AMIA Jt Summits Transl Sci Proc, 2016. CRI: p. 105-110.

The Drug Data to Knowledge Pipeline:

Large-Scale Claims Data Classification for Pharmacologic Insight

Mark L. Homer, PhD[1,2], Nathan P. Palmer, PhD[1,2], Olivier Bodenreider, MD, PhD[3], Aurel Cami, PhD[1,2], Laura Chadwick, PharmD[1,2,4], Kenneth D. Mandl, MD, MPH[1,2]
[1]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA;
[2]Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA;
[3]National Library of Medicine, National Institutes of Health, Bethesda, MD, USA;
[4]MCPHS University, Boston, MA, USA

Abstract

In biomedical informatics, assigning drug codes to categories is a common step in the analysis pipeline. Unfortunately, incomplete mappings are the norm rather than the exception with coverage values less than 85% not uncommon. Here, we perform this linking task on a nationwide insurance claims database with over 13 million members who were dispensed, according to National Drug Codes (NDCs), over 50,000 unique product forms of medication. The chosen approach employs Cerner Multum's VantageRx and the U.S. National Library of Medicine's RxMix. As a result, 94.0% of the NDCs were successfully mapped to categories used by common drug terminologies, e.g., Anatomical Therapeutic Chemical (ATC). Implemented as an SQL database and scripts, the approach is generic and can be setup for a new data set in a few hours. Thus, the method is a viable option for large-scale drug classification.

Introduction

Across clinics and hospitals, patient information continuously streams into electronic health records. The databases are designed to handle clinical and billing requirements, but also have a secondary use where analysis of patterns and trends leads to medical insights[1, 2, 3, 4, 5, 6]. For example, consider the hypothetical situation of 10,000 individuals diagnosed with the same medical condition. Suppose there were two relevant classes of drugs, differing by mechanism of action, and about half were treated with one drug and half the other. After follow up, the data could give insight into which type of drug is more effective "in the wild." In combination with randomized clinical trials and expert panels, such predictive analytics promises to expand and refine clinical guidelines, thereby bettering medical care.

This promise can only be fulfilled, however, if we can make sense of the data. Here, our particular focus is on matching data values to meaningful concepts. In the example above, it's critical we know the drug type given to each patient, but unfortunately what is often recorded is a cryptic drug code and perhaps a non-standardized, textual description. The matching of drug codes/descriptions to active ingredients or drug categories is a typical step in biomedical informatics, but an agreed upon, consistent, effective method is still an active topic of research[7, 8, 9, 10].

National Drug Codes (NDCs) are a classification system used in the medication information supply chain. An NDC identifier is a string of 11 digits. The first 4-5 digits denote the FDA provided drug labeler's number, while the remaining are chosen by the labeler. An NDC is assigned to each variation of the labeled drug product, so there can be many NDCs for one active ingredient that differ by brand name, strength, route of administration, and/or packaging with other medications, e.g., drug pack. Unfortunately, there is no consistent subset of digits within an NDC to indicate the medication's active ingredients. For example, both "52959050506" and "00093202631" contain azithromycin.

NDCs can be classified using several terminology systems. For example, the National Drug File – Reference Terminology (NDF-RT), provided by the Veterans Health Administration, can group medications by mechanism of action[11]. The Anatomical Therapeutic Chemical (ATC) Classification System, provided by the WHO Collaborating Centre for Drug Statistics Methodology, operates more strictly on a hierarchy, with its second level organizing substances by therapeutic purpose[12]. The active ingredient functions as a central concept common to these ontologies. So in principle, once an NDC is linked to an active ingredient, we can choose the most appropriate categorization system for a given analysis.

Our method implements this idea by first assigning active ingredient(s) to each NDC, using the VantageRx commercial database sold by Cerner Multum[13]. Then, both VantageRx and RxMix, a web-based service developed by the National Library of Medicine[14], enable mapping to various categories in the Multum, NDF-RT, ATC, MESH[15], DAILYMED[16], and FDASPL[17] terminologies. Built within an SQL database, our solution can operate on large-scale data sets, with hundreds of thousands of NDCs. Unlike the practice of manually assembling custom dictionaries for each drug class, the constructed tables can be reused with a minimum of editing.

