development). We generate source code (in Java and Flex) from UML to support the construction of curation tools and convert to OWL to link our work to the broader community.

3 RESULTS

Table 1: Measurement variables indexed by paper

<table>
<thead>
<tr>
<th>Percent Specific lysis</th>
<th>CTL response</th>
<th>T cell count</th>
<th>Qualitative T cell count</th>
<th>Antibody titers</th>
<th>Antibody optical density</th>
<th>Antigenemia concentration</th>
<th>Cytokine concentration</th>
<th>Virus/Antigen titers</th>
<th>Antibody response</th>
<th>Antigenemia presence</th>
<th>Cytokine response</th>
<th>Virus/Antigen presence</th>
<th>Proviral DNA copies</th>
<th>Proviral DNA presence</th>
<th>Viral RNA copies</th>
<th>Viral load</th>
<th>Mean viral RNA copies</th>
<th>Mean viral load</th>
<th>ELISPOT number</th>
<th>ELISPOT response</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

 http://www.isi.edu/projects/ooevv/overview
Table 1 lists the eight papers based only on measurement variables used in each study. We found that the eight studies consistently used small sets of parameters that we grouped together as the combined input for assays. For example, the input parameters of a Cytotoxic T-Lymphocyte (CTL) assay include ‘epitope’ and ‘co-stimulant’. We defined ‘composite variables’ to represent this grouping (e.g., ‘CTL-parameters’) but will incorporate definitions of assays into future designs of OoEV for a more precise classification. Even within this small sample, there are a large number of different types of measurements with many studies using more specialized metrics. However, we also see several instances of standard metrics that were consistently used (i.e., T-cell count and viral load). Also present are examples where the same underlying quality is measured using different scales. For example, the measurement for neutralizing antibodies is encoded both by a binary scale (‘neutralizing index presence’ in Cafaro et al. 2001) and a numeric scale (‘neutralizing antibody concentration’ in Muthumani et al. 2003).

4 DISCUSSION

Our research is intended to support meta-analysis across vaccine protection studies. An important contribution of this work is the explicit treatment of variables that measure the same underlying quality with different scales. An example of the significance of such a representation is within the text of a typical results section, where the authors may use ordinal terminology (‘strong’, ‘weak’, etc.) to simplify numeric data, which may be presented as a graph. By representing both scales, a curator can choose how to best capture these findings to optimize either curation speed or the precision of the underlying data.

There are other ontologies that represent experimental variables. The Ontology of Biomedical Investigation (OBI) provides a complex representation of experimental design (Brinkman et al. 2010), including statistical tests used in vaccine protection studies (He 2010). The Experimental Factor Ontology (EFO) focuses on experimental factors relevant to gene-expression (Malone et al. 2010). OBI uses different representations for data in different contexts, based on discussions within a small dedicated group of specialized curators. We advocate another approach: using a simple design pattern with accompanying tools, we emphasize efficient curation and sharing with minimal ontological commitment, whilst using ontological standards to share identifiers and terminology wherever possible. We anticipate that this will improve coverage and usability of terms describing variables, scales and their accompanying data.

ACKNOWLEDGEMENTS

This research is funded by NSF (#0849977) and by NIH (GM083871, MH079068 and RR025736). We thank Brian Foley and Bob Murname for their support. We gratefully acknowledge the work of Jessica Turner, Cartic Ramakrishnan, Marcelo Tallis, Karthik Narasandra Manjunatha and Swati Raina.

REFERENCES


Profiling Structured Product Labeling with NDF-RT and RxNorm

Qian Zhu1, * Guoqian Jiang1 and Christopher G Chute1

1Department of Health Sciences Research, Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN, USA

ABSTRACT

The Structured Product Labeling (SPL) is a document markup standard approved by Health Level Seven (HL7) and adopted by United States Food and Drug Administration (FDA) as a mechanism for exchanging drug product information. The SPL includes rich information about FDA approved clinical drugs, however, lack of linkage to standard drug ontologies hinders its meaningful use. In this paper, we present a framework that intends to map SPL drug labels with existing drug ontologies, NDF-RT (National Drug File Reference Terminology) and RxNorm. We also use existing categorical annotations to classify SPL drug labels into corresponding classes. We established the classification and relevant linkage for SPL drug labels using the following 3 approaches. First, we retrieved NDF-RT categorical information from the External Pharmacologic Class (EPC) indexing SPLs. Second, we used the RxNorm and NDF-RT mappings to classify and link SPLs with NDF-RT categories. Third, we profiled SPLs using RxNorm term type information. In the implementation process, we employed a Semantic Web technology framework, in which we stored the data sets from RxNorm, NDF-RT and SPLs into a RDF triple store, and executed SPARQL queries to retrieve data from customized SPARQL endpoints.

1 INTRODUCTION

Structured Product Labeling (SPL) encodes rich clinical drug knowledge, however, it is not easy to integrate the SPL labels with other data sources due to proprietary formats, and different conceptual models. This is a common scenario occurring in the biomedical domain, where dozens of public resources involve laborious processes to manually annotate data. This is mostly because they are using heterogeneous code systems to represent their data. Hence, data normalization and building all possible linkages among these data sets will make data interoperation and integration feasible.

