A Framework for Characterizing Drug Information Sources

Mark Sharp, MA[1,2], Olivier Bodenreider, MD, PhD[1], Nina Wacholder, PhD[2]
[1]U.S. National Library of Medicine, Bethesda, MD; [2]School of Communication, Information and Library Studies, Rutgers University, New Brunswick, NJ

Abstract

Drug information is complex, voluminous, heterogeneous, and dynamic. Multiple sources are available, each providing some elements of information about drugs (usually for a given purpose), but there exists no integrated view or directory that could be used to locate sources appropriate to a given purpose. We examined 23 sources that provide drug information in the pharmacy, chemistry, biology, and clinical medicine domains. Their drug information content could be categorized with 39 dimensions. We propose this list of dimensions as a framework for characterizing drug information sources. As an evaluation, we show that this framework is useful for comparing drug information sources and selecting sources most relevant to a given use case.

Introduction

Drug information is complex, voluminous, heterogeneous, and dynamic. Despite attempts at ontological unification[1,2] multiple disparate sources are still the rule, each providing some elements of information about drugs (usually for a given purpose), but no integrated view and no common directory is available to locate sources appropriate to a given purpose. Drug information resources such as RxNorm, ChemIDplus, DrugInfo, and ClinicalTrials.gov (Table 1) integrate certain kinds of drug information relevant to broad usage domains including pharmacy, clinical practice, toxicology, chemistry, and clinical and preclinical research. In addition to direct data integration, some of these resources also employ cross-references to facilitate navigation to additional resources. However, these resources are not indexed by use case (e.g., finding indications for a given drug) or user type (patients, clinicians, pharmacists, researchers,…). Thus there remains a need for help in locating sources appropriate to a given need.

The objective of this paper is to propose a framework for characterizing drug information sources. We show that this framework is useful for comparing sources and selecting sources most relevant to a given use case.

Materials

We considered approximately 30 drug information sources identified through our experience in this area or referenced by ChemIDplus. This list was narrowed to 23 based on criteria such as electronic availability, presence of explicit data elements, and balancing domain and user-type coverage. The domains considered (pharmacy, chemistry, biology, and clinical medicine) were driven by a set of use cases that reflect the needs of three specific user types: consumers, physicians and pharmacists, and biomedical researchers. Examples include: finding equivalent drug names (e.g., generic version of a brand name, or the chemical name of the active ingredient); finding alternative drugs for a given indication (disease) 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. Purposely excluded were use cases from manufacturing, sales, marketing, or regulatory domains involving dimensions such as drug pricing, retailers, packaging, and patents. The 23 sources selected are listed in Table 1.

Method

Building the framework

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; e.g., *finasteride* (two brand names representing different dosages for different indications – *Proscar* 5 mg for benign prostatic hyperplasia; *Propecia* 1 mg for male-pattern baldness); *aspirin* (multiple formulations, combination products, therapeutic classes, and indications: pain, inflammation, fever, stroke risk, …); *paracetamol/acetaminophen* (synonyms); *Sinemet* (combination of carbidopa and levodopa for Parkinson's disease). Next we normalized the features into a set of "dimensions of drug information." For example,

"brand name"; "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 domain, which makes them functionally equivalent to hierarchical facets.

In addition, we assessed some of the technical characteristics of each source with implications for usage and integration, such as number of single-component approved generic names (GNs) covered, cost, database availability and update frequency, application programming interface (API) availability, and presentation (terms and relations, tables, free text, etc.).

Evaluating the framework

Grouping sources. Using the operational criteria presented earlier, we assessed each source according to the dimensions in our framework, which resulted in a matrix of mostly binary (1 or 0) scores (Table 2). We used the number of generic names and the number of chemical entities as weights for the relevant dimensions. Correspondence analysis[3] provides a method for representing both the row categories (dimensions of drug information) and the column categories (drug information sources) in the same space, so that the results can be visually examined for structure. To reduce dimensionality, only the first two or three axes of the new space are plotted. In the two-dimensional graphical display, the overall quality of representation of the points can be expressed as a proportion of the total variation (called *inertia* in correspondence analysis parlance). The statistical package MVSP was used to perform the correspondence analysis.

Selecting sources. The description of the sources provided by the matrix was also used to select sources for the following hypothetical use case. A user (consumer, clinician, pharmacist, or biomedical researcher) wants to find the drugs corresponding to a given indication or vice versa. This use case minimally requires coverage of the GN 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, albeit indirectly, in determining indications or *possible* indications.

Results

The list and grouping of the 39 dimensions in our framework are shown in Table 2, along with the assessment of the 23 sources. GN coverage is shown in Table 1. The other technical assessments are not shown due to space limitations.

