

©2011

Mark E. Sharp

ALL RIGHTS RESERVED

DIMENSIONS OF DRUG INFORMATION

By

MARK E. SHARP

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Communication, Information and Library Studies

written under the direction of

Professor Nicholas Belkin

and approved by

---

---

---

---

New Brunswick, New Jersey

January 2011

## ABSTRACT OF THE DISSERTATION

### Dimensions of Drug Information

by MARK E. SHARP

Dissertation Director

Professor Nicholas Belkin

The high number, heterogeneity, and inadequate integration of drug information resources constitute barriers to many drug information usage scenarios. In the biomedical domain there is a rich legacy of knowledge representation in ontology-like structures that allows us to connect this problem both to the very mature field of library and information science classification research and the very new field of ontology matching/merging (OM). We argue for a broad view of OM that makes room not only for the "pre-formal" phase/type of multi-ontology integration exemplified by RxNorm and the UMLS Metathesaurus, but also for an even earlier phase/type when "What is there?" in a domain has to deal with implicit and poorly structured "ontologies" that barely qualify as such. Such is the case in the drug domain. We introduce *dimensions of drug information* as an approach to early, pre-formal OM in the drug domain that draws inspiration and incorporates principles from facet analysis, domain analysis, and Semantic Web research on linked data and mashups. By surveying 23 publically available drug information resources, we identified 39 dimensions relevant to four drug (sub)domains - pharmacy, chemistry, biology, and clinical medicine - and mapped them to the resources. An arbitrary four-domain, monohierarchical classification of the dimensions produced, by extension, a reasonable four-domain resource classification. Correspondence analysis and hierarchical cluster analysis also produced evidence of its partial validity. Detailed analysis of information on nine parent drug compounds from 15 resources refined this high-level dimensional mapping and identified hundreds of subdimensions which could be expressed as a six-level hierarchy. Based on these

dimensions, we integrated this information in an experimental database and showed that it was useful (1) as a training set for automating the normalization of additional raw data from the same 15 sources, bringing the important goal of building an integrated, comprehensive (all drugs) database within reach, and (2) for satisfying a variety of use cases, some quite complex, derived from published literature representing the user types corresponding to our domain focus.

## **Acknowledgements**

I would like to thank my dissertation director, Nicholas Belkin, for attracting me to the program, teaching me the fundamentals of library and information science, being the perfect role model for a successful researcher and committed educator, shepherding me through the dissertation process, many stimulating insights and discussions, and eleven years of unswerving support, understanding, and friendship.

Special thanks also to the other members of my dissertation committee, Tefko Saracevic, Marija Dalbello, and Olivier Bodenreider, for their many contributions to this work, and to Dr. Bodenreider for sponsoring my guest researchship at the U.S. National Library of Medicine where much of it was done.

I would also like to thank the members of my qualifying examination committee, Nina Wacholder, Paul Kantor, Alex Borgida, and Anselm Spoerri, for teaching me about natural language processing, information retrieval, ontologies, and information visualization, and for sponsoring my many pre-dissertation research projects, all of which contributed to this work.

In addition, I would like to thank Craig Scott, Marie Radford, Xiaojun Yuan, and Lorraine Gray for facilitating my re-entry into the Ph.D. program after an absence, and Joan Chabrak for her support, encouragement, and friendship.

I would also like to thank my employer, Merck & Co., Inc., for paying my tuition throughout my doctoral studies, and my many colleagues at Merck for their encouragement and interest, especially my managers Ann Jenkins, Judy Labovitz, Doris Schlichter, David Henderson, and Karen Marakoff, and my information science mentors, Patrick Perrin and Eric Minch.

Finally, I would like to thank my family for their inspiration and support, especially my wife, Celia Sharp.

## Table of Contents

Abstract .....	ii
Acknowledgements.....	iv
List of Figures .....	ix
List of Tables .....	x
Chapter 1. Introduction .....	1
1.1 Problem Statement.....	1
1.2 Rationale.....	2
1.3 Research Questions and Strategy .....	4
Chapter 2. Literature Review .....	7
2.1 Classification .....	7
2.2 Facet Analysis .....	9
2.3 Domain Analysis .....	12
2.4 Ontologies.....	14
2.5 Bio-ontologies .....	16
2.6 Drug Ontologies .....	17
2.7 Ontology Matching/Merging (OM).....	22
2.8 Semantic Web.....	25
2.9 Linked Data .....	26
2.10 Mashups.....	27
2.11 Pharmaceutical Discovery Research.....	30
2.12 Drug Information User and Resource Research .....	36
2.13 Basis for Current Work.....	37
Chapter 3. Methods.....	39
3.1 Q1: What are the Dimensions of Drug Information? .....	39
3.1.1 User warrant / domain focus. ....	39

3.1.2 Literary warrant.....	39
3.1.3 Resource survey. ....	40
3.1.4 Experimental database.....	44
3.2 Q2: Do Dimensions Lead to Valid Groupings of Resources? .....	51
3.2.1 Face validity.....	51
3.2.2 Correspondence analysis. ....	51
3.2.3 Cluster analysis. ....	51
3.3 Q3: Can Dimensions Facilitate Integration/OM Tasks?.....	52
3.3.1 Classifying sources.....	52
3.3.2 Selecting sources appropriate to a given information need.....	52
3.3.3 Pooling data from different sources. ....	52
3.3.3.1 Data reduction.....	52
3.3.3.2 Automatic normalization of additional raw data.....	53
3.3.3.3 Satisfying use cases.....	55
3.3.3.3.1 Criteria for usefulness.....	55
3.3.3.3.2 Health care and related personnel.....	57
3.3.3.3.3 Pharmaceutical discovery researchers. ....	59
3.3.3.3.4 Consumers. ....	61
Chapter 4. Results .....	65
4.1 Q1: What are the Dimensions of Drug Information? .....	65
4.1.1 Resource survey. ....	65
4.1.2 Experimental database.....	70
4.2 Q2: Do Dimensions Lead to Valid Groupings of Resources? .....	72
4.2.1 Face validity.....	72
4.2.2 Correspondence analysis. ....	73
4.2.3 Cluster analysis. ....	74

4.2.3.1 Initial resource survey.....	74
4.2.3.2 Experimental database. ....	77
4.3 Q3: Can Dimensions Facilitate Integration/OM Tasks?.....	81
4.3.1 Classifying sources.....	81
4.3.2 Selecting sources appropriate to a given information need.....	81
4.3.3 Pooling data from different sources. ....	82
4.3.3.1 Data reduction.....	82
4.3.3.2 Automatic normalization of additional raw data.....	86
4.3.3.3 Satisfying use cases.....	89
4.3.3.3.1 Health care and related personnel.....	89
4.3.3.3.2 Pharmaceutical discovery researchers. ....	106
4.3.3.3.3 Consumers. ....	138
Chapter 5. Discussion .....	150
5.1 What are the Dimensions of Drug Information? .....	150
5.2 Do Dimensions Lead to Valid Groupings of Resources? .....	151
5.3 Can Dimensions Facilitate Integration/OM Tasks?.....	153
5.3.1 Data reduction. ....	153
5.3.2 Automatic normalization of additional data.....	153
5.3.3 Satisfying use cases.....	155
5.4 Limitations.....	156
Chapter 6. Conclusions .....	160
6.1 Contributions to Drug Information.....	160
6.2 Contributions to Library and Information Science.....	161
6.3 Contributions to Semantic Web.....	162
6.4 Further Research.....	163
Appendix A. Potential Applications of a Drug Information System .....	165

Appendix B. Potential Test Cases for System Evaluation .....	168
Appendix C. Resource Evaluation Checklist.....	170
Appendix D. Resource Evaluation Summaries.....	175
Appendix E. Normalization .....	185
Appendix F. Use Case Adaptation and Query Execution.....	189
Appendix G. Dimensions Found in Experimental Database - 6-Level Hierarchy.....	200
References.....	215
Curriculum Vitae .....	224

## List of Figures

Figure 1. RxNav as of October 4, 2007. ....	21
Figure 2. Differences between portals and mashups.....	28
Figure 3. High-level illustration of a mashup. ....	30
Figure 4. High-level illustration of a mashup-like drug discovery tool.....	31
Figure 5. Drug-target network derived from a mashup-like drug discovery tool. ....	32
Figure 6. Another drug-target network derived from a mashup-like drug discovery tool. ....	34
Figure 7. Single target network derived from a web-based drug discovery tool. ....	35
Figure 8. Test queries to evaluate drug databases (Kupferberg & Jones Hartel, 2004). ....	59
Figure 9. Correspondence analysis between drug information sources and dimensions. ....	74
Figure 10. Cluster analysis of initial survey dimensions. ....	75
Figure 11. Cluster analysis of initial survey resources. ....	76
Figure 12. Cluster analysis of experimental database dimensions (2-level hierarchy). ....	79
Figure 13. Cluster analysis of experimental database sources. ....	80
Figure 14. Data reduction in the experimental database by hierarchical aggregation. ....	85
Figure 15. Online INN excerpt (top) and resulting copy-and-paste ASCII text format (bottom). .....	180

## List of Tables

Table 1. Semantic types of concepts in RXNORM. ....	19
Table 2. Relationships in RXNORM. ....	20
Table 3. Initial resource evaluations. ....	42
Table 4. Resources for systematic dimension analysis. ....	43
Table 5. Experimental database schema with structured data examples. ....	49
Table 6. Experimental database schema with free text examples. ....	50
Table 7. UMLS vs. DailyMed indications for ticlopidine hydrochloride. ....	63
Table 8. UMLS vs. DailyMed contraindications for ticlopidine hydrochloride. ....	64
Table 9. Dimensions of drug information by resources. ....	66
Table 10. Dimensions of drug information definitions. ....	67
Table 11. Generic name overlap estimates for some sources. ....	69
Table 12. Dimensions found in experimental database - 2-level hierarchy. ....	71
Table 13. Distribution of data and dimensions by top term and hierarchical level. ....	72
Table 14. Classifying resources by domains. ....	73
Table 15. Data reduction in the experimental database. ....	84
Table 16. Probabilities of correct automatic dimension normalization by source. ....	88
Table 17. Data reduction in query results for Health Use Case A. ....	92
Table 18. Target biological correlates for Research Use Cases A and B. ....	108
Table 19. Drug biological correlates for Research Use Case A. ....	109
Table 20. Research use cases dimensional coverage. ....	111
Table 21. Drug indications ("disease") for Research Use Case B. ....	116
Table 22. Drug biological correlates ("phenotype") for Research Use Case B. ....	117
Table 23. WHO-ATC cardiovascular classes with disease/bioprocess equivalents. ....	118
Table 24. Hypothetical new indications and therapeutic classes for dutasteride (Research Use Case C). ....	122

Table 25. Hypothetical new indications for terazosin (Research Use Case C).....	123
Table 26. Hypothetical new therapeutic classes for terazosin (Research Use Case C). ....	125
Table 27. Hypothetical new target biological correlates for dutasteride (Research Use Case C). .....	125
Table 28. Chemical characteristics of tamsulosin-like compounds in our database (Research Use Case D). ....	131
Table 29. Chemical characteristics of tamsulosin-like compounds (Research Use Case D). ....	134
Table 30. Deviation of chemical characteristics from those of tamsulosin by similar compounds (Research Use Case D). ....	135
Table 31. Side effects of Ticlid (Consumer Use Case H). ....	146
Table 32. Data reduction in use case query results. ....	149
Table 33. DrugBank vs. UMLS structured molecular target data.....	184

## Chapter 1. Introduction

### 1.1 Problem Statement

A central problem in information science, systems, and technology has been and remains the integration of data, information, and knowledge represented in different ways. Early information retrieval (IR) systems counted on human indexers to translate the free text of documents into concise sets of keywords or their codes, as much because of purely technical storage and search limitations as because of the uncontrolled semantics of free text. This is a sensible, productive integration solution and therefore continues to be widely used in practice. However, it has limited scalability, imperfect consistency across indexers, and even less consistency across keyword lists, indexes, and the like. Advances in IR research and computer technology have opened the way to more automated solutions, raising the possibility that human indexes and indexing could become obsolete. Yet a central tenet of the Semantic Web, widely regarded as the latest major IR breakthrough, is the assignment of meaning to web objects (including documents) by keyword-like tags in a human-like (if not literally human-mediated) way. For better or worse, this development seems to signify a kind of surrender to the inevitability of human-generated semantic codes (keyword lists, indexes, and the like) and judgments (document interpretation). Thus the problems associated with manual indexes and indexing remain important targets for research and technology development.

We investigated a new approach to this general issue of resolution between different representations of "reality" that can be considered a type of *ontology matching/merging (OM)*. The connection to the preceding paragraph consists of the historical and conceptual links between ontologies and traditional IR keyword lists, indexes, and the like, especially in the biomedical domain. We focused on drug information for this reason, because drug information is important, and because of our extensive experience with it. Existing drug ontologies are generally not as "formal" as their counterparts in most OM research, reflecting the poor state of drug information integration across resources. Assuming such "informal" ontologies can be "formalized," they can

be seen as "pre-formal." By extension, since many practical drug information resources do not employ any kind of explicit ontology, yet the way they organize or display their information *implies* one, their representations can be seen as being in an even earlier stage of ontology formalization.

We assert that the problem of *early, pre-formal OM* within a domain is a general one. Our research aims to address this problem by developing a specific method for resolution between different representations of "reality" across information resources in practical use within the drug domain. At its core this method depends on an ontology-like representation we call *dimensions of drug information*. Our specific objectives were to provide a plausible, empirical definition of the dimensions of drug information, and to test its validity and usefulness. Test methods included professional subjective judgments of information quality, correspondence analysis, cluster analysis, and building a model database that satisfied realistic use cases better than other existing resources according to objective, quantitative IR performance measures. We claim that our results will advance not only the state of drug information, but also provide a general framework for addressing the larger problem of early, pre-formal OM of which the drug information case is but one example.

## 1.2 Rationale

An *ontology* is an agreed upon, formal specification of a conceptualization within a domain; it is an answer to the question "What is there?"<sup>1</sup> in a domain. The "more formal"<sup>2</sup> an ontology is, the more powerful it is, not just at organizing and representing knowledge, but for automating tasks like integration and discovery. In many respects, ontologies resemble keyword

---

<sup>1</sup> Attributed to Willard Quine (1908-2000), Harvard professor of philosophy and mathematics. There does not seem to be a canonical published reference.

<sup>2</sup> The issue of "formal" vs. "informal" ontologies is examined in detail in Section 2.4. Here we offer a brief preview by quoting Mika, Iosif, Sure, and Akkermans (2004, pp. 461-462): "The term formal indicates the grounding of representation in some sort of well understood logic; i.e., ontologies go beyond simple vocabularies (terminologies) by providing definitions of concepts based on how they relate to other concepts and relationships. Lastly, formal also refers to the fact that ontologies may be expressed in machine processable formats, which makes them applicable to information systems."

lists, indexes, thesauri, faceted hierarchies, and other such traditional IR indexing tools. Like ontologies, these tools can become valuable information artifacts in their own right, organizing and representing vast amounts of raw text, bringing order and heightened comprehensibility to entire knowledge domains. Like ontologies, generating them can involve analyzing the *structure* of knowledge in a domain. In fact, many traditional thesauri and quasi-thesauri are increasingly being called "ontologies"; this is especially true in the biomedical domain.

Like their IR predecessors, ontologies are still overwhelmingly human-made and therefore diverse, even within domains. OM seeks to overcome the resulting barriers to integration, not just of the ontologies themselves but of the information bases they represent. But OM is hard, so practical shortcuts such as *linked data* and *mashups* are being investigated. However, even these require some semblance of ontological formality. What about domain knowledge that is not yet expressed in this way? What about the earlier, "pre-formal"<sup>3</sup> state when "What is there?" might have to be answered by informal ontologies or even free text? We propose that OM be understood to include such approaches.

In the case of drug information, multiple disparate resources are still the rule. "Disparate" refers to the resources' coverage (which drugs), scope (what kinds of drug information), presentation tactics (terms and relations, data tables, free text, ...), and intended users (patients, clinicians, pharmacists, researchers, ...). None present a comprehensive, fully integrated view, and no common directory is available to locate resources appropriate to a given purpose or user type. The high number, heterogeneity, and inadequate integration of drug information resources constitute barriers to many drug information usage scenarios.

One way to conceptualize this problem is that there is a need for OM in the drug domain. That is, the disparities across drug information resources can be viewed as differences in what they consider to be drug information/knowledge and how they represent it ontologically, whether

---

<sup>3</sup> This term derives from the foregoing footnote and the idea that "informal" ontologies can be "formalized" also reviewed in Section 2.4.

explicitly or implicitly. Examples of explicit ontologies include the Medical Subject Headings (MeSH), RxNorm, National Drug Formulary Reference Terminology (NDFRT), and World Health Organization Anatomic-Therapeutic-Chemical (WHO-ATC) drug classification.<sup>4</sup> An example of part of an implicit ontology would be one database's practice of populating a column named *brand name* with values such as "Tylenol"; "Bayer Aspirin"; "Proscar"; and "Viagra"; while another database puts the same values into a column named *trademark*.<sup>5</sup> Another example is the drug relationships to indications, contraindications, interactions, side effects, mechanisms, and chemistry specified as free text in their package inserts. If these disparate ontologies (and we will address the controversy of calling them that) could be somehow rationalized - unified, merged, cross-mapped, reconciled, or otherwise integrated - it would help to lower drug information usage barriers.

In the drug and related biomedical domains there is a rich legacy of a pre-formal, less-technical-than-conventional-OM phase or type of multi-ontology integration exemplified by RxNorm and the UMLS Metathesaurus. We assert an even earlier phase/type when "What is there?" is more about surveying the knowledge in a domain than formalizing it. Here ontologies may be implicit and poorly structured yet contain valuable knowledge that, as a consequence, resists integration. This "early 'what-is-there?'" phase/type of OM is what is now needed in the drug domain.

### 1.3 Research Questions and Strategy

We introduce the idea of *dimensions of drug information* as an approach to early OM in the drug domain. Like *facets* and other traditional classification constructs, dimensions' overarching practical mission is to bring order, or at least some measure of consistency, to knowledge abstraction, organization, representation, and integration. However, we wish to define

---

<sup>4</sup> Details and references for these and other resources will be supplied later in this document; see, e.g., Table 4.

<sup>5</sup> In this document we will follow this convention of italicizing the names of semantic types, variables, categories, dimensions, etc., while expressing examples of their values or instances in quotation marks.

it primarily by *extension* in the ontological sense; i.e., by what we find "out there" in the world of drug information. Therefore, our first challenge was to *identify* the dimensions of drug information. Restated as a research question,

Q1. What are the<sup>6</sup> dimensions of drug information?

To answer this question we inventoried a variety of information resources by examining database schemas, web pages, and query results. Their drug-related data elements (intensional content) and values (extensional content) were normalized into a set of dimensions of drug information. For example, the categories *brand name* and *trademark* and sets of values such as {"Proscar", "Propecia", "Bayer Aspirin", "Tylenol", ...} are all evidence of a source's coverage of the *trade names* dimension.

Next we *tested* the set of dimensions we identified. We used two kinds of tests: adequacy/face validity, and usefulness. To test the adequacy/face validity of our research answer to Q1, we investigated this research question:

Q2. Do dimensions lead to valid groupings of resources?

To answer this question, we manually classified the Q1 dimensions into four domains (pharmacy, chemistry, biology, and clinical medicine) and submitted the dimensions-by-resources mapping/matrix to correspondence analysis. We also submitted the unclassified dimensions-by-resources mapping/matrix to cluster analysis for comparison.

To test the usefulness of our research answer to Q1, we investigated this research question:

Q3. Can dimensions facilitate these integration/OM tasks?

A. Classifying sources

---

<sup>6</sup> While striving for generality we recognize the impossibility of universality in answering this question. This question should be understood as shorthand for "What comprehensive set of dimensions of drug information can we discover given our limited focus, and how might they be qualified to illuminate, in a general way, the bigger universe of *all* dimensions of drug information?"

This is basically a usefulness version of Q2; i.e., are the resource groupings suggested by the correspondence analysis and cluster analysis useful as well as valid? The criterion here is professional judgment.

*B. Selecting sources appropriate to a given information need*

Given the dimensions-by-resources mapping/matrix from Q1 and an equivalent mapping/matrix of dimensions to information needs, resources can be mapped, via dimensions, to information needs. This was demonstrated and its usefulness assessed by professional judgment.

*C. Pooling data from different sources*

We built a model database of integrated drug information from diverse sources on a small sample of drugs. The basis of integration is the normalized dimensions and their values, and we demonstrate how this facilitates data pooling by (1) quantifying, at several levels of granularity, the scatter of raw data relative to normalized dimensions and values (an example being the *trade name* example given earlier); (2) demonstrating how additional raw data could be added to the database and automatically normalized based on mechanized pattern-matching; and (3) testing the effectiveness of the database for satisfying a variety of “real world” use cases representing three user types: health care professionals, drug discovery research scientists, and consumers. Some of these use cases represent more complexity than pooling *per se*, such as cross-source search, clustering, cross-referencing, and interface reduction (querying one source and making use of the knowledge in others). The criteria here include objective measures of query efficiency (numbers of commands, queries, interfaces, keystrokes, ...) and retrieval quality (volume, consistency/variety, ...).

Our model database is valuable not just for defining (Q1) and demonstrating the validity (Q2) and usefulness (Q3) of dimensions of drug information, but also as a model or prototype for a much larger database of something closer to *all* drug information capable of satisfying many *kinds of* information needs. This is arguably the most important potential future extension of this work.

## Chapter 2. Literature Review

A central problem in information science, systems, and technology has been and remains the integration of data, information, and knowledge represented in different ways. We will examine library and information science (LIS) *classification* research which led to the practice of indexing information from human-created keyword/term/code lists including *thesauri*, a practice which remains widespread despite known drawbacks. One such drawback is the contradiction between the ever-changing nature of data, information, and knowledge, versus the practical need to keep the thesaurus stable. Post-coordinate thesauri address this problem by limiting terms to fundamental concept representations that can then be combined to represent more complex topics. *Facet analysis* builds on this insight by providing theory and methodology for organizing such terms in a consistent, logical, powerful way, resulting in *faceted thesauri*. *Ontologies* purport to take terminology to an even higher level by mapping it to concepts and relationships<sup>7</sup> that can represent the *meaning* of information artifacts more precisely, and can be computed upon to discover new knowledge. Since human world views and information system functional requirements are inherently variable, there will always be differing ontologies which need to be merged or otherwise integrated. Ontologies are only one means to the end of integration, however, so we also examine others including the *Semantic Web*, *linked data*, and *mashups*. Throughout these surveys we will review both general and biomedical/drug-domain-specific work.

### 2.1 Classification

Our work can be seen as an extension of LIS classification research. Classification fits into the grand scheme of LIS under Ranganathan's (1957) fourth law: *save the time of the reader*. By creating a concise index of all the books in the library (or, nowadays, other information

---

<sup>7</sup> In this document, "relationship" means the predicate in a subject-predicate-object triple such as (aspirin,*treats*,pain) while "relation" means the entire triple. Thus a relation (usually) contains two concepts and their relationship, and often a *back* or *inverse* relation contains the same two concepts reversed and a *back* or *inverse* relationship such as (pain,*is treated by*,aspirin).

objects in other repositories, such as web pages on the web) classified by their topics, one empowers users to find the information objects they want more rapidly by consulting the index, rather than having to depend on cross-references, grope around at random, or, potentially, look at every information object in the repository. (See Renear and Palmer, 2009, for a current, science-centered treatment.) Furthermore, the index can become a valuable information artifact in its own right, a kind of abstract, summary, or schema of the entire repository's information contents, bringing order and heightened comprehensibility. (See Soergel, 1999, for an extension of this idea to specific applications.) Thus a good index accurately *represents* and *organizes*.

Like libraries, early information retrieval (IR) systems counted on human indexers to translate the free text of documents into concise sets of keywords or their codes, as much because of purely technical storage and search limitations as because of the uncontrolled semantics of free text. The goal of indexing in IR is to group like objects (by assigning common index terms/codes) while differentiating them (with unique terms/codes or combinations) enough for individual retrieval. The *thesaurus* (Joyce & Needham, 1958; Foskett, 1980; Aitchison & Clarke, 2006) became the gold standard for structuring such term/code lists in a consistent, powerful way to accomplish these goals. Perhaps the best known biomedical thesaurus is the U.S. National Library of Medicine's (NLM) *Medical Subject Headings* (MeSH; NLM, 2009), which is essentially the modern descendent of the index to the first hardcopy *Index Medicus* (Billings, 1879, as cited by Wyman, 1999, p. 67), and is still used for manual indexing of NLM's massive Medline literature database.

This approach's problems soon became apparent (e.g., Shera, 1970, pp. 90-91). They include lack of scalability (too much information, too few indexers), lack of consistency (different document interpretations by different indexers), and a stubborn tendency of "standard"<sup>8</sup> thesauri and other quasi-indexes to mutate, proliferate, and clash (different human world views

---

<sup>8</sup> "The great thing about [vocabulary] standards is that there are so many of them." - Doris Schlichter, Merck vocabulary manager, late 1990's.

and preferences). Advances in computer technology opened the way to full text storage and search, permitting large-scale application of automation schemes, some of them quite old, including keyword extraction (Luhn, 1961), indexing (Doyle, 1961, 1962), relevance ranking (Maron & Kuhns, 1960; Salton & Buckley, 1988), machine learning of user preferences (Rocchio, 1966; Salton & Buckley, 1990), and natural language processing (Rau, 1988; Allen, 1995; Jurafsky & Martin, 2000).

These advances raised the possibility that human indexes and indexing could become obsolete, despite well-informed skepticism about the machine-centric approach to IR (Saracevic, 1975; Belkin, 1978; Salton, 1987; Swanson, 1988). Yet a central tenet of the Semantic Web, widely regarded as the latest major IR breakthrough, is the assignment of meaning to web objects (including documents) by keyword-like tags in a human-like (if not literally human-mediated) way. For better or worse, this development seems to signify a kind of surrender to the inevitability of human-generated semantic codes (thesauri, indexes, keyword lists, etc.) and judgments (document interpretation). Thus the problems associated with manual indexes and indexing remain important targets for research and technology development.

## **2.2 Facet Analysis**

Early thesauri were designed for "post-coordinate" indexing. That is, an attempt was made to limit terms to fundamental concept representations that could then be combined to represent more complex topics. Ironically, this seemed to make the terms harder to organize hierarchically.

A subject overview, or systematic display, if it existed at all, was of secondary importance in most thesauri in the early days. A detailed classified arrangement, as in an enumerative classification scheme, was considered too complex and outdated to have a role in postcoordinate information retrieval, where clear and simple terms were needed to be used in combination for optimum results... An exception was ... MeSH, that from early editions in the 1960s, placed value on its tree structures. (Aitchison & Clarke, 2006, pp. 9-10)

The need for detailed pragmatic hierarchies may have been one factor pulling later thesauri toward enumerative, "pre-coordinated" terminology.

Concurrently, Ranganathan's ideas about faceted classification were being refined, primarily by the Classification Research Group in England (CRG; e.g., B. C. Vickery) (Spiteri, 1998; Aitchison & Clarke, 2006). Facet analysis (FA) retains the insight of the early post-coordinate thesaurus designers (that there are inherent advantages to limiting terms to fundamental concept representations that can then be combined to represent more complex topics) and attempts to solve the classification problem by rigorous semantic analysis.

The technique analyses complex subjects into constituent categories of the same inherent type. These fundamental categories include actions, comprising processes and operations; entities, such as natural objects, products, materials; agents, including personnel and equipment; and time and place. In the field of education, for instance, there would be a teaching methods facet, arising from the operations category; an educational personnel facet from the personnel category; a teaching aids facet from the equipment category, and so on. The facets are mutually exclusive, and the terms within each facet share a common characteristic. A 'facet indicator,' or 'node label,' is often inserted to name that common characteristic. The terms so grouped tend to be short and simple, and may then be used in combination with other simple terms to express compound subjects, either postcoordinately when indexing using a thesaurus, or as precoordinated class marks in the context of a faceted classification scheme. (Aitchison & Clarke, 2006, p. 11)

The two threads intersected in 1969 as *Thesaurofacet*, the first thesaurus to use FA.

Unlike some faceted thesauri to follow, *Thesaurofacet* managed to reconcile the practical need for polyhierarchality (concepts with multiple broader concepts) with the FA maxim of mutual exclusivity.

Kwasnik (1999) implicitly ties FA to ontologies by situating the former within knowledge representation structures based on entities and their relationships, with the goal of new "knowledge discovery and creation" (p. 22). She describes the *process* of classification as the "clustering of experience" (p. 24), one of many processes that contribute to knowledge accumulation, representation, and expansion.

The process of classification can be used in a formative way and is thus useful during the preliminary stages of inquiry as a heuristic tool in discovery, analysis, and theorizing... A good classification [*structure*] functions in much the same way that a theory does, connecting concepts in a useful structure. (p. 24)

FA yields one type of such a structure; a given instance is "good" if it is useful; i.e., "descriptive, explanatory, heuristic, fruitful, and perhaps also elegant, parsimonious, and robust" (p. 24).

Other parallels between FA and ontologies are striking. Both are concerned with abstracting, representing, organizing, and formalizing knowledge, especially the *essence* of "what things are" (e.g., *blue* is a *color*). Both acknowledge the domain-dependence of knowledge and the consequent need for diversity and flexibility while striving for unification. Both have practical goals which often conflict with the yearning for order, logical rigor, theoretical coherence, and universality. Both communities aspire to have their models become "the basis of all" - IR in the case of faceted classification (Broughton, 2006, p. 49), web-based integration in the case of ontologies. Both have a classical, specific meaning (and its defenders) which is threatened (or even, at least in the case of ontologies, already eroded) by popular usage. The CRG backed away from Ranganathan's universal top categories (Personality, Matter, Energy, Space, and Time [where "Personality" should be understood as "Entities" {Broughton, 2006}]) but universal "upper ontologies" are still quite respectable in some quarters (Sowa, n.d.; Pease, 2009). For that matter, "a classification by any other name is still a classification" (Soergel, 1999, p. 1120) and ontologies in some ways represent a wasteful reinvention of what that larger field of LIS has already learned (Soergel, 1999). Vickery (1997) and Broughton (2006) explicitly address some FA-ontology parallels from a CRG perspective.

Given these parallels, it is surprising how rarely FA is mentioned in the ontology literature and *vice versa*. The obvious boolean *AND* search produced zero hits on PubMed<sup>9</sup> and only one hit (Tudhope & Binding, 2008) on ISI Web of Knowledge<sup>10</sup> in September, 2009. Exceptions (discovered using Google Scholar<sup>11</sup>) include (Vickery, 1997; Soergel, 1999; Aitchison & Clarke, 2006; Broughton, 2006).

Spiteri (1998) summarizes Ranganathan's highest level description of FA, the *Three Planes of Work*. In the *Idea Plane* one analyzes a subject field into its component parts. In the *Verbal Plane* one chooses appropriate terminology for the component parts. The *Notational*

---

<sup>9</sup> <http://www.ncbi.nlm.nih.gov/pubmed/>

<sup>10</sup> <http://isiwebofknowledge.com/>

<sup>11</sup> <http://scholar.google.com/>

*Plane* deals with expressing the component parts using a "notational device." Clearly these are tasks we undertook with regard to drug information. In that general sense, and in the exploratory knowledge engineering spirit conveyed by Kwasnik, we perhaps did a kind of facet analysis. But consider the following excerpt from the NISO standard for thesaurus construction:

Facet analysis is particularly useful for:

- new and emerging fields where there is incomplete domain knowledge, or where relationships between the content objects are unknown or poorly defined;
- interdisciplinary areas where there is more than one perspective on how to look at a content object or where combinations of concepts are needed;
- vocabularies where multiple hierarchies are required but can be inadequate due to difficulty in defining their clear boundaries; or
- classifying electronic documents and content objects where location and collocation of materials is not an important issue. (NISO, 2005, p. 14)

This sounds even more like our situation with drug information, yet Broughton (2006) implies that NISO ("rather typically US" p. 61) has misunderstood facet analysis, perhaps further implying that we are not doing it.

Facet analysis is the topic of a current ARIST review (La Barre, 2010).

### **2.3 Domain Analysis**

In the information science (IS) context, domain analysis was put forward by Hjørland and Albrechtsen (1995) as the methodological correlate of a more sociological alternative to other IS metatheories, paradigms, and viewpoints.

The domain-analytic paradigm in information science (IS) states that the best way to understand information in IS is to study the knowledge-domains as thought or discourse communities, which are parts of society's division of labor. Knowledge organization, structure, cooperation patterns, language and communication forms, information systems, and relevance criteria are reflections of the objects of the work of these communities and of their role in society... The domain-analytic paradigm is thus firstly a social paradigm... The domain-analytic paradigm is secondly a functionalist approach, attempting to understand the implicit and explicit functions of information and communication and to trace the mechanisms underlying informational behavior from this insight. Thirdly it is a philosophical-realistic approach, trying to find the basis for IS in factors that are external to the individualistic subjective perceptions of the users... (p. 400)

Among their intellectual forerunners the authors credit Saracevic (1975) whose analysis of relevance included a "subject literature view of relevance" and "subject knowledge view of

relevance," the latter being "fundamental to all other views of relevance, because subject knowledge is fundamental to communication of knowledge" (Saracevic, 1975, p. 333). Hjørland and Albrechtsen go on to note that "the focus on domain-specific cognitive functioning represents a very strong current tendency" (p.405) and "[t]here is neither a simple dichotomy between structure and content nor between relevant and irrelevant information" (p.406). The following example (credited to Putnam) is illustrative:

Everyone to whom gold is important for any reason has to acquire the word "gold"; but he does not have to acquire the method of recognizing if something is or is not gold. He can rely upon a special subclass of speakers. The features that are generally thought to be present in connection with a general name - necessary and sufficient conditions for membership in the extension, ways of recognizing if something is in the extension ("criteria"), etc. - are all present in the linguistic community considered as a collective body; but that collective body divides the labour of knowing and employing these various parts of the "meaning" of "gold." This division of linguistic labour rests upon the division of non-linguistic labour (Putnam, 1975, 245). (Hjørland and Albrechtsen, 1995, p. 408)

That is, "classifications of knowledge domains cannot be regarded as independent of knowledge claims" (Hjørland and Albrechtsen, 1995, p.409) and "subject representation/classification (the inner side of information systems)" is the "best example of applications" of the domain-specific view (p. 412).

According to the domain-analytic framework, the meaning of a term can only be understood from the context in which it appears. The meaning of a term such as gold can only be understood by an interpretation of the discourse in which that term appears. Gold has at least one chemical meaning (a heavy metal, difficult to dissolve by acids, electrical leading, etc.), one economic meaning (conventional economic measurement and reserve), one fictional meaning (related to wealth, happiness, the half kingdom and princess), etc. What other terms would be related to gold in a thesaurus depends entirely on the function served by a particular thesaurus. Whether documents retrieved by that term in an algorithm would be relevant to a question depends entirely upon whether that term has one or another of its possible meanings. (p. 413)

One line of research should occupy itself with the use of language in different domains. What kind of culture exists concerning the form of titles, the pattern of citations, etc.? What are the consequences for the informational value of titles, subject terminology, descriptors, and citations in IR? What important transdisciplinary tendencies and concepts exist in the disciplines? (p. 419)

We did not (intentionally) do sociological research, but many of the concepts and ideas mentioned by Hjørland and Albrechtsen are relevant, including "[k]nowledge organization,

structure" (p. 400); "language and communication forms" (p. 400); "implicit and explicit functions of information" (p. 400); "factors that are external to the individualistic subjective perceptions of the users" (p. 400); "[no] simple dichotomy between structure and content" (p. 406); "necessary and sufficient conditions for membership in the extension, ways of recognizing if something is in the extension" (p. 408); classifications' dependence on knowledge claims; thesaurus relations dependence on function; "informational value of titles, subject terminology, descriptors, and citations" (p. 419); and "transdisciplinary tendencies and concepts" (p. 419). Domain analysis was used as the basis for a recent ARIST review on pharmaceutical information (Bawden & Robinson, 2010).

Regarding Saracevic (1975), certainly we are implicitly concerned with relevance (of domains to usage scenarios, of dimensions to domains, of resources to domains and usage scenarios, of query results to information needs), and our methodology depends on both my "subject knowledge" and the "subject literature" represented by the drug information we examined.

## 2.4 Ontologies

The most widely cited definition of *ontology* in this context is that given by Gruber (1993): "An ontology is an explicit specification of a conceptualization" (p. 199).

"Conceptualization" in this context is at least quasi-synonymous with "knowledge"; hence ontologies are a kind of *knowledge representation* (KR). However, *ontology* has a long and varied history and lately has become a buzzword in information technology, so other definitions abound. Three basic senses may be distinguished in the literature:

1. Ontologies as sets of *categories* (also called *types*, *classes*, or *concepts*). An *ontology* is one such set; *ontology* is the study of such sets (Sowa, n.d.). Categories are distinct from the *symbols* (e.g., words) that represent them, and also from the *instances* or *objects* that they in turn represent. The set of categories in an ontology are related to each other by *relationships*, usually pairwise, that form a *lattice*, *tree structure*, or *hierarchy* conveying subdivision and inheritance of

properties. This sense of ontologies is the narrowest and the oldest, having changed little from its origins in nineteenth century western philosophy. It is the sense employed by computer scientists who reserve the term “ontology” for the highest (i.e., most general, fundamental, immutable, consequential) levels of knowledge specification (e.g., Powers, 2004).

2. “Formal ontologies.” There is nothing informal about the classical sense #1 but it can be made more powerful by adding such information as class and relationship *attributes*, *constraints*, and *compositional syntax* for creating new categories by logical combination of existing categories. There is clearly an overlap between this sense of ontologies and the computer science notion of *conceptual models* (CMs), although at some point a CM clearly can become more than an ontology (e.g., after the addition of instances or event definitions). This sense of ontologies is represented in the writings of Franconi. Two views of CMs are given by Boman, Bubenko, Johannesson, and Wangler (1997) and Borgida and Brachman (2002). The appellation “formal ontologies” may be odious to those who disagree that anything “informal” should be called an ontology, but it is becoming common to refer to this sense in that way.

3. Terminological ontologies (Sowa, n.d.). Justly or not, the term “ontologies” is being co-opted and applied increasingly to what used to be called thesauri, controlled vocabularies, subject heading lists, and just about any other form of organized terminology (e.g., Bard & Rhee, 2004). This sense is distinct from sense #2 in that the so-called ontologies were not constructed to support conventional computerized logical reasoning and probably in most cases are not adequately specified to do so (e.g., Kashyap & Borgida, 2003). However, calling them ontologies anyway conveys respect for their knowledge content and a kind of faith that they *can* be “formalized” (e.g., Hahn, 2003; Williams & Anderson, 2003).

A pragmatic centrist view was offered by Mika, Iosif, Sure, and Akkermans (2004).

By their most common definition, ontologies represent a shared and formal understanding of a domain. In this definition, the term shared refers to the fact that ontologies embody a consensus among members of a given community. The term formal indicates the grounding of representation in some sort of well understood logic; i.e., ontologies go beyond simple vocabularies (terminologies) by providing definitions of concepts based

on how they relate to other concepts and relationships. Lastly, formal also refers to the fact that ontologies may be expressed in machine processable formats, which makes them applicable to information systems. (pp. 461-462).

We can incorporate these ideas into Gruber's definition as follows: an ontology is an *agreed upon, formal*, explicit specification of a conceptualization *within a domain*. *Formal* signifies both rigorous standardization and machine processability, while *agreed upon* and *within a domain* clarify the human, social (i.e., arbitrary) nature of any conceptualization. The key idea here is that ontologies should go beyond "simple vocabularies" (including traditional thesauri) by specifying *how* concepts relate to other concepts.

## 2.5 Bio-ontologies

In a high-profile review, Bard and Rhee (2004) coined the term "bio-ontologies" and gave strong voice to a growing movement to refer to traditional pragmatic biomedical thesauri, controlled vocabularies, and classification systems as ontologies despite their obvious differences from the more formal ontologies of philosophy, logic, and computer science. Debate over this tendency (Smith & Welty, 2001; Ceusters, Smith, & Flanagan, 2003; Bodenreider, Smith, & Burgun, 2004; Goble & Wroe, 2004; Stevens, Wroe, Lord, & Goble, 2004; Soldatova & King, 2005; Charlet, 2007) seems to be subsiding as it becomes more entrenched (Cimino & Zhu, 2006; Mabee et al., 2007; Noy et al., 2009) and historically rationalized (Bodenreider & Stevens, 2006). Another research thread attempts to reconcile the differences by attempting to "formalize" the knowledge contained in bio-ontologies. Such efforts have targeted the UMLS Metathesaurus (Hahn & Schulz, 2004), UMLS Semantic Network (Kashyap & Borgida, 2003), Gene Ontology (Williams & Andersen, 2003), National Cancer Institute Thesaurus (Golbeck et al., 2003), MeSH (Soualmia, Golbreich, & Darmoni, 2004), International Classification of Diseases (Heja, Surjan, Lukacsy, Pallinger, & Gergely, 2007), Systematized Nomenclature of Medicine Clinical Terms (Bodenreider, Smith, Kumar, & Burgun, 2007), and Foundational Model of Anatomy (Golbreich, Zhang, & Bodenreider, 2006). Some of these and other ontologies have been collected under an umbrella named Open Biomedical Ontologies (Smith et al., 2007).

The premier (biggest and most widely known, researched, and used) bio-ontology is the U.S. National Library of Medicine's (NLM) Unified Medical Language System® (UMLS) (NLM, 2007b). Although commonly called an ontology, UMLS is actually three distinct tools for dealing with biomedical terminology. The Semantic Network comes closest to being a formal ontology but contains only 189 (2009AA version) very high level concepts. The Specialist Lexicon is not an ontology at all, but rather a set of natural language processing tools tailored to biomedical text. The Metathesaurus is what most users mean by the "UMLS ontology." It covers names and inter-relationships for millions of concepts but is more properly viewed as an integrated cross-mapping of over 100 distinct biomedical terminological ontologies. NLM, therefore, by producing, maintaining, and expanding the UMLS Metathesaurus, has nearly two decades of practical experience with "pre-formal" ontology matching/merging (OM).

## **2.6 Drug Ontologies**

The "unique nature of pharmaceutical information ... [its] breadth of scope, plus its economic and social importance, lends pharmaceutical information a unique place within information science" (Bawden & Robinson, 2010, pp. 63, 66). "[B]ecause of its numerous and diverse users and sources, its technical advances, and its economic and social significance, [it] has played a major role in advancing information science itself" (p. 94). The advance of ontology-intensive information systems from bio-ontologies into the pharmaceutical domain has been much anticipated (Meyer, 2002; Hug et al., 2004; Aronson & Ferner, 2005; Feldman, Dumontiera, Linga, Haider, & Hogue, 2005; Gardner, 2005; Mendrick, 2006).

NLM supports such research in the context of development of drug data interchange standards and information resources. Its best-known product is RxNorm (Bodenreider & Nelson, 2004; Liu, Ma, Moore, Ganesan, & Nelson, 2005; Zeng, Bodenreider, Kilbourne, & Nelson, 2006; Zeng, Bodenreider, Kilbourne, & Nelson, 2007). RxNorm is a compendium of drug information (primarily terminology) from different sources such as First DataBank, Micromedex, Medi-Span, and Multum. It is designed to mediate messages between computerized systems that

use different drug vocabularies so as to ease inter-operability in recording or processing data dealing with clinical drugs. RxNorm contains standard names for clinical drugs, both branded and generic, with cross-references to their active ingredients, drug components, related brand names, National Drug Codes (NDCs), and nomenclature from other drug vocabularies (NLM, 2007a). This RxNorm (mixed case) must be distinguished from RXNORM (all caps), the "Level 0" (unrestricted) source of normalized drug names in UMLS. RxNav is NLM's web-based RXNORM browser and is freely available at <http://mor.nlm.nih.gov/download/rxnav/>.

RXNORM can be thought of as an ontology of drug concepts and their relationships. "An [RXNORM] description can be understood as a graph whose nodes are UMLS concepts corresponding to ingredients, drug forms, dose forms, etc. The relationships among these elements [are the] edges in the graph" and RxNav employs this graphical representation to facilitate browsing of RXNORM (Bodenreider & Nelson, 2004, p. 1530). The semantic types of the concepts found in RXNORM are shown in Table 1 and the relationship types in Table 2, and a RxNav screenshot in Figure 1.

A current research focus at NLM is on adding more (than RXNORM terminology) drug information to RxNav, including pharmacologic action, drug-drug interactions, indications, contraindications, and adverse reactions (Zeng et al., 2007). This is a nucleus around which may crystallize our conceptualization of dimensions of drug information; that is, examples of such dimensions include *terminology*, *pharmacologic action*, *drug-drug interactions*, *indications and contraindications*, and *adverse reactions*. Furthermore, this NLM research effort can be viewed as enhancing RxNav based on dimensions, evidence that our thesis addresses an important topic.

**Table 1. Semantic types of concepts in RXNORM.**

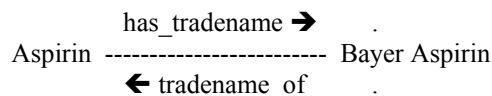
The semantic types are expressed as (combinations of) dimensions we found in an initial study (Table 9, Table 10), showing that only IN, DF, and BN map one-to-one to them. There does not appear to be a pure *combination generic name* semantic type in RXNORM.

<i>Semantic type</i>	<i>Concept example</i>	<i>Liu et al. (2005) term<sup>a</sup></i>	<i>Zeng et al. (2007) count</i>
generic name	Aspirin	IN	5,604
dose form	Oral Solution	DF	140
generic name + dose	Aspirin 100 MG	SCDC	13,509
generic name + dose form	Aspirin Oral Solution	SCDF	8,311
generic name + dose + dose form	Aspirin 100 MG Oral Solution	SCD	17,726
trade name	Platet AA&C	BN	11,363
trade name + dose	Platet 100 MG	SBDC	13,460
trade name + dose form	Platet Oral Solution	SBDF	11,033
generic name + dose + dose form + trade name	Aspirin 100 MG Oral Solution [Platet]	SBD	14,064
combination generic name			
combination generic name + dose			
combination generic name + dose form	Acetaminophen / Aspirin / Caffeine Oral Tablet		
combination generic name + dose + dose form	Aspirin 100 MG / Caffeine 32 MG Oral Tablet		
combination generic name + dose + trade name	Aspirin 400 MG / Caffeine 32 MG [AA&C]		
combination generic name + dose form + trade name	Acetaminophen / Aspirin / Caffeine Oral Tablet [Excedrin Geltab]		
combination generic name + dose + dose form + trade name	Aspirin 400 MG / Caffeine 32 MG Oral Tablet [AA&C]		

<sup>a</sup> • IN ingredient. This is a compound or moiety that gives the drug its distinctive clinical properties. Examples: Fluoxetine, Insulin, and Isophane. • DF dose form. Example: Oral Solution. • SCDC semantic clinical drug component. This represents the ingredient plus strength. Example: Fluoxetine 4 MG/ML. • SCDF semantic clinical drug form. This represents the ingredient plus dose form. Example: Fluoxetine Oral Solution. • SCD semantic clinical drug. This represents the ingredient plus strength and dose form. Example: Fluoxetine 4 MG/ML Oral Solution. • BN brand name. This is a proprietary name for a family of products containing a specific active ingredient. Example: Prozac. • SBDC semantic branded drug component. This represents the branded ingredient plus strength. Example: Prozac 4 MG/ML. • SBDF semantic branded drug form. This represents the branded ingredient plus dose form. Example: Prozac Oral Solution. • SBD semantic branded drug. This represents the ingredient, strength, and dose form, plus brand name. Example: Fluoxetine 4 MG/ML Oral Solution [Prozac].

**Table 2. Relationships in RXNORM.**

In this document, "relationship" means the middle part of a declarative triple while "relation" means the entire triple; e.g., for the relation (*Aspirin*, *has\_tradename*, *Bayer Aspirin*), the relationship is (*has\_tradename*). Note that relations and relationships are usually directional (exceptions include "has\_synonym") and come in inverse pairs such as

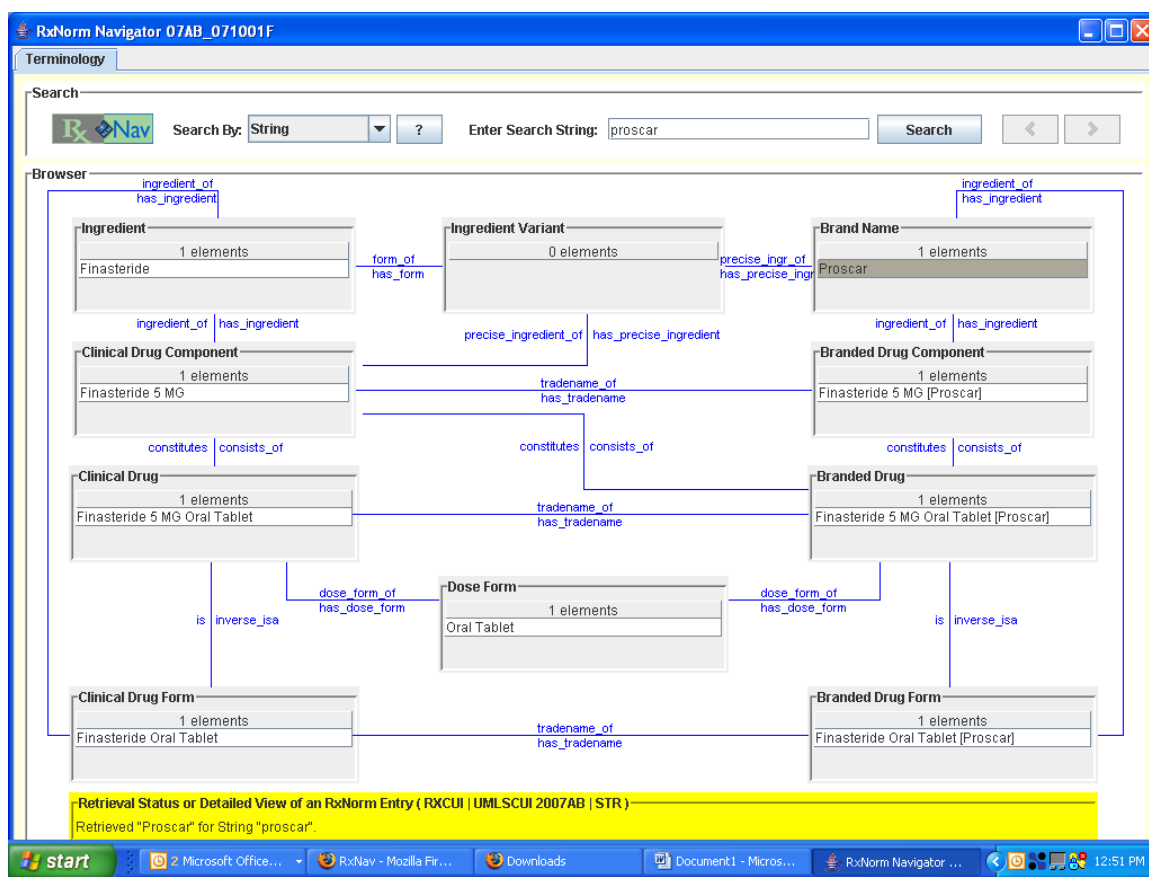


This terminology is distinct from the UMLS terms "relationship" (*rel*) for traditional thesaurus relationships and "relationship type" (*rela*) for the more specific relationships shown in this example. RxNav uses only the latter. The numbers count both directions ( $\rightarrow$  and  $\leftarrow$ ) and were computed using UMLS 2007AC. RN related narrower, RB related broader, RO related other, SY synonym.

<i>rel</i> $\rightarrow$	<i>rela</i> $\rightarrow$	$\leftarrow$ <i>rel</i>	$\leftarrow$ <i>rela</i>	<i>count</i>
RO	reformulation_of	RO	reformulated_to	58
SY				160
RN	form_of	RB	has_form	3030
RO	has_precise_ingredient	RO	precise_ingredient_of	1586
RN	has_precise_ingredient	RB	precise_ingredient_of	4990
RN	isa	RB	inverse_isa	98856
RB	has_tradename	RN	tradename_of	156248
RO	has_dose_form	RO	dose_form_of	156732
RO	constitutes	RO	consists_of	182974
RO	has_ingredient	RO	ingredient_of	199606

NLM is experimenting with three additional RxNav sources: DailyMed, MedMaster, and the National Drug Formulary Reference Terminology (NDFRT, a UMLS source terminology).

The first two use a mostly free text knowledge representation that differs markedly from RXNORM's standardized relationships that RxNav is designed to visualize, so these will be integrated (or "pseudo-integrated") by a "linkout" strategy whereby clicking on a drug term in the RxNav display allows the user to navigate to the DailyMed or MedMaster webpage about that drug. NDFRT uses standardized relationships (Carter et al., 2006) and so is amenable to knowledge base (KB) warehouse integration and RxNav directed graph display, but screen clutter may force it to be displayed separately (via a "tab" to another screen view) from the RXNORM knowledge (O. Bodenreider, personal communications, 2007-2008).



**Figure 1. RxNav as of October 4, 2007.**

"Ingredient" and "Ingredient Variant" correspond to the *generic name* dimension. Other dimensions represented here include *dose* ("5 MG"), *dosage form* ("Oral Tablet"), and *trade name* ("Proscar").

Aside from RXNORM and RxNav, precedent for a comprehensive drug ontology is rather thin. The University of Manchester group behind the GALEN medical ontology also developed a separate Drug Ontology based on information from the *British National Formulary* (Solomon et al., 1999) but it is no longer being maintained or promoted on the web. I attempted to extract a drug ontology from the UMLS with limited success (Sharp, 2005). The science publisher Elsevier has sponsored a Drug Ontology Project for Elsevier (DOPE) (de Waard, Fluit, & van Harmelen, 2007). NDFRT is the only drug ontology among 166 bio-ontologies listed by BioPortal<sup>12</sup> (Noy et al., 2009). Specialized chemical ontologies are beginning to appear (Ben-

<sup>12</sup> <http://bioportal.bioontology.org/ontologies> as of October 1, 2009.

Miled et al., 2002; de Matos et al., 2004; Feldman et al., 2005; Taylor et al., 2006; Degtyarenko et al., 2008) but their acceptance and utility remain open questions.

## 2.7 Ontology Matching/Merging (OM)

Multiple ontologies are inevitable both within and across domains due to the variability of human world views and information system functional requirements. This is not only a logical assertion but also a lesson of decades of incomplete-to-nonexistent ontological hegemony.<sup>13</sup> To obtain additive benefits one must find ways to integrate multiple differing ontologies; that is, combine or merge them so that common concepts and relations are represented in a unified way while the non-overlapping pieces "go where they should" and conflicts are resolved democratically, if not to everyone's satisfaction. In other words, one hopes to approach the ideal result where the merged ontology combines all the knowledge, usefulness, and consensus of its component ontologies without losing any of their coherence or logical rigor. We see this generic process as a legitimate view of OM which can include efforts such as our own, RxNorm, and the UMLS Metathesaurus, in addition to the more algorithmic approaches that dominate the current literature on ontology (or schema) matching, mapping, merging, or alignment *per se*.

Hameed, Preece, and Sleeman (2004) use the term *ontology reconciliation*. This article stands out from the rest of the OM literature in giving serious consideration to "why people and organizations will tend to use different ontologies, and why the pervasive adoption of common ontologies is unlikely" (p. 231). The authors' reasons include multiple competing ontologies that purport to serve the same purpose, legacy data/systems that depend on diverse ontologies, differing human "perspective[s] on the world" (p. 232) and their multiple levels (personal, corporate, national, etc.), and the need for growth and change (i.e., ongoing struggle to preserve whatever reconciliation can be achieved now). To this list I would add proprietary knowledge that corporations need in their ontologies but don't want to divulge to outsiders (regardless of

---

<sup>13</sup> e.g., the U.S. National Cancer Institute's *NCI Thesaurus*, itself a breakaway derivative of MeSH, still has, by NCI's own admission, no less than 14 competing cancer-related ontologies ([http://bioportal.nci.nih.gov/ncbo/faces/pages/ontology\\_list.xhtml](http://bioportal.nci.nih.gov/ncbo/faces/pages/ontology_list.xhtml)).

shared world views), and the usual politics of anything human (protecting jobs, reputations, bureaucratic turf, etc.).

Lambrix and Tan (2006) give a particularly good example of the OM literature in the biomedical domain. They review and classify existing OM systems using the following framework (1-4) and similarity-computing ("matcher") strategies (a-f):

1. Automated computation of similarity values between terms.
2. Automated suggestion of term-term alignments based on (1.).
3. Human review and acceptance/rejection of (2.).
4. Automated conflict/redundancy checking of (3.).

Step 1 employs various similarity-computing ("matcher") strategies, including:

- a. *Linguistic* approaches make use of textual descriptions of the concepts and relations such as names, synonyms and definitions. Similarity is measured by string matching (n-gram, stemming, edit distance, etc.) and sometimes frequency counting or other IR tactic. Most systems use these strategies.
- b. *Structure-based* approaches use the graph structure or equivalent (*is-a*, *part-of*, or other relations). The similarity of concepts is based on their environment (e.g., two concepts with the same hierarchical parents and/or children would have high similarity).
- c. *Constraint-based* approaches are based on *axioms*, a characteristic of the most formally specified ontologies. For example, similarity may be defined by common *domain/range* (what can be a superclass/subclass; e.g. *female/mother*) or *disjointness* (nonoverlap; e.g., *female/male*). This approach has limited power and applicability but is used by a few systems.
- d. *Instance-based*: For example, Lambrix and Tan's system, SAMBO, uses the co-occurrence of terms and annotations in biomedical literature (PubMed) to suggest alignments between Gene Ontology (GO) and Signal Ontology (SigO) terms. However,

there is often a gray area between instances and subclasses, potentially making this approach a subtype of (b.).

