Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

A graph-based approach to auditing RxNorm

Olivier Bodenreider [a,*], Lee B. Peters [a]

[a] Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD, USA

ARTICLE INFO

Article history:
Received 20 June 2008
Available online 24 April 2009

Keywords:
Biomedical terminologies
Auditing methods
RxNorm
Quality assurance
Graphs

ABSTRACT

Objectives: RxNorm is a standardized nomenclature for clinical drug entities developed by the National Library of Medicine. In this paper, we audit relations in RxNorm for consistency and completeness through the systematic analysis of the graph of its concepts and relationships.

Methods: The representation of multi-ingredient drugs is normalized in order to make it compatible with that of single-ingredient drugs. All meaningful paths between two nodes in the type graph are computed and instantiated. Alternate paths are automatically compared and manually inspected in case of inconsistency.

Results: The 115 meaningful paths identified in the type graph can be grouped into 28 groups with respect to start and end nodes. Of the 19 groups of alternate paths (i.e., with two or more paths) between the start and end nodes, 9 (47%) exhibit inconsistencies. Overall, 28 (24%) of the 115 paths are inconsistent with other alternate paths. A total of 348 inconsistencies were identified in the April 2008 version of RxNorm and reported to the RxNorm team, of which 215 (62%) had been corrected in the January 2009 version of RxNorm.

Conclusion: The inconsistencies identified involve missing nodes (93), missing links (17), extraneous links (237) and one case of mix-up between two ingredients. Our auditing method proved effective in identifying a limited number of errors that had defeated the quality assurance mechanisms currently in place in the RxNorm production system. Some recommendations for the development of RxNorm are provided.

Published by Elsevier Inc.

1. Introduction

Terminology development in biomedicine largely relies on the manual work of human editors (sometimes called modelers) (e.g. [1,2]). Although sometimes facilitated by the use of knowledge representation formalisms such as description logics, this process is known to be error-prone [3,4]. Many approaches have been proposed for analyzing large biomedical terminologies, based on the property of their terms [3–6], on their structure [7–10] and on their semantics [3,4,11,12]. Most approaches focus on auditing hierarchical relations, which form the backbone of biomedical terminologies [7,8,10]. Many terminology developers include quality assurance and quality control processes as part of the development cycle [13]. However, such mechanisms fail to capture many errors and independent researchers and the user community play an important role in identifying and reporting errors in biomedical terminologies.

From a structural perspective, most biomedical terminologies can be seen as directed graphs in which nodes are concepts and links are semantic relationships. A path between two concepts can be characterized by the sequence of relationships that need to be traversed in order to reach a target concept from a source concept. While broad terminologies (e.g., SNOMED CT, NCI Thesaurus) usually have a complex model of meaning (or T-Box in description logics-based terminologies), specialized terminologies such as RxNorm (presented in detail later) only define a few major categories (or types) in the domain and their interrelations. In such terminologies, most of the assertions hold among instances of these categories (rather than among the categories themselves) and are associative rather than hierarchical. The graph of types provides a model against which graphs of instances can be validated. For example, all paths defined in the model are expected to be instantiated, allowing checking for completeness (of nodes and links at the instance level). Similarly, alternate paths between two types known to be consistent at the type level are also expected to be consistent at the instance level.

The objective of this study is to audit relations in RxNorm for consistency and completeness through the systematic analysis of the graph of its concepts and relationships (at the instance level, in reference to the type level). More specifically, we hypothesize that the traversal of equivalent paths yielding different results is indicative of errors in the graph of instances, including missing links and erroneous links, and possibly missing nodes.

[*] Corresponding author. National Library of Medicine, 8600 Rockville Pike, MS 3841 (Bldg. 38A, Rm B1N28U), Bethesda, MD 20894, USA. Fax: +1 301 480 3035.
E-mail address: olivier@nlm.nih.gov (O. Bodenreider).

1532-0464/$ - see front matter Published by Elsevier Inc.
doi:10.1016/j.jbi.2009.04.004

2. Background: RxNorm

RxNorm is a standardized nomenclature for clinical drug entities developed by the National Library of Medicine [14]. RxNorm is one of a suite of designated standards for use in US Federal Government systems for the electronic exchange of clinical health information. RxNorm has been used as part of a mediation strategy to exchange medication data between the Veterans Affairs (VA) and the Department of Defense (DoD) clinical information systems [15] and as a drug vocabulary for personal health records [16]. It is also expected to become an enabling resource for applications such as e-prescribing [17] and medication reconciliation [18].

2.1. RxNorm categories

The RxNorm data set is organized around eight major categories, called "term types" in RxNorm parlance (presented in **bold, sans serif typeface**, while instances of these categories are shown in *italic typeface*). There are four categories for generic drugs and four equivalent categories for branded drug entities. The four categories for generic drugs (referred to hereafter as generic concepts) are for ingredient alone (**ingredient**), ingredient plus strength (**clinical drug component**), ingredient plus dose form (**clinical drug form**) and ingredient plus strength and dose form (**clinical drug**). Analogously, the four categories for branded drug entities (referred to hereafter as branded concepts) are **brand name** (alone), **branded drug component** (brand name plus strength), **branded drug form** (brand name plus dose form) and **branded drug** (brand name plus strength and dose form).[1] Table 1 lists the eight major categories[2] and some instances. The dataset under investigation in this study (April 1, 2008) comprises, after excluding obsolete data, 3460 **ingredients** (ignoring specific salts), 9740 **brand names**, 13,362 **clinical drug components**, 13,868 **branded drug components**, 18,097 **clinical drugs**, 14,539 **branded drugs**, 8160 **clinical drug forms** and 11,376 **branded drug forms**.

2.2. RxNorm relations

As shown in Fig. 1, relations are defined among branded concepts and among generic concepts. For each **brand name** concept, there exists one or more **branded drug components**, **branded drugs** and **branded drug forms**. Each **ingredient** is associated with one or more **clinical drug components**, **clinical drugs** and **clinical drug forms**. Moreover, the RxNorm drug entities are related to each other by a well-defined set of named relationships (presented in *italic, sans serif typeface*). For example, **brand name** concepts are related to **branded drug component** concepts by the relationships *ingredient_of* and *has_ingredient*, the latter being the inverse relationship. Examples of relations at the instance level include:

- **branded drug → branded drug component**

  Zyrtec 5 MG Oral Tablet consists_of Cetirizine 5 MG [Zyrtec]

- **clinical drug component → clinical drug**

  Cetirizine 5 MG constitutes Cetirizine 5 MG Oral Tablet

Table 1
RxNorm major categories.

Category	Abbreviation	Instance
Ingredient	IN	Cetirizine
Brand name	BN	Zyrtec
Clinical drug component	SCDC	Cetirizine 5 MG
Branded drug component	SBDC	Cetirizine 5 MG [Zyrtec]
Clinical drug name	SCD	Cetirizine 5 MG Oral Tablet
Branded drug	SBD	Zyrtec 5 MG Oral Tablet
Clinical drug form	SCDF	Cetirizine Oral Tablet
Branded drug form	SBDF	Cetirizine Oral Tablet [Zyrtec]

Fig. 1 shows all relationships between the various kinds of drug entities. It must be noted that the relationship *isa* defined between **branded drug** and **branded drug form** and between **clinical drug** and **clinical drug form** does not have the usual semantics of the subsumption relation of the same name (e.g., as defined in [19]), but simply links an entity with **ingredient** (resp. **brand name**), strength and dose form to the corresponding entity with **ingredient** (resp. **brand name**) and dose form, but no strength.

In addition to relations among branded concepts and among generic concepts, RxNorm also defines relations between branded concepts and generic concepts. As illustrated in Fig. 1, most relations are between entities at the same level (e.g., **ingredient** plus strength to **brand name** plus strength). This relationship is called *tradename_of* from branded concepts to generic concepts, the inverse relationship being *has_tradename*. Additionally, RxNorm defines the relationship *consists_of* between **branded drugs** and **clinical drug components**, with *constitutes* as its inverse. Examples of relationships at the instance level include:

- **ingredient → brand name**

  Cetirizine has_tradename Zyrtec

- **branded drug → clinical drug**

  Zyrtec 5 MG Oral Tablet tradename_of Cetirizine 5 MG Oral Tablet

It should be noted that all relations in RxNorm are systematically mirrored by inverse relations. As shown in Fig. 1, for each link between two type nodes (e.g., *ingredient_of* between **ingredient** and **clinical drug component**), there is an inverse link (e.g., *has_ ingredient* between **clinical drug component** and **ingredient**). At the instance level, all relations in RxNorm are also represented bidirectionally, i.e., for each relation (e.g., *Cetirizine ingredient_of Cetirizine 5 MG*), the inverse relation (i.e., *Cetirizine 5 MG has_ingredient Cetirizine*) is also recorded in the RxNorm dataset. For this reason we often represent the links between drug entities as undirected (instead of bidirectional). The (undirected) representation of *Zyrtec 5 MG Oral Tablet* is shown in Fig. 2.

While all branded concepts stand in a relation to some generic drug concepts, some generic drug concepts are not linked to any branded concepts. For example, there is no branded concept corresponding to *Cetirizine 10 MG Extended Release Tablet*, which means that this particular ingredient, strength and dose form combination is not commercialized under a particular brand, but rather available as a generic drug.

For single-ingredient drugs there is a strict correspondence in RxNorm between branded and generic drug entities. To each branded drug entity (e.g., **brand name**) corresponds one generic drug entity of the equivalent type (e.g., **ingredient**). Additionally, as shown in Fig. 2, each branded drug entity is related to only one **brand name** and similarly each generic drug entity is related to only one **ingredient**. In contrast, there is no such correspondence between generic and branded concepts for multi-ingredi-

[1] RxNorm provides a "semantic normal form" for drug names and the RxNorm documentation refers to the eight major categories as ingredient (IN), semantic clinical drug component (SCDC), semantic clinical drug form (SCDF), semantic clinical drug (SCD), brand name (BN), semantic branded drug component (SBDC), semantic branded drug form (SBDF) and semantic branded drug (SBD). For readability in this article, we drop the "semantic" qualifier from these names.

