Published in final edited form as:
J Biomed Inform. 2017 September ; 73: 30–42. doi:10.1016/j.jbi.2017.07.013.

Quality Assurance of Chemical Ingredient Classification for the National Drug File – Reference Terminology

Ling Zheng[1], **Hasan Yumak**[2], **Ling Chen**[2], **Christopher Ochs**[1], **James Geller**[1], **Joan Kapusnik-Uner**[3], and **Yehoshua Perl**[1]

[1]Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102

[2]BMCC, CUNY, New York, NY 10007

[3]First Databank Inc., San Francisco, CA 94080

Abstract

The National Drug File – Reference Terminology (NDF-RT) is a large and complex drug terminology consisting of several classification hierarchies on top of an extensive collection of drug concepts. These hierarchies provide important information about clinical drugs, e.g., their chemical ingredients, mechanisms of action, dosage form and physiological effects. Within NDF-RT such information is represented using tens of thousands of roles connecting drugs to classifications. In previous studies, we have introduced various kinds of Abstraction Networks to summarize the content and structure of terminologies in order to facilitate their visual comprehension, and support quality assurance of terminologies.

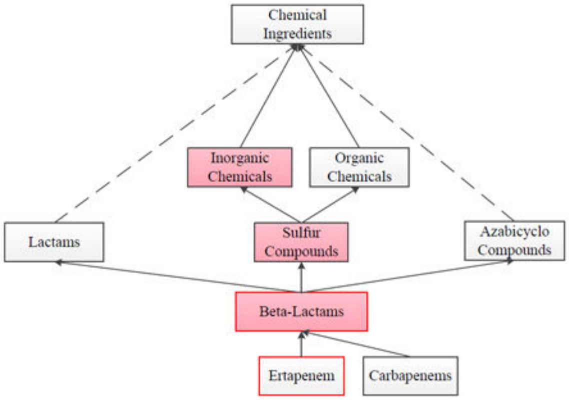
However, these previous kinds of Abstraction Networks are not appropriate for summarizing the NDF-RT classification hierarchies, due to its unique structure. In this paper, we present the novel Ingredient Abstraction Network (IAbN) to summarize, visualize and support the audit of NDF-RT's *Chemical Ingredients* hierarchy and its associated drugs. A common theme in our quality assurance framework is to use characterizations of sets of concepts, revealed by the Abstraction Network structure, to capture concepts, the modeling of which is more complex than for other concepts. For the IAbN, we characterize drug ingredient concepts as more complex if they belong to IAbN groups with multiple parent groups. We show that such concepts have a statistically significantly higher rate of errors than a control sample and identify two especially common patterns of errors.

Graphical abstract

Corresponding Author: Ling Zheng, Department of Computer Science, New Jersey Institute of Technology, Newark, NJ 07102, Tel #: (862) 371-0988, Fax #: (973) 596-5777, lz265@njit.edu.

Competing interests
The authors declare that they have no competing interests.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimersthat apply to the journal pertain.



Keywords

National Drug File – Reference Terminology; NDF-RT Quality Assurance; Auditing NDF-RT; Drug Terminology; Chemical Ingredients Hierarchy; Auditing Chemical Ingredients

1 Introduction

Quality assurance **(QA)** of biomedical terminologies is important. Errors in terminologies may propagate to the biomedical information systems that use them. User concerns about errors in terminologies have been reported, for example, by Elhanan et al. [1]. The aim of terminology quality assurance is to improve the modeling of a terminology (e.g., by identifying incorrect and/or inconsistent relationships between concepts, a process we will refer to as "auditing"). Zhu et al. [2] reviewed methods for quality assurance of terminologies, appearing in the first special issue of a journal on "auditing of biomedical terminologies," edited by Geller et al. [3].

The aim of this paper is to develop a QA methodology for the National Drug File – Reference Terminology (NDF-RT) [4], specifically its large *Chemical Ingredients* (CI) hierarchy. NDF-RT is a drug terminology developed and maintained by the U.S. Department of Veterans Affairs (VA), Veterans Health Administration (VHA). NDF-RT is a formal representation of the VHA National Drug File (NDF) [5], which is a drug classification hierarchy used to group orderable drug products into one of 579 drug classes. NDF-RT is used to support clinical applications at the VHA's clinical centers. NDF-RT is a large terminology composed of over 43,000 concepts connected by 67,000 IS-A links and 73,000 other links (called "roles"), e.g., *has_Ingredient* and *has mechanism of action*.

The effects of drugs depend mostly on their chemical ingredients and each drug in NDF-RT is linked to its chemical ingredients via *has_Ingredient* roles (see Figure 1). Improving the modeling of NDF-RT's chemical ingredient concepts will improve the modeling of NDF-RT's drug concepts.

In a long range research program the SABOC team [6] has developed a framework that combines summarization, visualization and quality assurance (SVQA) into a sequence of well-ordered steps. The overall aim of the SVQA paradigm is to identify sets of concepts in a terminology that are expected to have a higher error rate than reference concepts. The identification of theses sets is based on *summaries* derived from the terminology's structure and semantics. Limited QA resources can be applied to the concepts in such a set, improving the rate of error detection and correction. The aim of this paper is to develop a method within the SVQA paradigm for quality assurance of the NDF-RT.

We will now elaborate on the summarization component of SVQA. Our approach to terminology summarization is based on algorithmically processing the network of concepts and hierarchical IS-A links to produce a smaller network of "nodes" and *child-of* links. We refer to the resulting smaller network as an Abstraction Network [7]. We have derived different kinds of Abstraction Networks to support QA for various terminologies and ontologies (see Section 2.2). One type of Abstraction Network we previously applied for terminology QA is the *partial-area taxonomy* [8], which is derived according to the role relationships emanating from a hierarchy's concepts.

However, it is impossible to derive *partial-area taxonomies* for large portions of NDF-RT; the structures of several concept hierarchies do not contain enough information to perform such a derivation. Seven of the NDF-RT hierarchies have no roles emanating from their concepts (i.e., their concepts only have hierarchical relationships). All of the roles in NDF-RT emanate from concepts in the *Pharmaceutical Preparations* (*PP*) hierarchy to concepts in the other hierarchies (Figure 1). Hence, the only hierarchy of the NDF-RT that lends itself to deriving the partial-area taxonomy (a kind of Abstraction Network) is the *PP* hierarchy.

Thus, to apply the SVQA process to NDF-RT's *Chemical Ingredients* (*CI*) hierarchy, a new summarization process is needed. Such a process will need to identify groups of chemically similar concepts that serve as targets of *has_Ingredient* roles. The challenge is that NDF-RT, in its source format, does not support the explicit identification of such concepts. Thus, as part of this study we have developed an Abstraction Network that groups similar chemical ingredient concepts together. Using this Abstraction Network we can characterize sets of concepts that are structurally complex. One can expect a higher rate of errors among these concepts, since their modeling is more challenging for a terminology designer.

In summary, in this paper: **(1)** We introduce an Abstraction Network called an *Ingredient Abstraction Network* (IAbN) to summarizes NDF-RT's chemical ingredients and their associated drug concepts. The IAbN enables visual comprehension and quality assurance of NDF-RT, completing the components of our SVQA paradigm. **(2)** Using an IAbN we identify characterizations of complex concepts that can be used to support QA of NDF-RT's chemical ingredient concepts. **(3)** We report on a consensus-based QA review of a sample of NDF-RT concepts that was performed by three chemistry domain experts to evaluate the effectiveness of the SVQA methodology when applied to NDF-RT's *Chemical Ingredients* hierarchy, and to find errors in the NDF-RT.

2 Background

2.1 NDF-RT

NDF-RT uses a description logic-based reference model to define drugs in the *Pharmaceutical Preparations* (*PP*) hierarchy according to multiple aspects [9]. These aspects include: the *Chemical Ingredients* (*CI*) hierarchy describing the chemical ingredients of drugs, the *Cellular or Molecular Interactions* (*MoA*) hierarchy describing the drug effects at molecular, subcellular, or cellular levels, the *Physiological Effects* (*PE*) hierarchy describing drug effects at tissue, organ, or system levels, the *Clinical Kinetics* (*PK – from Pharmacokinetics*) hierarchy describing the absorption, distribution, and elimination of drugs, and the *Therapeutic Categories* (*TC*) hierarchy, which is an experimental hierarchy exclusively used to model FDA established pharmacologic class concepts to describe general therapeutic intents of drugs. Two more hierarchies are the *Diseases, Manifestations or Physiologic States* (*Disease*) hierarchy describing the therapeutic, preventative, or diagnostic indications of drugs, and the *Dose Forms* hierarchy describing the dose forms of drugs.

The *MoA, PE* and *CI* hierarchies were initially created by matching VHA drug ingredient names to terms from the National Library of Medicine's Medical Subject Headings (MeSH) [10]. Specifically, the *CI* hierarchy was derived from MeSH's *Chemicals and Drugs Category* and the *MoA* and *PE* hierarchies were created by extending and restructuring selected Pharmacologic Actions associated with ingredients in MeSH. Concepts in the *Disease* hierarchy were included from MeSH's *Diseases Category* [9, 11]. The purpose of developing the MeSH was to support the classification of biomedical publications in the PubMed system [12] of the National Library of Medicine.

NDF-RT is available for download in Apelon DTS format at the NCI's Enterprise Vocabulary Services (EVS) website [13]. NDF-RT is also released as part of the UMLS [4] and it is available for download on the NCBO BioPortal [14]. NDF-RT organizes concepts around the *PP* hierarchy (the triangle in Figure 1), which is the largest hierarchy in NDF-RT with 25,759 concepts (59.4% of the 43,397 concepts in the June 2015 version). The root concept of the *PP* hierarchy is *Pharmaceutical Preparations*. Besides IS-A relationships, concepts in the *PP* hierarchy can have role relationships (represented by the arrows in Figure 1) pointing to concepts in the other hierarchies (the seven rectangles in Figure 1). Role relationships are used to define drugs according to their various aspects. Drug-disease relationships were mined from co-occurrence data in the Unified Medical Language System (UMLS) [15] (see Figure 1). The *TC* hierarchy is exclusive for concepts established by the FDA, so there are no NDF-RT asserted roles between the *PP* hierarchy and the *TC* hierarchy and the arrow in Figure 1 is not labeled with any NDF-RT asserted role.

For example, in Figure 2, the drug preparation *ASPIRIN* in the *PP* hierarchy has the role relationship *has_Ingredient* pointing to a chemical ingredient *Aspirin* in the *CI* hierarchy, the second largest hierarchy in NDF-RT with 10,145 concepts. The role relationships of drug classes and drug preparations are inherited by orderable drug products, e.g., *ASPIRIN 300MG TAB* (a VA Product) inherits the role relationship *has_Ingredient* and its target concept *Aspirin* from its parent drug preparation *ASPIRIN*.

