A Semantic Normal Form for Clinical Drugs in the UMLS: Early Experiences with the VANDF *

Stuart J. Nelson, MD [1], Steven H. Brown, MD [2], Mark S. Erlbaum, MD [3],
Nels Olson [3],Tammy Powell [1], Brian Carlsen [3], John Carter [3],
Mark S. Tuttle [3], and William T. Hole, MD [1]
1. National Library of Medicine, Bethesda, MD
2. Department of Veterans Affairs and Vanderbilt University, Nashville, TN
3. Apelon, Inc, Alameda, CA

Abstract

A semantic normal form (SNF) for a clinical drug, designed to represent the meaning of an expression typically seen in a practitioner's medication order, has been developed and is being created in the UMLS Metathesaurus. The long term goal is to establish a relationship for every concept in the Metathesaurus with semantic type "clinical drug" with one or more of these semantic normal forms. First steps have been taken using the Veterans Administration National Drug File (VANDF). 70% of the entries in the VANDF could be parsed algorithmically into the SNF. Next steps include parsing other drug vocabularies included in the UMLS Metathesaurus and performing human review of the parsed vocabularies. After machine parsed forms have been merged in the Metathesaurus Information Database (MID), editors will be able to edit matched SNFs for accuracy and establish relationships and relationship attributes with other clinical drug concepts.

Introduction

National Library of Medicine's (NLM) Unified Medical Language System® (UMLS®) project is a long-term research and development effort to design and build and maintain knowledge sources to be used by computer programs to overcome barriers to effective information retrieval [1]. The UMLS Metathesaurus® is one of the chief products of that project. Since the first version was released in 1990, the UMLS Metathesaurus has grown to include 776,940 concepts and 2.1 million concept names in over 60 different biomedical source vocabularies, some in multiple languages [2].

In late 2001, the NLM and the Veterans Administration (VA) began an experiment in modeling clinical drugs in the UMLS Metathesaurus. There were several motives for doing so: there was a suspicion that in the Metathesaurus there was considerable missed synonymy in naming of clinical drugs; the traditional methodologies of recognizing missed synonymy in the UMLS [3] did not seem to be effective for this category of concepts; there was hope that developing a new method might lead to improved interoperability of drug terminology [4]; the area of clinical drugs was seen as important in the growing issues of patient safety; and there was a growing consensus in the HL7 Vocabulary Technical Committee of what a model for clinical drugs should be. Importantly, the pharmacy knowledge base vendors, who spend considerable effort tracking NDC code changes and by necessity must maintain a terminology for pharmaceuticals, participated in the discussions and encouraged the efforts towards developing this standard. The HL7 model was based on what a clinician might order, and what type of order might be sent to the pharmacy. The dose form would be the form in which a drug was administered to a patient, as opposed to the form in which the manufacturer had supplied it. It was clearly distinct from the choices the pharmacy might make in fulfilling that order.

The form of the NLM-VA experiment with a clinical drug model was to define a Semantic Normal Form (SNF) to represent orderable drugs. Our hypotheses were that medications from "real world" information systems could be modeled by an SNF, that clinical drug concepts from disparate vocabularies with considerable naming variation could be declared synonymous (or found to be closely related) if they had identical SNF data structures, and that creation and maintenance of the SNFs would be a manageable task with the resources available.

We elected to begin testing our hypotheses using the Department of Veterans Affairs National Drug File (VANDF). The VANDF is a centrally maintained electronic formulary used by each of VHA's 172 medical centers. Facilities use the VANDF to check drug interactions, to manage orders, and to send outpatient prescriptions (57 million in 2001) to 7 regional automated mail-out pharmacies.

SNF Drug Component (SCDC)

CUI|ShortName|ActiveIngredient|PreciseIngredient|Basis|Strength|Units|Notes

Examples

C0111111|APAP|Acetaminophen|Acetaminophen|B|325|MG|Component example#1

C0123456|Codeine|Codeine Phosphate|Codeine|P|30|MG|Component example#2

SNF Clinical Formulation (SCD)

CUI|MetaID|ShortName|Component1/Component2/...|OrderableDoseForm|Notes

Example

C0654321|Codeine w/apap tablet|C0111111/C0123456|Oral Tablet|CFexample

Figure 1. Semantic Normal Forms for Clinical Drugs

Methods

SNFs for clinical drugs are canonical representations of clinical drugs, as defined by their active ingredients, strengths, and orderable dose forms. SNFs make explicit and/or normalize every active ingredient, strength, unit of measurement, and dosage form for a given clinical drug preparation. Employing both relationships between concepts and attribute-value pairs, the data represent the semantics of a clinical drug concept. SNFs for clinical drugs use standardized tokens for ingredient names, for units, and for dose forms, and a set of rules for expressing strength.