[2] Other categories in RxNorm include drug forms (DF), generic pack (GPCK) and branded pack (GPCK). We deliberately focus on the eight major categories, which represent more than 99% of all RxNorm entities at the instance level.

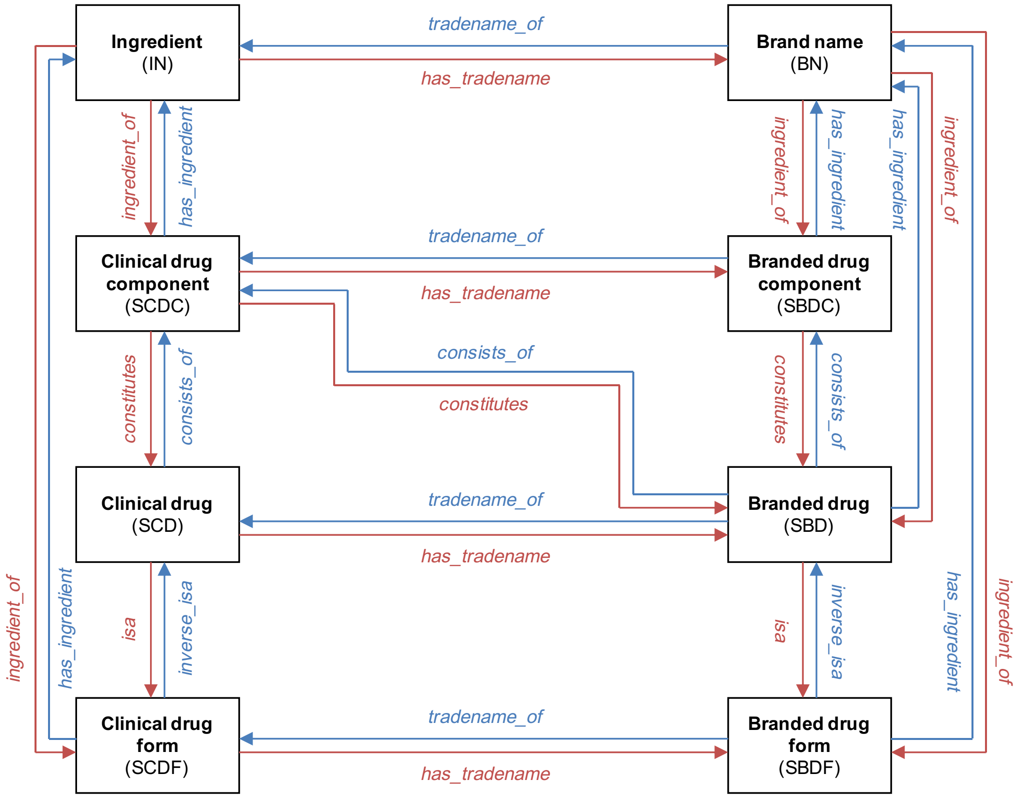

Fig. 1. Graph of the eight major categories in RxNorm and their interrelations.

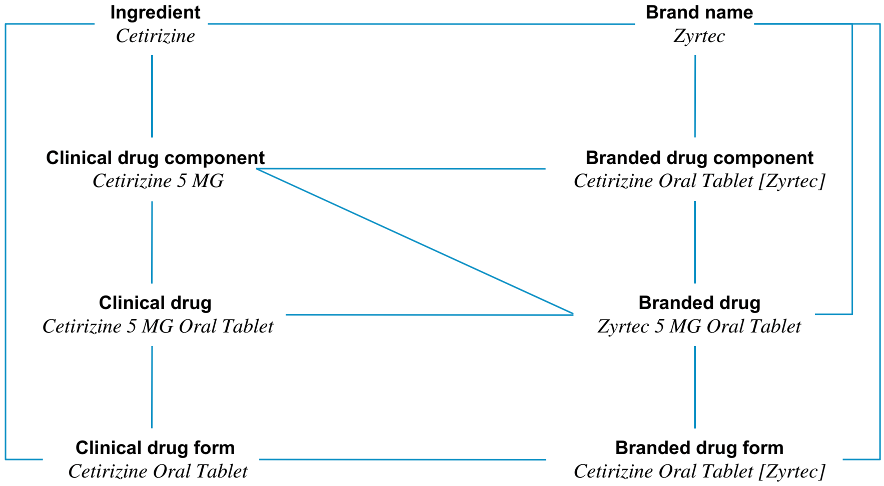

Fig. 2. Representation of *Zyrtec 5 MG Oral Tablet* in RxNorm with its interrelations to other clinical and branded drug entities.

ent drugs. Namely, while each multi-ingredient **branded drug**, **branded drug component** and **branded drug form** is related to only one **brand name**, multi-ingredient **clinical drugs** and **clinical drug forms** are related to multiple **ingredients** and **clinical drug components**. In addition, multi-ingredient **brand names** are related to multiple **ingredients**, multi-ingredient **branded drug components** and **branded drugs** are related to multiple **clinical drug components**.

As shown in Fig. 3, the **branded drug** (*Sulfamethoxazole 400 MG/Trimethoprim* 80 *MG Oral Tablet [Bactrim]*) is linked to one **clinical drug** (*Sulfamethoxazole 400 MG/Trimethoprim* 80 *MG Oral Tablet*). However, the **branded drug** is linked to one **branded drug component** (*Sulfamethoxazole 400 MG/Trimethoprim* 80 *MG [Bactrim]*), whereas the corresponding **clinical drug** is linked to two **clinical drug components** (*Sulfamethoxazole 400 MG* and *Trimethoprim 80 MG*), one for each ingredient (*Sulfamethoxazole* and *Trimethoprim*) of this multi-ingredient drug.

The number of relations asserted at the instance level in the dataset under investigation in this study (April 1, 2008) is listed in Table 2. The counts are given in reference to the normalized representation described in Section 3.1, so that the number of inconsistencies can be related to these counts.

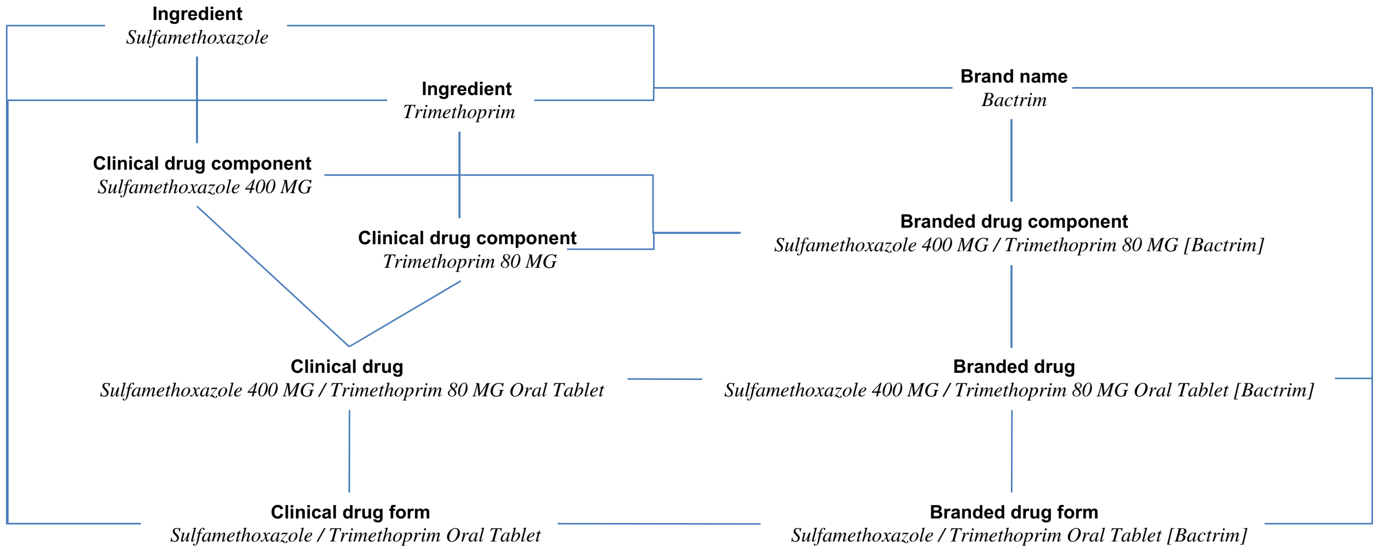
Ingredient
Sulfamethoxazole
Ingredient
Trimethoprim
Brand name
Bactrim
Clinical drug component
Sulfamethoxazole 400 MG
Clinical drug component
Trimethoprim 80 MG
Branded drug component
Sulfamethoxazole 400 MG / Trimethoprim 80 MG [Bactrim]
Clinical drug
Sulfamethoxazole 400 MG / Trimethoprim 80 MG Oral Tablet
Branded drug
Sulfamethoxazole 400 MG / Trimethoprim 80 MG Oral Tablet [Bactrim]
Clinical drug form
Sulfamethoxazole / Trimethoprim Oral Tablet
Branded drug form
Sulfamethoxazole / Trimethoprim Oral Tablet [Bactrim]

Fig. 3. Original representation of multi-ingredient drugs in RxNorm (SCDC-SBD relations are omitted for clarity).

Table 2
RxNorm major relations.