The objective of the present study is to map SPL drug labels using two major standard drug ontologies: the Veterans Administration’s (VA) NDF-RT1 and the National Library of Medicine’s (NLM) RxNorm2. Our investigation was guided by answering the following research questions: (1) how SPL drug labels are covered and connected by RxNorm and NDF-RT; (2) how to utilize RxNorm/NDF-RT drug resources to map SPL drug labels from the drug class perspective; (3) how to leverage Semantic Web technology to accomplish the implementation task.

The paper is organized into the following sections. First, we introduce background information for SPL, NDF-RT, RxNorm and Semantic Web technology in the Background section; Second, in the Methods section, we introduce three main parallel approaches on SPL drug label profiling; Third, we illustrate our results generated from each step in the Results section, and then followed by Discussion and Conclusion.

2 BACKGROUND

2.1 Structured Product Labeling (SPL)

The Structured Product Labeling (SPL)3 is a document markup standard approved by Health Level Seven (HL7)4 and adopted by FDA as a mechanism for exchanging product information. SPL defines the human readable label documents that contain structured content of labeling (all text, tables and figures) for a product, along with additional machine readable information (i.e., drug listing data elements including information about the product and the packaging). SPL labels used in this study were extracted from NLM DailyMed website5.

2.2 National Drug File Reference Terminology (NDF-RT)

NDF-RT1 is used for modeling drug characteristics including ingredients, chemical structure, dose form, physiologic effect, mechanism of action, pharmacokinetics, and related diseases.

In support of SPL initiative, a non-hierarchical collection of External Pharmacologic Class (EPC) concepts has been added to NDF-RT in parallel and analogous with the VA Drug Classification hierarchy. These concepts are distinguished by an “[EPC]” tag suffixed to their preferred names. Role relationships originating from these EPC concepts target concepts from the NDF-RT MoA, PE, and CI hierarchies that are selected by the FDA to index their EPC for SPL purposes6.

1 To whom correspondence should be addressed: zhu.qian@mayo.edu
4 HL7: http://www.hl7.org/
2.3 RxNorm

RxNorm\(^7\) provides normalized names for clinical drugs and links its names to many of the drug vocabularies commonly used. Also it makes links to a number of vocabularies, such as SPL, NDF-RT, MeSH, and etc. The “SAB” code is defined by RxNorm to differentiate the different sources aggregated into RxNorm. For example, “SAB=MTHSPL” indicates that the corresponding concept is absorbed from SPL and “SAB=NFDR” indicating the source from NDF-RT. These two sources were used in this study. RxNorm defines term type “TTY” to indicate the role an atom plays in its source. The term types are assigned based on source documentation or NLM understanding of the source. Table 1 shows a list of term types “TTYS” used in this study with their names and descriptions.

<table>
<thead>
<tr>
<th>TTY</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SBD</td>
<td>Semantic Branded Drug</td>
<td>Ingredient + Strength + Dose Form + Brand Name</td>
</tr>
<tr>
<td>SCD</td>
<td>Semantic Clinical Drug</td>
<td>Ingredient + Strength + Dose Form</td>
</tr>
<tr>
<td>IN</td>
<td>Ingredients</td>
<td>A compound or moiety that gives the drug its distinctive clinical properties.</td>
</tr>
<tr>
<td>PIN</td>
<td>Precise Ingredient</td>
<td>A specified form of the ingredient that may or may not be clinically active.</td>
</tr>
<tr>
<td>BPCK</td>
<td>Brand Name Pack</td>
<td>(#(Ingredient Strength Dose Form) / #(Ingredient Strength Dose Form)) Pack</td>
</tr>
<tr>
<td>GPCK</td>
<td>Generic Pack</td>
<td>[(#(Ingredient + Strength + Dose Form) / #(Ingredient + Strength + Dose Form))] Pack</td>
</tr>
<tr>
<td>BN</td>
<td>Brand Name</td>
<td>A proprietary name for a family of products containing a specific active ingredient.</td>
</tr>
<tr>
<td>MIN</td>
<td>Multiple Ingredients</td>
<td>Two or more ingredients appearing together in a single drug preparation, created from SCDF.</td>
</tr>
<tr>
<td>SY</td>
<td>Synonym</td>
<td>Synonym of another TTY, given for clarity.</td>
</tr>
<tr>
<td>TMSY</td>
<td>Tall Man Lettering Synonym</td>
<td>Tall Man Lettering synonym of another TTY, given to distinguish between commonly confused drugs.</td>
</tr>
</tbody>
</table>

Table 1. A list of RxNorm term types “TTYS” with names and descriptions (Source from RxNorm Documentation\(^7\))

In this study, we used the following two files that were downloaded from RxNorm website in April 9, 2012. 1) RXNCONSO.RRF. The file includes all connections with source vocabularies. 2) RXNSAT.RRF. The file includes all source vocabulary attributes that do not fit into other categories to search for the information about drug categories and connections among RxNorm, NDF-RT and SPL.