- e. *Use of auxiliary information* such as dictionaries, thesauri, intermediate ontologies, or prior OM experience that provide knowledge to interpret the meaning of the concepts and relations to be aligned. Many systems use auxiliary information. SAMBO uses the UMLS Metathesaurus' conceptual ("CUI") conflation of terms from different biomedical ontologies.
- f. *Combination approaches*: "Although most systems combine different approaches, not much research is done on the applicability and performance of these combinations" (p. 199).

An exception to the foregoing statement (f) was presented by Zhang, Mork, Bodenreider, and Bernstein (2007). They evaluated two approaches to aligning two anatomy ontologies. Both approaches used a combination of lexical and structural techniques. In addition, the first approach used supplementary domain knowledge, while the second employed principles of schema matching. Lessons for improvement were learned, but only 33% of the possible one-to-one concept matches were identified by the two approaches together. "New directions need to be explored in order to handle more complex matches" (p. 227). Performance evaluation was also addressed by Giunchiglia, Yatskevich, Avesani, and Shvaiko (2009) who created a large (thousands) test dataset called TaxME2 by combining the Google, Yahoo, and Looksmart web directories using "almost two-dozen state-of-the-art ontology matching systems" (p. 137). Ontology alignment evaluation is institutionalized as the Ontology Alignment Evaluation Initiative (OAEI; <http://oaei.ontologymatching.org/>). The "OM bible" is (Euzenat & Shvaiko, 2007) (O. Bodenreider, personal communication, September 14, 2009).

Thus, the OM of the OM literature is a phase or subtype of our more generic view of OM. The former is characterized by heavy algorithmic/computational assistance, suitable for dealing

with explicit, well-structured ontologies where the domain is well defined and benchmarks (UMLS, Google, etc.) are available for reference and result evaluation. Logically, this leaves room for an earlier, less algorithmic/computational phase/subtype exemplified by the UMLS Metathesaurus and RxNorm. We assert an even earlier phase/subtype when "What is there?" is more about surveying the knowledge in a domain than formalizing it. Here ontologies may be implicit and poorly structured ("informal") yet contain valuable knowledge that, as a consequence, resists integration. A clear example is the drug relationships to indications, contraindications, interactions, side effects, mechanisms, and chemistry specified as free text in their package inserts. This is the *early, pre-formal* phase/type of OM that is now needed in the case of drug ontologies, and to which we see our work contributing.

## 2.8 Semantic Web

The Semantic Web (SW) is a vision (widely credited to Berners-Lee, Hendler, and Lasilla, 2001) of a future World Wide Web (WWW) where IR will be enhanced by being based more on *meaning* (semantics) than it is today. The construct has been attacked (Shirky, 2003) and defended (Bray, 2003; Wright, 2003) and continues to stimulate interest and work, including in the biomedical domain (e.g., Schroeder & Neumann, 2006, and accompanying articles) and debate (Legg, 2006; Lenz, Beyer, & Kuhn, 2007). The biomedical domain effort is now represented by the Semantic Web Health Care and Life Sciences (HCLS) Interest Group (<http://www.w3.org/blog/hcls>).

The SW involves many technologies that are beyond the scope of this review. A comprehensive overview of the biomedical SW is given by Bodenreider (2009). We have already alluded to the irony that the SW seems to be re-discovering manual indexing from human-created thesauri. SW-related ontology research is in part motivated by this irony; i.e., it hopes to improve upon traditional thesauri. Our work is part of this thread in the "early 'what is there?'" sense of ontologies. Two other SW technologies are relevant: *linked data* and *mashups*.

## 2.9 Linked Data

In contrast to the WWW, where links are between documents/webpages based on hypertext (HTML), the idea behind linked data is to link "arbitrary things described by RDF."<sup>14</sup> The arbitrary things can be any kind of object or concept; like webpages, they are represented by URIs.<sup>15</sup> "With linked data, when you have some of it, you can find other, related, data" (Berners-Lee, 2006) without having to wade through the irrelevant parts of the webpages where the data resides. That means machines (software "agents") can find the related data in a more straightforward and reliable way, potentially leading to advances in automation of tasks such as "plan and set up my vacation" or "invent new drugs for disease X." A centralized information resource is <http://www.linkeddata.org>.

For example, RxNav's developers would like to be able to link the RXNORM drug names to their approved indications, contraindications, interactions, side effects, mechanisms, and chemistry in the same way that it links generic names to trade names within RXNORM. Then a user could, for example, start from drug A, retrieve drug A's indication, then retrieve the names of all the other drugs approved for that indication. However, these relationships are currently available only as specified in free text in the drugs' package inserts (this is one of my findings). The closest RxNav can come is to link each drug name to its entire package insert via the latter's URI on the DailyMed website, which does not link to any other drug names. What is needed is to translate the insert text into RDF "triples" such as {"aspirin"; *has\_indication*; "headache"}, where "aspirin"; *has\_indication*; and "headache" are all represented by URIs that may include other information such as synonyms, definitions, permitted relationships, etc.

Ideally, NLM would like to map the DailyMed data to a ready-made, universally accepted drug ontology made up of (or convertible to) such RDF triples based on (and linked to) all other drug information information resources. Since such an ontology does not exist, a

---

<sup>14</sup> RDF = resource description framework, a formal ontology language.

<sup>15</sup> URI = universal resource identifier; i.e., a web address; e.g., <http://www.rutgers.edu>.

reasonable compromise would be to make up a new ontology adequate to cover the knowledge in DailyMed. This would permit RxNav to improve upon its current functionality (users could then find all the drugs *in DailyMed* with the same indication), but in a limited way (they would miss drugs *not* in DailyMed with the same indication). Such pragmatic gradations allow developers to balance tangible progress with universality. That is, data linking, like traditional bio-ontologies, does not entirely avoid the formalization problem in OM, but allows for pragmatic flexibility so that useful advances can be made incrementally. Data linking can be considered a type of early, pre-formal OM (O. Bodenreider, personal communication, December 21, 2010).

## 2.10 Mashups

[A mashup is] a web page or application that combines data or functionality from two or more external sources to create a new service. The term mashup implies easy, fast integration, frequently using open APIs<sup>16</sup> and data sources to produce results that were not the original reason for producing the raw source data.<sup>17</sup>

Mashups are thought to be a potential replacement for portals, the current leading content aggregation technology. Some differences are shown in Figure 2.

Of interest to us is the ability of mashups to operate on pure XML<sup>18</sup> content, and their "melting pot" as opposed to "salad bar"/side-by-side presentation style. Using the foregoing example, if DailyMed could translate its free text information into an XML ontology, a DailyMed-RxNav mashup could result in all their common drugs' relationships "melted" together in a virtual queryable database, as contrasted to the current, minimally integrated, side-by-side display which cannot be queried by *indications*, *contraindications*, *side effects*, etc.

---

<sup>16</sup> API = application programming interface, a kind of software for extracting information from websites.

<sup>17</sup> [http://en.wikipedia.org/wiki/Mashup\\_\(web\\_application\\_hybrid\)](http://en.wikipedia.org/wiki/Mashup_(web_application_hybrid))

<sup>18</sup> XML = eXtensible markup language, a formal ontology language, more versatile than RDF.

	Portal	Mashup
Classification	Older technology, extension to traditional Web server model using well defined approach	Using newer, loosely defined "Web 2.0" techniques
Philosophy/Approach	Approaches aggregation by splitting role of Web server into two phases: markup generation and aggregation of markup fragments	Uses APIs provided by different content sites to aggregate and reuse the content in another way
Content dependencies	Aggregates presentation-oriented markup fragments (HTML, WML, VoiceXML, etc.)	Can operate on pure XML content and also on presentation-oriented content (e.g., HTML)
Location dependencies	Traditionally content aggregation takes place on the server	Content aggregation can take place either on the server or on the client
Aggregation style	"Salad bar" style: Aggregated content is presented 'side-by-side' without overlaps	"Melting Pot" style - Individual content may be combined in any manner, resulting in arbitrarily structured hybrid content
Event model	Read and update event models are defined through a specific portlet API	CRUD operations are based on REST architectural principles, but no formal API exists
Relevant standards	Portlet behaviour is governed by standards JSR 168, JSR 286 and WSRP, although portal page layout and portal functionality are undefined and vendor-specific	Base standards are XML interchanged as REST or Web Services. RSS and Atom are commonly used. More specific mashup standards are expected to emerge.

**Figure 2. Differences between portals and mashups.**

Source: Wikipedia ([http://en.wikipedia.org/wiki/Mashup\\_\(web\\_application\\_hybrid\)](http://en.wikipedia.org/wiki/Mashup_(web_application_hybrid))) September 18, 2009.

The IEEE First International Workshop on Socio-Technical Aspects of Mashups was held in April, 2010. Its announcement stated, in part,

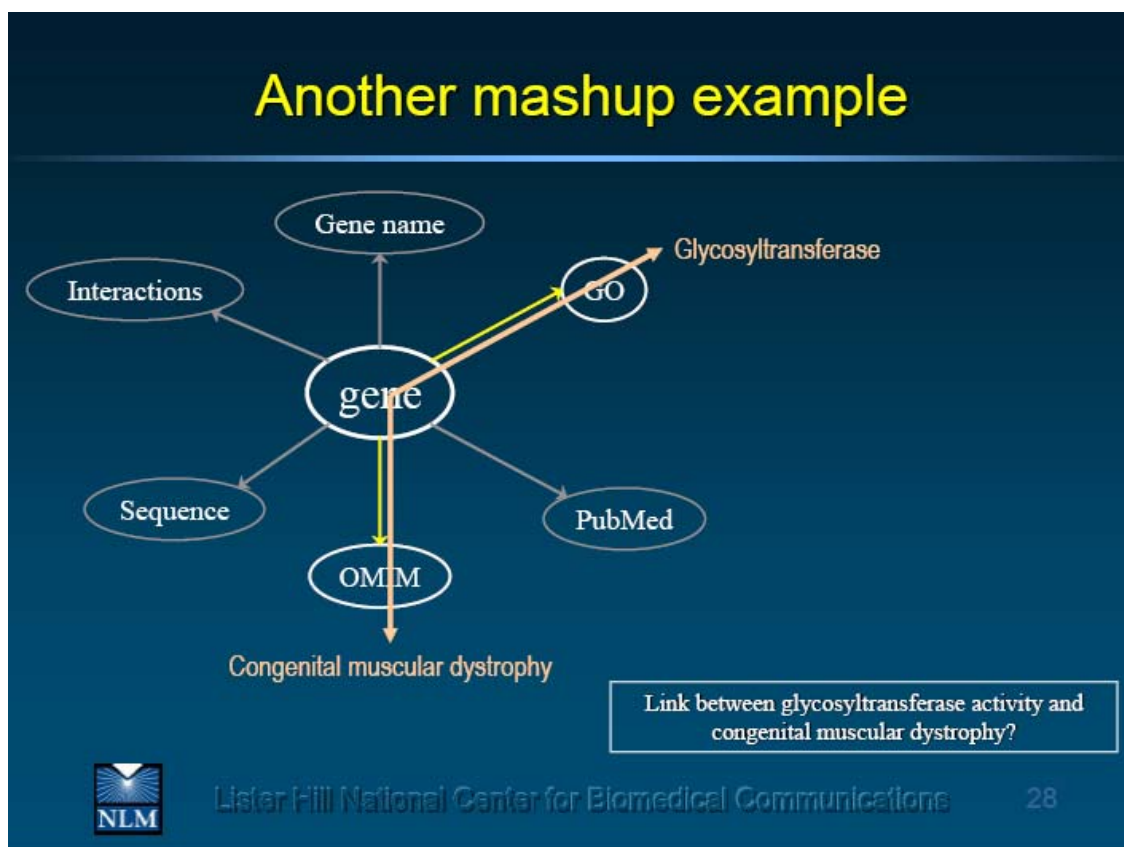
Mashups can bring data and functionality together in different ways and for different purposes. Many would argue that their greatest potential is for addressing transient problems for specific groups of users in dynamically changing business, social or political contexts. Mashups can be created by people, who may or may not be skilled in programming, to test out 'self-service' whenever needed through integrating heterogeneous information across the boundaries of different organizations over the Web. The implication of such an end-user development approach is far-reaching and hence deserves extensive scientific investigation. However, beyond all the hype, studies of the actual development and use of mashups for delivering business, social or political value are extraordinarily rare. While previous studies have focused on the technical side of constructing the Mashups infrastructure, little has been reported to demonstrate the real value or identify the problems, practicalities and pitfalls of their construction. Essentially, we need to understand how mashups emerge and change, succeed or fail, in settings where people, policies, systems, and data are intertwined with each other, forming a complex yet dynamic system...<sup>19</sup>

Thus mashups, like data linking, are (1) an approach to achieving practical increments in knowledge integration in advance of full formalization of domain ontologies, and (2) a technology for implementing the kind of "early 'what is there?'" OM we did in the drug domain.

<sup>19</sup> <http://www.aina2010.curtin.edu.au/workshop/stamashup/>

In the biomedical domain, mashups were the topic of a March 2008 special issue of the *Journal of Biomedical Informatics* (volume 41, number 5). Cheung et al. (2008) pointed out that mashup toolkits such as Dapper, Google Maps, and Yahoo! Pipes do not actually perform "most of the system integration" (p. 683); that is, the semantic part. "There is a need for creating mashups that better enable computers to help people achieve more powerful and complex data integration involving semantic mappings across multiple information models, terminologies, and ontologies. The term for such machine-based integration of data is 'semantic mashups.'" (p. 683).

Another term for a semantic mashup is *semantically integrated resources* (Sahoo, Bodenreider, Rutter, Skinner, & Sheth, 2008), precisely our goal. The mashup presented by Sahoo et al. (2008) could identify "hub genes" whose gene products participate in many pathways or interact with many other gene products. This utility of this is illustrated in Figure 3 (from Bodenreider, 2009). The user is able to discover a mechanistic connection between a disease represented in one ontology, database, or information resource (OMIM = Online Mendelian Inheritance in Man), and an enzyme represented in another (GO = Gene Ontology), due to their common association with the same gene. If Sahoo et al.'s mashup "works" for a particular instance of this use case, it is because the two ontologies happen to use the same name for the hub gene, and because their gene name *dimensions* (by whatever name) are immediately identifiable as equivalent. Thus "ontology-driven integration represents a flexible, sustainable and extensible solution to the integration of large volumes of information. Additional resources, which enable the creation of mappings between information sources, are required to compensate for heterogeneity across namespaces" (p. 752). That is, if the two ontologies in Figure 3 *do not* happen to use the same gene name value for the hub gene, or their gene name dimensions are *not* immediately identifiable as equivalent, the mashup needs these additional mappings.



**Figure 3. High-level illustration of a mashup.**

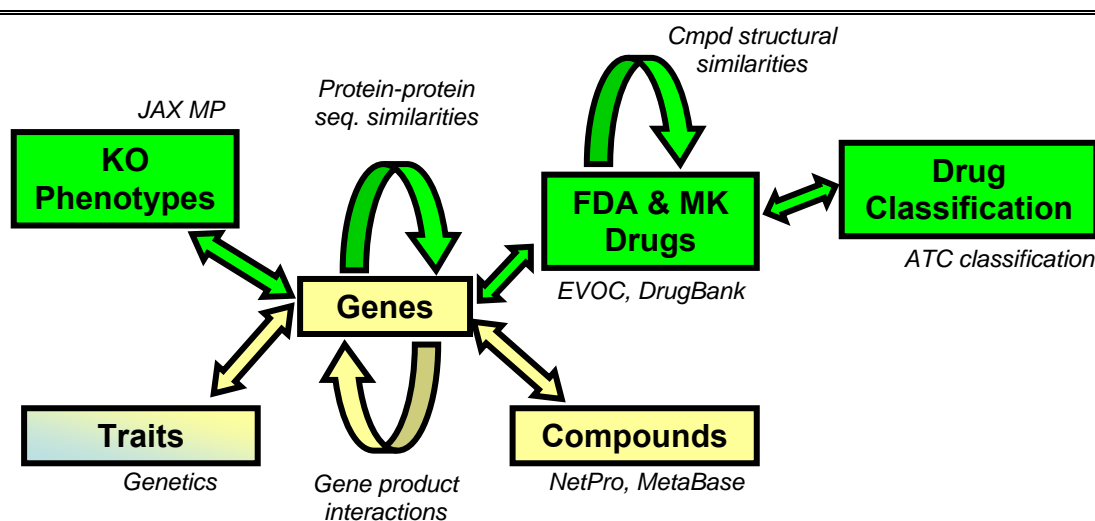
OMIM = Online Mendelian Inheritance in Man; GO = Gene Ontology. Source: Bodenreider (2009).

Our research aims to provide such additional resources. The gene that a drug interacts with (often, in pharmacology, lumped together with enzymes, receptors, and other such *molecular targets*) could be a dimension of drug information. Mechanistic connections between *molecular targets* and diseases (which can be *indications*, *contraindications*, or *side effects*) drive much pharmaceutical discovery research. To supplement traditional chemical and biological "wet bench" laboratory approaches, such research is beginning to experiment with ontology-driven knowledge base integration and discovery prototypes that resemble mashups in many respects.

### 2.11 Pharmaceutical Discovery Research

Castle, Shah, Avila, Derry, and Rohl (2007) enhanced a Target and Gene Information Network Analysis Visualization (TGI-NAV) tool to better support drug discovery research; specifically to help "connect diseases/phenotypes to genes, and, through genes, disease states to

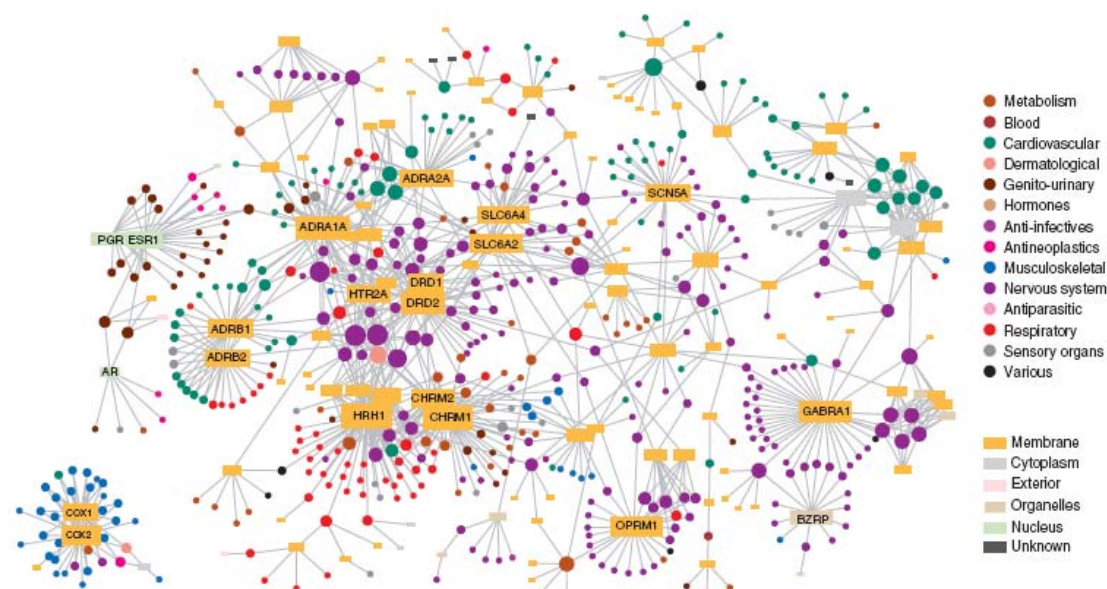
existing drugs." NAV graphically displays information from the TGI knowledge base, a synthesis of public and Merck proprietary information. Castle et al. added three new sets of nodes (phenotypes/disease states, drugs, and drug therapeutic activity classifications) and five sets of edges (phenotype-to-gene mappings, drug-to-target gene mappings, and drug-to-therapeutic activity classifications, plus compound-to-compound structural and protein-to-protein sequence similarity mappings). The high-level ontology is illustrated in Figure 4. The similarity to a mashup (Figure 3) is striking. Castle et al. showed the utility of their enhancements by discovering a possible relationship between anti-nauseant drug activity and abnormal pain threshold, and that the target gene set associated with cardiovascular drugs is enriched in phenotypes associated with heart disease. Two of Castle et al.'s resources, DrugBank and WHO-ATC, are included in our work as well.



**Figure 4. High-level illustration of a mashup-like drug discovery tool.**

Conceptual diagram of data types available in TGI NAV, where green boxes and arrows denote new data types described in Castle et al. (2007). JAX MP KO: Jackson Labs Mammalian Phenotype Knockouts; FDA: U.S. Food and Drug Administration; MK EVOC: Merck Electronic Vocabularies; ATC: WHO's Anatomical Therapeutic Chemical drug classification. Phenotypes are associated with genes based on sequelae when the gene is mutated (knocked out). Gene-gene relationships reflect sequence similarities while compound-compound associations reflect chemical similarities. Associations between genes and drugs include the intended targets of FDA-approved and FDA-experimental drugs (DrugBank) and Merck (MK) drug candidates. Drugs classification relationships are based on ATC. Gene-trait relationships are from Genetics Bayesian Networks. From Castle et al. (2007).

Almost identically, Yildirim, Goh, Cusick, Barabási, and Vidal (2007) built a bipartite graph of drugs and proteins linked by drug-target binary associations from DrugBank and drug therapeutic classifications from WHO-ATC. The resulting network (Figure 5) was considered a model for the "global set of relationships between protein targets of all drugs and all disease-gene products in the human protein-protein interaction or 'interactome' network" (p. 1119). It showed a strong local clustering of drugs of similar types according to WHO-ATC. Topological analyses of this network showed an overabundance of "follow-on" drugs (drugs that target already targeted proteins) but, by including drugs currently under investigation, a trend toward more functionally diverse targets ("polypharmacology") was seen. Etiological (disease-mechanism-based) and palliative (symptom-based) drugs were independently differentiated by a shortest distance measure.



**Figure 5. Drug-target network derived from a mashup-like drug discovery tool.**

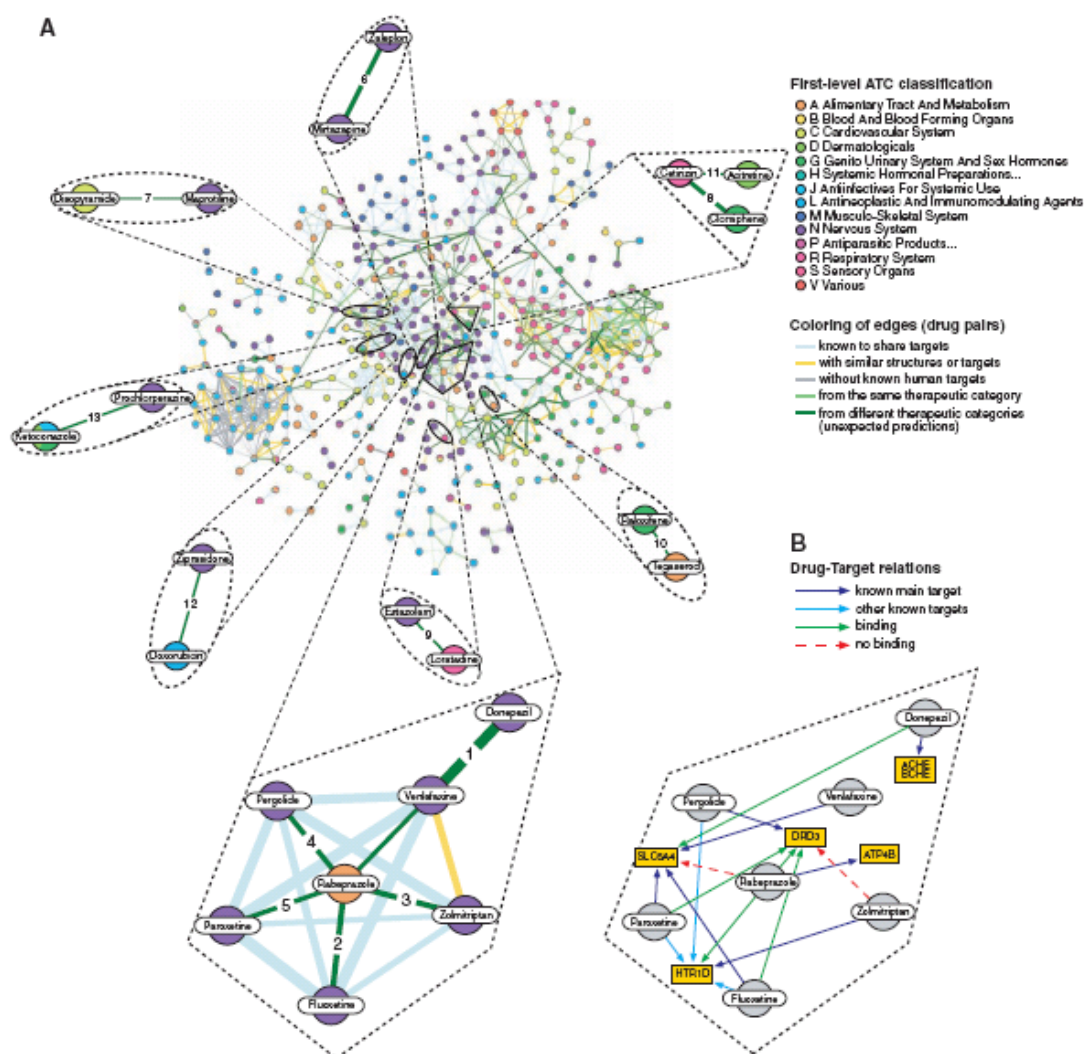
The network is generated by using the known associations between FDA-approved drugs and their target proteins (DrugBank). The area of the drug node (circles) is proportional to the number of targets (rectangles) that the drug has and vice versa. Drug nodes are colored according to their WHO-ATC classification, and the targets according to their cellular component obtained from the Gene Ontology database. From Yildirim et al. (2007).

In a third study of this type, Campillos, Kuhn, Gavin, Jensen, and Bork (2008) used graphical cluster analysis to discover new indications and therapeutic classes for existing drugs based on their common side effects by inferring (from side effect clusters) unknown molecular targets that are unexpected from indication, therapeutic class, or chemical structure similarity. By effectively "mashing up" (my phrase) DrugBank, WHO-ATC, UMLS, and other drug information resources according to what we call the dimensions of *generic name*, *therapeutic class*, *indication*, *side effect*, *molecular target*, and *chemical structure/similarity*, Campillos et al. created a database of 1018 "side effect driven drug-drug relations" for 746 marketed drugs and discovered 261 unexpected target predictions (Figure 6). Some of these predictions were tested by biochemical and physiological wet bench experiments, confirming 13/20 and 9/11, respectively.

These three papers represent a "warehouse" integration approach. In this approach, drug information in the form of terminologically controlled drug relations with other entities (targets, therapeutic classes, diseases, ...) is proactively copied from other sources' KB's and integrated into a local KB which can then be queried, visualized graphically, or cluster analyzed. Thus they do not fully follow the *distributed data* ideal of the Semantic Web, data linking, or mashups, which prefers to use APIs and other techniques to assemble the data dynamically at query time.

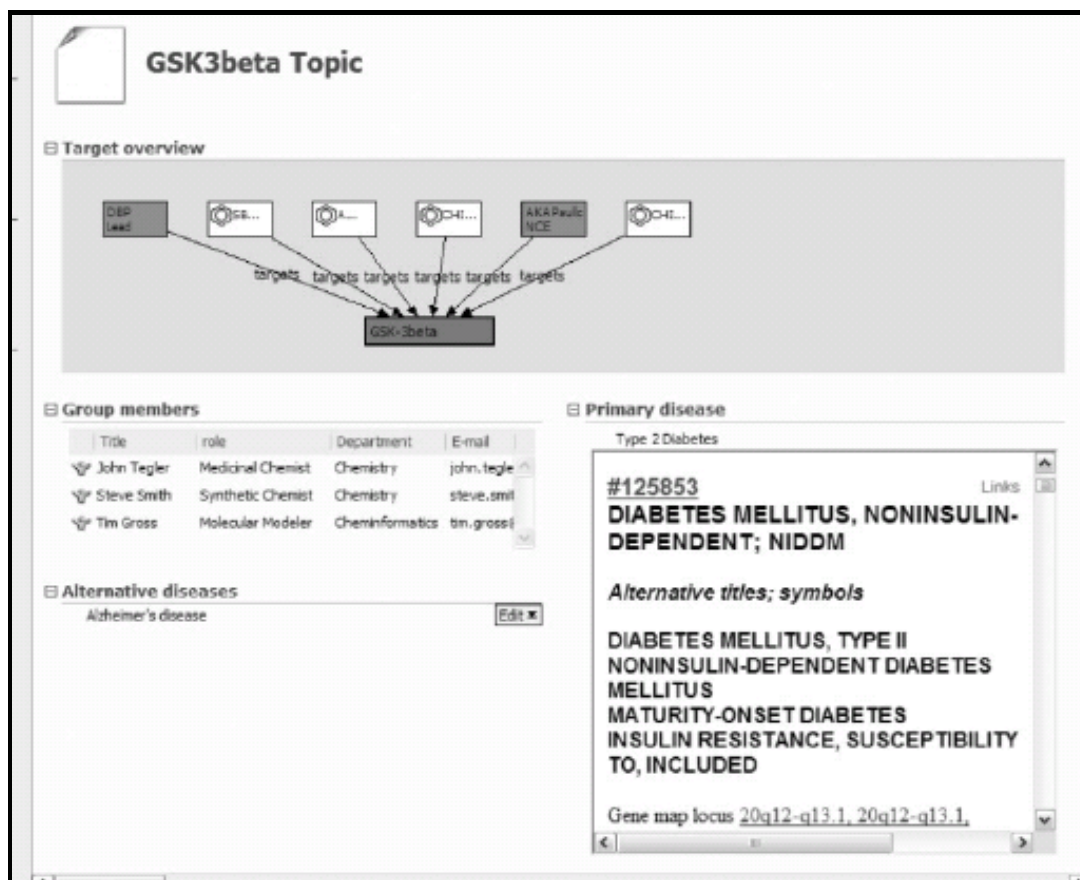
In contrast, Quan (2007) advocated a true Semantic Web approach and built a prototype drug discovery "dashboard" called "BioDash" as an example. Rather than controlled terminology and relation KBs, BioDash uses RDF "to access data from various heterogeneous data sources as if they had come from a single source. The user interface taps into this unified view of these data sources to display diagrams that show the cross-connections that exist within these data" (p. 175). To avoid information overload, "semantic lenses" (a form of computer-assisted query refinement) group together pieces of information that are relevant to a specific task. An example might be [my interpretation, not Quan's]: "find physical properties of X" translates to "find all Y where X p

Y and  $p = \{has\_melting\_point, has\_boiling\_point, \dots\}$ ." The semantic lenses are then assembled into "powerful information displays" (Figure 7). Quan's semantic lenses (*target overview*, *primary disease*, *alternative diseases*, and *group [people] members*), like the network node and edge types of the prior three papers, are comparable to our dimensions.



**Figure 6. Another drug-target network derived from a mashup-like drug discovery tool.**

(A) 424 drugs (nodes) form 1018 pairs with strong side-effect similarity and above 25% probability of sharing a target (edges, width proportional to probability). Drug subnetworks around the antiulcer drug rabeprazole and other experimentally confirmed predictions are magnified. (B) Selected drug-target relations in the subnetwork around rabeprazole. Predicted drug-target relations that were experimentally validated are shown with green arrows; dashed red arrows indicate that the predicted targets could not be confirmed. The confirmed relations are sufficient to prove the predicted drug-drug relations in the rabeprazole subnetwork. From Campillos et al. (2008).



**Figure 7. Single target network derived from a web-based drug discovery tool.**

Drug, disease, and human stakeholder relations are shown in this "topic view" of the molecular target GSK3beta from Bio-Dash corresponding to "semantic lenses" (*target overview, primary disease, alternative diseases, and group members*) which can be magnified to show greater detail (Quan, 2007).

The application to finding new indications for existing drugs was highlighted by Boguski, Mandl, and Sukhatme (2009). The authors call this goal "repurposing" and propose that it can play a major role in the "major overhaul of the R&D paradigm" (p. 1394) that many believe is needed in the drug industry. Past examples include successful "off-label" uses of drugs discovered by serendipity or, in some cases, using knowledge about the biological pathways and mechanisms of drug effects and diseases.

Because of our increasingly sophisticated understanding of human biology and the molecular pathways of disease, one would expect there to be increasing opportunities for expanding off-label use based on fully elucidated *pathways* and *mechanisms of action*, a situation that has been called a 'new grammar of drug discovery.' (p. 1394)

The authors call for a more concerted, systematic approach to repurposing discovery involving a new use of postmarketing surveillance information, consumer-driven data, and advanced information technology to discover beneficial as well as adverse *side effects*. In the foregoing two sentences, italics are added to highlight potential dimensions of drug information. In the case of *side effects*, Boguski et al. are in effect calling for a comprehensive national or international database comparable to, but much larger than, the manually created prototype of Campillos et al. (2008).

## **2.12 Drug Information User and Resource Research**

User studies in IR system research date back to the dawn of modern IR itself (Wilson, 1981). Nevertheless, giving equal consideration to "the human in the loop" (Kantor, n.d.) was considered a "user-centered" alternative to traditional, technology-centered IR research when Belkin (1978) and others began to develop it in the 1970's. This view later evolved into a so-called "cognitive approach" to IR which attempts to integrate the user- and system-oriented traditions. This approach places user-system interaction (rather than query-artifact matching) at the center of its IR situation model (Dervin & Nilan, 1986; Belkin, 1993; Saracevic, 1996).

Even within the biomedical domain, communities differ significantly in their IR needs and practices with special regard to negotiating meaning, provenance and ownership, group identity, and common knowledge. Advancing biomedical knowledge depends on accommodating these differences to allow inter-community knowledge sharing and "standing on shoulders" (Neumann & Prusak, 2007, p. 145).

A task-centric approach can be considered an extension of the user-centric approach.

The Web succeeded in large part because it allows users to download information from an ever-broadening range of sources through a single tool, the web browser... However, one area in which the Web is still lacking is in enabling users to consume information in aggregate... A more task-centric approach to information retrieval is required in order for our ability to consume information on the Web to scale with the growth of the Web itself... The naive approach to aggregation is to simply take all available information and put it onto one page. This approach may work for small information spaces, but for most life science problems, the naive approach readily leads to information overload, since much of this information is bound to be extraneous to the task at hand. The key to

eliminating extraneous information - and hence addressing information overload - is to make use of knowledge of the task at hand. (Quan, 2007, p. 172)

Specific user-type-centric drug information resource evaluations have included clinical decision support (Clauson, Marsh, Polen, Seamon, & Ortiz, 2007; Clauson, Polen, & Marsh, 2007), pharmacy students, faculty, and librarians (Kupferberg & Jones Hartel, 2004), and consumers (Plovnick & Zeng, 2004; Scott-Wright, Crowell, Zeng, Bates, & Greenes, 2006; Zeng, Crowell, Plovnick, Kim, Ngo, & Dibble, 2006; Keselman, Logan, Arnott Smith, Leroy, & Zeng-Treitler, 2008). We use the query sets developed by Kupferberg and Jones Hartel (2004) and Plovnick and Zeng (2004) in some of our evaluations.

### **2.13 Basis for Current Work**

The foregoing introduction and literature review situate *dimensions of drug information* as an approach to early OM, and thus a legitimate topic for LIS as well as drug informatics. That is, our research fills a gap in both literatures. The rationale can be summarized as follows.

- Ontologies (an offshoot, in some respects, of LIS classification research), are a preferred tool for integrating the kind of heterogeneous resources that now characterize drug information.
- There is no single, comprehensive, generally accepted drug ontology and may never be one due to drug information's large volume, rapid turnover, and diverse user needs and viewpoints.
- Drug discovery researchers are attempting to integrate disparate information resources based on the partial ontologies implicit in databases such as DrugBank and WHO-ATC, but this approach fails to capture canonical knowledge contained in such resources as DailyMed.
- *Dimensions of drug information* are, in some ways, an attempt to extract a comprehensive drug ontology from available drug information resources. The ways that these resources

represent drug information are "ontology-like" to varying degrees (hence unifying them is a kind of OM) but only in a primitive way (hence "early, pre-formal OM").

- This approach can be seen as a kind of domain analysis in the sense of analyzing how the drug information community of practice "sees" drug information, as suggested by their resources' early, pre-formal ontologies. The connection between domain analysis and OM is novel, as far as we know.
- Following most of the bio-ontology and drug informatics literature, Semantic Web, mashups, and linked data, we have a practical focus on information integration, as opposed to formal ontology development.
- Therefore our evaluations focus on the adequacy/face validity and usefulness of dimensions from an integration perspective.

## Chapter 3. Methods

For a brief overview of methods, please see the Research Strategy section (1.3).

### 3.1 Q1: What are the Dimensions of Drug Information?

**3.1.1 User warrant / domain focus.** We focus on a subset of drug information relevant to four domains: pharmacy, chemistry, biology, and clinical medicine. This focus was driven by a set of queries, applications, and information needs (Appendices A-B) that reflect the needs of specific user types - consumers, clinicians (physicians, nurses, etc.), pharmacists, and biomedical researchers - known to us through decades of study and professional experience. Examples include: finding equivalent *drug names* (e.g., *generic name* version of a *brand name*, or the *chemical name* of the *active ingredient*); finding alternative drugs for a given *indication* or vice versa; identifying drug *contraindications*, *precautions*, *warnings*, *side effects*, and *interactions*; and finding other drugs with the same or related *chemical properties* or *biological mechanisms*. Not considered are queries, applications, and information needs from manufacturing, marketing, legal, regulatory, financial, and other domains involving dimensions such as drug *pricing*, *retailers*, *packaging*, and *patents*. The italicized terms in this paragraph and in Appendices A-B represent our user-driven preconceptions of dimensions of drug information. Our breakdown of drug information user types and domains closely and independently parallels that of Bawden and Robinson (2010, pp. 65-66).

**3.1.2 Literary warrant.** In Appendix A, it can be seen that some of the applications are represented in published literature, giving them an additional "real world" legitimacy independent of our professional experience. The "proto-dimensions" in Appendix A are evident because they conform to our intuitive intensional definition: like *facets*, *categories*, *features*, and other traditional classification constructs, dimensions' overarching practical mission is to bring order, or at least some measure of consistency, to knowledge abstraction, organization, representation, and integration. The proto-dimensions similarly evident from our literature review can be summarized as follows.

- RxNorm (Bodenreider & Nelson, 2004; Liu et al., 2005; Zeng, Bodenreider, et al., 2006; Zeng et al., 2007): *terminology, generic names, active ingredients, drug components, brand names, National Drug Codes (NDCs), ingredients, drug forms, dose forms*. Also see Table 1, Table 2, and Figure 1.
- Zeng et al. (2007): *pharmacologic action, drug-drug interactions, indications, contraindications, adverse reactions*.
- Castle et al. (2007): *diseases/phenotypes, genes, disease states, drug [names], drug therapeutic activity classifications [WHO-ATC], compound structure, protein [target] sequence*.
- Yildirim et al. (2007): *drug [names], proteins, drug-target associations, drug therapeutic classifications [WHO-ATC], diseases, etiological drugs, palliative drugs*.
- Campillos et al. (2008): *generic names, side effects, molecular targets, indication, therapeutic class, chemical structure*.
- Quan (2007): *target, primary disease, alternative diseases, group [people] members*.
- Boguski et al. (2009): *indications, off-label indications, biological pathways, mechanisms of drug effects/action, mechanisms of diseases, molecular pathways of disease, adverse side effects*.

**3.1.3 Resource survey.** We considered approximately 30 drug information sources (Table 3) identified through our experience or their cross-references. In addition to dimensions, we assessed some of the technical characteristics of each source with implications for usage and integration, such as cost, database availability and update frequency, presentation (terms and relations, tables, free text, etc.), integration options such as application programming interfaces (API), and number and overlap of single-component generic names covered (Appendix C-D). Overlap was determined by lexical matching using the Merck drug name dictionary and

autoencoding system (which includes generic name, trade name, synonym, and Chemical Abstracts Service [CAS] registry number relations).

This resource set was narrowed to 23 (Table 4) based on criteria such as electronic availability, presence of explicit data elements, and balancing our desired domain and user-type coverage. We inventoried the 23 sources' features, derived from drug-related data elements (intensional content), or from their values (extensional content). This was done by examining database schemas, web pages, and query results. These tests often consisted of probing the source with a term representing some prototypical drug with certain expected results. Next we normalized the features into a set of dimensions of drug information. For example, categories such as *brand name* and *trademark* and sets of values such as {"Proscar", "Propecia", "Bayer Aspirin", "Tylenol", ...} are all evidence of a source's coverage of the *trade names* dimension. Finally, we grouped the dimensions by the four domains of interest and mapped them to the sources as a matrix of mostly binary (1 or 0) scores.

**Table 3. Initial resource evaluations.**

"Coverage" refers to the number of drug concepts equivalent to an RxNorm "ingredient" (~generic name). "Info+" is a preliminary judgment of whether the source offers a desirable (for our purposes) information increment over UMLS and the other sources. "Extract" and "link" are preliminary judgments of whether it is feasible and desirable to integrate the information into RxNav explicitly (requiring extracting it from the source) or via some kind of API or hyperlink. See Table 4 for website references.

<i>source</i>	<i>version</i>	<i>coverage</i>	<i>cost(\$)</i>	<i>avail.</i>	<i>info+</i>	<i>extract</i>	<i>link</i>
RXNORM	Zeng et al. (2007)	5,604	0	yes			
UMLS	Nov-2007 (AC)	tbd	0	yes	n/a	yes	n/a
DailyMed/SPL	11/12/2007	3,440 <sup>a</sup>	0	yes	yes	no	yes
Drugs@FDA	11/18/2007	1,689	0	yes	no		
DrugDigest	11/18/2007	>5,000 <sup>b</sup>	?	?	yes	yes	yes
Medline+/MedMaster	11/18/2007	?	?	?	no		
WHO-ATC	2005	~2,800	116	yes	yes	yes	n/a
WHO-DRUG	9/7/2007	9,899	13,446	yes	yes	?	n/a
Int'l Pharmacopoeia	4 <sup>th</sup> ed. (2006)	420	180	yes	no		
INN's	1953-2007	>8,000 <sup>c</sup>	0	yes	no		
EphMRA	2006	0	0	yes	no		
USP/USAN	Oct-2007	3,968	5,000?	?	no		
PubChem	Nov-2007	?	0	yes	yes	no	yes
ChemiIDplus	11/15/2007	95,640 <sup>d</sup>	0	yes	yes	no	yes
DrugBank	Feb-2006	~4,300	0	yes	yes	some	yes
NDFRT 2007 data	10/3/2007	0	0	yes	no		
KEGG	Nov-2007	Tbd	0	yes	tbd	tbd	tbd

<sup>a</sup> package inserts (mix of generic & trade names)

<sup>b</sup> "drugs and herbals"

<sup>c</sup> "names" in Latin, English, French, and Spanish

<sup>d</sup> "Drug / Therapeutic Agent"

**Table 4. Resources for systematic dimension analysis.**

FDA [U.S.] Food and Drug Administration; WHO World Health Organization; ATC Anatomic-Therapeutic-Chemical [classification]; Phar.Int. International Pharmacopoeia; INN International Nonproprietary Names; USP U.S. Pharmacopeia; USAN U.S. Adopted Names; AMA American Medical Association; MeSH Medical Subject Headings; MH Main Headings; UMLS Unified Medical Language System; ChEBI Chemical Entities of Biological Interest; KEGG Kyoto Encyclopedia of Genes and Genomes; C consumers; CP clinical/pharmacy workers; R researchers, # INs number of drugs equivalent to a RXNORM "ingredient" (IN; single-compound approved generic name).

<i>source name</i>	<i>Website</i>	<i>users</i>	<i># INs</i>
MedMaster	<a href="http://www.nlm.nih.gov/medlineplus/druginformation.html">http://www.nlm.nih.gov/medlineplus/druginformation.html</a>	C	?
DrugDigest	<a href="http://www.drugdigest.org/DD/Home">http://www.drugdigest.org/DD/Home</a>	C	~1,000
DailyMed	<a href="http://dailymed.nlm.nih.gov">http://dailymed.nlm.nih.gov</a>	C,CP	1,117
ClinicalTrials.gov	<a href="http://clinicaltrials.gov/">http://clinicaltrials.gov/</a>	C,CP	924
DrugInfo	<a href="http://druginfo.nlm.nih.gov/">http://druginfo.nlm.nih.gov/</a>	C,CP,R	">12,000"
RXNORM	<a href="http://mor.nlm.nih.gov/download/rxnav/">http://mor.nlm.nih.gov/download/rxnav/</a>	CP	5,592
Drugs@FDA	<a href="http://www.accessdata.fda.gov/scripts/cder/drugsatfda/">http://www.accessdata.fda.gov/scripts/cder/drugsatfda/</a>	CP	1,689
WHO-ATC	<a href="http://www.whocc.no/atcddd/">http://www.whocc.no/atcddd/</a>	CP	~3,000
WHO-DRUG	<a href="http://www.ume-products.com/DynPage.aspx?id=2829&amp;mn=1107">http://www.ume-products.com/DynPage.aspx?id=2829&amp;mn=1107</a>	CP	9,899
Phar.Int.	<a href="http://www.who.int/medicines/publications/pharmacopoeia/en/index.html">http://www.who.int/medicines/publications/pharmacopoeia/en/index.html</a>	CP	420
INN	<a href="http://www.who.int/medicines/services/inn/en/index.html">http://www.who.int/medicines/services/inn/en/index.html</a>	CP	~2,000
USP Dictionary	<a href="http://www.uspusan.com/usan/login">http://www.uspusan.com/usan/login</a>	CP	>4,317
USAN via AMA	<a href="http://www.ama-assn.org/ama/pub/category/2956.html">http://www.ama-assn.org/ama/pub/category/2956.html</a>	CP	689
MeSH MH	<a href="http://www.nlm.nih.gov/mesh/meshhome.html">http://www.nlm.nih.gov/mesh/meshhome.html</a>	CP,R	~2,000
MeSH all	<a href="http://www.nlm.nih.gov/mesh/meshhome.html">http://www.nlm.nih.gov/mesh/meshhome.html</a>	CP,R	~5,000
UMLS	<a href="http://www.nlm.nih.gov/research/umls/">http://www.nlm.nih.gov/research/umls/</a>	CP,R	~9,000
PubChem	<a href="http://pubchem.ncbi.nlm.nih.gov/">http://pubchem.ncbi.nlm.nih.gov/</a>	R	?
ChemiDplus	<a href="http://chem.sis.nlm.nih.gov/chemidplus/">http://chem.sis.nlm.nih.gov/chemidplus/</a>	R	?
ChEBI	<a href="http://www.ebi.ac.uk/chebi/">http://www.ebi.ac.uk/chebi/</a>	R	>7,000
DrugBank	<a href="http://redpoll.pharmacy.ualberta.ca/drugbank/">http://redpoll.pharmacy.ualberta.ca/drugbank/</a>	R	1,835
KEGG DRUG	<a href="http://www.genome.jp/kegg/drug/">http://www.genome.jp/kegg/drug/</a>	R	6,848
Reactome	<a href="http://www.reactome.org/">http://www.reactome.org/</a>	R	<100?
HumanCyc	<a href="http://humancyc.org/">http://humancyc.org/</a>	R	20

**3.1.4 Experimental database.** We built a database containing information on nine fundamental drug compounds extracted from 15 resources: MedMaster, DrugDigest, DailyMed, DrugInfo, RXNORM, ClinicalTrials.gov, Drugs@FDA, WHO-ATC, MeSH, UMLS, PubChem, ChemiIDplus, ChEBI, DrugBank, and KEGG DRUG.<sup>20</sup> The nine drugs are:

1. doxazosin / doxazosin mesylate
2. dutasteride
3. finasteride
4. leuprolide / leuprolide acetate
5. prazosin / prazosin hydrochloride
6. saw palmetto
7. tamsulosin / tamsulosin hydrochloride
8. terazosin / terazosin hydrochloride
9. ticlopidine / ticlopidine hydrochloride

The various salts, derivatives, formulations, trade names, combination products, etc., of these nine drugs (in the single-component generic name parent compound sense) populate the corresponding dimensions, so the database covers many more drugs in that larger sense. For comparison, the universe of U.S.-approved single-component generic name parent compound drugs may number about 5,000.<sup>21</sup> Also shown in this list are the salts (mesylate, acetate, hydrochloride) distinguished by separate records in the resources we used, bringing the number of drugs in that sense up to 15.

---

<sup>20</sup> Referenced in Table 4.

<sup>21</sup> This number is based on the coverages of RXNORM and MeSH. There is no exact consensus on what constitutes the universe of drugs (Table 11). In earlier work on this thesis topic, we attempted to make construction of a 5000-drug database feasible by narrowing the scope to sources with pre-normalized data applicable to a narrow set of use cases related to drug-indication relations (Appendix B). One problem with this was that the considered dimensions had little or no generality across the considered sources. Another was that the only available normalized *indication* data appeared to be a poor reflection of canonical approved indications (Table 7). The same was found to a lesser degree for *contraindications* (Table 8).

These drugs were selected as follows: Finasteride was chosen for its interesting split into two *trade names* ("Propecia" and "Proscar") corresponding to two unit doses (1 mg and 5 mg) for two different indications (male pattern hair loss and benign prostatic hyperplasia). Ticlopidine hydrochloride was chosen at random to compare some data in UMLS to equivalent data in DailyMed. The other drugs were selected to support demonstration of utility for the general application "Find multiple drugs for the same indication"; specifically, one of finasteride's indications, benign prostatic hyperplasia (BPH). Their names were obtained from the UMLS/NDFRT *Other related/may\_treat* relations for "Prostatic Hypertrophy" which is NDFRT's preferred term for BPH.

The database was built in a Microsoft Excel spreadsheet. The raw data was loaded by manually copying and pasting individual character strings (words, phrases, sentences, paragraphs, lists, etc.) from the source's display (usually a web page) into designated columns of a structured Excel spreadsheet (see next paragraph). Normalized translations of the raw data were then added to separately designated columns. The normalization process is described in Appendix E. The potential for automation of this laborious process is addressed below under Q3.C.2. Excel's built-in search, sort, string matching, and other functions were used to facilitate and enhance normalization, to extract an expanded set of normalized dimensions (Q1) for cluster analysis (Q2), and to simulate a human-computer IR interface for purposes of testing the database's effectiveness in use cases as a function of its dimension-based integration (Q3).

The Excel columns are designated as follows. Example cell values are given in quotes.

- A. Original sort order - line numbers to recover original context if needed.
- B. Normalized generic name - e.g., "finasteride"
- C. Normalized source - e.g., "DailyMed"
- D. Raw drug (record) name - e.g., "Propecia (Finasteride) Tablet, Film Coated  
[Merck & Co., Inc.]"

- E. Raw drug (record) URI - e.g.,  
["http://dailymed.nlm.nih.gov/dailymed/drugInfo.cfm?id=6926"](http://dailymed.nlm.nih.gov/dailymed/drugInfo.cfm?id=6926)
- F. Raw dimension name - e.g., "Brand Names"
- G. Raw dimension URI if available.
- H. Raw value - e.g., "PROPECIA"
- I. Raw value URI if available.
- J. Dimension-dimension clue. This and the other "clue" fields were added to clarify, systematize, and document complex normalizations. An example of a non-complex dimension normalization is from the raw "Brand Names" to *trade name*.<sup>22</sup> An example of a complex dimension normalization is from the raw "What side effects may I notice from this medicine?" to *side effect*. In the second case, it may be helpful to know that the clue is the raw substring "side effects". Other examples will be given in the Results section.
- K. Dimension-value clue. Sometimes raw dimension-value pairs were normalized in ways where raw dimension clues ended up in the normalized value and/or raw value clues ended up in the normalized dimension. For example, the raw dimension "Precautions - Nursing Mothers" and value "Finasteride is not indicated for use in women. It is not known whether finasteride is excreted in human milk." were normalized to *precaution - contraindication* and "breast feeding". Thus the dimension-dimension clue is "Precautions"; the dimension-value clue is "Nursing Mothers"; and the value-dimension clue is "not indicated".
- L. Value-dimension clue. See K.

---

<sup>22</sup> In the database, the normalized indication terms include their domain classification; e.g., *pharmacy - trade name*, *clinical - indication*, *biology - biological effect*. For better readability, these will be omitted from this narrative unless specifically relevant.

- M. Flag for whether the value-value clue (N) has been parsed in any way. That is, does it differ<sup>23</sup> from the raw value (H)? This field was added for consistency checking purposes.
- N. Value-value clue. Except for UMLS, most of the raw values are free text. Thus the need for normalization substrings is even greater than for the dimensions (J).
- O. Normalized dimension. See J and K.
- P. Flag for whether the value-value clue (N) has been normalized in any way. That is, does it differ<sup>24</sup> from the normalized value (Q)? This field was added for consistency checking purposes.
- Q. Normalized value. See J and K.
- R. Linked to dimension(s). See S.
- S. Linked to value - Sometimes there are interlinkages of data between a given source's dimension-value pairs about a given generic drug. For example, in some sources, within *generic name*="finasteride", *dose*="5 mg" is specific to *indication*="benign prostatic hyperplasia" and *trade name*="Proscar", while *dose*="1 mg" is specific to *indication*="male pattern hair loss" and *trade name*="Propecia".

Table 5 exemplifies the incorporation of data from a source such as UMLS/NDFRT which employs knowledge representation (KR) based on local ontology structure and controlled terminology. Such KR usually enables non-complex normalizations, hence most of the clue and flag fields (J-N, P) are blank in these examples. The UMLS/NDFRT relationship *Other related/may\_be\_treated\_by* is normalized to the dimension *indication - treatment*<sup>25</sup> (meaning

---

<sup>23</sup> Case-insensitive

<sup>24</sup> *ibid.*

<sup>25</sup> See footnote 7 on page 7. One difference between *dimensions* and *relationships* is that we are not dealing with directionality because the drug is always the subject. The NDFRT relationship *may\_be\_treated\_by* reported here is somewhat of an artifact of our UMLS database. Our dimension *indication - treatment* actually corresponds better to the NDFRT reverse relationship *may\_treat*.

*treatment* is a subtype of *indication*). In contrast to the example in Table 6 (next paragraph), UMLS/NDFRT does not distinguish between approved and other indications. The raw values - "Prostatic Neoplasms"; "Hirsutism"; "Alopecia"; and "Prostatic Hyperplasia" - are simply converted to dictionary case and singular form. The data was extracted from downloaded UMLS files so there are no URIs (E, G, I). UMLS/NDFRT does not link these indications to specific drug doses, trade names, etc., hence the linkage fields (R, S) are blank.

Table 6 exemplifies the incorporation of data from a source such as MedMaster which employs free text KR. Free text KR usually requires complex normalizations, hence the clue field values, highlighted to show their context in the natural language text. MedMaster indications can be assumed to be approved, hence the additional subtype qualifier (*indication - treatment - approved*). Note the discrepancies with the UMLS/NDFRT indication values for finasteride (Table 5). Two of these involve granularity ("male pattern hair loss" is a subtype of "alopecia" and "benign prostatic hyperplasia" is a subtype of "prostatic hyperplasia"). The other two UMLS/NDFRT values might be considered non-approved indications but there is no independent (i.e., other than comparing them to known approved indications) way to distinguish or grade these or infer the original (pre-normalized) term, which probably also is of finer granularity (e.g., "Prostatic Neoplasms" probably refers to "prostate cancer" which finasteride has been postulated to help prevent based on its mechanism of action). Table 6 also illustrates the URI fields (E, G) and the linkage of one dimension/value to another dimension/value within a given source/drug (R, S).

**Table 5. Experimental database schema with structured data examples.**

For display clarity, the columns and rows (1449, 1451, ...) are inverted relative to the actual database (<http://comminfo.rutgers.edu/~msharp/XKB/DB6.xls>).

<i>Column</i>	<i>column name</i>	<i>example 1</i>	<i>example 2</i>	<i>example 3</i>	<i>example 4</i>
A	original sort order	1449	1451	1454	1455
B	normalized generic name	finasteride	finasteride	finasteride	finasteride
C	normalized source	UMLS/NDFRT	UMLS/NDFRT	UMLS/NDFRT	UMLS/NDFRT
D	raw drug (record) name	Finasteride	Finasteride	Finasteride	Finasteride
E	raw drug (record) URI				
F	raw dimension name	Other related/ may_be_ treated_by	Other related/ may_be_ treated_by	Other related/ may_be_ treated_by	Other related/ may_be_ treated_by
G	raw dimension URI				
H	raw value	Prostatic Neoplasms	Hirsutism	Alopecia	Prostatic Hyperplasia
I	raw value URI				
J	dimension- dimension clue	treated	treated	treated	treated
K	dimension-value clue				
L	value-dimension clue				
M	value parse flag				
N	value-value clue				
O	normalized dimension	clinical - indication - treatment	clinical - indication - treatment	clinical - indication - treatment	clinical - indication - treatment
P	value normalization flag	N			
Q	normalized value	prostatic neoplasm	hirsutism	alopecia	prostatic hyperplasia
R	linked to dimension(s)				
S	linked to value				

**Table 6. Experimental database schema with free text examples.**

For display compaction and clarity, repeated identical values are merged for columns B-I, columns (A-S) and rows (3, 4, ...) here are inverted relative to the actual database (<http://comminfo.rutgers.edu/~msharp/XKB/DB6.xls>).

