ADEpedia 2.0: Integration of Normalized Adverse Drug Events (ADEs) Knowledge from the UMLS

Guoqian Jiang, MD, PhD[1], Hongfang Liu, PhD[1], Harold R. Solbrig[1], Christopher G. Chute, MD, DrPH[1]
[1]Department of Health Sciences Research, Division of Biomedical Statistics & Informatics, Mayo Clinic College of Medicine, Rochester, MN

Abstract

A standardized Adverse Drug Events (ADEs) knowledge base that encodes known ADE knowledge can be very useful in improving ADE detection for drug safety surveillance. In our previous study, we developed the ADEpedia that is a standardized knowledge base of ADEs based on drug product labels. The objectives of the present study are 1) to integrate normalized ADE knowledge from the Unified Medical Language System (UMLS) into the ADEpedia; and 2) to enrich the knowledge base with the drug-disorder co-occurrence data from a 51-million-document electronic medical records (EMRs) system. We extracted 266,832 drug-disorder concept pairs from the UMLS, covering 14,256 (1.69%) distinct drug concepts and 19,006 (3.53%) distinct disorder concepts. Of them, 71,626 (26.8%) concept pairs from UMLS co-occurred in the EMRs. We performed a preliminary evaluation on the utility of the UMLS ADE data. In conclusion, we have built an ADEpedia 2.0 framework that intends to integrate known ADE knowledge from disparate sources. The UMLS is a useful source for providing standardized ADE knowledge relevant to indications, contraindications and adverse effects, and complementary to the ADE data from drug product labels. The statistics from EMRs would enable the meaningful use of ADE data for drug safety surveillance.

1 Introduction

Adverse drug events (ADEs) are a well-recognized cause of patient morbidity and increased health care costs in the United States. Traditionally, spontaneous reporting is used as the main source for drug safety surveillance. Notably, the US Food and Drug Administration (FDA) uses an adverse event reporting system (AERS) to monitor for new adverse events and medication errors that might occur with all approved drug and therapeutic biologic products [1]. Various data mining approaches that use AERS reports have been developed to detect signals identifying associations between drugs and ADEs. For example, Vilar et al developed an approach that combines existing data mining algorithms with chemical information to enhance initial ADE signals generated from AERS [2]. However, the low reporting rate has been evidenced as one of main limitations from the spontaneous reporting approach [3].

To facilitate the ADE reporting and detection, there is emerging interest in secondary use of clinical data from the electronic medical records (EMRs). For examples, Linder et al implemented an EMR-based system to automatically send electronic ADE reports to the FDA in real-time [4]. Reisinger et al developed a common data model (CDM) to enable systematic analyses across disparate observational healthcare databases for the purpose of drug safety surveillance [5]. Chazard et al performed data mining to generate ADE rules using the data extracted from several hospitals' EMRs in a European project [6]. In a Clinical and Translational Research project at Mayo Clinic, a preliminary effort has been taken in building algorithms on identifying clinical phenotypes relevant to symptoms and findings associated with ADEs from the EMRs [7].

While these efforts on ADE detection from the EMR data and other ADE data sources look promising, however, a fundamental challenge is that the community lacks a publicly-available, standardized ADE knowledge base that encodes known ADE information for drug surveillance. The known ADE information is very useful for avoiding over-alerting of ADE signals detected by data mining algorithms [6], and also for reducing pharmacovigilance study noise levels, so that only novel signals are considered for further exploration [8]. To deal with this challenge, a few of studies are emerging to semantically annotate known ADE information, particularly using the FDA Structured Product Labels (SPLs) [9]. For example, Duke et al [10] developed a system called ADESSA, in which 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. In our previous study, for another example, we proposed a comprehensive framework for building a standardized ADE knowledge base known as *ADEpedia* (http://adepedia.org) through combining ontology-based approaches with Semantic Web technology [11].

The Unified Medical Language System (UMLS), developed by the National Library of Medicine (NLM), intends to promote creation of more effective and interoperable biomedical information system and services [12]. One of the

main knowledge resources in the UMLS is the Metathesaurus (Meta). The Meta contains information about biomedical and health related concepts, their synonyms and the relationships among them. Currently, it has integrated more than 160 source vocabularies, in which the major federal medication terminologies (e.g. RxNorm, NDF-RT, etc.) and other ADE relevant vocabularies (e.g. MedDRA, SNOMED CT, ICD, CTCAE, etc) are included. Although many of the ADE relevant source vocabularies have been utilized widely in different ADE detection projects, the integration of known ADE knowledge across the source vocabularies has not been explored in a systematic approach. We consider that a systematic review and organization of known ADE knowledge from the UMLS would be a good starting point to facilitate the integration of known ADE knowledge and ultimately form a comprehensive ADE knowledge base.