2.4 Semantic Web Technology

The Resource Description Framework (RDF)\(^8\), a W3C recommendation, is a directed, labeled graph data format for representing information in the Web. SPARQL is a query language for RDF graphs (Clement et al. 2009). Triple store is a database for the storage and retrieval of RDF metadata, ideally through standard SPARQL query language. The Web Ontology Language (OWL) is a standard ontology language for the Semantic Web\(^9\). Semantic Web technology supports flexible, extensible and evolvable knowledge transfer and reuse. It has been widely used in biomedical domains to formalize and model medical and biological systems (Neumann et al. 2004; O’Connor et al. 2007b; David et al. 2011). In the present study, we adopted it as the core technology in our implementation step. The SPL data used in this study is stored in a RDF triple store and by executing SPARQL queries to retrieve the desirable information.

3 METHODS

3.1 System Architecture

There are four primary modules in the system, comprising of 1) data transformation module; 2) data persistence module; 3) SPL profiling module, in which SPL drug labels are profiled by EPC, NDF-RT and RxNorm; 4) standardized drug/drug class network module. Figure 1 shows system architecture of the four modules.

![Fig.1. A diagram illustrating our system architecture. For the data transformation module, data reformatting steps were performed for SPL, NDF-RT and RxNorm individually prior to loading the data into a RDF triple store and a MySQL database since they have different data formats - original SPLs in XML format, NDF-RT available in OWL (Antoniou et al. 2004), and RxNorm in the UMLS Rich Representation Format (RRF). A XML2RDF sub module\(^10\) takes the input rendered in the XML format, and outputs the result in the RDF format through a transparent transformation service. NDF-RT in OWL was loaded into RDF store directly. Although the RxNorm data is available in RRF files, RxNorm provides a MySQL script to help load the data into a MySQL database. In this study, we queried against a MySQL database for RxNorm data. For the persistence module, we implemented an open source RDF store “4Store” that is developed by Garlik\(^11\) to host the SPL and NDF-RT data. After loading RDF triples into the RDF store, we implemented a SPARQL endpoint that

\(^7\) RxNorm Documentation: http://www.nlm.nih.gov/research/umls/rxnorm/overview.html#

\(^8\) OWL: http://www.w3.org/TR/owf-features/

\(^9\) XML2RDF: rhizomik.net/redefer/xml2rdf/

\(^10\) 4Store: http://4store.org/
provides standard SPARQL query service against the RDF store. For the drug/drug class network module, we incorporated SPL profiling results with our previous standardized drug work [Zhu et al. 2012]. We will integrate more drug / drug class resources, like PharmGKB12, DrugBank13, or National Drug Code (NDC)14 into this network, which will be visualized by Cytoscape15, a general platform for complex network analysis and visualization.

### 3.2 Profiling by EPC Classes

The EPC indexing SPLs were stored in a RDF triple store. We executed SPARQL query against the triple store to extract the setIds (unique identifier of the SPL), NUI (unique identifier of the NDF-RT), and the relevant role relationships describing and defining concepts according to their relationships with other concepts, mapped by NDF-RT from each SPL. Figure 2 shows the SPARQL query used to extract the EPC indexing SPL and EPC classes with related NDF-RT concepts associated with their role relationships.

```
PREFIX epc: <urn:hl7:org:v3#>
SELECT ?setId ?nui ?displayName 
{ GRAPH <http://fda.gov/spl/epc> 
  ?xml epc:setId ?setIdSection . 
  ?setIdSection epc:root ?setId . 
  ?subject epc:identifiedSubstance ?sub . 
  ?sub epc:identifiedSubstance ?sub1 . 
  ?sub1 epc:asSpecializedKind ?kind . 
  ?kind1 epc:code ?code . 
  ?code epc:displayName "displayName" . 
  FILTER (regex("displayName", "EPC", "i")) || 
  regex("displayName", "PE", "i") || 
  regex("displayName", "Chemical", "i") || 
  regex("displayName", "tty", "i") 
  FILTER (regex("nui", "i", "i") 
  )}
```

Fig.2. SPARQL query to extract EPC classes and NDF-RT concepts with their role relationships

The NDF-RT concepts whose names replicate each EPC string are mapped via role relationships to appropriate MoA, PE, or Chemical/Ingredient concepts in NDF-RT. An EPC indexing SPL label is possibly corresponding to multiple NDF-RT EPC classes, and for each EPC class, multiple role relationships might be mapped to.

### 3.3 Profiling by RxNorm and NDF-RT mapping

RxNorm reflects and preserves the meanings, concept names, and relationships from different copyright holders. Hence, the objective of this step was to use existing annotations in RxNorm to categorize SPL drug labels, and to make connections between SPLs and RxNorm/NDF-RT. As RxNorm (RXNCONSO and RXNSAT) were pre-loaded in a MySQL database, we executed the SQL clauses to extract the data. We used two RxNorm-integrated resources, SPL and NDF-RT in this step.

For the source SPL (i.e. “SAB = MTHSPL”), we executed SQL queries to extract the SPL concepts along with “TTY” and RxCUI (RxNorm unique identifier) from RXNCONSO table. In order to connect RxNorm with SPL labels, we queried RXNSAT table to retrieve corresponding SPL setIds for each RxCUI.