<i>column</i>	<i>column name</i>	<i>example 1</i>	<i>example 2</i>	<i>example 3</i>	<i>example 4</i>
A	original sort order	3	4	5	6
B	normalized generic name	finasteride			
C	normalized source	MedMaster			
D	raw drug (record) name	Finasteride			
E	raw drug (record) URI	<a href="http://www.nlm.nih.gov/medlineplus/druginfo/meds/a698016.html">http://www.nlm.nih.gov/medlineplus/druginfo/meds/a698016.html</a>			
F	raw dimension name	Why is this medication prescribed?			
G	raw dimension URI	<a href="http://www.nlm.nih.gov/medlineplus/druginfo/meds/a698016.html#why">http://www.nlm.nih.gov/medlineplus/druginfo/meds/a698016.html#why</a>			
H	raw value	<p>Finasteride (Proscar) is used alone or in combination with another medication (doxazosin [Cardura]) to treat benign prostatic hypertrophy (BPH, enlargement of the prostate gland). Finasteride improves symptoms of BPH such as frequent and difficult urination and may reduce the chance of acute urinary retention (suddenly being unable to pass urine). It also may decrease the chance of needing prostate surgery. Finasteride (Propecia) is also used to treat male pattern hair loss (a common condition in which men have gradual thinning of the hair on the scalp, leading to a receding hairline or balding on the top of the head.) Finasteride (Propecia) has not been shown to treat thinning hair at the temples and is not used to treat hair loss in women or children. Finasteride is in a class of medications called 5-alpha reductase inhibitors. Finasteride treats BPH by blocking the body's production of a male hormone that causes the prostate to enlarge. Finasteride treats male pattern hair loss by blocking the body's production of a male hormone in the scalp that stops hair growth.</p>			
I	raw value URI				
J	dimension-dimension clue	Why is this medication prescribed?	Why is this medication prescribed?		
K	dimension-value clue				
L	value-dimension clue	Is used to treat	is used to treat	is in a class of medications	treats by
M	value parse flag	P	P	P	P
N	value-value clue	benign prostatic hyperplasia	male pattern hair loss	5-alpha reductase inhibitors	blocking the body's production of a male hormone
O	normalized dimension	clinical - indication - treatment - approved	clinical - indication - treatment - approved	biology - therapeutic class	biology - biological effect
P	value normalization flag		N	N	N
Q	normalized value	benign prostatic hyperplasia	male pattern alopecia	5-alpha reductase inhibitor	androgen synthesis decrease
R	linked to dimension(s)	trade name	trade name		
S	linked to value	Proscar	Propecia		

### 3.2 Q2: Do Dimensions Lead to Valid Groupings of Resources?

"Information resources may be categorized in various ways, so that an extensive set of diverse resources, as certainly exists for the pharmaceutical domain, may be better understood and organized" (Bawden & Robinson, 2010, p. 79). Of the ways listed by these authors, our approach corresponds most closely to by "subject"; "type of material"; and "intended audience" (p. 80).

**3.2.1 Face validity.** Using the initial survey domain-dimension-resource matrix (Table 9 in Results), domain scores were computed for each resource by summing the matrix scores ( $\bullet=1$ ;  $\pm=0.5$ ). Each resource thus has a score for each of the four domains consisting of the number of that domain's dimensions covered by the source, divided by the total number of dimensions covered by the source, expressed as a percentage. The sources can then be grouped by their domain scores according to any desired criteria (highest, lowest, >50%, most equitable, etc.), and the validity of these groupings evaluated. For example, do all the sources with "Chem" in their names group together under *chemistry*?

**3.2.2 Correspondence analysis.** Correspondence analysis (Greenacre, 1984) provides a method for representing both the row and column categories of the domain-dimension-resource matrix in the same space, so that the results can be visually examined for structure. To reduce dimensionality, only the first two axes of the new space are plotted. The overall quality of representation of the points is expressed as a proportion of the total variation ("inertia"). Weighting schemes reflect the sources' differing numerical coverage of generic names (Table 4) and chemical entities (not shown) and give credit for partial coverage. We used the statistical software package MVSP for this analysis.

**3.2.3 Cluster analysis.** We also submitted the resource-by-dimension matrix without the domain groupings to hierarchical cluster analysis using the statistical software package SPSS (Norusis, 2005). Such an analysis was expected to yield a more objective picture of how the resources and dimensions cluster for comparison to our manual domain-dimension groupings;

that is, to evaluate our four-domain classification hypothesis. Cluster analysis was also applied to the expanded set of dimensions from the experimental database.

### 3.3 Q3: Can Dimensions Facilitate Integration/OM Tasks?

**3.3.1 Classifying sources.** This is basically a usefulness version of Q2; i.e., are the classifications implied by the grouping, correspondence, and clustering results for Q2 useful as well as valid?

**3.3.2 Selecting sources appropriate to a given information need.** Given a dimensions-by-resources matrix (Table 9) and a mapping of the same dimensions to usage scenarios,<sup>26</sup> the matrix can be used to select the resources most likely to satisfy the user's information need because they cover the relevant dimensions. For example, a user (consumer, clinician, pharmacist, or biomedical researcher) wants to find the drugs corresponding to a given indication or vice versa. This scenario minimally requires coverage of the *generic names* and *indications* dimensions. In addition, dimensions such as *therapeutic class*, *mechanism of action*, *biological effect*, *molecular target*, *experimental applications*, and *chemical superclass*, may also be useful in determining alternative or possible indications. Each resource's hypothetical effectiveness score for this scenario is obtained by summing its matrix scores for these dimensions (Table 9; ●=1; ±=0.5).

**3.3.3 Pooling data from different sources.** The experimental database represents an application of dimensions to pooling data from different sources. Its usefulness was evaluated from three perspectives: data reduction, automatic normalization of additional raw data, and satisfying use cases.

**3.3.3.1 Data reduction.** The normalization process (Appendix E) was designed to conflate different strings representing the same drugs, dimensions, or values in the raw data as illustrated in Table 5 and Table 6. Thus we expected that the number of unique raw

---

<sup>26</sup> A scenario is a description of interactions between types of users and the system.  
[http://en.wikipedia.org/wiki/Scenario\\_\(computing\)](http://en.wikipedia.org/wiki/Scenario_(computing))

representations would always be greater than the number of unique corresponding normalized representations within any cross-section of the database, and that the size of this difference would be a good measure of the effectiveness of our method at integrating the data. Ratios of unique<sup>27</sup> normalized to unnormalized representations in the experimental database, expressed as percentages, were computed for drugs (columns D versus B), dimensions (F versus O), values (H versus Q), and drug-dimension-value "triples" across the entire database and within each source.<sup>28</sup> In addition, the effect of dimensional hierarchical aggregation (Appendix G) was computed for selected dimensions such as *pharmacy - generic name* where the dimension and its sub-dimensions all take values of the same semantic type. Finally, case-specific examples of data reduction were computed from the use case results.

**3.3.3.2 Automatic normalization of additional raw data.** The experimental database constitutes a potential training dataset for automatic addition and integration of more data. Three scenarios were considered: (i) addition of more data about the same drugs from the same resources; (ii) addition of more data about the same drugs from different resources; and (iii) addition of more data about different drugs from the same resources. (i) was ruled out because we essentially loaded all the data available about our chosen sample of drugs in our chosen sample of resources. (ii) was ruled out for three reasons. The first is that our chosen sample of resources comes close to exhausting the universe of relevant resources which are freely available for study. Secondly, even if more were available, the time required for preliminary analysis makes this goal impractical. Thirdly, the model adopted for addition of data about more drugs (below) depends on our specific mapping of raw to normalized dimensions, which does not generalize well across the sources we examined (that is, they tend to have diverse representations

---

<sup>27</sup> The number of unique values in an Excel column is given by copying the column to column A of a scratch worksheet, sorting it, entering "0" into B1 and the formula " $\text{=if}(A2=A1,1,0)$ " in cell B2, copying B2 to all B cells to the end of the A data, copying and pasting "special" the values in B, sorting all on B, and recording the last row number with a "0" in B.

<sup>28</sup> The unique normalized dimensions were taken from the fourth hierarchical level for comparability across sources. These calculations did not exclude data representing unparsed links and summaries, database IDs and cross-references, compound dimensions, or *information for the patient* subtypes.

of the same dimension) and so would not be expected to generalize well to other resources.

To integrate additional data about more drugs from the same resources, one needs to

- a. Parse the raw data into source-drug-dimension-value "quadruples." Each quadruple defines a new record (row) in the database and the raw strings go into columns C, D, F, and H. For each new row,
- b. Normalize the raw drug name and load the normalized drug name into column B.
- c. Normalize the dimension name and load the normalized dimension name into column O.
- d. Normalize the value name and load the normalized value name into column P.

An example of a raw quadruple is: ChEBI-"tamsulosin hydrochloride"-"Brand Names - Source"-"Flomax - KEGG DRUG". The corresponding normalized quadruple is: ChEBI-"tamsulosin hydrochloride"-"*pharmacy - trade name*"-"Flomax".

Substantively addressing automation of steps a, b, and d is out of scope for this dissertation. Step a is basically a data access issue, while steps b and d are basically autoencoding (mechanized mapping of uncontrolled to controlled terminology) issues.

For step c, we derived a table of probabilities based on the experimental database. Based on the data we manually processed, any hypothetical new row of data from one of our resources has a "prior" probability of mapping to one of our normalized dimensions which is equal to the fraction of all rows with the same raw source-dimension (columns C and F) pair mapped to that normalized dimension (column O). For example, all the rows with the raw (C,F) pair ChEBI-"Brand Names - Source" have the normalized dimension "*pharmacy - trade name*", hence the probability is 1.0 (100%). All the data we processed is consistent with that mapping, so we expect all additional ChEBI data to follow the same pattern. In contrast, for the raw (C,F) pair ChemIDplus-"Names and Synonyms - Synonyms", only 100/262 (0.38) of the data rows are mapped to "*pharmacy - trade name*"; the probability is 38% that a new row from that section of a ChemIDplus page would be correctly normalized this way. These probabilities can also be viewed as a kind of precision score for the query "Find all cases (drug-value pairs) of the

normalized dimension when the latter is specified only as the raw source-dimension pair."

ChEBI's precision for retrieving *trade names* via "Brand Names - Source" is 100%, while ChemIDplus' via "Names and Synonyms - Synonyms" is 38% (the false 62% being *generic synonyms, abbreviations, chemical names, etc.*).<sup>29</sup>

For cases with probability less than 100%, additional probabilities were computed based on dimensional hierarchical aggregation and/or identification of clues within the value. For example, the raw (C,F) pair ChemIDplus-"Names and Synonyms - Synonyms" also maps to a variety of *pharmacy - generic name* subtypes (see Appendix G), none with a probability greater than 6%. However, if they are all aggregated "up" to "*pharmacy - generic name*" the combined probability is 31%. If the clue "[INN-Latin]" in the value "Ticlopidinum [INN-Latin]" is added to the (C,F) pair (making it a C,F,H-clue triple), its probability of mapping correctly to "*pharmacy - generic name - INN/Latin*" jumps from 5% to 100%. Of course, using the whole value (i.e., the C,F,H triple) guarantees 100% precision across the entire database, and there is no formal boundary between our clues and their whole values in our ad hoc prototype. Formalization and automation of value clue derivation go along with formalization and automation of value normalization, which we leave to future extensions. We merely wish to note its potential relevance to dimension normalization.

### 3.3.3.3 *Satisfying use cases.*

#### 3.3.3.3.1 *Criteria for usefulness.*

The use cases were adapted from "real world" information needs represented in published literature as described below. Therefore the criteria for usefulness are comparative: our

---

<sup>29</sup> The same raw dimension-value pair may be mapped to multiple normalized dimension-value pairs. This happens frequently with nonspecific raw dimensions such as "Description" or "Scope Note" and values in free text format. Therefore these probabilities should be understood as the probability of picking a given normalized dimension at random from among all the normalized dimensions associated with a given raw dimension in the database, *not* the probability that a given raw dimension's family of normalized dimensions *includes* a particular normalized dimension. The latter may be of interest as well as a kind of semantic recall measure. Later (in Results) we make the distinction between the *semantic* precision measured by our dimension probabilities, versus *pharmaceutical* precision, which entails the values.

experimental database should perform as well or better than the system in the literature or, if that cannot be assessed, better than the individual, unintegrated resources in the paper's system or ours.

1. Comprehensive coverage. A mapping of these papers' classes and resources of drug information (Section 3.1.2) to our equivalent dimensions and resources should show that our database covers the aggregate set of dimensions better (higher %) than the paper's collection if comparable, otherwise by any one resource alone, and integrates additional information (dimensions, values, and resources) that the paper's system or single resources do not cover.

2. Literary warrant fidelity. Each specific use case should represent, with minimal value substitution, an example of a need or test query expressed in these papers.

3. IR performance. Retrieval based on searching, clustering, etc., the normalized generic name, dimension, and value fields (B,O,Q) of the database should be larger, more robust, and more efficient than what could be achieved using the paper's system, if comparable, otherwise the raw data fields (D,F,H) or the original disparate, scattered information resources. Operationally, "larger" means more records. "More robust" means representing more information and its consistency (and inconsistency, which can also be informative) across sources; that is, if the answer is based on information from multiple sources, it is more likely to pass a truth test based on one or a subset of them. Robustness also refers to linkages to additional information (e.g., *indication:molecular target*). "More efficient" refers to the effort and complexity of the search process (number of databases, commands, queries, time, etc.) operationalized as data reduction (reduction in the number of unique strings representing the same concept).

Note that we are not using precision/recall or other measures of truth/accuracy. The reason is that such measures do not reflect on the dimensions *per se* so much as the accuracy of the sources (which is beyond our control and not our goal to evaluate) and our specific mappings. For example, UMLS/NDFRT's "incorrect" indications for ticlopidine (Table 7) depend on our mapping of the local *may\_treat* relationship to our *indication* dimension. This does not invalidate

the *indication* dimension, just (perhaps) our mapping of *may\_treat*. Following Plovnick and Zeng's "gold standard" evaluation model would entail evaluating additional resources or enlisting independent human subject experts. The marginal relevance of truth/accuracy measures to our main thesis was judged not worthy of the required effort and expense.

*3.3.3.3.2 Health care and related personnel.* This set of use cases is based on Liu et al. (2005), Zeng et al. (2007), and Kupferberg and Jones Hartel (2004). Liu et al. (2005) defined RXNORM's intended users as "health care personnel including prescribing physicians and nurses, and hospital personnel involved with drug ordering, inventory management, recording dose adjustments, checking drug interactions, or pharmacy management." Zeng et al. (2007) identified the general problem that RXNORM does not cover all the dimensions of drug information of interest to these users "such as pharmacologic action, drug-drug interactions, indications, contraindications, and adverse reactions." In our terms, these two papers suggest that health care, hospital, and pharmacy personnel wish to search, cluster, and/or distinguish drugs based on their values for these dimensions at the same level of detail and comprehensiveness as RXNORM's coverage of *generic name*, *trade name*, *dose*, and *dosage form*. The core of the unmet information need is of two types: where *X* is a dimension not covered by RXNORM,

Health Use Case A. For a given value of *generic name* find alternate values of *X*,

Health Use Case B. For a given value of *X* find alternate values of *generic name*.

To develop this set of use cases, we simply need to plug into *X* our normalized dimensions corresponding to Zeng et al.'s wish list, as given by our database columns O and F respectively.

- pharmacologic action: *clinical - therapeutic class ...*
- drug-drug interactions: *clinical - precaution - drug interaction ...*
- indications: *clinical - indication ...*
- contraindications: *clinical - precaution - contraindication*
- adverse reactions: *clinical - precaution - side effect ...*

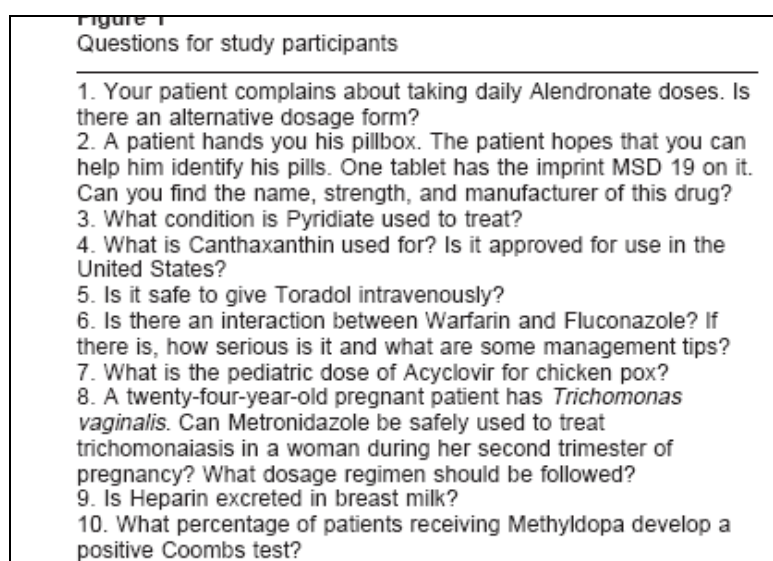
"..." signifies that all normalized dimensions should be understood to include their hierarchical sub-dimensions (Appendix G); retrieving these will allow finer granularity searches such as for approved indications (*clinical - indication - ... - approved*). An example of query type (1) is "Find all indications for finasteride" which translates to **generic name** = "finasteride" and **X** = "*clinical - indication ...*". Executing this query to our database (Appendix F) yields the normalized values "benign prostatic hyperplasia" and "male pattern alopecia". Keeping **X** = "*clinical - indication ...*" and plugging "benign prostatic hyperplasia" into query type (2) as "value of **X**" yields "finasteride"; "dutasteride"; and 13 of the other 15 unique normalized generic names in the database as "values of **generic name**" (since 13/15 of our drug sample was selected on that basis). The results could then be refined by *indication* sub-dimension and/or used to formulate follow-up queries to compare the different drugs' pharmacologic actions, drug-drug interactions, contraindications, or adverse reactions, according to the above mapping, other dimension-specific values, or combinations thereof.

Kupferberg and Jones Hartel (2004) enlisted "pharmacy students, faculty, and librarians" to help develop a list of 10 test queries to compare evaluations of five full-text drug databases (Figure 8). Our model database does not support these 10 specific test queries due to its small generic drug sample. However, we can simulate most of them using other specific values. Numbers 3, 4, 5, and 9 require only substituting one of our nine normalized generic parent names (finasteride, dutasteride, doxazosin, ...) for the drug name. Others can be adjusted as follows, where ~~strikeout~~ signifies deleted original text and **bold** our substitution or addition.

1. Your patient complains about taking daily ~~Alendronate~~ **leuprolide** doses. Is there an alternative dosage form **where frequency of administration < 1/day?**
7. What is the ~~pediatric dose~~ **dosing regimen** of ~~Acyelovir~~ **{finasteride, dutasteride, doxazosin, saw palmetto, ...}** for ~~chicken-pox~~ **{BPH, hypertension}?**

8. A 24-year-old pregnant woman has ~~Trichomonas vaginalis~~ **alopecia**. Can ~~Metronidazole~~ **{finasteride, dutasteride, doxazosin, saw palmetto, ...}** be safely used?
10. What percentage of patients receiving ~~Methyldopa~~ **doxazosin mesylate** develop a ~~positive Coombs test~~ **hypotension**?

These eight queries constitute our Health Use Cases C, D, E, F, G, H, I, and J.



**Figure 8. Test queries to evaluate drug databases (Kupferberg & Jones Hartel, 2004).**

---

*3.3.3.3 Pharmaceutical discovery researchers.* This set of use cases is based on Castle et al. (2007), Vogel (2007), Campillos et al. (2008), and Boguski et al. (2009). Castle et al.'s examples involve finding clusters of chemically related drug compounds and their molecular targets. The corresponding biological correlates (of the drugs and the targets) are then mined for co-occurrences that suggest plausible, novel, interesting, testable hypotheses. One example concerns phenotype:disease relations. These are what Vogel (2007) is looking for, where the disease is cancer and the phenotypes are faster endpoints than overt cancer and death, to speed and reduce the cost of anti-cancer drug research. Campillos et al. (2008) basically did the same

thing as Castle et al. but added a *side effect* dimension to the initial clustering. This paper was specifically motivated by prediction of novel hypothetical indications and therapeutic classes for existing drugs, which Boguski et al. (2009) called "repurposing."

Where our model database can improve, in principle, on these prototypes is in the richness of the set of drug biological correlates. These prototypes' requirement for already-normalized data limited them in this regard to WHO-ATC. In contrast, both prototypes used DrugBank's rich set of molecular target biological correlates. Our database has much more on the drug side. In addition to WHO-ATC, it has alternative *therapeutic class* values from many of our 14 other resources, plus additional drug *indication*, *biological effect*, and *mechanism of action* correlates. The volume of this data is equivalent to over 1.4 million rows of data if extrapolated from our small sample to our estimate of the U.S.-approved generic drug universe.<sup>30</sup>

The following use cases can be inferred from these papers. (See Appendix F for content adaptations and query formulations.)

Research Use Case A. A cluster of structurally similar compounds targeting the TACR1 gene product (known to be associated with abnormal pain threshold ) was found that points to the WHO-ATC class "antiemetics and antinauseants", suggesting that TACR1 modulation may produce antinauseant activity, and/or that there is a possible connection between antinauseant activity and abnormal pain threshold (Castle et al., 2007).

Research Use Case B. The WHO-ATC class "cardiovascular system" points to a list of cardiovascular drugs whose gene targets map to a smaller list of phenotypes. The highest ranking phenotype is "decreased heart rate" which is consistent with the WHO-ATC class. This suggests that other WHO-ATC→drug→gene target→phenotype mappings might be mined for phenotype:disease hypotheses (Castle et al., 2007).

---

<sup>30</sup> 2548 rows of data on these dimensions and their sub-dimensions, times 5000/9 = 1.42 million. The comparable figure for WHO-ATC (all four levels) is 88 x 5000/9 = 0.05 million. For all *molecular target* sub-dimensions it is 1239 x 5000/9 = 0.7 million. For the specific biological correlate sub-dimensions *general function*, *specific function*, *pathway*, and *GO biological process*, it is 180 \* 5000/9 = 0.1 million.

Research Use Case C. Campillos et al. (2008) extracted specific sets of drugs with common side effects but different WHO-ATC therapeutic classes, and used the drugs' molecular target and chemical structure/similarity values to predict previously unknown shared targets, which were tested by *in vitro* and cell assays. The validated shared targets predict novel hypothetical indications and therapeutic classes for existing drugs. For example, a set of nervous system drugs was found to have side effects in common the antiulcer drug rabeprazole. Four of their targets were predicted to bind rabeprazole, and two - the dopamine receptor DRD3 and the serotonin receptor HTR1D - were validated. This suggests that rabeprazole may be therapeutic for the indications of zolmitriptan (migraine), pergolide (Parkinson's disease), and paroxetine and fluoxetine (psychiatric disorders<sup>31</sup>).

Research Use Case D. Boguski et al. (2009) also address finding novel hypothetical indications and therapeutic classes for existing drugs ("repurposing") but do not present a prototype on which to base a use case, so we made this one up. A researcher wonders if any existing drugs might be "repurposed" to prevent prostate cancer. She searches ClinicalTrials.gov and gets a list of clinical trials which link the *Condition* "Prostate Cancer" to various *Interventions* including drug names.<sup>32</sup> She thinks this is a good start, but what she really needs is to find other, chemically related drugs and chemicals which are *not* on this list or already approved for prevention of prostate cancer.

3.3.3.3.4 *Consumers.* Actual examples of consumer/patient drug information queries are surprisingly hard to find, given the proliferation of consumer-oriented resources such as MedMaster and DrugDigest, U.S. government patient health information empowerment efforts,<sup>33</sup> and scholarly literature on health information seeking behavior and medical informatics. One

---

<sup>31</sup> fluoxetine: depression, obsessive-compulsive disorder, some eating disorders, panic attacks, premenstrual dysphoric disorder; paroxetine: depression, panic disorder, social anxiety disorder, obsessive-compulsive disorder, generalized anxiety disorder, posttraumatic stress disorder, premenstrual dysphoric disorder. Source: MedMaster.

<sup>32</sup> <http://clinicaltrials.gov/ct2/results?term=prostate+cancer>

<sup>33</sup> E.g., <http://www.hhs.gov/healthit/healthnetwork/background/>

reason for this may be that consumers' drug information queries are generally ill-formed and ineffective (Plovnick & Zeng, 2004; Scott-Wright et al., 2006; Zeng, Crowell, et al., 2006; Keselman et al., 2008).

Plovnick and Zeng (2004) collected consumer queries and search goals from patients and visitors recruited from public areas of a large hospital. Subjects described their health-information needs to an interviewer and were then given the opportunity to search the internet on a laptop. The subjects' queries were recorded for further analysis. We adapted these queries to our database's model content much like we did the Kupferberg and Jones Hartel (2004) queries.

1. Are there any natural **[herbal]** substitutes for the ~~hormone replacement~~ **BPH** therapy agent ~~Prempo~~ **{Proscar, Flomax, Avodart, Ticlid, Viadur ...}** ?
6. How are ~~arrhythmias~~ **BPH** treated?
9. Is there treatment for ~~restless-legs-syndrome~~ **baldness** ?
10. What are scientifically validated **[approved]** treatments for ~~cancer~~ **BPH**?
10. What are ~~scientifically-validated~~ **experimental** treatments for **prostate** cancer?
12. What are the side effects of ~~Lexapro~~ **{Proscar, Flomax, Avodart, Ticlid, Viadur ...}** ?
14. What foods should be avoided to prevent ~~cavities-in-children~~ **interactions with alpha blockers** ?

These seven queries constitute our Consumer Use Cases A, B, C, D, E, F, and G. In addition, we invented this one to highlight human-computer interaction issues.

Consumer Use Case H. A patient is taking Ticlid (ticlopidine hydrochloride) to prevent blood clotting on an implanted coronary stent. She is having difficulty breathing and wonders if it might be a side effect of the drug. She looks up Ticlid on MedMaster but the monograph section "What side effects can this medication cause?"<sup>34</sup> does not say anything about respiratory problems. She wishes that MedMaster had a "Search More Resources" button next to each section heading. (See Appendix F for database mappings.)

<sup>34</sup> <http://www.nlm.nih.gov/medlineplus/druginfo/meds/a695036.html#side-effects>

**Table 7. UMLS vs. DailyMed indications for ticlopidine hydrochloride.**

UMLS covers the DailyMed indication "thrombotic stroke" (bold) but incorrectly as treatment rather than prevention, and has 13 other indications not verified by DailyMed. Normalizing the other DailyMed indication "subacute stent thrombosis" via UMLS leads to the wrong concept ("Vascular stent thrombosis" [crossed out] rather than "Coronary stent thrombosis" neither of which is covered as a ticlopidine indication). DailyMed's finer grained indication restriction information for ticlopidine is not covered at all by UMLS.

## 1. Treatment.

<i>DailyMed extracted value</i>	<i>DailyMed normalized value (source)</i>	<i>UMLS (NDFRT) related concept</i>
(none)	(none)	<b>may treat</b> <ul style="list-style-type: none"> <li>Bacterial Vaginosis</li> <li>Dermatomycoses</li> <li>Eye Infections, Bacterial</li> <li>Eye Infections, Viral</li> <li><b>Intracranial Embolism and Thrombosis [~"thrombotic stroke"]</b></li> <li>Leg Ulcer</li> <li>Radiation Injuries</li> <li>Renal tubular acidosis</li> <li>Skin Diseases, Bacterial</li> <li>Skin Diseases, Parasitic</li> <li>Skin Diseases, Viral</li> <li>Staphylococcal Infections</li> <li>Surgical Wound Infection</li> <li>Urination Disorders</li> </ul>

## 2. Prevention.

<i>DailyMed extracted value</i>	<i>DailyMed normalized value (source)</i>	<i>UMLS (NDFRT) related concept</i>
<ul style="list-style-type: none"> <li>reduce the risk of thrombotic stroke</li> <li>reduce the incidence of subacute stent thrombosis</li> </ul>	<ul style="list-style-type: none"> <li>Thrombotic stroke (<i>Meddra</i>, <i>SNOMED</i>, <i>DXplain</i>)</li> <li><del>Vascular stent thrombosis (<i>Meddra</i>)</del></li> <li>Coronary stent thrombosis (<i>Meddra</i>)</li> </ul>	<b>may prevent</b>  (none)

## 3. Restrictions.

<i>DailyMed extracted value</i>	<i>DailyMed normalized value (source)</i>	<i>UMLS (NDFRT) related concept</i>
<ul style="list-style-type: none"> <li>thrombotic stroke - patients who have experienced stroke precursors</li> <li>thrombotic stroke -- patients who have had a completed thrombotic stroke.</li> <li>subacute stent thrombosis -- patients undergoing successful coronary stent implantation</li> <li>[general] -- for patients who are intolerant or allergic to aspirin therapy or who have failed aspirin therapy</li> </ul>	<ul style="list-style-type: none"> <li>(nothing for "stroke precursors"?)</li> <li>Thrombotic stroke (<i>Meddra</i>, <i>SNOMED</i>, <i>DXplain</i>)</li> <li>Insertion of coronary artery stent (<i>SNOMED</i>)</li> <li>Aspirin allergy (<i>SNOMED</i>)</li> </ul>	(none)

**Table 8. UMLS vs. DailyMed contraindications for ticlopidine hydrochloride.**

Of the 15 DailyMed contraindications, warnings, and precautions, five are covered by UMLS/NDFRT *has\_contraindication* relations (checked boxes). UMLS/NDFRT has four other such relations (question marks) not verified by DailyMed.

<i>DailyMed extracted value</i>	<i>DailyMed normalized value (UMLS)</i>	<i>UMLS (NDFRT) related concept</i>
<u><i>Contraindications</i></u> <ul style="list-style-type: none"> <li>• Hypersensitivity to the drug</li> <li>• Hematological Adverse Reactions / hematopoietic disorders</li> <li>• Neutropenia</li> <li>• Thrombocytopenia</li> <li>• hemostatic disorder</li> <li>• active pathological bleeding / bleeding risk [e.g. surgery]</li> <li>• bleeding peptic ulcer</li> <li>• intracranial bleeding</li> <li>• severe liver impairment / hepatically impaired</li> <li>• Thrombotic Thrombocytopenic Purpura (TTP)</li> <li>• Aplastic Anemia</li> </ul>	<ul style="list-style-type: none"> <li><input checked="" type="checkbox"/> Drug Allergy</li> <li>• Hematological Disease</li> <li>• Neutropenia</li> <li>• Thrombocytopenia</li> <li><input checked="" type="checkbox"/> Blood Coagulation Disorders</li> <li><input checked="" type="checkbox"/> Hemorrhage</li> <li>• Peptic Ulcer Hemorrhage</li> <li>• Intracranial Hemorrhages</li> <li><input checked="" type="checkbox"/> Liver diseases</li> <li>• Purpura, Thrombotic Thrombocytopenic</li> <li>• Aplastic Anemia</li> <li>• Anticoagulants</li> <li>• Hypercholesterolemia</li> <li>• Gastrointestinal Hemorrhage</li> <li><input checked="" type="checkbox"/> Kidney Failure</li> </ul>	<ul style="list-style-type: none"> <li><input checked="" type="checkbox"/> Drug Allergy</li> <li><input checked="" type="checkbox"/> Blood Coagulation Disorders</li> <li><input checked="" type="checkbox"/> Hemorrhage</li> <li><input checked="" type="checkbox"/> Liver diseases</li> <li>? Addison's disease</li> <li>? Anuria</li> <li>? Dehydration</li> <li>? hyperkalemia</li> <li><input checked="" type="checkbox"/> Kidney Failure</li> </ul>
<u><i>Additional Warnings</i></u> <ul style="list-style-type: none"> <li>• Anticoagulant Drugs</li> <li>• Cholesterol Elevation</li> </ul>		
<u><i>Additional Precautions</i></u> <ul style="list-style-type: none"> <li>• GI Bleeding</li> <li>• Renally Impaired Patients</li> </ul>		

## Chapter 4. Results

### 4.1 Q1: What are the Dimensions of Drug Information?

**4.1.1 Resource survey.** Initial resource survey results are shown in Table 3 and Appendix B. The 23 resources selected for further dimensional analysis with their websites and generic name coverages are shown in Table 4. The initial 4-domain-by-39-dimension-by-23-resource matrix is shown in Table 9 and the 39 dimensions are defined in Table 10. A generic name coverage overlap analysis is shown in Table 11. The overlap analysis was undertaken to give more meaning to the coverage estimates shown in Table 3 and Table 4, but also addresses a corollary research question, "What is the size of the drug universe?"; i.e., "How many fundamental drug compounds are there?" which involves "What is a drug?" Surprisingly, RXNORM and MeSH, despite covering approximately 5,000 generic names each (the source of our canonical generic drug universe estimate), only overlap each other about 60%. Furthermore they do not cover all the drugs in three resources with much smaller coverages, and RXNORM only covers 33% of USANs. However, the much larger (~16,000) Merck generic names dictionary does not cover all of RXNORM or even DrugBank (~1,800). This means that drug information resources do not agree on extensionally defining even the most fundamental dimension, *generic name*. That is, they do not agree on the most fundamental ontological question, "What is a drug?"

**Table 9. Dimensions of drug information by resources.**

Domain	Dimension	Source	MedMaster	DrugDigest	DailyMed	ClinicalTrials.gov	DrugInfo	RXNORM	Drugs@FDA	WHO-ATC	WHO-DRUG	Int'l Pharm.	INN	USP Dictionary	USAN via AMA	MeSH MH	MeSH all	UMLS	PubChem	ChemDplus	ChEBI	DrugBank	KEGG DRUG	Reactome	HumanCyc
pharmacy	trade names		•	•	•	±	•	•	•		•			•	•	•	•	•	•	•	•	•	±	±	
	dose/form			±	•	±		•	•	±	±						•	•				±	±		
	combo products		•	•	•	•	•	•	•	•	•			•				•	•	•		•	•		
	manufacturer				•				•					•	•				•						
	manuf. code name													•	•	•	•	•	•	•					
	approval info.								•													•			
chemistry	chemical name			•								•	•	•	•	•	•	•	•	•	•	•			•
	CAS#											•		•	•	•	•	•	•	•	•	•	•		
	structure graphic			•								•	•	•	•				•	•	•	•	•		•
	empirical formula			•								•	•	•	•				•	•	•	•	•		•
	InChI																		•	•	•	•			
	SMILES																		•	•	•	•			•
	similar structures																		•	•		•	•		
	H bond donors																		•						
	H bond acceptors																		•						
	molecular weight			•								•			•				•		•	•	•		
	solubility			•								•							•	•		•			
	chem. superclass						•			•						•	•	•			•				
	physical descr.											•										•			
	melting point											•								•		•			
	pKa											•										•			
	other chemistry											•							•	•	•	•	•	•	
biology	molecular target			•						±								•				•	•	±	
	mech. of action			•						±								•				•			
	biological effect			•						±								•				•			•
	metabolism			•														•				•			•
	other ADME			•																		•			
	toxicity																			•		•			
	anatomy									•	•							•							
	bioassay																	•							
	pathways																							•	
clinical	therapeutic class		•	•	•	•	•			•	•	•		•	•	•	•	•	•	•	•	•	•		
	indication		•	•	•	•												•				•			
	contraindication		•	•	•													•				•			
	sideeff/prec/warn		•	•	•													±							
	drug interactions		•	•	•													•				•			
	patient info		•	•	•																	•			
	research lit.				•																	•			
	experimental app's				•										±			•				•			

**Table 10. Dimensions of drug information definitions.**

	generic names	the most common general purpose drug names after they become public knowledge, usually corresponding to single chemical compounds as opposed to specific products such as a bottle of pills. Other definitions include the words "nonproprietary", "official", or "approved" but these do not always apply. We exclude chemical names but admittedly there is a gray area (e.g., acetic acid). Aspirin is a case in point of the nuances since it started out as a proprietary trademark. Synonyms include <i>common name</i> and <i>ingredient</i> .
pharmacy	trade names	proprietary names for marketed drug products; synonyms include <i>trademark</i> and <i>brand name</i> ; e.g., "Bayer Aspirin" and "Bufferin" are both trade names for the generic name "aspirin."
	dose/form	a combination of two dimensions usually specified together. Dose is the quantity (e.g., 50 mg) of active ingredient in one unit (e.g., a tablet), and "form" (short for "dosage form" or "formulation") is the physical unit or medium, often including route of administration information (e.g., oral tablet, oral liquid).
	combo products	more than one active ingredient specified by generic name; e.g., aspirin + caffeine
	manufacturer	company that manufactures the product; e.g., Merck.
	manuf. code name	e.g., L-644,128 and MK-733, Merck code names for the compound that eventually became simvastatin (Zocor).
	approval info.	lumps several different dimensions since approval can apply to name, indication, usage, marketing, and country/agency, and it can have temporal boundaries.
chemistry	chemical name	one of (usually) many ways to express a chemical name for the same compound; standards include CAS (Chemical Abstracts Service) and IUPAC (International Union of Pure and Applied Chemistry); e.g., simvastatin's CAS name is "[1S-[1(alpha),3(alpha),7(beta),8(beta)(2S*,4S*), 8a(beta)]]-2,2-dimethylbutanoic acid 1,2,3,7,8,8a-hexahydro-3,7-dimethyl-8-[2-(tetrahydro-4-hydroxy-6-oxo-2H-pyran-2-yl)ethyl]-1-naphthalenyl ester."
	CAS#	the Chemical Abstracts Service number uniquely assigned to the generic name; e.g., 79902-63-9.
	structure graphic	2D or virtual 3D representation of the spatial arrangement of atoms and bonds in a molecule.
	empirical formula	shows the number of atoms of each element in a compound; e.g., C <sub>28</sub> H <sub>38</sub> O <sub>5</sub>
	InChI	IUPAC Chemical Identifier; enables coding of structures as ASCII text without graphics or arbitrary chemical names.
	SMILES	Simplified Molecular Input Line Entry Specification; another way to code structures as ASCII text.
	similar structures	other chemical compounds with similar structure based a matching algorithm involving atomic composition, 3D bond structure, polarity, etc.
	H bond donors	Lipinsky's Rule of Five is a rule of thumb in drug development. It posits that, for reasons of absorption, distribution, metabolism, and excretion (ADME), a good drug candidate compound must have not more than 5 hydrogen (H) bond donors, not more than 10 H bond acceptors, molecular weight under 500 daltons, and good aqueous solubility (octanol-water partition coefficient log P of less than 5). The name comes from the recurrence of multiples of 5.
	H bond acceptors	
	molecular weight	
	solubility	
	chem. superclass	typically a substructure; e.g., Valium (diazepam; 7-chloro-1,3-dihydro-1-methyl-5-phenyl-2H-1,4-benzodiazepin-2-one) is a <b>benzodiazepine</b> .
	physical descr.	e.g., "white crystalline powder at room temperature."
	melting point	temperature at which the pure substance melts, determination of which is a common initial method of chemical identification.
	pKa	acid dissociation constant, a measure of a compound's acidity.
	other chemistry	includes other identification information, synthesis recipes, references, ...

**Table 10. Dimensions of drug information definitions (continued).**

biology	molecular target	primary endogenous site of action of a drug, usually a macromolecule it binds to and inhibits or stimulates, such as an enzyme, receptor, or gene.
	mech. of action	how the drug works; e.g., by inhibiting the enzyme HMG-CoA reductase, simvastatin reduces liver production of cholesterol which lowers blood cholesterol and thereby reduces cardiovascular risk. These various levels may overlap the molecular target, biological effect, anatomy, pathway, and/or therapeutic class dimensions, and/or be inferable from one of their ontologies such as WHO-ATC.
	biological effect	the more macroscopic end of the mechanism of action continuum; e.g., "lowers blood cholesterol."
	metabolism	macroscopic site (liver, etc.) and molecular pathway information (enzymes, derivatives, ...) about chemical transformation of the drug in body.
	other ADME	other absorption, distribution, metabolism, and excretion (ADME) info
	toxicity	doses and interactions that are toxic, and in what way.
	anatomy	macroscopic or organ system site of action of the drug; e.g., "dermatological" (skin).
	bioassay	bio-identification and -quantitation methods and criteria.
	pathways	metabolic pathways in which the drug's molecular target are involved; often useful for linking the pharmacology of two drugs with different targets.
clinical	therapeutic class	what the drug does, often expressed in anatomic or indication terms, and sometimes chemical; e.g., Valium (diazepam) is a neuropharmacologic agent, an anxiolytic, and a benzodiazepine. See also <i>mechanism of action</i> .
	indication	disease or other medical reason for taking the drug; an important subtype is <i>approved indications</i> . Often semantically equivalent to therapeutic class; e.g., aspirin is an analgesic, meaning it's indicated for pain.
	contraindication	disease or other medical reason for <i>not</i> taking the drug; "pregnancy" is a common one.
	side eff/prec/warn	side effects, precautions, and warnings are distinguished on package inserts and should probably be considered separate dimensions.
	drug interactions	drug effects, usually undesired, resulting from taking two or more drugs concurrently.
	patient info	"Information for Patients" appears to be a mandated package insert section with the five foregoing clinical dimensions populated with information in lay language.
	research lit.	research literature references behind the other clinical assertions.
	experimental app's	potential, not-yet-approved indications and other uses; e.g., use of taxol (an approved antineoplastic) to inhibit microtubule formation in cell biology experiments.

**Table 11. Generic name overlap estimates for some sources.**

Results are from preliminary detailed cross-mapping of the sources' generic name (GN) coverages as described in the text. Yellow background: source matrix; green background: GN coverage count of each source; white background: overlapping GN count between column source and row source; red font: includes Merck proprietary data; ♦ trivial cell (column source = row source). The DrugBank data shown represents the sum of "approved" and "experimental" counts, RXNORM covering 1,276/1,484 and 63/351 respectively.

	Source	U{RXNORM,DrugBank,Merck}	Merck single GN (incl.USAN)	Merck USAN	RXNORM	MeSH D tree; MH only	MeSH A1.4	UMLS A1.4	DrugBank	MedMaster	DrugDigest	DailyMed	ClinicalTrials.gov
Source	#GNs	18,506	15,929	4,317	5,592	2,000	5,000	9,000	1,835	1,000	1,000	1,117	924
RXNORM	5,592			1,432	♦		3,141	5,592	1,339			970	680
MeSH MH D tree	2,000	1,910				♦							
MeSH A1.4	5,000	4,675			3,141		♦					793	873
UMLS A1.4	9,000	8,163			5,592			♦					
DrugBank	1,835		1,187	663	1,339				♦				
MedMaster	1,000									♦			
DrugDigest	1,000		1,000								♦		
DailyMed	1,117		1,107	706	970		793					♦	
ClinicalTrials.gov	924			328	680		873						♦

**4.1.2 Experimental database.** The experimental database described in the Methods section may be downloaded here <http://comminfo.rutgers.edu/~msharp/XKB/DB6.xls>. Without separator lines and headings, the database contains 17,901 rows, meaning each of the 15 compounds is represented by approximately 1,200 rows of data on average. The database schema and some examples are displayed in a more legible way in Table 5 and Table 6 as discussed in Methods. The set of 550 normalized dimensions we found by this method are given by the unique values in column O of the database.<sup>35</sup> Excluded from further Q1 analysis were dimensions representing unparsed links and summaries, database IDs and cross-references, compound dimensions (e.g., *generic name + unit dose + dosage form*), and *information for the patient* subtypes.<sup>36</sup> The remaining 375 are shown in Appendix G as a six-level hierarchy, with the four domains (*pharmacy, chemistry, biology, clinical*) comprising the top level. The second level contains 54 dimensions comparable to the 39 found in the initial study (Table 12). The *biology* domain, despite covering only 17% of the database rows, accounts for 42% of the unique dimensions, including almost all of the fifth and sixth level splitting (Table 13). This is primarily due to the tabular toxicity data in ChemIDplus, which includes route and species subdimensions, the numerous subtypes of ADME measures mentioned in the DailyMed "Pharmacokinetics" and "Pharmacodynamics" sections, and the various parameters associated with drug metabolizing enzymes in DrugBank. At the opposite extreme, the *clinical* top-level dimension accounts for 35% of the data but only 14% of the dimensions.

---

<sup>35</sup> Extracted by copying column O to a scratch worksheet, sorting, and removing duplicate values. The duplicate removal method was: For a sorted list in column A starting on row 1, enter "0" into cell B1 and "=if(A2=A1,1,0)" into B2. Then copy B2 into all B cells down to the end of the A values. Then copy all the B values, right-click B1, and "Paste Special > Values". Then sort both columns on B and delete all rows with "1" in B.

<sup>36</sup> These excluded dimensions were identified by the Excel function "=search(X,A#)" in an adjacent column, where X is a substring of the value in cell A# (for example, " ID"), or by visual scanning. The coding and number of unique dimensions excluded were: unparsed links and summaries: REF/LINK, 11; database IDs and cross-references: ID, 85; compound dimensions: COMPOUND, 77; and *information for the patient*: INFO, 4. The list is included on the "dimensions" sheet of the database. After their removal, two dimensions had to be added to the working list because they were only represented in compound dimensions in the data.

**Table 12. Dimensions found in experimental database - 2-level hierarchy.**

Red signifies coverage at a lower hierarchical level; e.g., *pharmacy - generic name - combination product*.  
 Blue signifies results shown in the database but not in Appendix G.

<i>Initial results (Table 9, Table 10)</i>		<i>Experimental database results (Appendix G)</i>	
		<i>Equivalent to initial results</i>	<i>New 2nd level dimensions</i>
pharmacy	generic names	pharmacy - generic name	pharmacy - administration
	trade names	pharmacy - trade name	pharmacy - DEA schedule
	dose/form	pharmacy - dose pharmacy - dosage form	pharmacy - drug type
	combo products	pharmacy - generic name	pharmacy - generic availability
	manufacturer	pharmacy - approval info	pharmacy - herbal source biology
	manuf. code name	pharmacy - manufacturer code	pharmacy - inactive ingredient
	approval info.	pharmacy - approval info	pharmacy - lexical class
			pharmacy - packaging
			pharmacy - product type
			pharmacy - storage conditions
			pharmacy - unit appearance
biology	molecular target	biology - molecular target	biology - organism affected
	mech. of action	biology - mechanism of action	
	biological effect	biology - biological effect	
	metabolism	biology - ADME	
	other ADME	biology - ADME	
	toxicity	biology - toxicity	
	anatomy	clinical - therapeutic class	
	bioassay		
	pathways	biology - pathway	
chemistry	chemical name	chemistry - chemical name	chemistry - atmospheric OH rate constant
	CAS#	chemistry - CAS number	chemistry - charge
	structure graphic	chemistry - formula	chemistry - chemical class
	empirical formula	chemistry - formula	chemistry - chemical complexity
	InChI	chemistry - formula	chemistry - chemical type
	SMILES	chemistry - formula	chemistry - covalently bonded unit count
	similar structures	chemistry - formula	chemistry - heavy atom count
	H bond donors	chemistry - Lipinski	chemistry - Henry's law constant
	H bond acceptors	chemistry - Lipinski	chemistry - isoelectric point
	molecular weight	chemistry - Lipinski	chemistry - isotope atom count
	solubility	chemistry - solubility	chemistry - polarity
	chem. superclass	chemistry - chemical superclass	chemistry - related chemical
	physical descr.	chemistry - physical properties	chemistry - rotatable bond count
	melting point	chemistry - physical properties	chemistry - stereocenter count
	pKa	chemistry - pKa	chemistry - tautomer count
	other chemistry		chemistry - vapor pressure

**Table 12. Dimensions found in experimental database - 2-level hierarchy (continued).**

Red signifies coverage at a lower hierarchical level; e.g., *clinical - precaution - contraindication*.

Blue signifies results shown in the database but not in Appendix G.

<i>Initial results</i> (Table 9, Table 10)		<i>Experimental database results (Appendix G)</i>	
		<i>Equivalent to initial results</i>	<i>New 2nd level dimensions</i>
clinical	therapeutic class	clinical - therapeutic class	clinical - clinical trial comparison therapy
	indication	clinical - indication	clinical - clinical trial co-therapy
	contraindication	clinical - precaution	clinical - lab test
	side eff/prec/warn	clinical - precaution	clinical - storage conditions
	drug interactions	clinical - precaution	
	patient info	clinical - info for patients	
	research lit.		
	experimental app's	clinical - indication	

**Table 13. Distribution of data and dimensions by top term and hierarchical level.**

<i>domain</i>	<i>data rows</i>	<i>unique dimensions at hierarchical level =</i>					
		<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>total</i>
biology	3079	7	41	42	54	13	157
chemistry	1910	17	23	20	0	0	60
clinical	6208	6	34	10	2	0	52
pharmacy	6705	15	55	29	6	0	105
total	17902	45	153	101	62	13	374
biology	17%	16%	27%	42%	87%	100%	42%
chemistry	11%	38%	15%	20%	0%	0%	16%
clinical	35%	13%	22%	10%	3%	0%	14%
pharmacy	37%	33%	36%	29%	10%	0%	28%

## 4.2 Q2: Do Dimensions Lead to Valid Groupings of Resources?

**4.2.1 Face validity.** Table 14 illustrates a computationally simple extension of the domain classification from the dimensions to the resources. It can be seen that the lexical test described in the Methods section (ChEBI, PubChem, and ChemIDplus being classified under *chemistry*) is passed. One might also interpret the trend toward equitable all-domain coverage (no domain >50%) by the richest (highest number of dimensions and/or database records)

**Table 14. Classifying resources by domains.**

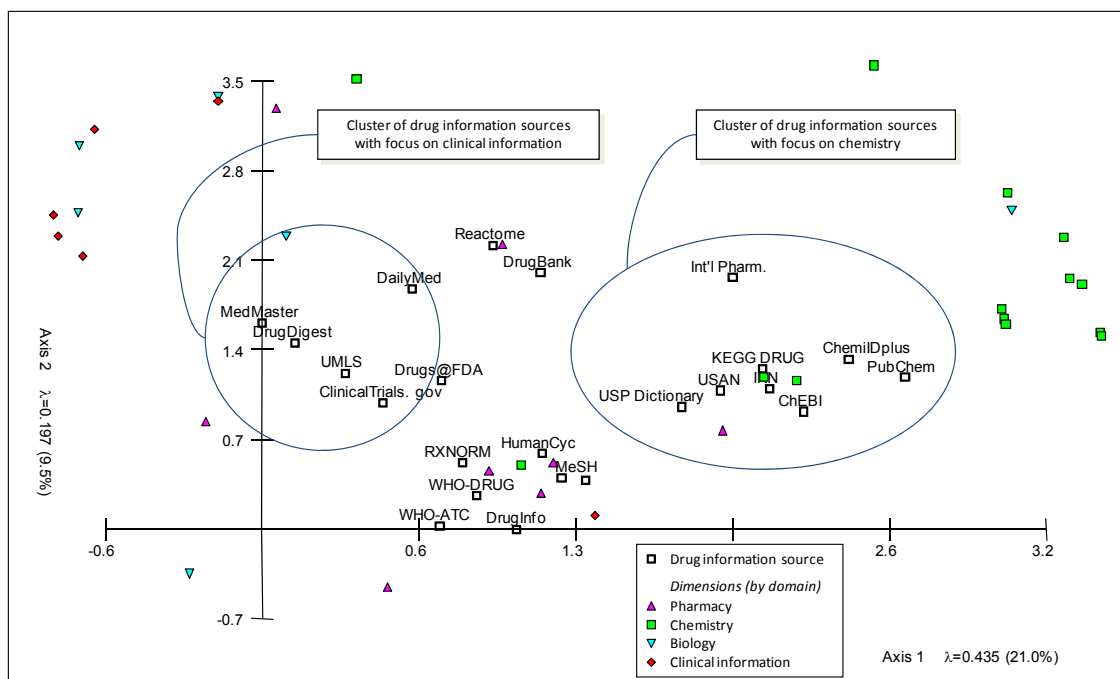
Table 9 data. The percentages represent how many of the dimensions covered by the source ("# dimensions") were grouped by each domain. Sources have been sorted and color-coded to highlight their predominate domain.

<i>source</i>	<i># dimensions</i>	<i>pharmacy</i>	<i>chemistry</i>	<i>biology</i>	<i>clinical</i>
RXNORM	3	100%			
Drugs@FDA	5	100%			
WHO-DRUG	4.5	56%		22%	22%
DrugInfo	4	50%	25%		25%
USP Dictionary	9	44%	44%		11%
MeSH all	7	43%	43%		14%
INN	3		100%		
Int'l Pharm.	11		91%		9%
ChEBI	11	9%	82%		9%
HumanCyc	5		80%	20%	
PubChem	17	18%	71%	6%	6%
ChemIDplus	15	20%	67%	7%	7%
KEGG DRUG	9.5	16%	63%	11%	11%
USAN via AMA	9.5	32%	53%		16%
MeSH MH	6	33%	50%		17%
DrugBank	29.5	12%	44%	20%	24%
Reactome	4	13%	25%	63%	
WHO-ATC	6	25%	17%	42%	17%
MedMaster	7	29%			71%
DrugDigest	8.5	29%			71%
ClinicalTrials. gov	5	40%			60%
DailyMed	21	19%	24%	24%	33%
UMLS	17.5	23%	17%	29%	31%

"all-purpose" resources (DrugBank, WHO-ATC, DailyMed, and UMLS) as being a kind of validation. On the other hand, RXNORM's 100% *pharmacy* classification is inconsistent with Liu et al.'s (2005) claim of utility for "health care personnel including prescribing physicians and nurses" which would seem to correspond better to our *clinical* domain.

**4.2.2 Correspondence analysis.** In the correspondence analysis, the first two principal axes accounted only for about 30% of the total inertia, which means that some points may not have been correctly represented with respect to these two axes. Regardless of the particular weighting schemes used, there was a consistent distinction between *clinical* and *chemistry* domain-classified dimensions (i.e., sources tending to cover clinical dimensions tended not to also cover chemistry dimensions), while the *pharmacy* and *biology* were more diffuse (i.e., they

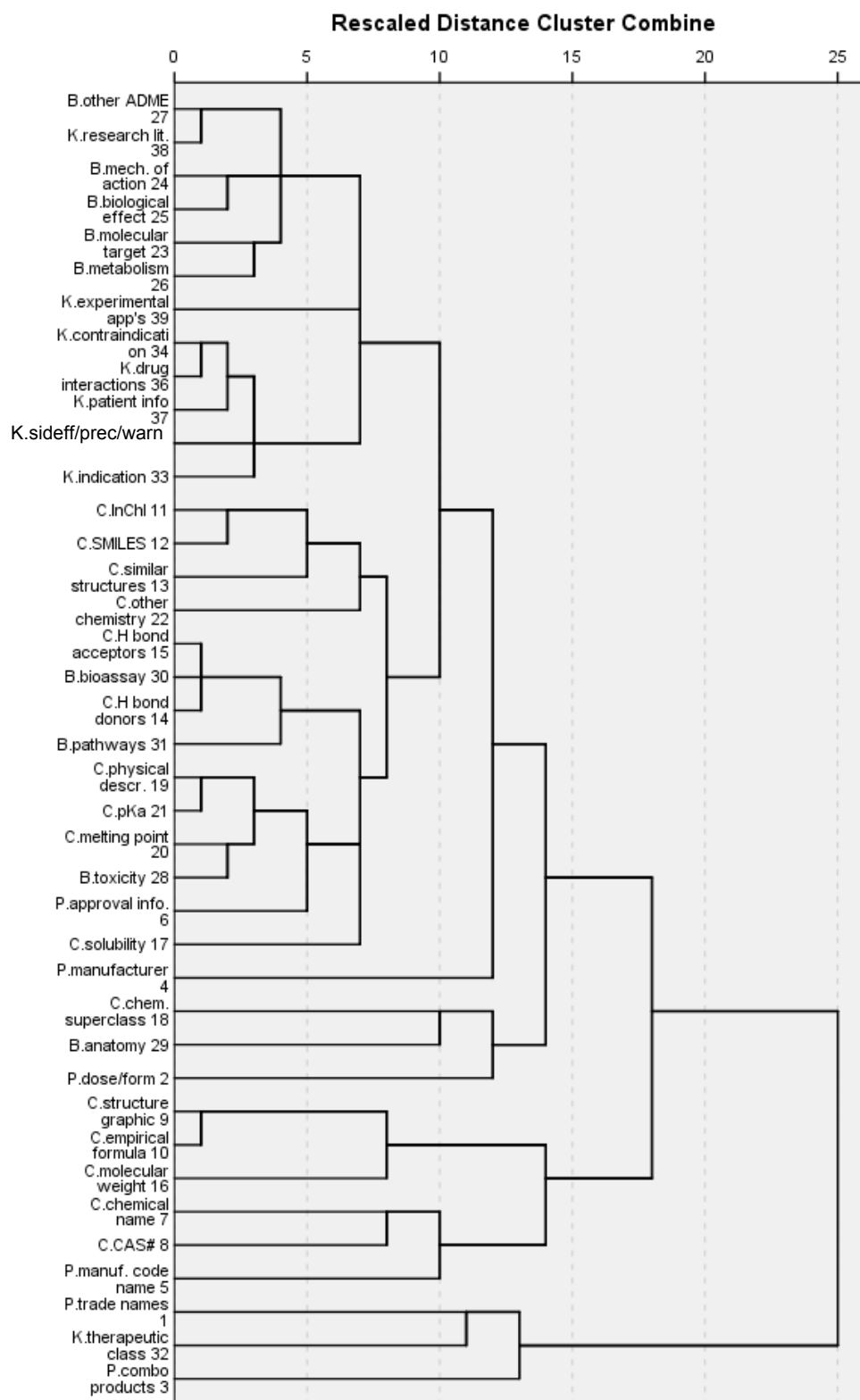
lacked discriminating power). In a joint plot (Figure 9), chemistry-oriented sources (e.g., ChEBI) can be seen polarized to the right and clinical-oriented sources (e.g., MedMaster) to the left. Sources tending to cover all four domains (e.g., DrugBank and WHO-DRUG) are not effectively categorized, and therefore tend toward the center. Similarly, the *therapeutic class* dimension is close to the center because it tends to be covered by the resources we examined regardless of their domain leanings; this might be interpreted as suggesting that *therapeutic class* is important in all four domains.



