Requirements and Strategy for the Development of a Pediatric Drug Ontology

Rachel Richesson[1,*], Jyotishman Pathak[2], Wendy McLeod[3], Ginger Blackmon[4], Kendra Vehik[3]
[1]Department of Informatics, Duke University School of Nursing, Durham, NC, USA
[2]Department of Health Sciences Research, Mayo Clinic College of Medicine, Rochester, MN, USA
[3]Pediatrics Epidemiology Center, University of South Florida College of Medicine, Tampa, FL, USA
[4]Wellhealth Pharmacy, Jacksonville, FL, USA

ABSTRACT
Clinical and epidemiological researchers across all medical specialties need tools and knowledge representations to support the classification, aggregation, and analysis of medication data. The National Drug File Reference Terminology (NDF-RT), a named standard for classifying medications, is developed by the US Department of Veterans Affairs (VA) as an extension to their National Drug File, which is the master list of drugs prescribed to VA patients, which are adults. NDF-RT is organized as a multi-axial hierarchy with additional relations between ingredients, medications, chemical structures, mechanism of action, and therapeutic indications. We describe our experience applying NDF-RT to a dataset of encoded medications that were collected from an international cohort of over 8,000 children. Our data-driven approach allows us to extract selected NDF-RT sub-classes of a researcher-provided concept of "antibiotics". We believe that a subset of concepts and relationships from NDF-RT will be sufficient to support pediatric research analyses involving classes and properties of medications, and that an NDF-RT subset relevant to pediatrics will be more easily adopted by clinical investigators and epidemiologists, thereby promoting standardization of drug classifications. Researchers from all domains would benefit from informatics tools utilizing ontologies to support data cleaning and analysis that is explicit, valid, and repeatable. We predict that a pediatric drug ontology view can be extracted from the NDF-RT reference ontology, and we hope for feedback from the ontology community on ways to advance this idea.

1 INTRODUCTION

Large multi-site data-rich research projects for complex diseases involve numerous data analyses conducted by different investigators and analysts associated with the studies. Likely, the data sets that are generated by these large research studies will be shared (in a de-identified manner) as publicly available data resources after the studies are completed. All of the data analysts affiliated with the studies and future data users from the community would benefit greatly from resources that can support the consistent classification of data for aggregated analyses. Despite the potential for ontologies to support consistency, quality and efficiency of data analyses, they are not yet widely applied in most research analysis settings.

We are currently evaluating the use of an existing reference terminology (the National Drug File Reference Terminology, NDF-RT) to enable consistent and reproducible approaches to analyzing medication data. Our previous work suggests that while NDF-RT is a suitably comprehensive drug classification ontology for pediatric medications, it is too complex for routine or single disease-specific analytic needs. We assert that the use of ontologies to support data analysis will require context-specific subsets of entities that relate to a given dataset, and relationships that relate to a given analysis plan or analytic approach.

2 BACKGROUND

2.1 International Diabetes Research

The incidence of Type 1 diabetes (T1DM) is increasing worldwide. The reason(s) for this increase remain unknown. (Vehik, Hamman et al. 2007) Researchers propose that multiple risk factors, including genetic predisposition, diet, body size, seasonality, infectious agents (primarily viruses) and geography, in addition to autoimmunity, are involved in the etiologic mechanism. The roadmap to understanding this complex disease entails: 1) identifying early life risk factors associated with autoimmunity and progression to T1DM; 2) investigating how changes in identified risk factors over time contribute to the changing incidence of T1DM; and 3) exploring hypothesized gene-environment interactions. (Vehik, Cuthbertson et al. 2011) The collection and analysis of medication data is an essential aspect for all of these research foci.