Here, we explain our methodology, and describe an evaluation procedure, based on a large-scale insurance claims data set. Our results yield a 94.0% mapping coverage rate for the Multum categorization system. This is a step forward in performance, since the percentages in existing studies can be 80% or lower[7]. Furthermore, the whole classification process for a new data set is estimated to take only a few hours using a commodity server. While we cover the work's limitations and point out the additional work needed to fully develop and vet the technique, we view the progress to date as an advancement in large-scale drug classification from NDCs and, thus, a significant contribution to clinical analytics.

Drug Classification Methods

Figure 1 gives an overview of how the solution, implemented in a Microsoft SQL Server database and scripts, is used and works. The NDCs one wishes to classify are directly linked to a main Multum drug code, i.e. identifier. For those NDCs not directly linked, an attempt is made to match on their associated description strings. Once linked to a Multum drug code, classification can be accomplished via either VantageRx's own categories or a map generated by RxMix. The end result is a lookup table, which indicates the classification(s) assigned to each NDC.

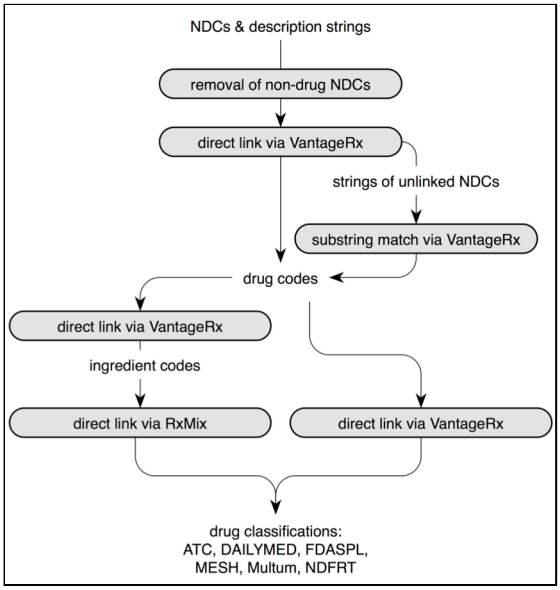

Figure 1. Conceptual overview of large-scale drug classification method.

Most of the linking leverages the VantageRx database by Multum Cerner. One purpose of the tables is to link NDCs. Other features include drug-drug interactions, synonyms for drug names, as well as therapeutic categories, the latter of which we also employ. Our version of the database is from 2011. In 2009, a study evaluated various databases' ability to link NDCs from 13 sources (both inpatient and outpatient, # records >190,000) and found the Multum tables to be one of the better choices[7]. Multum and RxNorm[18] tied for the #1 spot overall at 84.1% of the NDCs covered, on average. As an initial check, both Multum and RxNorm were used to link the NDCs in our test case using one of the standard procedures offered in their documentation. Multum's coverage was 77.8%, where as RxNorm's was 60.0%. For these reasons, VantageRx forms our method's core.

Per Figure 1, the first use of VantageRx is to filter out known non-medication NDCs. The database has tables pertaining specifically to medical supplies and their associated NDCs. Additionally, we make use of the provided description string to filter out known supplies, such as the substring "NEEDLE." Once accomplished, the next step is an inner join on a central VantageRx table to secure a main drug code in the Multum lexicon.

The overwhelming majority of NDCs are linked this way, but those not recognized can sometimes be identified by their associated description string. Multum's primary name for each drug is compared to the description string of each unmatched NDC, e.g. NDC = "60429078545" with Description = "WARFARIN TAB 2MG". The procedure exploits the fact that some of the description strings follow a known convention. For example, in our Warfarin case, the first word in the string denotes the drug's active ingredient. Since the Multum name typically follows the convention of a label followed by a strength and route of administration, we extract the only the label for comparison. Note, as in our example, that the substring matching only connects NDCs with active ingredients or brand names, not dosing strength nor route of administration. Since the objective here is drug classification, rather than medication reconciliation or dose adjustment, however, this is typically not a concern.

