Proceedings of the 17th ISMB'2014 SIG meeting "Bio-Ontologies" 2014.
(electronic proceedings: https://drive.google.com/file/d/0B6AZUdKw1IVHeGVtTVZCRXJhbFk/).

# *Evaluating the consistency of inferred drug-class membership relations in NDF-RT*

*Rainer Winnenburg and Olivier Bodenreider**

*National Library of Medicine, Bethesda, Maryland, USA { rainer.winnenburg|olivier.bodenreider }@nih.gov*

**ABSTRACT**

**Objectives:** To evaluate the consistency of inferred drug-class membership relations in NDF-RT (National Drug File Reference Terminology). **Methods**: We use an OWL reasoner to infer the drug-class membership relations from the class definitions and the descriptions of drugs and compare them to asserted relations. **Results**: The inferred and asserted relations only match in about 50% of the cases. **Conclusions**: This investigation quantifies and categorizes the inconsistencies between asserted and inferred drug classes and illustrates issues with class definitions and drug descriptions. **Supplementary figure**: Overview of the methods, available at:
http://mor.nlm.nih.gov/pubs/supp/2014-bioonto-rw/index.html.

## 1 INTRODUCTION

The National Drug File-Reference Terminology (NDF-RT) is a drug ontology created as an extension to the formulary used by the Veterans Administration and developed using a description logic (DL) formalism. It has provided a rich description of drug classes in reference to drug properties, such as mechanism of action, physiologic effect, chemical structure and therapeutic intent. However, instead of logical definitions for these drug classes (i.e., necessary and sufficient conditions), only necessary conditions are provided. As a consequence, a DL reasoner cannot identify drugs as members of a given drug class, even when they are described in terms of the same properties.

In previous work, we showed that, after creating necessary and sufficient conditions for the drug classes, we could effectively infer drug-class membership (Bodenreider, et al., 2010). We demonstrated the use of a modified version of NDF-RT for clinical decision purposes (patient classification). One limitation of this work is that we did not evaluate the inferred drug-class membership relations beyond our proof-of-concept application.

NDF-RT recently integrated authoritative drug-class membership assertions extracted from the Structured Product Labels (package inserts) by the Food and Drug Administration (FDA), along with a description of the drugs in terms of the same properties used for defining the classes.

The objective of the present work is to evaluate the consistency of the drug-class membership relations that can be inferred from the class definitions and drug descriptions, against the asserted, authoritative drug-class membership relations. This evaluation is also an indirect contribution to the assessment of the class definitions and the drug descriptions in terms of completeness and consistency (i.e., agreement between information sources).

## 2 BACKGROUND

### 2.1 NDF-RT drugs and classes

The National Drug File Reference Terminology (NDF-RT) is a resource developed by the Department of Veterans Affairs (VA), Veterans Health Administration, as an extension of the VA National Drug File (Lincoln, et al., 2004). Like other modern biomedical terminologies, NDF-RT is developed using description logics and is available in native XML format. The version used in this study is the latest version available, dated April 11, 2014, downloaded from http://evs.nci.nih.gov/ftp1/NDF-RT/, from which we derived our OWL representation.

This version covers 7,287 active moieties (DRUG_KIND, level = ingredient), as well as 543 Established Pharmacologic Classes (EPCs) defined in reference to some of the properties of the active moieties. NDF-RT now contains several sources of relations between drugs and their properties. The April 2014 version of NDF-RT introduced a new set of relations between drugs and their properties originating from the class indexing file released as part of DailyMed, identified by the suffix “FDASPL”. Moreover, this version also introduced authoritative drug-class membership assertions from the same source. Finally, NDF-RT also provides a description of the EPCs in reference to the same properties used for describing the drugs themselves, provided by “Federal Medication Terminologies subject matter experts” and identified by the suffix “FMTSME”. In this work, we focus on the drug-property assertions from FDASPL, class-property assertions from FMTSME, and drug-class assertions provided by the FDA.