The Clinical Drug SNF Model

We have created two different types of SNF concepts of semantic type "Clinical Drug" within the Metathesaurus. The two SNF forms created are shown in Figure 1. The first is that of the drug component, referred to as SCDC, consisting of an ingredient and a strength. The second form is that of the clinical formulation, referred to as SCD, consisting of component(s) and a dose form.

In order to deal with the frequent use of different salts of the same active ingredient, it is necessary to indicate both the active (base) ingredient as well as the precise ingredient in a drug component. Because of variation in the specification for strength, being sometimes for the base ingredient and other times for the salt, it is also necessary to indicate the basis (whether base ingredient or precise) of the indicated strength.

The released components will contain only ingredients named generically. Values for the precise and active ingredient fields will be Metathesaurus concepts. The relationship of the ingredient to the component will be ingredient_of.

The Component Field can be repeated an arbitrary number of times until all the active components are named, as indicated by the ellipsis in Figure 1. The relationship of the components to the clinical formulation is that of constitutes. The OrderableDoseForm is a Metathesaurus concept with source "Proposed HL7 Orderable Dose Forms."

VANDF Representation in SNF

The VANDF files were received by the NLM in September; namely: 1) an National Drug Code (NDC)-level file of packaged clinical drug preparations (each with an official VA Product Name,) and 2) a file of ingredients and strengths keyed to each distinct VA Product Name. For the most part, the ingredients listed were those that were active. Each record contained the semantically important data elements deconstructed into separate fields (e.g., active ingredient, strength, units, route of administration, drug dose form). Rather than attempting to parse these elements from the often-abbreviated VA Product Name, we decided to build SNFs from the fielded data elements. File formats, idiosyncrasies, referential integrity problems, omissions, and certain data errors were then identified and analyzed.

To implement the SNF conversion for VANDF, we devised an algorithm to determine the "base ingredient" from a precise ingredient SNF, or from a VANDF active ingredient name. In doing so, we used partial matching to Metathesaurus terms having the semantic type "Pharmacologic Substance". If no shorter "base ingredient" (e.g., codeine) of a VANDF ingredient name (e.g., codeine phosphate) could be found in the Metathesaurus, the SNF base ingredient was defaulted to the SNF precise ingredient, in this case, the VANDF active ingredient. These ingredient concepts were then assigned ingredient_of relationships to an SNF clinical drug component concept (SCDC), which also contained normalized strengths in standardized units of measurement.

SNF clinical drugs (SCDs) were instantiated with consists_of relationships to one or more SCDCs. Each SCD also had a dose_form_of relationship to an HL7 OrderableDoseForm concept. To implement the latter, MSE manually mapped most empirically determined combinations of VANDF route of administration and VANDF dose form, each to a single standard OrderableDoseForm. Where the meaning of the VANDF route or dose form could not be determined, no mapping was performed.

Results

VANDF Conversion

The VANDF file contained 93,029 records. The file was by most standards a clean, well-maintained file. Rigorous examination did find a few minor problems. We excluded from further analysis 1,706 inactive records, 711 exact duplicate records, and 3047 medical supplies records, leaving a total of 87,565 records that underwent algorithmic processing into SNF. A total of 11,345 distinct clinical drugs (of the 87,565 NDC level records) were identified. From these, 10,178 SNF drug components could be produced algorithmically from the 87,565 VANDF records.

A separate file listed active ingredients. Active ingredients, which often included the designation of the salt (e.g., codeine phosphate) numbered 3,301. An additional 778 base ingredients without a salt (e.g, codeine) were derived algorithmically. Well over 99% of 87,500 NDC-level VANDF drug records could be mapped to an HL7 dose form. Incomplete SNF drug components and SNF clinical formulations were discarded. Out of 502 route-form combinations in the source data, 428 were successfully mapped to the proposed HL7 concepts.

VANDF data was provided at the NDC level, but aggregated by shared VA Product Name for Metathesaurus inversion and SNF creation. Underlying VANDF data errors may lead to aggregations of different conceptual entities resulting from incorrect mixes of ingredients, routes, and dose forms. Missing data (e.g., units of measurement) in VANDF ingredient records caused incomplete, hence discarded drug component names. Of the 29,246 lines in the file of ingredients for VA products, 21,774 (74%) had all the information needed to generate drug component name, resulting in 10,178 distinct SCDCs. Of the remaining lines from which an SCDC could not be generated, 7,390 (99%) were cases where the "Strength" and "Units" fields were blank. There were 1414 distinct values of the "Ingredient name" field represented in these 7,390 lines with no strength and units.

Lack of one or more drug component names, referential integrity errors with the master VANDF drug file, and missing dose form mappings, caused partial, and therefore discarded SNF clinical formulations. Of 11,345 distinct VA product names (clinical drugs), 7,997 (70.5%) had all the information needed to generate a clinical formulation algorithmically, 2,148 (19%) had incomplete ingredient names, 337 (3%) lacked an unambiguous HL7 dose-form mapping, and 224 (2%) had no entry in the file of ingredients for VA products.