Using the main drug name, the medication can be mapped to Multum defined drug categories, or for additional classification systems, we can link to Multum ingredient codes. The power of the ingredient codes is that they are included in RxMix, a web-based service provided by the National Library of Medicine (NLM). RxMix permits one to setup a cascade of calls to RxNorm, RxTerms, RxClass, NDF-RT, and related APIs. Our solution has three calls. First, it links Multum ingredient IDs to their corresponding RxNorm concept unique identifiers (CUIs), using the "findRxcuiById" function. Second, those CUIs are linked to related ingredient concepts, such as precise ingredient names by the "getRelatedByType" function. Finally, categories within the ATC, MESH, NDFRT, DAILYMED, FDASPL systems are found via the "getClassByRxNormDrugId" function. Specifically, ATC provides a categorization system with four levels. MESH gives the "MeSH Pharmacologic Actions" categories. DAILYMED and FDASPL provide three different ways to categorize drugs: "Established Pharmacologic Class", "Mechanism of Action", and "Physiologic Effect". Finally, NDF-RT delivers four category sets: "Diseases, Manifestations or Physiologic States", "Cellular or Molecular Interactions", "Physiological Effects", and "Clinical Kinetics." Note that the final output is a one-to-many relationship. A single NDC can map to multiple ingredients, each of which can then map to many categories across several classification systems. For example, NDC 63304058730 contains atorvastatin and amlodipine besylate. The ingredients map to different ATC categories (HMG CoA reductase inhibitors and Dihydropyridine derivatives respectively) and as well as two Multum classes (antihyperlipidemic combinations and antihyperlipidemic combinations).

Evaluation Methods

In order to evaluate the classification method, we applied it to prescription fills in a large insurance claims data set. These records were supplied from a nationwide data warehouse, and cover the period January 2010 to May 2013. The data was considered large-scale by several counts (Table 1). Of particular interest, there were the 51,490 unique NDCs, which is on par with other large-scale studies[7]. They, along with their respective description strings, were extracted from the data set and organized in a table within the MS SQL server database. We ran the SQL scripts in Microsoft SQL Server 2012 on a Dell R610 server with 48GB of RAM and a 6 core Xeon CPU running Windows Server 2012. The final and intermediate outputs took the form of SQL tables.

Table 1. Counts of key attributes in the evaluation data set.

# records	348,033,664
# members	13,044,428
# pharmacies	60,863
# unique NDCs	51,490

To evaluate these results, the primary metric was percent coverage of each terminology, i.e. the fraction of NDCs in the evaluation data set mapped to each system. For the mapped NDCs, error checking was done by randomly selecting 500 entries for manual inspection. For each entry, we checked whether the correct ingredient and ATC category was assigned to the NDC. We recognize the sample is not large enough to ascertain a complete picture of the method's accuracy. Instead, the inspection was to provide some evidence of the procedure's performance. Inspection tools included the NDDF BioPortal website[19], the National Library of Medicine's DailyMed website[20], Lexicomp, a tool used by the Boston Children's Hospital formulary, and the World Health Organization's ATC online search query tool[21].

Results

Using two 2.4GHz processors, the method took less than 3 minutes to generate a classification table with 983,339 rows (Table 2 gives an excerpt). An NDC, Trizivir tablet, is shown along with the respective description strings and assigned categories. Apparent from the table are the one-to-many relationships, going from the NDC to its categories. The branching is because Trizivir contains three active ingredients: Abacavir, Lamivudine, Zidovudine. Each active ingredient, in turn, can be represented in more than one classification system.

Table 2. Excerpt from output table.