Grouping sources. The correspondence analysis was performed using a weighting scheme reflecting the coverage of GNs and chemical entities, and giving credit for partial coverage. The first two principal axes account only for about 30% of the total inertia, which means that some points may not be correctly represented with respect to these two axes. Regardless of weighting schemes, there is a consistent distinction in Table 2 between clinical and chemistry dimensions (i.e., most sources exhibiting clinical features do not exhibit chemistry features), while pharmacy and biology are more diffuse.

The correspondence analysis joint plot (Figure 1) provides a visual rendering of the information in Table 2. Here, the horizontal axis is polarized between features corresponding to clinical information (red diamonds) and biology (blue, down-pointing triangles) on the left, and chemistry features (green squares) on the right. In contrast, pharmacy features (purple, up-pointing triangles) are mostly at the center, which means that they lack discriminating power. Such polarization of the horizontal axis helps interpret the grouping of drug information sources. Sources clustering to the right focus on chemistry (e.g., ChEBI), while sources from the group on the left (e.g., MedMaster) focus on clinical and biology information. The group of sources at the center (including DrugBank and WHO-DRUG) corresponds mostly to drug information having features from most domains, which prevents them from being effectively categorized. Similarly, *therapeutic class* (red diamond in the middle) is close to the center, because this feature is shared by most drug information sources.

Selecting sources. By adding up the sources' matrix scores (●=1; ○ =0.5) for the dimensions required to satisfy the hypothetical use case described earlier, we identified UMLS[1] (7), DrugBank (6), DailyMed (5), WHO-ATC (3.5), and ClinicalTrials.gov (3) as the best candidates to satisfy such a use case.

Discussion

Comparative generic name coverage. Comparative GN coverage (Table 1) is important since GNs are the

[1] The UMLS integrates about 150 biomedical terminologies, including several drug information sources, which is the reason why it receives the highest score. The single most important source from the UMLS for this hypothetical use case is the Veterans Health Administration (VHA) National Drug File-Reference Terminology (NDF-RT).

most common way of naming fundamental drug entities. However, getting comparable numbers is difficult. The sources do not always distinguish GNs from other ways of naming and counting drug entities. There may also be ambiguity about whether a source's GNs all correspond to approved drugs or also include experimental drugs. Finally, database currency (update frequency) varies widely, from one day to several years. Ultimately, comparative drug coverage requires detailed cross-mapping of the sources' drug terminologies and record identifiers.

Overlapping dimensions. 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*. We generally gave a source credit for *therapeutic class* first and others only if the source gave additional information of this type. An exception is WHO-ATC, which is designed to embed anatomical and chemical superclass information in the therapeutic class term, and so to it we gave additional full credit for those dimensions plus partial credit for three related biology dimensions.

Our framework was designed to be generic enough to represent all dimensions found in the drug information sources we explored. From the perspective of clustering sources, overlapping dimensions could be eliminated. However, overnormalizing the dimensions would likely make it a less effective tool for describing and matching sources and use cases.

Applications. The matrix constitutes a mapping of sources to dimensions that, given a similar mapping of dimensions to use cases, can be used to map sources to use cases and compare the sources' likelihood of effectiveness for satisfying user needs. A standard framework for describing drug information sources is a necessary step towards improving the discoverability of such resources by humans and agents.

Limitations. The limitations of this study start with the arbitrary source list. We focused on sources we thought would pragmatically balance authority, comprehensiveness, appropriateness to our chosen domains and user types, and free, electronic, integration-friendly availability. Thus we did not consider additional commercial sources such as the Physician's Desk Reference (PDR)[7], Martindale[8], or the Merck Index[9] even though they might well have bested some our considered sources in these regards. Similarly, other open-source bioinformatics knowledge sources that were left out due to lack of explicit drug focus might well have done better in our evaluation than Reactome or HumanCyc.

Additional limitations have been alluded to: uncertainty of mappings and equivalence due to inter-source variation in scope, specification, and knowledge representation (including terminology); instabilities due to varying database currency and update frequency; difficulty of pragmatic quality assessment (i.e., unreliability of few- or single-instance comparisons versus high effort/benefit ratio of statistical rigor); and a pragmatic as opposed to systematic approach that may have overlooked some important dimensions.

Conclusion

We analyzed 23 drug information sources and extracted 39 dimensions of drug information relevant to four major domains. We demonstrated that this framework is useful for comparing drug information sources and selecting sources most relevant to a given use case.

Acknowledgements

This research was supported in part by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM).

