Desiderata for Drug Classification Systems for their Use in Analyzing Large Drug Prescription Datasets

Fabrício S. P. Kury, MD[1], Olivier Bodenreider, MD, PhD[1]
[1]Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, Bethesda, MD, USA

Abstract

Background: *Information about the billions of prescriptions written for patients each year is collected in large datasets. Drug classification systems (DCSs) are key to analyzing these datasets. However, their ability to support such analyses has not been studied.* ***Methods:*** *We identified six desirable features for drug classification systems (DCSs) from the perspective of analyzing large drug prescription datasets. In addition to offering operational definitions for these desiderata, we also applied them to clinically significant drugs in RxNorm for six DCSs, and used them to assess the impact of these DCSs on the analysis of a large drug prescription dataset from Medicare Part D claims.* ***Results:*** *Based on these desiderata, we could determine that ATC, VAC and EPC seem to be better suited for the analysis of large drug prescription datasets, because of their coverage and granularity, and because ATC and VAC support aggregation.*

Introduction and Background

Prescription drugs accounted for $297.7 billion in the U.S. in 2014, or 9.8% of the national health expenditures in that year[1]. Approximately 3 in 5 American adults affirm to be currently taking at least one prescription medication, a proportion that continuously increased from 1999 to 2012[2], and cost $858 per capita in 2013[3]. In year 2015, over four billion drug prescriptions were filled at U.S. pharmacies[4]. Drug therapy is one of the pillars of U.S. health care, and large drug prescription datasets can potentially support clinical, public health, and health policy analyses.

It is often convenient to refer to drugs not as individual ingredients or products, but rather as sets of drugs that share particular characteristics, i.e. drug classes. Medications can be grouped according to different perspectives, and different drug classification systems (DCSs) have been developed for various use cases. However, it might not be easy for researchers to select a given DCS for their study. In this investigation, we review some of the characteristics of six publicly available DCSs and outline desiderata for their use in analyzing large drug prescription datasets.

The DCSs under investigation are the Anatomical Therapeutic Chemical (ATC) classification system[5] developed by the World Health Organization (WHO) Collaborating Centre for Drug Statistics Methodology; the Established Pharmacological Classes (EPC) from the U.S. Food and Drug Administration (FDA); and four DCSs from the U.S. Veterans Health Administration' (VHA) National Drug File - Reference Terminology (NDF-RT)[6], namely Mechanism of Action (MoA), Physiological Effect (PE), Chemical Ingredient (Chem), and the Veteran Affairs' Drug Classes (VAC).

Prior research has investigated the desirable characteristics of medical terminologies (e.g. Cimino, 1998[7]). Some desiderata identified for medical terminologies are particularly relevant to DCSs, including coverage, the existence of a hierarchical structure, granularity, and non-ambiguity. The specific contribution of our work is to apply Cimino's desiderata specifically to DCSs and extend them where appropriate. Moreover, we provide operational definitions for these desiderata, apply them to clinically significant drugs in RxNorm for six DCSs, and use our desiderata to assess the impact of these DCSs on the analysis of 1 billion Medicare Part D claims for 4.8 million Medicare beneficiaries over nine years.

Materials

We heavily leverage the RxNorm and RxClass application programming interfaces (APIs)[8], developed by the U.S. National Library of Medicine, for selecting drugs of interest and associating them with their corresponding classes. Providing a detailed explanation of how we create these associations is beyond the scope of this paper, which focuses on the desirable characteristics of DCSs. Salient features are briefly mentioned. Of the six DCSs, only VAC associates drug products with classes. The other DCSs link ingredients to classes. We use RxNorm to link drug products to their ingredients, including for multi-ingredient drugs, so that they can be further linked to classes. For ATC, we also take

into account the dose form of the drug product to prevent inaccurate associations (e.g., association between a systemic drug product and a topic drug class).