NDC	Description	Class System	Class Code	Class Description
49702021718	TRIZIVIR TAB	ATC	J05AF	Nucleoside and nucleotide reverse transcriptase inhibitors
49702021718	TRIZIVIR TAB	ATC	J05AR	Antivirals for treatment of HIV infections, combinations
49702021718	TRIZIVIR TAB	MESH	D000963	Antimetabolites
49702021718	TRIZIVIR TAB	MESH	D018894	Reverse Transcriptase Inhibitors
49702021718	TRIZIVIR TAB	MESH	D019380	Anti-HIV Agents
49702021718	TRIZIVIR TAB	Multum	327	antiviral combinations

In contrast, Table 3 shows a random sample of 8 NDCs that did not match. Some of the NDCs have drug descriptions that appear to be truncated, e.g. "HYDROCO" likely for hydrocodone. Others, such as "CVS ALLERGY TAB 180MG", have descriptions that do not directly indicate the active ingredients and have NDCs that could not be found in RxNorm. However, in an analysis of 20 unmapped NDCs, 9 did appear in the RxNorm ontology. 3,124 out of 51,490 NDCs were filtered out as being non-drug related. 45,508 out of the remaining 48,366 NDCs (94.0%) were assigned to one or more Multum categories. Multum had the highest coverage because VantageRx ensures that every Multum drug identifier has at least one assigned category in the Multum system. Figure 2 details unique counts of NDCs, drug codes, ingredients, and categories at major steps. Table 4 lists each classification system's coverage. Since each NDC can map to many ingredients and each ingredient be linked to several categories across the different classification systems, each NDC was typically associated with many drug categories. Subsequently, each NDC, on average, was associated with 21.6 categories over the six drug classification systems.

Table 3. Examples of unmatched NDCs.

NDC	Description
68115057200	METHADONE TAB 5MG
68258903601	XELODA TAB 500MG
43353072153	HYDROCO/APAP TAB 7.5-325
50428847260	CVS ALLERGY TAB 180MG
43353076780	BUPROPION TAB 100MG
21695085701	NECON TAB 1/35
00403107920	AUGMENTIN TAB 875MG
54868535502	ETOPOSIDE CAP 50MG

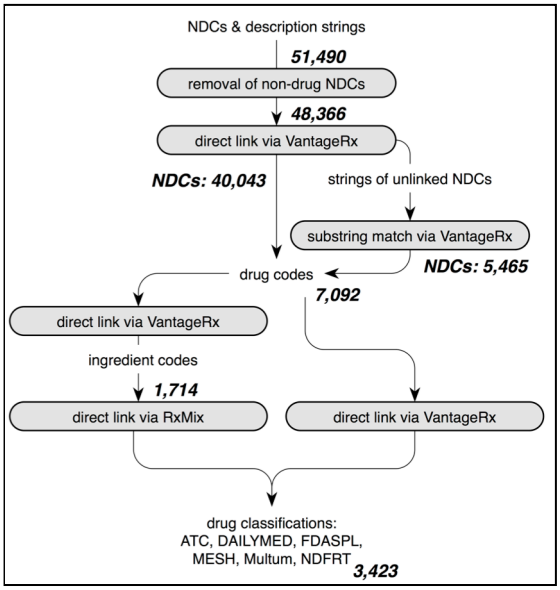

Figure 2. Evaluation results where numbers denote unique counts.

Table 4. Coverage results for each classification system.

Coverage	Number of NDCs (Percentage)
Total in Data Set	48,366 (100%)
ATC	42,780 (88%)
DAILYMED	37,852 (78%)
FDASPL	39,451 (82%)
MESH	41,416 (86%)
Multum	45,508 (94%)
NDFRT	42,578 (88%)

Two errors were found (0.4% error rate), during the visual inspection and manual search of 500 randomly sampled entries in the generated NDC-to-drug-classification table. NDC 51079094701 – "DILTIAZEM CAP 120MG ER" was correctly mapped to the ingredient diltiazem hydrochloride and NDC 67544064045 = "LORAZEPAM TAB 1MG" was correctly mapped to the ingredient lorazepam, but both were also associated with dextrose. Dextrose is associated with another packaging form and route of administration.