For the source NDFRT, (i.e., “SAB = NDFRT”), SQL queries were executed to extract the NDF-RT concepts along with NUI and preferred name. Each preferred name includes additional “KIND” information from NDF-RT. For example, an entry with [RxCUI = “4278”] corresponds to the [NUI = “N0000006373”] with preferred name “Famotidine [Chemical/Ingredient]”. We grouped the entries into NDF-RT kind “Chemical/Ingredient”.

It is worthy to note that RxCUIs are available in “RCXUI” column from both RXNCONSO and RXNSAT tables. NUI is stored in “SCUI (Source asserted concept identifier)” column, while no setId available in SCUI column. Hence, to establish the linkage between SPL and RxNorm, we searched RXNSAT to retrieve a list of setIds for each given RxCUI.

### 4 RESULTS

Total number of SPLs from DailyMed is 36,568. 1,247 EPC indexing SPLs were downloaded as of April 12, 2012 and loaded into RDF triple store. Meanwhile, we have RxNorm snapshot downloaded at the same time. “RXNCONSO.RRF” with 965,968 concepts and “RXNSAT.RRF” with 6,221,513 entries were loaded into a MySQL database.

#### 4.1 Results from EPC Classes

Each EPC indexing SPL can be mapped to multiple NDF-RT concepts with different or same role relationships. For example, a EPC indexing SPL “MUMPS VIRUS STRAIN B LEVEL JERYL LYNN LIVE ANTIGEN” with setId “433534fd-7318-45ec-84b2e-9f35e7a7140e” has EPC class “Live Mumps Virus Vaccine”, which is mapped to three NDF-RT concepts by role relationships, “has_PE” with “Actively Acquired Immunity”, “has_Chemical_Structure” with “Vaccines, Attenuated”, and “has_Chemical_Structure” with “Mumps Vaccine”.

The mapping results are listed in Table 2. The coverage shown in brackets in the Table 2 is calculated by number of concepts from NDF-RT and SPL with each role relationship divided by the total number of concepts from these two resources. Total 497 EPC classes are found in the NDF-RT

---

12 PharmGKB: www.pharmgkb.org
13 DrugBank: www.drugbank.ca
14 NDC: http://www.fda.gov/drugs/informationondrugs/ucm142438.htm
15 Cytoscape: http://www.cytoscape.org
RDF repository, indicating that 71.2% (354 out of 497) EPC classes have been integrated into SPL.

<table>
<thead>
<tr>
<th>Category</th>
<th>Num. of unique NUIs</th>
<th>Num. of unique setIds</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPC</td>
<td>354 (0.8%)</td>
<td>853 (2.3%)</td>
</tr>
<tr>
<td>Chemical/Ingredient</td>
<td>154 (0.3%)</td>
<td>342 (0.9%)</td>
</tr>
<tr>
<td>PE</td>
<td>70 (0.2%)</td>
<td>201 (0.6%)</td>
</tr>
<tr>
<td>MoA</td>
<td>70 (0.01%)</td>
<td>10 (0.03%)</td>
</tr>
<tr>
<td>Total</td>
<td>585 (1.2%)</td>
<td>858 (2.4%)</td>
</tr>
</tbody>
</table>

Table 2. Mapping results by EPC classes (Category: NDF-RT categories extracted from EPC indexing SPL; NUI: identifier of NDF-RT; setId: identifier of SPL)

4.2 Results from RxNorm vs. NDF-RT mappings

We extracted 41,343 unique RxNorm entries from the source “NDFRT” and their corresponding NUIs and preferred names. We identified 6,611 unique NUIs that correspond to 35,094 unique SPL setIds retrieved from RXNSAT file, which includes the connections between RxCUI and setIds. In this step, we utilized the NDF-RT category information to profile the SPLs. Only three categories from NDF-RT (VA Product, Chemical/Ingredient, and EPC) have been included in RxNorm and NDF-RT mappings, hence, we extracted these three categories along with NUIs and setIds. The statistical results are listed in Table 3. Comparing with the 36,568 SPL labels from DailyMed and 47,075 concepts from NDF-RT, 96% SPLs have been covered by RxNorm and NDF-RT mappings, and only 14.0% NDF-RT concepts have been linked to the SPL labels.

<table>
<thead>
<tr>
<th>Category</th>
<th>Num. of unique NUIs</th>
<th>Num. of unique setIds</th>
</tr>
</thead>
<tbody>
<tr>
<td>VA Product</td>
<td>4,800 (10.4%)</td>
<td>20,937 (57.3%)</td>
</tr>
<tr>
<td>Chemical/Ingredient</td>
<td>1,730 (3.7%)</td>
<td>34,788 (95.1%)</td>
</tr>
<tr>
<td>EPC</td>
<td>1,000 (0.02%)</td>
<td>14 (0.04%)</td>
</tr>
<tr>
<td>Total</td>
<td>6,611 (14.0%)</td>
<td>35,094 (96.0%)</td>
</tr>
</tbody>
</table>

Table 3. Mapping results by RxNorm/NDF-RT mappings (Category: NDF-RT categories from NDF-RT/RxNorm mappings; NUI: identifier of NDF-RT; setId: identifier of SPL)