We leverage drug properties in RxNorm to select a set of clinically significant drugs, to which our investigation is restricted, as they are arguably more important than over-the-counter drugs. More specifically, we only consider drug products listed in the prescribable subset of RxNorm and for which there is at least one association with a National Drug Code (NDC), denoting presence on the U.S. market. There are 18,015 such clinically significant RxNorm drug products, among the 30,205 drug products in RxNorm. We use the RxNorm API (https://rxnav.nlm.nih.gov/) to map the NDC identifiers found in the 1,015,371,392 Medicare Part D claims in our dataset to RxNorm products. This dataset corresponds to all Part D claims produced from 2006 to 2014 for a cohort of 10% of all Medicare beneficiaries randomly sampled from 1999 to 2013, totaling 4,842,281 beneficiaries. Virtually all claims in our dataset (99.45%) can be associated with a clinically significant drug product in RxNorm.

Methods and Results

From our experience in analyzing large drug prescription datasets, we propose six desiderata. When appropriate (i.e., for the last four desiderata), we apply them to clinically significant drugs in RxNorm for six DCSs, and use them to assess the impact of these DCSs on the analysis of a large drug prescription dataset from Medicare Part D claims. (The source code developed to assess impact is freely available under an *Attribution-NonCommercial-ShareAlike 4.0 International*[9] license at https://github.com/fabkury/ddcs.) The tables and figures that are not near their mention in the text below can be found at the end of this paper after the References section.

A. Administrative desiderata. As medical vocabularies in general, DCSs are expected to be authoritative, i.e., developed by a reputable organization and free of conflicts of interest (A1), and to reflect current medical knowledge (A2). Publicly available resources are also preferred as they are easier to use and share in the community (A3). For the six DCSs under investigation, the originating institution, frequency of update and licensing status and cost are listed in Table 1.

Table 1. Desideratum A: *Administrative desiderata.*

	ATC	EPC	MoA	PE	Chem	VAC
A1. Developer	WHO	FDA/VHA	VHA	VHA	VHA	VHA
A2. Updates	Yearly[9]	NDF-RT is published 10 times a year.[6]				
A3. License	Free*	NDF-RT is publicly available and free of charge.[6]				

*: ATC costs €200, however its hierarchy is freely available through the Unified Medical Language System Metathesaurus.

B. Hierarchical structure. One main requirement for analyzing large drug prescription datasets is support for aggregating drugs into classes at various levels of granularity. DCSs are expected to arrange classes hierarchically (B1), i.e., from more specific classes to more generic classes. Moreover, as for medical vocabularies in general, such hierarchies are expected to be acyclic and can also support multiple parents (B2). The level of granularity should not be limited by *fiat*, but rather be adapted to the analytical requirements (B3). Finally, drug classification can arguably be more consistent if drugs are only attached to leaf classes, i.e., at the bottom of the hierarchy, while more generic classes are used only for aggregation purposes (B4). For the six DCSs under investigation, the existence of a hierarchical structure (with single or multiple parents) and the number of levels in the hierarchy are listed in Table 2. Additionally, we also examine whether drugs are only attached to leaf classes or can be attached at any level.

Table 2. Desideratum B: *Hierarchical structure.*

	ATC	EPC	MoA	PE	Chem	VAC
B1. Existence of a hierarchical structure	Yes	No	Yes	Yes	Yes	Yes
B2. Hierarchy supports multiple parents	No	N/A*	Yes	Yes	Yes	No
B3. Number of levels	4	N/A*	1 to 8	1 to 9	1 to 13	3
B4. Drugs are always at the leaf level	Yes	N/A*	No	No	No	No

*: Not applicable because EPC has no hierarchical structure.

C. Coverage. No meaningful analysis of a large drug prescription dataset can be conducted if the drugs in this dataset are not sufficiently covered by the classification system. We assess the proportion of clinically significant drugs in RxNorm (59.6% of all drug products in RxNorm) covered by each of the six DCSs. Moreover, since not all drugs have the same frequency of prescription, we assess the proportion of prescriptions (claims) from our Medicare Part D

dataset covered by each of the six DCSs. The results, listed in Table 3, show that ATC and VAC provide the best coverage for RxNorm drugs and Medicare prescriptions, respectively. PE and Chem provide insufficient coverage.