Relationship	Path	Count	Total
has_tradename	IN → BN	9723	49,506
	SCDC → SBDC	13,868	
	SCD → SBD	14,539	
	SCDF → SBDF	11,376	
inverse_isa	SBDF → SBD	14,539	32,636
	SCDF → SCD	18,097	
consists_of	SCD → SCDC	18,097	47,175
	SBD → SCD	14,539	
	SBD → SCDC	14,539	
has_ingredient	SBDC → BN	13,868	62,456
	SBD → BN	14,539	
	SBDF → BN	11,376	
	SCDC → IN	14,513	
	SCDF → IN	8160	

2.3. RxNorm Web Services API

A browser called RxNav[3] was developed in 2004 to access the RxNorm dataset and display graphically all related concepts and the relations between them [16]. RxNav uses web services to access the RxNorm data. In early 2008 the web services that access the RxNorm data were enhanced and made available publicly [20]. The current application programming interface (API) comprises functions for resolving drug names and codes into RxNorm identifiers, for accessing the properties of drug concepts, and for getting the related concepts of RxNorm entities. Here, we take advantage of the latter set of functions for exploring the RxNorm graph computationally.

3. Methods and results

The methods used in this study can be summarized as follows. We start by creating a normalized representation of multi-ingredient drugs. Then, we identify all meaningful paths between two categories, for all the instances of the source category. Finally, we assess the consistency of alternate paths between pairs of categories by comparing sets of instances reached through the various alternate paths. These three steps are presented in detail below.

3.1. Normalizing the representation of multi-ingredient drugs

As explained in the last paragraphs of Section 2.2, the representation of multi-ingredient drugs differs in RxNorm for generic concepts compared to branded concepts. For example, as shown in Fig. 2 for single-ingredient drugs and in Fig. 3 for multi-ingredient drugs, each multi-ingredient **branded drug**, **branded drug component** and **branded drug form** is related to only one **brand name**, whereas multi-ingredient **clinical drugs** and **clinical drug forms** are related to multiple **ingredients** and **clinical drug components**.

This representation is adapted to common uses of RxNorm as there is no such thing in practice as a combination of ingredients. However, we found this difference to be a hindrance to our auditing endeavor. Instead of using different algorithms for auditing single- and multi-ingredient drugs, we chose to modify the schema of RxNorm so that the same algorithm could be used on both single- and multi-ingredient drugs.

The normalization process we propose only affects multi-ingredient drugs. As illustrated by the differences between Figs. 3 and 4, normalization occurs at the level of **ingredients** and **clinical drug components** and their relations to other generic concepts, namely **clinical drugs** (for **clinical drug components**) and **clinical drug forms** (for **ingredients**), as well as to the corresponding branded concepts, namely **brand names** (for **ingredients**), and **branded drug components** and **branded drugs** (for **clinical drug components**). The normalization process simply reifies multi-ingredient entities (i.e., transforms multi-ingredient entities into single-ingredient-like entities).

In practice the normalization process creates new **ingredient** concepts for combinations of **ingredients** and new **clinical drug component** concepts for combinations of **clinical drug components**. For example, as shown in Fig. 4, the two **ingredients** of the **brand name** *Bactrim*, *Sulfamethoxazole* and *Trimethoprim*, are grouped into the new **ingredient** concept *Sulfamethoxazole/Trimethoprim*. Similarly, the two **clinical drug components** of the **branded drug component** *Sulfamethoxazole 400 MG/Trimethoprim 80 MG [Bactrim]*, *Sulfamethoxazole 400 MG* and *Trimethoprim 80 MG*, are grouped into the new **clinical drug component** concept *Sulfamethoxazole 400 MG/Trimethoprim 80 MG*. The relations of the newly created concepts are

[3] http://mor.nlm.nih.gov/download/rxnav/.

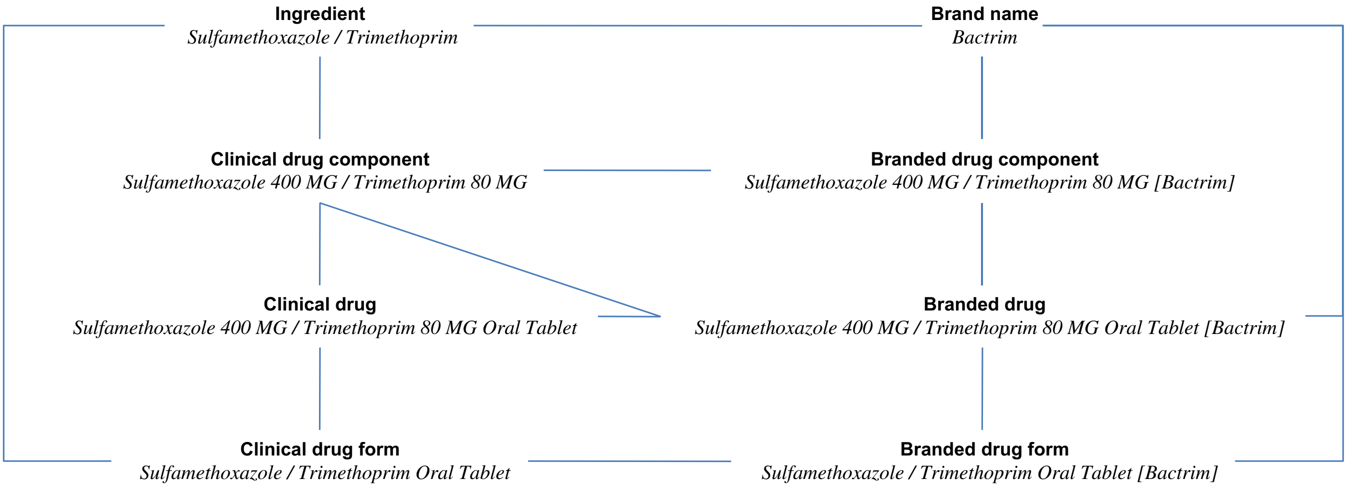
Ingredient
Sulfamethoxazole / Trimethoprim
Brand name
Bactrim
Clinical drug component
Sulfamethoxazole 400 MG / Trimethoprim 80 MG
Branded drug component
Sulfamethoxazole 400 MG / Trimethoprim 80 MG [Bactrim]
Clinical drug
Sulfamethoxazole 400 MG / Trimethoprim 80 MG Oral Tablet
Branded drug
Sulfamethoxazole 400 MG / Trimethoprim 80 MG Oral Tablet [Bactrim]
Clinical drug form
Sulfamethoxazole / Trimethoprim Oral Tablet
Branded drug form
Sulfamethoxazole / Trimethoprim Oral Tablet [Bactrim]

Fig. 4. Normalized representation of multi-ingredient drugs in RxNorm.

adapted accordingly. A single link is created from the new **ingredient** *Sulfamethoxazole/Trimethoprim* to both the **clinical drug form** *Sulfamethoxazole/Trimethoprim Oral Tablet* and the **brand name** *Bactrim*. Similarly, a single link is created from the new **clinical drug component** *Sulfamethoxazole 400 MG/Trimethoprim 80 MG* to both the **clinical drug** *Sulfamethoxazole 400 MG/Trimethoprim 80 MG Oral Tablet* (*ingredient_of*) and the **branded drug component** *Sulfamethoxazole 400 MG/Trimethoprim 80 MG [Bactrim]* (*tradename_of*). Finally, a single link is also created between the new **ingredient** and the new **clinical drug component** (*ingredient_of*). All links are represented bidirectionally. The original links are removed, and so are the original **ingredients** and **clinical drug components** if they do not participate in any other single- or multi-ingredient drug entities.

3.2. Identifying all meaningful paths

A path between two drug concepts can be characterized by the sequence of relationships that need to be traversed in order to reach a target drug concept from a source drug concept. For example, one path between **clinical drug component** (SCDC) and **branded drug component** (SBDC) is SCDC → SCD → SBD → SBDC, through the relationships *constitutes*, *has_tradename* and *consists_of*. Because all relations are mirrored with inverse relations in RxNorm, an inverse path can be found between SBDC and SCDC (i.e., SBDC → SBD → SCD → SCDC), traversing the inverse relationships in reverse order, i.e., going through the relationships *constitutes*, *tradename_of* and *consists_of*.

Moreover, after normalization of the representation of multi-ingredient drugs, the exploration of any path is functionally equivalent to the exploration of the inverse path. For example, auditing the path SCDC → SCD → SBD → SBDC from *Cetirizine 5 MG* is equivalent to auditing the path SBDC → SBD → SCD → SCDC from *Cetirizine 5 MG [Zyrtec]*. For this reason, of the 56 pairs of drug entities, only half of them (28) need to be considered for auditing purposes (Fig. 5).

For these 28 pairs of drug entities in RxNorm, we want to explore all paths between source and target drug concepts at the instance level. Most of the paths between a source and a target drug concept are expected to be equivalent. For example, as shown in Fig. 2, there are multiple possible paths between the **ingredient** *Cetirizine* and the **clinical drug form** *Cetirizine Oral Tablet*, including IN → SCDF and IN → SCDC → SCD → SCDF. These two paths are expected to be equivalent, i.e., to reach the same set of **clinical drug form** target concepts from the source **ingredient** concept.

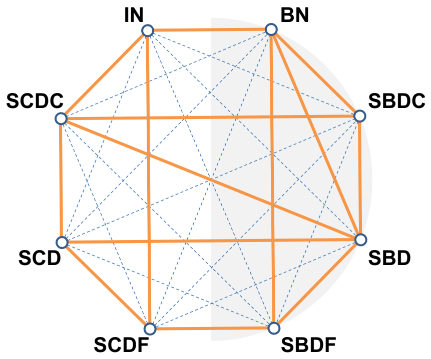

Fig. 5. The 28 paths among the 8 major drug entities in RxNorm. (Solid links join pairs of entities with direct relations in RxNorm; dotted links join pairs of entities without direct relations. The clear portion of the graph corresponds to generic concepts, the shaded portion to branded.)