### 2.2 Related work

In addition to being used as a framework for building ontologies, description logics (DL) has been shown to be useful

[*] To whom correspondence should be addressed.

for reasoning with biomedical entities, including protein phosphatases (Wolstencroft, et al., 2006) and penetrating injuries (Rubin, et al., 2005). However, to our knowledge, DL reasoning has not yet been applied to the automatic classification of drugs, except for our previous work on anticoagulants (Bodenreider, et al., 2010).

NDF-RT is frequently used as a resource for standardizing drug classes (e.g., (Wang, et al., 2013; Zhu, et al., 2013)). However, investigators generally use the drug properties as classes (e.g., drugs that have the physiologic effect "decreased coagulation activity" for anti-coagulants), rather than the Established Pharmacologic Classes. Moreover, only asserted relations are used in most investigations, as opposed to inferred drug-class relations.

The specific contribution of this paper is to leverage the logical definitions of drug classes in NDF-RT to automatically infer drug-class relations using a DL reasoner. We substantially extend our previous work on anticoagulants, by generalizing it to all drug classes and providing a comparison to authoritative drug-class relations from the FDA.

## 3 MATERIALS AND METHODS

Our approach to evaluating inferred drug-class membership relations in NDF-RT can be summarized as follows. Before we can leverage a description logic (DL) reasoner to infer the drug-class membership relations from the class definitions and the descriptions of drugs, we need to convert the NDF-RT data from their original format (XML) to a description logic format (OWL). In fact, we create two OWL datasets, one containing the asserted drug-class relations used as our gold standard, and one from which they have been removed, so that only inferred drug-class relations will be present in this one after the reasoner has been applied. Finally, we compare inferred and asserted drug-class relations from the perspective of drugs and from that of classes.

### 3.1 Converting NDF-RT XML to OWL

In order to produce the two OWL datasets used for comparing asserted and inferred drug-class relations, we start by creating a "baseline" OWL representation from the original XML dataset, which we will use as our asserted dataset (dataset "A"). Here, as previously described in (Bodenreider, et al., 2010), we transform the primitive classes for external pharmacologic classes into defined classes by specifying a set of necessary and sufficient conditions for each class (adding an `owl:equivalentClass` (≡) axiom). For the purpose of this work, we only consider definitional the three properties used for the description of the drugs (mechanism of action, physiologic effect and chemical structure).

We further modify this OWL file in order to create the inferred dataset (dataset "I") by applying the following transformations, required for enabling the inference mechanism. In practice, we harmonize the names of roles used in the definition of the classes (e.g., *has_MoA_FMTSME*) with those used in the description of the drugs (e.g., *has_MoA_FDASPL*) by creating `owl:equivalentProperty` axioms between them. The following equivalences are created:

- *has_MoA_FMTSME* ≡ *has_MoA_FDASPL* (for mechanism of action),
- *has_PE_FMTSME* ≡ *has_PE_FDASPL* (for physiologic effect), and
- *has_Chemical_Structure_FMTSME* ≡ *has_Chemical_Structure_FDASPL*.

### 3.2 Inferring relations between drugs and EPCs

We can now leverage an OWL reasoner to infer the drug-class membership relations from the class definitions and the descriptions of drugs. From the necessary and sufficient conditions we created for the classes, an OWL reasoner infers a subclass relation between a drug and a drug class, when the properties of the drug match those of the drug class. For example, the drug class *beta2-Adrenergic Agonist [EPC]* (*N0000175779*) is defined as equivalent to (*'Pharmaceutical Preparations' and (has_MoA_FMTSME some 'Adrenergic beta2-Agonists [MoA]')*). The drug *albuterol* (*N0000147099*) has the property *has_MoA_FDASPL some 'Adrenergic beta2-Agonists [MoA]'*, and is therefore inferred as being a subclass of *beta2-Adrenergic Agonist [EPC]*. (The inference will also occur if the property of the drug is a subclass of the property used in the definition of the class).