4.3 Results from RxNorm

We first identified 35,480 unique SPL entries from RxNorm using the source “SPL”. And then we identified 15,615 unique RxCUIs that correspond to the 35,480 unique SPL setIds by searching for RXNSAT table. We used the term types “TTYs” to classify SPL labels into ten categories. Each category with the number of unique RxCUIs and unique setIds and their coverage has been listed in the Table 4. Comparing with the 36,568 SPL labels from DailyMed and 965,968 concepts from RxNorm, 97.0% SPLs have been covered by RxNorm and NDF-RT mappings, and only 1.6% RxNorm codes have been linked to the SPL labels. It is worthy to note that SY and TMSY denote synonyms of another TTY, so there are overlaps between SY/TMSY and other TTYs.

<table>
<thead>
<tr>
<th>TTY</th>
<th>Num. of unique RxCUIs</th>
<th>Num. of unique setIds</th>
</tr>
</thead>
<tbody>
<tr>
<td>SBD</td>
<td>6714 (0.70%)</td>
<td>7087 (19.38%)</td>
</tr>
<tr>
<td>SCD</td>
<td>5981 (0.6%)</td>
<td>21127 (21.9%)</td>
</tr>
<tr>
<td>IN</td>
<td>1,834 (0.2%)</td>
<td>34,038 (93.1%)</td>
</tr>
<tr>
<td>PIN</td>
<td>773 (0.08%)</td>
<td>18,498 (50.6%)</td>
</tr>
<tr>
<td>BPCK</td>
<td>261 (0.03%)</td>
<td>259 (0.7%)</td>
</tr>
<tr>
<td>GPCK</td>
<td>48 (0.005%)</td>
<td>150 (0.04%)</td>
</tr>
<tr>
<td>BN</td>
<td>20,000 (0.02%)</td>
<td>5,001 (0.01%)</td>
</tr>
<tr>
<td>MIN</td>
<td>20,000 (0.02%)</td>
<td>3,000 (0.008%)</td>
</tr>
<tr>
<td>SY*</td>
<td>11,489 (1.2%)</td>
<td>21,438 (58.6%)</td>
</tr>
<tr>
<td>TMSY*</td>
<td>1,810 (0.2%)</td>
<td>4,208 (11.5%)</td>
</tr>
<tr>
<td>Total</td>
<td>15,615 (1.6%)</td>
<td>35,480 (97.0%)</td>
</tr>
</tbody>
</table>

Table 4. Mapping results from RxNorm (TTY denotes term types from RxNorm; RxCUI denotes identifier of RxNorm; setId denotes identifier of SPL. * denotes the concepts with SY and TMSY have overlaps with the concepts in other TTYs)

4.4 Network Visualization

The SPL profiles using RxNorm and NDF-RT not only demonstrate the connections among these three resources, but also help to establish a drug/drug class network among them. Within this network, the nodes are consisted of concepts from SPLs, RxNorm or NDF-RT as target or source nodes; the edges are consisted of the category information retrieved in the above steps. We are exploring to utilize the Cytoscape15 as a visualization tool to visualize the network.

5 DISCUSSION

Since SPL labels contain a large portion of clinical drugs, and chemical/ingredients, along with other possible drug categories, they have been demonstrated as a very useful drug knowledge resource, especially in clinical drug applications such as Adverse Drug Events (ADEs) detection from electronic medical records (EMRs). Actually a number of studies are emerging recently to semantically annotate ADEs information using the SPL labels for the purpose of drug safety surveillance. For example, in a project called SIDER, a public, computer-readable side effect resource that connects 888 drugs to 1450 side effect terms was developed using the SPL labels (Kuhn et al. 2010). In a system called ADESSA, for another example, the ADEs were extracted from the SPL labels and mapped to the MedDRA terms and concepts, then utilized the UMLS to generate mappings between the MedDRA terms and the SNOMED CT concepts (Duke et al. 2010). In a project at Mayo Clinic, the SPL labels were used in a framework for building a standardized ADE knowledge base known as ADEpedia16 through combining ontology-based approaches with Semantic Web technology (Jiang et al. 2011). In addition, Schadow conducted some other studies to evaluate the impact of SPL for medication knowledge management.

15 ADEpedia: http://adepedia.org

16 ADEpedia: http://adepedia.org
(Schadow 2007). And Schadow also had successfully aligned SPL with associated terminologies to make drug-intolerance (allergy) decision support in computerized provider order entry (CPOE) systems in 2008 (Schadow 2008).

In this paper, we conducted the study with SPLs from drug and drug classification perspective. We have successfully mapped SPL labels with NDF-RT and RxNorm using NDF-RT drug/drug class information and clinical drug identification information from RxNorm. For the SPLs and NDF-RT mappings, 96.0% of SPL drug labels are linked with NDF-RT categories; whereas for the SPLs and RxNorm mappings, 97.0% of SPL drug labels are linked to RxNorm codes.