D. Granularity: While the optimal number of drug classes is impossible to determine outside a particular use case, there should generally be a sufficient number of classes to support distinctions among drugs. As shown in Table 4, the total number of classes ranges from 486 in VAC to 10,239 in Chem. Of note, the proportion of "empty classes", i.e., classes with no directly associated clinically significant drugs from RxNorm, ranges from 22.5% in EPC to 98.1% in Chem. The small proportion of empty classes in EPC reflects the lack of a hierarchy, i.e., the absence of aggregation classes, in addition to reasonable coverage. Conversely, the high proportion in Chem reflects not only a rich hierarchy, but also poor coverage as mentioned earlier.

E. Non-ambiguity. Good classification systems have the property of being jointly exhaustive and mutually exclusive. For DCSs, however, this is not always the case for several reasons. As noted earlier, most DCSs link drug ingredients to classes, while a given ingredient can often be used for multiple therapeutic purposes. For example, the beta-blocker *timolol* can be used as a cardiovascular drug or as an anti-glaucoma drug. For this reason, *timolol* is associated with multiple classes in ATC. In contrast, drug products containing *timolol* are unambiguously associated with the corresponding anti-ophthalmic or cardiovascular classes in VAC. Moreover, multi-ingredient drugs are not always represented in DCSs (i.e., associated with a single combination class). Instead, each constituent ingredient often ends up being associated with a class. To assess the level of ambiguity, we analyze the distribution of number of classes per drug across the six DCSs. As shown in Table 5, there is virtually no ambiguity with VAC (as expected), whereas the level of ambiguity ranges from 7.81% to 16.81% for the other DCSs.

F. Dispersion. Even if a DCS provides a sufficient number of classes, it is still possible that many of the classes are not directly associated with any drugs at all (or with very few drugs), leaving a large proportion of the drugs concentrated within a relatively small number of classes. Such a situation will likely impede the meaningful analysis of the drugs. Ideally, drugs should spread over a sufficient proportion of the classes to avoid such concentration. To measure the "dispersion" of the drugs across the range of classes in a DCS, we plot the proportion of RxNorm products (Fig. 1) and Medicare Part D claims (Fig. 2) covered by a given proportion of classes in each DCS. To avoid double counts, a drug product or claim associated with multiple classes is only assigned to the largest class. The classes within each DCS in the plots are sorted from largest to smallest (thereby ending in empty classes that do not increase coverage). The dispersion is represented by the initial slope of these graphs. For example, in Fig. 1, the concentration of drug products into a small number of classes for Chem results into a sharply ascending curve. In contrast, the curve for ATC, VAC and EPC is ascending less rapidly, reflecting a better repartition of drugs among classes.

Discussion

Findings and significance

Selecting a drug classification system (DCS) remains difficult as the features of DCSs are not always explicitly listed. Moreover, to our knowledge, no list of desirable features has been established for specific use cases. This investigation is our attempt to clarify the characteristics of DCSs and list which ones are desirable for a relatively generic use case, namely the analysis of large prescription datasets. In our experience, the most important among our desiderata are coverage, the existence of a hierarchical structure, and granularity. Coverage is a strong limiting factor for PE and Chem, and, to a lesser extent, also for MoA. Of note, coverage in terms of RxNorm drugs significantly underestimates coverage for actual prescriptions, as illustrated by the difference between Fig. 1 and 2. In other words, ATC, VAC and EPC achieve over 92% coverage for prescriptions, despite having a 58-73% coverage for RxNorm drugs. EPC has the disadvantage of not providing a hierarchical structure to support aggregation. (The lack of a hierarchical structure among EPCs is actually being addressed through mapping to SNOMED CT classes). The advantage of VAC over ATC is to link drug products to classes directly, which is particularly more important for multi-ingredient drugs because it is commonly the case that the distinct ingredients belong to distinct classes from the perspective of ATC. The indirect link through ingredients in ATC results in higher ambiguity as shown in Table 5. For these reasons, it is not surprising that the i2b2 research community has adopted VAC for analyzing prescriptions in their clinical data warehouses. On the other hand, the fixed numbers of levels of ATC somewhat facilitates the interpretation of its classes.