Concepts in each hierarchy are organized in a generalization hierarchy; top level concepts are more general than bottom level concepts. Some concepts may have multiple parents. For example, *ASPIRIN* in the *PP* hierarchy and *Salicylates* in the *CI* hierarchy each have two parents in Figure 2.

Concepts in the *PP* hierarchy may have different types of role relationships to concepts in the *CI* hierarchy. These role relationships are introduced at a drug class level or a drug preparation level. For example, the role relationship *has_Chemical_Structure* that describes the chemical structure of an FDA established pharmacologic class is introduced at a drug class level while the roles *has_Ingredient, CI_ChemClass*, and *has_active_metabolites* are introduced at a drug preparation level.

Another terminology that contains classifications of active ingredients of drugs according to different dimensions is the Anatomical Therapeutic Chemical (ATC) Classification System. The ATC classifies drugs into groups at five different levels, the anatomical main group at Level 1, the therapeutic main group at Level 2, the therapeutic/pharmacological subgroup at Level 3, the chemical/therapeutic/pharmacological subgroup at Level 4, and the chemical substance at Level 5. However, each drug is connected to the classifications of different dimensions. The difference between ATC and NDF-RT include that each classification level is presented as only one layer in the ATC and all levels are in one hierarchy. The NDF-RT has a separate whole hierarchy for each dimension.

2.2 Abstraction Networks and Terminology Auditing

It is difficult for users and even creators of a large biomedical terminology to get the "big picture" of the content of that terminology, due to the large number of concepts and roles. An *Abstraction Network* of a terminology [7] is a compact network, that consists of nodes and of links connecting the nodes, summarizing the terminology. Each node in the Abstraction Network summarizes a group of similar concepts from the terminology. The concept "grouping criterion" for a terminology is based on the internal structural characteristics of this terminology and is different for different terminologies. Nodes in the Abstraction Network are organized in a hierarchy by *child-of* links that are derived from the IS-A links of the terminology.

A well designed Abstraction Network can summarize a large terminology in a compact way, helping users and curators to get a "big picture" of its content. In our previous research, we have designed several kinds of Abstraction Networks for various terminologies and ontologies. We have shown that Abstraction Networks can be used to support terminology quality assurance by reducing the QA resources that need to be expended to uncover erroneous concepts.

For example, we utilized the *area taxonomy* and *partial-area taxonomy* [8, 16, 17] Abstraction Networks to enable QA of SNOMED CT [18], the National Cancer Institute thesaurus (NCIt) [19], and the Gene Ontology [20]. The *disjoint partial-area taxonomy* [21] and the *tribal abstraction network* [22] Abstraction Networks were also shown to support QA of SNOMED CT. We have used other abstraction networks [23-27] to support QA of the Ontology of Clinical Research (OCRe) [28], the Cancer Chemoprevention Ontology

(CanCo) [29], the Sleep Domain Ontology (SDO) [30], and the Drug Discovery Investigations Ontology [31].

2.3 NDF-RT Related Work

Extensive research has been reported on NDF-RT, e.g., on its content coverage, the adequacy of representation, drug normalization and classification, etc. Rosenbloom et al. [32] investigated the adequacy of representation in the *Physiologic Effect* hierarchy. Carter et al. [33] studied drug class names from three sources to understand how drugs were classified. They further evaluated NDF-RT's semantic coverage. Zhu et al. [34] normalized drug data in PharmGKB [35] by mapping extracted drugs and drug classes to NDF-RT. Pathak et al. [36] investigated drug-disease relationships in NDF-RT and PharmGKB to make both more robust and integratable. Pathak et al. [37] also evaluated the applicability of RxNorm [38] and NDF-RT for classification of medication data extracted from electronic health records.

3 Methods

In Section 3.1 we design the *Ingredient Abstraction Network* (IAbN) for the *CI* hierarchy of NDF-RT. The basic approach of this study was to use IAbN support to identify concepts belonging to concept groups with especially complicated structural properties, as explained in Section 3.3 below. Then we chose samples from those groups and from a control group and used a two-level audit by three expert auditors to count the numbers of errors as a function of structural complexity. We reviewed the errors reported by the expert auditors and showed that the complex concepts have statistically significantly higher error rate compared to the control concepts.

3.1 Ingredient Abstraction Network

An *Ingredient Abstraction Network* (IAbN) is an Abstraction Network where the nodes summarize (1) the ingredients in the *Chemical Ingredients* hierarchy and (2) those drug concepts in the *Pharmaceutical Preparations* hierarchy that have no dosage information but that do have at least one *has_Ingredient* role to a drug ingredient in the *Chemical Ingredients* hierarchy.

Drug ingredients are chemical ingredients that are used in prescription drugs. We define five categories of concepts in the *Chemical Ingredients* (*CI*) hierarchy. The right side of Figure 3(a) illustrates the following categories of drug concepts for an excerpt of 14 *CI* concepts.

> Definition 1: A *drug ingredient concept* is a concept in the *Chemical Ingredients (CI)* hierarchy that is the target of *has_Ingredient* role(s) from concepts in the *Pharmaceutical Preparation* hierarchy.
>
> Definition 2: A *classification ingredient concept* is a concept in *CI* that "organizes" other drug ingredient concepts below it. In other words, it has drug ingredient concepts as children. It may or may not be itself a target of a *has_Ingredient* role.
>
> Definition 3: A *dual ingredient concept* is both a drug ingredient concept and a classification ingredient concept in *CI*. Such a concept is a target of a *has_Ingredient* role and has children that are drug ingredient concepts.

> Definition 4: A *strict classification ingredient concept* is a classification ingredient concept that is not also a drug ingredient concept. That is, it is not a target of a *has_Ingredient* role.
>
> In other words, a *classification ingredient concept* is either a dual ingredient concept or a strict classification ingredient concept.
>
> Definition 5: An *uncategorized ingredient concept* is a concept in the *CI* hierarchy that is neither a drug ingredient concept nor a classification ingredient concept. Such concepts are not used in the theory developed in this paper.

The design of an Abstraction Network for the *CI* hierarchy poses a challenge for several reasons: (1) A lack of roles emanating from *CI* concepts prevents the derivation of a commonly used Abstraction Network called a partial-area taxonomy [8] that can be derived for many other description logic-based terminologies. (2) The need to distinguish between *drug ingredient concepts* and *classification ingredient concepts* is further complicated by the existence of *dual ingredient concepts*. (3) To obtain a "big picture" of the *Chemical Ingredients* hierarchy there is a need to summarize the drug concepts, which in NDF-RT are parts of the *PP* hierarchy, according to their ingredient concepts in *CI*, as was illustrated by Ochs et al. [39].

The derivation algorithm for an IAbN will now be explained. It begins with identifying all of the drug concepts in the *PP* hierarchy that have a *has_Ingredient* role but no *has_DoseForm* role. *PP* concepts with dosage information are ignored since an ancestor concept, typically a parent (a *PP* generic drug ingredient) introduces the *has_Ingredient* role, which is inherited to such concepts. Hence, there is no need for direct summarization of *PP* concepts with dosage information, since such a summary is offered indirectly through the summarization of the other *PP* concepts. All the *PP* concepts in Figure 3(a), except *Pharmaceutical Preparations*, have one *has_Ingredient* role to a concept in the *CI* hierarchy. Different drug concepts in the *PP* hierarchy can have a *has_Ingredient* role to the same *CI* concept, e.g., both *Aspirin* and *Acetylsalicylate Sodium* in Figure 3(a) have the ingredient *Aspirin. PP* concepts may also have multiple *has_Ingredient* roles, e.g., *Aspirin/Caffeine* has distinct *has_Ingredient* roles to both *Aspirin* and *Caffeine*.

In the next step, drug ingredient concepts (see Definition 1 above) are identified by collecting the target concepts of all the *has_Ingredient* roles. Classification ingredient concepts (see Definition 2 above) are identified by analyzing the parent concept(s) of each drug ingredient concept. Next, for each drug ingredient concept, the lowest ancestor(s) that are a strict classification ingredient concept(s) (see Definition 4 above) are identified, with the intention of finding groups of drug ingredient concepts. (Common ancestors will be used in the next step to define groups.)

For example, for the *Aspirin CI* concept, the lowest ancestor that is a strict classification ingredient concept is *Salicylates. Salicylates* is the lowest common ancestor for *Aspirin*, *Magnesium Salicylate*, and *Diflunisal*. For *Warfarin Sodium* its parent concept, *Warfarin*, is a classification ingredient concept but it is also a drug ingredient concept (i.e., it is a dual ingredient concept; see Definition 3 above). Thus, the lowest ancestor of *Warfarin Sodium* that is a strict classification ingredient concept is *Warfarin*'s parent, *4-Hydroxycoumarins*.

Many *CI* hierarchy concepts have multiple parents, thus, a given drug ingredient concept may have more than one lowest ancestor that is a strict classification ingredient concept.

In the next step of deriving the Abstraction Network, the drug ingredient concepts are grouped together according to their common ancestor(s) that are strict classification ingredient concepts. For example, *Aspirin*, *Magnesium Salicylate*, and *Diflunisal* share *Salicylates* as a lowest common ancestor. Similarly, *Warfarin*, *Warfarin Sodium* and *Phenprocoumon* share *4-Hydroxycoumarins* as a lowest common ancestor. Figure 3(b) models the right side of Figure 3(a) and shows the "drug ingredient groups" induced by the lowest common ancestors. Color coding in Figure 3(b) helps to keep the groups apart: Every group has its own color.

In the following step, each strict classification ingredient concept is made into a *root* for its ingredient group. Roots of a group are shown with solid fill in Figure 3(b). Thus *Salicylates* becomes the root of the group with *Aspirin*, *Magnesium Salicylate*, and *Diflunisal* in it. The *CI* root concept, *Chemical Ingredients*, is also a root. Roots represent groups of *CI* concepts in the IAbN (Figure 3(c)). The text line "3 Ingredients" under *Salicylates* in Figure 3(c) indicates how much information is summarized by this box. Ingredient groups are not disjoint; drug ingredient concepts with multiple parents may be summarized by multiple ingredient groups. With this step we have created a summary (Figure 3(c)) of the "right side" (the *Chemical Ingredients* hierarchy) of Figure 3(a). In the next step, we include information from the left (*PP*) side of Figure 3(a) into Figure 3(c).