Discussion

The significant result from this preliminary experiment addressed the first and third of our hypotheses. The model appeared to be adequate for expressing most of the orderable drugs. Some areas do remain more problematic. Most multi-component ingredients fit into the model (though finding a suitable short name for generic multivitamins is a challenge) but others, such as additives for intravenous alimentation solutions, will need further work. Other problematic areas are orderable materials used in tests (e.g., allergenic extracts), contrast media, and radiopharmaceuticals.

The task of addressing all of the clinical drugs in this manner appears to be manageable. The indication that 70% or more of the work can be done algorithmically with human review reduces the amount of labor involved to a reasonable level.

The base ingredient algorithm will certainly produce many false positives and negatives. In the best of all possible worlds, it is still an approximation that may properly assign an incorrect base ingredient name. It may also fail semantically due to incorrect VANDF data elements or errors of omission in the Metathesaurus.

Mapping of VANDF route of administration plus dose form to HL7 canonical dose forms was often questionable or imprecise. Furthermore, in cases where the VANDF dose form field reflects a manufactured dose form which differs from the administered dose form (e.g., powders to be dissolved or suspended), mappings and SCDs may be incorrect, hence causing false positive synonymy downstream.

Future Plans

Considerations

Most of the drug vocabularies currently in the Metathesaurus have the semantically important elements of a clinical drug concept deconstructed into individual database fields (e.g., active ingredient, strength, units, route of administration, drug dose form). However, sources differ in the degree of decomposition, and may still require parsing and analysis of text strings (e.g., abbreviated clinical drug name, strength plus units for one or more ingredients, etc.) to acquire the missing elements. Some of the desired elements may already be present in the Metathesaurus as source attributes. Due to abbreviations or truncations of the drug name in the original databases, the clinical drug name in the Metathesaurus has often, in the past, been reconstructed or assembled de novo from its individual data elements during source inversion. The creation of the SCDs should obviate this step.

Vocabulary-specific route plus dose form combinations require mapping to the HL7 dose forms. Because each vocabulary is different in its expressions, this step must be done separately for each vocabulary. Similarly, ingredient names are not canonicalized or standardized. Since they are derived de novo from each candidate vocabulary, algorithmic determination of SNFs precise and base ingredient names will likely be imperfect or inconsistent.

This model for the SNFs of clinical drugs is intended to be useful for representing pharmaceuticals given to patients. It is possible that the model will be extensible to include such things as allergenic extracts, over-the-counter preparations, including herbal preparations and multivitamins, alimentation solutions, radioactive substances, and contrast media. However, it is not certain exactly how these will be approachable with this model. Further investigation will be required.

Additionally, there are devices containing drugs that may have more than one clinical drug in them (e.g., kits, oral contraceptive packs). Some of these cases may well be dealt with by establishing them as medical devices with a relationship attribute of contains to the SNF clinical drug.