Our investigation extends previous research on desiderata for medical vocabularies, of which DCSs can be thought of as a special case. However, many of Cimino's desiderata[7] are not particularly relevant to DCSs, i.e., will not directly impact the analysis of prescription datasets. For example, while the use of *meaningless identifiers* is part of Cimino's desiderata, the fact that ATC is not compliant with this desideratum does not make it a "bad" DCS, nor does it make

the other DCSs, which all comply with it, better DCSs for our generic use case. This investigation is also complementary to our previous work done on the comparison of drug classes between DCSs[11].

Finally, the graphs we created to visualize dispersion (Fig. 1 and 2) also reflect other desiderata and can be used for an overall assessment of the DCSs. In addition to dispersion, represented by the slope of the curves, coverage is indicated by the maximum value reached by the DCS in the vertical axis, and the length of the horizontal plateau of each line reflects the proportion of classes of the DCS with no drugs directly attached to them, i.e., the proportion of empty classes.

Limitations and future work

Our list of desiderata only addresses to one relatively generic use case, namely the analysis of large drug prescription datasets. Other use cases would probably call for different or additional desiderata as originally pointed out by Cimino[7]. For example, the analysis of multi-ingredient drugs would be facilitated by the presence of combination classes for such drugs in DCSs. Conversely, the existence of combination classes for multi-ingredient drugs makes it more difficult to find all patients with any prescription of a drug from a given class, because both the individual class and the combination class/classes need to be queried.

We only assessed the impact of the DCSs on the number of prescriptions in each class. In future work, we also want to assess their impact on the number of beneficiaries with prescriptions of the various classes of a DCS, which is a common metric for drug utilization. Moreover, when analyzing empty classes, we have not distinguished between empty leaf classes and empty intermediary classes. While intermediary classes are most often empty, and always empty by design in ATC, empty leaf classes might denote extraneous classes. We will study this distinction in our next investigation.

Finally, our analysis of the Part D claims has two limitations. First, the population of Medicare beneficiaries is skewed towards older adults, with a small proportion of younger individuals with disability and/or renal disease. Therefore, our assessment of the impact of DCSs on the analysis of this dataset may not generalize to other prescription datasets. Second, given Medicare reimbursement criteria, it is not surprising to find that the vast majority of claims (99.45%) could be linked to clinically significant drug products in RxNorm. This proportion may be lower in other prescription datasets, which are likely to include over-the-counter drugs.

Conclusion

We presented six desirable features for drug classification systems (DCSs) from the perspective of analyzing large drug prescription datasets. In addition to offering operational definitions for these desiderata, we also applied them to clinically significant drugs in RxNorm for six DCSs, and used them to assess the impact of these DCSs on the analysis of a large drug prescription dataset from Medicare Part D claims. Based on these desiderata, we could determine that ATC, VAC and EPC seem to be better suited for the analysis of large drug prescription datasets, because of their coverage and granularity, as well as their support for aggregation. Additional desiderata may be required for specific use cases.

Acknowledgments

This work is supported by the National Institutes of Health intramural funding and in part by the Intramural Research Program of the U.S. National Library of Medicine (NLM) and an appointment to the NLM Research Participation Program administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S Department of Energy and the NLM.

References

1. National Center for Health Statistics. Health, United States, 2015: With Special Feature on Racial and Ethnic Health Disparities. Hyattsville, MD. 2016. Table 94. Accessed from http://www.cdc.gov/nchs/data/hus/hus15.pdf#094 on September 30th, 2016.
2. Kantor, E. D., Rehm, C. D., Haas, J. S., Chan, A. T., Giovannucci, E. L., GT, S., … PD, S. (2015). Trends in Prescription Drug Use Among Adults in the United States From 1999-2012. *JAMA*, *314*(17), 1818. http://doi.org/10.1001/jama.2015.13766
3. Kesselheim, A. S., Avorn, J., Sarpatwari, A., SP, K., BS, Q., R, L., … NK, C. (2016). The High Cost of Prescription Drugs in the United States. *JAMA*, *316*(8), 858. http://doi.org/10.1001/jama.2016.11237
4. The Henry J. Kaiser Family Foundation. 2015. Total Number of Retail Prescription Drugs Filled at Pharmacies. Accessed from http://kff.org/other/state-indicator/total-retail-rx-