As it is the case for graph traversal in general [21], we allow each node of the RxNorm graph to be traversed only once in order to avoid infinite recursion (e.g., SCDC → SCD → SCDC → SCD → · · ·). More importantly, the traversal of the RxNorm graph also is influenced by the nature of drug information. The following elements restrict how the RxNorm graph may be traversed. First, some generic concepts do not have any associated branded concepts (e.g., there is no **branded drug** corresponding to the **clinical drug** *Cetirizine 10 MG Extended Release Tablet*). Second, some generic concepts are associated with several branded concepts (e.g., *Coumadin*, *Jantoven*, *Marfarin* and *Warfin* are **brand names** for the **ingredient** *Warfarin*). Third, only a limited number of strength and dose form combinations exists for a given ingredient or branded drug (e.g., 1 MG/ML is an appropriate strength for the dose form oral solution, but 10 MG is not). And fourth, not all brands produce all strengths and dose forms of a given drug (e.g., *Warfarin* is available in various strengths for the dose form *Oral Tablet*, but the only strength available for the **brand name** *Marfarin* is 4 MG). For these four reasons, we know that some paths will predictably be different from paths with the same source and target concepts. In the auditing process, we want to ignore such predictable differences and focus on identifying discrepancies among paths expected to be equivalent.

Based on our knowledge of the subject matter and our experience in defining rules for traversing the RxNorm graph in RxNav, we defined *a priori* four constraints that allow us to avoid processing meaningless (predictably inconsistent) paths.

- **Constraint 1.** The path shall only cross once between the generic and branded drug entities.

One reason for this constraint is that *some generic concepts do not have any associated branded concepts*. Therefore, it would be predictably inconsistent to go, for example, from one **clinical drug component** to the corresponding **clinical drugs** through branded concepts (i.e., crossing over twice between the generic and branded drug entities). Because there is no **branded drug** corresponding to the **clinical drug** *Cetirizine 10 MG Extended Release Tablet*, paths from the **clinical drug component** *Cetirizine 10 MG* to **clinical drugs** on the generic side will correctly find *Cetirizine 10 MG Extended Release Tablet*, while paths going through the branded concepts will not (Fig. 6).

The other reason is that *some generic concepts are associated with several branded concepts*. Therefore, it would also be predictably inconsistent to go, for example, from one **branded drug** to the corresponding **branded drug component** through the generic concepts **clinical drug component**, as the specificity of the original brand would be lost, by design, on the generic side. For example, starting from the **branded drug** *Coumadin 1 MG Oral Tablet* and traversing to the generic side leads to the **clinical drug component** *Warfarin 1 MG*. Crossing back over to the brand side appropriately leads to the **branded drug component** *Warfarin 1 MG [Coumadin]*, but also incorrectly to the **branded drug component** *Warfarin 1 MG [Jantoven]*. By doing so, the brand specificity (*Coumadin*) has been lost in the **branded drug components** retrieved from the **branded drug** (Fig. 7).

- **Constraint 2**. For paths starting on the brand side and crossing over to the generic side, any property (strength or dose form) of the target entity, if acquired along the path, shall be acquired on the brand side.

This constraint also comes from the fact that *some generic concepts do not have any associated branded concepts*. Therefore, acquiring the property (strength or dose form) on the generic side might result in reaching generic concepts that do not correspond to the source (brand) entity. For example, when going from the **branded drug component** *Cetirizine 10 MG [Zyrtec]* to **clinical drugs** through the **clinical drug component** *Cetirizine 10 MG* (i.e., acquiring the dose form property on the generic side), the **clinical drug** *Cetirizine 10 MG Extended Release Tablet* will be found, although there is no branded equivalent for this **clinical drug**. In contrast, going through the **branded drugs** *Zyrtec 10 MG Chewable Tablet* and *Zyrtec 10 MG Oral Tablet* (i.e., acquiring the dose form property on the brand side) will appropriately lead to the **clinical drugs** *Cetirizine 10 MG Chewable Tablet* and *Cetirizine 10 MG Oral Tablet* (Fig. 8).

- **Constraint 3**. The entities **ingredient** and **brand name** shall not be traversed from and to any other entities bearing strength or dose form properties.

The entities **ingredient** and **brand name** do not contain any strength or dose form information. Because *only a limited number of strength and dose form combinations exists for a given ingredient or branded drug*, going through an entity without strength or dose form information from an entity that contains strength information to an entity that contains dose form information results in wrongly associating every strength of the source entity with every dose form of the target entity. For example, going from the **clinical drug component** *Cetirizine 10 MG* (bearing strength) to **clinical drug forms** (bearing dose form) through the **ingredient** *Cetirizine* (bearing none) would result in the inappropriate association of *Cetirizine 10 MG* with *Cetirizine Oral Solution*, because 10 MG is never a valid strength for oral solutions (Fig. 9).

- **Constraint 4**. For paths starting on the generic side and crossing over to the brand side, any property (strength or dose form) of the source entity, if removed along the path, shall be removed on the brand side.

The reason for this constraint is that *not all brands produce all strengths and dose forms of a given drug*. Therefore, removing strength or dose form on the generic side might result in reaching branded concepts that do not correspond to the specific strength or dose form of the source (generic) entity. For example, when going from the **clinical drug** *Warfarin 1 MG Oral Tablet* to **branded drug forms** through the **clinical drug form** *Warfarin*

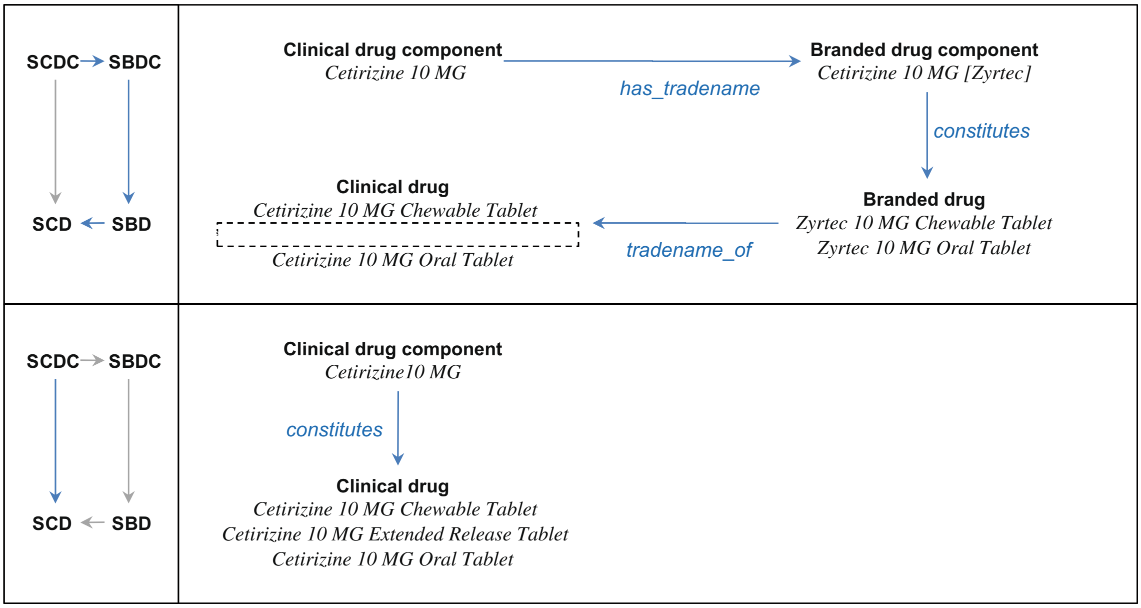

Fig. 6. Contrasting two paths (top: SCDC → SBDC → SBD → SCD and bottom: SCDC → SCD). The path at the top violates the constraint of not crossing between generic and branded drugs several times (Constraint 1) and fails to identify the clinical drug *Cetirizine 10 MG Extended Release Tablet* (dashed box), for which there is no corresponding branded drug.

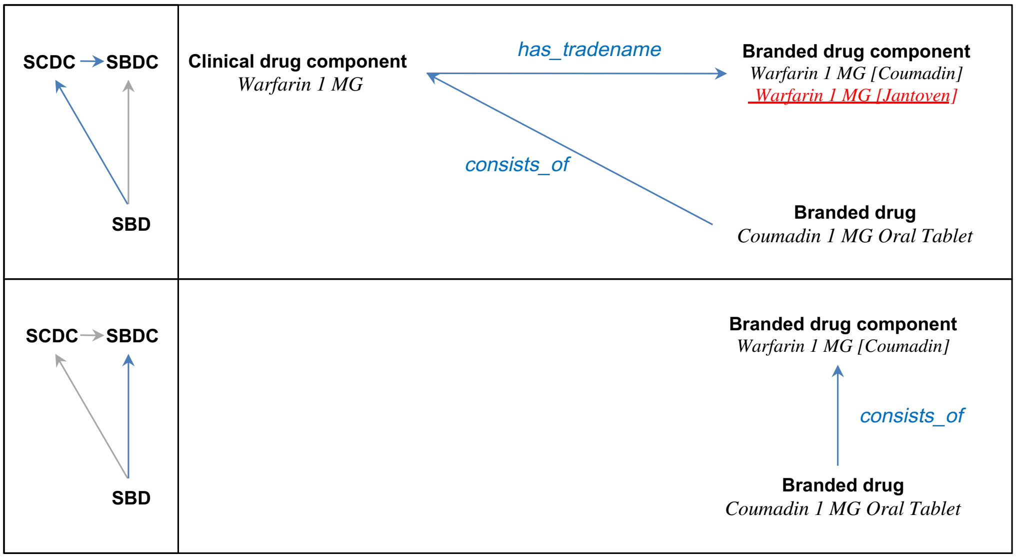

Fig. 7. Contrasting two paths (top: SBD → SCDC → SBCD and bottom: SBD → SBDC). The path at the top violates the constraint of not crossing between generic and branded drugs several times (Constraint 1) and retrieves additional branded drug components (e.g., *Warfarin 1 MG [Jantoven]*, underlined), wrongly associated with the brand *Coumadin*.