In addition, co-occurrence is one of essential statistics that has been widely used in text mining and knowledge acquisition from biomedical and clinical documents. For example, Wright et al. used co-occurrence-based association rule mining technique to identify a large number of clinical accurate associations that may be useful for identifying probable gaps in the problem list [13]. Chen et al. applied a co-occurrence-based method to compute and evaluate the strength of association for acquiring disease-drug knowledge from Medline abstracts and patient records [14]. Liu Y, et al demonstrated an approach that statistically significant co-occurrence of drug-disease mentions in the clinical notes can be used to detect ADE signals [15].

The objectives of this study are 1) to extract and normalize ADE knowledge from the UMLS and to integrate the ADE knowledge with the ADEpedia; and 2) to enhance the knowledge base with the drug-disorder co-occurrence data from a large-scale EMR system. We use the version of UMLS 2011 AB. We profile all drug concepts (from the semantic group Chemicals & Drugs), all potential ADE concepts (from the semantic group Disorders), and all relationships between the two concept sets. We manually review all the relationship types and group them into four categories: indications, contraindications, adverse effects and other associations. We obtain the co-occurrence data for all the UMLS drug-disorder concept pairs from a 51-million-document corpus of Mayo Clinic clinical notes. We perform a preliminary utility evaluation to demonstrate the usefulness of the system.

2 Materials and Methods

2.1 Materials

2.1.1 Knowledge Resources for ADE Integration

We used the following three knowledge resources. 1) UMLS 2011AB: We downloaded and installed the version of UMLS 2011AB. 2) Semantic groups: We used 2 UMLS semantic groups “Chemicals & Drugs (CHEM)” and “Disorders (DISO)”, and their corresponding semantic types [16]. 3) ADE data from the ADEpedia. In the knowledge base, the medication data are represented by RxNorm codes and the ADE data are represented by SNOMED CT and MedDRA codes.

2.1.2 Electronic Medical Records (EMRs) Data Sources

We used the UMLS CUI co-occurrence data that was extracted from a 51-million-document corpus of clinical text covering Mayo Clinic clinical notes between 1/1/2001-12/31/2010 in a previous study [17]. The clinical notes were originally retrieved from the Mayo’s Enterprise Data Trust (EDT) [18], which is a comprehensive snapshot of Mayo Clinic’s service areas and includes structured data, unstructured text, and Clinical Notes Indexing (CNI)-produced annotations [19]. Additionally, a broad range of note types at Mayo were represented, including Clinical Note, Hospital Summary, Post-procedure Note, Procedure Note, Progress Note, Tertiary Trauma, and Transfer Note.

2.2 Methods

2.2.1 Profiling the Drug-Disorder Pairs and Relationships in the UMLS

We developed a Perl script for extracting the ADE data set from the UMLS data files. And we also profiled the drug-disorder pairs and relationships asserted in the UMLS. First, we extracted a subset from the MRREL file that contains only drug-disorder or disorder-drug concept pairs and their relationships. The drug concepts are those CUIs assigned one of semantic types from the semantic group “Chemicals & Drugs” whereas the disorder concepts are those CUIs assigned one of semantic types from the semantic group “disorders”. The semantic type information is obtained from the MRSTY file and the concept names and source information are obtained from the MRCONSO file. Second, we analyzed the coverage of both drug and disorder concept sets that have drug-disorder relationships asserted in the UMLS. We also profiled the contributing source vocabularies for the asserted drug-disorder relations. Third, we profiled all relationships between the two concept sets and grouped them into the following four categories: indications, contraindications, adverse drug effects and other associations.

2.2.2 Enriching the UMLS ADE data with the EMR co-occurrence statistics

We extracted the co-occurrence data from the corpus of Mayo Clinic's clinical notes for all drug-disorder concept pairs identified from the section 3.3.2 above. We also extracted the co-occurrence data for all drug-disorder concept pairs from the ADEpedia. We used the UMLS CUIs and retrieved the corresponding CUIs from the MRCONSO file for all ADEs that originally are represented by SNOMED CT and MedDRA codes in the ADEpedia, and also for all drugs that are originally represented by RxNorm RxCUIs.