A secondary benefit of the classification with an OWL reasoner is to create a hierarchy of the drug classes themselves, based on their logical definitions. For example, *beta2-Adrenergic Agonist [EPC]* (*N0000175779*) is inferred to be a subclass of *beta-Adrenergic Agonist [EPC]* (*N0000175555*), because the definition of *beta2-Adrenergic Agonist [EPC]* shown earlier is more specific than that of *beta-Adrenergic Agonist [EPC]* (*'Pharmaceutical Preparations' and (has_MoA_FMTSME some 'Adrenergic beta-Agonists [MoA]')*). For this reason, we reclassify both OWL datasets, although no inferred drug-class relation will be generated in dataset "A".

### 3.3 Comparing asserted and inferred drug-class relations

We compare asserted (dataset "A") and inferred (dataset "I") drug-class relations from the perspective of drugs and drug classes, respectively. In both cases, we issue queries against the OWL datasets (after reclassification). For each drug, we query its set of drug classes in each dataset and determine which classes are common to both datasets vs. specific to one dataset. For example, the drug *albuterol* (*N0000147099*) has the same class in both datasets, *beta2-Adrenergic Agonist [EPC]* (*N0000175779*). In contrast, the drug *hydrochlorothiazide* (*N0000145995*) had an asserted relation to *Thiazide Diuretic [EPC]* (*N0000175419*), but an inferred relation to *Thiazide-like Diuretic [EPC]*

(*N0000175420*). For each drug class, we query its set of drugs in each dataset and determine which drugs are common to both datasets vs. specific to one dataset. In order to consider higher-level classes to which no drugs may be direct members, we use the transitive closure of the hierarchical relation `rdfs:subClassOf`. As a consequence, a given class will have as members not only its direct drugs, but also the members of all its subclasses. Moreover, because salt ingredients are represented as "subclasses" of the corresponding base ingredients, both salt and base ingredient will be members the class of which the base ingredient is a member. For example, in both the "A" and "I" datasets, the class *beta-Adrenergic Agonist [EPC]* has the base ingredient *albuterol* as an indirect member through its subclass class *beta2-Adrenergic Agonist [EPC]*. It also has the salt ingredient *albuterol sulfate* as a member (through the base ingredient *albuterol*).

### 3.4 Implementation

The modifications described above were implemented into the OWL file using an XSL (eXtensible Stylesheet Language) transformation. The resulting OWL file was classified with HermiT 1.2.2 (University of Oxford - Information Systems Group, 2010). Protégé 4.3 was used for visualization purposes (Stanford Center for Biomedical Informatics Research, 2014). The OWL file containing the inferences computed by the reasoner was loaded in the open source triple store Virtuoso 7.10 (OpenLink Software, 2014). The query language SPARQL was used for querying drug-class relations.

## 4 RESULTS

### 4.1 Asserted and inferred drug-class relations

***Drugs***. Of the 7,287 drugs (at the ingredient level) in NDF-RT, 1,540 have at least one relation to a drug class (EPC). As shown in Table 1, all but two drugs (1,538) have asserted drug-class relations and 1,000 drugs have inferred relations. 998 drugs have both asserted and inferred relations.

***Drug classes***. Of the 543 drug classes (EPC) in NDF-RT, 471 have relations to drugs (462 are directly related to a drug and 9 are related indirectly through their subclasses). Of the 462 classes with direct relations to drugs, all but 12 (450) have asserted relations and 299 have inferred relations. As shown in Table 2, of the 471 classes with direct or indirect relations to drugs, all but three (468) have asserted relations and 309 have inferred relations. In total, 306 of these 471 classes have both asserted and inferred relations to drugs.