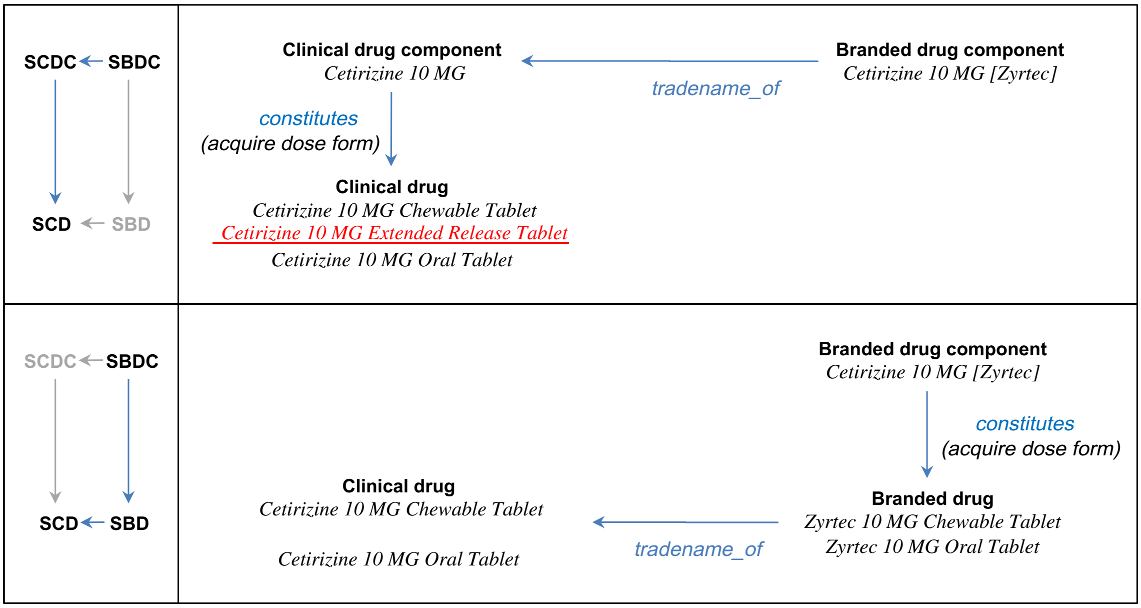

Fig. 8. Contrasting two paths (top: SBDC → SCDC → SCD and bottom: SBDC → SBD → SCD). The path at the top violates the constraint of acquiring strength (or dose form) only on the brand side in paths crossing over to the generic side (Constraint 2) and leads to irrelevant the SCD instance *Cetirizine 10 MG Extended Release Tablet* (underlined), for which there is no corresponding branded drug.

Oral Tablet (i.e., removing the strength property on the generic side), four **branded drug forms** will be found (*Warfarin Oral Tablet [Coumadin]*, *Warfarin Oral Tablet [Jantoven]*, *Warfarin Oral Tablet [Marfarin]* and *Warfarin Oral Tablet [Warfin]*). This is incorrect, because there are only two **branded drugs** with a strength of 1 MG for this form (*Coumadin 1 MG Oral Tablet* and *Jantoven 1 MG Oral Tablet*). In other words, the two other brands, *Marfarin* and *Warfin*, should not be retrieved, because the strength of their oral tablet preparations is not 1 MG, but 4 MG for *Marfarin* and 10 MG for *Warfin*. In contrast, going through the **branded drugs** *Coumadin 1 MG Oral Tablet* and *Jantoven 1 MG Oral Tablet* (i.e., removing the strength property on the brand side) will correctly lead to the **clinical drug components** *Warfarin Oral Tablet [Coumadin]* and *Warfarin Oral Tablet [Jantoven]*, since the corresponding **branded drugs** have the same strength (1 MG) as the starting **clinical drug** *Warfarin 1 MG Oral Tablet* (Fig. 10).

These constraints were easily implemented through a regular expression applied on the sequence of transitions for a given path and to the sequence of states (i.e., lists of properties) for all the nodes in a path, emulating a finite state automaton.

A total of 230 meaningful paths remain after all constraints have been applied. Since all relations in RxNorm are bidirectionally recorded, there exist 115 pairs of inverse paths. Only one copy needs to be explored for each path pair. As shown in Table 3, these 115 paths can be grouped into 28 classes with respect to source and target nodes in the path.

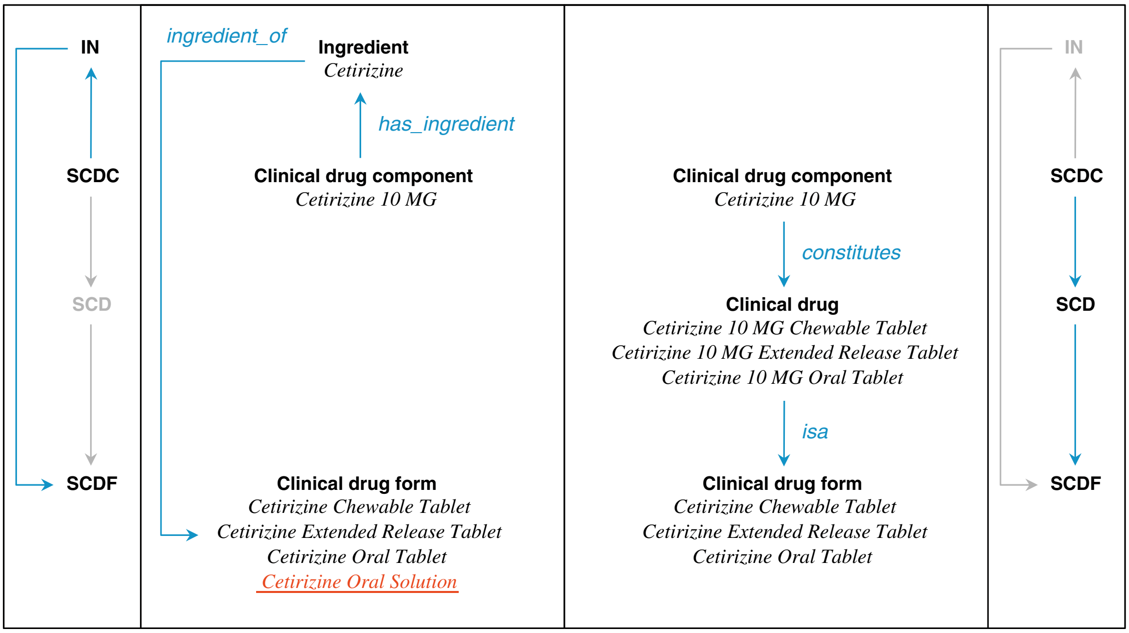

Fig. 9. Contrasting two paths (left: SCDC → IN → SCDF and right: SCDC → SCD → SCDF). The path on the left violates the constraint of not traversing IN (or BN) from and to entities bearing strength or dose form (Constraint 3) and leads to the irrelevant SCDF instance *Cetirizine Oral Solution* (underlined), because 10 MG is never a valid strength for oral solutions.

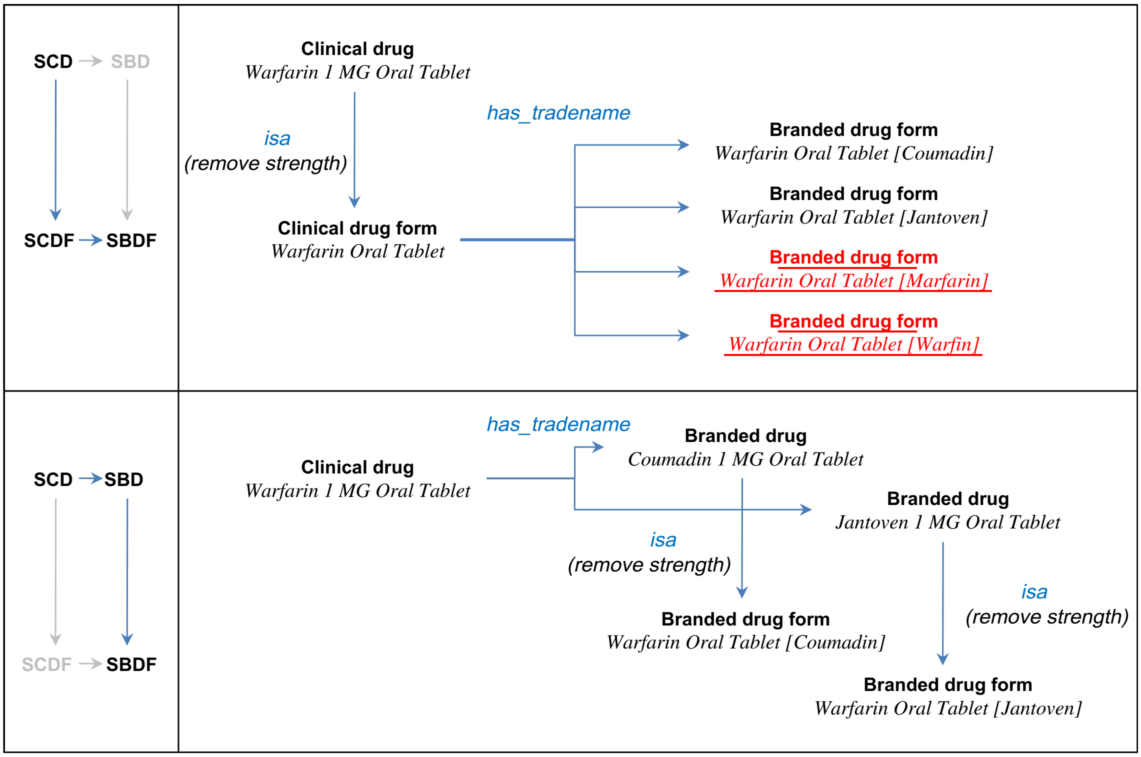

Fig. 10. Contrasting two paths (top: SCD → SCDF → SBDF and bottom: SCD → SBD → SBDF). The path at the top violates the constraint of removing strength (or dose form) only on the brand side in paths crossing over to the brand side (Constraint 4) and leads to the irrelevant SBDF instances *Warfarin Oral Tablet [Marfarin]* and *Warfarin Oral Tablet [Warfin]* (underlined), for which the strength 1 MG does not exist.