For each ingredient group, the *PP* drug concepts that have a *has_Ingredient* role to a drug ingredient concept in the ingredient group are identified. For example, the *Aspirin* and *Acetylsalicylate Sodium* drug concepts in *PP* both have *Aspirin* in *CI* as the target of their *has_Ingredient* roles. The *Aspirin* drug ingredient concept belongs to the *Salicylates* ingredient group, thus, the *Aspirin* and *Acetylsalicylate Sodium* drug concepts from *PP* are also summarized by the *Salicylates* ingredient group. This is expressed by the text line "4 Drugs" under *Salicylates* in Figure 3(c). (The other two drug concepts are *Magnesium Salicylate* and *Diflunisal*). Since ingredients may belong to multiple ingredient groups, a given *PP* drug concept may be represented by multiple ingredient groups.

Within the IAbN, ingredient groups are organized into a hierarchy according to *child-of* links derived from the underlying IS-A hierarchy. An ingredient group **A** is a *child-of* another ingredient group **B** if **A**'s root has **B**'s root as an ancestor in the *CI* hierarchy and there are no other roots of the IAbN on any path from **A**'s root to **B**'s root in the *CI* hierarchy. An ingredient group may be a *child-of* multiple ingredient groups. In summary, Figure 3(c) shows the IAbN derived from NDF-RT excerpt in Figure 3(a).

In the visualization of an IAbN it is necessary to organize the ingredient groups in a way that helps the summary reflect the "big picture." Thus, ingredient groups *may be* organized into color coded levels according to the length of the longest *child-of* path to the root ingredient group (*Chemical Ingredients*). This will be shown later in Figure 4. Figure 3(c) does not use this color level encoding.

We note that *Ethyl Biscoumacetate* is an uncategorized ingredient concept (see Definition 5 above), as shown in Figure 3(a). This occurs when an ingredient is modeled in *CI* but no *PP* drug concept has a *has_Ingredient* role to this ingredient. For the current research, such concepts are not summarized by any ingredient group and are not considered part of the IAbN. In the Discussion Section we propose methods for extending the IAbN to include uncategorized ingredient concepts and drug concepts with dosage information in the summary.

3.2 Software for Creating and Browsing IAbNs

In previous work, we created the Ontology Abstraction Framework (OAF) [40], an open source software system and tool for deriving Abstraction Networks. As part of this research project, we created an IAbN module for the OAF, enabling the derivation of IAbNs for NDF-RT (and other terminologies). The OAF system can create interactive IAbN displays that can be browsed by a user. The OAF, which includes a reference implementation of the IAbN derivation in Java, is available at https://njitsaboc.github.io/.

3.3 Quality Assurance Techniques for Auditing the CI Hierarchy

In this paper, we perform a QA study of the *Chemical Ingredients* (*CI*) hierarchy of NDF-RT, which provides a chemistry-oriented classification of the drug ingredients. Such a classification is, for example, relevant to Drug-Drug Interactions (DDIs), since in many cases drugs that are chemically similar tend to have similar interactions [41]. As already mentioned in the Introduction, the *CI* hierarchy was imported into NDF-RT from MeSH.

The complication is that the *CI* hierarchy was designed with a chemistry orientation rather than a pharmacology orientation. As we will see, those two orientations do not always coincide, and NDF-RT is a drug terminology rather than a chemical terminology such as ChEBI [42]. As may be expected, such differences will play an important role when performing the quality assurance review of the *CI* hierarchy.

The Abstraction-Network-based framework we have developed can be summarized as follows. First, an Abstraction Network is developed to summarize the specific terminology [7]. An algorithm is described and implemented to computationally derive the Abstraction Network from the terminology. Based on the Abstraction Network, we identify characterizations of sets of concepts of the terminology that are expected to display a higher percentage of errors, compared to a control sample [24, 43]. Those sets of concepts can be computationally retrieved [40, 44], because the characterizations of such sets of concepts are based on structural features.

One of the recurring themes in such characterizations is that there are concepts that are more complex than “arbitrary” concepts of the terminology. Examples of characterizations of complex concepts include overlapping concepts [21, 45, 46] and multiple inheritance regions [8, 47]. Complex concepts are typically more error-prone. While those characterizations are based on deriving a Partial-Area Taxonomy [8, 16, 23] their complexity stems from concepts having multiple generalizations through multiple parents, reflecting an entity that is simultaneously “this and that.” Not surprisingly, the modeling of such concepts is more challenging and a higher ratio of errors can be expected for them.

The characterization of concepts that we are testing in this study on the NDF-RT *CI* hierarchy is "drug ingredients belonging to only one ingredient group with multiple parent ingredient groups" in the IAbN. Such concepts fit the above theme of complex concepts being "this and that" and are expected to have higher error rates.

Hypothesis 1: Among drug ingredients belonging to only one ingredient group, those in an ingredient group with multiple parent ingredient groups are more likely to have errors than those in an ingredient group with only one parent ingredient group.

The drug ingredients from those ingredient groups that have multiple *parent ingredient groups* inherit multiple classifications. The more classifications the drug ingredients belong to, the more complex those ingredients are, which increases the possibility that the classifications may have errors. We will formulate this as Hypothesis 2.

Hypothesis 2: Among drug ingredients belonging to only one ingredient group, those in an ingredient group with more than two parent ingredient groups are more likely to have errors than those with exactly two parent ingredient groups.

To test the above hypotheses, a sample of drug ingredient concepts within only one ingredient group was reviewed by two chemistry domain experts, coauthors LC and HY. Table 1 shows the distribution of NDF-RT's all drug ingredients appearing in exactly one ingredient group according to their group's number of parent ingredient groups. We picked 263 drug ingredients from the ingredient groups that have multiple parent ingredient groups as study concepts as follows. The study concepts included 118 randomly selected drug ingredients with two parent ingredient groups plus all drug ingredients with three (118), four (25) or five (2) parent ingredient groups. Thus, in total there were 263 study concepts. We randomly chose 170 drug ingredients from the ingredient groups that have only one parent ingredient group as control concepts, achieving statistical significance. Hence, the total number of reviewed drug ingredients in the study is 433.

LC and HY were blind to the hypotheses and the sampling methodology. The concepts were presented in alphabetical order. There were three steps of the review process. First, each of the reviewers studied the sample individually and submitted an error report that consisted of identified errors with corresponding corrections.

The domain experts were instructed to review the hierarchical relationships of each concept for correctness and to mark those they considered incorrect. The individual error reports from the domain experts were combined into a single anonymized list of unique errors. In the second step, the list of combined errors was sent back to the domain experts who had to obtain a consensus. Each reviewer marked 'agree' or 'disagree' for each error in the list.

In the third step, an additional evaluation of the consensus result was performed by JKU (a pharmacologist who is leading First DataBank's drug vocabulary standards initiatives). Only the errors agreed upon by both LC and HY were sent to JKU for the third round review. JKU recorded those concepts for which she agreed that there was an error in a hierarchical relationship. Thus, in this study a concept was considered erroneous only if all three domain experts (LC, HY, and JKU) agreed on the error.

If the above hypotheses are confirmed, they will guide the focus of the NDF-RT team to sets of concepts that are more likely to have errors. Considering the limited resources typically available for terminology QA, this is important, because no comprehensive QA effort for a terminology as the size of the NDF-RT is likely to be budgeted. Thus, the IAbN approach will allow NDF-RT's curators to achieve a higher error yield, as measured by the ratio of the number of errors corrected to the number of concepts reviewed.

4 Results

4.1 Abstraction Networks

We derived an IAbN for the June 2015 inferred version of NDF-RT's *Chemical Ingredients* (*CI*) hierarchy, which consists of 10,145 concepts. This IAbN consists of 860 ingredient groups which summarize 2,664 drug ingredients and 6,872 *Pharmaceutical Preparation* hierarchy drug concepts. We define the *abstraction ratio* of the IAbN to be the average number of drug ingredients per ingredient group. The abstraction ratio of the June 2015 IAbN is 3.07 (=2,664/860). There are 813 drug ingredient concepts summarized by more than one ingredient group (total 535 ingredient groups), and each such drug ingredient is summarized by an average of 1.52 (=813/535) ingredient groups. The average number of *PP* drug concepts summarized by each ingredient group is 7.99.

Figure 4 shows an excerpt of 128 of the IAbN's ingredient groups, as the IAbN is too large to fit on a single page. (In the Discussion Section we will mention methods to overcome this issue.) By reviewing the ingredient groups of the IAbN one can see the major types of drug ingredients used in NDF-RT's drugs. For example, the *Polymers* group (Level 2: green) summarizes 26 ingredients and 81 drugs, *Piperidines* (Level 3: blue) summarizes 47 ingredients and 76 drugs, *Tetracyclines* summarizes 17 ingredients and 35 drugs, *Ethanolamines* summarizes 45 ingredients and 231 drugs, and *Penicillins* summarizes 34 ingredients and 64 drugs.

4.2 Quality Assurance

As noted above, the IAbN has 860 ingredient groups, which summarize 2,664 drug ingredients. Each of these drug ingredients may be summarized by either exactly one ingredient group or multiple ingredient groups. There are 1,851 drug ingredients appearing in exactly one ingredient group (total 597 ingredient groups) and 813 drug ingredients appearing in multiple ingredient groups (total 535 ingredient groups). Note that this is a partition of the drug ingredients into two disjoint sets (1,851+813=2,664), but not a partition of the 860 ingredient groups into disjoint sets (597+535≠860). The reason for the latter is that the same group may contain some ingredients belonging solely to this group and some ingredients belonging to multiple groups.

In this study we concentrated on the 1,851 drug ingredients appearing in exactly one ingredient group. From these 1,851 drug ingredients, a total of 433 drug ingredients were included in this study (see Table 2). The respective numbers of drug ingredients with one, two, three, four, and five parent ingredient group(s) are 170, 118, 118, 25, and two. The sample of 433 drug ingredients and their paths to the root concept *Chemical Ingredients* were presented to auditors LC and HY (domain experts and chemistry professors). They focused on auditing the hierarchical relationships of a drug ingredient concept and the hierarchical relationships of all of its ancestors. The complete hierarchical relationships for each concept were reviewed because some errors may occur due to transitivity between ingredient classifications, which may be further apart in the hierarchy, as illustrated in the following section.

In the two initial auditing reports, the two reviewers agreed on 100 erroneous drug ingredients; 19 drug ingredients were judged as erroneous by one reviewer. We compiled a new data set including all the errors reported by any of the two auditors, without the name of the originator of the error, and sent this new data set back to the two auditors for generating a consensus report. The two auditors gave their responses (agree or do not agree) to all the errors listed in the new dataset, which were compiled into a consensus report including 119 errors that both reviewers agreed to. Then the consensus report was reviewed by coauthor JKU. Only when an error of a concept listed in the consensus report of LC and HY was confirmed by JKU, then this concept was labeled “erroneous,” i.e., a consensus of three reviewers was achieved for these concepts in this two-stage study. In fact, JKU confirmed all consensus errors reported by LC and HY. Appendix 1 lists all of the errors confirmed in the consensus study.