***Drug-class relations***. There are 1,787 asserted and 1,047 inferred direct drug-class relations, of which 872 are in common. Of the asserted relations, 915 could not be inferred, whereas 175 inferred relations are not present in the asserted set. Considering the transitive closure of the hierarchical relation `rdfs:subClassOf`, we obtain 4,169 asserted and 2,378 inferred drug-class relations, of which 2,310 are in common. Of the asserted relations 1,859 could not be inferred, whereas 68 inferred relations are not present in the asserted set.

### 4.2 Perspective of drugs

For each drug, we compare the set of (direct) drug classes in datasets "A" and "I". The various types of differences observed between asserted and inferred drug-class relations are presented in Table 1. The largest category corresponds to drugs with identical sets of asserted and inferred drug-class relations (46%). For example, the drug *imatinib* has the same class *Kinase Inhibitor [EPC]* in both datasets. Drugs with asserted drug-class relations, but lacking inferred drug-class relations represent 35% of the cases. For example, the drug *losartan* has the class *Angiotensin 2 Receptor Blocker [EPC]* in dataset "A", but no class in dataset "I".

**Table 1.** Drug-class relations (direct), drug perspective

| Drugs related to drug classes | # | % |
|---|---|---|
| Drugs with identical sets of classes for the asserted and inferred drug-class relations | 703 | 45.65 |
| Drugs with compatible sets of classes (each class from the asserted is identical to or hierarchically related to a class in the inferred set) | 130 | 8.44 |
| Drugs with additional drug-class relations in the asserted set only | 133 | 8.64 |
| Drugs with additional drug-class relations in the inferred set only | 16 | 1.04 |
| Drugs with additional drug-class relations in both the asserted and inferred set | 16 | 1.04 |
| Drugs with asserted drug-class relations only (no inferred relations) | 540 | 35.06 |
| Drugs with inferred drug-class relations only (no asserted relations) | 2 | 0.13 |
| **Total number of related drugs** | **1540** | **100.00** |

**Table 2.** Drug-class relations (direct and indirect), class perspective

| Drug classes related to drugs | # | % |
|---|---|---|
| Classes with identical sets of drugs for the asserted and inferred drug-class relations | 243 | 51.59 |
| Classes with additional drug-class relations in the asserted set only | 38 | 8.07 |
| Classes with additional drug-class relations in the inferred set only | 20 | 4.25 |
| Classes with additional drug-class relations in both the asserted and inferred set | 5 | 1.06 |
| Classes with asserted drug-class relations only (no inferred relations) | 162 | 34.39 |
| Classes with inferred drug-class relations only (no asserted relations) | 3 | 0.64 |
| **Total number of related classes** | **471** | **100.00** |

### 4.3 Perspective of drug classes

For each drug class, we compare the set of (direct and indirect) drug members in datasets "A" and "I". The various types of differences observed between asserted and inferred drug-class relations are presented in Table 2. As we observed for drugs, the largest category corresponds to drug classes with identical sets of asserted and inferred drug-class relations (52%). For example, the class *Monoamine Oxidase Inhibitor [EPC]* has the same nine drugs in both datasets, including *isocarboxazid* and *rasagiline*. Drug classes with asserted drug-class relations, but lacking inferred drug-class relations also represent about 35% of the cases. For example, the class *Antimalarial [EPC]* has 16 drugs in dataset “A”, including *chloroquine* and *proguanil*, but no members in dataset “I”.

# 5 DISCUSSION

## 5.1 Inconsistencies between asserted and inferred drug-class relations

***Missing inferences***. As mentioned in the results, the largest category of inconsistencies is represented by missing inferred drug-class relations, including cases where there are no inferred relations at all and cases where inferred relations only cover part of the asserted relations. Missing inferences should not be interpreted as an inherent failure of the OWL reasoner to identify drug-class relations, but rather as issues with the completeness and quality of class definitions and drug descriptions (see below for details). For example, the reason why the drug *trazodone* has an asserted, but not inferred drug-class relation to *Serotonin Reuptake Inhibitor [EPC]* (unlike *citalopram* that has both inferred and asserted relations) is because the mechanism of action of *trazodone* (*Serotonin Uptake Inhibitors [MoA]*) is not described in the dataset.