3.3. Exploring and comparing meaningful paths

3.3.1. Exploring paths

The RxNorm API was used to explore the paths. In particular, the function *getRelatedByRelationship()* was used for querying the instances of a given type that could be reached from a given RxNorm entity (instance) through a given link.

Each of the 115 meaningful paths (of categories) was explored as follows. Starting from the category corresponding to the first node in the path (source category), all instances of this

Table 3
Consistency among alternate paths. (The total number of paths is given for each category of paths. For each set of equivalent paths, only one typical path is shown. The number of paths equivalent to the typical path is given. Reference paths are indicated in bold font. The number of target nodes for each path was computed in April 2008 and in January 2009. The types of inconsistency refer to Section 4.2. Solid bullets show the origin of the inconsistency, while clear bullets indicate that the inconsistency has percolated to other paths. * indicates differences in set composition compared to the reference path, even when cardinalities are the same.)

					Target nodes		Types of inconsistency				
Start node	End node	Tot. paths	Typical paths	Equiv. paths	April 2008	Jan. 2009	1 a	1 b	1 c	2	3
BN	SBDC	3	**BN→SBDC**	2	**13,868**	**14,108**					
BN	SBD	3	**BN→SBD**	2	**14,539**	**14,848**					
BN	SBDF	3	BN→SBDF	0	11,376	11,620	●				
			BN→SBD→SBDF	1	**11,340**	**11,600**					
SBDC	SBD	1	**SBDC→SBD**	0	**14,539**	–					
SBDC	SBDF	1	**SBDC→SBD→SBDF**	0	**14,469**	–					
SBD	SBDF	1	**SBD→SBDF**	0	**14,539**	–					
IN	SCDC	2	**IN→SCDC**	1	**14,513**	**14,576**					
IN	SCD	2	**IN→SCDC→SCD**	0	**18,097**	**18,213**					
			IN→SCDF→SCD	0	(*) 18,097	18,213					●
IN	SCDF	2	IN→SCDF	0	8,160	8,184		●			
			IN→SCDC→SCD→SCDF	0	**8,104**	**8,139**					
SCDC	SCD	1	**SCDC→SCD**	0	**18,097**	–					
SCDC	SCDF	1	**SCDC→SCD→SCDF**	0	**17,556**	–					
SCD	SCDF	1	**SCD→SCDF**	0	**18,097**	–					
IN	BN	25	IN→BN	0	9,723	10,035				●	
			IN→SCDC→SBD→BN	16	**9,830**	**10,059**					
			IN→SCDF→SBDF→BN	0	9,864	10,076	○	○	○		
			IN→SCDC→SCD→SCDF→SBDF→BN	0	9,857	10,076	○		○		
			IN→SCDC→SCD→SCDF→SBDF→SBD→BN	1	9,831	10,059					●
			IN→SCDF→SCD→SBD→BN	2	(*) 9,830	10,059					●
IN	SBDC	11	IN→BN→SBDC	2	13,838	14,109				○	
			IN→SCDC→SBDC	5	**13,868**	**14,108**					
			IN→SCDC→SCD→SCDF→SBDF→SBD →SBDC	0	13,869	14,108					●
			IN→SCDF→SCD→SBD→SBDC	0	(*) 13,868	14,108					●
IN	SBD	11	IN→BN→SBD	2	14,509	14,849				○	
			IN→SCDC→SBD	6	**14,539**	**14,848**					
			IN→SCDC→SCD→SCDF→SBDF→SBD	0	14,540	14,848					●
IN	SBDF	11	IN→BN→SBDF	0	11,355	11,620	○			○	
			IN→BN→SBD→SBDF	1	11,319	11,600				○	
			IN→SCDC→SBD→SBDF	4	**11,340**	**11,600**					
			IN→SCDC→SCD→SCDF→SBDF	0	11,369	11,620			○		
			IN→SCDF→SBDF	0	11,376	11,620		○	○		
			IN→SCDF→SCD→SBD→SBDF	0	(*) 11,340	11,600					●
SCDC	BN	9	**SCDC→SBDC→BN**	8	**13,868**	**14,108**					
SCDC	SBDC	3	**SCDC→SBDC**	2	**13,868**	**14,108**					
SCDC	SBD	3	**SCDC→SBD**	2	**14,539**	**14,848**					
SCDC	SBDF	3	**SCDC→SBD→SBDF**	2	**14,469**	**14,733**					
SCD	BN	3	**SCD→SBD→BN**	2	**14,539**	**14,848**					
SCD	SBDC	1	**SCD→SBD→SBDC**	0	**14,539**	–					
SCD	SBD	1	**SCD→SBD**	0	**14,539**	–					
SCD	SBDF	1	**SCD→SBD→SBDF**	0	**14,539**	–					
SCDF	BN	6	SCDF→SBDF→BN	0	11,376	11,620	○		○		
			SCDF→SBDF→SBD→BN	4	**11,340**	**11,600**					
SCDF	SBDC	2	**SCDF→SCD→SBD→SBDC**	1	**14,469**	**14,733**					
SCDF	SBD	2	**SCDF→SBDF→SBD**	1	**14,539**	**14,848**					
SCDF	SBDF	2	SCDF→SBDF	0	11,376	11,620			●		
			SCDF→SCD→SBD→SBDF	0	**11,340**	**11,600**					

node were retrieved. For each instance of the source category, we recorded the set of instances of the target category which could be reached, following the links indicated in the path of categories. The complete set of instances reached for a given path is the union of the sets of target instances reached from each source instance.

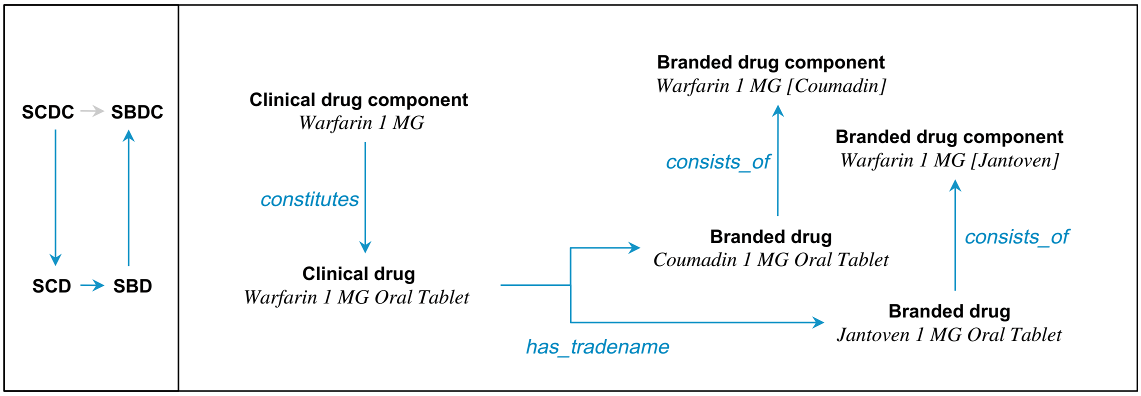

Fig. 11. Exploring the path SCDC → SCD → SBD → SBDC from the SCDC instance *Warfarin 1 MG*.

For example, the path SCDC → SCD → SBD → SBDC is explored as follows. The list of instances of SCDC (source instances) includes *Warfarin 1 MG*. As shown in Fig. 11, the only SCD instance that can be reached from *Warfarin 1 MG* through the relationship *constitutes* is *Warfarin 1 MG Oral Tablet*. From this SCD instance, following the relationship *has_tradename*, two SBD instances can be reached: *Coumadin 1 MG Oral Tablet* and *Jantoven 1 MG Oral Tablet*. The SBD *Coumadin 1 MG Oral Tablet* leads to the SBDC *Warfarin 1 MG [Coumadin]* (target category) through the relationship *consists_of*. Similarly, the SBD *Jantoven 1 MG Oral Tablet* leads to the SCDC instance *Warfarin 1 MG [Jantoven]*. In summary, the source SCDC instance *Warfarin 1 MG* leads to two target SBDC instances *Warfarin 1 MG [Coumadin]* and *Warfarin 1 MG [Jantoven]* through the path SCDC → SCD → SBD → SBDC. The source SCDC instance *Warfarin 1 MG* therefore contributes two target SBDC instances to the path SCDC → SCD → SBD → SBDC. Overall, this path yields 13,868 target instances.

3.3.2. Comparing paths

Alternate (meaningful) paths between a given source entity and a given target entity are expected to be equivalent. Alternate paths are equivalent if the same set of target instances is reached from a given set of source entities. A set of alternate paths is consistent if all alternate paths in the set are equivalent.

For example, there are three alternate paths between **clinical drug component** (SCDC) and **branded drug component** (SBDC), through entities including **clinical drugs** (SCD) and **branded drugs** (SBD):

1. SCDC → SBDC
2. SCDC → SBD → SBDC
3. SCDC → SCD → SBD → SBDC

The three alternate paths between SCDC and SBDC yield the same sets of 13,868 target instances and are deemed equivalent. The set of paths between SCDC and SBDC is deemed consistent.