Table 2 shows the error distribution of the 433 audited drug ingredients. We can see that the percentage of erroneous concepts increases with the number of parent ingredient groups (except for the small number with five parents). Column 5 shows the percentage of erroneous concepts per parent ingredient group. The values in Column 5 are around 13% per parent ingredient group.[1]

However, one cannot interpret this result assuming independence of the errors for each parent ingredient group. In some cases, the errors of a concept are classification errors occurring at an ancestor level. Several such errors may occur at ancestors of the same parent ingredient group, for example, *Ertapenem* in Figure 5 is in the *Beta-Lactams* ingredient group with four parent ingredient groups *Heterocyclic Compounds, Amides, Sulfur Compounds* and *Azabicyclo Compounds.* (Only the latter two groups, related to errors, are shown in the figure.) The classification errors reported for *Ertapenem* are transitively derived only from the parent ingredient group *Sulfur Compounds* and its parent *Inorganic Chemicals*.

Table 3 shows the contingency table for the control and study concepts to calculate the P value for Hypothesis 1. The two-tailed P value is less than 0.0001 by Fisher’s exact test [48], which means that the drug ingredients from the ingredient groups that have multiple parent ingredient groups are statistically significantly more likely to have errors than those from the ingredient groups that have one parent ingredient group.

In order to test Hypothesis 2, we compared the error counts of drug ingredients from the ingredient groups that have more than two parent ingredient groups with those that have

[1] We thank an anonymous reviewer for this observation.

exactly two parent ingredient groups. Table 4 shows the contingency table for the concepts with two and more than two parent ingredient groups to calculate the P value. The two-tailed P value equals 0.0002 by Fisher's exact test, which means that Hypothesis 2 was confirmed.

Overall, there are 119 concepts (119/433 =27.5%) with errors. Some errors appear at the parent level, while other errors were introduced at higher levels (up to several levels above the erroneous concept), which will be discussed in the end of the Results Section. The types of errors are summarized in Table 5. The sets of erroneous concepts for the different error types in Table 5 are not disjoint, since one concept may have multiple errors. Row 1 in Table 5 shows that eight concepts in this study are assigned wrong parents, e.g., *Loracarbef* was erroneously defined as child of *Cephalosporins,* while the direct parent of *Loracarbef* should be *Carbacephem*.

We identified two especially prevalent error patterns in this study. Row 2 and Row 3 show these two most common errors in the study that cover most of the erroneous concepts. In Row 2 an organic (or inorganic) concept is assigned to both *Organic Chemicals* and *Inorganic Chemicals* due to the inheritance from its ancestor classification, which can be either organic or inorganic. For example, *Sulfur Compounds* appear as inorganic or organic compounds that contain sulfur as an integral part of the molecule according to the definition. *Rabeprazole* actually is an organic chemical while it is assigned under both *Organic Chemicals* and *Inorganic Chemicals* because the parent of its grandparent is *Sulfur Compounds.* Row 3 indicates that the specified chemical ring structures of a concept which is a *Heterocyclic Compound[s]* are contradicting each other. For example, a concept is assigned several classifications out of the set *R* = {"*Heterocyclic Compounds, 1-Ring,*" "*Heterocyclic Compounds, 2-Ring,*" "*Heterocyclic Compounds, 3-Ring*" and "*Heterocyclic Compounds with 4 or More Rings*"}. That is due to inheritance from its ancestor classifications, i.e., a concept may have several ancestor classifications (at a very general level) and each of its ancestor classifications may be under one of the four choices in *R*. For example, *Alosetron* is a 4-ring structure with three fused rings. Its parent is *Carbolines* with a 3-ring structure, the parents of which are, *Pyridines* with a 1-ring structure and *Indoles* with a 2-ring structure. Hence due to the transitivity, *Alosetron* is a *Heterocyclic Compounds, 1-Ring*, and also a *Heterocyclic Compounds, 2-Ring*, and even a *Heterocyclic Compounds, 3-Ring*.

Row 4 represents the other types of erroneous classifications, happening above the parent level, which cover 26 concepts (21.8%). For example, *Hydrogen Peroxide* does not belong to *Electrolytes* because it is a molecule without ions, and it is not an electrolyte.

We will start by discussing one specific example about Organic/Inorganic Chemical misclassification in detail, to illustrate the encountered modeling problems and potential remodeling options to overcome these problems.

As an example, in Figure 5, *Ertapenem* was found to be misclassified. It should be classified under *Carbapenems*, which could be solved easily. However, there are additional subtle problems in Figure 5, not expressed by a wrong parent, but implied by a sequence of relationships that, by themselves, are correct.

1. *Beta-Lactams* has indeed a sulfur atom. However, in pharmacology it is not referred to as a "sulfur compound." Thus, the link from *Beta-Lactams* to *Sulfur Compounds* is debatable.

2. *Beta-Lactams* is definitely not considered an inorganic chemical. However, the link from *Beta-Lactams* to *Sulfur Compounds* together with the link from *Sulfur Compounds* to *Inorganic Chemicals* leads to a wrong conclusion by transitivity, again making the link from *Beta-Lactams* to *Sulfur Compounds* problematic. The same wrong conclusion would be drawn with respect to *Ertapenem*.

3. The connection from *Beta-Lactams* to *Lactams* is correct. However, the class *Lactams* is of questionable use in pharmacology. A high-quality pharmacology ontology should not contain classes for which there is no pharmacologic use case. It is hard to imagine a pharmacologic situation in which a user would query or search for *Lactams*. This observation contradicts the "idealistic content notion of an ontology" [49] that could be expressed as "ontologies contain domain truths with no regard to the use and usefulness of these truths." An ontology that is "cluttered up" with too many truths that are not useful will be held in low regard by its users.

Problem 3 above can be solved by omitting *Lactams* completely. Problems 1 – 2 above can be solved by introducing a new concept *Organic Sulfur Compounds* instead of the concept *Sulfur Compounds.* The latter is changed to *Inorganic Sulfur Compounds.* Figure 6 reflects these changes and solves the problems listed above. Problem 1 is solved because there is no more link from *Beta-Lactams* to (what used to be) *Sulfur Compounds* and is now *Inorganic Sulfur Compounds*. Problem 2 is solved because there is no more path from *Beta-Lactams* to *Inorganic Chemicals*.

As a more general conclusion to be drawn from this example we note that whenever a pharmacology ontology is audited, it is essential *not to just look for pairs of concepts and the connection of each pair, but to start with a drug and follow the whole path towards the root to determine whether every assignment implied by transitivity is actually correct from the pharmacologic point of view*.

Figure 5 is also "strange" from a knowledge representation point of view. Figure 5 reads as "Every instance of Sulfur Compund(s) is an instance of Organic Chemical(s)" and "Every instance of Sulfur Compund(s) is an instance of Inorganic Chemical(s)." However, chemicals are assumed to be either organic or inorganic, but not both. Thus, independent of pharmacological considerations, Figure 6 also solves a formal knowledge representation issue of Figure 5.

5 Discussion

The development of the IAbN (Ingredient Abstraction Network) enables the summarization of NDF-RT's drug concepts according to their various classifications. The IAbN makes it possible to compactly visualize the major types of ingredients, and the drugs that contain them, as they currently exist in NDF-RT. One reason that the IAbN derivation methodology works well for the *CI* hierarchy is that the majority of drug ingredient concepts (refer back to Definition 1) in NDF-RT are leaves (i.e., have no children) (see Table 6).

Not every distinction between more complex concepts and less complex concepts necessarily indicates a higher rate of errors in the former group. In the IAbN there are 1,851 concepts that appear in exactly one ingredient group in 597 such ingredient groups. In contrast, there are 813 concepts that appear in multiple ingredient groups in 535 such ingredient groups. It would seem that concepts that appear in multiple ingredient groups are more complex due to the duality of their groupings and therefore have higher error rates. However, in an auxiliary study we found that the error rates of concepts in these two sets were indistinguishable. In hindsight, the errors we reported for the *Chemical Ingredients* hierarchy of NDF-RT are mostly between ingredient groups rather than between an ingredient drug and its group. Hence, the distinction that matters for the error rate is the one we reported in this study, namely between ingredient groups with one parent ingredient group and ingredient groups with multiple parent ingredient groups.

5.1 Limitations

One limitation of the IAbN-based QA methodology is the amount of initial manual effort required to evaluate the effectiveness of the methodology. Two domain experts had to manually review several hundred NDF-RT concepts and another round of consensus building was required to confirm the results, involving a third domain expert. However, once the initial evaluation study was completed, and the methodology was shown to identify sets of concepts that are statistically significantly more erroneous than other concepts, it can then be applied to other concepts. Additionally, if the methodology is applied by a curator of NDF-RT, it may reduce the overall amount of effort, as he or she may be familiar with the modeling of the concepts.

As stated above, the aim of our work is to provide a compact summary of the NDF-RT. While the IAbN is significantly smaller than the *CI* hierarchy that it summarizes, (860 nodes versus 10,145 concepts), it is a limitation of this work that the IAbN is still large and the size impedes a useful visualization. In previous work, Ochs et al. [39] introduced the Aggregate IAbN, which provides a *secondary summarization mechanism*. This mechanism gives the user a certain degree of control over the size of the final Abstraction Network. However, for the purpose of QA of the NDF-RT *CI* hierarchy, we are not using the Aggregate IAbN, and thus it is not relevant to the current paper.

A third limitation is that not every concept in the *CI* hierarchy is summarized by an ingredient group. In our previous Abstraction Networks, each concept was always summarized by at least one Abstraction Network node [7]. However, if a *CI* hierarchy concept is an *uncategorized ingredient concept* (Definition 5), e.g., *Ethyl Biscoumacetate* in Figure 3, it is not reflected at all in the IAbN. In fact, the majority (7,481; 73.7%) of concepts in the *CI* hierarchy are not summarized by any ingredient group. This situation occurs for several reasons. The *CI* hierarchy was primarily imported from MeSH and many of the concepts from MeSH are very general and do not represent ingredients that are used in drugs. Other ingredient concepts may not be relevant to the drugs in NDF-RT, as no drug in the *PP* hierarchy of NDF-RT includes them as an ingredient.

Lastly, we note that drug concepts with dosage information are not associated with any ingredient group, i.e., not summarized. If the need arises, these drug concepts can also be summarized by the implied ingredient group(s).