**Figure 9. Correspondence analysis between drug information sources and dimensions.**

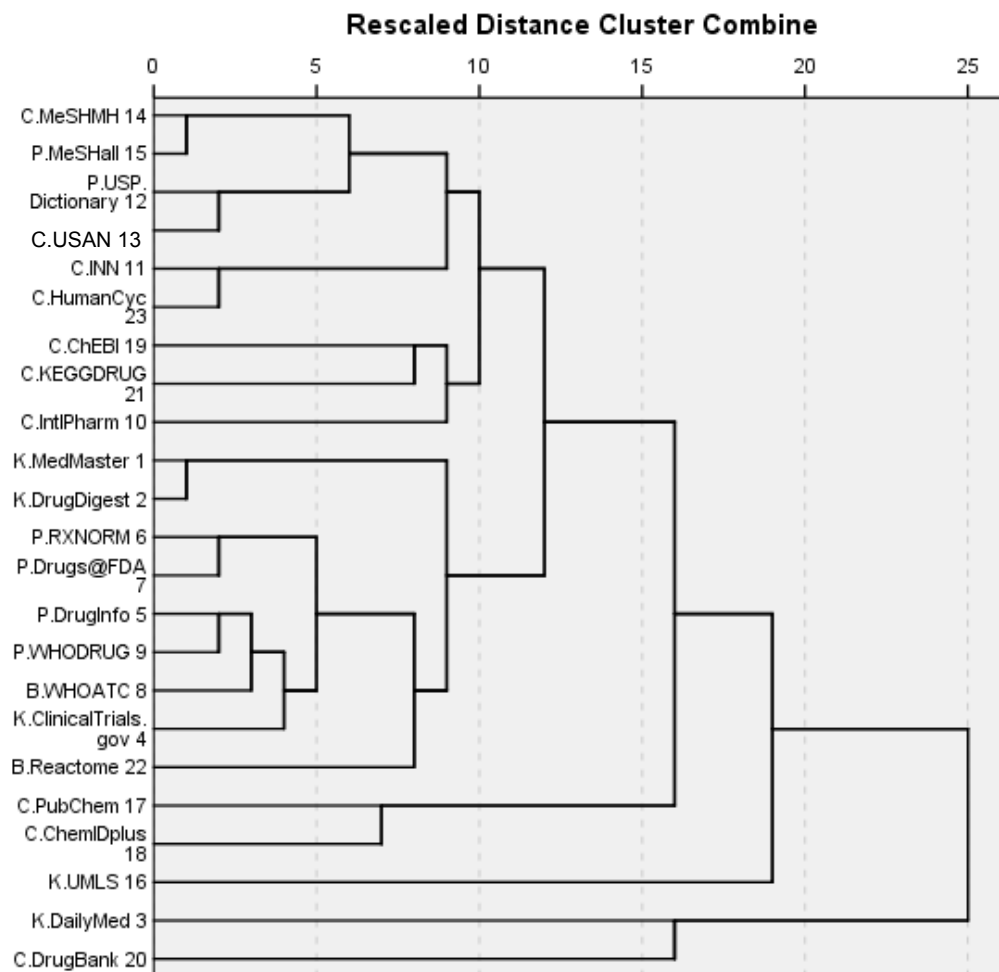
### 4.2.3 Cluster analysis.

**4.2.3.1 Initial resource survey.** To test our four-domain classification of the dimensions and resources (Table 9, Table 14), the Table 9 matrix was subjected to hierarchical cluster analysis (Norusis, 2005). Table 9 was converted to a SPSS dataset with the sources constituting the cases, the dimensions constituting the variables, and scores of 1 (●), 0.5 (±), or 0 (blank). The resulting dendrograms are shown in Figure 10 and Figure 11.



**Figure 10. Cluster analysis of initial survey dimensions.**

Nominal (Table 9) domain classifications are indicated by the leading letter: B biology, C chemistry, K clinical, P pharmacy.



**Figure 11. Cluster analysis of initial survey resources.**

Nominal (Table 14) domain classifications are indicated by the leading letter: B biology, C chemistry, K clinical, P pharmacy.

In Figure 10, at a source co-occurrence similarity distance of five or less, the top cluster - *other ADME, research literature, mechanism of action, biological effect, molecular target, metabolism* - corresponds closely (5/9) to the *biology* domain dimensions of Table 9. A second such cluster - *contraindication, drug interaction, patient info, side effect/precaution/warning, and indication* - includes 5/8 of Table 9's *clinical* dimensions. Both of these clusters coalesce in a super-cluster at similarity distance seven which also includes two more *clinical* dimensions - *experimental applications* and *research literature*. The eighth *clinical* dimension - *therapeutic*

*class* - does not strongly cluster with any other dimension, supporting the correspondence analysis' implication that it is important to all four domains. The next two clusters at distance seven coalesce into a super-cluster at distance eight which contains 10/16 of Table 9's *chemistry* dimensions, along with three *biology* dimensions (*bioassay*, *pathways*, and *toxicity*) and one *pharmacy* dimension (*approval info*). Another small cluster at distance eight contains three more *chemistry* dimensions. Of the four domain-dimension groupings hypothesized in Table 9, only *pharmacy* failed to be supported by this analysis.

In Figure 11, four source clusters may be distinguished at a dimension co-occurrence similarity distance of nine or less. The first one includes MeSH, USP Dictionary, USAN, INN, and HumanCyc, a mixture of *pharmacy* and *chemistry* resources according to Table 14. The second includes ChEBI, KEGG DRUG, and Int.Pharm., all nominal *chemistry* resources. The third includes MedMaster, DrugDigest, RXNORM, Drugs@FDA, DrugInfo, WHO-DRUG, WHO-ATC, ClinicalTrials.gov, and Reactome, a mixture of *pharmacy*, *biology*, and *clinical*. The fourth includes PubChem and ChemIDplus, both *chemistry*. The close similarity (distance of three or less) of MeSH MH and MeSH all, USP and USAN, MedMaster and DrugDigest, RXNORM and Drugs@FDA, and WHODRUG and WHO-ATC, is consistent with each pair's organizational overlap, scope, and/or mission. PubChem and ChemIDplus also form a sensible cluster, albeit at distance seven. The close clustering of INN and HumanCyc, and DrugInfo and WHODRUG, is more perplexing. The nonclustering resources - UMLS, DailyMed, and DrugBank - are distinguished in Table 14 by their high number of dimensions and lack of a dominant (>50%) domain; that is, both analyses support their distinction as "all-purpose" resources.

**4.2.3.2 Experimental database.** We also subjected the resource-by-dimension matrix implied by the experimental database to hierarchical cluster analysis. Unlike the (mostly) binary (0 or 1) scores of the Table 9 matrix, raw scores in the database matrix are how many times the normalized dimension (column O) is represented with a given source (column C) in the same row

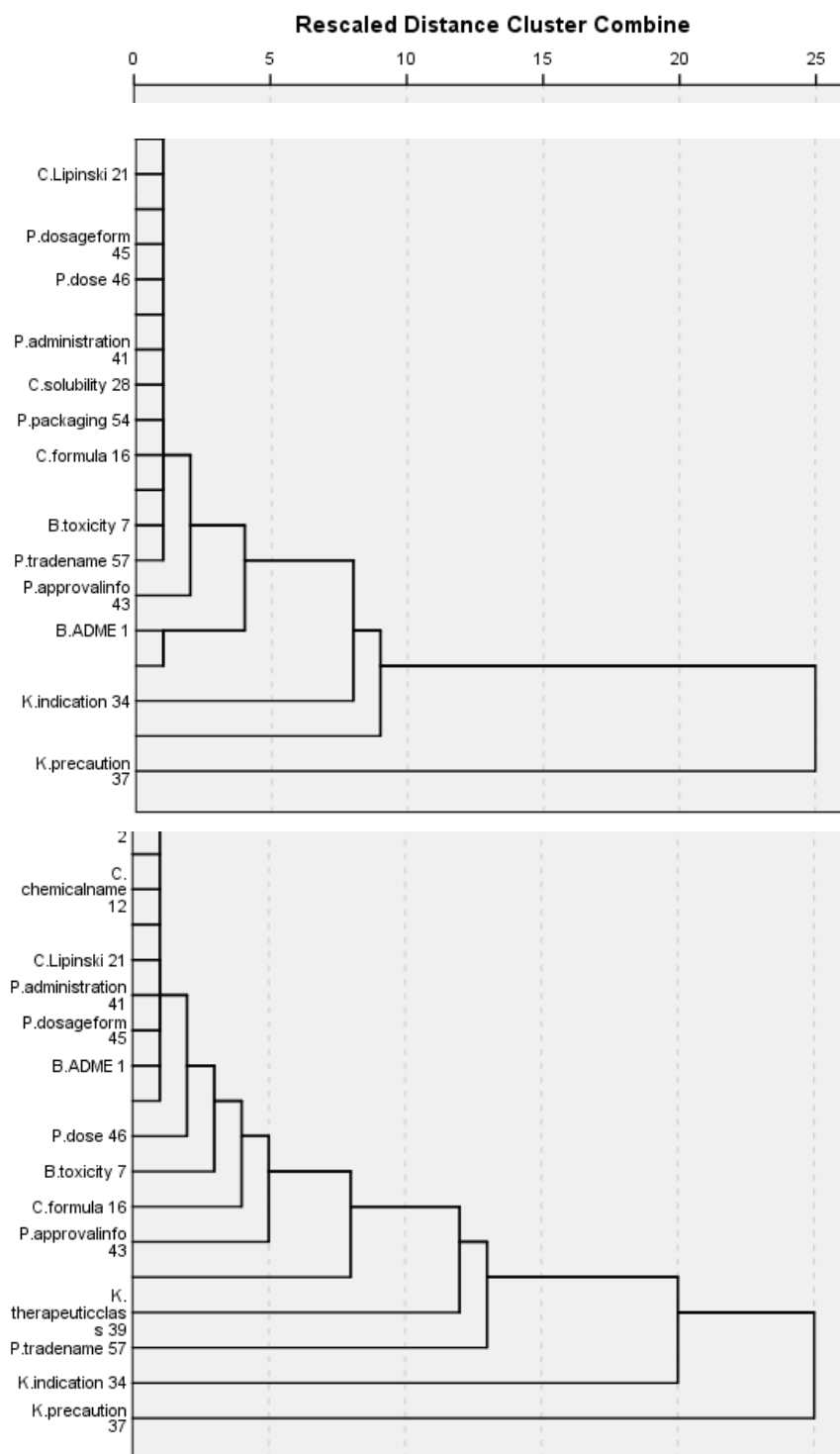
in the database.<sup>37</sup> Cluster analysis was performed on the raw score matrix and also, to correct for possible effects of varying resource richness, scores as a percentage of each resource's total number of records in the database. Compound and other dimensions of uncertain semantics were eliminated (see Section 4.1.2), leaving a set of 399 similar to the 375 shown in Appendix G, and a second level compression of 57 similar to the 54 shown in Table 12.

The second-level dimension dendrograms are shown in Figure 12. (With 399 dimensions, the full six-level dendrograms cannot be legibly displayed on a single page; they are roughly the same "shape" as the second-level dendrograms.) In marked contrast to Figure 10, both the raw score and percentage versions show almost all of the 57 second-level dimensions in a single, flat, tail-like cluster at the top of the dendrogram, and just a few "outsiders" at the bottom, notably *indication* and *precaution*. It is interesting that, in Figure 10, *contraindication*, *drug interaction*, *side effect/precaution/warning*, and *indication* form a tight, purely *clinical*, almost exclusive cluster. In the experimental database, *contraindication*, *drug interaction*, *side effect*, and *warning* are subsumed as third-level dimensions under *precaution*. That is, the database cluster analysis result is the opposite of the initial survey result for *indication* and *precaution*.

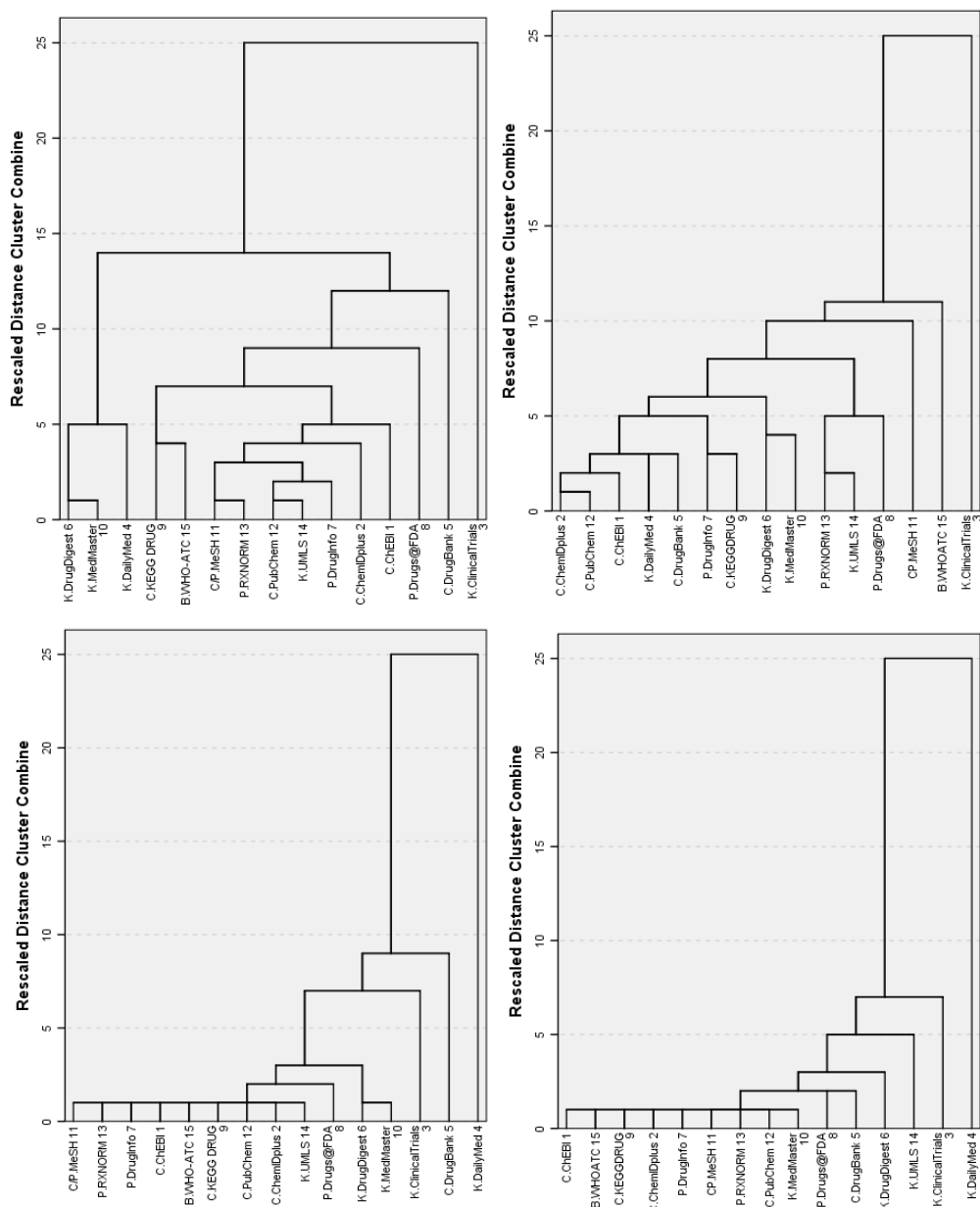
The source dendrograms are shown in Figure 13. The raw score dendrograms have somewhat the same skewed appearance as the dimensions clusters in Figure 12 but it is not as extreme. The nominal (Table 14) biology, chemistry, and pharmacy sources tend to cluster tightly in the long "tail" at a distance of two or less, while the nominal clinical sources (and DrugBank) tend not to cluster. The percentage source dendrograms are not so skewed; their overall "shape" is more like that of the Table 9 analysis (Figure 11). The aforementioned sensible, pairwise, close clustering of DrugDigest and MedMaster is apparent, but not that of RXNORM and Drugs@FDA. The six-level percentage dendrogram shows the tightest clustering

---

<sup>37</sup> Computed using Excel by sorting on C and O, summing consecutive duplicates, and converting this 3-column format to a matrix using the Data > Pivot Table function.



**Figure 12. Cluster analysis of experimental database dimensions (2-level hierarchy).** Nominal (Table 12) domain classifications are indicated by the leading letter: B biology, C chemistry, K clinical, P pharmacy. Upper panel: raw scores. Lower panel: percentage of each source's records. The tops have been cropped for legibility.



**Figure 13. Cluster analysis of experimental database sources.** Nominal (Table 14) domain classifications are indicated by the leading letter: B biology, C chemistry, K clinical, P pharmacy. Upper panels: 2-level dimension hierarchy. Lower panels: raw scores. Right panels: percentage of each source's records.

of nominal chemistry sources. In all four Figure 13 dendrograms, ClinicalTrials.gov's lack of similarity to other sources is strikingly unlike its position in Figure 11, perhaps due to its unusual pairing of few dimensions with large row counts in the database.

### 4.3 Q3: Can Dimensions Facilitate Integration/OM Tasks?

**4.3.1 Classifying sources.** As stated in Methods, this is basically a usefulness version of Q2; i.e., are the classifications implied by the grouping, correspondence, and clustering results for Q2 useful as well as valid? The Table 9 "framework" was judged to be "useful for comparing resources" by Sharp, Bodenreider, and Wacholder (2008). These authors' reporting of the results of the correspondence analysis (closely paraphrased above) implies that the framework was effective at classifying predominantly *chemistry*, *clinical*, and all-four-domain resources, but ineffective at classifying predominantly *biology* and *pharmacy* resources.

Cluster analysis of Table 9 supported its hypothetical domain classification of *biology*, *chemistry*, and *clinical* dimensions, but not *pharmacy*. An exception is *therapeutic class*, nominally a *clinical* dimension, which does not strongly cluster with any other dimension, supporting the notion that it is important to all four domains. The *biology* and *clinical* clusters formed a sensible super-cluster corresponding the well-accepted *biomedical* domain. Source clustering in the Table 9 analysis did not, in general, follow Table 9's (via Table 14) predictions, but did confirm, based on our dimensions, six sensible pairs - MeSH MH and MeSH all, USP and USAN, MedMaster and DrugDigest, RXNORM and Drugs@FDA, WHODRUG and WHO-ATC, and PubChem and ChemIDplus - consistent with each pair's organizational overlap, scope, and/or mission. The nonclustering resources - UMLS, DailyMed, and DrugBank - are distinguished by their high number of dimensions and lack of a dominant (>50%) domain; that is, cluster analysis supports their distinction as "all-purpose" resources.

**4.3.2 Selecting sources appropriate to a given information need.** Using the Table 9 matrix as described in Methods to select the best resources to satisfy the general *indications* usage scenario, we identified UMLS (7), DrugBank (6), DailyMed (5), WHO-ATC (3.5), and

ClinicalTrials.gov (3) as having the highest scores and thereby being the best candidates to satisfy the user's information need. This result was judged to have "demonstrated that this framework is useful for ... selecting sources most relevant to a given use case" by Sharp et al. (2008).

### **4.3.3 Pooling data from different sources.**

**4.3.3.1 Data reduction.** We expected that the number of unique raw representations would always be greater than the number of unique corresponding normalized representations within any cross-section of the database, and that the size of this difference would be a good measure of the effectiveness of our method at integrating the data. However, because diverse raw data formats (terms, relations, items in a list, whole paragraphs, etc.) were loaded into the spreadsheet the same way, there is an antagonism between conflation of short strings representing a single concept (which lowers the ratio of unique normalized to unnormalized representations), versus parsing of longer strings into multiple dimension-value pairs (which raises the ratio).

Our observed ratios of unique normalized to unnormalized representations, expressed as percentages, are shown in Table 15. Overall, the number of unique normalized drug-dimension-value "triples" is 70% of the number of unique raw triples. Thus overall data reduction was achieved, even by this conflicted measure. Individual sources varied from a high of 234% (i.e., 2.34 times as many unique normalized triples as unique raw triples) for MedMaster to 72% for UMLS, basically reflecting a gradient of low-to-high raw dimensionality and high-to-low free-text formatting. (That is, MedMaster has few raw dimensions and lots of free text, while UMLS has many raw dimensions and little free text.) The 50% score for DailyMed, which has lots of free text, appears to contradict this pattern, but is suspect due to internal and practical inconsistencies (not all values were thoroughly parsed and normalized). Note that the 70% overall figure relative to the 72%-234% range suggests that single-concept normalization across sources would have been more dramatically illustrated by this measure without the antagonistic parsing artifact.

When one focuses on the three parts of the triples individually, for the drug part the overall figure is 12% and the range is 22%-117%, basically reflecting how closely each source adheres to standard generic drug naming conventions for defining a set of related drug information. The outliers are PubChem (117%) which has more of a chemical compound view, and DailyMed (22%) which has a commercial product view. The overall dimension figure is 71% and the range is 54%-900%, basically reflecting a gradient of high-to-low raw dimensionality and low-to-high free text. The exceptions to this pattern are WHO-ATC (75%) which has only 8 dimensions but they are not well differentiated by their values' semantic types, and PubChem (358%) which has many well-differentiated *chemistry* dimensions but also some free text and undifferentiated "synonym" lists. The overall value figure is 76% and the range is 69%-181%. All but three sources fell within 88%-108%, suggesting that their value terminology is at least internally consistent. The three exceptions are: ClinicalTrials.gov (69%) and DailyMed (76%), reflecting the diverse authorship of their free text content (study titles in the case of ClinicalTrials.gov); and MedMaster at 181% .

The raw primary drug names tend to be terms, index entries, web page titles, and the like, and so are not as vulnerable to the antagonistic parsing artifact as triples, dimensions, or values. Hence 12% is probably a better estimate than 70, 71%, or 76% for the degree of single-concept normalization across sources we achieved.

Additional data reduction of the dimensions was achieved by hierarchical aggregation (Figure 14 top). At the second hierarchical level (Table 12), the overall normalization ratio is 22%. Except for the outliers DrugInfo and ClinicalTrials.gov, which have only three unnormalized dimensions each, the most dramatic single-source hierarchical effect is seen on ChemIDplus, apparently due to the fifth and sixth level chemistry and toxicology splitting mentioned earlier. However, the same hierarchical aggregation does not lead to significant data reduction of the drug-dimension-value triples by this measure (Figure 14 bottom).

Case-specific examples of data reduction are perhaps more indicative of the effectiveness of our integration method. These will be given in the use case results below.

**Table 15. Data reduction in the experimental database.**

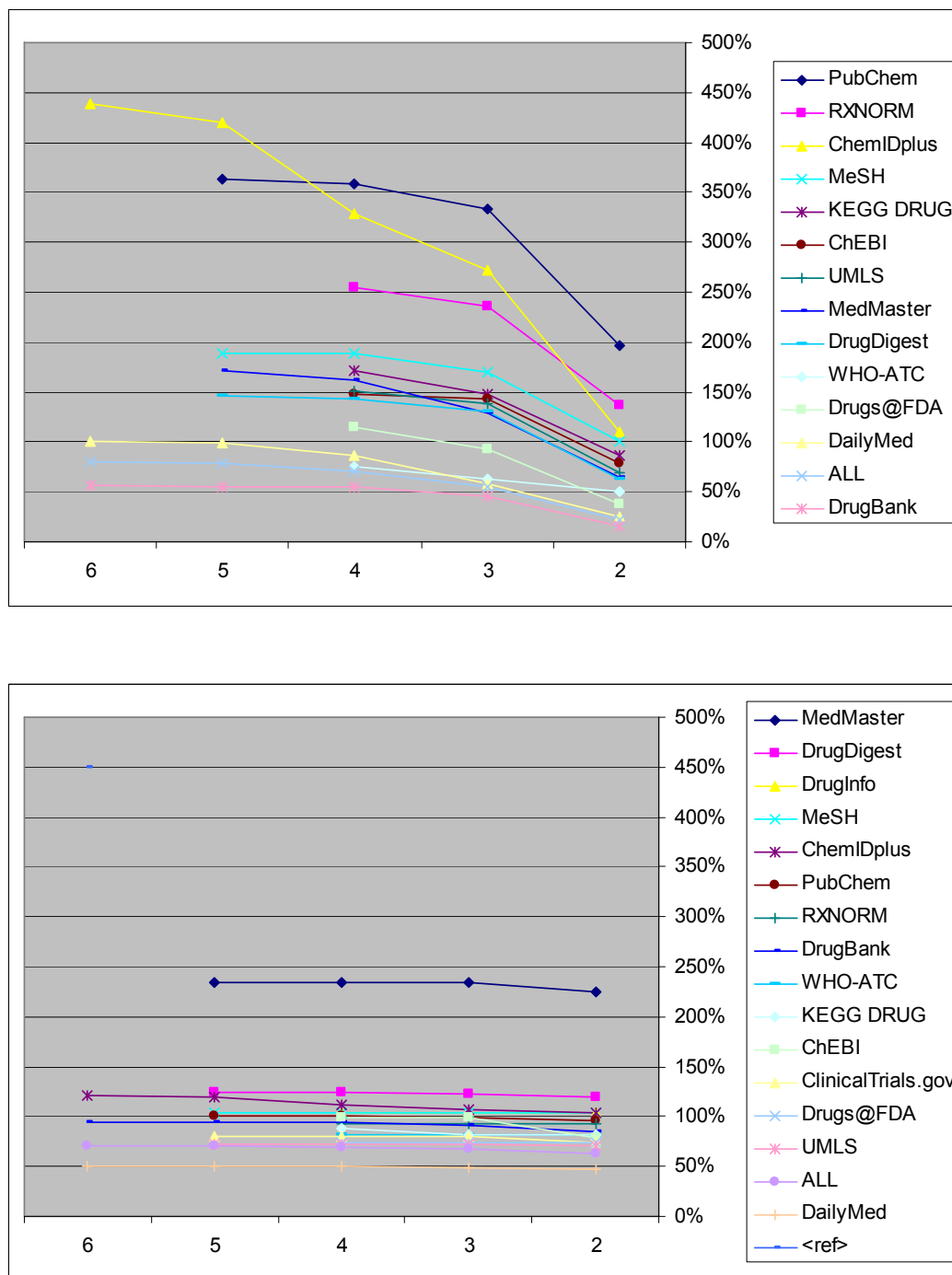
The unique normalized dimensions were taken from the fourth hierarchical level for comparability across sources (cf. Figure 14 and Appendix G). Unlike Appendix G, these calculations did not exclude data representing unparsed links and summaries, database IDs and cross-references, compound dimensions, or information for the patient subtypes.

<i>source</i>	<i>drug-dimension-value triples</i>		
	<i>unnorm</i>	<i>norm</i>	<i>% n/u</i>
ALL	12421	8671	70%
MedMaster	301	705	234%
DrugDigest	639	795	124%
ChemIDplus	618	694	112%
DrugInfo	268	280	104%
MeSH	280	290	104%
PubChem	1105	1115	101%
ChEBI	274	271	99%
DrugBank	1988	1860	94%
RXNORM	296	273	92%
KEGG DRUG	406	357	88%
WHO-ATC	80	66	83%
ClinicalTrials.gov	884	713	81%
Drugs@FDA	428	318	74%
UMLS	1442	1033	72%
DailyMed	3425	1717	50%

<i>source</i>	<i>drug</i>		
	<i>unnorm</i>	<i>norm</i>	<i>% n/u</i>
ALL	123	15	12%
PubChem	12	14	117%
ChEBI	11	11	100%
ClinicalTrials.gov	9	9	100%
DrugBank	8	8	100%
KEGG DRUG	12	12	100%
MedMaster	9	9	100%
MeSH	9	9	100%
UMLS	16	15	94%
DrugInfo	15	14	93%
WHO-ATC	9	8	89%
ChemIDplus	16	14	88%
DrugDigest	16	11	69%
RXNORM	21	10	48%
Drugs@FDA	25	9	36%
DailyMed	36	8	22%

<i>source</i>	<i>dimension</i>		
	<i>unnorm</i>	<i>norm</i>	<i>% n/u</i>
ALL	690	490	71%
DrugInfo	3	27	900%
ClinicalTrials.gov	3	19	633%
PubChem	36	129	358%
ChemIDplus	21	69	329%
RXNORM	11	28	255%
MeSH	17	32	188%
KEGG DRUG	21	36	171%
MedMaster	31	50	161%
UMLS	58	88	152%
ChEBI	23	34	148%
DrugDigest	30	43	143%
Drugs@FDA	13	15	115%
DailyMed	172	150	87%
WHO-ATC	8	6	75%
DrugBank	256	139	54%

<i>source</i>	<i>value</i>		
	<i>unnorm</i>	<i>norm</i>	<i>% n/u</i>
ALL	6517	4978	76%
MedMaster	231	417	181%
DrugDigest	322	348	108%
RXNORM	259	259	100%
ChemIDplus	485	484	100%
MeSH	272	266	98%
DrugInfo	214	206	96%
DrugBank	980	935	95%
ChEBI	230	217	94%
Drugs@FDA	248	227	92%
PubChem	835	761	91%
UMLS	819	736	90%
KEGG DRUG	222	199	90%
WHO-ATC	43	38	88%
DailyMed	1574	1199	76%
ClinicalTrials.gov	820	568	69%



**Figure 14. Data reduction in the experimental database by hierarchical aggregation.**

X axis: dimension hierarchical level. Y axis: percent unique normalized/unnormalized. Top: dimensions alone. Bottom: drug-dimension-value triples. Not shown in top are DrugInfo and ClinicalTrials.gov curves which are offscale (>600% at level 4) due to having only 3 unnormalized dimensions each. The <ref> point in the bottom is to force comparable Y axis ranges. For terminological representation of the dimension hierarchy see Appendix G (all levels) and Table 12 (level 1-2).

**4.3.3.2 Automatic normalization of additional raw data.** The table of probabilities of dimension normalization based on source, unnormalized dimension, and value clues can be viewed at [http://comminfo.rutgers.edu/~msharp/XKB/dimension\\_prediction.xls](http://comminfo.rutgers.edu/~msharp/XKB/dimension_prediction.xls) . The "table" sheet has 1679 rows (not counting the headers in row 1) and 10 columns derived from the experimental database columns as follows:

- A. source (C)
- B. raw dimension (F)
- C. value-dimension clue (L)
- D. normalized dimension (O)
- E. fraction (probability) of A,B (C,F) pairs associated with D (O),  $p(C,F \rightarrow O)$
- F. fraction (probability) of A,B,C (C,F,L) triples associated with D (O),  $p(C,F,L \rightarrow O)$
- G. fraction (probability) of A,B (C,F) pairs associated with hierarchically aggregated D ( $O_{\text{hier}}$ ),  $p(C,F \rightarrow O_{\text{hier}})$
- H. hierarchical level of aggregation used for G
- I. fraction (probability) of A,B (C,F) pairs associated with another hierarchically aggregated D ( $O_{\text{hier}}$ ),  $p(C,F \rightarrow O_{\text{hier}})$
- J. hierarchical level of aggregation used for I

The full database content (all columns) corresponding to the "table" sheet is given in the "data" sheet, so that examples of the other columns' corresponding content may be viewed.

Many of the observed  $p(C,F \rightarrow O)$ 's are 1.0, reflecting the preponderance of semantically well-differentiated raw dimensions in our sources, and their influence on our choice of normalized dimensions. That is, we could not improve upon the semantic differentiation of a raw dimension such as "Molecular Weight", so we essentially imported it wholesale (as *chemistry - molecular weight*) and made sure that this normalization was consistent throughout the database. At the opposite extreme (low  $p(C,F \rightarrow O)$ ) are raw dimensions which seem well-specified but sometimes contain values belonging to a different dimension. For example, KEGG DRUG's

"Activity" dimension contains 16 values, 15 of which are consistent with *clinical - therapeutic class* ( $p(C,F \rightarrow O) = .94$ ), but one of which is "Treatment of benign prostatic hyperplasia", clearly a *clinical - indication - treatment* value ( $p(C,F \rightarrow O) = .06$ ). In between are general "catch-all" raw dimensions such as DailyMed's "Description" or MeSH's "Scope Note" which may contain a wide variety of kinds of information about a drug. Nearly all of these low-to-mid-range  $p(C,F \rightarrow O)$ 's can be dramatically raised (usually to 1.0) based on clues in the value (e.g., "Treatment" in the foregoing example) ( $p(C,F,L \rightarrow O)$ ). Alternatively, smaller but useful gains can be obtained by dimensional hierarchical aggregation ( $p(C,F \rightarrow O_{hier})$ ).

These results could be used to identify a subset of high-precision source-dimension pairs for rapid expansion of the experimental database to encompass the corresponding subset of information (e.g., molecular weights) for all the drugs covered by our resources. However, it does not seem meaningful to count or average these scores across source-dimension pairs as a measure of overall source or dimensional precision because such a measure would not reflect the varying relevance of dimensions to use cases, provenance issues, difficulties around parsing, etc.

Instead, in Table 16 we present a small sample of these results relating to the *clinical - indication* dimension which has been identified as of special interest by prior research in this area. The best precision (as  $p(C,F \rightarrow O)$ ) for this dimension and its hierarchical sub-dimensions *clinical - indication - treatment* and *clinical - indication - treatment - approved* was exhibited by DrugDigest's "Learn how <this drug> is/are used to treat:"; UMLS's "Other Related/may\_be\_treated\_by"; DailyMed's "Indications and Usage"; MedMaster's "Other uses for this medicine"; and DrugBank's "Indication" dimensions. Following these are an assortment of nonspecific "catch-all" dimensions such as "Description"<sup>38</sup> in the 44-59% range, followed by a variety of lower precision source-dimension pairs, the lowest being KEGG DRUG's "Activity" as described above. The best-precision DrugDigest and UMLS dimensions also have the advantage

---

<sup>38</sup> Values for DrugInfo's "Description"; PubChem's "Compound Summary"; MeSH's "Scope Note"; and ChemIDplus's "Notes - Note" are the same for any given drug. DrugBank's "Description" and DrugDigest's "What is/are <this drug>?" are independent.

of having values in controlled terminological format (i.e., no parsing needed), but UMLS does not cover the fourth (*approved*) hierarchical level.

**Table 16. Probabilities of correct automatic dimension normalization by source.**

This example shows  $p(C, F \rightarrow O)$  for the normalized dimension *clinical - indication - treatment - approved* at the three levels of hierarchical aggregation shown, as explained in the text. Blank cells mean the source does not cover those *indication* subdimensions for any of our sample of drugs.

<i>source</i>	<i>raw dimension</i>	<i>normalized dimension hierarchical level</i>		
		2	3	4
		<i>indication</i>	<i>treatment</i>	<i>approved</i>
DrugDigest	Learn how <this drug> is/are used to treat:	1.00	1.00	1.00
UMLS	Other Related/may be treated by	1.00	1.00	
DailyMed	Indications and Usage	1.00	0.68	0.68
MedMaster	Other uses for this medicine	1.00	0.50	
DrugBank	Indication	0.93		
DrugInfo	Description	0.59	0.35	0.02
DrugDigest	What is/are <this drug>?	0.59	0.46	0.46
PubChem	Compound Summary	0.58	0.40	0.05
MeSH	Scope Note	0.56	0.32	
ChemIDplus	Notes - Note	0.54	0.38	0.04
DrugBank	Description	0.44	0.28	
MedMaster	Why is this medication prescribed?	0.43	0.39	0.39
MeSH	Note	0.33	0.33	0.11
MeSH	Indexing Information	0.29		
DrugBank	Pharmacology	0.23		
MedMaster	About your treatment	0.20	0.20	0.20
DrugDigest	Who is this for? - Uses	0.11	0.11	
MedMaster	Background	0.08	0.08	
KEGG DRUG	Activity	0.06	0.06	

However, these data do not take into account the superior provenance of the DailyMed information. Disagreements between the associated DrugDigest, UMLS, and DailyMed values (e.g., Table 7 and Table 8) remind us that  $p(C, F \rightarrow O)$  is really measuring *semantic* (as opposed to *pharmaceutical*) precision. That is, all UMLS values associated with "Other Related/may be treated by" may be valid treatment indications for some drug, but they are not necessarily the correct indications for the specific drug at the other end of the triple (Table 7). This is not necessarily an indictment of UMLS' data quality; it could simply mean that we erred in equating "Other Related/may be treated by" with *clinical - indication - treatment*. On the

other hand, the low  $p(C, F \rightarrow O)$ 's for MedMaster's "Why is this medication prescribed?" (the quintessential definition of *indication*) do not necessarily mean that the approved indications in MedMaster are inconsistent with those in DailyMed (this was not investigated), rather that other kinds of information can also be extracted from that section of the MedMaster drug pages. This may help to explain the MedMaster's anomalously high ratio of normalized to unnormalized drug-dimension-value triples and values (preceding section and Table 15).

Although the pattern-matching algorithms used here are primitive compared to true natural language processing, this exercise demonstrated the important principle of leveraging mechanization to expand the database to a truly practical size for professional drug information users.

#### 4.3.3.3 *Satisfying use cases.*

##### 4.3.3.3.1 *Health care and related personnel.*

##### Health Use Case A. "Find all indications for finasteride."

This query translates to  $\{(B) = \text{finasteride}; (O) = \text{clinical - indication} \dots\}$ <sup>39</sup> which retrieved 101 rows<sup>40</sup> containing 25 clinical trial IDs (see below) and 21 other unique normalized values (Q): alopecia, benign prostatic hyperplasia, chronic central serous chorioretinopathy, healthy,<sup>41</sup> hematospermia, hematuria, hirsutism, idiopathic hirsutism, infertility, male hypogonadism, male pattern alopecia, muscle atrophy, prostate cancer, prostatic disorder, prostatic hyperplasia, prostatic hypertrophy, prostatic neoplasm, retinal disease, sarcopenia, sexual dysfunction, and transurethral resection of prostate. The two approved indications -

---

<sup>39</sup> To summarize the relevant Methods discussion, query translation was accomplished by *ad hoc* manual parsing of the natural language query and mapping its components first to the database schema (columns), and then to specific values within those columns. In this case "indications" maps to the normalized dimension (column O) value "clinical - indication ..." (i.e., any string starting with "clinical - indication ") and "finasteride" maps to the normalized generic name (column B) value "finasteride." Value mappings may be accomplished in Excel by sorting/browsing and/or string searching. See Appendix F for details on selected use cases. The bracketed expression is a shorthand for a search in our database for a boolean AND co-occurrence of values on the same row; here "finasteride" in column B and any string starting with "clinical - indication" in column O. See

Table 5 and Table 6 for column definitions and value examples.

<sup>40</sup> [http://comminfo.rutgers.edu/~msharp/XKB/Health\\_usecaseA.xls](http://comminfo.rutgers.edu/~msharp/XKB/Health_usecaseA.xls)

<sup>41</sup> "Healthy" is a ClinicalTrials.gov *Condition* signifying normal control subjects.

benign prostatic hyperplasia and male pattern alopecia - can be identified by their row co-occurrence with  $\{(O) = \textit{clinical} - \textit{indication} - \textit{treatment} - \textit{approved}\}$ .

### *Criteria for usefulness*

1. Comprehensive coverage. RXNORM's dimensional coverage maps to ours as follows: "Ingredient":*generic name*; "Brand Name":*trade name*; "Dose Form":*dosage form*.<sup>42</sup> In addition, the equivalent of our *unit dose* can be mapped to substrings of RXNORM terms; e.g., the "100 MG" in "Aspirin 100 MG" (Table 1, Table 2, and Figure 1). These mappings account for RxNorm's coverage of "generic names"; "brand names"; and "dose forms" (Bodenreider & Nelson, 2004; Liu et al., 2005; Zeng, Bodenreider, et al., 2006; Zeng et al., 2007). "Terminology" roughly corresponds to our *pharmacy* domain. "Active ingredients"; "drug components"; and "ingredients" appear to be quasi-synonyms for "generic names" and likewise "drug forms" for "dose forms." Only RxNorm's "National Drug Codes (NDCs)" is unequivocally not covered by our UMLS-based integration of RXNORM. In addition, Zeng et al. (2007) mention five *clinical* dimensions mappable to our *therapeutic class*, *indication*, *drug interaction*, *contraindication*, and *side effect* (see Methods). Our experimental database covers all these at the second or third dimensional hierarchical level, plus

- 12 additional second-level *pharmacy* dimensions,
- 5 additional second-level *clinical* dimensions,
- 13 additional third-level *clinical - precaution* dimensions (along with *contraindication*, *drug interaction*, and *side effect*),
- 7 second-level *biology* dimensions,
- 23 second-level *chemistry* dimensions.

---

<sup>42</sup> We follow RXNORM's convention of including *route of administration* in *dosage form* whenever possible; e.g., "oral tablet".

In addition, our database has hundreds of sub-dimensions representing finer granularity classifications, including 88 under the dimensions mappable to RXNORM's and Zeng et al.'s (2007) (Appendix G).

The resources covered by RXNORM and our database seem to be disjoint (no overlap) except, of course, for RXNORM itself, implying that we have 15 times as many resources, but this is not relevant to this use case since RXNORM does not cover indications. If it did, the most likely source would be the UMLS/NDFRT *may\_be\_treated\_by* relations, so our advantage would still be 15-to-1 overall, 10-to-1 for indications. RXNORM's drug coverage is, of course, much higher ( $5592/9 = 621$  times) but does not account for all known drugs (Table 11).

2. Literary warrant fidelity. "Indications" is one of the desired RXNORM enhancements named by Zeng et al. (2007). This paper gives no specific value examples.

### 3. IR performance.

a. Larger retrieval. Restricting the retrieval to rows with the substrings "finasteride" and "indication" in the raw drug name (D) and dimension (F) columns, as opposed to the normalized equivalents (B and O), reduces it from 101 to 8 rows (8%).

b. More robust. The same (D,F)-based retrieval excludes information from 9 of the 10 (B,O)-based retrieval's sources, leaving only DailyMed (10%) and the two approved treatment indications (10%). The linkages to clinical trials are lost. These results apply to the query "Find all indications for finasteride." For the query "Find all approved [in the U.S.] indications for finasteride" DailyMed, DrugDigest, and MedMaster are more competitive with our database. The other sources, including RXNORM and UMLS, would fail completely.

c. More efficient. Data reduction (reduction in the number of unique strings representing the same concept) exhibited by the initial 101 row retrieval was: drug name (B/D) 1/12 (8%); dimension name (O/F) 6/16 (38%); value (Q/H) 21/48 (44%); drug-dimension-value triple 33/54 (61%). (The value and triple figures do not include the clinical trial IDs.) The number of databases was reduced from 10 to one (10%), implying an even greater reduction in

the number of commands, queries, keystrokes, and time. These results apply to the query "Find all indications for finasteride." For the query "Find all approved [in the U.S.] indications for finasteride" MedMaster and DrugDigest remain competitive with our database, but DailyMed compares unfavorably due to the need to read the entire free-text "Indications and Usage" sections of eight different package inserts retrieved by the query "finasteride". Data reduction for the "approved" query is shown as the retrieval in Table 17.

**Table 17. Data reduction in query results for Health Use Case A.**

In all cases the normalized generic name is "finasteride" and the normalized dimension is *clinical - indication - treatment - approved*. "..." indicates additional free text not shown here.

<i>source</i>	<i>un-normalized drug name</i>	<i>un-normalized dimension</i>	<i>un-normalized value</i>	<i>normalized value</i>
MedMaster	Finasteride	Why is this medication prescribed?	Finasteride (Proscar) is used alone or in combination with another medication (doxazosin [Cardura]) to treat benign prostatic hypertrophy (BPH, enlargement of the prostate gland). ....	benign prostatic hyperplasia
MedMaster	Finasteride	Why is this medication prescribed?	... Finasteride (Propecia) is also used to treat male pattern hair loss (a common condition in which men have gradual thinning of the hair on the scalp, leading to a receding hairline or balding on the top of the head.) ...	male pattern alopecia
DrugDigest	Finasteride Tablets (Alopecia)	What is/are <this drug>?	FINASTERIDE (fi NAS teer ide) is used to treat male pattern baldness in men only. This medicine is not for use in women. This medicine may be used for other purposes; ask your health care provider or pharmacist if you have questions.	male pattern alopecia
DrugDigest	Finasteride Tablets (Alopecia)	Learn how <this drug> is/are used to treat:	Benign Prostatic Hyperplasia (BPH)	benign prostatic hyperplasia
DrugDigest	Finasteride Tablets (Benign Prostatic Hyperplasia)	What is/are <this drug>?	FINASTERIDE (fi NAS teer ide) is used to treat benign prostatic hyperplasia (BPH) in men. This is a condition that causes you to have an enlarged prostate....	benign prostatic hyperplasia
DailyMed	Finasteride (Finasteride) Tablet, Film Coated [Actavis Elizabeth LLC.]	Indications and Usage	Finasteride is indicated for the treatment of symptomatic benign prostatic hyperplasia (BPH) in men with an enlarged prostate to: -Improve symptoms ; -Reduce the risk of the need for surgery including transurethral resection of the prostate (TURP) and prostatectomy.	benign prostatic hyperplasia

**Table 17. Data reduction in query results for Health Use Case A (continued).**

<i>source</i>	<i>un-normalized drug name</i>	<i>un-normalized dimension</i>	<i>un-normalized value</i>	<i>normalized value</i>
DailyMed	Finasteride (Finasteride) Tablet, Film Coated [Aurobindo Pharma Limited]	Indications and Usage	Finasteride tablets are indicated for the treatment of symptomatic benign prostatic hyperplasia (BPH) in men with an enlarged prostate to: - Improve symptoms - Reduce the risk of the need for surgery including transurethral resection of the prostate (TURP) and prostatectomy.	benign prostatic hyperplasia
DailyMed	Finasteride (Finasteride) Tablet, Film Coated [Mylan Pharmaceuticals Inc.]	Indications and Usage	Finasteride tablets are indicated for the treatment of symptomatic benign prostatic hyperplasia (BPH) in men with an enlarged prostate to: - Improve symptoms - Reduce the risk of the need for surgery including transurethral resection of the prostate (TURP) and prostatectomy.	benign prostatic hyperplasia
DailyMed	Finasteride (Finasteride) Tablet, Film Coated [Teva Pharmaceuticals USA]	Indications and Usage	Finasteride tablets are indicated for the treatment of symptomatic benign prostatic hyperplasia (BPH) in men with an enlarged prostate to: - Improve symptoms - Reduce the risk of the need for surgery including transurethral resection of the prostate (TURP) and prostatectomy.	benign prostatic hyperplasia
DailyMed	FINASTERIDE Tablet, Film Coated [Dr.Reddy's Laboratories Limited]	Indications and Usage	Finasteride 5 mg Tablets, USP are indicated for the treatment of symptomatic benign prostatic hyperplasia (BPH) in men with an enlarged prostate to: Improve symptoms Reduce the risk of acute urinary retention Reduce the risk of the need for surgery including transurethral resection of the prostate (TURP) and prostatectomy....	benign prostatic hyperplasia
DailyMed	FINASTERIDE Tablet, Film Coated [Northstar Rx LLC]	Indications and Usage	Finasteride is indicated for the treatment of symptomatic benign prostatic hyperplasia (BPH) in men with an enlarged prostate to: - Improve symptoms ; -Reduce the risk of the need for surgery including transurethral resection of the prostate (TURP) and prostatectomy.	benign prostatic hyperplasia
DailyMed	Propecia (Finasteride) Tablet, Film Coated [Merck & Co., Inc.]	Indications and Usage	PROPECIA is indicated for the treatment of male pattern hair loss (androgenetic alopecia) in MEN ONLY. Safety and efficacy were demonstrated in men between 18 to 41 years of age with mild to moderate hair loss of the vertex and anterior mid-scalp area ...	male pattern alopecia
DailyMed	Proscar (Finasteride) Tablet, Film Coated [Merck & Co., Inc.]	Indications and Usage	PROSCAR is indicated for the treatment of symptomatic benign prostatic hyperplasia (BPH) in men with an enlarged prostate to: - Improve symptoms -Reduce the risk of acute urinary retention -Reduce the risk of the need for surgery including transurethral resection of the prostate (TURP) and prostatectomy. ...	benign prostatic hyperplasia

### *Summary of Health Use Case A*

The query could be efficiently translated to Excel operations to retrieve the relevant database rows based on the normalized generic name and dimension. The corresponding normalized values (answers to the query) included 25 clinical trial IDs and 21 indications, mostly experimental/unapproved. The two approved indications - benign prostatic hyperplasia and male pattern alopecia - could be identified by exploiting the hierarchical details of the normalized dimension  $\{(O) = \text{clinical} - \text{indication} - \text{treatment} - \text{approved}\}$ . Our system satisfied the criteria for usefulness as follows: more dimensions and resources than the reference system (RXNORM and Zeng et al.'s (2007) wish list); fidelity to reference's information need; larger retrieval and more resources with normalized than raw value search; and data reduction.

### Health Use Case B. "Find all drugs indicated for benign prostatic hyperplasia."

This query translates to  $\{(O) = \text{clinical} - \text{indication} \dots; (Q) = \text{benign prostatic hyperplasia}\}$  which retrieved 187 rows<sup>43</sup> containing 10 unique normalized drug names (B): doxazosin, doxazosin mesylate, dutasteride, finasteride, prazosin, saw palmetto, tamsulosin, tamsulosin hydrochloride, terazosin, and terazosin hydrochloride. Three of the other five normalized drug names in our database (leuprolide, leuprolide acetate, prazosin hydrochloride) are not retrieved because, unlike finasteride, their related term "Prostatic Hypertrophy" from UMLS/NDFRT (the reason for their inclusion in our database) does not correspond to "benign prostatic hyperplasia" in any of our other sources. The other two (ticlopidine and ticlopidine hydrochloride) were included for other reasons. The dissociation of prazosin and prazosin hydrochloride is noteworthy; is it an artifact or due to a rare substantive pharmacological effect of hydrochloridation? Of the ten normalized generic names retrieved, all but saw palmetto are approved, and it is for treatment  $\{(O) = \text{clinical} - \text{indication} - \text{treatment} - \text{approved}\}$ . Interestingly, the saw palmetto hit comes not only from (C) ClinicalTrials.gov, but also

---

<sup>43</sup> [http://comminfo.rutgers.edu/~msharp/XKB/Health\\_usecaseB.xls](http://comminfo.rutgers.edu/~msharp/XKB/Health_usecaseB.xls)

MedMaster, DrugDigest, DrugInfo, MeSH, PubChem, and ChemIDplus.<sup>44</sup> DailyMed does not cover saw palmetto.

If the intent of "drugs" in the query includes trade names and other types of drug names, a few can be inferred from the retrieval's content in the raw drug name column (D): Cardura, Avodart, Proscar, Serenoa repens, Permixon, Flomax, and Hytrin. Many more could be obtained by the follow-up whole database query  $\{(O) = \textit{pharmacy - trade name} \dots; (B) = [\textit{doxazosin}, \textit{doxazosin mesylate}, \textit{dutasteride}, \textit{finasteride}, \textit{prazosin}, \textit{saw palmetto}, \textit{tamsulosin}, \textit{tamsulosin hydrochloride}, \textit{terazosin}, \textit{terazosin hydrochloride}]\}$ .<sup>45</sup> Drugs involved in clinical trials involving BPH are signified in the retrieval by  $\{(O) = \textit{clinical - indication} \dots \textit{clinical trial condition}\}$ . Although the involvement is not necessarily the (B) drug as a BPH treatment, users seeking experimental treatment options may wish to pursue these leads. This task is facilitated by our clinical trial ID links in column S which can be used to hyperlink to additional clinical trial information on ClinicalTrials.gov; such hyperlinks are of the form <http://clinicaltrials.gov/ct2/show/NCT00736645> where the bold italics signify an ID.

#### *Criteria for usefulness*

In addition to the relevant subset of the foregoing data on Health Use Case A...

a. Larger retrieval. Restricting the retrieval to rows with the substrings "benign prostatic hyperplasia" and "indication" in the raw value (H) and dimension (F) columns, as opposed to the normalized equivalents (Q and O), reduces it from 187 to 28 rows (15%).

b. More robust. In general (not just for benign prostatic hyperplasia), within our resource sample, only ClinicalTrials.gov, DrugDigest, and UMLS enable searching for drugs by indication.<sup>46</sup> UMLS does not link any drugs to "benign prostatic hyperplasia" (only to "Prostatic

<sup>44</sup> DrugInfo, PubChem, and ChemIDplus re-use MeSH's "Scope Note" or "Note" as their "Description", "Summary", and "Note".

<sup>45</sup> In this shorthand the semicolon signifies boolean AND and the commas signify boolean OR.

<sup>46</sup> WHO-ATC is self-defined as a *therapeutic class* classification system, and we have honored that, even though some of the classes are quasi-indication dimensions, including "G04C drugs used in benign

Hypertrophy"). These results apply to the query "Find all drugs indicated for benign prostatic hyperplasia." For the query "Find all drugs approved [in the U.S.] as indications for benign prostatic hyperplasia" only DrugDigest succeeds at all, retrieving 8/10 normalized generic names (80%), missing only tamsulosin hydrochloride and terazosin hydrochloride.

Restricting the retrieval to rows with the substrings "benign prostatic hyperplasia" and "indication" in the raw value (H) and dimension (F) columns, as opposed to the normalized equivalents (Q and O), excludes information from 8/10 sources, leaving only DailyMed and DrugBank (20%). However, only three normalized generic names (doxazosin, prazosin, and saw palmetto) are lost, leaving the other seven (70%). The linkages to clinical trials are lost.

Like {(O) = *pharmacy - trade name ...*; (B) = [...]} for trade names, other follow-up queries could select on other dimension-value pairs to, for example, eliminate products with certain *dosage forms, inactive ingredients, or precautions (side effects, contraindications, warnings, drug interactions, food interactions, ...)*. This type of query will be illustrated in the next use case.

c. More efficient. Data reduction exhibited by the initial 187-row retrieval was: drug name (B/D) 10/50 (20%); dimension name (O/F) 7/23 (30%); value (Q/H) 1/95 (1%); drug-dimension-value triple 31/150 (20%). The number of databases was reduced from 10 to five (50%), implying an even greater reduction in the number of commands, queries, keystrokes, and time.

#### *Summary of Health Use Case B*

The query could be efficiently translated to Excel operations to retrieve the relevant database rows based on the normalized dimension and value. The corresponding normalized generic names (answers to the query) included 10 of the 15 in the database because the sample was largely chosen to satisfy this use case based on UMLS/NDFRT's broader category of

---

prostatic hypertrophy" which can be used to identify dutasteride, finasteride, tamsulosin, and terazosin (and, by extension, tamsulosin hydrochloride and terazosin hydrochloride).

prostatic hypertrophy. Three of the other five are not retrieved apparently because they are for some other kind of prostatic hypertrophy, and the other two were put into the database for other reasons. A parent/salt dissociation was noted which could have pharmacodynamic implications. The only unapproved drug - saw palmetto - could be identified by exploiting the hierarchical details of the normalized dimension. The query could be easily expanded to retrieve associated trade names and clinical trials. Our system satisfied the criteria for usefulness as follows: more dimensions and resources than the reference system (RXNORM and Zeng et al.'s (2007) wish list); fidelity to reference's information need; larger retrieval and more resources with normalized than raw value search; and data reduction.

Health Use Case C. "Your patient complains about taking daily ~~Alendronate~~ **leuprolide** doses. Is there an alternative dosage form **where frequency of administration < 1/day?**"

This query translates to two database searches. The first is  $\{(B) = \text{leuprolide } \dots; (O) = \text{pharmacy} - \text{administration} - \text{frequency}\}$  which retrieved two rows. The corresponding raw drug names (D) and normalized values (Q) were  $\{(D) = \text{LUPRON DEPOT (leuprolide acetate) injection, powder, lyophilized, for suspension [Abbott Laboratories]; (Q) = 1/month}\}$  and  $\{(D) = \text{Viadur (leuprolide acetate) [Bayer Pharmaceuticals Corporation]; (Q) = 1/yr}\}$ . The second search is  $\{(O) = \text{pharmacy} - \text{dosage form}; (D) = [\text{LUPRON DEPOT } \dots, \text{Viadur } \dots]\}$  which retrieved two rows with the value  $\{(Q) = \text{implant}\}$  for both. That is, the answer is "yes, there is an implantable form of leuprolide (acetate) available as a once a month (Lupron Depot) or once a year (Viadur) implant."

To be sure they are approved for the same indications as other forms of leuprolide, one may search on  $\{(B) = \text{leuprolide } \dots; (O) = \text{clinical} - \text{indication} - \text{treatment} - \text{approved}\}$ . The resulting values (Q) are central precocious puberty, endometriosis, prostate cancer, and uterine fibroids.  $\{(D) = \text{LUPRON DEPOT } \dots\}$  co-occurs only with endometriosis and uterine fibroids, and  $\{(D) = \text{Viadur } \dots\}$  co-occurs only with prostate cancer. Therefore, if age and gender of the

patient in question are known, they can be used to qualify the initial result. If the patient is a child, the indication must be central precocious puberty, so the answer changes to "no or unknown." If the patient is an adult female, the indication must be endometriosis or uterine fibroids, therefore "yes, Lupron Depot monthly implant." If the patient is an adult male, the indication must be prostate cancer, therefore "yes, Viadur yearly implant."