3.3.3. Results

The results of the exploration of the 115 meaningful paths are summarized in Table 3. In order to reduce the amount of information in this table, we only display one typical path for each set of equivalent paths. For example, from the three equivalent paths presented above for the start node SCDC and end node SBDC, column 3 confirms that there are indeed three paths, although only one of them (SCDC → SBDC) is actually listed in column 4. In fact, column 5 indicates that there are two other unlisted equivalent paths for this path. Column 6 lists the number of target instances reached for each set of equivalent paths in the April 2008 version of RxNorm. The remaining columns present information pertaining to the evaluation and will be discussed later. Each of the 28 rows of Table 3 presents the list of alternate paths between a given pair of start and end nodes and shows which alternate paths contain inconsistencies. Paths free of inconsistencies – one for each group – are called reference paths and are indicated in bold.

Of the 28 groups of alternate paths expected to be consistent, 9 groups contain only one path and could not be checked for inconsistencies. Of the 19 groups having more than one path, all alternate paths are equivalent in 10 groups (53%), while 9 groups (47%) exhibit inconsistencies. Overall, 28 paths (represented by 20 typical paths) are not equivalent to the reference path from the same group. These 28 inconsistent paths represent 24% of the 115 meaningful paths.

3.3.4. Evaluation

All inconsistencies identified by our method were reported in September 2008 to our NLM colleagues in charge of RxNorm, who provided feedback on our findings. Their assessment is presented in the section below, along with the analysis of inconsistencies. Additionally we repeated the experiment on the January 2009 version of RxNorm in order to determine whether any of the inconsistencies reported had been corrected (Table 3).

4. Discussion

4.1. Summary of inconsistencies

The analysis of Table 3 reveals that inconsistencies in four paths (BN → SBDF, IN → SCDF, SCDF → SBDF and IN → BN) are actually responsible for the inconsistencies observed in the 12 of the 20 inconsistent (typical) paths. The reason for this is that these four paths are included as proper subpaths in the other eight paths. For example, SCDF→SBDF is a proper subpath of IN → SCDC → SCD → SCDF → SBDF from the group IN–SBDF.

The degree of inconsistency observed among alternate paths (i.e., the difference in number of target nodes reached, compared to the reference path) was generally small. For example, for the path IN–SCDF, the reference path yields 8104 target instances, while the inconsistent alternate path yields 8160 target instances. The 56 differences represent 0.7% of the target instances for this path.

4.2. Analysis of inconsistencies

Through manual analysis of the inconsistencies observed among alternate paths, this study revealed three major types of issues at the origin of the inconsistencies. The various types of inconsistency identified in the paths are presented in Table 3.

4.2.1. Type 1 inconsistencies

These inconsistencies involved **clinical drug form** (or **branded drug form**) entities linked to some **ingredient** (resp. **brand name**), but not linked to a **clinical drug** (resp. **branded drug**) entity. A total of 93 such inconsistencies were identified, affecting nine of the 20 paths exhibiting inconsistencies.

According to the RxNorm team, these inconsistencies do not necessarily violate the RxNorm editorial rules and can be justified by the fact that these **clinical drug forms** and **branded drug forms** are active concepts in at least one of the source vocabularies integrated in RxNorm. However, these entities might have an active status only in those source vocabularies updated with a lesser frequency (compared to most source vocabularies updated on a monthly basis). Therefore, we believe these **clinical drug forms** and **branded drug forms** should have a special status as they are part of incomplete RxNorm graphs and might cause problems in applications (e.g., in computerized prescription systems).

The three following subtypes of inconsistency can be distinguished based on the analysis of inconsistent paths.

- **Type 1a**. We found 36 cases of **branded drug form** concepts having no relation to any **branded drug** concept, but linked to some **branded name** concept. In this case, the direct path BN → SBDF is inconsistent with the alternate path BN → SBD → SBDF. (We consider BN → SBD → SBDF to be the reference path as it ensures that each SBDF is linked to some SBD.) For example, the **branded drug form** *Carbidopa/Levodopa Oral Tablet [Sinemet]* is related to the **brand name** *Sinemet*, but neither concept is linked to any **branded drug** concept. Inconsistencies of this type propagate directly to four other paths of which the path BN → SBDF (or the reverse path SBDF → BN) is a proper subpath (e.g., SCDF → SBDF → BN). Overall, this type of inconsistency affects five of the 20 paths exhibiting inconsistencies. Of the 36 cases, 16 had been corrected in the January 2009 version of RxNorm, including*Pseudoephedrine/Triprolidine Oral Tablet [Sudafed Plus]*, removed from the list of valid **branded drug form** concepts.
- **Type 1b**. We found 57 cases of **clinical drug form** concepts having no relation to any **clinical drug** concept, but linked to some **ingredient** concept. In this case, the direct path IN → SCDF is inconsistent with the alternate path IN → SCDC → SCD → SCDF. (We consider IN → SCDC → SCD → SCDF to be the reference path as it ensures that each SCDF is linked to some SCD.) For example, the **clinical drug form** *Papain Chewable Tablet* is related to the **ingredient** *Papain*, but neither concept is linked to any **clinical drug** concept. Inconsistencies of this type propagate directly to two other paths of which the path IN → SCDF is a proper subpath (e.g., IN → SCDF → SBDF). Overall, this type of inconsistency affects three of the 20 paths exhibiting inconsistencies. Of the 57 cases, 12 had been corrected in the January 2009 version of RxNorm, including *Sodium Fluorescein Injectable Solution*, removed from the list of valid **clinical drug form** concepts.
- **Type 1c**. We found 36 cases of **clinical drug form** concepts linked to some **branded drug form** concept having no relation to any **clinical drug** or **branded drug** concept. In this case, the direct path SCDF → SBDF is inconsistent with the alternate path SCDF → SCD → SBD → SBDF. (We consider SCDF → SCD → SBD → SBDF to be the reference path as it ensures that each SCDF is linked to some SCD and each SBDF to some SBD.) For example, the **clinical drug form** *Reserpine Oral Tablet* is related to the **branded drug form** *Reserpine Oral Tablet [Serpasil]*, but, while the **clinical drug form** concept is linked to **clinical drug** concepts (e.g., *Reserpine 1 MG Oral Tablet*), the **branded drug form** concept is not linked to any **branded drug** concept. Inconsistencies of this type propagate directly to five other paths of which the path SCDF → SBDF is a proper subpath (e.g., IN → SCDC → SCD → SCDF → SBDF), unless SBDF is followed by SBD in the path, which ensures the existence of an SBD concept for the SDCF concept. Overall, this type of inconsistency affects six of the 20 paths exhibiting inconsistencies. Of note, the 36 inconsistencies observed here are the same as the 36 inconsistencies reported under Type 1a, where they are identified through other paths.

Overall, of the 93 inconsistencies of type 1, 28 had been corrected in the January 2009 version of RxNorm.

4.2.2. Type 2 inconsistencies

These inconsistencies involved **ingredient** and **brand name** concepts linked to one another in a manner different from that used to relate the corresponding **clinical drug** and **branded drug** concepts. A total of 254 such inconsistencies were identified, affecting five of the 20 paths exhibiting inconsistencies.

In all three cases, the direct path IN → BN is inconsistent with alternate paths, such as IN → SCDC → SBD → BN (We consider IN → SCDC → SBD → BN to be the reference path as it ensures that there is some SCD, through the SCDC, or SBD linked to the IN and BN.) According to the RxNorm team, these inconsistencies correspond to errors and are in the process of being corrected, when they have not been corrected already.

The three following subtypes of inconsistency can be distinguished based on the analysis of inconsistent paths.

- **Type 2a**. In 17 cases, a **brand name** entity has no relation to any **ingredient** entities. Examples include *Sochlor*, not linked directly to its ingredient, *Sodium chloride*. All 17 cases had been corrected in the January 2009 version of RxNorm.
- **Type 2b**. In 129 cases, a **brand name** entity has extraneous relations to some **ingredient** entities. Examples include *Histex PD*, inappropriately linked to the ingredients *Hydrocodone* and *Pseudoephedrine*, when its actual ingredient is only *Carbinoxamine*. Of the 129 cases, 98 had been corrected in the January 2009 version of RxNorm, including *Benadryl Allergy Sinus*, no longer linked to the ingredient *Acetaminophen*.
- **Type 2c**. In 108 cases, a **brand name** entity refers to several combinations of ingredients depending on the dose form or strength of the drug. In such cases, a **brand name** entity is linked to multiple **clinical drug components**, each of which is linked to different sets of **ingredients**. For example, the **brand name** *Relagard* is linked to two **branded drug component** entities: *Acetic Acid 0.0092 MG/MG [Relagard]* and *Acetic Acid 0.009 MG/MG/Oxyquinoline* Sulfate 0.00025 *MG/MG [Relagard]*. Although sharing the same **brand name**, the former has only one **ingredient** (*Acetic Acid*), while the latter has 2 (*Acetic Acid* and *Oxyquinoline Sulfate*). Of the 108 cases, 71 had been corrected in the January 2009 version of RxNorm, including *Dimetapp*, whose **branded drug component** entities all refer to the same combination of ingredients (*Brompheniramine* and *Phenylpropanolamine*).

Overall, of the 254 inconsistencies of type 2, 186 had been corrected in the January 2009 version of RxNorm.

4.2.3. Type 3 inconsistencies

What looks like a mix-up between two ingredients causes one inconsistency that is reflected in seven of the 20 paths exhibiting inconsistencies. In this case, although the alternate paths sometimes exhibit the same numbers of target instances, the sets of target instances are actually different. The two ingredients involved in the mix-up are *Omega-3 Acid Ethyl Esters (USP)* and *Fatty Acids, Omega-3*. As nothing general is to be learned from this error, we do not report it here in detail. This problem had been corrected in the January 2009 version of RxNorm.