The Environmental Determinants of Diabetes in the Young (TEDDY) epidemiologic study of T1DM is funded by a half dozen organizations [see acknowledgement] to explore genetic-environmental interactions in relation to the development of T1DM. (TEDDY Study Group 2008) Over 8,600 newborns identified to be at genetic risk for T1DM are being followed for 15 years for the appearance of diabetes-associated autoantibodies and T1DM, with documentation of early childhood diet, child and maternal medications, infections, vaccinations, and psychosocial stressors. Study subjects are recruited across six clinical centers worldwide (Finland, Germany, Sweden and three in North America). Stool samples are collected monthly until 4 years and then biannually thereafter to measure bacterial, viral, dietary, chemical, and pharmaceutical biomarkers. The TEDDY study is in its 6th year - just recently completing the 5-year recruitment phase.

* To whom correspondence should be addressed:
rachel.richessson@dm.duke.edu

Multiple investigators have begun to analyze TEDDY data. The TEDDY study consists of 8 Principal Investigators and more than 60 study investigators from 4 nations organized into 9 subcommittees that address the 12 primary research questions and dozens of concurrent analyses on various research questions and topics of interest on TEDDY. This has already led to multiple duplicative and error-prone efforts by TEDDY working groups manually classifying reported medications into various drug classes. In the absence of a standard classification system for aggregating finely coded instances of medication data, we are seeing ad-hoc classifications by multiple TEDDY working groups. This is inefficient for the study as a whole, makes analyses difficult to replicate, and provides no guidance for the infinite number of secondary users of these data when they become a public resource at the end of the study. A standard ontology for drug classification, such as NDF-RT (Brown, Elkin et al. 2004), can enable standardized approaches to medication data grouping and analysis, thereby supporting comparability across studies, interpretation/synthesis of research findings, and meta-analysis. We explore and characterize the use of a subset of NDF-RT to support an explicit ancillary research question using TEDDY study data.

2.2 Standards for Naming and Classifying Drugs in the TEDDY Study

As of this writing, the TEDDY study has more than 2 million data points on 8,677 infants and children. The number will grow during the next ten years of the study. Of these data, there are approximately 200,000 instances of reported medications, coded using RxNorm, representing over 300 unique ingredients.

RxNorm is a nomenclature for clinical drugs produced by the U.S. National Library of Medicine (NLM). (Nelson, Zeng et al. 2011) RxNorm contains the names of prescription and many nonprescription formulations approved for human use (primarily in the USA). An RxNorm clinical drug name reflects the active ingredients, strengths, and dose form comprising that drug. When any of these elements vary, a new RxNorm drug name is created as a separate concept identified by a concept unique identifier (RxCUI). Consequently, to distinguish between such drug entities, RxNorm uses 'term types' (TTYs) that represent categories for generic and branded drugs. While it does provide extensive coverage for drug entities, RxNorm does not offer a sensible way to aggregate or classify clinical drugs or active ingredients for analysis. Despite this limitation, RxNorm was chosen as the coding system for the TEDDY study because of its inclusion of pediatric medications, regular maintenance by the NLM, including daily updates from the US Food and Drug Administration (FDA) and linkages to commercial pharmacy management information system knowledge bases.

Similar to RxNorm, NDF-RT includes lists of medications (ingredients and packaged products), but these are limited to those medications in the VA formulary, which does not serve pediatric populations. NDF-RT, however, does contain a multi- axial hierarchical knowledge structure for organizing drug classes. In particular, NDF-RT uses a description logic-based formal reference model that groups drugs and ingredients into the high-level classes for Chemical Structure (e.g., Acetanilides), Mechanism of Action (e.g., Prostaglandin Receptor Antagonists), Physiological Effect (e.g., Decreased Prostaglandin Production), drug-disease relationship describing the Therapeutic Intent (e.g., Pain), Pharmacokinetics describing the mechanisms of absorption and distribution of an administered drug within a body (e.g., Hepatic Metabolism), and legacy VA-NDF classes for Pharmaceutical Preparations (VHA Drug Class; e.g., Non-Opioid Analgesic). (Nelson, Brown et al. 2002)