The SPL mappings reported in this study reveal the SPL coverage by NDF-RT and RxNorm, which can potentially facilitate the future integration among SPL, NDF-RT and RxNorm for normalizing drug information as much as possible. In addition, the SPL mappings also enable the integration of SPL drug labels into a drug and drug class network developed in our previous study (Zhu et al. 2012). Semantic Web technology plays a key role in the implementation of our system. We represented the drug data in RDF triples and hosted the RDF triples in a RDF triple store to make the data integration, data management more feasible. In addition, running SPARQL queries against RDF store through SPARQL endpoint simplified our efforts on linking SPL drug labels to NDF-RT by using the predicates defined in the RDF triples. In the future, we will employ D2R (Bizer et al. 2003) server for the RxNorm RDF transformation through describing mappings between the RDF-based relational database schema and a RDF data model.

6 CONCLUSION

In this study, we have successfully mapped SPL drug labels with RxNorm and NDF-RT. In total, 99.4% SPL drug labels have been linked with two drug ontologies: RxNorm and NDF-RT. The coverage for individual RxNorm and NDF-RT is 97.0% and 96.0%. Since NDF-RT provides rich drug/drug class information, we were able to map SPLs to NDF-RT from drug class perspective. Meanwhile, we utilized the clinical meaningful drug term types from RxNorm to profile SPLs.

We will continue the following investigations in the future. 1) Since existing SPL drug labels have been classified into a number of categories, including over-the-counter (OTC), prescription, animal, and so on, we will explore to integrate the information into the current drug network. 2) We will build a backbone drug network based on NDF-RT and integrate the network with more drug resources, like DrugBank, NDC, PharmGKB, etc. 3) We will explore to build linkages between the drug/drug class and relevant phenotype/genotype.

ACKNOWLEDGEMENTS

This work was supported by the Pharmacogenomic Research Network (NIH/NIGMS - U19 GM61388) and the SHARP Area 4: Secondary Use of EHR Data (90TR000201).

REFERENCES


Extraction of Adverse Drug Effects from Medical Case Reports
Harsha Gurulingappa¹, Abdul-Mateen Rajput ², and Luca Toldo ²∗

¹Molecular Connections Pvt. Ltd., Basavanagudi, Bangalore 560004, India
²Merck KGaA, Frankfurterstraße 250, Darmstadt 64293, Germany

ABSTRACT
A sheer amount of information about adverse effects of drugs are published in medical case reports that pose major challenges for drug safety experts to perform timely monitoring. Efficient strategies for identification and extraction of information about adverse drug effects from free-text resources are needed to support pharmacovigilance research and decision making. Therefore, this work focuses on the adaptation of a machine learning-based relation extraction system for the identification and extraction of drug-related adverse effects from MEDLINE case reports. It relies on a high quality corpus that was manually annotated, using ontology-driven methodology. Qualitative evaluation of the system show robust results.

1 INTRODUCTION
Adverse effects of drugs is a bothersome issue that confronts drug manufacturers, healthcare providers, and regulatory authorities. Stringent measures for capturing the risks associated with drug usage are established in forms of spontaneous reporting systems that are timely analyzed to ensure safe use of drugs (Hauben and Bate, 2009). Amongst various data sources used by drug safety experts to perform the safety monitoring, case reports published in the scientific biomedical literature represent an important resource due to their abundant existence, rapid rate of generation, and valuable information enclosed (Vandenbroucke, 2001). Due to their unstructured nature, however, manual analysis of the scientific literature is challenging, cumbersome, and labor intensive.

In recent years, development of automatic natural language processing (NLP) and information extraction (IE) techniques have gained immense popularity. They include identification of biomedical named entities, relations between the entities, or events associated with them. Noticeable efforts have been invested on mining the adverse effects in different forms of free-text data. Examples include Wang et al., 2009 who applied the MedLEE system on discharge summaries to identify medication events and entities that could be potential adverse entities that were detected using the strength of statistical association based on their co-occurrences. Leaman et al., 2010 proposed a lenient NLP model for extracting adverse effects of drugs from social media such as blogs. Gurulingappa et al., 2011 developed a machine learning-based system for classifying the sentences in MEDLINE case reports that assert adverse effects of drugs. However, according to the author’s knowledge, there is a limited focus on identification of semantic relationships between drugs and adverse effects in text. This is partly due to the unavailability of suitable public corpora that could be used for technology development and benchmarking. Extracting relations between drugs and adverse effects can facilitate appropriate indexing, precise searching, visualization, and faster information tracing. The use of ontology of adverse drug reactions for automated signal generation in pharmacovigilance has already been proposed (Henegar et al., 2006) and its application to information retrieval has been exploited by the same group few years later, in the VIGITERMES project (Delamarre et al., 2010), where the OntoEIM adverse event ontology have been used to extend the dictionary of adverse event entities, normalize queries, and consolidate annotations, delivering 29% precision and 67% recall of MEDLINE abstracts. Automatic extraction of adverse drug effects from clinical records is an active area of research (Aramaki et al., 2010). Mining social internet message boards to identify adverse drug reactions has been reported (Benton et al., 2011), whereby in that work the extraction of event - drug pairs was determined only using co-occurrence of terms within a window of 20 tokens apart, and the use of machine learning systems was only focused on deidentification for privacy protection. This work reports on the adaptation of a machine learning-based system for identifying the relations between drugs and adverse effects in MEDLINE case reports, that relies on an ontology-driven manually annotated corpus, that strictly follows semantic annotation guidelines developed for clinical text (Roberts et al., 2009). The system has been qualitatively evaluated and studied for its ability of support real time pharmacovigilance studies.