2.2.3 Evaluating the Utility of the UMLS ADE Data

The goal of the utility evaluation is to demonstrate the relevance and usefulness of the ADE data extracted from the UMLS for its integration in the ADEpedia. We compared the UMLS data set with the ADE data set from the ADEpedia. We identified the overlapping UMLS CUIs between the two ADE data sets and performed a case study to demonstrate added value of ADE knowledge from the UMLS for a specific drug.

4 Results

4.1 Profiling Results of ADE Knowledge in the UMLS

In the file MRREL of the UMLS 2011 AB, there are 844,993 unique concepts (i.e. CUIs) that have the semantic types from the semantic group "Chemicals & Drugs" and 538,851 unique concepts that have the semantic types from the semantic group "Disorders". We extracted 266,832 concept pairs between the two concept sets, covering 14,256 (1.69%) concepts from the group "Chemicals & Drugs" and 19,006 (3.53%) concepts from the group "Disorders". We profiled the contribution of source vocabularies to the concept pairs. In total, there are 47 source vocabularies having the contribution. Not surprisingly, NDF-RT (70.4%) and SNOMED CT (7.4%) are leading sources for the contribution. We also profiled all relationships (i.e. predicates) between the two concept sets. There are totally 102 relationships identified. We manually organized the relationships into 4 categories: 1) Indications covering 38.0% drug-disorder concept pairs; 2) Contraindications covering 24.3% concept pairs; 3) Adverse drug effects covering 15.1% concept pairs; and 4) Other associations covering 22.6% concept pairs.

4.2 Co-occurrence Enrichment Using the EMRs

We extracted the co-occurrence data for all drug-disorder concept pairs identified from the UMLS. Out of 266,832 concept pairs, 71,626 concept pairs co-occurred in the EMRs, accounting for 26.8%. Through mapping the drug and disorder concepts into the UMLS CUIs, we obtained 299,476 drug-disorder concept pairs represented in the UMLS CUIs. We also extracted the co-occurrence data for all drug-disorder concept pairs from the ADEpedia. Of 299,476 drug-disorder concept pairs, 100,577 concept pairs co-occurred in the EMRs, accounting for 33.6%.

4.3 Utility Evaluation and Case Study

The RxNorm drugs in the ADEpedia are represented by 3,904 distinct RxNorm Codes (i.e. RxCUIs), which are mapped to 3,832 distinct UMLS CUIs. Through comparing with the ADE data set extracted from the UMLS, we identified 2,164 CUIs overlapping across the two data sets, accounting for 56.5% of RxNorm Drugs in the ADEpedia. The 2,164 CUIs linked with 66,058 drug-disorder pairs from the UMLS.

As a case study, we selected a drug called "Digoxin", which is indicated for the treatment of congestive heart failure. "Digoxin" is represented by a RxCUI 3407 in the ADEpedia. Using its corresponding UMLS CUI C0012265 and querying the ADE data set extracted from the UMLS, we had the results of 32 drug-disorder pairs. Table 1 shows the results of additional ADE knowledge extracted from the UMLS for the drug "Digoxin|C0012265" and their co-occurrences from the EMRs. The relationships are also grouped into 4 categories: Indications, Contraindications, Adverse Drug Effects and Other Associations.

5 Discussion

In this study, we intended to integrate existing ADE knowledge from disparate resources to achieve a comprehensive ADE knowledge base. We envision that the ADE data extracted from FDA SPL labels (as we did in ADEpedia) and the asserted ADE knowledge in the UMLS are two main resources for known ADE knowledge that are complementary.

In the present study, the profiling results of the UMLS clearly demonstrated that the drug-disorder associations in the UMLS are largely underspecified. Only 1.69% (14,256) concepts from the group "Chemicals & Drugs" and 3.53% (19,006) concepts from the group "Disorders" in the UMLS had drug-disorder associations asserted. The main contributing vocabulary sources, not surprisingly, are NDF-RT and SNOMED CT, which contributed 70.4% and 7.4% of total drug-disorder pairs asserted in the UMLS. From the perspective of ADE detection application, we

classified the asserted relationships between the drug-disorder pairs into 4 categories, which are aligned with the original content model of NDF-RT. For example, as suggested in the NDF-RT documentation [29], we have put the relationships: may_treat, mayo_prevent, may_diagnose, induces and their inverse relations into the category Indications. We consider that this kind of categorization would provide aggregation capability for the knowledge source and improve its utility for the purpose of drug surveillance.