NDF-RT is freely and publicly available through the NLM Unified Medical Language System (UMLS) (Bodenreider 2004) and the NCBO BioPortal. (Noy, Shah et al. 2009)

Our previous research (Richesson, Smith et al. 2007) using data from 2004-5 showed that RxNorm included codes for virtually all of the unique active ingredients (282/284 = 99%) from over 5,000 medications reported for over 1,200 children. This demonstrates the utility of RxNorm as a coding scheme for pediatric drugs, and the high coverage of RxNorm for pediatric and international medications validated the choice to use RxNorm in TEDDY, despite its limitations in organizing and classifying medications. Approximately 12% of unique drug ingredients reported in the TEDDY study did not have RxNorm codes, and hence could not be automatically mapped to NDF-RT classes using the UMLS mappings. As of December 2011, the TEDDY study data contained more than 200,000 instances of reported medications on 8,111 study participants.

Recognizing the important and different functions of RxNorm and NDF-RT and the need to navigate between them, the US NLM provides and updates mappings between these two systems as part of its UMLS. The mappings between RxNorm and NDF-RT are specified primarily between ingredients and clinical drugs. A graphical representation of the underlying RxNorm and NDF-RT information models, including multiple-inheritance reference hierarchies and named sets of medication concepts at different levels of abstraction can be found in Pathak and Richesson (2010). Their graphical representation also depicts the mapping relationship between RxNorm and NDF-RT systems using the ingredients and clinical drug linkages. Additional details about how the mappings between different information entities were traversed in this work can be found in Pathak, Murphy, et al. (2011).

3. APPROACH

3.1 Case study – Antibiotic medications and diabetes

Our work is an early evaluation of the coverage and feasibility of using NDF-RT *classes* to aggregate medication data coded at the ingredient level. Our assessment of NDF-RT is in the context of an explicitly defined ancillary research question and a specific dataset that was generated for TEDDY data to answer the question of whether early exposure to antibiotics is related to the presence and taxa of intestinal bacteria. The analysis data set includes 90 TEDDY subjects enrolled from all 6 clinical centers. These subjects were part of the highest HLA risk group in the TEDDY study and had provided stool samples from 3-18 months of age (making them eligible for the ancillary study by virtue of complete data). Among other variables (e.g., identifiers, patient demographics and characteristics, laboratory data on organisms present in stool culture, and laboratory data related to seroconversion to pre-diabetes), the data set for the ancillary study includes medications (as reported by parents at quarterly visits) encoded using RxNorm at the ingredient level. The investigators wanted to examine whether or not exposure to drugs with antibiotic properties impacts the diversity of intestinal bacteria found in TEDDY patients in different countries, as well as determine whether or not exposure to antibiotics changed the patterns of bacteria from specific functional groups over time. Modified Chi square tests and Poisson models have been used to support the analyses of stool sample and clinical data to identify differences within subgroups of TEDDY subjects. Using the selected hierarchical relationships from the NDF-RT ontology, we have transformed medication data into a dichotomous variable that can be fed into these and subsequent analysis by TEDDY investigators.

A total of 203 unique medication products were included in the analysis data set; 143 had RxNorm codes. For mapping the RxNorm ingredients to NDF-RT classes we developed a simple algorithm leveraging the RxNav and NDF-RT web services provided by the NLM. Simplistically, we explored the linkages between the clinical drug and ingredient concepts of RxNorm and NDF-RT. However, such a traversal is not trivial due to issues around misspelled drug names, lack of explicit relationships between RxNorm term types, as well as gaps in coverage across both the drug terminologies. (Chute, Pathak 2010) To address this issue, our algorithm adopts a 2-stage approach: in the first stage, it traverses the direct linkages between RxNorm SCD (Semantic Clinical Drug) and IN (Ingredient) concepts to corresponding clinical and drug and ingredient concepts in NDF-RT for identifying an appropriate VHA Drug Class. However, if this step fails either due to lack of mappings, or corresponding concepts, Stage II of the algorithm is pursued. In this second stage, the algorithm leverages chemical ingredient(s) information available for a particular drug product to assign NDF-RT drug classes. In particular, for a given drug product in RxNorm, this stage first identifies all the RxNorm and NDF-RT ingredient concepts for the drug product. The method then determines the drug product(s) in NDF-RT that contain only those NDF-RT ingredient concepts identified from the first step by traversing the child and sibling nodes in the hierarchy, and extracts the corresponding VA Drug Classes. The specifics of the algorithm used are described elsewhere. (Pathak, Murphy 2011)