5.2 Future Work

There are several ways to improve the effectiveness of our methodology. As shown in the Results section, errors typically occurred at a general level (i.e., close to the root concept *Chemical Ingredients*). If these general concepts are audited first then the corrections to errors would propagate down the hierarchy. In a future study, we will review these general concepts when they are summarized by only one ingredient group that has multiple parent ingredient groups. In a recent study [50] we showed that a similar methodology, based on a different Abstraction Network, was effective for the NCIt's *Neoplasm* subhierarchy. We will also evaluate the effectiveness of "group-based auditing" [50] in the *Chemical Ingredients* hierarchy. In group-based auditing, when a single concept in an ingredient group is found to be erroneous, all other concepts in the ingredient group will also be reviewed for errors. The other concepts in the ingredient group may have the same, or similar, errors. Hence, more errors could be uncovered with minimal extra work.

The IAbN derivation approach is applicable beyond the *CI* hierarchy. For example, it is possible to apply the IAbN derivation methodology to the *Cellular or Molecular Interactions* (MOA) hierarchy, summarizing NDF-RT's drugs according to their mechanisms of action, rather than their chemical ingredients. In a future study, we will investigate the structural properties of IAbNs derived from NDF-RT's other classification hierarchies.

In a future study, we will investigate ways of summarizing uncategorized ingredient concepts (see limitations above). One idea is to associate each such concept with the closest classification ingredient above it. Using such an approach, *Ethyl Biscoumacetate*, for example, would be summarized by the *4-Hydroxycoumarins* ingredient group. An IAbN with drug concepts including dosage information is a potential extension of this work.

In Summer 2016, the VA announced plans to replace NDF-RT by the Medication Reference Terminology (MED-RT) [51]. The various classification hierarchies will be preserved in different ways. For example, instead of importing the *Chemical Ingredients* (*CI*) hierarchy from MeSH, as in NDF-RT, the Chemical Ingredient roles from the *PP* hierarchy of MED-RT will point to the corresponding concepts in MeSH. This change is in line with the trend of referring to concepts in the hierarchies of other terminologies (e.g., GO Plus [52, 53] has relationships to classes in external ontologies, e.g., the ChEBI ontology).

The transition to MED-RT creates challenges for correcting errors in *CI*, since the corrections would need to be applied to MeSH. This is made difficult by the structure, orientation and ownership of MeSH: MeSH uses hierarchical relationships that do not always conform to the ontological rules for IS-A links. MeSH was not designed with the pharmacology viewpoint in mind. The purpose of MeSH is to support the categorization of scientific papers in PubMed and, in the case of the *CI* component, the categorization of chemistry-oriented publications. Hence, errors reported in this study for *CI* may not be considered errors in MeSH. Lastly, MeSH is not owned by the Veteran's Health Administration, which creates additional barriers to implementing corrections.

The preferred solution would be for the VA to create a pharmacology-oriented hierarchy of chemical ingredients and their classifications for MED-RT. Given that, as mentioned above, 7481 (73.1%) of the *CI* hierarchy concepts are uncategorized in the IAbN, since no drug contains them as a chemical ingredient, it appears that MeSH, with its chemistry orientation, is improper for representing chemical ingredients in a drug terminology. The design of such a new hierarchy, which will likely include a significantly smaller number of relevant chemicals, should involve domain experts (pharmacologists), who would provide the needed pharmacology viewpoint for such a *CI* hierarchy within MED-RT.

We are currently preparing to submit our error report to the NDF-RT editorial team for feedback. We plan to coordinate with the NDF-RT team in future studies to improve the modeling of NDF-RT using IAbN-based auditing techniques.

6 Conclusions

In this paper, we introduced the application of the Ingredient Abstraction Network (IAbN) to summarize the concepts in NDF-RT's *Chemical Ingredients* hierarchy. The IAbN summary was used for visualizing the contents of the NDF-RT and for quality assurance of NDF-RT's chemical ingredient concepts. The IAbN was shown to highlight sets of concepts that are more likely to have errors. The most common errors involved incorrect direct classifications (= wrong parents), erroneous organic/inorganic classifications and *Heterocyclic Compounds* with rings: Concepts were erroneously assigned to several different, contradictory atomic ring structures. Several concepts that were correct and even correctly placed were proposed for elimination from NDF-RT, because they are not useful from the point of view of pharmacology experts.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Michael Lincoln of the VHA and Mark Erlbaum of Apelon for sharing with us their insights into NDF-RT structure and its history. Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Number R01CA190779. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Elhanan G, Perl Y, Geller J. A survey of SNOMED CT direct users, 2010: impressions and preferences regarding content and quality. J Am Med Inform Assoc. 2011; 18(Suppl 1):i36–44. [PubMed: 21836159]
2. Zhu X, Fan JW, Baorto DM, Weng C, Cimino JJ. A review of auditing methods applied to the content of controlled biomedical terminologies. J Biomed Inform. 2009; 42(3):413–25. [PubMed: 19285571]
3. Geller J, Perl Y, Halper M, Cornet R. Special issue on auditing of terminologies. J Biomed Inform. 2009; 42(3):407–11. [PubMed: 19465342]

4. National Drug File - Reference Terminology (NDF-RT). Available from: http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/NDFRT/
5. National Drug File (NDF). Available from: https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/VANDF/
6. Structural Analysis of Biomedical Ontologies Center (SABOC). Available from: http://saboc.njit.edu/
7. Halper M, Gu H, Perl Y, Ochs C. Abstraction networks for terminologies: Supporting management of "big knowledge". Artif Intell Med. 2015; 64(1):1–16. [PubMed: 25890687]
8. Wang Y, Halper M, Min H, Perl Y, Chen Y, Spackman KA. Structural methodologies for auditing SNOMED. J Biomed Inform. 2007; 40(5):561–81. [PubMed: 17276736]
9. U.S. Department of Veterans Affairs, Veterans Health Administration. National Drug File – Reference Terminology (NDF-RT) Documentation February 2015 Version . Available from: http://evs.nci.nih.gov/ftp1/NDFRT
10. Medical Subject Headings (MeSH). Available from: http://www.ncbi.nlm.nih.gov/mesh
11. Chute CG, Carter JS, Tuttle MS, Haber M, Brown SH. Integrating pharmacokinetics knowledge into a drug ontology: as an extension to support pharmacogenomics. AMIA Annu Symp Proc. 2003:170–4. [PubMed: 14728156]
12. PubMed. Available from: http://www.ncbi.nlm.nih.gov/pubmed/
13. Federal Medication Terminologies. Available from: https://www.cancer.gov/research/resources/terminology/fmt
14. National Drug File - Reference Terminology on BioPortal. Available from: http://purl.bioontology.org/ontology/NDFRT
15. Carter JS, Brown SH, Erlbaum MS, Gregg W, Elkin PL, Speroff T, et al. Initializing the VA medication reference terminology using UMLS metathesaurus co-occurrences. AMIA Annu Symp Proc. 2002:116–20.
16. Min H, Perl Y, Chen Y, Halper M, Geller J, Wang Y. Auditing as part of the terminology design life cycle. J Am Med Inform Assoc. 2006; 13(6):676–90. [PubMed: 16929044]
17. Ochs, C., Perl, Y., Halper, M., Geller, J., Lomax, J. Gene ontology summarization to support visualization and quality assurance; Proceedings of the 7th International Conference on Bioinformatics and Computational Biology; 2015. p. 167-74.
18. Stearns MQ, Price C, Spackman KA, Wang AY. SNOMED clinical terms: overview of the development process and project status. AMIA Annu Symp Proc. 2001:662–6.
19. Sioutos N, de Coronado S, Haber MW, Hartel FW, Shaiu WL, Wright LW. NCI Thesaurus: a semantic model integrating cancer-related clinical and molecular information. J Biomed Inform. 2007; 40(1):30–43. [PubMed: 16697710]
20. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000; 25(1):25–9. [PubMed: 10802651]
21. Wang Y, Halper M, Wei D, Perl Y, Geller J. Abstraction of complex concepts with a refined partial-area taxonomy of SNOMED. J Biomed Inform. 2012; 45(1):15–29. [PubMed: 21878396]
22. Ochs C, Geller J, Perl Y, Chen Y, Agrawal A, Case JT, et al. A tribal abstraction network for SNOMED CT target hierarchies without attribute relationships. J Am Med Inform Assoc. 2015; 22(3):628–39. [PubMed: 25332354]
23. Ochs C, Agrawal A, Perl Y, Halper M, Tu SW, Carini S, et al. Deriving an abstraction network to support quality assurance in OCRe. AMIA Annu Symp Proc. 2012:681–9. [PubMed: 23304341]
24. He Z, Ochs C, Agrawal A, Perl Y, Zeginis D, Tarabanis K, et al. A family-based framework for supporting quality assurance of biomedical ontologies in BioPortal. AMIA Annu Symp Proc. 2013:581–90. [PubMed: 24551360]
25. Ochs, C., He, Z., Perl, Y., Arabandi, S., Halper, M., Geller, J. Choosing the granularity of abstraction networks for orientation and quality assurance of the Sleep Domain Ontology; Proceedings of the 4th International Conference on Biomedical Ontology; 2013. p. 84-9.

26. He, Z., Ochs, C., Soldatova, L., Perl, Y., Arabandi, S., Geller, J. Auditing redundant import in reuse of a top level ontology for the Drug Discovery Investigations Ontology; Proceedings of the 2013 International Workshop on Vaccine and Drug Ontology Studies; 2013.

27. Ochs C, Perl Y, Geller J, Haendel M, Brush M, Arabandi S, et al. Summarizing and visualizing structural changes during the evolution of biomedical ontologies using a Diff Abstraction Network. J Biomed Inform. 2015; 56:127–44. [PubMed: 26048076]

28. Sim I, Tu SW, Carini S, Lehmann HP, Pollock BH, Peleg M, et al. The Ontology of Clinical Research (OCRe): an informatics foundation for the science of clinical research. J Biomed Inform. 2014; 52:78–91. [PubMed: 24239612]

29. Zeginis D, Hasnain A, Loutas N, Deus HF, Fox R, Tarabanis K. A collaborative methodology for developing a semantic model for interlinking Cancer Chemoprevention linked-data sources. Semantic Web. 2014; 5(2):127–42.