4.2.4. Summary of inconsistencies

Overall, the major types of inconsistency identified in the RxNorm dataset include extraneous nodes (type 1 inconsistencies), missing relations (type 2a inconsistencies) and extraneous relations (type 2b inconsistencies). Of the 348 inconsistencies identified in the April 2008 version of RxNorm, 215 (62%) had been corrected in January 2009.

4.3. Significance and limitations

The number of inconsistencies identified among alternate paths and the number of inconsistencies identified through their analysis is relatively modest (348 for 92,602 drug entities and 192,773 relations), which is a testimony to the high quality and careful curation of the RxNorm database. However, we believe this study is significant, because the underlying errors are difficult to identify. In fact, these inconsistencies had obviously defeated the quality assurance mechanisms currently in place in the RxNorm production system and had not been reported to (and acted upon by) the RxNorm team by the user community in the several years RxNorm has been available. We believe that only a systematic, principled analysis can identify such errors in a large dataset. The list of inconsistencies we identified was shared with the RxNorm developers.

RxNorm relations link together the various kinds of drug entities. Exhaustiveness and correctness of the relations are important parameters if RxNorm is to be used in applications, such as electronic prescription systems and in conjunction with decision support systems. For example, in a prescription system, physicians should not be presented with ingredients for which no branded drugs are available. For decision support systems relying on links between brand names and ingredients to check drug interactions, it is critical that all necessary relations be consistently implemented.

The method we developed is fully automated and performs a systematic evaluation of the entire RxNorm dataset. The availability of the RxNorm API allowed us to reduce low-level programming to a minimum. Unlike other auditing methods, the graph-based process we developed for analyzing RxNorm characterizes inconsistencies and groups them in categories according to their origin. As a result, the inconsistencies reported to the RxNorm team can be processed in groups and the appropriate quality assurance can be added to the production system.

Because of the specificity of RxNorm among biomedical terminologies (limited domain, absence of hierarchical structure, strong underlying graph model), traditional approaches to auditing terminologies are not directly applicable to RxNorm. Conversely, the graph-based approach developed for auditing RxNorm is not easily applicable to other biomedical terminologies. However, this study illustrates the need for automated, scalable methods, applied systematically to a terminology by an independent group of researchers.

Other limitations include the need for modifying the schema of the RxNorm database prior to running the auditing experiment. However, we see this limitation as minor, because the transformation is fully automated and more importantly, it enables us to use the same simple algorithm for processing both single- and multi-ingredient drugs. Although applied to the entire RxNorm dataset, this study deliberately focuses on the eight major drug categories and ignores categories including drug forms (DF), generic pack (GPCK) and branded pack (BPCK). However, these eight major categories represent more than 99% of all RxNorm entities at the instance level. In future work, we plan to audit the remaining categories as well. Inherent to this method is the impossibility of auditing paths between pairs of entities for which only one single path is available.

Finally, other approaches could be used to address the same issue, including role composition in a description logic-based environment. However, RxNorm is not available in any native description logic representation format and we found RxNorm to be amenable to graph-based approaches.

4.4. Recommendations for the RxNorm development process

The normalization process developed for this study, which makes the representation of generic concepts compatible with that of branded concepts, is a critical element of our method. However, we do not recommend that the RxNorm developers change the current representation. In fact, reified combinations of **ingredients** and **clinical drug component** entities are artificial constructs, with no equivalent in the real world, and would therefore not be useful to most users of RxNorm.

Some of the inconsistencies detected in this study call for additional quality assurance processes to be implemented in the RxNorm production system. For example, it would be easy to check if a given **clinical drug form** with links to an **ingredient** is also linked to at least one **clinical drug**.

This study forced us to formalize what constitutes a meaningful path for traversing the RxNorm graph. Although a small number of constraints are required for ensuring meaningful traversal of the graph, we found it difficult to formulate these constraints. As the use of RxNorm increases, we suggest that guidance be added to the RxNorm documentation regarding traversal of the RxNorm graph.

Finally, the RxNorm graph contains some redundancy, but redundancy is not present systematically throughout the graph. On the one hand, it might be better to provide users with the minimal number of relations necessary for traversing the graph in a meaningful way. This option would call for removing the direct relation between **ingredient** and **clinical drug form**, for example, as it can be reconstructed through the path IN → SCDC → SCD → SCDF. On the other hand, it might be useful to some users to have a fully saturated set of relations. This option would call for adding a direct relation between **ingredient** and **clinical drug**, mirroring the relation between **brand name** and **branded drug** on the brand side.

5. Conclusions

Through the graph-based method we developed for auditing RxNorm and applied to the entire RxNorm dataset (April 2008), we identified 348 inconsistencies, including extraneous nodes (93), missing links (17), extraneous links (237) and one case of mix-up between two ingredients. We shared our findings with the RxNorm team. A large proportion of the underlying errors had been corrected in the January 2009 version of RxNorm and the remaining inconsistencies are under review. Our auditing method proved effective in identifying a limited number of errors that had defeated the quality assurance mechanisms currently in place in the RxNorm production system, despite the high quality and careful curation of the RxNorm dataset in general. Based on our analysis, we recommended some changes to the RxNorm quality assurance process, as well as additions to the RxNorm documentation.

This study illustrates the need for principled, automated, scalable methods, applied systematically to the entire content of a terminology by an independent group of researchers. The lessons learned from this auditing experiment can be summarized as follows. Auditing needs to be grounded in domain knowledge (e.g., the constraints defined for selecting meaningful paths). Because curation is a labor-intensive process, auditing methods need to have good specificity if they are to be used to focus the attention of the editors of the terminology on particular areas. It is also useful that auditing methods characterize the errors they identify in order to facilitate the work of the editors. Auditing methods need to be automated and scalable, so they can be repeatedly applied to the entire content of the terminology as necessary (e.g., when updates become available). Independent auditing is important, because close proximity to the production process – including its tools, constraints (e.g., time and resources), culture and traditions – makes it difficult to imagine or implement solutions that deviate from the production routine (e.g., modify the database schema for auditing purposes). Finally, the result of the auditing process should be used not only to identify areas of the content of the terminology in need of review, but, more importantly, to inform the quality assurance process implemented as part of the terminology production environment. In other words, quality assurance has to be thought of as a proactive, not reactive process in the life cycle of a terminology.

Acknowledgments

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM). The authors wish to thank the RxNorm development team at NLM. In particular, we wish to thank Tammy Powell for providing useful feedback on the inconsistencies identified in this study and Stuart Nelson for his support.

References

[1] Hartel FW, de Coronado S, Dionne R, Fragoso G, Golbeck J. Modeling a description logic vocabulary for cancer research. J Biomed Inform 2005;38(2):114–29.

[2] Stearns MQ, Price C, Spackman KA, Wang AY. SNOMED clinical terms: overview of the development process and project status. Proc AMIA Symp 2001:662–6.

[3] Ceusters W, Smith B, Goldberg L. A terminological and ontological analysis of the NCI Thesaurus. Methods Inf Med 2005;44(4):498–507.

[4] Ceusters W, Smith B, Kumar A, Dhaen C. Ontology-based error detection in SNOMED-CT. Medinfo 2004;11(Pt. 1):482–6.

[5] Ogren PV, Cohen KB, Acquaah-Mensah GK, Eberlein J, Hunter L. The compositional structure of gene ontology terms. Pac Symp Biocomput 2004:214–25.

[6] Ogren PV, Cohen KB, Hunter L. Implications of compositionality in the gene ontology for its curation and usage. Pac Symp Biocomput 2005:174–85.

[7] Cimino JJ, Min H, Perl Y. Consistency across the hierarchies of the UMLS Semantic Network and Metathesaurus. J Biomed Inform 2003;36(6):450–61.

[8] Gu H, Perl Y, Elhanan G, Min H, Zhang L, Peng Y. Auditing concept categorizations in the UMLS. Artif Intell Med 2004;31(1):29–44.

[9] 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.

[10] Bodenreider O. Strength in numbers: exploring redundancy in hierarchical relations across biomedical terminologies. AMIA Annu Symp Proc 2003:101–5.

[11] Cimino JJ. Auditing the Unified Medical Language System with semantic methods. J Am Med Inform Assoc 1998;5(1):41–51.

[12] Cornet R, Abu-Hanna A. Two DL-based methods for auditing medical terminological systems. AMIA Annu Symp Proc 2005:166–70.

[13] 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.

[14] Liu S, Ma W, Moore R, Ganesan V, Nelson S. RxNorm: prescription for electronic drug information exchange. IT Prof 2005;7(5):17–23.

[15] Bouhaddou O, Warnekar P, Parrish F, Do N, Mandel J, Kilbourne J, et al. Exchange of computable patient data between the Department of Veterans Affairs (VA) and the Department of Defense (DoD): terminology mediation strategy. J Am Med Inform Assoc 2008;15(2):174–83.

[16] Zeng K, Bodenreider O, Kilbourne JT, Nelson SJ. RxNav: towards an integrated view on drug information. Medinfo 2007:P386.

[17] Schade CP, Sullivan FM, de Lusignan S, Madeley J. e-Prescribing, efficiency, quality: lessons from the computerization of UK family practice. J Am Med Inform Assoc 2006;13(5):470–5.

[18] Cimino JJ, Bright TJ, Li J. Medication reconciliation using natural language processing and controlled terminologies. Medinfo 2007;12(Pt. 1):679–83.

[19] Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, et al. Relations in biomedical ontologies. Genome Biol 2005;6(5):R46.

[20] Peters L, Bodenreider O. Using the RxNorm web services API for quality assurance purposes. AMIA Annu Symp Proc 2008:591–5.

[21] Aho AV, Ullman JD. The graph data model. In: Foundations of computer science. New York: Computer Science Press; 1992.