We used NDF-RT January 12, 2010 release that has been synchronized with the RxNorm January 04, 2010 release. The mappings between RxNorm and NDF-RT entities between "Clinical Drug" and "Pharmaceutical Ingredient" were obtained from the respective source files using the unique identifiers for the concepts contained in the source files. We used a Microsoft Excel spreadsheet to document the classification, recoding, and expert review of the NDF-RT classifications. It is important to note that the grouping "antibiotic" has important clinical meaning and significance to TEDDY investigators, yet NDF-RT does not have a single class called 'antibiotics'. [NDF-RT does have several related classes such as 'antimicrobials', 'anti-infectives', and 'topical antibiotics' that must be combined to aggregate data into a clinically meaningful class called 'antibiotics'.] Using the NDF-RT Bioportal, we identified the classes of NDF-RT that could be considered as 'antibiotics'. The appropriateness of these classes was verified by TEDDY investigators, but it is worth noting that other investigators might construct different "antibiotic" groupings (for example, including or excluding topical anti-bacterials), based upon their particular research context and objectives.

Using a small set of 339 unique reported medications, we limited the number of NDF-RT classes that need to be considered as having antibiotic properties. Using these RxNorm-encoded medications and the UMLS mappings between RxNorm and NDF-RT, we extracted the associated NDF-RT parent classes in a particular hierarchy (the VA legacy class hierarchy) that is clinically oriented. We then used the NCBO Bioportal interface to traverse the hierarchy to determine if these data-driven classes are descendants of classes that we considered to have antibiotic properties. When the data-driven classes matched those NDF-RT antibiotic property classes, then we manually classified that medication ingredient as an antibiotic on our Excel spreadsheet.

We created a new dichotomous variable in the spreadsheet called "Antibiotic" (yes/no), and coded this as yes for all the RxNorm-coded medications that fell into the set of selected antibiotic-related classes. We then validated the resulting relationships by having a domain expert verify that each of the TEDDY reported medications that we classified as antibiotic could indeed be classified as such. We identified a domain expert who is a trained and license pharmacist prac-

ticing in a commercial setting. Resource constraints prevented us from using more than one expert reviewer. The reviewer was instructed to review a list of 143 unique medications on a spreadsheet and agree or disagree with our classification of the drug as an antibiotic.

Additionally, we asked the domain expert to view the list of 60 reported medications that were underspecified – e.g., "unknown antibiotic", "unspecified steroid" - to see if any could indeed be considered as antibiotics. (These underspecified medications are not precise enough to have RxNorm codes and hence were not mapped NDF-RT class or subsequent classification as an antibiotic.)

4. RESULTS

The 90 subjects in the data set for the proposed analysis on antibiotic use and intestinal bacteria diversity included 143 unique RxNorm ingredients which mapped to NDF-RT classes that are subclasses of antibiotic related classes. The NDF-RT antibiotic-related subclasses found in our sample are shown below.