Table 1. Results of additional ADE knowledge extracted from the UMLS for the drug "Digoxin|C0012265" enriched by the EMR co-occurrences. The relationships are grouped into 4 categories: Indications, Contraindications, Adverse Drug Effects and Other Associations

Relationship	UMLS CUI	Preferred Name (Disorders)	Co-Occurrence
Indications			
may_prevent	C0018802	Congestive heart failure	394794
may_treat	C0004239	Atrial Flutter	23514
may_treat	C0039240	Supraventricular tachycardia	17899
may_treat	C0007166	Low Cardiac Output	1485
may_treat	C0036980	Shock, Cardiogenic	18
Contraindications			
contraindicated_drug	C0004238	Atrial Fibrillation	277344
contraindicated_drug	C0013182	Drug Allergy	33787
contraindicated_drug	C0042514	Tachycardia, Ventricular	21123
contraindicated_drug	C0018794	Heart Block	7134
contraindicated_drug	C0002726	Amyloidosis	3375
contraindicated_drug	C0007194	Hypertrophic Cardiomyopathy	2609
contraindicated_drug	C0031048	Pericarditis, Constrictive	1932
contraindicated_drug	C0042510	Ventricular Fibrillation	1534
contraindicated_drug	C0043202	Wolff-Parkinson-White Syndrome	862
Adverse Drug Effects			
causative_agent_of	C0274727	Poisoning by digoxin	5161
causative_agent_of	C0571956	Digoxin allergy	324
causative_agent_of	C0573925	Digoxin overdose	57
causative_agent_of	C0573925	Digoxin overdose	57
causative_agent_of	C0570462	Digoxin adverse reaction	8
causative_agent_of	C0416684	Accidental overdose of digoxin	2
causative_agent_of	C0573926	Intentional digoxin overdose	1
causative_agent_of	C0569329	Accidental digoxin poisoning	0
causative_agent_of	C0569330	Intentional digoxin poisoning	0
causative_agent_of	C0569331	Digoxin poisoning of undetermined intent	0
causative_agent_of	C0573927	Digoxin overdose of undetermined intent	0
causative_agent_of	C0569329	Accidental digoxin poisoning	0
causative_agent_of	C0569330	Intentional digoxin poisoning	0
causative_agent_of	C0569331	Digoxin poisoning of undetermined intent	0
causative_agent_of	C0573927	Digoxin overdose of undetermined intent	0
Other Associations			
inverse_isa	C0740721	Drug problem	83672
(blank)	C0274727	Poisoning by digoxin	5161
(blank)	C0161601	Poisoning by cardiotonic glycosides and drugs of similar action	0

We enriched the knowledge base with the statistical co-occurrence information extracted from the EMRs. We consider the co-occurrence information extracted from a 51-million-document EMR system will be useful for validating ADE detection algorithm across clinical institutions and across text corpora. In future, we consider it will make sense to explore the characteristics of the UMLS term occurrences in a large-scale corpus analysis as demonstrated at [17], focusing on the ADE detection use case.

To demonstrate the relevance and usefulness of the ADE knowledge extracted from the UMLS, we performed a preliminary utility evaluation and a case study. Comparing with the ADE data available from the ADEpedia developed in our previous study, we found out that 56.5% (2,164) of RxNorm Drugs in the ADEpedia had corresponding drug-disease associations asserted in the UMLS. The result indicated that the UMLS is a useful resource from the perspective of its integration with the ADEpedia. The case study of a drug "Digoxin| C0012265"

provided a typical example demonstrating the added knowledge from the UMLS into the ADEpedia in terms of indications, contraindications and adverse drug effects.