### *Criteria for usefulness*

Health Use Cases C-J are pharmacy use cases (queries) adapted from Kupferberg and Jones Hartel (2004) as described in Methods. This satisfies the literary warrant criterion for usefulness for all of them. These authors ranked the overall performance on these queries of five drug information resources but did not present the query results, hence we cannot compare them to ours. Their five resources<sup>47</sup> have no overlap with ours. The specificity of the queries usually resulted in much smaller retrievals from our database than for Health Use Cases A-B, making the quantitative data reduction results somewhat meaningless. Therefore we will present our observations on usefulness for Health Use Cases C-J in a briefer, less structured way. Any individual source not mentioned may be assumed to fail the use case completely.

For Health Use Case C, all our definitive data on Lupron Depot and Viadur came from DailyMed. Unlike our database, it is not possible to search DailyMed by dosage form, frequency of administration, or indication. One would have to read the free-text "Dosage and Administration" and "Indications and Usage" sections of all 11 package inserts retrieved by the query "leuprolide".

### *Summary of Health Use Case C*

The query could be efficiently translated to Excel operations to retrieve the relevant database rows based on the normalized generic name, dimension, and value. The corresponding raw drug names and values were equivalent to the answer "yes, there is an implantable form of

---

<sup>47</sup> AHFS Drug Information (STAT!Ref); DRUGDEX (Micromedex); eFacts (Drug Facts and Comparisons); Lexi-Drugs Online (Lexi-Comp); and the PDR Electronic Library (Micromedex).

leuprolide (acetate) available as a once a month (Lupron Depot) or once a year (Viadur) implant." The query could be easily expanded to verify that these products are approved for the same indications as other forms of leuprolide, leading to the refinement that, if the patient is a child, the answer changes to "no or unknown"; if the patient is an adult female, the answer is "yes, Lupron Depot monthly implant"; and if the patient is an adult male, the answer is "yes, Viadur yearly implant." Except for literary warrant, the criteria for usefulness could not be evaluated for Health Use Cases C-J because there was no reference system or results. For Health Use Case C, one may consider DailyMed to be the reference system since all our definitive data came from it. DailyMed cannot be searched on the relevant dimensions. Instead the user would have to read the free-text "Dosage and Administration" and "Indications and Usage" sections of all 11 package inserts retrieved by the query "leuprolide".

Health Use Case D. "What condition is ~~Pyridate~~ **{finasteride, dutasteride, doxazosin, saw palmetto, ...}** used to treat?"

Health Use Case E. "What is ~~Canthaxanthin~~ **{finasteride, dutasteride, doxazosin, saw palmetto, ...}** used for? Is it approved for use in the U.S.?"

These are essentially the same as Health Use Case A above and so would result in similar evidence of usefulness.

Health Use Case F. "Is there an interaction between Warfarin and ~~Fluconazole~~ **{finasteride, dutasteride, doxazosin, saw palmetto, ...}**?"

This query translates to {(O) = *clinical - precaution - drug interaction*; (Q) = warfarin} which retrieved 11 rows containing five normalized drug names (B): saw palmetto, tamsulosin, tamsulosin hydrochloride, ticlopidine, and ticlopidine hydrochloride. That is, the answer is "yes" for these five normalized generic names and "no or unknown" for the other 10 in our database. Extension of this result to trade and other alternative names could be done via {(O) = *pharmacy -*

*trade name ...; (B) = [...]* as discussed under Health Use Case B.

#### *Criteria for usefulness*

This result pools information from four sources. DrugDigest covers 4/5 of the normalized generic name hits (80%), MedMaster 3/5 (60%), DrugBank 2/5 (40%), and DailyMed 1/5 (20%). Only DrugDigest provides a comprehensive way to search for specific drug interactions with efficiency comparable to our database's design principle. Of course, due to our small drug sample, it might be possible to obtain more hits on other sources by searching on (in this case) "warfarin". But for the query "Which anti-BPH drugs interact with warfarin?" all would compare unfavorably to our database by our usefulness criteria since DrugDigest's indication and drug interaction search features are not integrated.

#### *Summary of Health Use Case F*

The query could be efficiently translated to Excel operations to retrieve the relevant database rows based on the normalized dimension and value. The corresponding normalized generic names were equivalent to the answer "yes" for five of them and "no or unknown" for the other 10 in our database. Extension of this result to trade and other alternative names could be done as in Health Use Case B. Our database pooled information from four sources, none of which covered all five warfarin interactions by itself, and only one of which permits comparably efficient dimension-based searching

Health Use Case G. "What is the ~~pediatric dose~~ **dosing regimen** of ~~Aceclovir~~ **{finasteride, dutasteride, doxazosin, saw palmetto, ...}** for ~~chicken pox~~ **{BPH, hypertension}**?"

This query translates to {(O) = *pharmacy - dose - dosing regimen ...; (Q) = [hypertension, benign prostatic hyperplasia]}*. The retrieval includes 18 rows with {(B) = *doxazosin mesylate; (O) = pharmacy - dose - dosing regimen - indication-specific; (Q) = [1-16 mg 1/day [hypertension]], 1-8 mg 1/day [benign prostatic hyperplasia]]}*. Each of the two

different (Q) values corresponds to the same nine DailyMed package inserts for the product names (D) Cardura and eight generic versions of doxazosin mesylate produced by different manufacturers. The same 18 rows account for all the *pharmacy - dose - dosing regimen ...* values for doxazosin and doxazosin mesylate in our database. The other drugs in our database require two queries: one to retrieve (B) with  $\{(O) = \text{clinical - indication - ... - approved}; (Q) = [\text{hypertension, benign prostatic hyperplasia}]\}$ , and one to retrieve (Q) with  $\{(B) = \langle \text{first query results} \rangle; (Q) = \text{pharmacy - dose - dosing regimen ...} \}$ .

#### *Criteria for usefulness*

All the  $\{(Q) = \text{pharmacy - dose - dosing regimen ...} \}$  data in our database comes from DailyMed but, unlike our database, it is not possible to search DailyMed by dosing regimen. One would have to first know which drugs to look up (Health Use Case B), then query DailyMed on them individually and read the free-text "Dosage and Administration" sections of all package inserts retrieved. For Cardura this would not be burdensome, but for doxazosin it would. For the other drugs in our database there would be the additional problem that it is not possible to search DailyMed by indication.

#### *Summary of Health Use Case G*

The query could be efficiently translated to Excel operations to retrieve the relevant database rows based on the normalized dimension and value. The corresponding normalized generic name and values were equivalent to the answer (for doxazosin mesylate only) "1-16 mg once a day for hypertension and 1-8 mg once a day for benign prostatic hyperplasia." The other drugs in our database require an additional query for indications. Again using DailyMed as a reference system, it is not possible to search DailyMed by dosing regimen. One would have to first know which drugs to look up (Health Use Case B), then query DailyMed on them individually and read the free-text "Dosage and Administration" sections of all package inserts retrieved. For all the relevant drugs besides doxazosin mesylate there would be the additional problem that it is not possible to search DailyMed by indication.

Health Use Case H. "A 24-year-old pregnant woman has ~~Trichomonas vaginalis~~ alopecia. Can Metronidazole {finasteride, dutasteride, doxazosin, saw palmetto, ...} be safely used?"

This query translates to two searches, the first to identify drugs indicated for alopecia, the second to evaluate their safety for pregnant women. The first search {(O) = *clinical - indication - ...*; (Q) = alopecia} retrieved three rows, one with {(B) = dutasteride; (C) = ClinicalTrials.gov} and two with {(B) = finasteride; (C) = UMLS}, thus identifying dutasteride and finasteride as the only two relevant drugs in our database. The string "safely used" does not occur in our database, but a health care or pharmacy professional would presumably be able to map this query to our dimension *contraindication* or its higher-level dimension *precaution*. Therefore the second search is {(B) = [dutasteride, finasteride]; (O) = *clinical - precaution - contraindication*; (Q) = pregnancy}. It retrieved 28 rows, 8 with {(B) = dutasteride} and 20 with {(B) = finasteride}. Thus the answer is "no"; finasteride and dutasteride are both contraindicated for pregnant women, which means they cannot be safely used.

#### *Criteria for usefulness*

UMLS is the only one of our sources that covers both the alopecia (as opposed to male pattern alopecia) indication and pregnancy contraindication relations of finasteride, but it does not cover either one for dutasteride. ClinicalTrials.gov only covers the alopecia indication for dutasteride. DailyMed, DrugBank, and DrugDigest cover the pregnancy contraindication for both drugs but not the alopecia indication. DrugBank allows the most straightforward lookup by generic name and its raw dimension "Contraindications" but this only works for finasteride; for dutasteride DrugBank lists pregnancy under "Interactions" instead. DrugDigest's raw dimension that lists pregnancy is consistent for both drugs but its name is even more cryptic: "What should I tell my health care providers before I take this medicine?" DailyMed consistently has this information under "Contraindications" but has its usual disadvantage of having to retrieve and read multiple package inserts for the same normalized generic name (8 in the case of finasteride).

DailyMed, DrugBank, and DrugDigest all require dealing with free text and only DrugDigest allows searching by indication (but not contraindication).

#### *Summary of Health Use Case H*

The query translates to two searches, the first to identify drugs indicated for alopecia (dutasteride and finasteride), the second to evaluate their safety for pregnant women. The string "safely used" does not occur in our database, but a health care or pharmacy professional would presumably be able to map this query to our dimension *contraindication* or its higher-level dimension *precaution*, producing the answer "no"; finasteride and dutasteride are both contraindicated for pregnant women, which means they cannot be safely used. Our database performed this use case more completely and efficiently than any contributing individual source (UMLS, ClinicalTrials.gov, DailyMed, DrugBank, or DrugDigest) alone.

Health Use Case I. "Is ~~Heparin~~ {**finasteride, dutasteride, doxazosin, saw palmetto, ...**} excreted in breast milk?"

Searching on "breast milk" throughout the database hit on the unnormalized value (H) "Studies in lactating rats given a single oral dose of 1 mg/kg of [2-14C]-CARDURA indicate that doxazosin accumulates in rat breast milk with a maximum concentration about 20 times greater than the maternal plasma concentration. It is not known whether this drug is excreted in human milk. Because many drugs are excreted in human milk, caution should be exercised when CARDURA is administered to a nursing mother." This (H) value corresponded to {(O) = *clinical - precaution*; (Q) = breast feeding}. The search {(O) = *clinical - precaution* ...; (Q) = breast feeding} retrieved 58 rows, 24 with {(O) = *clinical - precaution*} and 32 with {(O) = *clinical - precaution - contraindication*}. The combined list of corresponding normalized generic names (B) included 14 of the 15 normalized generic names in our database, missing only tamsulosin hydrochloride.

Here the user had to do some reading and interpretation of the 58 raw (H) values. Only two, both pointing to prazosin hydrochloride, specifically said that the drug "has been shown to be excreted in small amounts in human milk." Others, like the first search hit, specifically affirmed the query for rat breast milk; these pointed to  $\{(B) = [\text{doxazosin mesylate, ticlopidine hydrochloride}]\}$ . These and others stated that it is not known whether the drug is excreted in human milk. Others simply advised caution or non-use by breast feeding women. Any of these details could be the intent of the query and so define which of the 15 retrieved normalized generic names are true hits. In any case, our database successfully answered the query in a few minutes.

#### *Criteria for usefulness*

The initial "breast milk" to  $\{(O) = \text{clinical - precaution; (Q) = breast feeding}\}$  mapping came from DailyMed, as did the specific narrowly affirmative results for prazosin hydrochloride, doxazosin mesylate, and ticlopidine hydrochloride. The remaining data came from DrugDigest (12 normalized generic names), MedMaster (6), DailyMed (2), UMLS (2), and DrugBank (1). Only UMLS enables searching on "breast feeding" as a contraindication but offers no way to map "excreted in breast milk" to it. DailyMed's contribution was specifically identifiable as coming from its raw sub-dimension (F) "Precautions - Nursing Mothers" but has its usual disadvantage of having to deal with free text from multiple package inserts for the same normalized generic name.

Data reduction exhibited by the 58-row retrieval was: drug name (B/D) 14/53 (26%); dimension name (O/F) 2/10 (20%); value (Q/H) 1/23 (4%); drug-dimension-value triple 20/58 (34%). The number of databases was reduced from five to one (20%), implying an even greater reduction in the number of commands, queries, keystrokes, and time.

#### *Summary of Health Use Case I*

Searching on "breast milk" throughout the database hit on an unnormalized value mapping the query to the normalized dimension-value pair *precaution*-"breast feeding" which could be used to retrieve 14 of the 15 normalized generic names in our database. Then the user would have to read and interpret the 58 raw free-text values. Only two specifically said that the

drug has been shown to be excreted in human milk. Others specifically affirmed the query for rat breast milk, others stated that it is not known whether the drug is excreted in human milk, and others simply advised caution or non-use by breast feeding women. Any of these details could be the intent of the query and so define which of the 14 retrieved normalized generic names are relevant. Our database performed this use case more completely and efficiently than any contributing individual source (DailyMed, DrugDigest, MedMaster, UMLS, or DrugBank) alone.

Health Use Case J. "What percentage of patients receiving ~~Methyldopa~~ **doxazosin mesylate** develop a ~~positive Coombs test~~ **hypotension**?"

This query translates to {(O) = *clinical - precaution - side effect*; (Q) = hypotension} which retrieved eight rows, all with {(B) = doxazosin mesylate; (C) = DailyMed; (F) = Adverse Reactions; (H) = Hypotension 1.7%\* 0.0%}. Using the hyperlink in column G of any of these rows, the user could link to the original DailyMed webpage's Adverse Events section and ascertain that the "0.0%" is the corresponding placebo score, the asterisk means " $p \leq 0.05$  for treatment differences," and that the data came from clinical trials.<sup>48</sup>

#### *Criteria for usefulness*

This clinical trial figure of 1.7% does not really answer the query about patients, which requires post-marketing surveillance data, but it's the best DailyMed or our database can do for the prevalence of hypotension as a side effect of doxazosin mesylate. DailyMed might well be able to answer the same query better for other drugs, but our database cannot since many of the long text and table content of the DailyMed package inserts was not loaded and/or normalized due to the time/effort burden; these are signified by a blank (Q) value. Some of these might yield results based on string searches in the (H) values (e.g., for "hypotension" and "%"); this was not investigated. DailyMed would also be expected to be more robust than our database for different side effects other than hypotension for the same reason.

<sup>48</sup> <http://dailymed.nlm.nih.gov/dailymed/drugInfo.cfm?id=6702#n1m34084-4>

Our database, DailyMed, MedMaster, DrugDigest, DrugBank, and to some extent UMLS are much better at answering the true/false query "Does drug X have side effect Y?"; Health Use Case A-B-like queries "Find side effects for drug X"<sup>49</sup> and "Find drugs that exhibit side effect Y"; and the latter's negative "Find drugs that *do not* exhibit side effect Y." In addition, our database, DailyMed, MedMaster, and DrugDigest make some attempt to sub-classify side effects as common, rare, major, serious, etc.

#### *Summary of Health Use Case J*

The query could be efficiently translated to Excel operations to retrieve the relevant database rows based on the normalized generic name, dimension, and value. The corresponding raw values were equivalent to the answer "1.7% of subjects in clinical trials of doxazosin mesylate developed hypotension." This does not really answer the query about *patients*, which requires post-marketing surveillance (rather than clinical trial) data, but it's the best DailyMed or our database can do for the prevalence of hypotension as a side effect of doxazosin mesylate. DailyMed might well be able to answer the same query better for other drugs and/or different side effects. Our database, DailyMed, MedMaster, DrugDigest, DrugBank, and to some extent UMLS are much better at answering the true/false query "Does drug X have side effect Y?"; Health Use Case A-B-like queries "Find side effects for drug X" and "Find drugs that exhibit side effect Y"; and the latter's negative "Find drugs that *do not* exhibit side effect Y."

#### *4.3.3.3.2 Pharmaceutical discovery researchers.*

Research Use Case A. A cluster of structurally similar compounds targeting the TACR1 gene product (known to be associated with abnormal pain threshold ) was found that points to the WHO-ATC class "antiemetics and antinauseants", suggesting that TACR1 modulation may produce antinauseant activity, and/or that there is a possible connection between antinauseant activity and abnormal pain threshold (Castle et al., 2007).

---

<sup>49</sup> See Consumer Use Case F below.

*Adaptation:* A cluster of structurally similar (**quinazoline**) compounds targeting the ~~TACR1~~ **alpha1 adrenergic receptor** gene product (known to be associated with ~~abnormal pain threshold~~ **12 unique target biological correlates**) was found that points to the WHO-ATC class ~~"antiemetics and antinauseants"~~ **127 independent drug biological correlates**, suggesting that ~~TACR1~~ **alpha1 adrenergic receptor** modulation may ~~produce antinauseant activity~~ **affect the latter**, and/or that there ~~is a~~ **are** possible **unknown** connections ~~between antinauseant activity and abnormal pain threshold~~ **among the 12 x 127 pairs**.

We searched on {(O) = *chemistry - chemical superclass*} and identified a cluster of 38 rows with {(Q) = *quinazoline*} pointing to six normalized generic names (B): doxazosin, doxazosin mesylate, prazosin, prazosin hydrochloride, terazosin, and terazosin hydrochloride.<sup>50</sup> A follow-up search {(B) = [doxazosin..., prazosin..., terazosin...]; (O) = *biology - molecular target*} retrieved {(Q) = [alpha1A adrenergic receptor, alpha1B adrenergic receptor, alpha1C adrenergic receptor, alpha1D adrenergic receptor]}. These targets' biological correlates were then identified by searching on {(S) = [alpha1A adrenergic receptor, alpha1B adrenergic receptor, alpha1C adrenergic receptor, alpha1D adrenergic receptor]; (O) = [*biology - molecular target - general function, biology - molecular target - specific function, biology - molecular target - GO biological process, biology - molecular target - pathway*]}}, resulting in 12 unique values<sup>51</sup> (Table 18).

For the next step, we first reproduced exactly the method of Castle et al. (2007) by using {(B) = [doxazosin..., prazosin..., terazosin...]; (O) = *clinical - therapeutic class - WHO-ATC 5th level code*}, resulting in {(Q) = *antihypertensives*; (B) = [doxazosin..., prazosin...]} and {(Q) = *drugs used in benign prostatic hypertrophy*; (B) = terazosin...}.<sup>52</sup> Since both of these are well-known correlates of alpha1 adrenergic receptor activity, we cast a wider net for drug biological

<sup>50</sup> [http://comminfo.rutgers.edu/~msharp/XKB/Research\\_usecaseA\\_cluster.xls](http://comminfo.rutgers.edu/~msharp/XKB/Research_usecaseA_cluster.xls)

<sup>51</sup> [http://comminfo.rutgers.edu/~msharp/XKB/Research\\_usecaseA\\_bio\\_target.xls](http://comminfo.rutgers.edu/~msharp/XKB/Research_usecaseA_bio_target.xls)

<sup>52</sup> The actual Q values were C02CA and G04CX; "antihypertensives" and "drugs used in benign prostatic hypertrophy" are the decodes.

correlates based on {(B) = [doxazosin..., prazosin..., terazosin...]; (O) *biology - biological effect, biology - mechanism of action, clinical - indication ..., clinical - therapeutic class*}, resulting in 127 unique drug biological correlates<sup>53</sup> (Table 19). Thus, our system's "answers" in this use case are these 127 hypothetical effects of alpha1 adrenergic receptor modulation, and the 1,524 (=127 x 12) hypothetical connections between them and Table 18's list of 12 target biological correlates.

---

**Table 18. Target biological correlates for Research Use Cases A and B.**

alpha1 adrenergic receptor activity
carbohydrate transport and metabolism
cell communication
cell surface receptor linked signal transduction
cellular process
extracellular calcium influx
G protein coupled receptor protein signaling pathway
G protein mediated activation of a phosphatidylinositol calcium second messenger system
G(11) protein mediated activation of a phosphatidylinositol calcium second messenger system
G(q) protein mediated activation of a phosphatidylinositol calcium second messenger system
phosphatidylinositol-calcium second messenger system
signal transduction

---

<sup>53</sup> [http://comminfo.rutgers.edu/~msharp/XKB/Research\\_usecaseA\\_bio\\_drug.xls](http://comminfo.rutgers.edu/~msharp/XKB/Research_usecaseA_bio_drug.xls)

**Table 19. Drug biological correlates for Research Use Case A.**

acute urinary retention	catheter ablation	peripheral vasodilator
adrenergic agent	chronic heart failure	pheochromocytoma
adrenergic antagonist	chronic hepatitis C	platelet aggregation inhibitor
alcohol craving	cirrhosis	portal hypertension
alcohol dependence	cirrhosis complications	postsynaptic adrenergic
alcoholism	cocaine abuse	inhibition
allergic rhinitis	cocaine craving	post-traumatic stress disorder
alpha adrenergic antagonist	cocaine dependence	post-traumatic stress disorder
alpha blocker	combat disorder	- civilian
alpha1 adrenergic antagonist	combat stress symptoms	post-traumatic stress disorder
alpha1 adrenergic contraction	complicated hypertension	- combat trauma
antagonist	congestive heart failure	post-traumatic stress disorder
alpha1 adrenergic pressor	coronary heart disease	- noncombat trauma
antagonist	diabetes mellitus	prostatic disorder
alpha1A adrenergic	diabetic nephropathy	prostatic hyperplasia
antagonist	erectile dysfunction	prostatic hypertrophy
alpha1B adrenergic	essential hypertension	psychomotor agitation
antagonist	falling	Raynaud disease
alpha1D adrenergic	female voiding dysfunction	Raynaud syndrome
antagonist	fibrosis	resistant hypertension
alpha2A adrenergic	G protein coupled receptor	rhinitis medicamentosa
antagonist	ligand	rhodopsin family amine
alpha2B adrenergic	gastrointestinal hemorrhage	receptor ligand
antagonist	heart disease	scorpion envenomation
alpha2C adrenergic	heart failure	sleep disorder
antagonist	hepatitis C - chronic	smooth muscle relaxation
Alzheimer disease	hypercholesterolemia	spinal cord injury
Alzheimer disease-related	hyperhidrosis	stress-induced
agitation	hypertension	cocaine/alcohol craving
anti-benign prostatic	hypertension - mild	and relapse
hyperplasia agent	hypotensive	stroke
antidepressant induced	hypotensive agent	supine hypertension in
excessive sweating	insomnia	autonomic failure
antihypercholesterolemic	kidney failure	systemic vascular resistance
agent	lower urinary tract symptoms	decrease
antihypertensive	lowers serum cholesterol	tachyphylaxis
antineoplastic	microvascular angina	uncontrolled hypertension
anxiety disorder	microvascular angina pectoris	unknown
arterial vasodilation	mood disorder	untreated hypertension
atrial fibrillation	morning surge	ureterolithiasis
autonomic dysreflexia	myocardial infarction	urethral resistance decrease
benign prostatic hyperplasia	myocardial ischemia	urinary obstruction
bladder smooth muscle	nephrolithiasis	urinary retention
relaxation	neurogenic bladder	urological agent
bladder sphincter tone	neurotransmitter agent	variceal bleeding
decrease	nightmare	vascular disease
blood pressure decrease	nocturia	vascular smooth muscle
cardiovascular disease	orthostatic hypotension	inhibition
cardiovascular disorder	overactive bladder	vasodilation
catecholamine	peripheral adrenergic	vasodilator
vasoconstrictor	antagonist	venous vasodilation
inhibition	peripheral vasodilation	

### *Criteria for usefulness*

#### 1. Comprehensive coverage.

Table 20 shows the dimensions covered by our model database that are equivalent to those used or mentioned by Castle et al. (2007), Yildirim et al. (2007), Campillos et al. (2008), Quan (2007), and Boguski et al. (2009). Of the composite set of 17 dimensions, our model covers 14 (82%) compared to 23-47% by these papers. Those that our model does not cover are etiological vs. palliative drugs (Yildirim et al., 2007) and "group members" (Quan, 2007). Conversely, our model covers 46 additional second-level dimensions (4 *biology*, 4 *clinical*, 16 *pharmacy*, and 22 *chemistry*; Table 12) and hundreds of sub-dimensions (Appendix G) not covered by these papers.

Of the nine data resources (other than for drug chemical structure/similarity) used by Castle et al. (2007), Yildirim et al. (2007), and Campillos et al. (2008), we used four: DrugBank, WHO-ATC, DailyMed, and UMLS. We did not use Castle et al.'s custom Merck development compound data, Jackson Labs Mammalian Phenotype Ontology, BLAST for target sequence, or Matador and PDSP-Ki for targets. We relied mainly on DrugBank (Yildirim et al. relied exclusively on it) for target information, and our diverse-source *clinical - indication ...* data for phenotypes. For drug chemical structure/similarity our database provides 16 sub-dimensions of *chemistry - formula ...* (see Appendix G) populated with values from ChemIDplus, ChEBI, DailyMed, DrugBank, KEGG DRUG, and PubChem, in contrast to Castle et al.'s and Campillos et al.'s single highly abstracted measures. In addition, our database provides 22 other second-level *chemistry* dimensions; *biology - ADME*, *- biological effect*, *- mechanism of action*, and *- pathway* data; and more diverse and robust *clinical - therapeutic class ...* data besides WHO-ATC. Most importantly, we provide the genuine *clinical - indication ...* data (for which all these papers are using WHO-ATC as a weak substitute) from diverse resources ranging from strictly approved-only (DailyMed) to experimental (UMLS; ClinicalTrials.gov), classified that way and by treatment vs. prevention.

**Table 20. Research use cases dimensional coverage.**

Semantically equivalent dimensions across models (row 1) are represented on the same row starting on row 3. Row 2 is the number of dimensions in the column.

<i>Our database</i>	<i>Castle et al. (2007)</i>	<i>Yildirim et al. (2007)</i>	<i>Campillos et al. (2008)</i>	<i>Quan (2007)</i>	<i>Boguski et al. (2009)</i>
14	8	6	6	4	7
pharmacy - generic name	drug [names]	drug [names]	generic names		
		etiological drugs			
		palliative drugs.			
chemistry - formula - structural formula OR ...	compound structure		chemical structure.		
clinical - indication	disease states	diseases	indication		indications
clinical - indication	diseases/ phenotypes				mechanisms of diseases
clinical - indication - ... - approved				primary disease	
clinical - indication [not approved]				alternative diseases	off-label indications
clinical - therapeutic class - WHO-ATC 5th level code	drug therapeutic activity classifications [WHO-ATC]	drug therapeutic classifications [WHO-ATC]	therapeutic class [WHO- ATC]		
clinical - precaution - side effect			side effects		adverse side effects
biology - molecular target	<protein [target]>	proteins; drug- target associations	molecular targets	target	
biology - molecular target - protein sequence	protein [target] sequence.				
biology - molecular target - gene name	genes				
biology - mechanism of action					mechanisms of drug effects/action
biology - pathway					biological pathways
biology - pathway					molecular pathways of disease
				group [people] members.	

2. Literary warrant fidelity. Our {(O) = *chemical superclass* ; (Q) = quinazoline} cut is a weak substitute for Castle et al.'s chemical similarity measure, but this was necessitated by our limited chemistry knowledge and our database's small drug sample size, not its dimensional coverage. As stated in the preceding paragraph, our database has numerous alternative *chemistry* data types which could be used by a subject expert to compute chemically similar drug clusters in ways more equivalent to Castle et al.'s, given an equivalent drug sample size. All our other adaptations/results are semantically equivalent to Castle et al.'s: "alpha1 adrenergic receptor" to "TACR1 gene product"; "antihypertensives" and "drugs used in benign prostatic hypertrophy" to "antiemetics and antinauseants"; Table 18's list of 12 target biological correlates to "abnormal pain threshold"; and Table 19's list of 127 drug biological correlates to "antinauseant activity."

### 3. IR performance.

a. Larger retrieval. Clearly, Table 19's list of 127 hypothetical effects of alpha1 adrenergic receptor modulation is larger than the single {TACR1:antinauseant activity} hypothesis generated in Castle et al.'s example. Some of these are trivial (e.g., alpha1 adrenergic antagonist) or well known (heart disease) but others might be productive (Alzheimer disease; spinal cord injury). Similarly, the same list of 127 drug correlates multiplied by Table 18's list of 12 target correlates produces 1,524 hypothetical connections, clearly more than the example's single {antinauseant activity:abnormal pain threshold}. Again, some are trivial ({alpha1 adrenergic antagonist:alpha1 adrenergic receptor activity}) or well known ({diabetes mellitus:carbohydrate transport and metabolism}) but others might be productive ({cocaine abuse:extracellular calcium influx}).

b. More robust. Our 12 target correlates, like Castle et al.'s one, all came from data in DrugBank, so it is likely an effect of our expanding the range of fields used from DrugBank's equivalent of *biology - molecular target - specific function* to also include *biology - molecular target - general function*, *biology - molecular target - GO biological process*, and *biology - molecular target - pathway*. In contrast, our 127 drug correlates came from 13 of our

sources (all but Drugs@FDA and RXNORM). If all but WHO-ATC are eliminated, the retrieval is reduced to 2 (<2%), comparable to Castle et al's one. If all but WHO-ATC and DrugBank are eliminated, this figure is 16 (13%). That is, even if Castle et al. had exploited DrugBank more fully as we did, our more diverse resource collection would still generate six times more hypotheses.

c. More efficient. Data reduction exhibited by the combined target (451 rows) and drug (367 rows) correlate retrieval was: drug name (B/D) 6/28 (21%); dimension name (O/F) 14/41 (34%); value (Q/H) 139/150 (93%); drug-dimension-value triple 348/372 (94%). (The value and triple figures do not include the clinical trial IDs.) The number of databases was reduced from 13 to one (8%), implying an even greater reduction in the number of commands, queries, keystrokes, and time. The 139 normalized values are shown in Table 20 and Table 21.

#### *Summary of Research Use Case A*

The goal here was to combine known drug-function, drug-target, and target-function relations to produce a set of hypothetical function-function and novel target-function relations. The required queries could be efficiently translated to Excel operations to retrieve the relevant database rows based on the normalized generic name, dimension, and value. Our system's answer was the 1,524 hypothetical function-function connections between Table 18 and Table 19 and the 127 hypothetical target-function relations between the alpha1 adrenergic receptor and Table 19. Our system satisfied the criteria for usefulness as follows: more dimensions and resources than the reference system; fidelity to reference's information need; larger retrieval and more resources with normalized than raw value search; and data reduction. Most importantly, we provide the genuine *indication* data for which the reference system used WHO-ATC as a weak substitute. Our *chemical superclass* cut is a weak substitute for the reference system's chemical similarity measure, but this was necessitated by our limited chemistry knowledge and our database's small drug sample size, not its dimensional coverage.

Research Use Case B. The WHO-ATC class "cardiovascular system" points to a list of cardiovascular drugs whose gene targets map to a smaller list of phenotypes. The highest ranking phenotype is "decreased heart rate" which is consistent with the WHO-ATC class. This suggests that other WHO-ATC → drug → gene target → phenotype mappings might be mined for phenotype:disease hypotheses (Castle et al., 2007).

The search {(O) = *clinical - therapeutic class - WHO-ATC 5th level code*; (Q) = C...} retrieved {(B) = [doxazosin, doxazosin mesylate, terazosin, terazosin hydrochloride]}. In our first simulation, these drugs' "gene targets" were identified by the search {(B) = [doxazosin..., terazosin...]; (O) = *biology - molecular target*} yielding {(Q) = [alpha1A adrenergic receptor, alpha1B adrenergic receptor, alpha1D adrenergic receptor]} and "phenotypes" was replaced by Table 18's list of 12 biological correlates of the alpha1 adrenergic receptor.

This simulation failed to produce the desired results. The target biological correlates (Table 18) were of a semantic type that might more accurately be called "bioprocesses" rather than phenotypes, and so resulted in a set of bioprocess:disease rather than phenotype:disease hypotheses (e.g., "heart rate:cardiovascular" rather than "decreased heart rate:cardiovascular"). Furthermore, all 12 of them had to do with processes which, unlike "decreased heart rate," are not specific to the cardiovascular system.<sup>54</sup> Therefore we performed a second simulation where "gene targets" was removed from the logic and "phenotypes" was replaced by the 74 drug biological correlates of doxazosin and terazosin as defined under Research Use Case A (Table 22).

---

<sup>54</sup> Ten of them have to do with cell-level processes not specific to the cardiovascular system. One (carbohydrate transport and metabolism) is an entirely different body system corresponding to parts of the A section of WHO-ATC. Only the trivial "alpha1 adrenergic receptor activity" maps to the cardiovascular section of WHO-ATC, and, unlike "decreased heart rate," it can also be mapped to other body systems, even by WHO-ATC's standards (G04 - urologicals ... G04CA - alpha-adrenoceptor antagonists; R - respiratory system ... R03AA - alpha- and beta-adrenoceptor agonists).

"Disease[s]" were identified by the search  $\{(B) = [\text{doxazosin...}, \text{terazosin...}]; (O) = \text{clinical - indication ...}\}$  yielding 42 unique (Q) values (Table 21).<sup>55</sup> Interestingly, these 42 drug:disease correlates also included non-cardiovascular concepts, perhaps suggesting that they should not be discounted from the phenotype simulation results, since the goal of generating credible hypotheses is not restricted to the cardiovascular system. For example, Table 21's "stroke" matches up credibly with Table 22's "platelet aggregation inhibitor." If all results are considered, the first simulation generated  $12 \times 42 = 504$  bioprocess:disease hypotheses (Table 18 x Table 21), and the second generated  $32 \times 42 = 1,344$  phenotype-like:disease hypotheses (Table 22 x Table 21).<sup>56</sup> This constitutes our system's "answer" in this use case.

#### *Criteria for usefulness*

1. Comprehensive coverage. Same as Research Use Case A.
2. Literary warrant fidelity. Our database was able to produce a list of cardiovascular drugs based on their WHO-ATC classifications and map them to their molecular targets, but it does not cover phenotypes *per se*. As a substitute for phenotypes, our first simulation used target biological correlates to remain as true as possible to the target-based approach of Castle et al. However, this simulation failed to produce any matches comparable to their "decreased heart rate:cardiovascular" match. Our second simulation used drug biological (effect, mechanism, and pathway) correlates as a substitute for phenotypes. This produced a set of matches which looked more like "decreased heart rate:cardiovascular" but in so doing it strayed from the target-based approach of Castle et al.

---

<sup>55</sup> Does not include "<negative>" or clinical trial IDs.

<sup>56</sup> The latter  $1,344 = 42 \text{ (Table 21)} \times 32$ , where  $32 = 74 \text{ (Table 22)} - 42 \text{ (Table 21)}$ . That is, since *indication* (Table 21's "disease" substitution) was also one of the drug biological correlates used for "phenotype" (Table 22), Table 21 is a subset of Table 22 and the  $42 \times 42$  self pairings in the matrix do not count as hypotheses.

**Table 21. Drug indications ("disease") for Research Use Case B.**

WHO-ATC body system classification: black: cardiovascular; red: not cardiovascular. "Kidney failure" maps to "C03 diuretics" while "nephrolithiasis" and lower urinary tract concepts map to "G04 urologicals."

acute urinary retention	prostatic hyperplasia	untreated hypertension
essential hypertension	cardiovascular disorder	coronary heart disease
nephrolithiasis	hypertension	microvascular angina pectoris
antidepressant induced	prostatic hypertrophy	ureterolithiasis
excessive sweating	catheter ablation	diabetes mellitus
heart disease	hypertension - mild	morning surge
nocturia	resistant hypertension	urinary obstruction
atrial fibrillation	cocaine abuse	diabetic nephropathy
heart failure	kidney failure	myocardial infarction
overactive bladder	stroke	urinary retention
benign prostatic hyperplasia	cocaine dependence	erectile dysfunction
hypercholesterolemia	lower urinary tract symptoms	myocardial ischemia
prostatic disorder	uncontrolled hypertension	vascular disease
cardiovascular disease	complicated hypertension	
hyperhidrosis	microvascular angina	

### 3. IR performance.

a. Larger retrieval. Castle et al. only report the highest ranking phenotype, so we do not know how many phenotypes their method retrieved or what fraction of them were of a cardiovascular nature. Our first simulation retrieved 12 target biological correlates, none of which are uniquely cardiovascular (Table 18). Our second simulation retrieved 74 drug biological correlates (Table 22). Of the latter, 34 are uniquely cardiovascular according to WHO-ATC (e.g., cardiovascular disease, heart disease, hypertension), 28 are not cardiovascular (e.g., antineoplastic, cocaine abuse, nephrolithiasis), and 12 are cardiovascular and other (e.g., adrenergic agent, neurotransmitter agent, smooth muscle relaxation). Thus our second simulation is more competitive with Castle et al.'s phenotype-based approach by this criterion.

For the disease axis, our multi-source, *indication*-based search produced 42 drug:disease correlates (Table 21), far more than the Castle et al.'s single cardiovascular [disease], and more even than the 20 semantically unique disease or bioprocess terms contained in the 163 WHO-

**Table 22. Drug biological correlates ("phenotype") for Research Use Case B.**

WHO-ATC body system classification: black: cardiovascular; red: not cardiovascular; blue: multi-system including cardiovascular. "Kidney failure" maps to "C03 diuretics" while "nephrolithiasis" and lower urinary tract concepts map to "G04 urologicals."

acute urinary retention	heart failure	bladder smooth muscle
cocaine dependence	resistant hypertension	relaxation
overactive bladder	anti-benign prostatic	microvascular angina
adrenergic agent	hyperplasia agent	urinary retention
complicated hypertension	hypercholesterolemia	bladder sphincter tone
peripheral adrenergic	smooth muscle relaxation	decrease
antagonist	antidepressant induced	microvascular angina pectoris
adrenergic antagonist	excessive sweating	urological agent
coronary heart disease	hyperhidrosis	blood pressure decrease
peripheral vasodilation	stroke	morning surge
alpha adrenergic antagonist	antihypercholesterolemic	vascular disease
diabetes mellitus	agent	cardiovascular disease
platelet aggregation inhibitor	hypertension	myocardial infarction
alpha blocker	systemic vascular resistance	vascular smooth muscle
diabetic nephropathy	decrease	inhibition
postsynaptic adrenergic	antihypertensive	cardiovascular disorder
inhibition	hypertension - mild	myocardial ischemia
alpha 1 adrenergic antagonist	uncontrolled hypertension	vasodilation
erectile dysfunction	antineoplastic	catecholamine
prostatic disorder	hypotensive	vasoconstrictor inhibition
alpha 1 adrenergic contraction	untreated hypertension	nephrolithiasis
antagonist	arterial vasodilation	vasodilator
essential hypertension	kidney failure	catheter ablation
prostatic hyperplasia	ureterolithiasis	neurotransmitter agent
alpha 1 adrenergic pressor	atrial fibrillation	venous vasodilation
antagonist	lower urinary tract symptoms	cocaine abuse
heart disease	urethral resistance decrease	nocturia
prostatic hypertrophy	benign prostatic hyperplasia	
alpha 1A adrenergic	lowers serum cholesterol	
antagonist	urinary obstruction	

ATC cardiovascular section classes (Table 23).<sup>57</sup> If all results are considered, the first simulation generated  $12 \times 42 = 504$  bioprocess:disease hypotheses and the second generated  $32 \times 42 = 1,344$  phenotype-like:disease hypotheses.

<sup>57</sup> The remaining 143 WHO-ATC "C" classes are of a chemical nature (e.g., "C01AA - digitalis glycosides), arbitrary subclasses and combinations (e.g., "C02L - antihypertensives and diuretics in combination"), or redundant (e.g., "C02K - other antihypertensives" given "C02 - antihypertensives").

**Table 23. WHO-ATC cardiovascular classes with disease/bioprocess equivalents.**

The remaining 143 WHO-ATC "C" classes are of a chemical nature (e.g., "C01AA - digitalis glycosides), arbitrary subclasses and combinations (e.g., "C02L - antihypertensives and diuretics in combination"), or redundant (e.g., "C02K - other antihypertensives" given "C02 - antihypertensives").

<i>WHO-ATC class</i>	<i>equivalent disease/bioprocess</i>
C - cardiovascular system	cardiovascular disease
C01 - cardiac therapy	heart disease
C01B - antiarrhythmics, class I and III	arrhythmia
C01D - vasodilators used in cardiac diseases	vasoconstriction
C02 - antihypertensives	hypertension
C02A - antiadrenergic agents, centrally acting	central adrenergic activity
C02B - antiadrenergic agents, ganglion-blocking	ganglionic adrenergic activity
C02C - antiadrenergic agents, peripherally acting	peripheral adrenergic activity
C02D - arteriolar smooth muscle, agents acting on	arteriolar smooth muscle activity
C03 - diuretics	diuresis
C04 - peripheral vasodilators	peripheral vasoconstriction
C05 - vasoprotectives	vascular disease
C05A - antihemorrhoidals for topical use	hemorrhoids
C05B - antivaricose therapy	varicose veins
C05C - capillary stabilizing agents	capillary disease
C07 - beta blocking agents	beta adrenergic activity [vasodilation]
C08 - calcium channel blockers	calcium channel activity
C09 - agents acting on the renin-angiotensin system	renin-angiotensin activity
C10 - serum lipid reducing agents	hyperlipidemia
C10A - cholesterol and triglyceride reducers	hypercholesterolemia / hypertriglyceridemia

**b. More robust.** In our first simulation, the 12 target correlates, like Castle et al's phenotypes, all came from a single source, DrugBank in our case, the Mammalian Phenotype Ontology in theirs. In our second simulation, the 74 drug biological correlates came from 13 of our sources (all but Drugs@FDA and RXNORM), in contrast to Castle et al's one (the Mammalian Phenotype Ontology). In both simulations, our 42 drug:disease correlates came from seven sources (ClinicalTrials.gov, DailyMed, DrugBank, DrugDigest, DrugInfo, MedMaster, and UMLS), in contrast to Castle et al's one (WHO-ATC). Not only was our contributing resource collection larger and more diverse, the relevant data are true *indication* values and thus represent a much richer lexicon that is semantically closer to "diseases" than the few WHO-ATC *therapeutic class* values that can be so mapped (e.g., Table 23)

c. More efficient. Data reduction exhibited by the Table 21 drug correlate ("disease") retrieval (207 rows) was: drug name (B/D) 4/11 (36%); dimension name (O/F) 6/9 (67%); value (Q/H) 42/50 (84%); drug-dimension-value triple 83/65 (128%). For the Table 22 drug correlate ("phenotype") retrieval (486 rows) it was: drug name (B/D) 4/20 (20%); dimension name (O/F) 9/26 (35%); value (Q/H) 74/86 (86%); drug-dimension-value triple 190/139 (137%). That is, in the case of the triples, the normalization effect was swamped out by the antagonistic multi-value parsing effect discussed previously. (The value and triple figures do not include the clinical trial IDs.) The Table 21 retrieval is the *indication* subset of Table 22's; the difference in the drug (B/D) and dimension (O/F) figures reflects the additional non-*indication* dimensions, sources, and source diversity that contributed to Table 22, producing a greater data reduction effect of normalization. The number of databases was reduced from seven to one (14%) and 13 to one (8%), respectively, implying an even greater reduction in the number of commands, queries, keystrokes, and time.

#### *Summary of Research Use Case B*

The goal here was to combine known drug-disease, drug-target, and target-phenotype relations to produce a set of hypothetical phenotype-disease relations. The required queries could be efficiently translated to Excel operations to retrieve the relevant database rows based on the normalized generic name, dimension, and value. In our first simulation, our system's answer was the 504 disease-phenotype relations between Table 18 and Table 21. This result was unsatisfactory because the Table 18 entities are of a semantic type that might more accurately be called bioprocesses rather than phenotypes, and because they are not specific to the cardiovascular system. Therefore we performed a second simulation where targets were removed from the logic and drug-phenotype relations were simulated with drug biological correlates, producing the 1,344 phenotype-disease connections between Table 22 and Table 21. Our system satisfied the criteria for usefulness as follows: more dimensions and resources than the reference system; larger retrieval and more resources with normalized than raw value search; and data

reduction. Fidelity to the reference's information need was compromised by having to substitute other semantic types for phenotypes. On the other hand, we provide the genuine *indication* data for which the reference system used WHO-ATC as a weak substitute.

Research Use Case C. Campillos et al. (2008) extracted specific sets of drugs with common side effects but different WHO-ATC therapeutic classes, and used the drugs' molecular target and chemical structure/similarity values to predict previously unknown shared targets, which were tested by *in vitro* and cell assays. The validated shared targets predict novel hypothetical indications and therapeutic classes for existing drugs. For example, a set of nervous system drugs was found to have side effects in common with the antiulcer drug rabeprazole. Four of their targets were predicted to bind rabeprazole, and two - the dopamine receptor DRD3 and the serotonin receptor HTR1D - were validated. This suggests that rabeprazole may be therapeutic for the indications of zolmitriptan (migraine), pergolide (Parkinson's disease), and paroxetine and fluoxetine (psychiatric disorders<sup>58</sup>).

Our database was able to support the general query for drugs with common side effects but different WHO-ATC therapeutic classes as described in Appendix F.<sup>59</sup> The result was all nine parent drugs in our database;<sup>60</sup> that is, our entire drug sample constitutes such a set of drugs.<sup>61</sup> Using  $\{(O) = \text{chemistry} - \text{chemical superclass}\}$  for chemical structure/similarity as in Research Use Case A, we identified two drug clusters:  $\{(B) = [\text{finasteride, dutasteride}]\}$  and  $\{(B)$

---

<sup>58</sup> fluoxetine: depression, obsessive-compulsive disorder, some eating disorders, panic attacks, premenstrual dysphoric disorder; paroxetine: depression, panic disorder, social anxiety disorder, obsessive-compulsive disorder, generalized anxiety disorder, posttraumatic stress disorder, premenstrual dysphoric disorder. Source: MedMaster.

<sup>59</sup> [http://comminfo.rutgers.edu/~msharp/XKB/Research\\_usecaseC\\_SE\\_TC.xls](http://comminfo.rutgers.edu/~msharp/XKB/Research_usecaseC_SE_TC.xls)

<sup>60</sup> All salts were lumped with their parents due to the initial WHO-ATC class cut, hence in this paragraph "terazosin" means "terazosin and terazosin hydrochloride"; etc.

<sup>61</sup> Since our entire drug sample was retrieved by the initial common-side-effect-different-therapeutic-class query, given that our sample was based on the drugs' common indication (benign prostatic hyperplasia), one wonders if that more straightforward dimension (*indication*) could be effectively substituted for the initial drug set selection in Campillos et al.'s method, given a robust database of normalized drug-indication relations. If so, Campillos et al.'s method constitutes evidence supporting our assertion of the latter's practical nonexistence.

= [prazosin..., terazosin..., doxazosin...]}.<sup>62</sup> Following Campillos et al.'s logic, the non-overlapping {(O) = *biology - molecular target*; (Q) = ...} values for the drugs within each cluster constitute hypothetical cross-targets; i.e., the targets of one drug not common to a second drug in the same cluster predict previously unknown targets of the second drug. These hypothetical targets were: {(B) = dutasteride; (O) = *biology - molecular target*; (Q) = [5-beta reductase, androgen receptor]} and {(B) = terazosin...; (O) = *biology - molecular target*; (Q) = alpha1C adrenergic receptor}.

These targets were then mapped to their existing known drugs' {(O) = [*clinical - indications ...*, *clinical - therapeutic classes ...*]} and the non-overlapping (Q) values for dutasteride and terazosin identified. In this way we generated for dutasteride 14 hypothetical new indications and 7 hypothetical new therapeutic classes (Table 24), and for terazosin 73 hypothetical new indications (Table 25) and 14 hypothetical new therapeutic classes (Table 26). Narrowing the resource collection to more closely simulate Campillos et al.'s system (i.e., eliminating the ClinicalTrials.gov, ChemIDplus, and KEGG DRUG results) gave two indications and five therapeutic classes for dutasteride, and seven indications and five therapeutic classes for terazosin.

In addition, we tried substituting the targets' biological correlates (as defined under Research Use Case A) for Campillos et al.'s target:drug relations for deriving target:indication and target:therapeutic class links. Our small database sample size prevented this approach from adding any new results to those given in the preceding paragraph. Only the androgen receptor was identified as a hypothetical new target (for dutasteride), and the only drug its biological correlates points to in our database is finasteride. In addition, these biological correlates (Table 27) are perhaps too general to be useful for this purpose. The exception was "HINK3 activation" which a cursory Google search linked to Down syndrome.<sup>63</sup>

---

<sup>62</sup> [http://comminfo.rutgers.edu/~msharp/XKB/Research\\_usecaseC\\_MT\\_CS.xls](http://comminfo.rutgers.edu/~msharp/XKB/Research_usecaseC_MT_CS.xls)

<sup>63</sup> <https://www.wikigenes.org/e/gene/e/10114.html>

**Table 24. Hypothetical new indications and therapeutic classes for dutasteride (Research Use Case C).**

Red: resources not used by Campillos et al. (2008).

<i>source</i>	<i>dimension</i>	<i>value</i>
UMLS	<i>clinical - indication - treatment</i>	hirsutism
UMLS	<i>clinical - indication - treatment</i>	prostatic neoplasm
ClinicalTrials.gov	<i>clinical - indication - treatment - clinical trial condition</i>	chronic central serous chorioretinopathy
ClinicalTrials.gov	<i>clinical - indication - treatment - clinical trial condition</i>	idiopathic hirsutism
ClinicalTrials.gov	<i>clinical - indication - clinical trial condition</i>	hematospermia
ClinicalTrials.gov	<i>clinical - indication - clinical trial condition</i>	hematuria
ClinicalTrials.gov	<i>clinical - indication - clinical trial condition</i>	infertility
ClinicalTrials.gov	<i>clinical - indication - clinical trial condition</i>	muscle atrophy
ClinicalTrials.gov	<i>clinical - indication - clinical trial condition</i>	prostatic disorder
ClinicalTrials.gov	<i>clinical - indication - clinical trial condition</i>	retinal disease
ClinicalTrials.gov	<i>clinical - indication - clinical trial condition</i>	sarcopenia
ClinicalTrials.gov	<i>clinical - indication - clinical trial condition</i>	sexual dysfunction
ClinicalTrials.gov	<i>clinical - indication - clinical trial condition</i>	transurethral resection of prostate
ChemIDplus	<i>clinical - therapeutic class</i>	alpha reductase inhibitor
UMLS	<i>clinical - therapeutic class</i>	androgen antagonist - synthetic
UMLS	<i>clinical - therapeutic class</i>	antineoplastic
WHO-ATC	<i>clinical - therapeutic class - WHO-ATC 5th level code</i>	D11AX10 [D11AX other dermatologicals]
DrugBank [WHO-ATC]	<i>clinical - therapeutic class - body system</i>	dermatological agent
ChemIDplus	<i>clinical - therapeutic class - body system</i>	reproductive agent
DrugBank	<i>clinical - therapeutic class - body system</i>	skin and mucous membrane agent

**Table 25. Hypothetical new indications for terazosin (Research Use Case C).**

MeSH contributions also appear in other NLM sources. In the lower table, the source is ClinicalTrials.gov for all data.

<i>source</i>	<i>dimension</i>	<i>value</i>
DrugBank	<i>clinical - indication - treatment</i>	chronic heart failure
MedMaster	<i>clinical - indication - treatment</i>	congestive heart failure
MeSH	<i>clinical - indication - treatment</i>	heart failure
MedMaster	<i>clinical - indication - treatment</i>	pheochromocytoma
MedMaster	<i>clinical - indication - treatment</i>	Raynaud disease
MeSH	<i>clinical - indication - treatment</i>	Raynaud syndrome
DrugBank	<i>clinical - indication - treatment</i>	urinary obstruction

<i>dimension</i>	<i>value</i>
<i>clinical - indication - treatment - clinical trial condition</i>	Alzheimer disease-related agitation
<i>clinical - indication - treatment - clinical trial condition</i>	cardiovascular disorder
<i>clinical - indication - treatment - clinical trial condition</i>	complicated hypertension
<i>clinical - indication - treatment - clinical trial condition</i>	female voiding dysfunction
<i>clinical - indication - treatment - clinical trial condition</i>	post-traumatic stress disorder - noncombat trauma
<i>clinical - indication - treatment - clinical trial condition</i>	supine hypertension in autonomic failure
<i>clinical - indication - treatment - clinical trial condition</i>	untreated hypertension
<i>clinical - indication - clinical trial condition</i>	alcohol craving
<i>clinical - indication - clinical trial condition</i>	alcohol dependence
<i>clinical - indication - clinical trial condition</i>	alcoholism
<i>clinical - indication - clinical trial condition</i>	allergic rhinitis
<i>clinical - indication - clinical trial condition</i>	Alzheimer disease
<i>clinical - indication - clinical trial condition</i>	anxiety disorder
<i>clinical - indication - clinical trial condition</i>	autonomic dysreflexia
<i>clinical - indication - clinical trial condition</i>	cardiovascular disease
<i>clinical - indication - clinical trial condition</i>	chronic hepatitis C
<i>clinical - indication - clinical trial condition</i>	cirrhosis
<i>clinical - indication - clinical trial condition</i>	cirrhosis complications
<i>clinical - indication - clinical trial condition</i>	cocaine abuse
<i>clinical - indication - clinical trial condition</i>	cocaine craving
<i>clinical - indication - clinical trial condition</i>	cocaine dependence
<i>clinical - indication - clinical trial condition</i>	combat disorder
<i>clinical - indication - clinical trial condition</i>	combat stress symptoms
<i>clinical - indication - clinical trial condition</i>	coronary heart disease
<i>clinical - indication - clinical trial condition</i>	diabetes mellitus
<i>clinical - indication - clinical trial condition</i>	diabetic nephropathy
<i>clinical - indication - clinical trial condition</i>	erectile dysfunction
<i>clinical - indication - clinical trial condition</i>	essential hypertension
<i>clinical - indication - clinical trial condition</i>	falling
<i>clinical - indication - clinical trial condition</i>	fibrosis
<i>clinical - indication - clinical trial condition</i>	gastrointestinal hemorrhage
<i>clinical - indication - clinical trial condition</i>	heart disease
<i>clinical - indication - clinical trial condition</i>	heart failure [clinical trial]
<i>clinical - indication - clinical trial condition</i>	hepatitis C - chronic

**Table 25. Hypothetical new indications for terazosin (Research Use Case C) (continued).**

<i>dimension</i>	<i>value</i>
<i>clinical - indication - clinical trial condition</i>	hypercholesterolemia
<i>clinical - indication - clinical trial condition</i>	insomnia
<i>clinical - indication - clinical trial condition</i>	kidney failure
<i>clinical - indication - clinical trial condition</i>	microvascular angina
<i>clinical - indication - clinical trial condition</i>	microvascular angina pectoris
<i>clinical - indication - clinical trial condition</i>	mood disorder
<i>clinical - indication - clinical trial condition</i>	morning surge
<i>clinical - indication - clinical trial condition</i>	myocardial infarction
<i>clinical - indication - clinical trial condition</i>	myocardial ischemia
<i>clinical - indication - clinical trial condition</i>	nephrolithiasis
<i>clinical - indication - clinical trial condition</i>	neurogenic bladder
<i>clinical - indication - clinical trial condition</i>	nightmare
<i>clinical - indication - clinical trial condition</i>	orthostatic hypotension
<i>clinical - indication - clinical trial condition</i>	overactive bladder
<i>clinical - indication - clinical trial condition</i>	portal hypertension
<i>clinical - indication - clinical trial condition</i>	post-traumatic stress disorder
<i>clinical - indication - clinical trial condition</i>	post-traumatic stress disorder - civilian
<i>clinical - indication - clinical trial condition</i>	post-traumatic stress disorder - combat trauma
<i>clinical - indication - clinical trial condition</i>	prostatic disorder
<i>clinical - indication - clinical trial condition</i>	psychomotor agitation
<i>clinical - indication - clinical trial condition</i>	resistant hypertension
<i>clinical - indication - clinical trial condition</i>	rhinitis medicamentosa
<i>clinical - indication - clinical trial condition</i>	scorpion envenomation
<i>clinical - indication - clinical trial condition</i>	sleep disorder
<i>clinical - indication - clinical trial condition</i>	spinal cord injury
<i>clinical - indication - clinical trial condition</i>	stress-induced cocaine/alcohol craving and relapse
<i>clinical - indication - clinical trial condition</i>	stroke
<i>clinical - indication - clinical trial condition</i>	tachyphylaxis
<i>clinical - indication - clinical trial condition</i>	uncontrolled hypertension
<i>clinical - indication - clinical trial condition</i>	ureterolithiasis
<i>clinical - indication - clinical trial condition</i>	variceal bleeding
<i>clinical - indication - clinical trial condition</i>	vascular disease

**Table 26. Hypothetical new therapeutic classes for terazosin (Research Use Case C).**

Red: resources not used by Campillos et al. (2008).

