Standardized Drug and Pharmacological Class Network Construction

Qian Zhu[1] *, Guoqian Jiang [1], Liwei Wang[2], Christopher G. Chute[1]
[1] Department of Health Sciences Research,
Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN, USA
[2] School of Public Health, Jilin University, Changchun, Jilin, China

ABSTRACT

Dozens of drug terminologies and resources capture the drug and/or drug class information, ranging from their coverage and adequacy of representation. No transformative ways are available to link them together in a standard way, which hinders data integration and data representation for drug-related clinical and translational studies. In this paper, we introduce our preliminary work for building a standardized drug and drug class network that integrates multiple drug terminological resources, using Anatomical Therapeutic Chemical (ATC) and National Drug File Reference Terminology (NDF-RT) as network backbone, and expanding with RxNorm and Structured Product Label (SPL). In total, the network consists of 39,728 drugs and drug classes. Meanwhile, we calculated and compared structure similarity for each drug / drug class pair from ATC and NDF-RT, and analysed constructed drug class network from chemical structure perspective.

1 INTRODUCTION

Drug classes are group names for drugs that have similar activities or are used for a same type of disease and disorder. There are different ways to classify drugs. One way is to group drugs based on their therapeutic use or class (e.g., antiarrhythmic or diuretic drugs) as used by Anatomical Therapeutic Chemical (ATC) [1]. Another way is to group drugs using their dominant mechanism of action as used by National Drug File Reference Terminology (NDF-RT) [2]. However, drug classes defined by different systems are not compatible. It is worth to compare and integrate them in a universal fashion in order to support clinical related studies better. For example, Mougin, et al. [3] conducted a study for comparing drug classes between ATC and NDF-RT focusing on the relations between drugs and pharmacological classes (i.e., drug-class membership relations), which will facilitate the integration of these two resources.

Drug terminologies define drug entities as well as relevant properties and relationships with pharmacological classes. Drug terminologies are usually developed and maintained by different institutions using site-specific drug coding systems. Heterogeneous drug representations across different systems make it difficult to navigate diverse drug resources. The lack of a transformative way to link heterogeneous drug resources hinders data integration and data representation for drug-related clinical and translational studies. To overcome this obstacle, we proposed to represent drug information from diverse resources in a standard and integrated manner.

ATC and NDF-RT are the proposed sources of drug classification information. In the present study, we developed an approach to map drug and drug class entities from ATC and NDF-RT to UMLS (Unified Medical Language System) [4] and generated these mappings as a drug network backbone. Furthermore, we extended such network with RxNorm [5] and Structured Product Labeling (SPL) [6] integration, benefited from the broad drug relevant knowledge provided by these two resources. RxNorm provides links among different vocabularies, e.g. NDF-RT. SPL contains full drug interaction information, such as drug and drug interaction, and adverse drug event, etc., which has been explored and implemented by investigators and relevant applications have been developed, such as LinkedSPLs [7], ADEPedia [8]. Additionally, to extend and compare the drug classes defined by ATC and NDF-RT from chemical structure point of view, we introduced chemical structure similarity with the assumption that similar molecules have similar activities.

The paper is organized in several sections. We introduce the background knowledge about the resources and tools used in material section; in the methods section, we introduce the workflow details for network construction; then followed by discussion and conclusion sections.

2 MATERIALS AND METHODS

NDF-RT is a well-known drug terminological resource, and snapshot of NDF-RT was downloaded as of Nov. 8, 2012. In ATC classification system, drugs are categorized into different groups at five different levels according to the organ or system on which they act and/or their therapeutic and chemical characteristics [9]. ATC with a released version on January 2012 was used in this study. RxNorm provides normalized names for clinical drugs and links them to several drug vocabularies differentiating by “SAB” label. For example, “SAB=MTHSPL” indicates the source from SPL and “SAB=NDFRT” from NDF-RT. Two files are used in this study: 1) RXNCONSO.RRF, including all connections with source vocabularies. 2) RXNREL.RRF including relationships among concepts. RxNorm used in this study was

[*] To whom correspondence should be addressed: zhu.qian@mayo.edu

the version of Oct. 2012. SPL contains structured content of labeling (all text, tables and figures), along with additional machine readable information. The mappings between SPL and RxNorm used in this study are extracted from RxNorm RXNCONSO files with SAB = MTHSPL.