6 Conclusion

In conclusion, we have built an ADEpedia 2.0 framework that integrates known ADE knowledge from disparate sources. The UMLS is a useful source for providing standardized ADE knowledge relevant to indications, contraindications and adverse effects, and complementary to the ADE data from drug product labels. The statistics enrichment from EMRs would potential enable the meaningful use of ADE data for drug safety surveillance. The EMR-enriched ADE data is available for download at the ADEpedia wiki (http://adepedia.org). In the future, we will explore: 1) building a collaborative curation platform for the quality of the ADE data; 2) generating ADE data sets for different applications (e.g., all ADEs for the drugs with semantic type Clinical Drug); 3) making the ADE data available in the Semantic Web friendly formats; 3) investigating the ADE detection applications that utilize the EMR-enriched ADE data and known ADE knowledge in the ADEpedia.

Acknowledgements: This study is supported in part by the SHARP Area 4: Secondary Use of EHR Data (90TR000201).

References

[1] FDA AERS system URL: http://www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm; last visited at February 11, 2012.

[2] Vilar S, Harpaz, Chase HS, Costanzi S, Rabadan R, Friedman C. Facilitating adverse drug event detection in pharmacoviglance databases using molecular structure similarity: application to rhabdomyolysis. J Am Med Inform Assoc. 2011; Suppl 1:i73-80.

[3] Lopez-Gonzalez E, Herdeiro MT, Figueiras A. Determinants of underreporting of adverse drug reactions: a systematic review. Drug Saf. 2009; 32(1):19-31.

[4] Linder JA, Hass JS, Iyer A, Labuzetta MA, Ibara M, Celeste M, Getty G, Bates DW. Secondary use of electronic health record data: spontaneous triggered adverse drug event reporting. Pharmacoepidemiology and Drug Safety. 2010; 19:1211-1215.

[5] Reisinger SJ, Ryan PB, O'Hara DJ, Powell GE, Painter JL, Pattishall EN, Morris JA. Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. J Am Med Inform Assoc. 2010 Nov-Dec;17(6):652-62

[6] Chazard E, Ficheur G, Bernonville, Luyckx, Beuscart R. Data mining to generate adverse drug events rules. IEEE Transactions on Information Technology in Biomedicine. 2011; Vol. 15. 823-829.

[7] THE ADE Phenotype Ontology project: http://informatics.mayo.edu/adephont/index.php/Main_Page. last visited at February 11, 2012.

[8] Nadkarni PM. Drug safety surveillance using de-identified EMR and claims data: issues and challenges. J Am Med Inform Assoc. 2010 Nov-Dec;17:671-674.

[9] The FDA Structured Product Labeling URL: http://www.fda.gov/ForIndustry/DataStandards/StructuredProductLabeling/default.htm. last visited February 11, 2012.

[10] Duke JD, Friedlin J. ADESSA: A Real-Time Decision Support Service for Delivery of Semantically Coded Adverse Drug Event Data. AMIA Annu Symp Proc. 2010 Nov 13;2010:177-81.

[11] Jiang G, Solbrig HR, Chute CG. ADEpedia: a scalable and standardized knowledge base of Adverse Drug Events using semantic web technology. AMIA Annu Symp Proc. 2011;2011:607-16. Epub 2011 Oct 22.

[12] The UMLS URL: http://www.nlm.nih.gov/research/umls/; last visited at February 11, 2012.

[13] Wright A, Chen ES, Maloney FL. An automated technique for identifying associations between medications, laboratory results and problems. J Biomed Inform. 2010 Dec;43(6):891-901. Epub 2010 Sep 25.

[14] Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. J Am Med Inform Assoc. 2008 Jan-Feb;15(1):87-98. Epub 2007 Oct 18.

[15] Liu Y, LePendu P, Iyer S, Shah NH. Using temporal patterns in medical records to discern adverse drug events from indications. AMIA Summits Transl Sci Proc. 2012;2012:47-56. Epub 2012 Mar 19.

[16] The UMLS Semantic Groups file URL: http://semanticnetwork.nlm.nih.gov/SemGroups/; last visited at February 21, 2012.

[17] Wu S, Liu H, Li D, Tao C, Musen M, Chute C, Shah N.UMLS Term Occurrences in Clinical Notes: A Large-scale Corpus Analysis. J Am Med Inform Assoc. 2012 Jun 1;19(e1):e149-e156. Epub 2012 Apr 4.

[18] Chute CG, Beck SA, Fisk TB, Mohr DN. The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. J Am Med Inform Assoc.2010;**17**(2):131.

[19] Savova G, Kipper-Schuler K, Buntrock J, Chute C. UIMA-based clinical information extraction system. Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP 2008;39.