<i>Source</i>	<i>dimension</i>	<i>value</i>
KEGG DRUG	<i>clinical - therapeutic class</i>	alpha1A adrenergic antagonist
KEGG DRUG	<i>clinical - therapeutic class</i>	alpha1B adrenergic antagonist
KEGG DRUG	<i>clinical - therapeutic class</i>	alpha1C adrenergic antagonist
KEGG DRUG	<i>clinical - therapeutic class</i>	alpha1D adrenergic antagonist
KEGG DRUG	<i>clinical - therapeutic class</i>	alpha2A adrenergic antagonist
KEGG DRUG	<i>clinical - therapeutic class</i>	alpha2B adrenergic antagonist
KEGG DRUG	<i>clinical - therapeutic class</i>	G protein coupled receptor ligand
DrugBank	<i>clinical - therapeutic class</i>	antihypercholesterolemic agent
UMLS	<i>clinical - therapeutic class</i>	hypotensive agent
WHO-ATC	<i>clinical - therapeutic class</i>	peripheral adrenergic antagonist
KEGG DRUG	<i>clinical - therapeutic class</i>	rhodopsin family amine receptor ligand
UMLS	<i>clinical - therapeutic class</i>	vasodilator
ChemIDplus	<i>clinical - therapeutic class - organism</i>	human
WHO-ATC	<i>clinical - therapeutic class - WHO-ATC 5th level code</i>	C02CA01 [C02CA - alpha-adrenoreceptor antagonists]

**Table 27. Hypothetical new target biological correlates for dutasteride (Research Use Case C).**

<i>dimension</i>	<i>value</i>
<i>biology - molecular target - general function</i>	DNA binding
<i>biology - molecular target - GO biological process</i>	regulation of biological process
<i>biology - molecular target - GO biological process</i>	regulation of cellular metabolism
<i>biology - molecular target - GO biological process</i>	regulation of metabolism
<i>biology - molecular target - GO biological process</i>	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolism
<i>biology - molecular target - GO biological process</i>	regulation of physiological process
<i>biology - molecular target - GO biological process</i>	regulation of transcription
<i>biology - molecular target - GO biological process</i>	regulation of transcription, DNA-dependent
<i>biology - molecular target - specific function</i>	cellular differentiation
<i>biology - molecular target - specific function</i>	cellular proliferation
<i>biology - molecular target - specific function</i>	eukaryotic gene expression
<i>biology - molecular target - specific function</i>	HIPK3 activation
<i>biology - molecular target - specific function</i>	steroid hormone receptor activity

*Criteria for usefulness*

1. Comprehensive coverage. Same as Research Use Case A.
2. Literary warrant fidelity. Our database was able to support Campillos et al.'s general query for drugs with common side effects but different WHO-ATC therapeutic classes, and our

methodology for populating these two dimensions was very similar to that of Campillos et al. Our *chemistry - chemical superclass* dimension is a weak substitute for their chemical structure/similarity measure, but this is not due to any fault in database design, as discussed under Research Use Case A. Furthermore, this simulation did produce two drug clusters which had enough relational data to predict three previously unknown shared targets, 87 novel hypothetical indications, and 21 novel hypothetical therapeutic classes for two existing drugs by following Campillos et al.'s logic. When the resources were restricted to those more like the ones used by Campillos et al., our simulation still produced three novel hypothetical targets, nine novel hypothetical indications, and ten novel hypothetical therapeutic classes, roughly equivalent to the 3-10 hypothetical new indications for rabeprazole that can be inferred from the four drugs Campillos et al. found to share targets with it. Of course, we are missing the *in vitro* and cell assay target validation step.

### 3. IR performance.

a. Larger retrieval. Of the nine parent compounds in our database, we generated for two of them: three novel hypothetical targets, 87 novel hypothetical indications, and 21 novel hypothetical therapeutic classes. Extrapolated to 5,000 known generic parent compounds, our results are equivalent to 7,500 novel hypothetical targets, 217,500 novel hypothetical indications, and 52,500 novel hypothetical therapeutic classes. If the ClinicalTrials.gov, ChemIDplus, and KEGG DRUG results are eliminated to more closely simulate Campillos et al.'s system, the numbers fall to nine (10%) indications and ten (48%) therapeutic classes for two drugs, roughly equivalent to the 3-10 hypothetical new indications for rabeprazole that can be inferred from the four drugs Campillos et al. found to share targets with it.

b. More robust. Our *clinical - precaution - side effect ...* dimension is populated with values from five sources (DailyMed, DrugDigest, MedMaster, UMLS, and ClinicalTrials.gov) compared to Campillos et al.'s one (DailyMed or equivalent). Our database's 1,255 {(O) = *clinical - precaution - side effect ...*} rows contain 431 unique {B,Q} parent

drug:normalized value pairs; DailyMed accounts for 52% of the rows and 31% of the unique {B,Q} pairs,<sup>64</sup> DrugDigest 30% and 41%, MedMaster 16% and 25%, UMLS 1% and 3%, and ClinicalTrials.gov 0.3% and 1%, giving credit in that order. In addition, we have about half of the side effects subclassified by whether they are *common*, *major*, or *minor*, which might help narrow the initial drug clusters to those which are most productive, given a larger drug sample.

Our simulation of this use case, like Campillos et al., used only WHO-ATC for the therapeutic class cut. However, we have additional therapeutic class data from ten other sources (ChEBI, ChemIDplus, DailyMed, DrugBank, DrugDigest, DrugInfo, KEGG DRUG, MedMaster, MeSH, PubChem, UMLS) which might help narrow (or widen) the initial drug clusters to those which are most productive, given a larger drug sample. Our database's 728 {(O) = *clinical - therapeutic class* ...} rows contain 168 unique {B,Q} parent drug:normalized value pairs; WHO-ATC accounts for 6% of the rows and 21% of the unique {B,Q} pairs.

For drug:target relations, Campillos et al. used the Matador<sup>65</sup> and PDSP Ki (Psychoactive Drug Screening Program inhibition constant)<sup>66</sup> databases in addition to DrugBank, so their hypothetical shared target step should be more stringent than ours. In addition, they further refined their shared target hypotheses to those that were supported by wet bench *in vitro* and cell assay results.

However, Campillos et al.'s empirical results stop at the shared target point. To complete this use case inferred from their introduction and discussion, one needs to map the drugs which share the target to their indications and therapeutic classes. For the latter they had WHO-ATC, but do not report using it for this purpose in the manner of Castle et al. (2007). Had they done so, our database would show a large robustness advantage not only for therapeutic classes *per se* (ten additional resources), but also for genuine indications as discussed under Research Use Case B.

---

<sup>64</sup> DailyMed's share of the unique drug:value pairs is underestimated because we did not parse and normalize all the raw data, so column Q is blank for those rows.

<sup>65</sup> <http://matador.embl.de>

<sup>66</sup> <http://pdsp.med.unc.edu/kidb.php>

In fairness to Campillos et al. and Castle et al., drug development research does not typically leap straight from hypotheses to direct testing of drug:disease efficacy, even in animal or *in vitro* models, but rather proceeds through the kind of molecular-level validation phase their systems support more robustly. However, mapping novel drug:target relations to disease endpoints is important from the perspective of program management, funding, public relations, etc.

c. More efficient. Data reduction exhibited by the initial  $\{(O) = [\textit{clinical} - \textit{precaution} - \textit{side effect} \dots, \textit{clinical} - \textit{therapeutic class} - \textit{WHO-ATC 5th level code}]\}$  retrieval (1,277 rows) was: drug name (B/D) 9/70 (13%); dimension name (O/F) 5/17 (29%); value (Q/H) 258/200 (129%); drug-dimension-value triple 524/467 (112%). It is possible that, in the case of the values and triples, the normalization effect was swamped out by the antagonistic multi-value parsing effect discussed previously. However, these numbers are distorted by the fact that not all the relevant DailyMed raw (H) values were loaded into the database. For the  $\{(O) = [\textit{biology} - \textit{molecular target}, \textit{chemistry} - \textit{chemical superclass}]\}$  retrieval (240 rows) the data reduction was: drug name (B/D) 8/58 (14%); dimension name (O/F) 2/29 (7%); value (Q/H) 58/127 (46%); drug-dimension-value triple 80/203 (39%). For a general  $\{(O) = [\textit{clinical} - \textit{indication} \dots, \textit{clinical} - \textit{therapeutic class} \dots]\}$  with salt terms (acetate, mesylate, hydrochloride) lumped with their parents in column B, the retrieval (2,272 rows) was: drug name (B/D) 9/96 (9%); dimension name (O/F) 26/55 (47%); value (Q/H) 380/799 (48%); drug-dimension-value triple 655/1340 (49%). (The value and triple figures do not include the clinical trial IDs.) For these three retrievals, the number of databases was reduced from six to one (17%), nine to one (11%), and 13 to one (8%), implying an even greater reduction in the number of commands, queries, keystrokes, and time.

It seems that the novel hypothetical indications generated by our method would be more useful than the novel hypothetical therapeutic classes, the latter tending to be too general (e.g., "dermatological agent") or inferable from known classes (e.g., "peripheral adrenergic antagonist"). On the other hand, most of the novel hypothetical indications came from ClinicalTrials.gov *Conditions* which means that (1) they include false positives (co-occurrences

of a drug and condition in a trial other than for treatment or prevention) and (2) even the true positives are "doubly hypothetical" in the sense that the target-sharing drug's efficacy has not been proven. Removing the therapeutic classes and ClinicalTrials.gov data resulted in smaller retrieval size more like that of Campillos et al.: two novel hypothetical indications for dutasteride (from UMLS) and seven for terazosin (from DrugBank, MedMaster, and MeSH).

#### *Summary of Research Use Case C*

The goal here was to combine drug-side effect, drug-therapeutic class, drug-target, and drug-chemical structure/similarity relations to predict novel drug-target, drug-indication, and drug-therapeutic class relations. Our database was able to support the general query for drugs with common side effects but different therapeutic classes, yielding all nine parent drugs in our database. This surprising result seems to imply that *indication* could be effectively substituted for the reference system's complicated same-side-effect-different-therapeutic-class clustering approach, further implying that a robust, normalized drug-indication database was not available to the authors, as we have asserted. Follow-up queries to our database produced for 87 hypothetical new indications and 21 hypothetical new therapeutic classes for two drugs. Our system satisfied the criteria for usefulness as for Research Use Case A. The reference system had a more robust collection of drug:target relations, but ours allowed a more empirical extension from drug:target relations to disease endpoints. It seems that the novel hypothetical indications generated by our method would be more useful than the novel hypothetical therapeutic classes. However, the novel hypothetical indications include false positives and "doubly hypothetical" values derived from ClinicalTrials.gov. Removing the therapeutic classes and ClinicalTrials.gov data from our results made the reference system compare better.

Research Use Case D. A researcher wonders if any existing drugs might be "repurposed" (Boguski et al., 2009) to prevent prostate cancer. She searches ClinicalTrials.gov and gets a list of clinical trials which link the *Condition* "Prostate Cancer" to various *Interventions* including

drug names. She thinks this is a good start, but what she really needs is to find other, chemically related drugs and chemicals which are *not* on this list or already approved for prevention of prostate cancer.

We retrieved data on 1,823 clinical trials on prostate cancer and extracted 723 raw drug names as described in Appendix F. Comparison of this list to the (D) raw drug names in our database yielded four normalized (B) generic parent names: dutasteride, finasteride; leuprolide, and tamsulosin. The chemical characteristics of these drugs were retrieved as the values (Q) corresponding to {(B) = [dutasteride, finasteride; leuprolide..., tamsulosin...]; (O) = [*chemistry - chemical complexity, chemistry - chemical superclass, chemistry - heavy atom count, chemistry - Lipinski ... , chemistry - physical properties - melting point, chemistry - polarity - TPSA, chemistry - rotatable bond count, chemistry - solubility ... , chemistry - stereocenter count ... , chemistry - tautomer count*]}.<sup>67</sup> The most parsimonious resource collection (C) that supplied this data was DrugBank, PubChem, and UMLS.

Of the four drugs, tamsulosin had the most typical values for these dimensions across all drugs in our database, so, hoping to find other drugs with similar values, we chose to make it our model prostate cancer drug.<sup>68</sup> We retrieved drugs (B) with similar values (Q) to those of tamsulosin for the above *chemistry ...* dimensions (O). Surprisingly, seven out of the nine parent drug compounds in our database qualified; in order of number of closest values to tamsulosin's, finasteride (8), prazosin (7), terazosin (6), doxazosin (4), dutasteride (4), leuprolide (1), and ticlopidine (1) (Table 28).<sup>69</sup> The difference between the four drugs in clinical trials and these seven (plus tamsulosin) constitutes our retrieval of tamsulosin-like compounds not currently in clinical trials on prostate cancer: prazosin, terazosin, doxazosin, and ticlopidine.

<sup>67</sup> [http://comminfo.rutgers.edu/~msharp/XKB/Research\\_usecaseD\\_chem\\_tamsu.xls](http://comminfo.rutgers.edu/~msharp/XKB/Research_usecaseD_chem_tamsu.xls) Sheet1

<sup>68</sup> [http://comminfo.rutgers.edu/~msharp/XKB/Research\\_usecaseD\\_chem\\_tamsu.xls](http://comminfo.rutgers.edu/~msharp/XKB/Research_usecaseD_chem_tamsu.xls) Sheet2

<sup>69</sup> [http://comminfo.rutgers.edu/~msharp/XKB/Research\\_usecaseD\\_chem\\_tamsu.xls](http://comminfo.rutgers.edu/~msharp/XKB/Research_usecaseD_chem_tamsu.xls) Sheet3 column G

**Table 28. Chemical characteristics of tamsulosin-like compounds in our database (Research Use Case D).**

Data are from DrugBank (rows 1-6), MeSH/UMLS (7), and PubChem (8-19). (DB) across whole database.

<i>dimension</i>	<i>parent compound</i>				<i>closest to tamsulosin (DB)</i>
	<i>tamsulosin</i>	<i>finasteride</i>	<i>prazosin</i>	<i>terazosin</i>	
<i>chemistry - physical properties - melting point [°C]</i>	227	250	279	273	242-250°C dutasteride; 250°C finasteride
<i>chemistry - solubility - logP - predicted</i>	3.06	3.53	1.93	1.12	2.53 doxazosin; 3.53 finasteride
<i>chemistry - solubility - logP hydrophobicity - experimental</i>	2.3	4.7	1.3	1	2.1 doxazosin
<i>chemistry - solubility - logS - predicted</i>	-4.79	-5.27	-2.74	-2.41	-4.57 leuprolide
<i>chemistry - solubility - water - experimental [mg/ml]</i>	sparingly soluble in water	0.012	0.5	0.024 <sup>a</sup>	
<i>chemistry - solubility - water - predicted [mg/ml]</i>	0.0066	0.022	0.69	1.5	.022 finasteride
<i>chemistry - chemical superclass</i>	sulfonamide	androstene; azasteroid	piperazine; quinazoline	piperazine; quinazoline	
<i>chemistry - chemical complexity</i>	539	678	544	544	544 prazosin; 544 terazosin
<i>chemistry - heavy atom count</i>	28	27	28	28	27 finasteride; 28 prazosin; 28 terazosin
<i>chemistry - Lipinski - H bond acceptor</i>	7	2	8	8	8 dutasteride; 8 prazosin; 8 terazosin
<i>chemistry - Lipinski - H bond donor</i>	2	2	1	1	2 doxazosin; 2 dutasteride; 2 finasteride
<i>chemistry - Lipinski - solubility logP octanol-water</i>	2.7	3	2	1.4	2.5 doxazosin; 3 finasteride
<i>chemistry - Lipinski - molecular weight [-average]</i>	408	372	383	387	383 prazosin; 387 terazosin
<i>chemistry - polarity - TPSA</i>	100	58	107	103	103 terazosin; 107 prazosin
<i>chemistry - rotatable bond count</i>	11	2	4	4	2 dutasteride; 2 finasteride; 4 prazosin; 4 terazosin

<sup>a</sup> DrugBank's value of 29.7 is clearly for the hydrochloride, so it has been replaced here with this value from the SRC PhysProp Database <http://esc.syrres.com/interkow/webprop.exe?CAS=63590-64-7>

**Table 28. Chemical characteristics of tamsulosin-like compounds in our database (Research Use Case D) (continued).**

<i>dimension</i>	<i>parent compound</i>				<i>closest to tamsulosin (DB)</i>
	<i>tamsulosin</i>	<i>finasteride</i>	<i>prazosin</i>	<i>terazosin</i>	
<i>chemistry - stereocenter count - defined atom</i>	1	7	0	0	
<i>chemistry - stereocenter count - defined bond</i>	0	0	0	0	0 all
<i>chemistry - stereocenter count - undefined atom</i>	0	0	0	1	0 finasteride; 0 prazosin; 0 ticlopidine
<i>chemistry - stereocenter count - undefined bond</i>	0	0	0	0	0 all

We attempted to compare these results to the "similar compound" searches available on PubChem, ChemIDplus, DrugBank, and KEGG DRUG. Surprisingly, they all retrieved completely different sets of top hits for tamsulosin. We wished to compare compounds identified by conventional generic names for equivalency to prazosin, terazosin, doxazosin, and ticlopidine. From DrugBank we selected three of the top four hits: dofetilide, bumetanide, and piretanide. From KEGG DRUG we selected the top three: amosulalol, formoterol, and isoxsuprine. PubChem's and ChemIDplus' tools offered no obvious, easy way to filter the results down to such compounds and so were not used.

Using the original resources' (DrugBank, UMLS, and PubChem) web interfaces, we looked up the values for the above *chemistry* dimensions for dofetilide, bumetanide, piretanide, amosulalol, formoterol, and isoxsuprine (Table 29).<sup>70</sup> To estimate the nine other compounds' chemical similarity to tamsulosin, we devised a similarity measure based on the deviation of a given drug's values from the corresponding values for tamsulosin. For example, given the melting points of 227°C for tamsulosin and 250°C for finasteride, the melting point deviation of finasteride is  $|(227-250)/227| = 10\%$ . For each drug we averaged the deviations over three groups of dimensions: physical behavior (melting point and solubility), chemical complexity (including the Lipinski parameters, polarity, and rotatable bonds), and stereocenter counts. The latter is

<sup>70</sup> [http://comminfo.rutgers.edu/~msharp/XKB/Research\\_usecaseD\\_chem\\_tamsu.xls](http://comminfo.rutgers.edu/~msharp/XKB/Research_usecaseD_chem_tamsu.xls) Sheet3 columns H-M

actually another measure of chemical complexity but we wanted to segregate the low raw scores (typically 1 or 0) and consequent high deviations (1 vs. 0  $\rightarrow$  100%). Finally, we averaged the three averages for each drug to obtain an overall measure of its similarity to tamsulosin (Table 30).<sup>71</sup>

In Table 30, it can be seen that finasteride has the highest overall deviation (i.e., lowest similarity to tamsulosin), 72%, due primarily to its high number of defined atom stereocenters characteristic of steroids. Next highest is isoxsuprine (66%). The other seven are fairly evenly spread between 17% and 40%, with one of our database's retrievals, prazosin, at 28%. The same general pattern holds for the chemical complexity and stereocenter subset averages: finasteride and two of the KEGG DRUG candidates having the highest deviations, all DrugBank candidates having low deviations, and prazosin and terazosin in between. The physical subset average, however, shows KEGG DRUG's single candidate for which data was available to have a lower deviation than any of our database's three candidates. Nevertheless, it seems clear that prazosin and terazosin are competitive with DrugBank's and KEGG DRUG's top tamsulosin-similar compounds by these measures. Moreover, these measures were able to clearly discriminate prazosin and terazosin from the very un-tamsulosin-like steroid finasteride.

---

<sup>71</sup> [http://comminfo.rutgers.edu/~msharp/XKB/Research\\_usecaseD\\_chem\\_tamsu.xls](http://comminfo.rutgers.edu/~msharp/XKB/Research_usecaseD_chem_tamsu.xls) Sheet3 rows 33-60



**Table 30. Deviation of chemical characteristics from those of tamsulosin by similar compounds (Research Use Case D).**  
Dimensions and sources are the same as in Table 28 but dimensions have been abbreviated here for compaction.

<i>dimension</i>	<i>parent compound</i>									
	<i>tamsulosin</i>	<i>finasteride</i>	<i>prazosin</i>	<i>terazosin</i>	<i>dofetilide</i>	<i>bumetanide</i>	<i>piretanide</i>	<i>amosulalol</i>	<i>formoterol</i>	<i>isoxsuprine</i>
<i>melting point [°C]</i>	227	10%	23%	20%		1%				
<i>solubility - logP - predicted</i>	3.06	15%	37%	63%	29%	13%	28%		28%	
<i>solubility - logP - experimental</i>	2.3	104%	43%	57%	9%	13%			17%	
<i>solubility - logS - predicted</i>	-4.79	10%	43%	50%	9%	13%	25%		18%	
<b>avg</b>		<b>35%</b>	<b>37%</b>	<b>47%</b>	<b>16%</b>	<b>10%</b>	<b>26%</b>		<b>21%</b>	
<i>chemical complexity</i>	539	26%	1%	1%	25%	2%	5%	5%	28%	45%
<i>heavy atom</i>	28	4%	0%	0%	4%	11%	11%	7%	11%	21%
<i>Lipinski - H bond acceptor</i>	7	71%	14%	14%	14%	0%	0%	0%	29%	43%
<i>Lipinski - H bond donor</i>	2	0%	50%	50%	0%	50%	0%	50%	100%	50%
<i>Lipinski - solubility logP</i>	2.7	11%	26%	48%	33%	4%	19%	56%	33%	4%
<i>octanol-water</i>										
<i>Lipinski - molecular weight</i>	408	9%	6%	5%	8%	11%	11%	7%	16%	26%
<i>polarity - TPSA</i>	100	42%	7%	3%	22%	27%	18%	19%	9%	38%
<i>rotatable bond</i>	11	82%	64%	64%	0%	27%	55%	18%	27%	36%
<b>avg</b>		<b>31%</b>	<b>21%</b>	<b>23%</b>	<b>13%</b>	<b>16%</b>	<b>15%</b>	<b>20%</b>	<b>32%</b>	<b>33%</b>
<i>stereocenter - defined atom</i>	1	600%	100%	100%	100%	100%	100%	100%	100%	100%
<i>stereocenter - defined bond</i>	0	0%	0%	0%	0%	0%	0%	0%	0%	0%
<i>stereocenter - undefined atom</i>	0	0%	0%	100%	0%	0%	0%	100%	0%	300%
<i>stereocenter - undefined bond</i>	0	0%	0%	0%	0%	0%	0%	0%	0%	0%
<b>avg</b>		<b>150%</b>	<b>25%</b>	<b>50%</b>	<b>25%</b>	<b>25%</b>	<b>25%</b>	<b>50%</b>	<b>25%</b>	<b>100%</b>
<b>grand avg</b>		<b>72%</b>	<b>28%</b>	<b>40%</b>	<b>18%</b>	<b>17%</b>	<b>22%</b>	<b>35%</b>	<b>26%</b>	<b>66%</b>

### *Criteria for usefulness*

1. Comprehensive coverage. Same as Research Use Case A.

2. Literary warrant fidelity. Our database was able to support Boguski et al.'s (2009) general objective of "repurposing" existing drugs; specifically, for use against prostate cancer with selection based on their chemical similarity to tamsulosin. Chemical similarity is not one of the dimensions suggested by Boguski et al. (Table 20), but we wanted to explore our database's ability to support a use case involving it more intensively than the other Research Use Cases, as warranted by Castle et al. (2007) and Campillos et al. (2008).

3. IR performance.

a. Larger retrieval. This criterion is not apt for this use case due to our model database's small compound sample. Of the nine parent compounds in our database, we were able to rank seven for chemical similarity to an eighth, tamsulosin. Of the top three, one (finasteride) is known to have therapeutic potential against prostate cancer, and the other two (prazosin and terazosin) are competitive with the top three tamsulosin-similar-structure-search retrievals from DrugBank and KEGG DRUG according to our measure of chemical similarity.

b. More robust. The *chemistry* dimensions we used for this use case were covered most parsimoniously by DrugBank, PubChem, and UMLS. In addition, our database integrates the same and other potentially useful *chemistry* information from ChEBI, ChemIDplus, ClinicalTrials.gov, DailyMed, DrugDigest, DrugInfo, KEGG DRUG, and MeSH, as well as toxicology data under  $\{(O) = \textit{biology} - \textit{toxicity} \dots\}$ . The latter is comparable to the *side effect* innovation of Campillos et al. (2008) with the additional advantage that such data is available for nondrug chemicals. That is, toxicological characteristics of known drug clusters (say, drugs for a given indication) could be determined using a database with our design and scope, then used (along with their chemical characteristics) to search for toxicologically (and chemically) similar nondrug chemicals on ChemIDplus and other sources which enable such searches.

c. More efficient. Data reduction exhibited by the *chemistry* subset used to compute chemical similarity {(O) = [*chemistry - chemical complexity, chemistry - chemical superclass, chemistry - heavy atom count, chemistry - Lipinski ..., chemistry - physical properties - melting point, chemistry - polarity - TPSA, chemistry - rotatable bond count, chemistry - solubility ..., chemistry - stereocenter count ..., chemistry - tautomer count*]}<sup>72</sup> (523 rows) was: drug name (B/D) 14/69 (20%); dimension name (O/F) 28/35 (80%); value (Q/H) 191/200 (96%); drug-dimension-value triple 370/363 (102%). For the entire {(O) = *chemistry ...*} subset of our database (1,909 rows) it was: drug name (B/D) 15/83 (18%); dimension name (O/F) 93/121 (77%); value (Q/H) 763/884 (86%); drug-dimension-value triple 1245/1529 (81%). For these two retrievals, the number of databases was reduced from eight to one (12%) and 11 to one (9%), implying an even greater reduction in the number of commands, queries, keystrokes, and time. Given the closeness of these two data reduction result sets, the numbers for intermediate *chemistry* subsets<sup>73</sup> would be expected to be similar.

Interestingly, tamsulosin also shares with prazosin, terazosin, and doxazosin its molecular targets, the A, B, and D isoforms of the alpha1 adrenergic receptor. If tamsulosin were truly being tested in clinical trials for treatment or prevention of prostate cancer,<sup>74</sup> and its relevant mechanism of action were through the alpha1 adrenergic receptor, this would be evidence for the usefulness of these results.

#### *Summary of Research Use Case D*

The goal here was to find novel drug candidates for an indication (prostate cancer) based on their chemical similarity to drugs already approved or under study for that indication. Our

<sup>72</sup> [http://comminfo.rutgers.edu/~msharp/XKB/Research\\_usecaseD\\_chem\\_tamsu.xls](http://comminfo.rutgers.edu/~msharp/XKB/Research_usecaseD_chem_tamsu.xls) Sheet1

<sup>73</sup> The purpose of the dimension set used was to focus on descriptive ("natural") as opposed to nominal values. Therefore other database's ID's, nomenclature, and formulas were not used. It could be argued that the latter also are naturally descriptive, but the requisite drill down, parsing, and clustering challenges exceed our Excel string-matching capabilities. We also eliminated dimensions with predominantly null or homogeneous values (e.g., *charge*; all values = 0) attributable solely to our small drug sample.

<sup>74</sup> Unfortunately, it appears that tamsulosin is a false hit in the original ClinicalTrials.gov search, which cannot discriminate between the various reasons a drug and a condition co-occur in a trial. In the case of tamsulosin and prostate cancer, it appears that tamsulosin is only being tested to relieve urinary side effects of brachytherapy.

database contained four normalized generic parent names in common with the 1,823 linked to clinical trials on prostate cancer, and one of these, tamsulosin, was deemed to be a suitable model compound based on our normalized *chemistry* dimensions and values. Searching for drugs with similar (to tamsulosin) normalized *chemistry* dimension-value pairs retrieved the other three already in clinical trials plus four new ones. Two of the latter compared well to the top six drugs retrieved by the similar structure searches available on DrugBank and KEGG DRUG according to an objective measure of tamsulosin similarity based on our *chemistry* dimensions and publicly available values. Moreover, these measures were able to clearly discriminate these two, prazosin and terazosin, from the very un-tamsulosin-like steroid finasteride. Interestingly, tamsulosin also shares with prazosin and terazosin its molecular targets, the A, B, and D isoforms of the alpha1 adrenergic receptor.

#### 4.3.3.3.3 Consumers.

Consumer Use Case A. "Are there any natural **[herbal]** substitutes for the ~~hormone~~ replacement BPH therapy agent ~~Prempro~~ **{Proscar, Flomax, Avodart, Ticlid, Viadur ...}**?"

Like Health Use Case B, the goal of this query is to find the names of drugs indicated for BPH (benign prostatic hyperplasia). This query adds a constraint that the drug must be an herbal product, and uses trade names to exemplify non-herbal BPH drugs. A search in our database for {(O) = *clinical - indication - ...*; (Q) = benign prostatic hyperplasia} and "herbal" in any column produced one row with {(B) = saw palmetto; (C) = MedMaster; (O) = *clinical - indication - herbal evidence grade A*<sup>75</sup>}. The co-occurrence of {(B) = saw palmetto; (O) = *clinical - indication - ...*; (Q) = benign prostatic hyperplasia} was confirmed by 15 additional data rows representing {(C) = [ChemIDplus, ClinicalTrials.gov, DrugDigest, DrugInfo, MedMaster, MeSH, PubChem]}. The herbal constraint was confirmed by 14 additional rows with {(B) = saw

<sup>75</sup> The meaning of "grade A" can be obtained from the MedMaster dimension link in column G: "\*Key to grades: A: Strong scientific evidence for this use ..."  
<http://www.nlm.nih.gov/medlineplus/druginfo/natural/patient-sawpalmetto.html#Evidence>

palmetto; (O) = *pharmacy - drug type*; (Q) = [alternative medicine, complementary medicine, herbal medicine, homeopathic preparation, medicinal plant, plant extract, plant product, western herb/natural substance]] representing {(C) = [ClinicalTrials.gov, MedMaster, MeSH, UMLS]}. Trade names comparable to Proscar, Flomax, Avodart, Ticlid, Viadur ... were found by {(B) = saw palmetto; (O) = *pharmacy - trade name ...*}; they were {(Q) = [Cobra Powerful Men's Performance Enhancer, Elusan Prostate, Herbal Breast Enhancement, Herbal Mens Performance Enhancer, Mens Herbal Enhancement, Permixon, Prosta Urogenin, Prostagutt, Prostaserine, Prostata, Strogen]] and came from {(C) = [ChemIDplus, DrugInfo, MedMaster, MeSH, PubChem, UMLS]}.

#### *Criteria for usefulness*

1. Comprehensive coverage. Plovnick and Zeng (2004) used UMLS to normalize these queries, then executed them on MedlinePlus<sup>76</sup> (of which MedMaster is the "Drugs & Supplements" subsite) and Google, and compared the retrieval quality against a "gold standard answer" based on MD-oriented information from Harrison's Online and MDConsult. Our database's overlap with this collection is only partial (UMLS and MedMaster). However, we do integrate information from an additional consumer-oriented source (DrugDigest), plus clinical professional-oriented information from DailyMed, RXNORM, and ClinicalTrials.gov, plus research professional-oriented information from DrugBank, PubChem, ChEBI, ChemIDplus, and KEGG DRUG (clinicians may also be researchers). The {(B) = saw palmetto} subset consists of 568 rows of data representing a wide diversity of *clinical*, *pharmacy*, *biology*, and *chemistry* dimensions which would support follow-up queries to obtain more information about saw palmetto (such as the trade name query described above). This will be true in general for all our Consumer Use Cases.

2. Literary warrant fidelity. Consumer Use Cases A-G were adapted from Plovnick and Zeng (2004) as described in Methods. This satisfies the literary warrant criterion for all of them.

<sup>76</sup> <http://www.nlm.nih.gov/medlineplus/>

### 3. IR performance.

a. Larger retrieval. Saw palmetto was retrieved from our database because it, along with finasteride, is one of UMLS/NDFRT's *Other related/may\_be\_treated\_by* relations for "Prostatic Hypertrophy" and it happens to be an herbal. Nevertheless, despite their coverage of over 900 more drug compounds (Table 11), DrugDigest and ClinicalTrials.gov (the only other two of our resources that enable searching for drugs by indication) produce the same answer; that is, our database performs as well as they do on this query. With regard to Plovnick and Zeng's system, UMLS does not link saw palmetto or any other drugs to BPH (as opposed to "Prostatic Hypertrophy") and so would retrieve nothing. In contrast, the top hit for the Google search "herbal BPH" listed "Western herbs: saw palmetto, pygeum, pumpkin seeds, nettle root; Chinese herbs: vacarria, plantago seed, rehmannia; Chinese formulas: Guizhi Fuling Wan, Niu Che Shenqi Wan, Jingui Shenqi Wan, Tonglong Tang, Dahuang Mudan Tang."<sup>77</sup>

b. More robust. Plovnick and Zeng used four resources compared to our 15. Of the latter, eight contributed to satisfying this specific use case (ChemIDplus, ClinicalTrials.gov, DrugDigest, DrugInfo, MedMaster, MeSH, PubChem, and UMLS) and a ninth (RXNORM) contributes additional information about saw palmetto. This suggests that the Google hit described above (a likely facsimile of what Plovnick and Zeng's system would produce), although larger in the sense of more drugs, would largely fail to be confirmed by their gold standard validity test.

c. More efficient. Data reduction exhibited by the  $\{(O) = \textit{clinical - indication} \dots; (Q) = \textit{benign prostatic hyperplasia}\}$  subset (178 rows) was: drug name (B/D) 15/50 (30%); dimension name (O/F) 5/21 (24%); value (Q/H) 1/93 (1%); drug-dimension-value triple 29/141 (21%). For the entire  $\{(O) = \textit{clinical - indication} \dots\}$  subset of our database (1,544 rows) it was: drug name (B/D) 15/81 (19%); dimension name (O/F) 19/31 (61%); value (Q/H) 620/904 (69%); drug-dimension-value triple 882/1087 (81%). For these two retrievals, the number of databases

---

<sup>77</sup> <http://www.itmonline.org/journal/arts/bph.htm>

was reduced from ten to one (10%) and 11 to one (9%), implying an even greater reduction in the number of commands, queries, keystrokes, and time.

#### *Summary of Consumer Use Case A*

Like Health Use Case B, the goal of this query is to find the names of drugs indicated for BPH (benign prostatic hyperplasia). This query adds a constraint that the drug must be an herbal product, and uses trade names to exemplify non-herbal BPH drugs. Our database produced an answer consisting of the normalized generic name "saw palmetto" and all of its related alternative names, including 11 trade names. Our system satisfied the criteria for usefulness as follows: more dimensions and resources than the reference system; fidelity to reference's information need; larger retrieval and more resources with normalized than raw value search; and data reduction. The reference system produced a larger retrieval because its sources included Google, but the provenance/validity of the result is suspect.

Consumer Use Case B. "How are ~~arrhythmias~~ **BPH** treated?"

Consumer Use Case C. "Is there treatment for ~~restless-legs-syndrome~~ **baldness**?"

Consumer Use Case D. "What are scientifically validated [**approved**] treatments for ~~cancer~~ **BPH**?"

Consumer Use Case E. "What are ~~scientifically-validated~~ **experimental** treatments for **prostate** cancer?"

Consumer Use Case F. "What are the side effects of ~~Lexapro~~ **{Proscar, Flomax, Avodart, Ticlid, Viadur ...}** ?"

Our database's ability to answer these queries was demonstrated and discussed under prior use cases. "Treat" and "treatment" map to  $\{(O) = \text{clinical} - \text{indication} - \text{treatment} \dots\}$  and "approved treatment" to  $\{(O) = \text{clinical} - \text{indication} - \text{treatment} - \text{approved}\}$ . "Experimental treatment" could be defined as the (B) drug name result set difference between them for any given value (here  $\{(Q) = \text{prostate cancer}\}$ ) or, more conservatively, as  $\{(O) = \text{clinical} - \text{indication}$

- *clinical trial condition*}. Non-drug (surgical, etc.) treatments are not covered by our dimensions of drug information, but a cluster of 11 rows was found with {(O) = [*clinical - clinical trial comparison therapy, clinical - clinical trial co-therapy*]; (Q) = [laser transurethral prostatectomy, luteal phase ganirelix, retrospective non-intervention, androgen ablation therapy, photoselective vaporization of the prostate, radiation therapy, testosterone replacement]}. These (Q) values could be mapped to their corresponding indications (BPH, baldness, prostate cancer, etc.) through their clinical trial ID links in column S. For Case F, "side effects" maps to {(O) = *clinical - precaution - side effect...*} and trade names to {(O) = *pharmacy - trade name*}.

#### *Criteria for usefulness*

1. Comprehensive coverage. Same as Consumer Use Case A. In addition, side effect coverage was discussed under Research Use Case C.

2. Literary warrant fidelity. Same as Consumer Use Case A.

3. IR performance. Not analyzed.

Consumer Use Case G. "What foods should be avoided to prevent ~~cavities in children~~ **interactions with alpha blockers** ?"

Searching on "food interaction" throughout the database found it mapped to three non-null normalized dimension-value sets: {(O) = [*biology - ADME - absorption - food effect - bioavailability*]; (Q) = 20% increase}, {(O) = *clinical - precaution - food interaction - administration with food*; (Q) = [optional, recommended]}, and {(O) = *clinical - precaution - food interaction*; (Q) = [alcoholic beverage, grapefruit juice, natural licorice]}. The first two (from {(C) = DrugBank; (F) = Food Interaction} and {(C) = DailyMed; (F) = Precautions - Food Interaction}) have to do with how oral administration of the drug with food affects its absorption into the bloodstream and so is recommended, not recommended, or optional; these are not relevant to the use case. The last set links {(B) = [doxazosin, prazosin]; (C) = DrugBank; (F) = Food Interaction; (H) = [Avoid alcohol., Avoid natural licorice.]} and {(B) = dutasteride; (C) =

MedMaster; (F) = What special dietary instructions should I follow?; (H) = Talk to your doctor about drinking grapefruit juice while taking this medicine.}. Searching on {(Q) = alpha blocker} found it mapped to {(B) = [doxazosin..., prazosin, tamsulosin, terazosin]; (O) = *clinical - therapeutic class*}. Thus dutasteride is not an alpha blocker, so grapefruit juice is eliminated, and our system's "answer" to the query is "alcohol and natural licorice."

#### *Criteria for usefulness*

1. Comprehensive coverage. Same as Consumer Use Case A.
2. Literary warrant fidelity. Same as Consumer Use Case A.
3. IR performance. Not analyzed.

#### *Summary of Consumer Use Case G*

Searching on "food interaction" throughout the database found it mapped to three normalized dimension-value sets, only one of which was relevant this use case; it produced the answer "alcohol and natural licorice." Of the criteria for usefulness, comprehensive coverage and literary warrant were satisfied and the others were not analyzed.

Consumer Use Case H. A patient is taking Ticlid (ticlopidine hydrochloride) to prevent blood clotting on an implanted coronary stent. She is having difficulty breathing and wonders if it might be a side effect of the drug. She looks up Ticlid on MedMaster but the monograph section "What side effects can this medication cause?"<sup>78</sup> does not say anything about respiratory problems. She wishes that MedMaster had a "Search More Resources" button next to each section heading.

We imagined a hypothetical MedMaster clone with such a button next to each section title, and that clicking on the button does a search in our database for the section title; in this case {(C) = MedMaster; (D) = Ticlid; (F) = What side effects can this medication cause?}. The resulting cluster of data rows map to {(O) = *clinical - precaution - side effect*}. Executing the

<sup>78</sup> <http://www.nlm.nih.gov/medlineplus/druginfo/meds/a695036.html#side-effects>

implied normalized search for all side effects of Ticlid across sources, {(D) = Ticlid...; (O) = *clinical - precaution - side effect*}, retrieved 94 rows; 64 from DailyMed and 30 from DrugDigest (Table 31). This data could be displayed to the user, perhaps via an option to subset by user type (DailyMed for clinicians, DrugDigest for consumers), and with hyperlinks (contained in columns E and G of our database) to the original source display.

The user would be able to see right away that we classified "breathing difficulty" as a "major" side effect of Ticlid ({(D) = Ticlid...; (O) = *clinical - precaution - side effect - major*}). With the hyperlinks, she could click on "breathing difficulty" in the data summary table and be taken to DrugDigest's Ticlid information page's section entitled "What side effects may I notice from this medicine?" There she would learn that "breathing difficulty" is one of Ticlid's "[s]ide effects that you should report to your doctor or health care professional as soon as possible" (which we normalized as {(O) = *clinical - precaution - side effect - major*}). Here on DrugDigest's Ticlid page, as well as in the data summary table (Table 31 or subset), the user could also be advised of numerous other major side effects of Ticlid to be watchful for.

The normalized side effect search could have been broadened from {(D) = Ticlid...; (O) = *clinical - precaution - side effect*} to {(B) = ticlopidine...; (O) = *clinical - precaution - side effect*}. Executing this search retrieved 397 rows: 320 from DailyMed, 60 from DrugDigest, 15 from MedMaster, 1 from ClinicalTrials.gov, and 1 from UMLS. In this case such a strategy was not needed to find "breathing difficulty" and did not add significantly to the list of side effects of Ticlid shown in Table 31. In other cases it might be a useful "Broaden the search" option.

#### *Criteria for usefulness*

1. Comprehensive coverage. Side effect coverage: see Research Use Case C.

2. Literary warrant fidelity. None.

3. IR performance.

a. Larger retrieval. Table 31 contains 88 unique side effects of Ticlid, compared to 15 that can be found via MedMaster. Broadening the normalized side effect search from {(D)

= Ticlid...; (O) = *clinical - precaution - side effect*} to {(B) = ticlopidine...; (O) = *clinical - precaution - side effect*} produces 95.

b. More robust. Our database's {(O) = *clinical - precaution - side effect*} subset (1,255 rows) is informed by five sources including the use case's imagined one (MedMaster) and the FDA-vetted gold standard for U.S.-approved drugs (DailyMed).

c. More efficient. This use case illustrates how our model database's design could facilitate search (i.e., make it more efficient), not just in the classical IR query-result sense, but also from a Semantic Web (linked data, mashup) viewpoint. Our database effectively linked (mashed up) data between MedMaster and DrugDigest, allowing the user to find relevant information in DrugDigest using the MedMaster interface; indeed, without prior awareness of DrugDigest. Data reduction was not analyzed for this use case due to the distorting effect of not parsing and normalizing all the long DailyMed raw side effect values (H) as discussed under Research Use Case C. Data reduction exhibited by all use case retrievals with more than 50 rows of data is summarized in Table 32.

#### *Summary of Consumer Use Case H*

The goal here was to expand a search for side effects of a given drug without having to interact with multiple resource interfaces. We showed that the MedMaster interface could be linked to our database such that the normalized generic name and dimension could be inferred from the MedMaster drug name and section heading, used to find additional corresponding values from other resources in our database, and (optionally) hyperlink to the other resource's relevant information *in situ* and/or broaden/narrow the search by telescoping up and down our dimensions hierarchy. This use case illustrates how our model database's design could facilitate search, not just in the classical IR query-result sense, but also from a Semantic Web (linked data, mashup) viewpoint. Our database effectively linked (mashed up) data between MedMaster and other resources, allowing the user to find relevant information in DrugDigest using the MedMaster interface; indeed, without prior awareness of DrugDigest.

**Table 31. Side effects of Ticlid (Consumer Use Case H).**

The "severity" values are the right-most lexical components of our subdimensions of *clinical - precaution - side effect*; e.g., "major" → *clinical - precaution - side effect - major*.

<i>side effect</i>	<i>severity</i>	<i>source</i>	<i>source's drug name</i>
abnormal liver function test		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
agranulocytosis		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
allergic pneumonitis		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
allergic reaction	major	DrugDigest	Ticlid
allergic reaction		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
anaphylaxis		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
angioedema		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
anorexia		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
aplastic anemia		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
appetite decreased	major	DrugDigest	Ticlid
arthropathy		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
asthenia		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
black tarry stools	major	DrugDigest	Ticlid
bleeding increased		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
bone marrow depression		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
breathing difficulty	major	DrugDigest	Ticlid
cholestatic jaundice		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
conjunctival hemorrhage		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
diarrhea	major	DrugDigest	Ticlid
diarrhea	common	DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
dizziness		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
dyspepsia	common	DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
ecchymosis		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
eosinophilia		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
epistaxis		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
erythema multiforme		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
exfoliative dermatitis		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]

**Table 31. Side effects of Ticlid (Consumer Use Case H) (continued).**

<i>side effect</i>	<i>severity</i>	<i>source</i>	<i>source's drug name</i>
facial swelling	major	DrugDigest	Ticlid
fever	major	DrugDigest	Ticlid
flatulence		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
gastrointestinal bleeding		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
gastrointestinal fullness		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
gastrointestinal pain	common	DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
headache	major	DrugDigest	Ticlid
headache		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
hematemesis	major	DrugDigest	Ticlid
hematuria	major	DrugDigest	Ticlid
hematuria		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
hepatitis		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
hepatocellular jaundice		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
hives	major	DrugDigest	Ticlid
hyponatremia		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
immune thrombocytopenia		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
intracerebral bleeding		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
itching	major	DrugDigest	Ticlid
jaundice	major	DrugDigest	Ticlid
joint pain	major	DrugDigest	Ticlid
joint swelling	major	DrugDigest	Ticlid
kidney failure		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
leukemia		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
lip swelling	major	DrugDigest	Ticlid
liver failure		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
liver necrosis		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
maculopapular rash		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
myositis		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
nausea	major	DrugDigest	Ticlid
nausea	common	DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
nephrotic syndrome		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]

**Table 31. Side effects of Ticlid (Consumer Use Case H) (continued).**

<i>side effect</i>	<i>severity</i>	<i>source</i>	<i>source's drug name</i>
neutropenia	common	DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
nosebleed	major	DrugDigest	Ticlid
pain		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
paleness	major	DrugDigest	Ticlid
pancytopenia		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
peptic ulcer		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
perioperative bleeding		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
peripheral neuropathy		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
pruritus		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
purpura	common	DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
rash	common	DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
reticulocytosis		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
sepsis		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
serum sickness		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
skin rash	major	DrugDigest	Ticlid
spontaneous posttraumatic bleeding		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
Stevens-Johnson syndrome		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
stomach pain	major	DrugDigest	Ticlid
sudden weakness	major	DrugDigest	Ticlid
systemic lupus (positive ANA)		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
thrombocytopenia		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
thrombocytosis		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
thrombotic thrombocytopenic purpura		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
tinnitus	major	DrugDigest	Ticlid
tinnitus		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
tongue swelling	major	DrugDigest	Ticlid
unusual bleeding	major	DrugDigest	Ticlid
unusual bruising	major	DrugDigest	Ticlid
urination difficulty	major	DrugDigest	Ticlid
urination pain	major	DrugDigest	Ticlid
urticaria		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]

**Table 31. Side effects of Ticlid (Consumer Use Case H) (continued).**

<i>side effect</i>	<i>severity</i>	<i>source</i>	<i>source's drug name</i>
urticarial rash		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
vasculitis		DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
vomiting	major	DrugDigest	Ticlid
vomiting	common	DailyMed	Ticlid (ticlopidine hydrochloride) Tablet, Film Coated [Roche Pharmaceuticals]
wheezing	major	DrugDigest	Ticlid

**Table 32. Data reduction in use case query results.**

Only analyses based on >50 rows of retrieved data are shown.

<i>Use case</i>	<i># rows</i>	<i>% normalized values/raw values</i>				
		<i>drug name (B/D)</i>	<i>dimension (O/F)</i>	<i>value (Q/H)</i>	<i>triple (BOQ/DFH)</i>	<i>database</i>
Health A	101	8	38	44	61	10
Health B	187	20	30	1	20	50
Health I	58	26	20	4	34	20
Research A	818	21	34	93	94	8
Research B "disease"	207	36	67	84	128	14
Research B "phenotype"	486	20	35	86	137	8
Research C side effect + ATC class	1,277	13	29	129	112	17
Research C target + chem. class	240	14	7	46	39	11
Research C indication + ther. class	2,272	9	47	48	49	8
Research D chem.subset	523	20	80	96	102	12
Research D chem.all	1,909	18	77	86	81	9
Consumer A indic.=BPH	178	30	24	1	21	10
Consumer A indic.all	1,544	19	61	69	81	9

## Chapter 5. Discussion

We investigated a new approach to drug information integration that can be considered a type of ontology matching/merging (OM). Existing practical drug information representations are both diverse and in an early, pre-formal stage of ontology development. Integrating them can therefore be seen as early, pre-formal OM. This is not only an urgent practical issue because of drug information's scientific, medical, and economic importance, but also a good test case for the general problem of early, pre-formal OM within a domain. Our method for resolving/integrating diverse representations of drug information depends on the ontology-like representation we call *dimensions of drug information*. Our specific objectives were to provide a plausible, empirical definition of the dimensions of drug information (5.1 below), and to test its validity (5.2) and usefulness (5.3).

### 5.1 What are the Dimensions of Drug Information?

We focused on a subset of drug information relevant to four domains: pharmacy, chemistry, biology, and clinical medicine. In our initial survey of 23 relevant information resources, we found 39 dimensions that corresponded well to those that we expected based on domain-specific (1) typical queries, applications, and information needs of users, and (2) literary warrant. In a corollary generic name coverage overlap analysis, we found 5,000 to be a reasonable estimate of the U.S.-approved generic parent drug name universe, but were surprised by the low degree of overlap between sources. This means that drug information resources do not agree on extensionally defining even the most fundamental dimension, *generic name*. That is, they do not agree on the most fundamental ontological question, "What is a drug?"

The more detailed analysis contained in our experimental database discovered over 500 normalized dimensions in 15 of the initial 23 resources. Even when narrowed to those most relevant to our use case focus, there were still 375 which could be expressed as a six-level hierarchy with the four domains (*pharmacy*, *chemistry*, *biology*, *clinical*) comprising the top level. This result is consistent with Bawden and Robinson's (2010) statement, "We may expect,

therefore, that pharmaceutical information will be of a diverse nature but will follow healthcare knowledge in having a generally clear and consistent structure, in the form of a complex hierarchy with many levels" (p. 65). The second level contains 54 dimensions comparable to the 39 found in the initial study. All of the initial 39 were confirmed at the second or third level of the database hierarchy. The *biology* domain seemed to account for a disproportionately high number of dimensions and hierarchical splitting, while *clinical* was disproportionately low. This could imply that our dimension hierarchy has an inter-domain imbalance in conceptual splitting, whether as a result of domain, source, or our own bias.

## 5.2 Do Dimensions Lead to Valid Groupings of Resources?

The matrix score extension of our arbitrary four-domain classification of the initial survey's 39 dimensions to its 23 resources produced a reasonable resource classification that could be considered passing the test of face validity. For example, ChEBI, PubChem, and ChemIDplus were classified under *chemistry* in accordance with their names, and the richest resources (DrugBank, WHO-ATC, DailyMed, and UMLS) had the most equitable all-domain coverage. Anomalous/unexpected classifications included USAN under *chemistry* rather than with USP under *pharmacy*, and the failure of the two MeSH sources to score at all under *biology*. RXNORM's 100% *pharmacy* classification is inconsistent with Liu et al.'s (2005) claim of utility for "health care personnel including prescribing physicians and nurses" which would seem to correspond better to our *clinical* domain; however, we regard this as a case of cross-domain usage, not evidence against our classification.

Correspondence analysis of the initial survey matrix also produced evidence of its partial validity. *Chemistry*- and *clinical*-oriented sources were effectively polarized in the joint plot because sources tending to cover *clinical* dimensions tended not to also cover *chemistry* dimensions. The *pharmacy* and *biology* dimensions lacked discriminating power. Like sources tending to cover all four domains, the *therapeutic class* dimension was least polarized. That is,

*therapeutic class* tends to be covered by all the resources we examined, suggesting that it is important in all four domains.

Hierarchical cluster analysis of the initial survey matrix also produced evidence of its partial validity. The top dimension cluster corresponded closely to our *biology* domain-classified dimensions, and a second tight cluster to our *clinical* dimensions. These two clusters coalesced at a close similarity distance with two more *clinical* dimensions, evoking the well-known super-domain of *biomedical* knowledge. The only remaining *clinical* dimension - *therapeutic class* - did not strongly cluster with any other dimension, supporting our correspondence analysis interpretation that it is important to all four domains. Cluster analysis also produced strong evidence for the validity of our initial *chemistry*-classified dimensions, along with the reasonable suggestion that *bioassay*, *pathways*, and *toxicity* should be considered *chemistry* instead of *biology*. Of our four initial hypothetical domain-dimension groupings, only *pharmacy* failed to be supported by this analysis.

When viewed "sideways" (i.e., by source), the matrix produced hierarchical source clusters supporting our initial *chemistry* classifications, but not those of the other domains. The close similarity of MeSH MH and MeSH all, USP and USAN, MedMaster and DrugDigest, RXNORM and Drugs@FDA, and WHODRUG and WHO-ATC, is consistent with each pair's organizational overlap, scope, and/or mission. PubChem and ChemIDplus also formed a sensible cluster, but it was missing ChEBI. Three of the four "all-purpose" resources - UMLS, DailyMed, and DrugBank - again were distinguished, in this analysis by failure to cluster with other sources. These small, sensible clusters, combined with the lack of support for three out of four domains, suggest that hierarchical cluster analysis of the initial survey matrix was statistically robust, but did not validate its extension from domain classification of dimensions to domain classification of sources. Cluster analysis of the resource-by-dimension matrix implied by the experimental database, on the other hand, did not produce any sensible dimension or source clusters, and so

may not be statistically robust due to inappropriate use of continuous scores as opposed to the mostly binary scores of the initial survey matrix.

### **5.3 Can Dimensions Facilitate Integration/OM Tasks?**

We provided some evidence that dimensions are useful (as well as valid) for classifying resources and selecting sources appropriate to a given information need. In general, however, we leave these questions up to further study requiring implementation of our data model in a more refined IR system and evaluation by human subject experts and other end users. We provide much more evidence that dimensions can facilitate pooling data from different resources.

#### **5.3.1 Data reduction.**

Logically, it seems almost trivial to assert that our normalization process should result in data reduction; that is, the number of unique raw representations should always be greater than the number of unique corresponding normalized representations within any cross-section of the database. However, because diverse raw data formats (terms, relations, items in a list, whole paragraphs, etc.) were loaded into the spreadsheet the same way, there was an antagonism between conflation of short strings representing a single concept (which lowers the ratio of unique normalized to unnormalized representations), versus parsing of longer strings into multiple dimension-value pairs (which raises the ratio). Therefore, although overall data reduction was achieved, it was much less than expected. Additional data reduction of the dimensions was achieved by hierarchical aggregation, but it did not lead to significant overall additional data reduction of the drug-dimension-value triples. However, use-case-specific examples of data reduction were generally more impressive.

#### **5.3.2 Automatic normalization of additional data.**

Our experimental database contains over 17,000 unnormalized source-dimension-value triples mapped to normalized dimension-value pairs which can be regarded as a training set for automating the normalization of additional raw data from the same 15 sources, bringing the important goal of building an integrated, comprehensive (all drugs) database within reach. The

probability of getting an accurate new normalization, based on "priors" (the current data),  $p(C, F \rightarrow O)$ , reflects how semantically well-differentiated are the raw dimensions in our sources, and their influence on our choice of normalized dimensions. Many of the  $\sim 1700$   $p(C, F \rightarrow O)$ 's are 1.0, and of those that are significantly lower, nearly all can be dramatically raised (usually to 1.0) based on further parsing of the raw value (e.g., "treatment" implies that the normalized dimension is *clinical - indication - treatment...*). Alternatively, smaller but useful gains can be obtained by dimensional hierarchical aggregation, which would not require free text parsing.

For the specific goal of integrating more *indication* data, we showed that the best precision was exhibited by DrugDigest's "Learn how <this drug> is/are used to treat."; UMLS's "Other Related/ may\_be\_treated\_by"; DailyMed's "Indications and Usage"; MedMaster's "Other uses for this medicine"; and DrugBank's "Indication" dimensions. DrugDigest and UMLS also have the advantage of having values in controlled terminological format (i.e., no parsing needed). However, these data do not take into account the superior provenance of the DailyMed information. Disagreements between the associated DrugDigest, UMLS, and DailyMed values remind us that  $p(C, F \rightarrow O)$  is really measuring *semantic* (as opposed to *pharmaceutical*) precision. That is, all UMLS values associated with "Other Related/ may\_be\_treated\_by" may be valid treatment indications for some drug, but they are not necessarily the correct indications for the specific drug at the other end of the triple.

Although the pattern-matching algorithms used here were primitive compared to true natural language processing, this exercise demonstrated the important principle of leveraging mechanization to expand the database to a truly practical size for real-world drug information users. Thus our model database is valuable not just for defining (Q1) and demonstrating the validity (Q2) and usefulness (Q3) of dimensions of drug information, but also as a model or prototype for a much larger database of something closer to *all* drug information capable of satisfying many *kinds of* information needs. This is arguably the most important potential future extension of our work.

### 5.3.3 Satisfying use cases.

Finally, we demonstrated that our experimental database could satisfy a variety of use cases derived from published literature representing the user types corresponding to our domain focus, and that this success depended on the dimensional basis of the data's multi-resource integration. The nature and degree of success varied by use case, and were reported in detail in the Results section. In each case, our database's performance was compared to the publication's information system if possible, otherwise to our individual unintegrated sources, according to specific criteria for usefulness: comprehensive coverage (more dimensions, values, and/or resources), literary warrant fidelity, and better IR performance defined as larger, more robust, and more efficient (data reduction). We did not evaluate precision/recall or other measures of truth/accuracy because such measures do not reflect on the dimensions *per se* so much as the accuracy of the sources (which is beyond our control and not our goal to evaluate) and our specific mappings. This we leave up to further study requiring implementation of our data model in a more refined IR system and evaluation by human subject experts and other end users. Of particular note were the following results.

In Research Use Case C, since our entire drug sample was retrieved by the initial common-side-effect-different-therapeutic-class query, given that our sample was based on the drugs' common indication (benign prostatic hyperplasia), one wonders if that more straightforward dimension (*indication*) could be effectively substituted for the initial drug set selection in Campillos et al.'s (2008) method, given a robust database of normalized drug-indication relations. If so, Campillos et al.'s method constitutes evidence supporting our assertion of the latter's practical nonexistence prior to our work.

In Research Use Case D, our system retrieved prazosin, terazosin, and doxazosin as novel drug candidates for treating prostate cancer on the basis of their similarity to tamsulosin. Prazosin and terazosin compared well by an objective measure of chemical similarity to tamsulosin to the top compounds retrieved by the DrugBank and KEGG DRUG similar structure

searches. Interestingly, all four share the same molecular targets, the A, B, and D isoforms of the alpha1 adrenergic receptor. If tamsulosin were truly being tested in clinical trials for treatment or prevention of prostate cancer, and its relevant mechanism of action were through the alpha1 adrenergic receptor, this would be evidence for the usefulness of these results.