In this paper, we introduce a drug and drug class network by utilizing multiple drug terminological resources: ATC, NDF-RT, RxNorm, and SPL. ATC and NDF-RT are used as the network backbone, from which we integrated RxNorm and SPL as extension. Meanwhile, we calculated structure similarity for drug pairs from ATC and NDF-RT, and clustered them by structural similarity. The details of each step conducted in this study are described in the following sections.

2.1 Mapping NDF-RT with ATC

To map NDF-RT with ATC via UMLS, we translated NUI, NDF-RT Numerical Unique Identifier, and ATC name to UMLS CUI, UMLS concept unique identifier.

3.1.1 ATC mapping to UMLS

ATC is not well integrated with other drug terminologies (e.g., NDF-RT), as it uses its own coding system to code drug entities. To map ATC with NDF-RT and present the drug network transformatively by using standard representation, UMLS, we employed NCBO annotator [10] to semantically annotate each ATC name. Among more than 200 ontologies from UMLS Metathesaurus and NCBO BioPortal [11], RxNorm and NDF-RT have higher priority in this study. To avoid unnecessary annotations by non-drug relevant ontologies, we limited UMLS semantic types [12] within "Chemicals & Drugs" semantic group [13]. We extracted ontology id and concept id, which are two mandatory input parameters to invoke NCBO BioPortal REST API [14] for searching UMLS CUI, from the annotation results.

3.1.2 NDF-RT mapping to RxNorm and UMLS

NDF-RT concepts are organized into different categories with corresponding category labels. For example, "N0000179008, 1,1,1-trichloroethane, [Chemical/Ingredient]" and "N0000175641, Autonomic Ganglionic Blocker, [EPC]" are chemical ingredient and EPC class respectively. In this study, we retrieved the concepts that are labeled as VA class, VA product, EPC, Chemical ingredient and generic ingredient combination.

SQL query was executed to search RxCUIs (RxNorm Concept Unique Identifier) from RxNorm RXNCONSO table that was pre-loaded into our local MySQL database for NUIs. We retrieved UMLS CUI by invoking NLM RxNav RESTful API [15] with each NUI as an input parameter.

2.2 Calculating structure similarity

To analyze and expand the drug and drug class network from chemical structure perspective, we calculated the structure similarity among the drug pairs from ATC and NDF-RT, and grouped them using the score of structure similarity as Tanimoto Coefficient, i.e., similarity between these pairs of descriptors [16]. The cutoff value of the structure similarity is set as the score greater than 0.85, as it exhibits similar biological activity between the two molecules.

We first converted NDF-RT drug name and ATC name to SIMILES (Simplified molecular-input line-entry system) [17] as chemical representation by invoking PubChem entrez web service [18] and NCI resolver [19] REST API. Then we translated SMILES to chemical fingerprint and calculated Tanimoto similarity by using the aforementioned CDK functions.

2.3 Integrating RxNorm and SPL mappings

Mappings among RxNorm, SPL and NDF-RT are provided by RxNorm and available in the RxNorm RXNCONSO table. Two steps were performed to retrieve these mappings. First, we obtained concepts labeled as "SAB=NDFRT" and "SAB=RXNORM", denoted as RxNorm and NDF-RT mappings. Then, we searched for the concepts with "SAB=MTHSPL" label from the concepts identified in the first step. Then the final list of concepts is the common concepts across the three resources.

The network has been expanded from NDF-RT nodes that have mappings with RxNorm and SPL. We extracted SPL identifier (setId) from RXNREL table and saved for future SPL relevant information, LinkedSPL integration.