Discussion

The method achieves 94.0% coverage, a noticeable achievement over our initial attempts using simple, direct applications of RxNorm (60.0%) or Multum (77.8%). For the large scale insurance claims data set we used in our evaluation, the 94.0% coverage allowed 98.2% of the prescription claims to be categorized. The high coverage suggests that the unmapped NDCs have a low prevalence in the dataset. Compared to prior studies, the achieved coverage is unusually high, although a rigorous comparison is difficult because the evaluation sets are different. One key difference is our classification technique relies on both an NDC and a description string. We leverage the fact that most operational databases contain such a description column to permit human readability.

The evaluation included a spot analysis, where promisingly, nearly half (9 out of 20) of the unassigned NDCs could be identified using RxNorm. In related work, 84.2% of NDCs were successfully mapped using RxNorm, when also considering obsolete NDCs present in earlier versions of RxNorm[22]. This finding suggests an appropriate next step in the development would be to link both through Multum and RxNorm. Better leveraging the drug descriptions associated with the NDCs is another direction for future research. We used a simple string processing approach for excerpting the first word in the description and attempting to directly match it with the first segment of the Multum name. Yet, some of the NDCs' description strings appear truncated from their commonly used names. Perhaps then, more sophisticated text processing algorithms would increase the string match success rate. Additionally, we used a 2011 version of VantageRx because it was available to us, even though the evaluation data set ranged from 2010 to 2013. Thus, merging with RxNorm, enhancing the string processing, and upgrading VantageRx might achieve further reductions in unassignment rates.

We also took a small sample (500 rows – 0.05%) of the output table, manually checked the assignments, and found only two errors. This provides evidence of a high degree of accuracy, but the sample size is admittedly small. Given the size of our data set, more checking, particularly among the string matched mappings would be desirable. However, with such a large number of relations, manually vetting a significant fraction of the records is intractable. In fact, previous studies have shown that errors are known to exist within formal ontology systems[23]. One possibility would be to run two other drug classifications procedures and inspect situations where these other two procedures agree with each other, but not the method presented here.

Using a dedicated server, our implementation generated drug classification assignments in under 3 minutes. The solution is nearly automated, with only two areas requiring modification when working with a new data set. There are two string manipulations, one to help to pre-filter out non-drug identifiers and the other to process the description strings for matching. This method can be reused for any analysis off of the data set, or even when new records are entered into the database in question.

Conclusion

This pipeline is a viable solution for classifying NDCs. Key to the high coverage rates is the performance of the string matching routine. The solution is designed for large-scale drug classification tasks and indeed the solution completed its processing in a timely manner, particularly considering that each NDC is mapped to many categories across several classifications system simultaneously. Thus, we feel that the technology is a useful option whenever NDCs within a large-scale database require classification. Indeed, few doubt the value of such databases; the method's contribution is a useful tool for converting that data into information from which to glean biomedical insights.

Acknowledgments

This work has been supported by R01 GM104303 - Instrumenting i2b2 for Improved Medication Research: Adding the Patient Voice (PI: Dr. Kenneth D. Mandl) and by T15LM007092 - Boston Area Research Training Program in Biomedical Informatics (PI: Dr. Alexa T. McCray).

References

1. Brownstein JS, Sordo M, Kohane IS, Mandl KD. The tell-tale heart: population-based surveillance reveals an association of rofecoxib and celecoxib with myocardial infarction. PLoS ONE. 2007;2(9):e840.
2. Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet. 2012 Jun;13(6):395–405.
3. Alvarez CA, Clark CA, Zhang S, Halm EA, Shannon JJ, Girod CE, et al. Predicting out of intensive care unit cardiopulmonary arrest or death using electronic medical record data. BMC Med Inform Decis Mak. 2013;13(1):28.
4. Huang SH, LePendu P, Iyer SV, Tai-Seale M, Carrell D, Shah NH. Toward personalizing treatment for depression: predicting diagnosis and severity. J Am Med Inform Assoc. 2014 Nov;21(6):1069–75.
5. Longhurst CA, Harrington RA, Shah NH. A “green button” for using aggregate patient data at the point of care. Health Aff (Millwood). 2014 Jul;33(7):1229–35.
6. Cami A, Manzi S, Arnold A, Reis BY. Pharmacointeraction network models predict unknown drug-drug interactions. Medina MA, editor. PLoS ONE. 2013;8(4):e61468.