2 METHODS
2.1 Corpus Preparation
The data set used for training and validation of the relation extraction system is the ADE corpus (Gurulingappa et al., 2012). The ADE corpus contains 2972 MEDLINE case reports that are manually annotated and harmonized by three annotators. The corpus contains annotations of 5063 drugs, 5776 conditions (e.g. diseases, signs, symptoms), and 6821 relations between drugs and conditions representing clear adverse effect implications. All annotations are confined to sentence level i.e. drugs and conditions representing adverse effects co-occurring only within individual sentences are annotated. Drugs and conditions that do not fall into adverse effect relations are not annotated. This was done in accordance to the annotation guidelines.

The ADE corpus contains annotations of relations between drugs and conditions that represent True relations. This represents a sparsely annotated dataset. For training a supervised classifier, it was essential to generate False relations i.e. drugs and conditions that do not fall into adverse effect relations. For this purpose, ProMiner, a dictionary-based named entity recognition system (Hanisch et al., 2005) was employed. ProMiner was incorporated with DrugBank (Knox et al., 2011) and MedDRA (Merrill, 2008) dictionaries for the identification of drugs and conditions respectively in the ADE corpus that were previously not annotated by human annotators. As a result of named entity recognition, new instances encompassing 2269
drugs and 3437 conditions were automatically annotated. Drug-condition pairs co-occurring within sentences that were previously not annotated by humans formed False relations. Altogether, 5968 False relations were automatically generated. The corpus enriched with machine annotated drugs, conditions, and relations between them is referred as ADE-EXT (indicating extended ADE corpus). Figure 1 shows an illustration of True and False relations between drug and conditions co-occurring within a sentence.

In the ADE-EXT corpus, 120 manually annotated True relations were not suitable for the NLP task. Examples include overlapping inter-related entities such as acute lithium toxicity where lithium is related to acute toxicity. After removal of nested annotations, the ADE-EXT corpus was decomposed into a training set (ADE-EXT-TRAIN) and a test set (ADE-EXT-TEST). Counts of entities and relations in subsets of ADE-EXT corpora is shown in Table 1.

### 2.2 Relation Extraction Workflow

For the identification and extraction of drug-condition entity pairs that fit into adverse effect relation, the Java Simple Relation Extraction (JSRE) system (Giuliano et al., 2007) was employed. JSRE provides a re-trainable and scalable supervised classification platform that uses Support Vector Machines (SVMs) (Burges, 1998) with different kernels specially designed for the NLP and relation extraction. All sentences in ADE-EXT-TRAIN and ADE-EXT-TEST containing drug-condition pairs labeled as either True or False were transformed into the SRE format before subjecting them to relation extraction. The SRE format is a unique way of representing data within the JSRE platform where tokens appearing in sentences are enriched with their parts-of-speech tags, lemmas, and flags indicating if a token is a part of named entity or not. Amongst different kernels available, the shallow linguistic kernel was thoroughly used since it has been widely applied and has shown success during similar relation extraction tasks (Tikk et al., 2010). The ADE-EXT-TRAIN was used as data for training and cross-evaluation of JSRE whereas the ADE-EXT-TEST was used an independent test set.

### 2.3 Mapping annotation ontology against Ontology of Adverse Events

The CLEF initiative (Roberts et al., 2007) investigated how to generate semantically annotated medical corpora for information extraction. As described (Gurulingappa et al., 2012) we adopted the standard established by the CLEF framework for the annotation workflow (Roberts et al., 2009) however we reshaped the annotation schema by using only two of the original entities (CONDITION, DRUG) and extended it with a third one (DOSAGE). None of the relationships used by the CLEF annotation schema could be reused for our work, since the CLEF annotation schema did not consider adverse drug reactions, instead we created two relations: DRUG-CAUSE-CONDITION, DRUG-HAS-DOSAGE. In this work we focused only on automating the detection of DRUG-CAUSE-CONDITION thus DOSAGE will not be mentioned further. The ADE corpus has been created using the Knowtator plugin for Protégé (Ogren, 2006), an ontology-driven corpus annotation tool also used for the creation of the CLEF corpus. Although we adopted the same tool used in CLEF and also adopted the standard established by the CLEF framework for the annotation workflow, we could not adopt the same annotation ontology since the latter was not able to capture the adverse drug relation and the drug dosing relation. The annotation ontology described above was therefore used to create the ADE corpus. Subsequent to the corpus creation, the realism-based biomedical ontology for representation of adverse events (AEO) has been published (Yongqun et al., 2011). AEO has been developed following the principles of Ontological Realism, thus is aligned with the Basic Formal Ontology and the Relation Ontology, and with the Open Biological and Biomedical Ontologies (OBO) Foundry principles of openness, collaboration and use of a common shared syntax. AEO has 484 representational units, annotated by means of 369 terms with specific identifiers and 115 terms imported from existing ontologies. The use of ontologies has proven of great value in biomedicine, also since it enable machine reasoning, abstraction and automatic hypothesis generation. We therefore had interest in investigating if the knowledge encoded in the annotations of the ADE
3 RESULTS

3.1 Performance Evaluation Criteria

The performance of relation extraction was evaluated by 10-fold cross-validation of the training data. During cross-validation of the training data and final evaluation over the test set, classification performances were assessed using the F-score over True-labeled relations since they denote adverse effect relations between drugs and conditions that denote a focused relation class being studied.