In addition, we performed a case study to demonstrate the usefulness of the drug and drug class network.

3 RESULTS

There are total 5,717 individual entities, which correspond to 4,483 distinct ATC names, i.e. one drug can be categorized into multiple therapeutic classes (more details described in the Discussion section).

Of 48,266 NDF-RT concepts, 34,011 concepts were used in this study, consisting of 15,857 VA Products, 486 VA classes, 9,960 Chemical/Ingredients, 7,184 Generic Ingredient Combinations, and 524 EPC. The child and parent relationships among these NDF-RT concepts are retrieved and stored from RxNorm RXNREL table via "CHD" (concept 1 is a child of concept 2) and "PAR" (concept 1 is a parent of concept 2) labels.

RxNorm, SPL and NDF-RT mappings were extracted from two RxNorm files: RXNCONSO and RXNREL, which were loaded into MySQL database.

3.1 Results for ATC and NDF-RT mappings

In order to build drug and drug class network with ATC and NDF-RT as backbone, first of all, we mapped ATC entities with NDF-RT concepts via UMLS, four steps involved.

4.1.1 ATC Annotated by NCBO

3,607 ATC entities including 3,152 drugs and 455 drug classes were mapped to UMLS CUIs by two ontologies, RxNorm and NDF-RT from NCBO BioPortal. Of these 3607 ATC mappings, 2180 ATC entities were exactly matched with the preferred names from RxNorm and NDF-RT. 866 ATC entities including 211 drug classes and 655 drugs were mapped to other ontologies available from NCBO. There are 1,244 ATC entities (21.8%) including 657 drugs and 587 drug classes failed to map to UMLS due to no annotations generated accordingly. We attempted to map these failed ATC names with RxNorm directly by invoking NLM RxNav RESTful API [20] with ATC names as input parameter, but none of them got mapping results. The failure reasons are discussed in the discussion section further.

4.1.2 NCBO annotation evaluation

The annotations were automated programmatically using NCBO Annotator Web Services API. We manually evaluated the annotation results. Of the 4,473 annotations with NDF-RT and RxNorm, 2,401 exact mappings were not further evaluated. The authors (QZ, LW) manually reviewed the rest of annotations (2,072 in total). As the evaluation results, 88.7% is correct, 10.3% is partial mappings, and 1.0% is incorrect. The precision was calculated as 99.5%, recall as 78.2% and F-measure as 87.4%, in which we counted exact mappings, partial mappings and correct mappings (4,453 in total) as true positive, 1,244 failed mappings as false negative and 20 incorrect mappings as false positive.

4.1.3 Mapping NDF-RT to RxNorm and UMLS

NDF-RT and RxNorm mappings exist in the RXNCONSO table with “SAB=NDFRT” label. Consequently, RxCUI corresponding to each NDF-RT concept can be retrieved from these mappings directly.

NDF-RT provides UMLS mappings. Hence, to retrieve UMLS for each NDF-RT concept, we called NLM NDF-RT RESTful API [9]. The searching results are shown in Table 1. 99.2% NDF-RT concepts have been mapped to UMLS.

NDF-RT Concepts	NUI	UMLS CUI
Chemical/Ingredient (9,960)	9,934	9,932
VA Class (486)	486	483
VA Product (15,857)	15,695	13,263
EPC (524)	480	478
Generic ingredient combination (7,184)	7,139	6,801
Total (34,011)	33,734	30,957

Table 1. UMLS CUI retrieval by RxNav NDF-RT API

4.1.4 ATC and NDF-RT mapping

In total, 3,850 distinct mappings between ATC and NDF-RT were generated, including 2,015 chemical/ingredients, 1,826 Generic Ingredient Combinations and 1 VA class. It includes distinct 2,226 ATC entities, covering 99 drug classes, and 2,127 individual drugs.