30. Arabandi S, Ogbuji C, Redline S, Chervin R, Boero J, Benca R, et al. Developing a Sleep Domain Ontology. AMIA Clinical Research Informatics Summit. 2010

31. Qi D, King RD, Hopkins AL, Bickerton GR, Soldatova LN. An ontology for description of drug discovery investigations. J Integr Bioinform. 2010; 7(3)

32. Rosenbloom ST, Awad J, Speroff T, Elkin PL, Rothman R, Spickard A, et al. Adequacy of representation of the National Drug File Reference Terminology Physiologic Effects reference hierarchy for commonly prescribed medications. AMIA Annu Symp Proc. 2003:569–78. [PubMed: 14728237]

33. Carter JS, Brown SH, Bauer BA, Elkin PL, Erlbaum MS, Froehling DA, et al. Categorical information in pharmaceutical terminologies. AMIA Annu Symp Proc. 2006:116–20. [PubMed: 17238314]

34. Zhu Q, Freimuth RR, Pathak J, Chute CG. PharmGKB Drug Data Normalization with NDF-RT. AMIA Jt Summits Transl Sci Proc. 2013:180. [PubMed: 24303334]

35. Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, et al. Pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther. 2012; 92(4):414–7. [PubMed: 22992668]

36. Pathak J, Weiss LC, Durski MJ, Zhu Q, Freimuth RR, Chute CG. Integrating VA's NDF-RT drug terminology with PharmGKB: preliminary results. Pacific Symposium on Biocomputing. 2012:400–9. [PubMed: 22174295]

37. Pathak J, Murphy SP, Willaert BN, Kremers HM, Yawn BP, Rocca WA, et al. Using RxNorm and NDF-RT to classify medication data extracted from electronic health records: experiences from the Rochester Epidemiology Project. AMIA Annu Symp Proc. 2011:1089–98. [PubMed: 22195170]

38. Liu S, Ma W, Moore R, Ganesan V, Nelson S. RxNorm: Prescription for Electronic Drug Information Exchange. IT Professional. 2005; 7(5):17–23.

39. Ochs C, Zheng L, Gu H, Perl Y, Geller J, Kapusnik-Uner J, et al. Drug-drug interaction discovery using abstraction networks for "National Drug File - Reference Terminology" Chemical Ingredients. AMIA Annu Symp Proc. 2015:973–82. [PubMed: 26958234]

40. Ochs C, Geller J, Perl Y, Musen MA. A unified software framework for deriving, visualizing, and exploring abstraction networks for ontologies. J Biomed Inform. 2016; 62:90–105. [PubMed: 27345947]

41. Blumenthal, DK., Garrison, JC. Pharmacodynamics: Molecular Mechanisms of Drug Action. In: Brunton, LL.Chabner, BA., Knollmann, BC., editors. Goodman and Gilman's The Pharmacological Basis of Therapeutics. 12. McGraw-Hill Education; 2011.

42. ChEBI. Available from: https://www.ebi.ac.uk/chebi/

43. Ochs C, He Z, Zheng L, Geller J, Perl Y, Hripcsak G, et al. Utilizing a structural meta-ontology for family-based quality assurance of the BioPortal ontologies. J Biomed Inform. 2016; 61:63–76. [PubMed: 26988001]

44. Geller J, Ochs C, Perl Y, Xu J. New abstraction networks and a new visualization tool in support of auditing the SNOMED CT content. AMIA Annu Symp Proc. 2012:237–46. [PubMed: 23304293]

45. Ochs C, Geller J, Perl Y, Chen Y, Xu J, Min H, et al. Scalable quality assurance for large SNOMED CT hierarchies using subject-based subtaxonomies. J Am Med Inform Assoc. 2015; 22(3):507–18. [PubMed: 25336594]

46. Wang Y, Halper M, Wei D, Gu H, Perl Y, Xu J, et al. Auditing complex concepts of SNOMED using a refined hierarchical abstraction network. J Biomed Inform. 2012; 45(1):1–14. [PubMed: 21907827]

47. Halper M, Wang Y, Min H, Chen Y, Hripcsak G, Perl Y, et al. Analysis of error concentrations in SNOMED. AMIA Annu Symp Proc. 2007:314–8. [PubMed: 18693849]

48. Good, PI. Permutation, Parametric, and Bootstrap Tests of Hypotheses: A Practical Guide to Resampling. 3. Springer; New York: 2005.

49. Schneider L. Foundational ontologies and the realist bias. KI workshop on reference ontologies and application ontologies. Sep 16.2003

50. Zheng L, Min H, Chen Y, Xu J, Geller J, Perl Y. Auditing National Cancer Institute thesaurus neoplasm concepts in groups of high error concentration. Applied Ontology. 2017; 12(2):113–130.

51. Introduction to MED-RT as the Replacement for NDF-RT. Available from: https://evs.nci.nih.gov/ftp1/NDF-RT/Introduction%20to%20MED-RT.pdf

52. Gene Ontology Consortium. The Gene Ontology in 2010: extensions and refinements. Nucleic Acids Res. 2010; 38(Database issue):D331–5. [PubMed: 19920128]

53. go-plus. Available from: http://geneontology.org/page/download-ontology

Highlights

- A new Ingredient Abstraction Network to summarize NDF-RT's Chemical Ingredients
- A new quality assurance methodology based on the Ingredient Abstraction Network
- Concepts in chemical ingredient groups with multiple parent groups have more errors
- Complex chemical ingredient groups have a higher error rate
- Two common patterns of errors were found for chemical ingredient concepts in NDF-RT

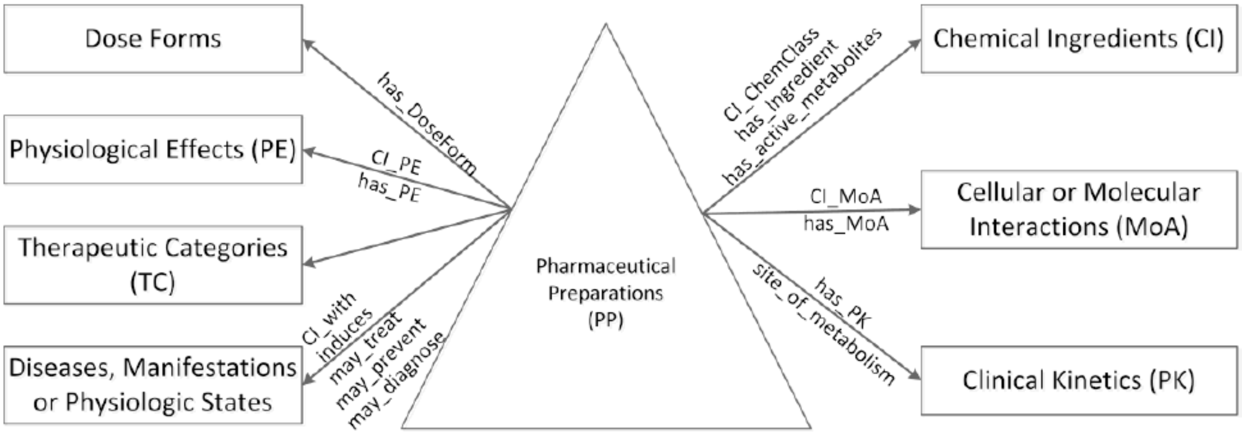

Dose Forms
Physiological Effects (PE)
Therapeutic Categories (TC)
Diseases, Manifestations or Physiologic States
has_DoseForm
CI_PE
has_PE
CI_with
induces
may_treat
may_prevent
may_diagnose
Pharmaceutical Preparations (PP)
CI_ChemClass
has_ingredient
has_active_metabolites
Chemical Ingredients (CI)
CI_MoA
has_MoA
Cellular or Molecular Interactions (MoA)
site_of_metabolism
has_PK
Clinical Kinetics (PK)

Figure 1.

Content Model of NDF-RT [9] (The “CI” in role names means contraindicated, not Chemical Ingredient)

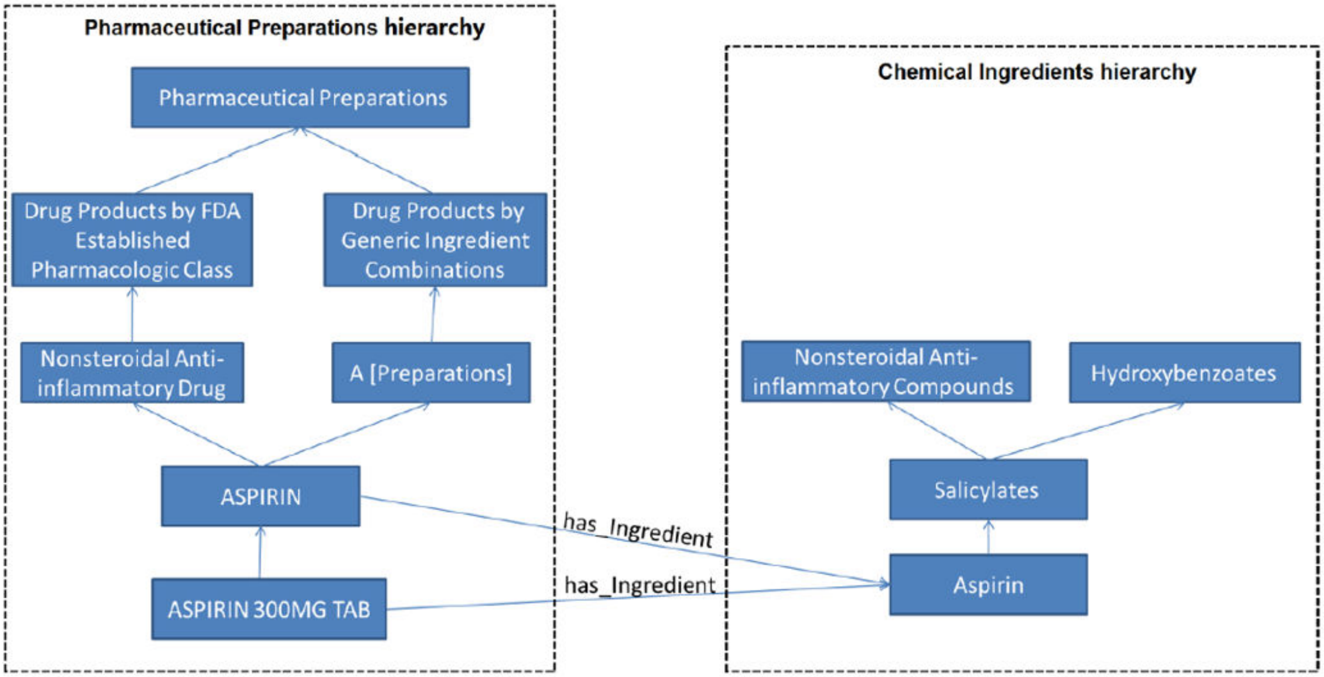

Pharmaceutical Preparations hierarchy
Pharmaceutical Preparations
Drug Products by FDA Established Pharmacologic Class
Drug Products by Generic Ingredient Combinations
Nonsteroidal Anti-inflammatory Drug
A [Preparations]
ASPIRIN
ASPIRIN 300MG TAB
has_Ingredient
has_Ingredient
Chemical Ingredients hierarchy
Nonsteroidal Anti-inflammatory Compounds
Hydroxybenzoates
Salicylates
Aspirin

Figure 2.