drugs/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D on September 30th, 2016.
5. World Health Organization (WHO) Collaborating Centre for Drug Statistics Methodology. ATC/DDD Index 2016. 2016. Accessed from http://www.whocc.no/atc_ddd_index/ on September 30, 2016.
6. U.S. Department of Veterans Affairs, Veterans Health Administration. 2015. National Drug File – Reference Terminology (NDF-RT™) Documentation. February 2015 Version. Accessed from https://evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT%20Documentation.pdf on September 30, 2016.
7. Cimino, J. J. (1998). Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of Information in Medicine*, *37*(4-5), 394–403. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9865037
8. U.S. National Library of Medicine. RxNav and RxClass. Accessed from https://rxnav.nlm.nih.gov/ on September 30, 2016.
9. http://creativecommons.org/licenses/by-nc-sa/4.0/. Does not replace legal code at http://creativecommons.org/licenses/by-nc-sa/4.0/legalcode. Both links accessed on October 20th, 2016.
10. WHO Collaborating Centre for Drug Statistics Methodology, Guidelines for ATC classification and DDD assignment 2013. Oslo, 2012. Accessed from http://www.whocc.no/filearchive/publications/1_2013guidelines.pdf on September 30, 2016.
11. Winnenburg, R., & Bodenreider, O. (2014). A framework for assessing the consistency of drug classes across sources. *Journal of Biomedical Semantics*, *5*, 30. http://doi.org/10.1186/2041-1480-5-30

Tables and figures

Table 3. Desideratum C: *Coverage.*

DCS	% RxNorm drug products	% Medicare Part D claims
ATC	73.0%	97.6%
EPC	57.7%	92.3%
MoA	30.1%	59.4%
PE	13.3%	15.5%
Chem	24.6%	28.1%
VAC	67.5%	98.9%

Table 4. Desideratum D: *Granularity.*

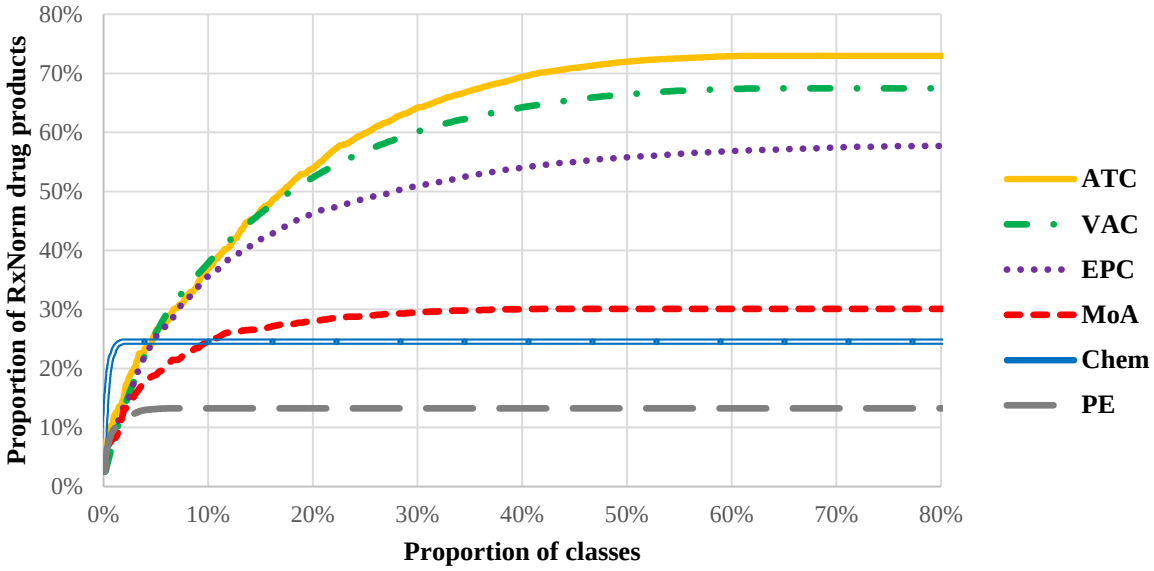
DCS	Total classes	Non-empty classes
ATC-1 to 4	1,257	545 (43,4%)
EPC	595	461 (77,5%)
MoA	608	254 (41,8%)
PE	1,866	112 (6%)
Chem	10,239	198 (1,9%)
VAC	486	308 (63,4%)