3.2 Results for structural similarity calculation

SMILES have been retrieved for all drugs from ATC and NDF-RT via PubChem Entrez web API and NCI Resolver web API. 2,618 ATC entities have gotten SMILES from NCI, 3,471 entries retrieved from PubChem. Combining NCI and PubChem searching results, total 3,487 ATC entries got SMILES, and 9,132 unique NDF-RT concepts got SMILES.

We calculated the Tanimoto coefficient as structure similarity for each pair of concepts from ATC and NDF-RT separately by converting SMILES to fingerprint. Then we got 8,513 pairs from ATC and 69,882 pairs from NDF-RT with Tanimoto coefficient greater than 0.85, and integrated them into the drug and drug class network.

3.3 Results for NDF-RT, RxNorm and SPL mapping

We integrated RxNorm and SPL mappings with NDF-RT. The mappings between RXNORM, NDF-RT and SPL resulted in 5,838 unique RxNorm concepts with 36,408 NDF-RT concepts and 41,188 SPL labels. The mappings mostly fall into two main categories according to term types defined by RxNorm, 3,056 are Semantic Clinical Drugs and 1,543 are Ingredients.

It is worthy to note that one RxNorm concept may be mapped to multiple NDF-RT and/or SPL concepts, for example, RxCUI “74” mapped to 3 NUIs in NDF-RT including N0000006481, N0000147349, N0000006481 and 11 set_ids in MTHSPL such as 0d65128b-8eb7-440b-870a-7e3be18152b3,1e6d6cd5-ab14-4258-a0fe-5f6a3cae437f.

4 DISCUSSION

In this study, we successfully built a drug and drug class network with 39,728 concepts from ATC and NDF-RT. All concepts were mapped to UMLS and labeled as UMLS CUIs accordingly. We also integrated RxNorm and SPL mappings, and extended the network with structure similarity calculation.

4.1 ATC to UMLS mapping

In total, 77.9% ATC terms have been mapped to UMLS. Comparing to 68.7% mapping results conducted by Merabti et al [21], our study shows the improvement of mappings from ATC to UMLS by leveraging NCBO annotator. However, 22.1% ATC terms failed to be mapped due to several reasons as follows, 1) Many of the ATC terms are combinations of multiple concepts, such as “calcium acetate and magnesium carbonate”, “combinations of sulfonamides and trimethoprim, including derivatives”; 2) The exclusions are embedded in the ATC names, such as “platelet aggregation inhibitors excluding heparin”, “nutrients without phenylalanine; 3) Non-standard representation is used by ATC though we corrected and expanded some abbreviations occurring in

ATC name. For example, “DIGESTIVES, INCL. ENZYMES” was corrected to “DIGESTIVES, INCLUDING ENZYMES”; 4) Non-drug terms are used, especially for drug classes in ATC, such as “VARIOUS”, “SENSORY ORGANS”. Above obstacles were the main reasons for mapping failure. In the future study, we will explore MMTx program that reported by Mougin et al. [3], and more NLP (Nature Language Processing) algorithms to parse ATC names for improving the mapping performance between the ATC and the UMLS.

4.2 Benefits from structure similarity integration

Structure similarity calculation applied in this study enables connections among the drug nodes sharing common similar chemical substructures. Beside the benefit shown in the case study, this integration also provides relevant clues for guiding clinical decision support system from the structure perspective as it offers a full profile of therapeutics for individual drugs. ATC classification system categorizes drugs according to its therapeutic classes; hence, one ATC drug can be grouped into multiple categories due to its diverse therapeutic functionalities. For instance, “Thonzylamine” is an antihistamine and anticholinergic used as an antipruritic and is grouped into two categories: “antiallergic agents” and “antihistamines for topical use” within the ATC hierarchy. The corresponding two ATC entities (R01AC06 and D04AA01) for “Thonzylamine” in two separate classes (“R” and “D”) are connected based on similarity score that is equal to 1. Thus, the entities within these two categories are connected, and physicians would be able to utilize such knowledge for Thonzylamine for their clinical decision from both therapeutics and structure point of view.