An excerpt from NDF-RT's *Pharmaceutical Preparations* and *Chemical Ingredients* hierarchies. Concepts are shown as blue boxes and hierarchical relationships are shown as upward directed blue arrows. The *has_Ingredient* roles linking the concepts in the two hierarchies are shown as labeled blue arrows.

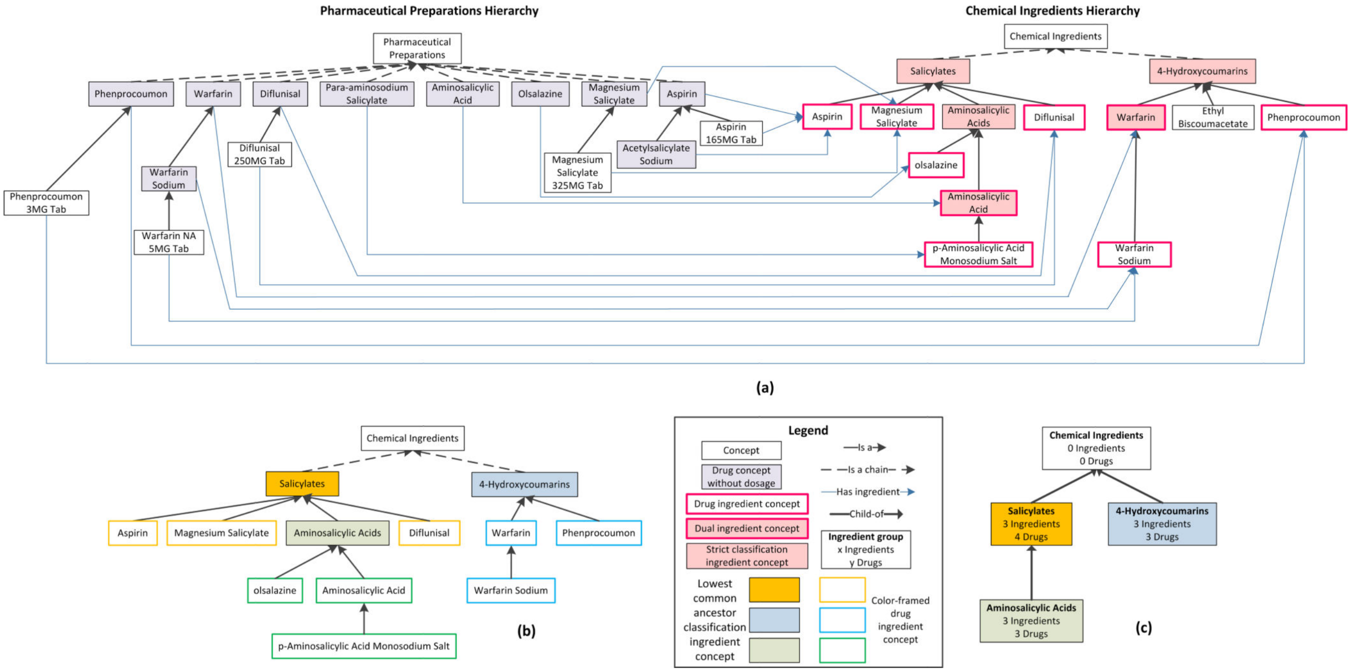


Figure 3.

(a) An excerpt of concepts from NDF-RT's *Pharmaceutical Preparations* (*PP*) and *Chemical Ingredients* (*CI*) hierarchies. On the left, drug concepts in the *PP* hierarchy with no dosage information have a shaded background. On the right, nine drug ingredient concepts have red borders and five classification ingredient concepts have a pink background. Two concepts, *Aminosalicylic Acid* and *Warfarin*, are both drug ingredient concepts and classification ingredient concepts, i.e., they are dual ingredient concepts. *Ethyl Biscoumacetate* is neither a drug ingredient concept nor a classification ingredient concept, i.e., it is an uncategorized ingredient concept. **(b)** *CI* grouped. Drug ingredient concepts are not shaded and their lowest common ancestor classification ingredient concepts are shaded. Each drug ingredient concept is color-framed according to its lowest common ancestor classification ingredient concept. **(c)** The IAbN for Figure 3(a). Ingredient groups are shown as boxes that are labeled with the name of the lowest common ancestor from 3(b). In each box are the total number of ingredient concepts summarized by the group, and the total number of drug concepts (without dosage information!) with *has_Ingredient* roles pointing to the *CI* hierarchy. *Child-of* links between ingredient groups are shown as upward directed bold arrows.