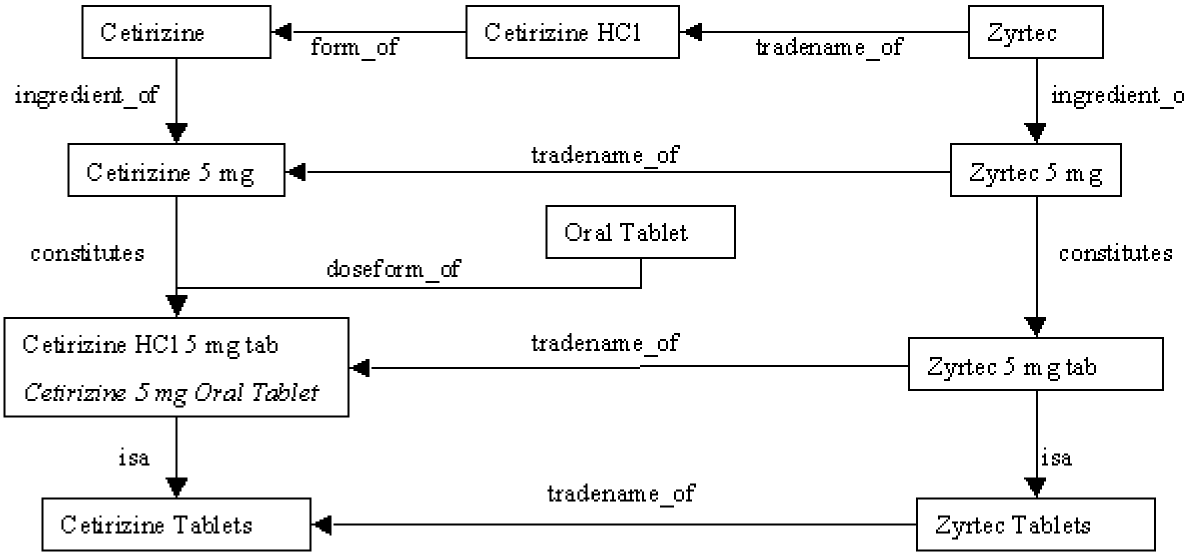
The Management Plan

Semantic normal forms (SNFs) for clinical drugs, both drug component (SCDC) and clinical formulation (SCD) will be made individually from each of the major sources of names of clinical drugs, the VANDF, Multum, Micromedex, First Databank, and Medispan, in that order. If parsing algorithms are unable to create SCDCs or SCDs, then a UMLS editor will do so. The SNFs will be individually edited before inserting them into the Metathesaurus Information Database, where UMLS editing is done. This first pass of editing is solely for the purpose of insuring that the SNF has been accurately produced, or to produce a SNF if one has not been created by parsing the name from the drug vocabulary. An editing interface which allows insertion or replacement of ingredients, precise ingredients, dose forms, or strength will be used by the editors. Listing of the SNFs for each source should consist of the full name in the source vocabulary, the lexical tag, the semantic type, followed by fields for the parsed drug components and dose form, separated by field delimiters. Editing of the fields is allowed through pick lists and through keyboarding. When keyboarded, validations check dose forms against the list of allowables, ingredients against concepts in the Metathesaurus with a chemical semantic type, and units against a check list of allowable units. The lexical tag field allows the notation of trade name (TRD), lab number (LAB), and short form (SFO) as well as the default of none (NON). An editor is able to change the semantic type or the lexical tag, as appropriate.

Lists of the SNFs for editing purposes are created according to the following criteria. For those clinical drugs which appear to have been successfully parsed (that is, an SCD has been created for them), the lists are printed out. Experience has taught us that review of material like this happens faster on paper than online. For processing in the editing interface , shorter work lists, up to 100 drugs at a time, are created for those whose parse was incomplete, and which require manual effort to successfully complete the parse.

Periodically, groups of the SNF will be inserted into the MID. At the time of insertion of the SNF into the MID, relationships and relationship attributes will be added, consistent with the relationship schema outlined above. Strengths will be normalized to a smaller number of allowable units. Once in the MID, matching of SNFs as well as the names will allow merging of multiple SNFs. Recognition that an ingredient is a trade name or lab number will allow the linkage of an SNF of a branded product to the generically named SNF with a relationship of trade_name_of.

Editing of concepts, concept-level relationships and relationship attributes can then proceed in the normal UMLS editing [5]. In that editing process, atoms whose SNF was incorrectly determined can be split out and the SNF altered. At the time the concept is approved, the SNF will become the preferred name for the concept. NLM will be designated as the source of the SNF by virtue of its support and responsibility for the automatic processing and human review that produced these forms.



Each box represents a UMLS concept
RxNorm italicized
Arrows represent labeled relationships

Figure 2
RxNorm Relationships in UMLS

After creation of the SNFs, the plan is to establish a set of relationships between the concepts in the Metathesaurus of semantic type "Clinical Drug," the ingredients, and one or more of the SNFs (either SCDC or SCD) for clinical drugs. The working name for the system of SNFs and the relationships is RxNorm. Figure 2 shows some of the relationships anticipated.

Objectives for the Spring Release

By the time of the spring release of the UMLS Metathesaurus in May, 2002, we should have an accurate assessment of how well the methodology of editing works, that is, is it fast, accurate, and reproducible. We should be able, during this process, to identify any difficulties with establishing the dose forms and with the rules for expressing strength. The complete model, including the relationship attributes for all appropriate clinical drugs in the Metathesaurus, should be instantiated for some number of frequently prescribed drugs.

Acknowledgements

Much of the conceptual work of developing this model came out of discussions at multiple HL7 Vocabulary Technical Committee meetings and conference calls. The work of all those participating in those discussions is gratefully acknowledged. References

1. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. *Methods Inf Med*. 1993 Aug;32(4):281-91.

2. UMLS Metathesaurus [Fact Sheet]. National Library of Medicine; Bethesda (MD); 2002 Feb 26.

3. Hole WT, Srinivasan S. Discovering Missed Synonymy In A Large Concept-Oriented Metathesaurus. *Proc AMIA Symp* 2000:354-8.

4. Broverman C, Kapusnik-Uner J, Shalaby J, Sperzel D. A Concept-Based Medication Vocabulary: An Essential Requirement for Pharmacy Decision Support. *Pharm Pract Manag* Q 18:1, Apr 1998.

5. Tuttle MS, Suarez-Munist ON, Olson NE, et al. Merging terminologies. *Medinfo* 1995;8 Pt 1:162-6.