4.3 Future work

Drug entity mapping algorithm will be modified to enable more connections detected; more human review will be expected to improve the accuracy of the mappings. Meanwhile, we will seek possible collaborations with external sites such as the NLM for improving such mapping algorithm development. We will integrate more drug related resources, such as Drugbank and PharmGKB, and drug interaction data, drug and adverse event data as shown in Figure 1. The entire data set generated in this project will be released to public once the proposed action items accomplished.

5 CONCLUSION

We successfully integrated NDF-RT, ATC, RxNorm and SPL and built a drug and drug class network using standardized identifier for representing drug and drug class entities. In addition, the network was expanded from chemical structure perspective by similarity calculation. More other drug terminological resources and drug interaction information will be integrated in the future study.

ACKNOWLEDGMENTS

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

[1] ATC: http://www.who.int/classifications/atcddd/en/. Accessed by Apr.11.2013.

[2] NDF-RT: http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NDFRT/. Accessed by Apr.11.2013.

[3] Mougin, F., Burgun, A., and Bodenreider, O. Comparing Drug-Class Membership in ATC and NDF-RT . Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, 2012:437-443.

[4] Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004, 32, 267–270

[5] RxNorm: www.nlm.nih.gov/research/umls/rxnorm. Accessed by Apr.11.2013.

[6] Structured Product Labeling: http://www.fda.gov/ForIndustry/DataStandards/StructuredProductLabeling/default.htm. Accessed by Apr.11.2013.

[7] Hassanzadeh O, Zhu Q, Freimuth R, Boyce R, Extending the "Web of Drug Identity" with Knowledge Extracted from United States Product Labels, submitted to AMIA Summit on Clinical Research Informatics, 2013

[8] Jiang G, Solbrig H. R, Chute C.G. ADEpedia: a scalable and standardized knowledge base of Adverse Drug Events using semantic web technology. AMIA Annu Symp Proc. 2011:607-16.

[9]http://en.wikipedia.org/wiki/Anatomical_Therapeutic_Chemical_Classification_System. Accessed by Apr.11.2013.

[10] Jonquet C., Shah N., Musen M. The Open Biomedical Annotator. AMIA Summit on Translational Bioinformatics; 2009: 56–60. The NCBO Annotator web service: http://www.bioontology.org/annotator-service. Accessed by Apr.11.2013.

[11] Noy, N., Shah, N., Dai, B., Dorf, M., Gri_th, N., Jonquet, C., Montegut, M., Rubin, D., Youn, C., Musen, M.: Bioportal: A web repository for biomedical ontologies and data resources. In: Demo session at 7th International Semantic Web Conference (ISWC 2008)

[12] Semantic Type: http://www.nlm.nih.gov/research/umls/META3_current_semantic_types.html. Accessed by Apr.11.2013

[13] Bodenreider O, McCray AT Exploring semantic groups through visual approaches. Journal of Biomedical Informatics 2003; 36(6):414-432.

[14] BioPortal REST services: http://www.bioontology.org/wiki/index.php/NCBO_REST_services. Accessed by Apr.11.2013.

[15] NDF-RT RESTful API: http://rxnav.nlm.nih.gov/NdfrtRestAPI.html#label:r24. Accessed by Apr.11.2013.

[16] Holliday JD, Hu CY, Willett P, Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bitstrings. Comb Chem High Throughput Screen, 2002, 5(2):155-66.

[17] SMILES: http://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system. Accessed by Apr.11.2013.

[18] PubChem Entrez: http://www.ncbi.nlm.nih.gov/books/NBK25500/. Accessed by Apr.11.2013.

[19] NCI resolver: http://cactus.nci.nih.gov/chemical/structure. Accessed by Apr.11.2013.

[20] RxNorm RESTful API: http://rxnav.nlm.nih.gov/RxNormRestAPI.html. Accessed by Apr.11.2013.

[21] Merabti et al, 2011, Stud Health Technol Inform. 2011;166:206-13