7. Simonaitis L, McDonald CJ. Using National Drug Codes and drug knowledge bases to organize prescription records from multiple sources. American Journal of Health-System Pharmacy. 2009 Sep 18;66(19):1743–53.
8. Saitwal H, Qing D, Jones S, Bernstam EV, Chute CG, Johnson TR. Cross-terminology mapping challenges: A demonstration using medication terminological systems. J Biomed Inform. Elsevier Inc; 2012 Aug 1;45(4):613–25.
9. Defalco FJ, Ryan PB, Soledad Cepeda M. Applying standardized drug terminologies to observational healthcare databases: a case study on opioid exposure. Health Serv Outcomes Res Methodol. 2013 Mar;13(1):58–67.
10. Zhou X, Murugesan S, Bhullar H, Liu Q, Cai B, Wentworth C, et al. An evaluation of the THIN database in the OMOP Common Data Model for active drug safety surveillance. Drug Saf. Springer International Publishing AG; 2013 Feb;36(2):119–34.
11. U.S. Department of Veterans Affairs, Veterans Health Administration. National Drug File – Reference Terminology (NDF-RT™) Documentation; 2015 Feb [cited 2015 Sep 21]. Available from: http://evs.nci.nih.gov/ftp1/NDF-RT/NDF-RT%20Documentation.pdf
12. WHO Collaborating Centre for Drug Statistics Methodology. WHOCC – Structure and principles; [updated 2011 Mar 25; cited 2015 Jan 29]. Available from: http://www.whocc.no/atc/structure_and_principles/
13. Cerner Multum™. VantageRx™ Database; [cited 2015 Sep 21]. Available from: http://www.multum.com/vantagerxdatabase.html
14. Peters L, Mortensen J, Nguyen T, Bodenreider O. Enabling complex queries to drug information sources through functional composition. Stud Health Technol Inform. 2013;192:692–6.
15. U.S. National Library of Medicine. Medical Subject Headings (MESH®) Fact Sheet; [updated 2013 Dec 9; cited 2015 Mar 1]. Available from: http://www.nlm.nih.gov/pubs/factsheets/mesh.html
16. U.S. National Library of Medicine. DailyMed; [cited 2015 Sep 21]. Available from: http://dailymed.nlm.nih.gov/dailymed/index.cfm
17. U.S. Food and Drug Administration. Structured Product Labeling Resources; [updated 2015 Jan 14; cited 2015 Mar 1]. Available from: http://www.fda.gov/ForIndustry/DataStandards/StructuredProductLabeling/ucm2005542.htm
18. United States National Library of Medicine. RxNorm; [updated 2014 Dec 1; cited 2014 Nov 13]. Available from:. http://www.nlm.nih.gov/research/umls/rxnorm/
19. The National Center for Biomedical Ontology. [cited 2015 Sep 21] Available from: http://bioportal.bioontology.org/
20. United States National Library of Medicine. DailyMed. [cited 2015 Sep 21] Available from: http://dailymed.nlm.nih.gov/
21. WHO Collaborating Centre for Drug Statistics Methodology. Anatomical Therapeutic Chemical (ATC) Classification System. [cited 2015 Sep 21] Available from: http://www.whocc.no/atc_ddd_index/
22. Peters L, Bodenreider O. Approaches to Supporting the Analysis of Historical Medication Datasets with RxNorm. AMIA Annu Symp Proc. 2015:1034-41.
23. Mortensen JM, Minty EP, Januszyk M, Sweeney TE, Rector AL, Noy NF, et al. Using the wisdom of the crowds to find critical errors in biomedical ontologies: a study of SNOMED CT. J Am Med Inform Assoc. 2014 Oct 23.