References

1. Carter JS, Brown SH, Bauer BA, Elkin PL, Erlbaum MS, Froehling DA, Lincoln MJ, Rosenbloom ST, Wahner-Roedler DL, Tuttle MS. Categorical information in pharmaceutical terminologies. AMIA Annu Symp Proc. 2006;:116-20.
2. Solomon WD, Wroe CJ, Rector AL, Rogers JE, Fistein JL, Johnson P. A reference terminology for drugs. Proc AMIA Symp. 1999;:152-5.
3. Greenacre MJ. Theory and applications of correspondence analysis. London: Academic Press; 1984.
4. IUPAC. The IUPAC International Chemical Identifier (InChI™). http://www.iupac.org/inchi/ 2008.
5. Simplified molecular input line entry specification. http://en.wikipedia.org/wiki/Simplified_molecular_input_line_entry_specification 2008.
6. Lipinski's Rule of Five. http://en.wikipedia.org/wiki/Lipinski's_Rule_of_Five 2008.
7. PDR.net. http://www.pdr.net/ 2008.
8. Sweetman S. Martindale: The Complete Drug Reference (35th Edition) 2007. London: Pharmaceutical Press.
9. The Merck Index - An Encyclopedia of Chemicals, Drugs, and Biologicals (14th Edition) (Ed. M. O'Neil et al) 2006. Whitehouse Station, NJ: Merck & Co., Inc.

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

Table 1. Sources selected for evaluation. (# INs number of single-compound approved generic name; nd not determined).

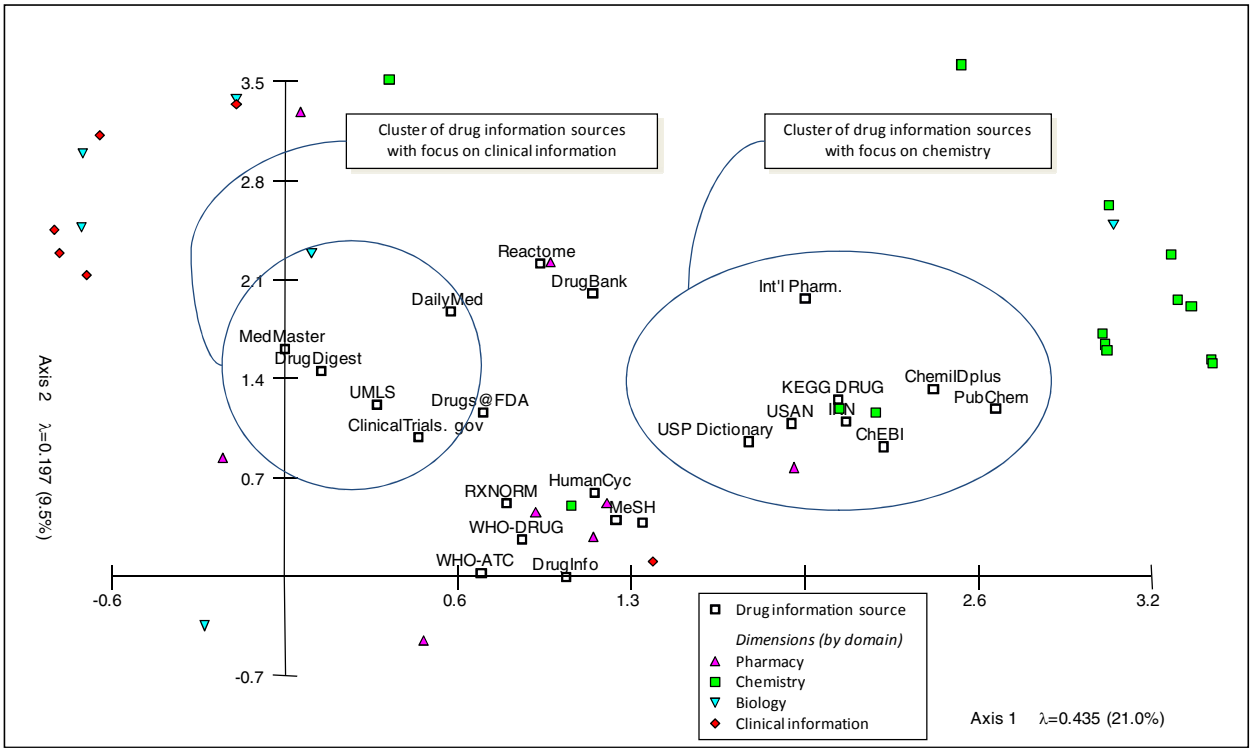

Figure 1. Correspondence analysis between drug information sources and dimensions of drug information (in four domains).

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	ChemiIDplus	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[4]																	●	●	●	●			
	SMILES[5]																	●	●	●	●			●
	similar structures																	●	●		●	●		
	H bond donors[6]																	●						
	H bond acceptors[6]																	●						
	molecular weight[6]			●							●			●				●		●	●	●		
	solubility[6]			●							●							●	●		●			
	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	●	●	●													●				●			
	side eff/prec/warn	●	●	●													○							
	drug interactions	●	●	●													●				●			
	patient info	●	●	●																	●			
	research lit.			●																	●			
	experimental app's				●									○			●				●			

Table 2. Dimensions of drug information for each drug information source. (coverage: ● = full, ○ = partial; CAS: Chemical Abstracts Service; ADME: absorption, distribution, metabolism, excretion).