Consumer Use Case H illustrates how our model database's design could facilitate search not just in the classical IR query-result sense, but also from a Semantic Web (linked data, mashup) viewpoint. Our database effectively linked (mashed up) data between MedMaster and DrugDigest, allowing the user to find relevant information in DrugDigest using the MedMaster interface; indeed, without prior awareness of DrugDigest.

#### 5.4 Limitations

Compression of the 550 normalized dimensions in the experimental database to the 375 in Appendix G involved weeding out "pseudo-dimensions" such as other database IDs. This needs reconsideration. Our thinking was that, at least for the purpose of OM, dimensions should be restricted to those whose values have some kind of natural or general significance. Database IDs would seem to be an example that does not meet this requirement. Yet the CAS number is a gold standard for distinguishing unique chemical compounds, and we left in patent number (as *pharmacy - approval info - patent number*) thinking that the patent could be parsed into valuable, natural, general information. Perhaps we should consider database IDs and other excluded dimensions to be placeholders for information not yet processed.

Contrary to the way we have represented both in the same morphosyntactic way in Appendix G, the monohierarchical mapping of the second dimensional level to *biology*, *chemistry*, *clinical*, or *pharmacy* represents an entirely different semantics than the breakdown of the second level into subtypes represented by the lower levels. In a sense, for example, *clinical - indication* means "indication-type clinical [information]" similarly to how *indication - treatment - approved* means "approved-treatment-type indication." But clinical users, professionals, resources, enterprises, etc., do not have a monopoly on the use of *indication* information anything

like the way that an approved treatment indication is inherently an indication. Analogous statements could be made even for highly specialized dimensions such as *biology - molecular target - gene name* or *chemistry - Lipinski's Rule of Five - molecular weight*. The face validity of the first-to-second-level mappings was only partly supported by correspondence analysis and cluster analysis, did not extend appreciably to classifying resources, and contributed little if anything to satisfying our test use cases. Perhaps the top level should be dropped.

The dimension-value boundary may be fuzzier than we have indicated. Appendix G does not show the dimension-value matchups. These are much more numerous, of course, and can be examined in the database. As noted previously, the same value may apply to widely different dimensions; e.g., medical conditions ("prostate cancer"; "thrombosis"; "pregnancy"; etc.) may be *indications*, *contraindications*, or *side effects*. We also know that some of the lower-level subtype qualifiers in Appendix G are also values for other dimensions. An obvious example is *approved*, a subtype of *clinical - indication* but a value for *pharmacy - approval status*. Others are *oral*, a subtype of *biology - toxicity* but a value for *pharmacy - route of administration*, and *rat*, *mouse*, and *monkey*, subtypes of *biology - toxicity* but potential values for *biology - organism affected*. Thus our normalized dimension-value pair

{dimension = *biology - toxicity - LD50 - oral - rat* ; value = "418 mg/kg"}

could be alternatively represented as

{dimension = *biology - toxicity - LD50* ; value = "418 mg/kg [oral; rat]"}

or even

{dimension = *biology - toxicity* ; value = "LD50=418 mg/kg [oral; rat]"}

This is beginning to look like facet analysis, where, rather than stringing together qualifiers to create hundreds of dimension subtypes, we would consider the data in terms of combinations of facets such as *approval status*, *organism/species*, *route of administration* (or even *body site*), and *indication* (or even *condition plus drug effect [prevents, treats, causes,*

*exacerbates, ...*). The question is, what will best support OM of the drug information we have sampled, operationalized as pooling data from different sources?

Our research use cases are perhaps more "professional grade" than our health care professional and consumer use cases, naturally reflecting our own prior interests and experience, but weakening our case for having provided a drug information integration strategy and database that can serve all three user types simultaneously. This is an area for improvement in follow-up research. In addition to recruiting appropriate human participants (e.g., focus groups to think up use cases), a possible resource is web blogs and forums targeted to physicians, pharmacists, patients, etc.

Research Use Case B had trouble with literary warrant fidelity. Our database was able to produce a list of cardiovascular drugs based on their WHO-ATC classifications and map them to their molecular targets, but it does not cover phenotypes *per se*. As a substitute for phenotypes, our first simulation used target biological correlates to remain as true as possible to the target-based approach of Castle et al. (2007). However, this simulation failed to produce any matches comparable to their "decreased heart rate:cardiovascular" match. Our second simulation used *drug* biological (effect, mechanism, and pathway) correlates as a substitute for phenotypes. This produced a set of matches which looked more like "decreased heart rate:cardiovascular" but in so doing it strayed from the target-based approach of Castle et al. However, it was able to produce over 1,000 phenotype-like:disease hypotheses based on a larger and more diverse resource collection and true *indication* values that are semantically closer to "diseases" than the few WHO-ATC *therapeutic class* values that can be so mapped.

In Research Use Case C, it seemed that the novel hypothetical indications generated by our method would be more useful than the novel hypothetical therapeutic classes, the latter tending to be too general or inferable from known classes. On the other hand, most of the novel hypothetical indications came from ClinicalTrials.gov *Conditions* which means that (1) they include false positives (co-occurrences of a drug and condition in a trial other than for treatment

or prevention) and (2) even the true positives are "doubly hypothetical" in the sense that the target-sharing drug's efficacy has not been proven. Removing the therapeutic classes and ClinicalTrials.gov data resulted in smaller retrieval size more like that of Campillos et al. (2008).

In our first public presentation, an interesting point was made concerning the over-representation of U.S. FDA-approved information in our work. We do include non-U.S. sources (WHO-ATC, ChEBI, DrugBank, KEGG, and others). The international aspect was one of our earliest proto-dimensions, since approval of drugs' names, indications, target populations, etc., can vary widely by country. Although our focus later turned elsewhere, the subdimension *approved* has independently come out of our data analysis under *indication*. One can imagine it also under *trade name*, among others, and further qualified by geopolitical qualifiers such as country names, *FDA*, *USAN*, *INN*, *BAN*, *JAN*, etc. This point deserves additional research. Overcoming the *international* heterogeneity of drug information, in fact, may be drug OM's biggest and most rewarding challenge.

## Chapter 6. Conclusions

This dissertation has investigated a new approach to the general issue of resolution between different representations of "reality"; that is, the different classifications, thesauri, indexing methods and similar schemes that have been developed to characterize one domain by different communities. We were particularly concerned with the situation in which the representations within a specific domain are not, in general, particularly well formed. The method that was investigated can be considered a type of ontology matching/merging (OM), specifically early, pre-formal OM. This method was tested in one specific domain, drug information. Because drug information representations are both diverse and in an early, pre-formal stage of ontology development, their resolution is not only an urgent issue in its own right, but also a good test case for the general problem of early, pre-formal OM within a domain. We claim that our results will advance not only the state of drug information, but also provide a general framework for addressing the larger problem of which the drug information case is but one example. Such extensions would follow the historical pattern of advances in pharmaceutical information "play[ing] a major role in advancing information science itself" (Bawden & Robinson, 2010, p. 94).

### 6.1 Contributions to Drug Information

We have demonstrated a novel, coherent, literature-grounded, useful technique for early, pre-formal OM in the drug domain. Our most important contribution was to provide a plausible, empirical answer to the question "What are the dimensions of drug information?" relevant to pharmacy, biology, chemistry, and clinical medicine. Such a "standard framework for describing drug information sources is a necessary step towards improving the discoverability of such resources by humans and agents." (Sharp et al., 2008, p. 664). We have tested our framework's usefulness for classifying resources, pooling data from diverse resources, and satisfying diverse use cases, with generally (but not universally) positive results. In so doing we made another potentially valuable contribution in the form of a model database with mappings/ probabilities for

its rapid expansion to a truly practical size ("all drugs") and "rough-cut" clues for seeding NLP-based improvements.

Our attempt to estimate the size of the generic name universe by resource coverage overlap analysis is novel, as far as we know. We found 5,000 to be a reasonable estimate of the U.S.-approved generic parent drug name universe, but this number is both very rough and inevitably dated. Our methods need to be refined, applied to larger resource collections, and automated so that updated estimates can be obtained over time. The low degree of overlap between sources was surprising, even alarming. Major public drug information resources do not agree on extensionally defining even the most fundamental dimension, *generic name*. That is, they do not agree on the most fundamental ontological question, "What is a drug?" This is a serious obstacle to development of a standard drug ontology.

## 6.2 Contributions to Library and Information Science

Ultimately, we could not clearly distinguish our approach from facet analysis or domain analysis. The fuzzy dimension-value boundary we observed could imply the need for a more facet analytic approach as discussed in Section 5.4. It also evokes the domain analytic principle that "[t]here is [no] simple dichotomy between structure and content" (Hjørland & Albrechtsen, 1995, p.406). Facet analysis would not have permitted the four-domain top level of our dimensions hierarchy (*molecular weight* is a facet of a chemical, not a facet of *chemistry*), consistent with our findings of only equivocal support for it by correspondence analysis and cluster analysis. Domain analysis similarly could have predicted our conclusion that evaluation by human subject experts is needed to reinforce some of our demonstrations of our dimensions' usefulness.

The long-string parsing artifact which compromised our data reduction metrics is likely to be a general problem whenever our method is applied to diverse raw data formats (terms, relations, items in a list, whole paragraphs, etc.) Some kind of preprocessing may be indicated.

If NLP is not feasible, perhaps some incremental improvement could be had with purely numerical approaches such as "n-grams" to break up long strings into arbitrary short ones.

Similar considerations apply to the "pseudo-dimensions" such as other database IDs that need reconsideration as placeholders for information not yet processed (also discussed in Section 5.4). There is, in fact, no clear boundary between terms, short strings, long strings, lists, and links to structured information; they all represent arbitrary (user- and/or system-sensitive) subdivisions of a continuum. Thus, like the dimension-value boundary and "what is a drug [or any other entity defining a domain]?", the distinction between integrating a single unit of information into a database or ontology and integrating whole databases or ontologies may be counterintuitively fuzzy.

### **6.3 Contributions to Semantic Web**

Extrapolating from our model database's sample of nine generic parent names to our estimate of all such drug names predicts a database of 9-18 million rows if limited to the 15 resources we employed. Expanding the resource collection, making it more international in scope, updating it over time, and improving the parsing of long strings and external links, would make it even bigger. Our data warehouse approach was adopted as an expedient, since the main challenge was to get a grip on "what is there?" and put some of it into a consistent human-read-write format. Once this prototyping phase is finished, our data warehouse model will have served its purpose. It was never our goal to replicate/integrate anything like "all" drug information in a centralized database.

Given web availability of the required resources, linked data and/or mashup models would seem to be preferable for a variety of reasons, both principled and pragmatic, and in general, not just for drug information. However, as discussed in Sections 2.9 and 2.10, these technologies do not entirely avoid the OM problem. "Additional resources, which enable the creation of mappings between information sources, are required to compensate for heterogeneity across namespaces" (Sahoo et al., 2008, p. 752). We see our model database precisely as one such

additional resource with immediate applicability to drug information. This was illustrated by Consumer Use Case H; our imaginary user was enabled to find relevant information in DrugDigest using the MedMaster interface; indeed, without prior awareness of DrugDigest. Future extensions may address the general applicability of our methodology to other domains with Semantic Web aspirations hindered by early, pre-formal ontologies.

#### **6.4 Further Research**

Further refinement of our model is needed based on a larger drug sample. The foregoing remarks notwithstanding, this can be accomplished most expeditiously by expanding our database with information on more drugs from the same 15 resources based on our table of mappings and probabilities at [http://comminfo.rutgers.edu/~msharp/XKB/dimension\\_prediction.xls](http://comminfo.rutgers.edu/~msharp/XKB/dimension_prediction.xls). The results should be evaluated periodically to recompute this table, leading to iterative improvements in the automated normalization of additional data.

The most dramatic improvements, however, await the application of advanced NLP techniques to the parsing and normalization of long free-text raw values. We offer our "rough-cut" clues in columns J-N for seeding such an effort. If NLP is not feasible, perhaps some incremental improvement could be had with purely numerical approaches such as "n-grams" to break up long strings into arbitrary short ones.

Our estimate of the size of the generic name universe by resource coverage overlap analysis is very rough and inevitably dated. Our methods need to be refined, applied to larger resource collections, and automated so that updated estimates can be obtained over time. Sustained, systematic study of this issue may encourage resources to improve their consensus on it.

We provided some evidence that dimensions are useful (as well as valid) for classifying resources and selecting sources appropriate to a given information need. In general, however, we leave these questions up to further study requiring implementation of our data model in a more refined IR system and evaluation by human subject experts and other end users.

Finally, we would like to see our work recognized and expanded upon in the larger pharmaceutical research, biomedical informatics, OM, and Semantic Web communities. We offer our model database as a resource with immediate applicability to data linking and mashups of drug information. Additional extensions may address the general applicability of our methodology to other domains with Semantic Web aspirations hindered by early, pre-formal ontologies.

## Appendix A. Potential Applications of a Drug Information System

1. Health care personnel including prescribing physicians and nurses, and hospital personnel involved with drug ordering, inventory management, recording dose adjustments, checking drug interactions, or pharmacy management (Liu et al., 2005).
  - 1.1. Applications that can be satisfied by the current RxNav (containing RXNORM only):
    - 1.1.1. Finding other *trade names* of the same *generic name* and vice versa.
    - 1.1.2. Finding alternative *dose/forms* of the same *generic name* or *trade name*.
    - 1.1.3. Finding alternative combination drugs (*combo products*) containing a *generic name* (say, to avoid an allergic reaction to a another component).
  - 1.2. Applications that cannot be satisfied by the current RxNav:
    - 1.2.1. Finding alternate *generic names* or *combo products* for the same *indication* (disease or other medical condition).
    - 1.2.2. Finding alternate *generic names* with a common *chemical structure* or *chemical superclass* (e.g., benzodiazepines).
    - 1.2.3. Finding alternate *generic names* with a common biological *mechanism of action*.
    - 1.2.4. Identifying *contraindications* (medical conditions that make use of a given drug dangerous).
    - 1.2.5. Identifying potentially dangerous or useful *drug interactions* when two drugs are given simultaneously to the same patient.
2. Consumers.
  - 2.1. Applications that can be satisfied by the current RxNav:
    - 2.1.1. Finding other *trade names* of the same *generic name* and vice versa. ("How can I get this drug cheaper?" "I can't find this drug by name; what are some equivalent alternatives?")
  - 2.2. Applications that cannot be satisfied by the current RxNav:
    - 2.2.1. Finding *indications*. ("Why am I being given this drug?")

- 2.2.2. Finding *side effects*. ("I have symptom X - is it caused by my drug?")
- 3. Pharma scientists wishing to discover new candidate chemical compounds for drug development, and/or new indications for existing drugs (Castle et al., 2007; Quan, 2007; Campillos et al., 2008).
  - 3.1. Applications that can be satisfied by the current RxNav:
    - 3.1.1. Clustering various *trade names*, *combo products*, *NDCs*, and other quasi-synonyms of the same *generic name*.
  - 3.2. Applications that cannot be satisfied by the current RxNav:
    - 3.2.1. Clustering other quasi-synonyms of the same *generic name* such as *chemical names*, *manufacturer code names*, and *CAS<sup>79</sup> numbers*.
    - 3.2.2. Clustering various *generic names* by *therapeutic class* (a combination of chemical and biological attributes; e.g., "topical anesthetic").
    - 3.2.3. Clustering generic drug compounds by indication, contraindication, or other medical attribute.
    - 3.2.4. Clustering generic drug compounds by chemical substructures or attributes (acidity, solubility, molecular weight, etc).
    - 3.2.5. Clustering generic drug compounds by biological targets (enzymes, receptors, genes, metabolic or regulatory pathways, etc).
- 4. Basic researchers wishing to see overall trends in pharma development in a drug-target network representing the "interactome" of "polypharmacology" (Yildirim et al., 2007). Applications same as #3
- 5. Cancer researchers interested in faster, cheaper endpoints (than mortality/morbidity in human clinical trials) for initial screening of anti-cancer drugs; for example, "inhibits cellular proliferation" or "induces apoptosis" (Vogel, 2007). Such knowledge is easily recognizable as following the drug-predicate-object "triple" syntax of the envisioned ontology-like KB.

---

<sup>79</sup> Chemical Abstracts Service

6. Vocabulary managers, ontology engineers, medical data coders, etc., seeking content or authoritative validation for drug terminology, dictionaries, ontologies, etc., in support of above applications. Chen, Perl, Geller, & Cimino (2007) found that the UMLS, which can be viewed as RXNORM with additional dimensions of drug information (among many others!), is widely used as a terminology even though it was not designed as one.

## Appendix B. Potential Test Cases for System Evaluation

### 1. Consumer.

- 1.1. Find some alternative approved drugs beyond interchangeable *trade names* for the same *indication* that meet specific requirements (e.g., local availability; price; foreign travel; dissatisfaction with current drug). Example test case: finasteride [5 mg oral tablet]  $\rightarrow^{80}$  *indication* = benign prostatic hypertrophy [*therapeutic class* = testosterone 5-alpha reductase inhibitors]  $\rightarrow$  {dutasteride [0.5 mg oral capsule]; epristeride [?form]; ...}.
- 1.2. Find some drugs for an *experimental indication* due to, e.g., no approved therapy; dissatisfaction with current therapy. Example test case: Bartter syndrome  $\rightarrow$  spironolactone.
- 1.3. Find an *open clinical trial* due to, e.g., last recourse for relief of illness. Example test case: Bartter syndrome / spironolactone  $\rightarrow$  Clinical Trial NCT00276289, "Spironolactone to Decrease Potassium Wasting in Hypercalciurics on Thiazides Diuretics" (ClinicalTrials.gov, 2008).

### 2. Clinical/pharmacy worker.

- 2.1. Same as (to support) consumer.
- 2.2. To support consumer in especially desperate situations clinicians may want to expand their search to *preclinical experimental therapies*. Example test case: pulmonary hypertension  $\rightarrow$  sildenafil (Liu, Liu, & Guan, 2007).
- 2.3. Find all alternative *approved* drugs to support clustering for, e.g., stock management; sales tracking/reporting.

### 3. Researcher.

---

<sup>80</sup>  $\rightarrow$  signifies "is related to"; square brackets signify optional details or additional information; curly braces signify a set of multiple values.

- 3.1. Find possible alternative therapies for an *indication*. Example test case: pulmonary hypertension → (*experimental drug*) sildenafil → (*approved indication*) erectile dysfunction → {avanafil, dasantafil, vardenafil, ...}.
- 3.2. Vice versa; i.e., find possible new *indications* for known drugs: {avanafil, dasantafil, vardenafil, ...} → ... → sildenafil → pulmonary hypertension.
- 3.3. Combine *therapeutic class* with other *chemical and biological relations* to identify clusters of related drug compounds and infer new knowledge. Example test case #1: structurally similar compounds targeting TACR1 gene product (known to be associated with abnormal pain threshold) → *WHO-ATC class* "antiemetics and antinauseants" → TACR1 modulation produces antinauseant activity → connection between antinauseant activity and abnormal pain threshold. Example test case #2: *WHO-ATC class* "cardiovascular system" → list of cardiovascular drugs → associated *gene targets* → *phenotype* gene sets → highest ranking phenotype gene set is "decreased heart rate" → gene targets of drugs with cardiovascular activity are enriched in phenotypes associated with heart disease (Castle et al., 2007).

## Appendix C. Resource Evaluation Checklist

"RxNavPlus" is a nickname for a hypothetical, experimental, integrated drug information system.

### 1. Name of source

#### 1.1. Intro

([website](#)) and other general info

#### 1.2. Initial/common/general questions

##### 1.2.1. How does this source compare to the others with regard to:

##### 1.2.1.1. Info quality

1.2.1.1.1. Semantic range / kinds of info

1.2.1.1.2. Targeted users

1.2.1.1.3. Accuracy

1.2.1.1.4. Currency / update frequency

1.2.1.1.5. Authority

1.2.1.1.6. Completeness / comprehensiveness

1.2.1.1.6.1. Number of drug concepts cf. RXNORM

1.2.1.1.6.1.1. Coverage  $\equiv$  Number of IN (N~5600) -like concepts  
(~single approved generic names)

1.2.1.1.7. Granularity

1.2.1.1.8. Format (technical)

1.2.1.1.9. Format (lexical/semantic)

##### 1.2.2. Is the info content ( $\equiv$ KB) available for my use?

1.2.2.1. At what monetary cost?

1.2.1. Other initial questions (may be unique to one or more sources)

### 1.3. Desired integration

1.3.1. What info do we want to extract from this source and integrate into RxNavPlus?

1.3.2. What sort of integration is desired? I.e., how do we want the info to "look & feel" in RxNavPlus?

1.3.2.1. Integration modes. For each source, choose one, both, or neither.

1.3.2.1.1. Simple linkout  $\equiv$  user can click on a concept term (e.g., a drug name) displayed on RxNavPlus to open a new web browser window on the other source's native webpage with info about that concept, or as close to it an API can get (might be a generic search page or even the source's home page; permission/licensing might also be a factor). The user is then "on his own" to navigate the other source in this window. This capability is already being added to RxNav for some sources and could easily be expanded to others.

1.3.2.1.2. Extract & import  $\equiv$  terms or other strings representing concepts and their relations are copied from the other KB, normalized to RXNORM or other preferred terminology, and displayed "seamlessly" on RxNavPlus like RXNORM relations on RxNav (as a directed graph with semantic types = boxes/nodes, concepts = terms in boxes, relationships = arcs between boxes; see Figure 1).

1.3.2.1.2.1. Primary extract & import access options. Choose one. See below for definitions and decision criteria.

1.3.2.1.2.1.1. Database option

1.3.2.1.2.1.2. Linking option

1.3.2.1.2.2. Secondary extract & import processing. See below for details.

1.4. Simple linkout

1.4.1. Do we want a simple linkout to this source?

1.4.2. If so, from which kinds of RxNavPlus display terms? (e.g., IN, BN, CD,...)

1.4.3. Which kinds of RxNavPlus display terms "work" in the following 2x2 matrix of linking/API strategies to linking levels?

1.4.3.1. Linking/API strategies

1.4.3.1.1. Search  $\equiv$  the API must use a search interface the same way a human user would; i.e., type a search term into a box, hit a "Search" button, select a hit, etc.

1.4.3.1.2. Direct-link URL  $\equiv$  a stereotyped URL pattern into which a term or equivalent ID number can be inserted to yield a working URL that, when pasted into a web browser's address box, will yield the desired webpage.

1.4.3.2. Linking levels

1.4.3.2.1. Hitlist  $\equiv$  webpage containing a list of other, multiple, potentially relevant webpages.

1.4.3.2.2. Drug info  $\equiv$  webpage containing more detailed info, typically on a single drug or other concept.

1.5. Extract & import: Database option

1.5.1. Is the KB available as a database that can be locally replicated and manipulated at will?

1.5.1.1. If so, what is the version/update frequency?

1.5.2. How feasible is this integration strategy? May be impacted by

1.5.2.1. Desired info

1.5.2.2. DB complexity, format, etc., effect on extractability of desired info

1.5.2.3. Version/update importance

1.5.2.4. Version/update frequency

1.5.3. For the desired integration, is the DB option the only feasible option (cf. linking), or is it preferable to linking? If so, supply

- 1.5.3.1. Website or other method for obtaining DB
- 1.5.3.2. Local DB update schedule.
- 1.5.3.3. Secondary extract & import processing
  - 1.5.3.3.1. Database file, field, value specs (DB option only) E.g., Perl scripts  
for UMLS, spreadsheet for WHO-ATC, ...
  - 1.5.3.3.2. Text processing / parsing / NLP
  - 1.5.3.3.3. Normalization
    - 1.5.3.3.3.1. Preferred terminology
    - 1.5.3.3.3.2. Relation triples
  - 1.5.3.3.4. RxNavPlus display
- 1.6. Extract & import: Linking option
  - 1.6.1. Is the KB available via a website or web service that can be queried dynamically  
(at RxNavPlus query time), thus obviating DB version/update concerns?
  - 1.6.2. How feasible is this integration strategy? May be impacted by
    - 1.6.2.1. Desired info
    - 1.6.2.2. Webpage/service complexity, format, etc., effect on extractability of  
desired info
  - 1.6.3. For the desired integration, is linking the only feasible option (cf. DB), or is it  
preferable to DB?
  - 1.6.4. Which kinds of RxNavPlus display terms "work" in the following 2x2 matrix of  
linking/API strategies to linking levels? (defined above)
    - 1.6.4.1. Linking/API strategies
      - 1.6.4.1.1. Search
      - 1.6.4.1.2. Direct-link URL
    - 1.6.4.2. Linking levels
      - 1.6.4.2.1. Hitlist

1.6.4.2.2. Drug info

1.6.4.3. Secondary extract & import processing

1.6.4.3.1. Text processing / parsing / NLP

1.6.4.3.1.1. HTML context (linking option only)

1.6.4.3.2. Normalization

1.6.4.3.2.1. Preferred terminology

1.6.4.3.2.2. Relation triples

1.6.4.3.3. RxNavPlus display

## Appendix D. Resource Evaluation Summaries

**UMLS.** The UMLS Metathesaurus® contains a great deal of drug information from several open sources including RXNORM, CSP (CRISP Thesaurus), MSH (Medical Subject Headings), NCI (National Cancer Institute Thesaurus), NDFRT (National Drug File Reference Terminology), PDQ (Physician Data Query), USPMG (United States Pharmacopeia Model Guidelines), and VANDF (Veterans Health Administration National Drug File), as well as limited-use sources such as SNOMEDCT (Systematic Nomenclature of Medicine Clinical Terms) and NDDF (National Drug Data File Plus Source Vocabulary). This information scores well on availability and integration potential, being already "owned" by NLM and terminologically parsed and normalized. However, there are accuracy, completeness, precision, and currency issues (see "Preliminary Results" below). Research question: Is UMLS information quality good enough for prototyping? How do the individual contributing sources compare? What do the limited-use sources add to the open sources? How can the Metathesaurus be subsetted to optimize information quality and open access?

**DailyMed/SPL.** (Structured Product Labels) (<http://DailyMed.nlm.nih.gov>) Also already owned by NLM, this is a database of about 3,600 drug labels (also called "package inserts") containing gold-standard-quality information about each drug's chemistry, biology, and medicine - exactly what we are looking for, as will be explained below. Unfortunately, the "structure" in SPL is at a very high level. The specific conceptual relations we would like to capture for UMLS-like integration with RXNORM are expressed as free text, presenting a formidable natural language processing (NLP) barrier to that level of integration. However, for a simple linkout to the unparsed label, SPL is already integrated into RxNav, as mentioned above. A possible attraction of adding such an "SPL dimension" to RxNav, besides the high information content and quality, is that SPL is also a new U.S. Food and Drug Administration (FDA) standardization initiative which may become mandatory for the industry. A possible integration

plus is reported work on translating the labels into SNOMEDCT terms *[ref.]*. But even if we can't integrate it that way, SPL can still play a role as gold standard for spot-checking the information quality from other sources. Research question: Is the SPL information quality increment (over UMLS) big enough to justify added integration effort of extracting descriptors from text? Note there are two factors here: information quality and effort of extracting descriptors from text. [Basically, our preliminary results already answered this: no. The quality is impeccable but the effort prohibitive. Using NLP to normalize SPL appears to be another Ph.D. thesis in its own right.] Other questions: How do UMLS or other sources compare to SPL quality-wise? [See the "Preliminary Results" section below.] Are there better ways to linkout from RxNav to SPL than the current right-click link based on a potentially inaccurate mapping of unique identifiers?

**Drugs@FDA.** (<http://www.fda.gov/cder/drugsatfda/datafiles/default.htm>) The poor match between SPL's and UMLS' indications (Table 7) and contraindications (Table 8) raises the urgency level of finding another source that can provide this information accurately and in a terminologically normalized way. ChemIDplus (see below) alerted me to two possibilities, Drugs@FDA and DrugDigest. This download site of Drugs@FDA offers nine well-structured tables covering 23,465 FDA-approved and -submitted drugs (1,689 unique single "activeingredient" values). The table descriptions tantalizingly include a field named "THER\_POTENTIAL" but I could not find any specific drug-indication relationships in the data. However, there may be useful links to NDA (New Drug Application) numbers of interest to some users.

**DrugDigest.** (<http://www.drugdigest.org/DD/Home>) "DrugDigest is the consumer health and drug information website of Express Scripts, Inc. (ESI), the nation's largest independent pharmacy benefit manager (PBM)." So there would be licensing issues. But the database clearly has enough structure (in addition to natural language text) to provide what we are looking for, at least at a consumer level of knowledge. Clicking on "Conditions & Treatments" > "Health Conditions" brings up a list of 71 indications such as Acne, Allergy, Alzheimer's Disease,

Anxiety, Arthritis, Asthma, Atrial Fibrillation, Attention Deficit- Hyperactivity Disorder (ADHD), Bacterial Infection, Benign Prostatic Hyperplasia (BPH), Bipolar Disorder, Breast Cancer, Cancer, ..., each hyperlinked to a text description with a "How is it treated" link that can be followed to one or more "Drug Classes" links and thence to drug names. For example, *Benign Prostatic Hyperplasia (BPH) > 5-Alpha Reductase Inhibitors > Finasteride (Proscar)* [note that *Finasteride* and *Proscar* are two separate links]. There is also a "database of more than 5,000 drugs and herbals and 11,500 potential interactions." Side effects and contraindications appear as bulleted free text suggesting possible terminological normalization; for example:

*What side effects may I notice from taking finasteride?*

- *breast enlargement or tenderness*
- *skin rash*
- *sexual difficulties (less sexual desire or ability to get an erection)*
- *small amount of semen released during sex*

*What should I tell my health care provider before I take this medicine?*

- *if you are female (finasteride is not for use in women)*
- *kidney disease or infection*
- *liver disease*
- *prostate cancer*
- *an unusual or allergic reaction to finasteride, other medicines, foods, dyes, or preservatives*

If the licensing issues prove prohibitive, perhaps ESI would at least permit us to hyperlink from an RxNav display page to their corresponding generic or trade name page following this URL pattern:

<http://www.drugdigest.org/DD/DVH/Uses/0,3915,262|Finasteride,00.html>

**MedlinePlus/MedMaster.** (<http://www.nlm.nih.gov/medlineplus/druginformation.html> )

MedMaster contains consumer information similar to DrugDigest's, but with less structure (one can only search on drug names; there are no other hyperlinks). So, in addition to licensing issues (the information is copyrighted by the American Society of Health-System Pharmacists), MedMaster would, like SPL, require extensive text mining to parse and normalize. Neither MedMaster nor DrugDigest list the finasteride drug Propecia, nor does DrugDigest list baldness or alopecia as an indication; as suspected, these consumer-oriented services could have a significant coverage gap relative to RXNORM and our other potential sources. Nevertheless, NLM is adding a right-click linkout to MedMaster to development versions of RxNav.

**WHO-ATC.** (<http://www.whooc.no/atcddd/> ) This is the widely used (by pharma)

World Health Organization Anatomic-Therapeutic-Chemical drug classification ontology.

Despite its ontological imperfections, there is interest in having it integrated with other chemical, biological, and medical knowledge for use by basic researchers (Castle et al., 2007; Yildirim et al., 2007) in addition to its traditional use by the pharma clinical/regulatory sector. The 2005 edition has 4,068 specific drug-by-class entries equivalent to about 2,800 unique, specific, single-ingredient generic names, the remainder being about equally split between duplicates (drugs that map to multiple classes) and other entries. These "other entries" warrant a closer look as they include specific combination and other terms which could be usefully mapped onto RXNORM, although many of them are out-of-scope (e.g., "Zinc bandages"). Even the "duplicates" may correspond to different doses or dosage forms of the same drug. In January I will have access to Merck's copy of the 2008 edition for pilot work. The full-price subscription is 160 euros (\$233), reduced by half "after July" for the January release, so cost and licensing for WHO-ATC are not problematic. WHO-ATC will be our first example of direct, UMLS-like, terminologically normalized integration of a non-UMLS source into RxNav.

**WHO-DRUG.** Properly referred to as the *WHO Drug Dictionary Enhanced*

(<http://www.umc-products.com/DynPage.aspx?id=2829&mn=1107> ), this is WHO's

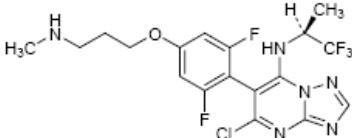
comprehensive but flawed and expensive cross-mapping of generic names ("Ingredients" in RXNORM), trade names, combination products, and ATC codes. For our purposes we only need the drug-ATC code relations of which WHO-DRUG appears to cover about four times as many as WHO-ATC, which appears to cover only about half of RXNORM's lexicon.<sup>81</sup> Getting an exact ATC-RXNORM coverage number is a research question. Research questions for WHO-DRUG: Can we get a free or cheap research license? Does additional (relative to WHO-ATC) coverage of RXNORM's lexicon justify additional cost and effort? The regular cost (85,000 Swedish kroners = \$13,446 for a single-user, one-shot copy) is prohibitive unless assumed by NLM or negotiated downward by >10x. I inquired on November 5, 2007, and have not yet received a reply.

#### **INN (International Nonproprietary Names).**

(<http://www.who.int/medicines/services/inn/en/index.html>) WHO oversees this effort to give standard generic names in several languages to all approved drugs. The database is claimed to cover more than 8,000 names and is available online for free in the form of 153 "lists" (58 lists of "recommended INN's" and 97 lists of "proposed INN's") in PDF format (<http://www.who.int/druginformation/general/innlists.shtml>). A screenshot of a sample record is shown in Figure 15 (top) where it can be seen that the data elements are the INN in Latin, English, French, and Spanish; the chemical name in English, French, and Spanish; the empirical formula; and the chemical structure graphic. INN's typically follow the British Approved Name (BAN) more often than the USAN, whereas RXNORM presumably follows USAN, so some mapping would be required. The biggest problem, of course, is the PDF format, which loses its structure when copied and pasted as ASCII text (Figure 15 bottom). So the free online access is useless for our purposes and the thinness of the metadata we are interested in (English chemical names only) does not seem to justify the required effort.

---

<sup>81</sup> As of September 7, 2007, WHO-DRUG claimed to contain 189,284 unique names; 1,123,194 different medicinal products (trade names, form and strength subtypes, etc.); and 9,899 different ingredients (generic names).

cevipabulinum	
cevipabulin	5-chloro-6-{2,6-difluoro-4-[3-(methylamino)propoxy]phenyl}- N-[(1S)-1,1,1-trifluoropropan-2-yl][1,2,4]triazolo[1,5-a]pyrimidin- 7-amine
cévipabuline	5-chloro-6-{2,6-difluoro-4-[3-(méthylamino)propoxy]phényl}- N-[(1S)-1,1,1-trifluoropropan-2-yl][1,2,4]triazolo[1,5-a]pyrimidin- 7-amine
cevipabulina	5-cloro-6-à2,6-difluoro-4-[3-(metilamino)propoxi]fenil)- N-[(1S)-1,1,1-trifluoropropan-2-il][1,2,4]triazolo[1,5-a]pirimidin- 7-amina
	C <sub>18</sub> H <sub>18</sub> ClF <sub>5</sub> N <sub>6</sub> O
	

**cevipabulinum**

cevipabulin 5-chloro-6-{2,6-difluoro-4-[3-(methylamino)propoxy]phenyl}-  
N-[(1S)-1,1,1-trifluoropropan-2-yl][1,2,4]triazolo[1,5-a]pyrimidin-  
7-amine

cévipabuline 5-chloro-6-{2,6-difluoro-4-[3-(méthylamino)propoxy]phényl}-  
N-[(1S)-1,1,1-trifluoropropan-2-yl][1,2,4]triazolo[1,5-a]pyrimidin-  
7-amine

cevipabulina 5-cloro-6-à2,6-difluoro-4-[3-(metilamino)propoxi]fenil)-  
N-[(1S)-1,1,1-trifluoropropan-2-il][1,2,4]triazolo[1,5-a]pirimidin-  
7-amina

C<sub>18</sub>H<sub>18</sub>ClF<sub>5</sub>N<sub>6</sub>O

N

N

N

N

HN

Cl

F

O F

HN H<sub>3</sub>C CF<sub>3</sub>

H CH<sub>3</sub>

**Figure 15. Online INN excerpt (top) and resulting copy-and-paste ASCII text format (bottom).**

---



---

### International Pharmacopoeia.

(<http://www.who.int/medicines/publications/pharmacopoeia/en/index.html>) This is another WHO publication which contains drug metadata of interest for our purposes (molecular weight, chemical name, chemical structure graphic, solubility, and therapeutic class ("Category")), in an

*Index Nominum*-like semi-structured monograph format. The drug names and "Category" terms would have to be normalized, and the solubility information would have to be parsed from free text and also normalized. Given these and potential licensing issues (although the standard cost is only \$180 for the CD-ROM), the low coverage (420 drugs) does not justify further consideration of this source.

### **eEphMRA Anatomical Classification of Pharmaceutical Products.**

(<http://www.ephmra.org/main.asp?page=465> November 9, 2007) This lesser known but free and WHO-ATC-like ontology is produced by the European Pharmaceutical Market Research Association and might be a serviceable substitute for WHO-ATC. However, I don't see any actual drug classifications (A-box) on their web site, just the "guidelines" (T-box). It has been mapped to WHO-ATC (EphMRA, 2007) so some combination of EphMRA T-box and WHO A-box is possible, but then what's the point? EphMRA also publishes a dosage form coding system called "New Form Codes" (NFC) (<http://www.ephmra.org/main.asp?page=466> ).

**USP/USAN.** The U.S. Pharmacopeia is the official U.S. drug naming authority, and the *USP Dictionary of USAN and International Drug Names* (USAN = United States Adopted Names) is its continually updated reference available for sale as a hardcopy or online book (<http://www.usp.org/products/dictionary/>) or (maybe) as a custom-formatted database.<sup>82</sup> This is the best source of U.S. generic name links to their corresponding chemical names, chemical structure graphics, and Chemical Abstract Service (CAS) registry numbers, but the "therapeutic claim" (~class) terminology is uncontrolled, almost chaotic. Chemical names and structure graphics are more easily available from MeSH, PubChem, and ChemIDplus, but we might revisit USP/USAN for CAS numbers and/or "lab number" manufacturer codes (details later).

**PubChem.** (<http://pubchem.ncbi.nlm.nih.gov/>) PubChem provides information on the biological activities of small molecules, and has been suggested as a candidate for UMLS and/or

---

<sup>82</sup> At least in 2001 Merck was able to negotiate a one-time purchase of the XML file from which the book was printed, and reformat it as a database.

RxNav integration by a Merck bioinformatics research group (J. Castle, personal communication, 2007). It is already owned by NLM but poorly structured (O. Bodenreider, personal communication, 2007). This information is readily extractable without NLP from the XML download format at the :PubChem FTP site. In addition, PubChem online links each compound to related "Substance"; "Substance category"; "Synonyms"; "Structure search / Related Structures"; "Bioassay / Assays"; "Protein3D"; "Rule of 5"; "Literature"; and "Other [non-NLM] Links" (e.g., Ingenuity Pathways Analysis, which is used extensively at Merck). All of this information is potentially interesting to pharma basic researchers for reasons discussed below under the SPL-UMLS comparison ("*Quality assessment (2)*"). However, the volume and complexity of it are daunting, so perhaps we can settle for simply linking to it rather than extracting and replicating it (with or without terminological normalization). Linking from an RXNORM "Ingredient" to the PubChem search result page is trivial, following the pattern

<http://www.ncbi.nlm.nih.gov/sites/entrez?cmd=search&db=pccompound&term=Finasteride>

To link directly to the "Compound" data page, we would need a map of RXNORM Ingredients to their corresponding PubChem CID's. Then the link becomes

<http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=57363>

**ChemIDplus.** This is a free, open-access web service from NLM Specialized Information Services (SIS) (<http://sis.nlm.nih.gov/>). There are two versions, Advanced and Lite (<http://chem.sis.nlm.nih.gov/chemidplus/chemidlite.jsp>) From either, the search result page has a "Full Record" option which follows the URL pattern

[http://chem.sis.nlm.nih.gov/chemidplus/ProxyServlet?objectHandle=Search&actionHandle=getAll3DMViewFiles&nextPage=jsp%2Fcommon%2FChemFull.jsp%3FcalledFrom%3Dlite&chemid=098319267&formatType=\\_3D](http://chem.sis.nlm.nih.gov/chemidplus/ProxyServlet?objectHandle=Search&actionHandle=getAll3DMViewFiles&nextPage=jsp%2Fcommon%2FChemFull.jsp%3FcalledFrom%3Dlite&chemid=098319267&formatType=_3D) where [098319267](#) is a reformatting of an RXNORM

Ingredient's Chemical Abstract Service (CAS) registry number, in this case finasteride's, 98319-26-7. Mappings of generic names to CAS numbers are readily available from MeSH and elsewhere. One cannot link directly to this URL as currently engineered, but presumably that

could be negotiated between the RxNav and SIS teams at NLM. ChemIDplus' structure similarity search is on a different page. Assuming both cover RXNORM Ingredients equally, I found ChemIDplus to be roughly equal to PubChem content-wise but much easier to use. So the ChemIDplus Full Record is a third candidate (in addition to the SPL label and the PubChem Compound page) for RXNORM Ingredient-specific hyperlinkage on RxNav.

**DrugBank.** (<http://redpoll.pharmacy.ualberta.ca/drugbank/>) This is a free, open-access database of biological data about drugs provided by the University of Alberta. It was used by both Castle et al. (2007) and Yildirim et al. (2007). Research questions: How does it compare to UMLS and PubChem for target and pathway coverage? What other kinds of drug information does it have and how well-structured is it for integration with RXNORM? DrugBank appears to be the ultimate "top of the hourglass" trying to cover *all* medical (SPL-like), chemical (PubChem- and ChemIDplus-like), and biological concepts related to a given drug, including a structure similarity search and outlinks to numerous other public databases. So it is certainly equally deserving of at least a hyperlink from RxNav. The direct-link URL pattern for the "Ingredient" search result is [http://redpoll.pharmacy.ualberta.ca/drugbank/cgi-](http://redpoll.pharmacy.ualberta.ca/drugbank/cgi-bin/webglimpse.cgi?ID=16&whole=ON&cache=yes&query=Finasteride)

[bin/webglimpse.cgi?ID=16&whole=ON&cache=yes&query=Finasteride](http://redpoll.pharmacy.ualberta.ca/drugbank/cgi-bin/webglimpse.cgi?ID=16&whole=ON&cache=yes&query=Finasteride)

and for the data display is

<http://redpoll.pharmacy.ualberta.ca/drugbank/cgi-bin/getCard.cgi?CARD=APRD00632.txt> where

[APRD00632](#) is the DrugBank accession number for Finasteride, so a map would be needed for the latter. As for extracting specific Ingredient relations to display in an enhanced RxNav, DrugBank has the same problems as SPL, PubChem, and ChemIDplus: volume, complexity, and need for text mining and terminological normalization. The only information which might be worth extracting and linking to "Finasteride" in RXNORM is which, compared to UMLS's is more precise, terminologically robust, and integratable with molecular biology and pathway databases (GenBank, SwissProt, IPA, etc.) used by pharma basic researchers.

**Table 33. DrugBank vs. UMLS structured molecular target data.**

DrugBank

<i>dimension</i>	<i>value</i>
Drug Target 1 Name	5-alpha reductase 1
Drug Target 1 Gene Name	SRD5A1
Drug Target 1 Synonyms	3-oxo-5-alpha-steroid 4-dehydrogenase 1
Drug Target 1 Synonyms	EC 1.3.99.5
Drug Target 1 Synonyms	Steroid 5- alpha-reductase 1
Drug Target 1 Synonyms	SR type 1
Drug Target 1 Synonyms	S5AR

UMLS

<i>source</i>	<i>rel</i>	<i>rela</i>	<i>value 1</i>	<i>value 2</i>
SNOMEDCT	PAR	inverse_isa	Finasteride	5-Alpha reductase inhibitor
NDFRT	RO	mechanism_of_action_of	Finasteride	5-Alpha reductase inhibitor

**NDFRT 2007 data.** (<ftp://ftp1.nci.nih.gov/pub/cacore/EVS/FDA/ndfirt/>) At NLM, the rap against the otherwise very impressive contribution of NDFRT to UMLS's drug knowledge is that it hasn't been updated since 2004. The tables available for download at this ftp site are dated June 2007. Research question: How do these tables compare to 2004 NDFRT data in UMLS with respect to structure/syntax, coverage, and quality? The data are just dictionaries linking abstract codes of the form "N0000000206" to various Mechanism of Action, Physiologic Effect, and Structural Class descriptors. There are no links to drug concepts. The latest update for Mechanism of Action (10/3/2007) contains 361 unique terms, as compared to 181 NDFRT "has\_mechanism\_of\_action" non-drug arguments in UMLS 2007AB. So clearly there has been substantial growth in these terminologies and, in addition, a close analysis shows that 31 UMLS Mechanism of Action terms have been expired or changed. But without the drug concept links these files are useless to us.

**KEGG (Kyoto Encyclopedia of Genes and Genomes).** This is a bioinformatics resource for linking genomes to biological systems and environments ([www.genome.jp/kegg/](http://www.genome.jp/kegg/)) including drugs (O. Bodenreider, personal communication, 2007); to be investigated.

## Appendix E. Normalization

We used an *ad hoc* normalization process that was informed by our past and current experience with the drug information literature ("literary warrant"), resources ("community warrant"), and professional work in biomedical terminology ("user warrant"). Along with addressing our research questions, the goal was to let the resources "speak for themselves" as much as possible and leave open a variety of paths forward, such as facet analysis, choosing a particular preferred target terminology (MeSH, SNOMEDCT, NDFRT, etc.), or employing automation aids such as the Merck autoencoder or NLM's MetaMap Transfer.<sup>83</sup>

**Intensional content.** For an example of intensional content normalization, UMLS/NDFRT's relationship *Other related/may\_be\_treated\_by* was normalized to the dimension *indication - treatment*. This dimension name is based on this line of reasoning.

- Indications are mentioned in the literature as a desirable class of drug information to integrate (literary warrant).
- There is a standard package insert section heading "Indications & Usage" (community warrant).
- Indications are usually dichotomized into *prevention* and *treatment* (community warrant);
- *Approved indications* are a distinct and crucial (medically, legally, marketing-wise, etc.) subset, hence we also inferred the more specific dimensions *indications - treatment - approved* and *indications - prevention - approved* (user warrant).
- However, UMLS/NDFRT does not distinguish approved from other indications (community warrant).

---

<sup>83</sup> <http://mmtx.nlm.nih.gov/>

- Finally, we impose a morphosyntactic preference for dictionary case, singular form, and rotation of subtype qualifiers separated by " - " to yield a pseudo-hierarchy (user warrant).

Other examples of raw relationships and hyperlink/headings which were normalized to *indication* and its subtypes are

- UMLS/NDFRT's *Other related/may\_be\_prevented\_by*
- ClinicalTrials.gov's *Condition*
- DailyMed's *Indications & Usage*
- DrugDigest's *Learn how <this drug> is used to treat:*
- MedMaster's *Why is this medication prescribed?*
- MedMaster's *Other uses for this medicine*

**Extensional content.** One purpose of normalizing the values is to validate the intensional dimension mappings. That is, for example, do all the values corresponding to ClinicalTrial.gov's *Condition* relationship represent concepts that can be considered indications? Our results suggests that they are. In future extensions of our work, if this pattern continues to hold for more drugs, it reinforces our confidence in the source-relationship-dimension mapping. Normalizing the value names facilitates their semantic typing by manual review and/or automated classification by, for example, UMLS semantic types (e.g., *Disease or Syndrome*, a kind of indication) or MeSH tree headings (e.g., *Diseases*).

Another purpose of normalizing the values is to support search, retrieval, and pooling of data. We want to be able to query the database across resources for drugs that have certain values for certain dimensions, or combinations of dimension-value sets, regardless of how the values (as well as dimensions) are expressed in the raw source data.

The third reason for normalizing the values is to extract dimensional information not given by nonspecific relationships/headings such as

- UMLS/MeSH's *isa*
- DailyMed's *Description*
- DrugDigest's *What is/are <this drug>?*

We know that these are nonspecific because of their lexical makeup and the fact that under them we have found a mixture of different dimensional types of values. In these cases, the relationship/heading provides inadequate clues for mapping the value to a dimension since. Normalization facilitates semantic typing, which may suffice. However, the same value/semantic type may map to multiple, very semantically different dimensions. Medical conditions, for example, can be *indications*, *contraindications*, or *side effects*; enzymes can be *molecular targets*, *metabolism - enzymes*, or (with "deficiency") medical conditions; "500 mg" can be a unit dose (amount of drug in the pill), daily dose, twice daily dose, lethal dose, etc. In such cases a "value-dimension clue" must also be extracted and normalized, as discussed in the Methods section (see also Table 6).

**Dimension-value mismatches.** What if a value is found under the "wrong" dimension in the raw source data? For example, the KEGG DRUG *Target* value for finasteride is "5alpha-reductase inhibitor [KO:K00250] [EC:1.3.99.5] [PATH:map00120] [PATH:map00150]." *Target* intensionally normalizes to *molecular target* but "5alpha-reductase inhibitor" normalizes to "5-alpha reductase inhibitor", finasteride's canonical *therapeutic class*. Should we normalize the value to "5-alpha reductase" and map it to *molecular target*, "5-alpha reductase inhibitor" and map it to *therapeutic class*, or both? Such semantic type fittings are reminiscent of the *range* concept in formal ontologies. Sharp et al. (2008) characterized this issue as a lack of independence among the dimensions.<sup>84</sup>

---

<sup>84</sup> "Some of the dimensions are not independent from each other. In particular, a drug's *therapeutic class*, *molecular target*, *mechanism of action*, and *biological effect* can often be inferred from a single one of them. For example, a drug whose *therapeutic class* is '5-alpha reductase inhibitor' has '5-alpha reductase' as its *molecular target* and '5-alpha reductase inhibition' as its *mechanism of action*" (Sharp et al., 2008, p. 664).

**Consistency.** Excel's built-in sort and string-matching functions were used to identify and correct normalization inconsistencies in the database. For example, since the data was initially loaded in "drug order," a given resource's dimension mappings were widely separated in time and space. By sorting on resource (column C), raw dimension (F), and normalized dimension (O), one may pull like dimension mappings together for examination across drugs. Such examinations may use string-matching formulae in a separate (dummy) column to highlight inconsistencies. Similarly, the consistency of raw value (H) clue parsing (N) and normalization (Q) can be checked across resources and drugs. Consistent clue annotation (J, K, L, N) and normalization (O, Q) are also amenable to this approach.

## Appendix F. Use Case Adaptation and Query Execution

Excel is convenient for prototyping because of its flexibility and transparency (no programming or formal query language needed), but it is not a particularly user-friendly database technology. We leave better database technology to future extensions. Following are examples of the use case adaptations and Excel query simulations underlying some of our research findings.

**Health Use Case A.** Where  $X$  is a dimension not covered by RXNORM,

- (1) for a given value of *generic name* find alternate values of  $X$ ,
- (2) for a given value of  $X$  find alternate values of *generic name*.

Example: (1) "Find all indications for finasteride" → *generic name* = "finasteride";  $X$  = "clinical - indication ...".

1. Sort on column B (normalized generic name) and column O (normalized dimension).
2. Search column B for "finasteride" and then column O for "clinical - indication".
3. Copy all the rows that have "finasteride" in column B and "clinical - indication ..." (... = any subtype) in column O (N=101) to a scratch worksheet.
4. The normalized indication values will be given in column Q of the scratch worksheet, and their intra-generic linkages, if any, in columns R and S.<sup>85</sup>

To refine the results to approved indications:

5. In the scratch sheet cell T1 write a formula to code for the occurrence of the string "approved" in column O: '=search("approved",O1)' without the single quotes.
6. Copy cell T1 to cells T2:T101.
7. Sort the scratch sheet on column T. This will bring all the "approved"/O hits (N=13) to the top.

**Health Use Case B.** The converse query -- (2) "Find all generics indicated for BPH" →  $X$  = "clinical - indication ..."; "value of  $X$ " = "benign prostatic hyperplasia" -- would be executed

---

<sup>85</sup> [http://comminfo.rutgers.edu/~msharp/XKB/Health\\_usecaseA.xls](http://comminfo.rutgers.edu/~msharp/XKB/Health_usecaseA.xls)

identically only substituting column Q for B and "benign prostatic hyperplasia" for "finasteride".<sup>86</sup>

**Research Use Case A.** A cluster of structurally similar compounds targeting the TACR1 gene product (known to be associated with abnormal pain threshold ) was found that points to the WHO-ATC class "antiemetics and antinauseants", suggesting that TACR1 modulation may produce antinauseant activity, and/or that there is a possible connection between antinauseant activity and abnormal pain threshold (Castle et al., 2007).

### *Clustering.*

1. Sort the database on column O (normalized dimension).
2. Copy all rows with O equal to "chemistry - chemical superclass" or "biology - molecular target" (N=240) to a scratch worksheet.
3. Sort the scratch worksheet on columns B (normalized generic name), O (normalized dimension), and Q (normalized value), and remove duplicates of those triples (remaining N=103).
4. Sort the scratch worksheet on Q. This brings together clusters of values independent of dimension. One such cluster is "quinazoline" corresponding to the *chemistry - chemical superclass* of a set of six normalized generic names (B): doxazosin, doxazosin mesylate, prazosin, prazosin hydrochloride, terazosin, and terazosin hydrochloride.<sup>87</sup> This is our substitution for Castle et al.'s chemical similarity measure.<sup>88</sup>
5. Sort the scratch worksheet on B and O and look up the *biology - molecular target* of these six drugs. In all cases it is an isoform of the alpha1 adrenergic receptor. This is our substitution for the TACR1 gene product.
6. Alternatively, in step 4 one could identify the cluster of alpha1 adrenergic receptor *molecular target* values and trace it through the same set of drugs to the *chemical superclass* quinazoline.

<sup>86</sup> [http://comminfo.rutgers.edu/~msharp/XKB/Health\\_usecaseB.xls](http://comminfo.rutgers.edu/~msharp/XKB/Health_usecaseB.xls)

<sup>87</sup> [http://comminfo.rutgers.edu/~msharp/XKB/Research\\_usecaseA\\_cluster.xls](http://comminfo.rutgers.edu/~msharp/XKB/Research_usecaseA_cluster.xls)

<sup>88</sup> Our database contains numerous representations of each drug's chemical structure under *chemistry - formula* and its sub-dimensions. Perhaps someone with more domain knowledge could devise a way to use this data for better chemical similarity clustering.

***Finding biological associations of the molecular target.***

7. These are given in the database under the normalized dimensions (O) *biology - molecular target - general function, biology - molecular target - specific function, biology - molecular target - GO biological process, biology - molecular target - pathway*. Copy these rows (N=180) to a new scratch worksheet.
8. Sort the scratch worksheet on the linked target value (column S) and remove all rows where S is not equal to one of the alpha1 adrenergic receptor isoforms (remaining N=112).<sup>89</sup>
9. Sort the scratch worksheet on the normalized value (Q) and remove duplicates. This will produce a list of 13 non-trivial target biological correlates. This list is comparable to "abnormal pain threshold" in the example.

***Finding biological associations of the drug.***

10. To simulate Castle et al.'s example using only WHO-ATC, sort the database on normalized dimension (O) and normalized generic name (B) and look up prazosin's, doxazosin's, and terazosin's *clinical - therapeutic class - WHO-ATC 5th level code's*. For prazosin and doxazosin the first five digits are C02CA. The first biological (as opposed to molecular) descriptor in the ATC hierarchy for this code is (C02) "antihypertensives." (This inference requires some WHO-ATC experience but can be discovered within the source (column C) = WHO-ATC records of our database, or on the web by various means.) For terazosin the code is G04CA, whose closest biological ATC descriptor is (G04C) "drugs used in benign prostatic hypertrophy." These two descriptors are comparable to "antiemetics and antinauseants" in the example.
11. A much richer list of these drugs' biological associations can be extracted from our database based on these normalized dimensions ("..." means including sub-dimensions): *biology - biological effect, biology - mechanism of action, clinical - indication ..., clinical - therapeutic class*. Sort the database on O and copy all such records to a new scratch worksheet (N=2430).

---

<sup>89</sup> [http://comminfo.rutgers.edu/~msharp/XKB/Research\\_usecaseA\\_bio\\_target.xls](http://comminfo.rutgers.edu/~msharp/XKB/Research_usecaseA_bio_target.xls)

12. Sort the scratch worksheet on B and delete all rows not containing one of the drugs of interest (prazosin, etc.) (remaining N=802).
13. Sort the scratch worksheet on B,O,Q and remove duplicate triples (remaining N=367).
14. Copy column Q to column V, sort it, and remove duplicates, "<negative>", and NCT numbers. This list of 127 drug biological correlates is comparable to "antinauseant activity" in the example.<sup>90</sup>

**Research Use Case B.** The WHO-ATC class "cardiovascular system" points to a list of cardiovascular drugs whose gene targets map to a smaller list of phenotypes. The highest ranking phenotype is "decreased heart rate" which is consistent with the WHO-ATC class. This suggests that other WHO-ATC → drug → gene target → phenotype mappings might be mined for phenotype:disease hypotheses (Castle et al., 2007).

Our *biology - molecular target* dimension can substitute for "gene targets," but we do not have phenotypes independently mapped to molecular targets. We can simulate this use case in two ways. The first is basically the same as the prior use case, only narrowing the drug biological correlates to *indication* and its sub-dimensions for "disease" and substituting the target biological correlates for "phenotypes". This would produce a set of bioprocess:disease rather than phenotype:disease hypotheses (e.g., "heart rate:cardiovascular" rather than "decreased heart rate:cardiovascular"). The second is to leave "gene targets" out of the loop and substitute the drug biological correlates for "phenotypes" and *indication* and its sub-dimensions for "disease." This would produce a set of hypotheses which would be closer semantically to phenotype:disease hypotheses (e.g., "decreased heart rate" could be a *biological effect*, *mechanism of action*, or *pathway* as well as a phenotype or *indication*).