ANTIBACTERIAL,TOPICAL
ANTIBACTERIALS,TOPICAL OPHTHALMIC
ANTI-INFECTIVES,OTHER
AMINOGLYCOSIDES
PENICILLINS,AMINO DERIVATIVES
NITROFURANS ANTIMICROBIALS
ERYTHROMYCINS/MACROLIDES
ANTIVIRALS
CEPHALOSPORIN 2ND GENERATION
CEPHALOSPORIN 1ST GENERATION
CEPHALOSPORIN 3RD GENERATION
CHLORAMPHENICOL
PENICILLIN-G RELATED PENICILLINS
NITROFURANS ANTIMICROBIALS
SULFONAMIDE/RELATED ANTIMICROBIALS
TETRACYCLINES

The domain expert agreed with our automatic antibiotic classification with all but 2 records of the 143 medications reported. One case was acyclovir, which is an antiviral. We had erroneously included this in our list of antibiotic classes. Similarly, the expert reviewer disagreed with our classification of Triclosan as an antibiotic. The review did consider it an antibiotic but clarified it as a topical, rather than systemic antibiotic. In addition, of the 60 reported medications that did not have RxNorm codes and that could not automatically be mapped to an NDF-RT class, the expert reviewer identified 5 that would be considered antibiotics and 2 that exhibited antibiotic properties.

5. DISCUSSION

This preliminary study builds upon our previous work and shows the potential of using existing tools to link precise coding systems to less granular reference terminologies to support a variety of secondary analyses and users. The use of a broad reference terminology as a medication domain ontology to support various research questions will require a systematic approach to assembling (data driven or expert selected) classes from the ontology to create groupings of significance to the end users. For example, the 'antibiotic' class is not reified in the NDF-RT and must be aggregated from subtypes that must be combined to aggregate data into a clinically meaningful class called 'antibiotics'. It is likely that not all experts would agree on such an aggregation for all diseases and contexts. Even in this small study, the expert consultant wished to make more fine grained distinctions in a few cases. Further experience and future automation of our methods could facilitate a standardized and consistent approach to a multitude of secondary data analyses across a variety of disease domains and research contexts.

Brewster et al. (2004) argue that a data corpus is the most accessible form of knowledge, and make the case for an ontology evaluation approach based upon data-driven evaluations. They propose several quantitative methods to evaluate the congruence of an ontology with a given corpus (or data set) in order to determine how appropriate it is for the representation of knowledge in that given domain. We argue that the re-creation and automation of our approach using many pediatric data sets can produce quantitative measures of NDF-RT 'fitness' as well as identify areas where the ontology should grow. We propose that a combination of multiple analysis questions and actual pediatric data can enable the extraction of data-driven views of NDF-RT, which (if broad enough in scope) can generate a broadly relevant Pediatric Drug Ontology view of NDF-RT. This paper presents a sensible strategy for extracting a Pediatric Drug Ontology from the NDF-RT and RxNorm resources. By reusing these resources to extract the view, interoperability can be maintained with other research efforts using these resources.

Our work also shows the importance of identifying the distinction between ontologies and reference terminologies. NDF-RT is a reference terminology, aimed at coverage of drug term usage, as such, it does not always obey good ontological principles (e.g., "catch-all" categories such as "Anti-Infectives, Other" and informal relations such as "may-treat"). As such, it is not clear at this time whether the extracted view would be a Pediatric Drug Ontology or a Pediatric Drug Reference Terminology, but we look forward to community feedback on the distinction and pros and cons of each.

For creating a full-fledged pediatric drug ontology "view" based on NDF-RT, we suggest the vSPARQL (Shaw, Landon, et al. 2011) ontology view creation platform. By extending the Semantic Web query language SPARQL, vSPARQL enables application of specific views over RDF

(http://www.w3.org/RDF/) and OWL (http://www.w3.org/2001/sw/wiki/OWL) data representations. Since NDF-RT is modeled in OWL, we can identify a core set of NDF-RT classes that are relevant to TEDDY, and create a sub-graph based on vSPARQL recursive querying capabilities. This sub-graph will provide the foundation for the pediatric drug ontology "view", which will be manually reviewed and refined, where necessary. This automated approach could be repeated using different datasets, and could also be used to develop quantitative metrics for evaluating ontology coverage or fitness of the NDF-RT for various data sets, study populations, and research contexts.