***Inferences with no corresponding asserted relations***. Although modest, the number of cases (38 drugs and 28 classes) where inferred drug-class relations are found when there is no asserted drug-class relation (or a different asserted drug-class relation) is interesting as it can help detect potentially missing asserted relations. For example, the drug *bupropion* has a single asserted relation to the structural class *Aminoketone [EPC]*. However, it has an inferred relation to *Norepinephrine Reuptake Inhibitor [EPC]* (through its mechanism of action *Norepinephrine Uptake Inhibitors [MoA]*). In this case, the set of asserted relations, which we use as our reference seems to be incomplete.

***Inconsistent drug-class relations due to granularity differences***. Drug-class relations from dataset “A” tend to associate drugs with more specific classes than in dataset “I”. For example, the antibiotic *amikacin* is associated with *Aminoglycoside Antibacterial [EPC]* (through asserted relations), but with the less specific *Aminoglycoside [EPC]* (through inferred relations). As shown in Table 1, we identified 130 drugs for which the classes in sets “A” and “I” are hierarchically related. Of these, there are only 4 cases with an inferred relation to a class that is more specific than the class involved in the asserted relation.

***Issues with class definitions and drug descriptions***. Some of the class definitions (e.g., *Antimalarial [EPC]*) refer to therapeutic intent (i.e, *may_treat*, *may_prevent*), which the FDA drug properties currently do not cover. Relations to such classes can therefore not be inferred from the current data. This issue accounts for 326 drugs with “missing” inferred relations. Moreover, 409 drugs are not described with any of the three properties used in the definition of the drug classes (e.g., the anticoagulant *rivaroxaban*). The majority of these cases involve salt ingredients (e.g., *albuterol sulfate*), which can only be associated with a class through the corresponding base ingredient, and allergenic extracts (e.g., *allergenic extract, bee*), for which drug descriptions are only inconsistently provided.

## 5.2 Limitations and future work

The analysis of the inconsistencies between asserted and inferred drug-class relations presented here is essentially quantitative. A detailed qualitative analysis does not fit within the confines of a short paper, but will be presented in a follow-up journal article.

Another limitation of our work is that it is not meant to capture cases where both the asserted drug-class relations and the drug description are missing (e.g., the antihypertensive drug *lisinopril*, which should be associated with the class *Angiotensin Converting Enzyme Inhibitor [EPC]*). Comparison with another drug classification, such as ATC, would help identifying such cases.

# ACKNOWLEDGEMENTS

This work was supported by the Intramural Research Program of the NIH, National Library of Medicine (NLM).

# REFERENCES

Bodenreider, O., Mougin, F. and Burgun, A. (2010) Automatic determination of anticoagulation status with NDF-RT. *13th ISMB'2010 SIG meeting "Bio-ontologies"*. pp. 140-143.

Lincoln, M.J., *et al.* (2004) U.S. Department of Veterans Affairs Enterprise Reference Terminology strategic overview, *Stud Health Technol Inform*, **107**, 391-395.

Virtuoso: http://virtuoso.openlinksw.com/

Rubin, D.L., Dameron, O. and Musen, M.A. (2005) Use of description logic classification to reason about consequences of penetrating injuries, *AMIA Annu Symp Proc*, 649-653.

Protégé: http://protege.stanford.edu/

HermiT: http://hermit-reasoner.com/

Wang, L., *et al.* (2013) Standardizing drug adverse event reporting data, *Stud Health Technol Inform*, **192**, 1101.

Wolstencroft, K., *et al.* (2006) Protein classification using ontology classification, *Bioinformatics*, **22**, e530-538.

Zhu, Q., *et al.* (2013) Standardized drug and pharmacological class network construction, *Stud Health Technol Inform*, **192**, 1125.