3.2 Assessment of Relation Extraction

Baseline experiments began with training and cross-validation of JSRE over the ADE-EXT-TRAIN corpus. Results of system’s performances are shown in Table 2. The system achieved an overall F-score of 0.87 after cross-validation. Upon the final test over ADE-EXT-TEST, the system attained F-score of 0.87 indicating a consistency in classification. A subset of instances misclassified during the cross-validation and testing were manually investigated to understand the common sources of errors. Limited context appeared to be one reason for misclassification. For example, the title Niacin maculopathy (PMID:3174043) infers maculopathy as an adverse effect of niacin that lacks contextual description to support machine classification. Distantly co-occurring inter-related entities constituted couple of errors. For example, in the sentence CASE SUMMARY: A 65-year-old patient chronically treated with the selective serotonin reuptake inhibitor (SSRI) citalopram developed confusion, agitation, tachycardia, tremors, myoclonic jerks and unsteady gait, consistent with serotonin syndrome, following initiation of fentanyl, and all symptoms and signs resolved following discontinuation of fentanyl (PMID:17381671); the relation between confusion and the last appearing drug name fentanyl was not correctly classified.

3.3 Impact of Size of the Training Set on the Performance

In order to study the impact of size of the training data on performance of classification, the ADE-EXT-TRAIN was decomposed into random subsets containing 10, 20, 50, 100, 200, 500, 1000, and 2000 documents. The JSRE was trained on these subsets independently in different rounds and subsequently applied on the ADE-EXT-TEST for performance evaluation. Table 3 shows that already using 500 documents one could achieve performances in the 80% range. Whereby, to reach a classifier with a standard deviation of 1%, one needs a substantially large training data.

3.4 Mapping the ADE Annotation Ontology to the Ontology of Adverse Events

As clearly shown in Figure 2, both the ADE annotation ontology and the Ontology of Adverse Events represent adverse drug reactions using formal ontological methods. Inspite of this common goal, the two ontologies use different naming for the two core entities: a CONDITION in the ADE annotation ontology coincide with a DRUG ADVERSE EVENT in AEO, a DRUG in the ADE annotation ontology coincide with a DRUG-ADMINISTRATION in AEO. The AEO ontology additionally introduce the entity DOSAGE, not specified in AEO at the time of its development since AEO originally focused on vaccines for which dosing is not an essential medical concept. Both ADE and AEO model a causal relationship between CONDITION or ADVERSE EVENT and DRUG OR MEDICAL INTERVENTION, with the latter being the causal source. The only entity shared by the CLEF annotation ontology with AEO and ADE is the DRUG-OR-DEVICE, that coincide with a DRUG OR MEDICAL INTERVENTION.

4 CONCLUSION

This work reports on the adaptation of a machine learning-based JSRE system for the identification and extraction of adverse effects of drugs in case reports. A methodology has been discussed to enrich a sparsely annotated corpus and its subsequent use to build a classification model. Evaluation of the system’s performance showed promising results. Performance of the system can be improved in several ways. In the current experiments, only the default features acceptable by JSRE were used. Optimization of feature representation to include additional features for instance from syntactic sentence parse trees may further improve the results. Development of additional strategies like post-processing to classify relations with missing contextual descriptions can help to recover more relations.

The reported experimental results denote research status on adverse drug effect identification from text. There are several strategies that will be immediately followed. The authors plan to benchmark the performances of several named entity taggers against the ADE corpus for the identification of drugs and conditions mentions in text.

The current experiments have been performed on the ADE corpus, since that was the only one available when this work was done, however while writing this report a new corpus has been published, namely the EU-ADR corpus (van Mulligen et al., 2012). It will be interesting to see if the performance of JSRE on the ADE corpus will be different compared to the EU-ADR corpus.
Similarly, benchmarking results of commercial and public relation extraction systems such as SemRep, Luxid MER Skill Cartridge®, RelEx, MedScan will be performed. The outcome of relation extraction from text to support signal detection and identity potentially novel or under-reported adverse effects will be studied.

The use of ontologies for driving information extraction has been reported (Wimalasuriya and Dou, 2010; Pandit and Honavar, 2010), we plan to explore the use of various available tools (e.g. ODIE, semantixs) using the AEO ontology and compare the performance of the ontology driven methods against the method presented here.

An outcome of the current work has demonstrated promising results and it has a potential to reduce the manual reading time, accelerate the signal tracking process, and therefore ensure safe use of drugs in the market.

5 ACKNOWLEDGEMENTS

This work has been partly supported by Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin, Germany and Bonn-Aachen International Center for Information Technology, Bonn, Germany.

REFERENCES