Table 5. Desideratum E: *Non-ambiguity.**

Classes per RxNorm drug product†	ATC4	EPC	MoA	PE	Chem	VAC
1 class	81.02%	91.53%	86.61%	69.51%	87.34%	99.04%
2 classes	10.98%	7.81%	7.99%	16.81%	10.04%	0.90%
3 to 4 classes	4.78%	0,65%	3.30%	13.59%	2.62%	0.06%
5 to 7 classes	1.88%	0	1,81%	0.09%	0	0
8 to 11 classes	1.33%	0	0.28%	0	0	0
12 or more classes	0	0	0	0	0	0

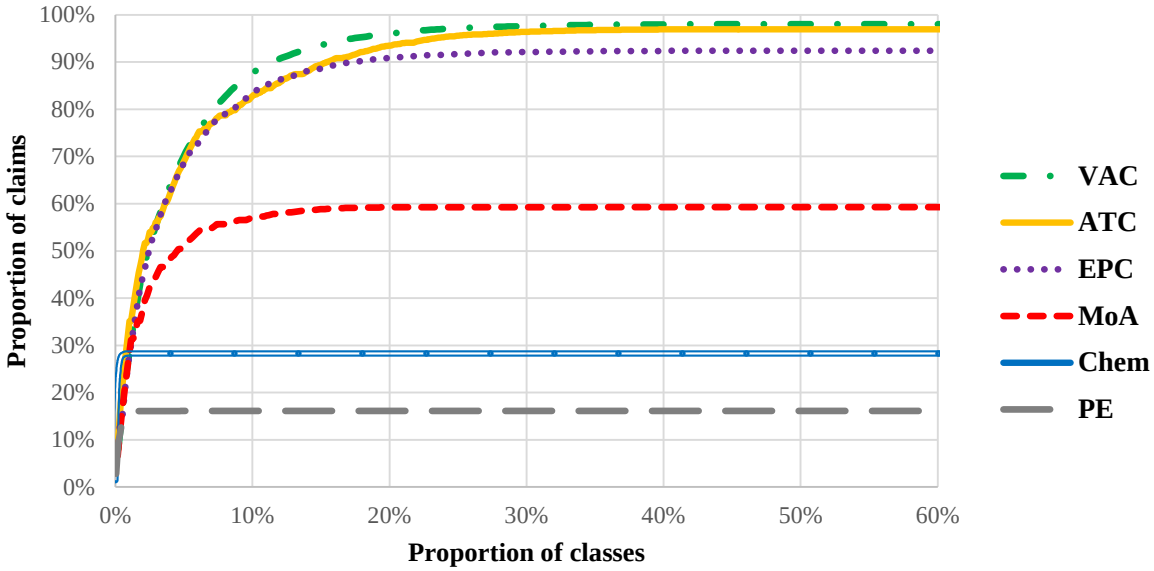
*: Some columns do not add up to exactly 100% due to rounding.
†: Only single-ingredient RxNorm drug products were considered.

Figure 1. Desideratum F: *Dispersion.* Proportion of RxNorm products covered by a given proportion of classes.*



*: Graph was trimmed to improve comparability – ATC attains maximum coverage at 62% classes, VAC at 64%, EPC at 78%, MoA at 42%, Chem at 2%, and PE at 6%.

Figure 2. Desideratum F: *Dispersion.* Proportion of Medicare Part D claims from 2006 to 2014 covered by a given proportion of classes.*



*: Graph was trimmed to improve comparability – VAC attains maximum coverage at 60% classes, ATC at 58%, EPC at 69%, MoA at 38%, Chem at 2%, and PE at 6%.