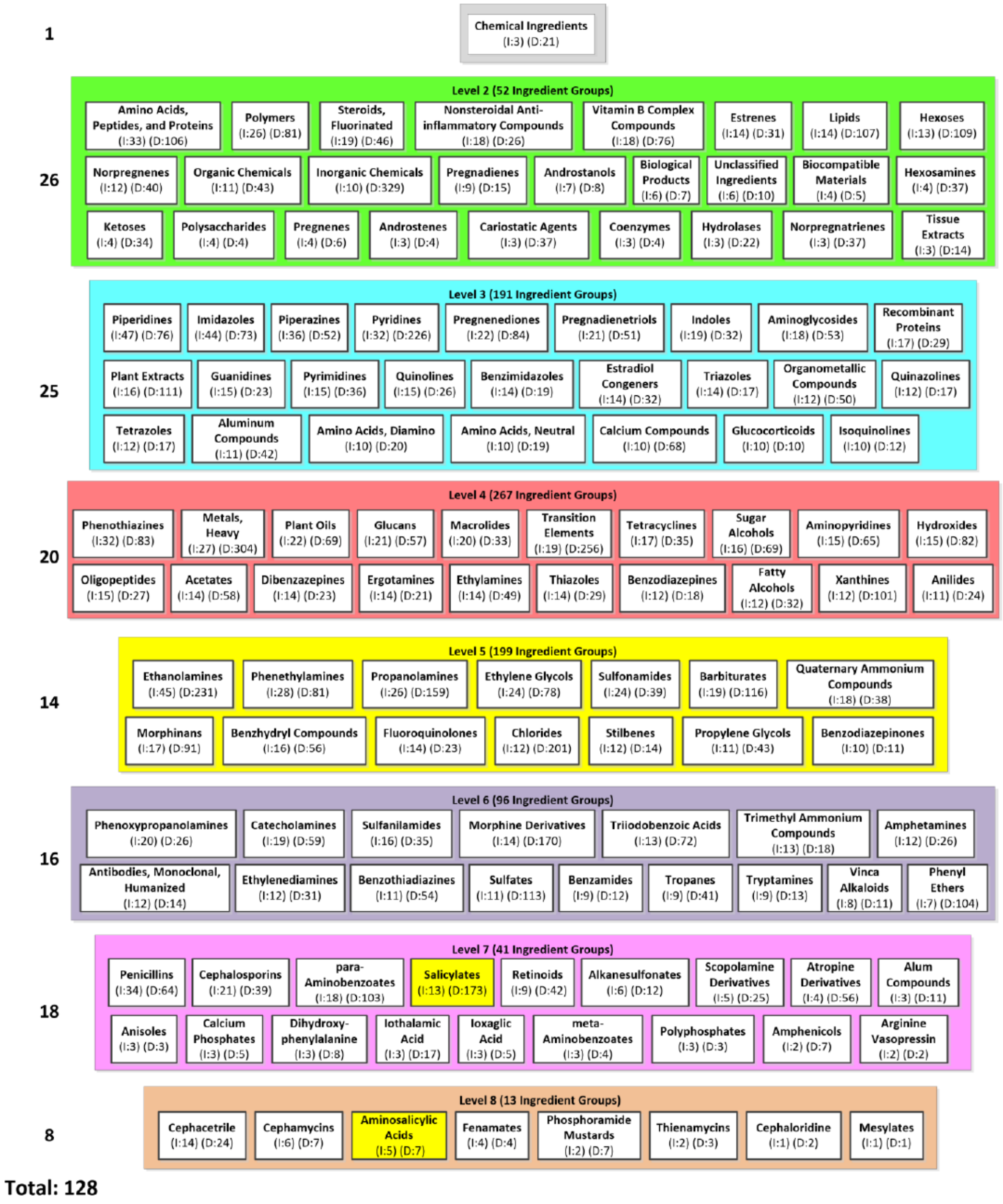

1
Chemical Ingredients (I:3) (D:21)
26
Level 2 (52 Ingredient Groups)
Amino Acids, Peptides, and Proteins (I:33) (D:106)
Polymers (I:26) (D:81)
Steroids, Fluorinated (I:19) (D:46)
Nonsteroidal Anti-inflammatory Compounds (I:18) (D:26)
Vitamin B Complex Compounds (I:18) (D:76)
Estrenes (I:14) (D:31)
Lipids (I:14) (D:107)
Hexoses (I:13) (D:109)
Norpregnenes (I:12) (D:40)
Organic Chemicals (I:11) (D:43)
Inorganic Chemicals (I:10) (D:329)
Pregnadienes (I:9) (D:15)
Androstanols (I:7) (D:8)
Biological Products (I:6) (D:7)
Unclassified Ingredients (I:6) (D:10)
Biocompatible Materials (I:4) (D:5)
Hexosamines (I:4) (D:37)
Ketoses (I:4) (D:34)
Polysaccharides (I:4) (D:4)
Pregnenes (I:4) (D:6)
Androstenes (I:3) (D:4)
Cariostatic Agents (I:3) (D:37)
Coenzymes (I:3) (D:4)
Hydrolases (I:3) (D:22)
Norpregnatrienes (I:3) (D:37)
Tissue Extracts (I:3) (D:14)
25
Level 3 (191 Ingredient Groups)
Piperidines (I:47) (D:76)
Imidazoles (I:44) (D:73)
Piperazines (I:36) (D:52)
Pyridines (I:32) (D:226)
Pregnenediones (I:22) (D:84)
Pregnadienetriols (I:21) (D:51)
Indoles (I:19) (D:32)
Aminoglycosides (I:18) (D:53)
Recombinant Proteins (I:17) (D:29)
Plant Extracts (I:16) (D:111)
Guanidines (I:15) (D:23)
Pyrimidines (I:15) (D:36)
Quinolines (I:15) (D:26)
Benzimidazoles (I:14) (D:19)
Estradiol Congeners (I:14) (D:32)
Triazoles (I:14) (D:17)
Organometallic Compounds (I:12) (D:50)
Quinazolines (I:12) (D:17)
Tetrazoles (I:12) (D:17)
Aluminum Compounds (I:11) (D:42)
Amino Acids, Diamino (I:10) (D:20)
Amino Acids, Neutral (I:10) (D:19)
Calcium Compounds (I:10) (D:68)
Glucocorticoids (I:10) (D:10)
Isoquinolines (I:10) (D:12)
20
Level 4 (267 Ingredient Groups)
Phenothiazines (I:32) (D:83)
Metals, Heavy (I:27) (D:304)
Plant Oils (I:22) (D:69)
Glucans (I:21) (D:57)
Macrolides (I:20) (D:33)
Transition Elements (I:19) (D:256)
Tetracyclines (I:17) (D:35)
Sugar Alcohols (I:16) (D:69)
Aminopyridines (I:15) (D:65)
Hydroxides (I:15) (D:82)
Oligopeptides (I:15) (D:27)
Acetates (I:14) (D:58)
Dibenzazepines (I:14) (D:23)
Ergotamines (I:14) (D:21)
Ethylamines (I:14) (D:49)
Thiazoles (I:14) (D:29)
Benzodiazepines (I:12) (D:18)
Fatty Alcohols (I:12) (D:32)
Xanthines (I:12) (D:101)
Anilides (I:11) (D:24)
14
Level 5 (199 Ingredient Groups)
Ethanolamines (I:45) (D:231)
Phenethylamines (I:28) (D:81)
Propanolamines (I:26) (D:159)
Ethylene Glycols (I:24) (D:78)
Sulfonamides (I:24) (D:39)
Barbiturates (I:19) (D:116)
Quaternary Ammonium Compounds (I:18) (D:38)
Morphinans (I:17) (D:91)
Benzhydryl Compounds (I:16) (D:56)
Fluoroquinolones (I:14) (D:23)
Chlorides (I:12) (D:201)
Stilbenes (I:12) (D:14)
Propylene Glycols (I:11) (D:43)
Benzodiazepinones (I:10) (D:11)
16
Level 6 (96 Ingredient Groups)
Phenoxypropanolamines (I:20) (D:26)
Catecholamines (I:19) (D:59)
Sulfanilamides (I:16) (D:35)
Morphine Derivatives (I:14) (D:170)
Triiodobenzoic Acids (I:13) (D:72)
Trimethyl Ammonium Compounds (I:13) (D:18)
Amphetamines (I:12) (D:26)
Antibodies, Monoclonal, Humanized (I:12) (D:14)
Ethylenediamines (I:12) (D:31)
Benzothiadiazines (I:11) (D:54)
Sulfates (I:11) (D:113)
Benzamides (I:9) (D:12)
Tropanes (I:9) (D:41)
Tryptamines (I:9) (D:13)
Vinca Alkaloids (I:8) (D:11)
Phenyl Ethers (I:7) (D:104)
18
Level 7 (41 Ingredient Groups)
Penicillins (I:34) (D:64)
Cephalosporins (I:21) (D:39)
para-Aminobenzoates (I:18) (D:103)
Salicylates (I:13) (D:173)
Retinoids (I:9) (D:42)
Alkanesulfonates (I:6) (D:12)
Scopolamine Derivatives (I:5) (D:25)
Atropine Derivatives (I:4) (D:56)
Alum Compounds (I:3) (D:11)
Anisoles (I:3) (D:3)
Calcium Phosphates (I:3) (D:5)
Dihydroxy-phenylalanine (I:3) (D:8)
Iothalamic Acid (I:3) (D:17)
Ioxaglic Acid (I:3) (D:5)
meta-Aminobenzoates (I:3) (D:4)
Polyphosphates (I:3) (D:3)
Amphenicols (I:2) (D:7)
Arginine Vasopressin (I:2) (D:2)
8
Level 8 (13 Ingredient Groups)
Cephacetrile (I:14) (D:24)
Cephamycins (I:6) (D:7)
Aminosalicylic Acids (I:5) (D:7)
Fenamates (I:4) (D:4)
Phosphoramide Mustards (I:2) (D:7)
Thienamycins (I:2) (D:3)
Cephaloridine (I:1) (D:2)
Mesylates (I:1) (D:1)
Total: 128

Figure 4.

An excerpt of 128 (15%) ingredient groups from the IAbN for June 2015 version of the *CI* hierarchy. The smaller ingredient groups have been hidden as follows. We show as many groups as possible in each level, in decreasing order by the number of ingredients in each group, while keeping the group names readable. *Child-of* links are hidden for readability. The number of ingredients and drugs summarized by each ingredient group is shown in parentheses and prepended with I: and D:, respectively. *Salicylates* and *Aminosalicylic Acids*, from Figure 3(c), are highlighted in yellow.

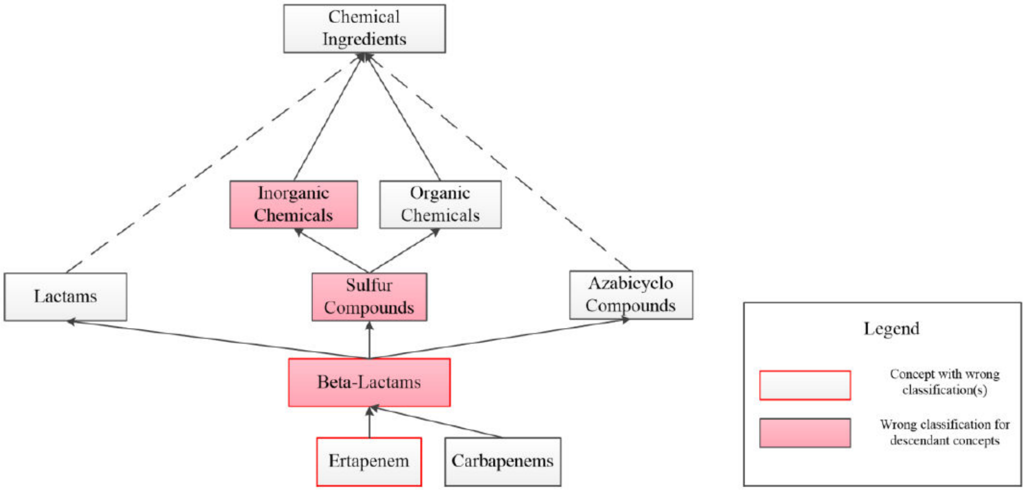


Figure 5.

A structure from NDF-RT that is problematic from a pharmacological point of view. *Lactams* is not a useful concept in pharmacology. Furthermore, even though *Beta-Lactams* contains a sulfur atom, it is *not* referred to as a “sulfur compound” in pharmacology and *Beta-Lactams* is not an Inorganic Chemical(s). Lastly, *Ertapenem* is misclassified.

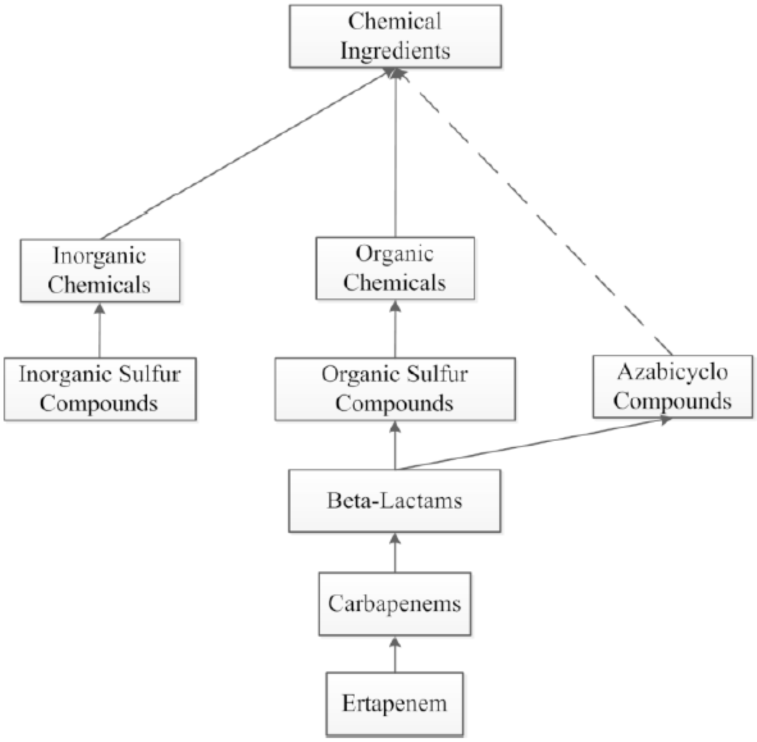


Figure 6.

The problems from Figure 5 are eliminated here. *Lactams* has been removed and two new concepts, *Inorganic Sulfur Compounds* and *Organic Sulfur Compounds*, are introduced. A new IS-A link from *Beta-Lactams* to *Organic Sulfur Compounds* is added. This avoids the wrong conclusion that *Beta-Lactams* is an Inorganic Chemical. The location of *Ertapenem* in the hierarchy has also been corrected.

Table 1

The distribution of the drug ingredients in exactly one ingredient group based on their number of parent ingredient groups

# of parent ingredient groups	# of drug ingredients	Percentage (Column 2/1851)
0	1	0.05%
1	1136	61.37%
2	569	30.74%
3	118	6.37%
4	25	1.35%
5	2	0.11%
Total:	1851	100.00%

Table 2

The statistical analysis of the auditing results of the 433 drug ingredients

# of parent ingredient groups	# of audited concepts	# of erroneous concepts	Percentage of erroneous concepts	Percentage of erroneous concepts per parent ingredient group
1	170	22	12.9%	12.9%
2	118	29	24.6%	12.3%
3	118	55	46.6%	15.5%
4	25	13	52.0%	13.0%
5	2	0	0.0%	0.0%
Total:	433	119	27.5%	

Table 3

The 2×2 contingency table for the control and study concepts

# of parent ingredient groups	# of erroneous concepts	# of concepts without errors
1	22	148
>1	97	166

Table 4

The 2×2 contingency table for the concepts with two and more than two parent ingredient groups

# of parent ingredient groups	# of erroneous concepts	# of concepts without errors
2	29	89
>2	68	77

Table 5

Examples of error types with counts

	Error type	# of erroneous concepts	Percentage (Column 2/119)	Examples
1	Incorrect direct classification (= wrong parent)	8	6.7%	Bisacodyl, Ertapenem, Loracarbef
2	Organic/Inorganic Chemicals classification	81	68.1%	Cyclomethicone, Oxyphenonium, Rabeprazole
3	Heterocyclic Compounds, X-Ring(s) (X is one of {1, 2, 3, 4 or more})	34	28.6%	Alosetron, Bilirubin, Ramipril
4	Other types of erroneous classifications	26	21.8%	Hydrogen Peroxide, Loracarbef, Levodopa

Table 6

Level distribution of drug ingredient concepts going upwards from the leaf level

Level (In this Table Level 0 Represents Leaves)	# of Drug Ingredient Concepts
Level 0	2020
Level 1	623
Level 2	20
Level 3	1