Though we conducted this research and demonstration manually, research on automating this technique would be of value to several communities. Future work could allow these selected relationships to be more readily implemented into statistical and analytical tools. Our approach can allow the views to be extended and collaboratively authored. The central storage (perhaps on the NCBO BioPortal) and automated access to the ontology view would allow the main ontology to grow and evolve as needed by the greater biomedical research community, and also allow the same methods and tools to be used to identify important drug class relationships that will facilitate future and repeated analyses. We are submitting proposals for funding of this approach. We look forward to the feedback of the ontology community on our strategy and results.

ACKNOWLEDGEMENTS

We wish to thank Lori Ballard and Jeff Krischer (University of South Florida, Tampa) from the TEDDY data center for their support, and Dr. Eric Triplet of the University of Florida, Gainsville, for his research question that inspired this demonstration. We also wish to thank Mike Haller and Helena Larsson from the TEDDY project, and Christopher Chute from the Mayo Clinic, for their helpful reviews and cooperation. TEDDY is funded by several NIH institutes, Juvenile Diabetes Research Foundation (JDRF), and Centers for Disease Control and Prevention (CDC). This work is also funded in part by the eMERGE grant.

REFERENCES

Bodenreider, O. (2004). "The Unified Medical Language System (UMLS): Integrating Biomedical Terminology." Nucleic Acids Research **32** (Database issue): D267-70.

Brewseter, C., H. Alani, et al. (2004). “Data Driven Ontology Evaluation.” In: International Conference on Language Resources and Evaluation (LREC 2004), 24-30. May 2004, Lisbon, Portugal.

Brown, S., P. Elkin, et al. (2004). "VA National Drug File Reference Terminology: A Cross-Institutional Content Cover-age Study." MedInfo: Studies in Health Technology and Informatics: 477-781.

Chute, C.G., J. Pathak. (2010). Analyzing categorical information in two publicly available drug terminologies: RxNorm and NDF-RT. J Am Med Inform Assoc. 2010;17(4):432-9.

Pathak J, S.P. Murphy, et al. (2011). Using RxNorm and NDF-RT to Classify Medication Data Extracted from Electronic Health Records: Experiences from the Rochester Epidemiology Project. AMIA Annu Symp Proc. 2011;2011:1089-98.

Pathak, J., R.L. Richesson, R.L. (2010). Use of standard drug vocabularies in clinical research: A case study in pediatrics. AMIA Annu Symp Proc. 2010 Nov 13;2010:607-11.

Nelson, S. J., S. H. Brown, et al. (2002). "A Semantic Normal Form for Clinical Drugs in the UMLS: Early Expe-riences with the VANDF." AMIA Annual Symposium: 557-561.

Nelson, S. J., K. Zeng, et al. (2011). "Normalized names for clinical drugs: RxNorm at 6 years." Journal of the American Medical Informatics Association**18**(3).

Noy, N., N. Shah, et al. (2009). "BioPortal: ontologies and integrated data resources at the click of a mouse." Nucleic Acids Research**37**(Suppl 2): 1-4.

Richesson, R., S. Smith, et al. (2007). "Achieving Standardized Medication Data in Clinical Research Studies: Two Approaches and Applications for Implementing RxNorm." Journal of Medical Systems.

Shaw, M., T. Landon, et al. (2011). vSPARQL: A View Definition Language for the Semantic Web. J Biomed Inform. 2011 February; 44(1): 102–117.

TEDDY Study Group (2008). "The environmental determinants of diabetes in the young (TEDDY) study." Annals of the New York Academy of Sciences**1150**: 1–13.

Vehik, K., D. Cuthbertson, et al. (2011). "Long-Term Outcome of Individuals Treated With Oral Insulin: Diabetes Prevention Trial-Type 1 (DPT-1) oral insulin trial." Diabetes Care.

Vehik, K., R. F. Hamman, et al. (2007). "Increasing incidence of type 1 diabetes in 0- to 17-year-old Colorado youth." Diabetes Care **30**(3): 503-509.