**Research Use Case C.** Campillos et al. (2008) extracted specific sets of drugs with common side effects but different WHO-ATC therapeutic classes, and used the drugs' molecular target and chemical structure/similarity values to predict previously unknown shared targets,

---

<sup>90</sup> [http://comminfo.rutgers.edu/~msharp/XKB/Research\\_usecaseA\\_bio\\_drug.xls](http://comminfo.rutgers.edu/~msharp/XKB/Research_usecaseA_bio_drug.xls)

which were tested by *in vitro* and cell assays. The validated shared targets predict novel hypothetical indications and therapeutic classes for existing drugs. For example, a set of nervous system drugs was found to have side effects in common the antiulcer drug rabeprazole. Four of their targets were predicted to bind rabeprazole, and two - the dopamine receptor DRD3 and the serotonin receptor HTR1D - were validated. This suggests that rabeprazole may be therapeutic for the indications of zolmitriptan (migraine), pergolide (Parkinson's disease), and paroxetine and fluoxetine (psychiatric disorders<sup>91</sup>).

***Finding drugs with common side effects but different therapeutic classes.***

1. The relevant mappings are "drugs":normalized generic name (column B) and these dimensions: "side effects":*clinical - precaution - side effect ...*, "therapeutic classes":*clinical - therapeutic class ...*, "molecular target":*biology - molecular target*, "chemical structure/similarity":*chemistry - chemical superclass*, and "indications":*clinical - indication ...*.
2. Sort the database on column O and copy all rows with "*clinical - precaution - side effect ...*" or "*clinical - therapeutic class - WHO-ATC 5th level code*" to a scratch worksheet (N=1277).
3. For this exercise we can pool the salts with their parents since in our sample they always have the same WHO-ATC classes. Highlight column B of the scratch worksheet and delete all ("replace all") occurrences of the following strings: " hydrochloride"; " mesylate"; " acetate" [note the leading blank].
4. Sort the scratch worksheet on B,Q and remove duplicate doubles (remaining N=440).
5. Sort the scratch worksheet on O and bring the 9 *WHO-ATC* rows to the top.
6. Delete the last two digits of all nine *WHO-ATC* values (e.g., "C02CA04" ➔ "C02CA").
7. Pool the two finasteride *WHO-ATC* values in one cell ("D11AX ; G04CB") and delete the other finasteride *WHO-ATC* row.

---

<sup>91</sup> fluoxetine: depression, obsessive-compulsive disorder, some eating disorders, panic attacks, premenstrual dysphoric disorder; paroxetine: depression, panic disorder, social anxiety disorder, obsessive-compulsive disorder, generalized anxiety disorder, posttraumatic stress disorder, premenstrual dysphoric disorder. Source: MedMaster.

8. Change the *WHO-ATC* rows' font color rows to red and copy their column Q contents to column U.
9. Sort the scratch worksheet on B and copy all the *WHO-ATC* codes in column U to the following set of empty U cells down to the cell preceding the next drug's *WHO-ATC* row.
10. Sort the scratch worksheet on O and delete all the *WHO-ATC* rows (remaining N=431).
11. Sort the scratch worksheet on Q,U and flag all adjacent rows with "common side effects but different therapeutic classes" (equal Q but unequal U<sup>92</sup>) (N=214).
12. Copy column B of the flagged rows to column W, sort, and remove duplicates. This will leave 9 "drugs with common side effects but different therapeutic classes"; i.e., all nine parent drugs, and therefore our whole database, constitute such a set of drugs.<sup>93</sup>

***Predicting previously unknown shared targets based on chemical structure/similarity.***

13. Sort the database on column O and copy all rows with *biology - molecular target* or *chemistry - chemical superclass* to a new scratch worksheet (N=240).
14. Pool the salts with their parents as above.
15. Sort the scratch worksheet on B,Q and remove duplicate doubles (remaining N=80).
16. Concatenate multiple Q values pertaining to the same B,O pair and put the concatenated string in the first U cell of each B,O pair.
17. Sort the scratch worksheet on U. This will bring the 16 concatenated values to the top.
18. Sort those 16 rows on O. To a chemist, the eight *chemical superclass* rows contain two chemical structure/similarity clusters: {finasteride, dutasteride} and {prazosin, terazosin, doxazosin}.<sup>94</sup>

---

<sup>92</sup> Enter "=if(or(and(q9=q10,u9<>u10),and(q9=q8,u9<>u8)),0,1)" into cell V9, then copy V9 to V10:V431, then copy and paste-special-values V9:V431, then sort on V.

<sup>93</sup> [http://comminfo.rutgers.edu/~msharp/XKB/Research\\_usecaseC\\_SE\\_TC.xls](http://comminfo.rutgers.edu/~msharp/XKB/Research_usecaseC_SE_TC.xls)

<sup>94</sup> [http://comminfo.rutgers.edu/~msharp/XKB/Research\\_usecaseC\\_MT\\_CS.xls](http://comminfo.rutgers.edu/~msharp/XKB/Research_usecaseC_MT_CS.xls)

19. Comparing the eight *molecular target* concatenated U values for each suggests the following hypothetical ("unknown") targets (e.g., finasteride targets not common to dutasteride would be hypothetical dutasteride targets):

dutasteride: 5-beta reductase; androgen receptor<sup>95</sup>

terazosin: alpha1C adrenergic receptor

20. Like Campillos et al., we do not have the targets mapped directly to indications or therapeutic classes, so we can either go through the targets' known drug ligands, as they did, or substitute the target biological correlates for indications and therapeutic classes.

***Finding novel hypothetical indications and therapeutic classes for existing drugs based on hypothetical targets known drugs.***

21. To do the first option, sort the database on {O,Q,B} and look up the normalized generic name (B) corresponding to *biology - molecular target* (O) = "5-beta reductase", "androgen receptor", or "alpha1C adrenergic receptor" (Q). For "5-beta reductase" and "androgen receptor" it is finasteride; for "alpha1C adrenergic receptor" it is {doxazosin, prazosin}.

22. Copy the *clinical - indication* ... rows to a scratch worksheet, remove duplicate (B,O,Q) tiples, delete the clinical trial IDs, then sort on (B).

23. Separate, pool, and sort the dutasteride and finasteride (B) rows on (Q). Flag any (Q) value for finasteride (B) which is not also a value for dutasteride (B). These are the hypothetical indications for dutasteride resulting from this simulation of Campillos et al. (N=14).

24. Repeat the prior step substituting {terazosin, terazosin hydrochloride} for dutasteride and {doxazosin, prazosin} for finasteride. The results are the hypothetical indications for {terazosin, terazosin hydrochloride} resulting from this simulation of Campillos et al. (N=73).<sup>96</sup>

25. Copy the *clinical - therapeutic class* ... rows from step 21 to a new scratch worksheet, remove duplicate (B,O,Q) tiples, then sort on (B).

---

<sup>95</sup> The nonoverlapping target pair {finasteride: 3-oxo-5-alpha-steroid 4-dehydrogenase 2} and {dutasteride: 5-alpha reductase type II} are synonyms.

<sup>96</sup> [http://comminfo.rutgers.edu/~msharp/XKB/Research\\_usecaseC\\_MT\\_drug\\_ind.xls](http://comminfo.rutgers.edu/~msharp/XKB/Research_usecaseC_MT_drug_ind.xls)

26. Separate, pool, and sort the dutasteride and finasteride (B) rows on (Q). Flag any (Q) value for finasteride (B) which is not also a value for dutasteride (B). These are the hypothetical therapeutic classes for dutasteride resulting from this simulation of Campillos et al. (N=7).

27. Repeat the prior step substituting {terazosin, terazosin hydrochloride} for dutasteride and {doxazosin, doxazosin mesylate, prazosin, prazosin hydrochloride} for finasteride. The results are the hypothetical therapeutic classes for {terazosin, terazosin hydrochloride} resulting from this simulation of Campillos et al. (N=14).<sup>97</sup>

***Finding novel hypothetical indications and therapeutic classes for existing drugs based on hypothetical targets' biological correlates.***

28. To do the second option of #20, sort the database on {O,Q,B} and the target biological correlates as defined in Research Use Case A Step 7 [(O) = {*biology - molecular target - general function, biology - molecular target - specific function, biology - molecular target - GO biological process, biology - molecular target - pathway*}]. Copy all such rows to a new scratch worksheet (N=180).

29. Sort the scratch worksheet on the linked-to value (S) and delete all rows where this value is not one of {5-beta reductase, androgen receptor, alpha1C adrenergic receptor} (N=14 remaining; 14 for androgen receptor and 0 for alpha1C adrenergic receptor).

30. All 14 of the androgen receptor correlates are mapped only to finasteride, so they constitute hypothetical new drug biological correlates for dutasteride according to this model. One is null ("*<no data>*"), leaving 13.

31. The corresponding indications and therapeutic classes would be obtained by mapping these 13 values (Q) to their normalized generic names (B) in the whole database, and then the resulting normalized generic names to their *clinical - indication ...* and *clinical - therapeutic class ...* (O)

---

<sup>97</sup> [http://comminfo.rutgers.edu/~msharp/XKB/Research\\_usecaseC\\_MT\\_drug\\_TC.xls](http://comminfo.rutgers.edu/~msharp/XKB/Research_usecaseC_MT_drug_TC.xls)

values (Q). In our database, the only (B) hit is finasteride, so the results are the same as in the first approach (#23 and #26).<sup>98</sup>

**Research Use Case D.** A researcher wonders if any existing drugs might be "repurposed" (Boguski et al., 2009) to prevent prostate cancer. She searches ClinicalTrials.gov and gets a list of clinical trials which link the *Condition* "Prostate Cancer" to various *Interventions* including drug names. She thinks this is a good start, but what she really needs is to find other, chemically related drugs and chemicals which are not on this list or already approved for prevention of prostate cancer.

***Finding drugs in clinical trials on prostate cancer.***

1. Go to <http://ClinicalTrials.gov> and search for "prostate cancer" without the quotes.
2. Download all 1823 retrieved trials in the default format (N=14,586 txt file lines).
3. Open the txt file with Excel and sort on column A.
4. Remove all rows that do not have the string "Drug: " (remaining N=1356).
5. Copy the data into a new Word document as text only.
6. Replace "Drug: " with "^pDrug: ^t" in all and "|" with "^t" in all.
7. Copy the data into a new Excel worksheet and sort on column A descending. This will bring all the drugs to the top of column B (N=1830). Delete the other rows and columns.
8. Sort and remove duplicates (remaining N=723). This is the desired list of drugs in clinical trials on prostate cancer.
9. Match this list with the column D raw drug names in the database. (This can be done by sorting and visually scanning, among other ways.) The results are Dutasteride [dutasteride], ELIGARD [leuprolide acetate], Finasteride [finasteride], Leuprolide [leuprolide], Leuprolide Acetate [leuprolide acetate], Leuprorelin [leuprolide], LUPRON [leuprolide acetate], and Tamsulosin [tamsulosin]. Reduced to normalized generic parent names (B): dutasteride, finasteride; leuprolide, and tamsulosin.

---

<sup>98</sup> [http://comminfo.rutgers.edu/~msharp/XKB/Research\\_usecaseC\\_MT\\_targ.xls](http://comminfo.rutgers.edu/~msharp/XKB/Research_usecaseC_MT_targ.xls)

***Finding chemical characteristics of drugs in clinical trials on prostate cancer.***

10. Sort the database on (O,B) and look up the values corresponding to (B) = {dutasteride, finasteride; leuprolide, tamsulosin} and informative<sup>99</sup> *chemistry ... dimensions*; i.e., (O) = {*chemistry - chemical complexity, chemistry - chemical superclass, chemistry - heavy atom count, chemistry - Lipinski ..., chemistry - physical properties - melting point, chemistry - polarity - TPSA, chemistry - rotatable bond count, chemistry - solubility ..., chemistry - stereocenter count ..., chemistry - tautomer count*}.<sup>100</sup> The most parsimonious resource collection that supplies this data is DrugBank, MeSH/UMLS, and PubChem.

11. Of the four drugs, tamsulosin had the most typical values for these dimensions across all drugs in our database, so, hoping to find other drugs with similar values, we chose to make it our model prostate cancer drug.<sup>101</sup>

***Finding drugs with chemical characteristics of drugs in clinical trials on prostate cancer.***

12. Continuing from #11, find drugs (B) with similar values (Q) to those of tamsulosin for the above *chemistry ... dimensions* (O). Surprisingly, seven out of the nine parent drug compounds in our database qualified; in order of number of closest values to tamsulosin's, finasteride (8), prazosin (7), terazosin (6), doxazosin (4), dutasteride (4), leuprolide (1), and ticlopidine (1).<sup>102</sup>

13. The difference between #10's four drugs and #12's seven (plus tamsulosin) constitutes our retrieval of tamsulosin-like compounds not currently in clinical trials on prostate cancer: prazosin, terazosin, doxazosin, and ticlopidine.

<sup>99</sup> The idea here was to focus on descriptive ("natural") as opposed to nominal values. Therefore other database's ID's, nomenclature, and formulas were not used. It could be argued that the latter also are naturally descriptive, but the requisite drill down, parsing, and clustering challenges exceed our Excel string-matching capabilities. We also eliminated dimensions with predominantly null or homogeneous values (e.g., *charge*; all values = 0) attributable solely to our small drug sample.

<sup>100</sup> [http://comminfo.rutgers.edu/~msharp/XKB/Research\\_usecaseD\\_chem\\_tamsu.xls](http://comminfo.rutgers.edu/~msharp/XKB/Research_usecaseD_chem_tamsu.xls) Sheet1

<sup>101</sup> [http://comminfo.rutgers.edu/~msharp/XKB/Research\\_usecaseD\\_chem\\_tamsu.xls](http://comminfo.rutgers.edu/~msharp/XKB/Research_usecaseD_chem_tamsu.xls) Sheet2

<sup>102</sup> [http://comminfo.rutgers.edu/~msharp/XKB/Research\\_usecaseD\\_chem\\_tamsu.xls](http://comminfo.rutgers.edu/~msharp/XKB/Research_usecaseD_chem_tamsu.xls) Sheet3 column G

14. Compare these results to the "similar compound" searches available on PubChem, ChemIDplus, DrugBank, and KEGG DRUG. Surprisingly, they all retrieved different top hits for tamsulosin. From DrugBank we obtained dofetilide, bumetanide, and piretanide. From KEGG DRUG we obtained amosulalol, formoterol, and isoxsuprine. PubChem's and ChemIDplus' utilities offered no obvious, easy way to filter the results down to such approved drugs comparable to prazosin, terazosin, doxazosin, and ticlopidine.
15. Using the original resources' (DrugBank, UMLS, and PubChem) web interfaces, look up the values for the dimensions given in #10 above for dofetilide, bumetanide, piretanide, amosulalol, formoterol, and isoxsuprine.<sup>103</sup>
16. Compute the nine other compounds' chemical (dis)similarity to tamsulosin as the percent deviation of a given drug's values from the corresponding value for tamsulosin. For example, given the melting points of 227°C for tamsulosin and 250°C for finasteride, the melting point deviation of finasteride is  $|(227-250)/227| = 10\%$ . For each drug, average the deviations over three groups of dimensions: physical behavior (melting point and solubility), chemical complexity (including the Lipinski parameters, polarity, and rotatable bonds), and stereocenter counts. The latter is actually another measure of chemical complexity but has outlier low raw scores (typically 1 or 0) and consequent high deviations (1 vs. 0  $\rightarrow$  100%). Finally, average the three averages for each drug to obtain an overall measure of its similarity to tamsulosin.<sup>104</sup>

<sup>103</sup> [http://comminfo.rutgers.edu/~msharp/XKB/Research\\_usecaseD\\_chem\\_tamsu.xls](http://comminfo.rutgers.edu/~msharp/XKB/Research_usecaseD_chem_tamsu.xls) Sheet3 columns H-M

<sup>104</sup> [http://comminfo.rutgers.edu/~msharp/XKB/Research\\_usecaseD\\_chem\\_tamsu.xls](http://comminfo.rutgers.edu/~msharp/XKB/Research_usecaseD_chem_tamsu.xls) Sheet3 rows 33-60

**Appendix G. Dimensions Found in Experimental Database - 6-Level Hierarchy**

biology - ADME

biology - ADME - absorption

biology - ADME - absorption - AUC

biology - ADME - absorption - AUC - accumulation ratio - first dose

biology - ADME - absorption - AUC - accumulation ratio - steady state

biology - ADME - absorption - bioavailability

biology - ADME - absorption - bioavailability - intravenous

biology - ADME - absorption - bioavailability - oral

biology - ADME - absorption - bioavailability - subcutaneous

biology - ADME - absorption - C(steady state)

biology - ADME - absorption - Cmax

biology - ADME - absorption - fasting

biology - ADME - absorption - food effect

biology - ADME - absorption - food effect - AUC

biology - ADME - absorption - food effect - bioavailability

biology - ADME - absorption - food effect - Cmax

biology - ADME - absorption - food effect - Tmax

biology - ADME - absorption - T(steady state)

biology - ADME - absorption - Tmax

biology - ADME - absorption - Tmax - evening dosing

biology - ADME - absorption - Tmax - morning dosing

biology - ADME - demographic interaction

biology - ADME - demographic interaction - geriatric

biology - ADME - demographic interaction - geriatric - absorption - AUC

biology - ADME - demographic interaction - geriatric - absorption - Cmax

biology - ADME - demographic interaction - hepatic impairment - absorption - AUC

biology - ADME - distribution

biology - ADME - distribution - crosses blood-brain barrier

biology - ADME - distribution - multiple dose accumulation

biology - ADME - distribution - plasma protein binding

biology - ADME - distribution - plasma protein binding - albumin

biology - ADME - distribution - plasma protein binding - alpha-1 acid glycoprotein

biology - ADME - distribution - semen

biology - ADME - distribution - steady-state volume

biology - ADME - dose proportionality

biology - ADME - excretion

biology - ADME - excretion - % in feces

biology - ADME - excretion - % in urine

biology - ADME - excretion - % unchanged drug

biology - ADME - excretion - plasma clearance

biology - ADME - excretion - predominant route

biology - ADME - half-life

biology - ADME - half-life - after repeated dosing

biology - ADME - half-life - elimination

biology - ADME - half-life - elimination - delayed release form

biology - ADME - half-life - elimination - immediate release form

biology - ADME - half-life - initial dose

biology - ADME - half-life - plasma

biology - ADME - half-life - range

biology - ADME - half-life - terminal

biology - ADME - half-life - terminal elimination

biology - ADME - half-life - terminal elimination - steady state

biology - ADME - metabolism

biology - ADME - metabolism - % metabolized

biology - ADME - metabolism - biotransformation

biology - ADME - metabolism - conjugate

biology - ADME - metabolism - enzyme

biology - ADME - metabolism - enzyme - phase 1

biology - ADME - metabolism - enzyme - phase 1 - gene name

biology - ADME - metabolism - enzyme - phase 1 - protein sequence

biology - ADME - metabolism - enzyme - phase 1 - SNP

biology - ADME - metabolism - enzyme - primary

biology - ADME - metabolism - enzyme - secondary

biology - ADME - metabolism - extent

biology - ADME - metabolism - interactions

biology - ADME - metabolism - mean systemic clearance

biology - ADME - metabolism - mechanism

biology - ADME - metabolism - metabolite

biology - ADME - metabolism - metabolite - major

biology - ADME - metabolism - metabolite activity

biology - ADME - metabolism - organ

biology - ADME - metabolism - relative peak metabolite plasma concentration - after repeated dosing

biology - ADME - metabolism - relative peak metabolite plasma concentration - initial dose

biology - ADME - metabolism - time to peak metabolite plasma concentration

biology - ADME - time to baseline after discontinuation

biology - ADME - time to maximal effect

biology - ADME - time to steady state

biology - ADME - time to substantial effect

biology - biological effect

biology - mechanism of action

biology - molecular target

biology - molecular target - cellular location

biology - molecular target - chromosome locus

biology - molecular target - chromosome number

biology - molecular target - essentiality

biology - molecular target - gene name

biology - molecular target - gene sequence

biology - molecular target - gene sequence - length

biology - molecular target - general function

biology - molecular target - GO biological process

biology - molecular target - GO cellular component

biology - molecular target - GO molecular function

biology - molecular target - molecular weight

biology - molecular target - number of residues

biology - molecular target - pathway

biology - molecular target - Pfam domain function

biology - molecular target - Pfam domain function code

biology - molecular target - protein sequence

biology - molecular target - reaction

biology - molecular target - signal

biology - molecular target - SNP

biology - molecular target - specific function

biology - molecular target - structure

biology - molecular target - structure - 3D

biology - molecular target - synonym

biology - molecular target - theoretical pI

biology - molecular target - transmembrane region

biology - molecular target - UniProtKB/Swiss-Prot name

biology - organism affected

biology - pathway

biology - toxicity

biology - toxicity - carcinogenicity

biology - toxicity - CNS

biology - toxicity - developmental

biology - toxicity - LD50 - intramuscular - mouse

biology - toxicity - LD50 - intramuscular - rat

biology - toxicity - LD50 - intraperitoneal - monkey

biology - toxicity - LD50 - intraperitoneal - mouse

biology - toxicity - LD50 - intraperitoneal - rat

biology - toxicity - LD50 - intravenous - mouse

biology - toxicity - LD50 - intravenous - rat

biology - toxicity - LD50 - oral - dog

biology - toxicity - LD50 - oral - monkey

biology - toxicity - LD50 - oral - mouse

biology - toxicity - LD50 - oral - mouse/rat

biology - toxicity - LD50 - oral - rat

biology - toxicity - LD50 - oral - rat/mouse

biology - toxicity - LD50 - rat

biology - toxicity - LD50 - subcutaneous - mouse

biology - toxicity - LD50 - subcutaneous - rat

biology - toxicity - LDLo - oral - human - man

biology - toxicity - LDLo - oral - mouse

biology - toxicity - lethal dose - oral - mouse

biology - toxicity - lethal dose - oral - rat

biology - toxicity - mutagenicity

biology - toxicity - reproductive

biology - toxicity - TDLo - oral - human

biology - toxicity - TDLo - oral - human - man

biology - toxicity - TDLo - oral - human - woman

biology - toxicity - toxic effect - intramuscular - mouse

biology - toxicity - toxic effect - intramuscular - rat

biology - toxicity - toxic effect - intraperitoneal - monkey

biology - toxicity - toxic effect - intraperitoneal - mouse

biology - toxicity - toxic effect - intraperitoneal - rat

biology - toxicity - toxic effect - intravenous - mouse

biology - toxicity - toxic effect - intravenous - rat

biology - toxicity - toxic effect - oral - dog

biology - toxicity - toxic effect - oral - human

biology - toxicity - toxic effect - oral - human - man

biology - toxicity - toxic effect - oral - human - woman

biology - toxicity - toxic effect - oral - monkey

biology - toxicity - toxic effect - oral - mouse

biology - toxicity - toxic effect - oral - mouse/rat

biology - toxicity - toxic effect - oral - rat

biology - toxicity - toxic effect - overdose symptom

biology - toxicity - toxic effect - subcutaneous - mouse

biology - toxicity - toxic effect - subcutaneous - rat

chemistry - atmospheric OH rate constant

chemistry - charge

chemistry - chemical class

chemistry - chemical complexity

chemistry - chemical name

chemistry - chemical name - CAS type 1

chemistry - chemical name - derivative

chemistry - chemical name - IUPAC

chemistry - chemical superclass

chemistry - chemical type

chemistry - covalently bonded unit count

chemistry - formula - amino acid sequence

chemistry - formula - empirical formula

chemistry - formula - InChI

chemistry - formula - InChIKey

chemistry - formula - SMILES

chemistry - formula - SMILES - canonical

chemistry - formula - SMILES - isomeric

chemistry - formula - structural formula

chemistry - formula - structural formula - 2D

chemistry - formula - structural formula - 3D

chemistry - formula - structural formula - Jmol

chemistry - formula - structural formula - KCF file

chemistry - formula - structural formula - KEGGdraw

chemistry - formula - structural formula - MOL file

chemistry - formula - structural formula - SDF file

chemistry - formula - structural formula - similar structure search

chemistry - heavy atom count

chemistry - Henry's law constant

chemistry - isoelectric point

chemistry - isotope atom count

chemistry - Lipinski - H bond acceptor

chemistry - Lipinski - H bond donor

chemistry - Lipinski - molecular weight

chemistry - Lipinski - molecular weight - average

chemistry - Lipinski - molecular weight - exact mass

chemistry - Lipinski - molecular weight - monoisotopic

chemistry - Lipinski - solubility logP octanol-water

chemistry - physical properties - melting point

chemistry - physical properties - physical state

chemistry - pKa

chemistry - polarity - TPSA

chemistry - related chemical - broader

chemistry - rotatable bond count

chemistry - solubility

chemistry - solubility - Caco2 permeability - experimental

chemistry - solubility - logP - predicted

chemistry - solubility - logP hydrophobicity - experimental

chemistry - solubility - logP octanol-water

chemistry - solubility - logS - experimental  
chemistry - solubility - logS - predicted  
chemistry - solubility - water  
chemistry - solubility - water - experimental  
chemistry - solubility - water - predicted  
chemistry - stereocenter count - defined atom  
chemistry - stereocenter count - defined bond  
chemistry - stereocenter count - undefined atom  
chemistry - stereocenter count - undefined bond  
chemistry - tautomer count  
chemistry - vapor pressure  
clinical  
clinical - clinical trial comparison therapy  
clinical - clinical trial co-therapy  
clinical - indication  
clinical - indication - clinical trial condition  
clinical - indication - herbal evidence  
clinical - indication - herbal evidence grade A  
clinical - indication - herbal evidence grade C  
clinical - indication - herbal evidence methodology  
clinical - indication - herbal summary  
clinical - indication - herbal untested  
clinical - indication - patient selection criteria  
clinical - indication - patient type  
clinical - indication - prevention  
clinical - indication - prevention - approved

clinical - indication - prevention - approved - combo

clinical - indication - prevention - clinical trial condition

clinical - indication - treatment

clinical - indication - treatment - approved

clinical - indication - treatment - approved - combination

clinical - indication - treatment - clinical trial condition

clinical - lab test - drug level

clinical - lab test - drug level in blood/serum/plasma

clinical - lab test - drug level in urine

clinical - precaution

clinical - precaution - contraindication

clinical - precaution - disease history

clinical - precaution - drug interaction

clinical - precaution - food interaction

clinical - precaution - food interaction - administration with food

clinical - precaution - food interaction - diet

clinical - precaution - GI retention time

clinical - precaution - handling

clinical - precaution - herbal

clinical - precaution - herbal/supplement interaction

clinical - precaution - in case of overdose

clinical - precaution - indication specification

clinical - precaution - lab test interaction

clinical - precaution - lab test interference

clinical - precaution - lab test monitoring

clinical - precaution - side effect

clinical - precaution - side effect - common  
clinical - precaution - side effect - major  
clinical - precaution - side effect - minor  
clinical - precaution - unproven indication  
clinical - precaution - warning  
clinical - precaution - warning - boxed  
clinical - storage conditions  
clinical - therapeutic class  
clinical - therapeutic class - body system  
clinical - therapeutic class - herbal mild property  
clinical - therapeutic class - historical  
clinical - therapeutic class - organism  
pharmacy - administration  
pharmacy - administration - frequency  
pharmacy - administration - route  
pharmacy - approval info - approval status  
pharmacy - approval info - company  
pharmacy - approval info - company - distributor country  
pharmacy - approval info - company - distributor name  
pharmacy - approval info - company - manufacturer country  
pharmacy - approval info - company - manufacturer name  
pharmacy - approval info - country  
pharmacy - approval info - FDA approval date  
pharmacy - approval info - FDA chemical type  
pharmacy - approval info - FDA drug type  
pharmacy - approval info - FDA review classification

pharmacy - approval info - marketing status - FDA

pharmacy - approval info - RLD

pharmacy - approval info - TE code

pharmacy - DEA schedule

pharmacy - dosage form

pharmacy - dosage form - dilution

pharmacy - dose - daily total

pharmacy - dose - dosing regimen

pharmacy - dose - dosing regimen - indication-specific

pharmacy - dose - dosing regimen - initial

pharmacy - dose - dosing regimen - maintenance - total daily

pharmacy - dose - dosing regimen - maximum

pharmacy - dose - dosing regimen - monotherapy

pharmacy - dose - dosing regimen - restart

pharmacy - dose - unit dose

pharmacy - dose - unit dose - by unit

pharmacy - dose - unit dose - by volume

pharmacy - dose - unit dose - free acid/base equivalent

pharmacy - dose - unit dose - herbal untested

pharmacy - dose - unit dose - herbal untested - daily total

pharmacy - drug type

pharmacy - generic availability

pharmacy - generic name

pharmacy - generic name - abbreviation

pharmacy - generic name - combination chemotherapy

pharmacy - generic name - combination product

pharmacy - generic name - derivative

pharmacy - generic name - derivative - tritiated

pharmacy - generic name - free acid/base

pharmacy - generic name - free acid/base - combination product

pharmacy - generic name - free acid/base - isomer

pharmacy - generic name - herbal physical form

pharmacy - generic name - herbal physical form - combination product

pharmacy - generic name - herbal synonym

pharmacy - generic name - herbal synonym - Danish

pharmacy - generic name - herbal synonym - French

pharmacy - generic name - herbal synonym - German

pharmacy - generic name - herbal systematic name - family

pharmacy - generic name - herbal systematic name - genus

pharmacy - generic name - herbal systematic name - species

pharmacy - generic name - herbal systematic name - species - misspelling

pharmacy - generic name - herbal systematic name - species - synonym

pharmacy - generic name - herbal systematic name - subspecies

pharmacy - generic name - hydrate

pharmacy - generic name - hydrate - synonym

pharmacy - generic name - INN

pharmacy - generic name - INN/BAN

pharmacy - generic name - INN/English

pharmacy - generic name - INN/French

pharmacy - generic name - INN/Latin

pharmacy - generic name - INN/Spanish

pharmacy - generic name - isomer

pharmacy - generic name - JAN

pharmacy - generic name - JAN/USP

pharmacy - generic name - JP15/USAN

pharmacy - generic name - misspelling

pharmacy - generic name - salt

pharmacy - generic name - salt - abbreviation

pharmacy - generic name - salt - abbreviation - misspelling

pharmacy - generic name - salt - French

pharmacy - generic name - salt - hydrate - synonym

pharmacy - generic name - salt - misspelling

pharmacy - generic name - Spanish

pharmacy - generic name - synonym

pharmacy - generic name - USAN

pharmacy - generic name - USAN/INN/BAN

pharmacy - generic name - USAN/JAN

pharmacy - generic name - USAN/JP15

pharmacy - generic name - USP/INN

pharmacy - generic name - word order variant

pharmacy - herbal source biology

pharmacy - inactive ingredient

pharmacy - lexical class

pharmacy - manufacturer code

pharmacy - manufacturer code - derivative

pharmacy - packaging

pharmacy - packaging - NDC package description

pharmacy - product type

pharmacy - storage conditions

pharmacy - trade name

pharmacy - trade name - combination chemotherapy

pharmacy - trade name - combination product

pharmacy - trade name - derivative

pharmacy - unit appearance

pharmacy - unit appearance - coating

pharmacy - unit appearance - color

pharmacy - unit appearance - imprint code

pharmacy - unit appearance - score

pharmacy - unit appearance - shape

pharmacy - unit appearance - size

pharmacy - unit appearance - symbol

## References

- Aitchison, J., & Clarke, S. D. (2006). The thesaurus: A historical viewpoint, with a look to the future. *Cataloging & Classification Quarterly*, 37(3), 5-21.
- Allen, J. (1995). *Natural language understanding, second edition*. Redwood City, CA: Benjamin/Cummings.
- Aronson, J. K., & Ferner, R. E. (2005). Clarification of terminology in drug safety. *Drug Safety*, 28, 851-870.
- Bard, J. L., & Rhee, S. Y. (2004). Ontologies in biology: Design, applications and future challenges. *Nature Review Genetics*, 5, 213-222.
- Bawden, D., & Robinson, L. (2010). Pharmaceutical information: A thirty year perspective on the literature. *Annual Review of Information Science and Technology*, 45, 63-119.
- Belkin, N. J. (1978). Information concepts for information science. *Journal of Documentation*, 34, 55-85
- Belkin, N. J. (1993). Interaction with texts: Information retrieval as information-seeking behavior. In G. Knorz, J. Krause, & C. Womser-Hacker (Eds.), *Information Retrieval '93. Von der Modellierung zur Anwendung* (pp. 55-66). Konstanz: Universitätsverlag Konstanz.
- Ben-Miled, Z., Webster, Y. W., Li, N., Bukhres, O., Nayar, A. K., Martin, J., & Oppelt, R. (2002). BAO, a biological and chemical ontology for information integration. *Online Journal of Bioinformatics*, 1, 60-73.
- Berners-Lee, T. (2006). Linked data. Retrieved from <http://www.w3.org/DesignIssues/LinkedData.html>
- Berners-Lee, T., Hendler, J., & Lasilla, O. (2001, May). The semantic web. *Scientific American*, 284(5), 34-43.
- Bodenreider, O. (2009, June 15). Challenges and promises of the Semantic Web in health care and life sciences [PowerPoint slides]. Paper presented at the Special Library Association Annual Conference, Biomedical & Life Sciences Division, Washington, DC. Retrieved from <http://mor.nlm.nih.gov/pubs/pres/090615-SLA.pdf>
- Bodenreider, O., & Nelson, S. J. (2004). RxNav: A semantic navigation tool for clinical drugs. In M. Fieschi et al. (Eds.), *MEDINFO 2004* (p. 1530). Amsterdam: IOS Press.
- Bodenreider, O., & Stevens R. (2006). Bio-ontologies: current trends and future directions. *Briefings in Bioinformatics*, 7, 256-274.
- Bodenreider, O., Smith, B., & Burgun, A. (2004). The ontology-epistemology divide: a case study in medical terminology. In: A. Varzi, & L. Vieu (Eds.), *Proceedings of the international conference on Formal Ontology in Information Systems* (pp. 185-195). Amsterdam: IOS Press.

Bodenreider, O., Smith, B., Kumar, A., & Burgun, A. (2007). Investigating subsumption in SNOMED CT: An exploration into large description logic-based biomedical terminologies. *Artificial Intelligence in Medicine*, 39, 183-195.

Boguski, M. S., Mandl, K. D., & Sukhatme, V. P. (2009). Repurposing with a difference. *Science*, 324, 1394-1395.

Boman, M., Bubenko Jr, J. A., Johannesson, P., & Wangler, B. (1997). *Conceptual modelling*. Upper Saddle River, NJ: Prentice Hall.

Borgida, A., & Brachman, R. J. (2002). Conceptual modelling with description logics. In F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, & P. F. Patel-Schneider (Eds.), *Description logic handbook* (pp. 359-381). Cambridge, England: Cambridge University. Retrieved from <http://www.inf.unibz.it/~franconi/dl/course/dlhb/dlhb-10.pdf>

Bray, T. (2003). Metadata, semantics, and all that. Retrieved from <http://www.tbray.org/ongoing/When/200x/2003/11/09/SemWebFirstStep>

Broughton, V. (2006). The need for a faceted classification as the basis of all methods of information retrieval. *Aslib Proceedings: New Information Perspectives*, 58, 49-72.

Campillos, M., Kuhn, M., Gavin, A.-C., Jensen, L. J., & Bork, P. (2008). Drug target identification using side-effect similarity. *Science*, 321, 263-266.

Carter, J. S., Brown, S. H., Bauer, B. A., Elkin, P. L., Erlbaum, M. S., Froehling, D. A., Lincoln, M. J., Rosenbloom, S. T., Wahner-Roedler, D. L., & Tuttle, M. S. (2006). Categorical information in pharmaceutical terminologies. *American Medical Informatics Association Annual Symposium Proceedings, 2006*, 116-120.

Carter, J. S., Brown, S. H., Erlbaum, M. S., Gregg, W., Elkin, P. L., Speroff, T., & Tuttle, M. S. (2002). Initializing the VA Medication Reference Terminology using UMLS Metathesaurus co-occurrences. *American Medical Informatics Association Annual Symposium Proceedings, 2002*, 116-120.

Castle, J., Shah, J., Avila, I., Derry, J., & Rohl, C. (2007, May). Molecular Informatics: Phenotype/disease, gene, drug, and therapeutic activity connections. Merck Research Laboratories internal monthly highlight report.

Ceusters, W., Smith, B., & Flanagan, J. (2003, May). Ontology and medical terminology: Why description logics are not enough. In *Towards an Electronic Patient Record (TEPR 2003)*, San Antonio, 10-14 May 2003. Boston, MA: Medical Records Institute (CD-ROM publication). Retrieved from <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.91.4053&rep=rep1&type=pdf>

Charlet, J. (2007). The management of medical knowledge: Between non-structured documents and ontologies. *Annals of Telecommunications*, 62, 808-826.

Chen, Y., Perl, Y., Geller, J., & Cimino, J. J. (2007). Analysis of a study of the users, uses, and future agenda of the UMLS. *Journal Of The American Medical Informatics Association*, 14, 221-231.

Cheung, K.-H., Kashyap, V., Luciano, J. S., Chen, H., Wang, Y., & Stephens, S. (2008). Semantic mashup of biomedical data. *Journal of Biomedical Informatics*, 41, 683-686.

Cimino, J., & Zhu, X. (2006). The practical impact of ontologies on biomedical informatics. *International Medical Informatics Association Yearbook of Medical Informatics; Methods of Information in Medicine*, 45, 124-35

Clauson, K. A., Marsh, W. A., Polen, H. H., Seamon, M. J., & Ortiz, B. I. (2007). Clinical decision support tools: Analysis of online drug information databases. *BMC Medical Information Decision Making*, 7, 7-13.

Clauson, K. A., Polen, H. H., & Marsh, W. A. (2007). Clinical decision support tools: performance of personal digital assistant versus online drug information databases. *Pharmacotherapy*, 12, 1651-1658.

de Matos, P., Ennis, M., Degtyarenko, K., Darsow, M., Guedj, M., Hermjakob, H., Rijnbeek, M., Kretschmann, E., Binns, D., & Apweiler, R. (2004, July). ChEBI: A dictionary of chemical compounds. Paper presented at the International Conference on Intelligent Systems for Molecular Biology, Glasgow, Scotland. Retrieved from [http://www.iscb.org/ismb2004/posters/pmatosATebi.ac.uk\\_329.html](http://www.iscb.org/ismb2004/posters/pmatosATebi.ac.uk_329.html)

Degtyarenko, K., De Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcantara, R., Darsow, M., Guedj, M., & Ashburner, M. (2008). ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36, D344-D350.

Dervin, B., & Nilan, M. (1986). Information needs and uses. *Annual Review of Information Science and Technology*, 21, 3-33.

de Waard, A., Fluit, C., & van Harmelen, F. (2007). Use Case: Drug Ontology Project for Elsevier (DOPE). Retrieved from <http://www.w3.org/2001/sw/sweo/public/UseCases/Elsevier/>

Digital Anatomist Project (2004). Interactive atlases. Retrieved from <http://www9.biostr.washington.edu/da.html>

Doyle, L. (1961). Semantic road maps for literature searchers. *Journal of the Association for Computing Machinery*, 8, 223-239.

Doyle, L. B. (1962). Indexing and abstracting by association. Part 1. SP-718/001/00. Santa Monica, CA: System Development Corporation.

Euzenat, J., & Shvaiko, P. (2007). *Ontology matching*. Berlin: Springer-Verlag.

Feldman, H. J., Dumontiera, M., Linga, S., Haider, N., & Hogue, C. W. V. (2005). CO: A chemical ontology for identification of functional groups and semantic comparison of small molecules. *FEBS Letters*, 579, 4685-4691.

Foskett, D. J. (1980). Thesaurus. In A. Kent, H. Lancour, & J. E. Daily (Eds.), *Encyclopaedia of library and information science* (Vol. 30, pp. 416-462). New York: Marcel Dekker.

Franconi, E. (n.d.). Description logics. Logics and ontologies. Retrieved from <http://www.inf.unibz.it/~franconi/dl/course/slides/modelling/modelling.pdf>

- Gardner, S. P. (2005). Ontologies and semantic data integration. *Drug Discovery Today*, 10, 1001-1007.
- Giunchiglia, F., Yatskevich, M., Avesani, P., & Shvaiko, P. (2009). A large dataset for the evaluation of ontology matching. *Knowledge Engineering Review*, 24, 137-157.
- Goble, C., & Wroe, C. (2004). The Montagues and the Capulets. *Comparative and Functional Genomics*, 5, 623-632.
- Golbeck, J., Fragoso, G., Hartel, F., Hendler, J., Oberthaler, J., & Parsia, B. (2003). The National Cancer Institute's thesaurus and ontology. *Journal of Web Semantics*, 1, 75-80.
- Golbreich, C., Zhang, S. M., & Bodenreider, O. (2006). The foundational model of anatomy in OWL: Experience and perspectives. *Journal of Web Semantics*, 4(3), 181-195.
- Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.
- Gruber, T. R. (1993). A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2), 199-220.
- Hahn, U. (2003). Turning informal thesauri into formal ontologies: a feasibility study on biomedical knowledge re-use. *Comparative and Functional Genomics*, 4, 94-97.
- Hahn, U., & Schulz, S. (2004). Building a very large ontology from medical thesauri. In S. Staab & R. Studer (Eds.), *Handbook on ontologies* (pp. 133-150). New York: Springer.
- Hameed, A., Preece, A., & Sleeman, D. (2004). Ontology reconciliation. In S. Staab & R. Studer (Eds.), *Handbook on ontologies* (pp. 231-250). New York: Springer.
- Heja, G., Surjan, G., Lukacsy, G., Pallinger, P., & Gergely, M. (2007). GALEN based formal representation of ICD10. *International Journal of Medical Informatics*, 76, 118-123.
- Hjørland, B. & Albrechtsen, H. (1995). Toward a new horizon in information science: Domain-analysis. *Journal of the American Society for Information Science*, 46, 400-425.
- Hug, H., Dannecker, R., Schindler, R., Bagatto, D., Stephan, A., Wess, R. A., & Gut, J. (2004). Ontology-based knowledge management of troglitazone-induced hepatotoxicity. *Drug Discovery Today*, 9, 948-954.
- Joyce, T., & Needham, R. M. (1958). The thesaurus approach to information retrieval. *American Documentation*, 9, 192-197.
- Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing: An introduction to natural language processing, speech recognition, and computational linguistics*. Upper Saddle River, NJ: Prentice Hall.
- Kantor, P. B. (n.d.). White paper. Retrieved from <http://www.scils.rutgers.edu/~kantor/dlmetric.html>

- Kashyap, V., & Borgida, A. (2003, October). Representing the UMLS semantic network using OWL (or "what's in a semantic web link?"). Paper presented at the Second International Semantic Web Conference, Sanibel Island, Florida. Retrieved from <http://cgsb2.nlm.nih.gov/%7Ekashyap/publications/ISWC%202003.pdf>
- Keselman, A., Logan, R., Arnott Smith, C., Leroy, G., & Zeng-Treitler, Q. (2008). Developing informatics tools and strategies for consumer-centered health communication. *Journal of the American Medical Informatics Association*, 15, 473-483.
- Kupferberg, N., Jones Hartel, L. (2004). Evaluation of five full-text drug databases by pharmacy students, faculty, and librarians: do the groups agree? *Journal of the Medical Library Association*, 92, 66-71.
- Kwasnik, B. (1999). Role of classification in knowledge representation and discovery. *Library Trends*, 48(1), 22-47.
- La Barre, K. (2010). Facet analysis. *Annual Review of Information Science and Technology*, 44, 243-286.
- Lambrix, P., & Tan, H. (2006). SAMBO - A system for aligning and merging biomedical ontologies. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4, 196-206.
- Legg, C. (2006). Ontologies on the Semantic Web. *Annual Review of Information Science and Technology*, 41, 407-451.
- Lenz, R., Beyer, M., & Kuhn, K. A. (2007). Semantic integration in healthcare networks. *International Journal of Medical Informatics*, 76, 201-207.
- Liu, S., Ma, W., Moore, R., Ganesan, V., & Nelson, S. (2005). RXNORM: prescription for electronic drug information exchange. *IT Professional*, 7, 17-23. Retrieved from <http://www.nlm.nih.gov/research/umls/RXNORM/RXNORM.pdf>
- Luhn, H. P. (1961). The automatic derivation of information retrieval encodements from machine-readable text. In A. Kent (Ed.), *Information retrieval and machine translation* (Vol. 3, pp. 1021-1028). New York: Interscience.
- Mabee, P. M., Ashburner, M., Cronk, Q., Gkoutos, G. V., Haendel, M., Segerdell, E., Mungall, C., & Westerfield, M. (2007). Phenotype ontologies: the bridge between genomics and evolution. *Trends in Ecology & Evolution*, 22, 345-350.
- Maron, M. E., & Kuhns, J. L. (1960). On relevance, probabilistic indexing and information retrieval. *Journal of the Association for Computing Machines*, 7, 216-244.
- Mendrick, D. L. (2006). Translational medicine: the discovery of bridging biomarkers using pharmacogenomics. *Pharmacogenomics*, 7, 943-947.
- Meyer, H. F. (2002). Streamlining the research and development pipeline by coupling of information technology and biology. *Drug Information Journal*, 36, 169-178.

- Mika, P., Iosif, V., Sure, Y., & Akkermans, H. (2004). Ontology-based content management in a virtual organization. In S. Staab & R. Studer (Eds.), *Handbook on ontologies* (pp. 455-475). New York: Springer.
- Neumann, E., & Prusak, L. (2007). Knowledge networks in the age of the Semantic Web. *Briefings in Bioinformatics*, 8, 141-149.
- NISO (2005). *ANSI/NISO Z39.19-2005. Guidelines for the construction, format, and management of monolingual controlled vocabularies*. Bethesda, MD: NISO Press. Retrieved from <http://www.niso.org/standards/resources/Z39-19-2005.pdf>
- NLM (2007a). RXNORM. Retrieved from <http://www.nlm.nih.gov/research/umls/RXNORM/index.html>
- NLM (2007b). Unified Medical Language System. Retrieved from <http://www.nlm.nih.gov/research/umls/>
- NLM (2009). Medical subject headings. Retrieved from <http://www.nlm.nih.gov/mesh/>
- Norusis, M. J. (2005). Cluster analysis. In *SPSS 13.0 statistical procedures companion* (pp. 361-391). Upper Saddle River, NJ: Prentice-Hall. Retrieved from [http://www.norusis.com/pdf/SPC\\_v13.pdf](http://www.norusis.com/pdf/SPC_v13.pdf)
- Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M. A., Chute C. G., & Musen, M. A. (2009). BioPortal: Ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, 37, W170-W173.
- O'Neil, M. J. (Ed.) (2006). *The Merck Index - An encyclopedia of chemicals, drugs, and biologicals* (14th ed.). Whitehouse Station, NJ: Merck & Co., Inc.
- Pease, A. (2009). SUMO publications. Retrieved from <http://www.ontologyportal.org/Pubs.html>
- Plovnick, R. M., & Zeng, Q. T. (2004). Reformulation of consumer health queries with professional terminology: A pilot study. *Journal of Medical Internet Research*, 6(3), e27. Retrieved from <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1550613/>
- Powers, J. (2004, March). Convera taxonomy training [series of classroom training sessions]. Vienna, VA.
- Quan, D. (2007). Improving life sciences information retrieval using semantic web technology. *Briefings in Bioinformatics*, 8, 172-182.
- Ranganathan, S. R. (1957). *The five laws of library science*. London: Blunt and Sons, Ltd.
- Rau, L. F. (1988). Conceptual information extraction and retrieval from natural language input. In *RIAIO 88* (pp. 424-437). Paris: Centre des Hautes Etudes Internationales d'Informatique Documentaire, 1997, General Electric, USA.
- Renear, A. H., & Palmer, C. L. (2009). Strategic reading, ontologies, and the future of scientific publishing. *Science*, 325, 828-832.

- Rocchio, J. J. (1966). Document retrieval systems - optimization and evaluation (Doctoral dissertation). Cambridge, MA: Harvard Computational Laboratory.
- Sahoo, S. S., Bodenreider, O., Rutter, J. L., Skinner, K. J., & Sheth, A. P. (2008). An ontology-driven semantic mash-up of gene and biological pathway information: Application to the domain of nicotine dependence. *Journal of Biomedical Informatics*, 41, 752-765.
- Salton, G. (1987). Historical note: The past thirty years in information retrieval. *Journal of the American Society for Information Science*, 39, 375-380.
- Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24, 513-523.
- Salton, G., & Buckley, C. (1990). Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41, 288-297.
- Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 29, 321-343.
- Saracevic, T. (1996). Interactive models in information retrieval (IR): progress, problems, proposal. *Proceedings of the American Society for Information Science*, 33, 3-9.
- Schroeder, M., & Neumann, E. (2006). Semantic web for life sciences. *Journal of Web Semantics*, 4(3), 167-167.
- Scott-Wright, A., Crowell, J., Zeng, Q., Bates, D. W., & Greenes, R. (2006). Analysis of information needs of users of MEDLINEplus, 2002-2003. *American Medical Informatics Association Annual Symposium Proceedings*, 2006, 699-703.
- Sharp, M. (2005, March 8). Extracting a practical drug ontology from the UMLS. Paper presented at the First Semantic Technology Conference, San Francisco, CA. Retrieved from <http://comminfo.rutgers.edu/~msharp/SharpSemtech2005.pdf>
- Sharp, M., Bodenreider, O., & Wacholder, N. (2008). A framework for characterizing drug information sources. *American Medical Informatics Association Annual Symposium Proceedings*, 6, 662-666. Retrieved from <http://mor.nlm.nih.gov/pubsv/alum/2008-sharp.pdf>
- Shera, J. (1970). *Sociological foundations of librarianship*. Bombay: Asia Publishing House.
- Shirky, C. (2003). The Semantic Web, syllogism, and worldview. Retrieved from [http://www.shirky.com/writings/semantic\\_syllogism.html](http://www.shirky.com/writings/semantic_syllogism.html)
- Smith, B., & Welty, C. (2001). Ontology: Towards a new synthesis. In *Formal Ontology in Information Systems* (pp. iii-x). Ogunquit, Maine: ACM Press.
- Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S. A., Scheuermann, R. H., Shah, N., Whetzel, P. L., & Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature Biotechnology*, 25, 1251-1255.

- Soergel, D. (1999). The rise of ontologies or the reinvention of classification. *Journal of the American Society for Information Science*, 50, 1119-1120.
- Soldatova, L. N., & King, R. D. (2005). Are the current ontologies in biology good ontologies? *Nature Biotechnology*, 23, 1095-1098.
- Solomon, W. D., Wroe, C. J., Rector, A. L., Rogers, J. E., Fistein, J. L., & Johnson, P. (1999). A reference terminology for drugs. *American Medical Informatics Association Annual Symposium Proceedings, 1999*, 152-155.
- Soualmia, L., Golbreich, C., & Darmoni, S. (2004). Representing the MeSH in OWL: Towards a semi-automatic migration. In U. Hahn (Ed.), *Proceedings of the KR 2004 Workshop on Formal Biomedical Knowledge Representation* (pp. 81-87). Retrieved from <http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-102/soualmia.pdf>
- Sowa, J. F. (n.d.). Guided tour of ontology. Retrieved from <http://users.bestweb.net/~sowa/ontology/guided.htm>
- Spiteri, L. (1998). A simplified model for facet analysis. *Canadian Journal of Information and Library Science*, 23, 1-30.
- Stevens, R., Wroe, C., Lord, P., & Goble, C. (2004). Ontologies in bioinformatics. In S. Staab & R. Studer (Eds.), *Handbook on ontologies* (pp. 635-657). New York: Springer.
- Swanson, D. (1988). Historical note: Information retrieval and the future of an illusion. *Journal of the American Society for Information Science*, 39, 92-98.
- Sweetman, S. (2007). *Martindale: The complete drug reference* (35th ed.). London: Pharmaceutical Press.
- Taylor, K. R., Essex, J. W., Frey, J. G., Mills, H. R., Hughes, G., & Zaluska, E. J. (2006). The Semantic Grid and chemistry: Experiences with CombeChem. *Journal of Web Semantics*, 4(2), 84-101.
- Tudhope, D., & Binding, C. (2008). Faceted thesauri. *Axiomathes*, 18, 211-222.
- Vickery, B. C. (1997). Ontologies. *Journal of Information Science*, 23, 277-286.
- Williams, J., & Andersen, W. (2003). Bringing ontology to the Gene Ontology. *Comparative and Functional Genomics*, 4, 90-93.
- Wilson, T. D. (1981). On user studies and information needs. *Journal of Documentation*, 37, 3-15.
- Wright, A. G. (2003). Shirky on the Semantic Web. Retrieved from <http://www.agwright.com/blog/archives/000787.html>
- Wyman, P. (1999). *Indexing specialties: Medicine*. Phoenix, AZ: American Society of Indexers. Google Books preview retrieved from <http://books.google.com/books?id=7KfGGH4OemIC>

- Yildirim, M. A., Goh, K.-I., Cusick, M. E., Barabási, A.-L., & Vidal, M. (2007). Drug-target network. *Nature Biotechnology*, 25, 1119-1126. Retrieved from <http://www.nature.com/nbt/journal/v25/n10/pdf/nbt1338.pdf>
- Zeng, K., Bodenreider, O., Kilbourne, J., & Nelson, S. (2007, August). RxNav: Towards an integrated view on drug information [poster]. MEDINFO, Brisbane, Australia. Retrieved from <http://lhncbc.nlm.nih.gov/lhc/docs/published/2007/pub2007046.pdf>
- Zeng, K., Bodenreider, O., Kilbourne, J., & Nelson, S. J. (2006). RxNav: Providing standard drug information [demonstration]. *American Medical Informatics Association Annual Symposium Proceedings, 2006*, 1156. Retrieved from <http://mor.nlm.nih.gov/pubs/pdf/2006-amia-kz-demo.pdf>
- Zeng, Q. T., Crowell, J., Plovnick, R. M., Kim, E., Ngo, L., & Dibble, E. (2006). Assisting consumer health information retrieval with query recommendations. *Journal of the American Medical Informatics Association*, 13, 80-90.
- Zhang, S., Mork, P., Bodenreider, O., & Bernstein, P. A. (2007). Comparing two approaches for aligning representations of anatomy. *Artificial Intelligence in Medicine*, 39, 227-236.

## Curriculum Vitae

Mark E. Sharp

### Education

<i>Dates</i>	<i>College Attended</i>	<i>Subjects Pursued</i>	<i>Degree Earned</i>
1968-1969	Boston University	Philosophy, psychology, biology	
1969-1972	Clark University	Psychology, biology, chemistry	B.A.
1972-1974	Duke University	Biochemistry	M.A.
2000-2011	Rutgers University	Information science	Ph.D.

### Principal Occupations

<i>Dates</i>	<i>Occupation</i>	<i>Position</i>
1976-1986	Research biochemist	Chemist, U.S. National Institutes of Health
1987-1994	Biomedical lexicographer	Technical Information Specialist, U.S. National Institutes of Health
1994-2005	Biomedical lexicographer	Research Information Associate, Merck & Co., Inc.
2005-2011	Biomedical lexicographer	Project Lead, Merck & Co., Inc.

### Publications

Schechter, N. M., Sharp, M. E., Reynolds, J. A., & Tanford, C. (1976). Erythrocyte spectrin: purification in deoxycholate and preliminary characterization. *Biochemistry*, 15, 1897-1899.

Marx, S. J., Spiegel, A. M., Sharp, M. E., Brown, E. M., Downs, R. W., Attie, M. F., & Stock, J. L. (1980). Adenosine 3'5'-monophosphate response to parathyroid hormone: familial hypocalciuric hypercalcemia versus typical primary hyperparathyroidism. *Journal of Clinical Endocrinology and Metabolism*, 50, 546-549.

Marx, S. J., Sharp, M. E., Krudy, A., Rosenblatt, M., & Mallette, L. E. (1981). Radioimmunoassay for the middle region of parathyroid hormone: studies with a radioiodinated synthetic peptide. *Journal of Clinical Endocrinology and Metabolism*, 53, 76-84.

Hughes, W. S., Aurbach, G. D., Sharp, M. E., & Marx, S. J. (1984). The effect of bicarbonate anion on serum ionized calcium concentration in vitro. *Journal of Laboratory and Clinical Medicine*, 103, 93-103.

Sharp, M. E., & Marx, S. J. (1985). Radioimmunoassay for the middle region of parathyroid hormone: comparison of two radioiodinated synthetic peptides. *Clinica Chimica Acta*, 145, 59-68.

Fitzpatrick, L. A., Norton, J., Martin, C., Sharp, M., & Aurbach, G. D. (1988). Effect of prostaglandin F2-alpha on human parathyroid adenomas: evidence for uncoupling of parathyroid hormone secretion and cAMP accumulation. *Journal of Bone and Mineral Research*, 3, 81-86.

Wacholder, N., Sharp, M., Liu, L., Song, P., & Yuan, X. (2003). Experimental study of index terms and information access. *Proceedings of the American Society for Information Science and Technology*, 40, 184-192.

Sharp, M. (2005, March 8). Extracting a practical drug ontology from the UMLS. Paper presented at the First Semantic Technology Conference, San Francisco, CA. Retrieved from <http://comminfo.rutgers.edu/~msharp/SharpSemtech2005.pdf>

Sharp, M., Bodenreider, O., & Wacholder, N. (2008). A framework for characterizing drug information sources. *American Medical Informatics Association Annual Symposium Proceedings*, 6, 662-666.