AMIA Annu Symp Proc 2015:1034-1041.

Approaches to Supporting the Analysis of Historical Medication Datasets with RxNorm

Lee B. Peters, M.S., Olivier Bodenreider, M.D., PhD
National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA

Abstract

Objective: *To investigate approaches to supporting the analysis of historical medication datasets with RxNorm.* ***Methods:*** *We created two sets of National Drug Codes (NDCs). One is based on historical NDCs harvested from versions of RxNorm from 2007 to present. The other comprises all sources of NDCs in the current release of RxNorm, including proprietary sources. We evaluated these two resources against four sets of NDCs obtained from various sources.* ***Results:*** *In two historical medication datasets, 14-19% of the NDCs were obsolete, but 91-96% of these obsolete NDCs could be recovered and mapped to active drug concepts.* ***Conclusion:*** *Adding historical data significantly increases NDC mapping to active RxNorm drugs. A service for mapping historical NDC datasets leveraging RxNorm was added to the RxNorm API and is available at https://rxnav.nlm.nih.gov/.*

Introduction

Many electronic medical systems identify a drug product by using a National Drug Code (NDC). For example, NDCs are used for identifying prescription drugs in the Medicare "Part D" database, as well as by many pharmacies, pharmacy benefit managers and health insurance companies [1-10]. NDCs represent not only the product's characteristics (dosage strength and form), but also manufacturer and packaging information. Because of their specificity, NDCs tend to be less stable identifiers compared to other drug vocabularies, such as RxNorm. For example, NDCs can become obsolete not only due to the product discontinuation for the usual safety reasons, but also due to discontinuation by a manufacturer for business reasons, or due to changes in packaging (e.g., pack size). In some cases, the drug may still be produced (with the same identifier in RxNorm), but by a different manufacturer or with a different pack size (i.e., under a different NDC).

Because NDCs are widely used drug identifiers, there is a great need and interest in mapping the NDC for a drug product into a standardized RxNorm name for use in electronic medical systems. In fact, the most used function in the RxNorm API is *findRxcuiById*, used to map a variety of drug identifiers to RxNorm, and NDC is the type of identifier most often converted to RxNorm with our API. More specifically, our API received 155 million *findRxcuiById* requests in 2014, 123 million (79%) of which specify NDCs.

In addition to its own identifiers, RxNorm maintains a collection of curated, up-to-date NDCs and therefore supports the mapping of current NDCs to RxNorm. However, in addition to mapping NDCs to RxNorm for current datasets or transactions from health information networks, researchers have shown interest in analyzing historical medication datasets. One such dataset is the Medicare "Part D" dataset. While RxNorm maintains history information for its own identifiers, it does not keep track of obsolete NDCs and their connection to the drugs they referred to is lost. Similarly, the NDC database maintained by the FDA only contains currently valid NDCs.

Our objective is to investigate approaches to supporting the analysis of historical medication datasets with RxNorm. More specifically, we explore NDCs collected from earlier versions of RxNorm and NDCs provided by drug vocabularies integrated in RxNorm (but not curated by RxNorm), in their coverage of several large datasets of NDCs collected from various sources and time periods. Ultimately, our goal is to support the development of a new API function for mapping legacy NDCs to RxNorm, informed by the findings of this investigation.

Background

National Drug Code (NDC). The NDC is a universal product identifier for human drugs in the United States. The Drug Listing Act of 1972 requires registered drug establishments to provide the Food and Drug Administration (FDA) with a current list of all drugs manufactured, prepared, propagated, compounded, or processed by it for commercial distribution. Drug products are identified and reported using the NDC.

The NDC is represented by a unique 10-digit, 3-segment number. The first segment of the NDC identifies the labeler (manufacturer, distributer or re-packager). The second segment is the product code, which identifies the specific strength, dosage form (i.e., capsule, tablet, liquid) and formulation of a drug for a specific manufacturer.

The third segment is the package code, which identifies package sizes and types. The first segment (labeler code) is assigned by the FDA and the second and third segments are provided by the labeler.

Examples of 10-digit, 3-segment NDCs:

54868-1048-1 (5-4-1 format)

55111-476-79 (5-3-2 format)

0179-0111-70 (4-4-2 format)

Many systems, including RxNorm, convert these forms into an 11-digit NDC derivative, which pads the labeler, product, or package code segments of the NDC with leading zeroes wherever they are needed to create a 5-4-2 format without the dashes. Examples of conversion:

54868-1048-1 => 54868104801

55111-476-79 => 55111047679

0179-0111-70 => 00179011170

RxNorm is a standardized nomenclature for medications produced and maintained by the U.S. National Library of Medicine (NLM) in cooperation with proprietary vendors [11]. RxNorm concepts are linked by NLM to multiple drug identifiers for each of the commercially available drug databases within the UMLS® Metathesaurus®. In addition to integrating names from existing drug vocabularies, RxNorm creates standard names for clinical drugs. RxNorm also contains NDC codes in the 11-digit NDC derivative format described above. The NDCs curated by RxNorm are derived from two terminologies – DailyMed and First Data Bank. Some other source vocabularies also contribute NDCs to RxNorm. However, these additional NDCs are not curated by RxNorm. RxNorm is updated monthly and each version only contains active NDCs, i.e., those NDCs which reflect the current state of availability of drug products. With each new version of RxNorm, some are added, while others that have become obsolete are removed. RxNorm does not keep track of obsolete NDCs.

The RxNorm API [12] provides functionality to access the RxNorm dataset, including mapping from NDCs to obtain the RxNorm concept identifier (RxCUI). It accepts the NDC in the 10-digit, 3-segment sequence or as the 11-digit derivative. Only active NDCs (curated by RxNorm in the latest release) can be mapped to RxNorm concepts.

Related work. Hanna *et al* developed a historical NDC dataset to use in the Drug Ontology [13]. While our objective is in part similar to theirs, their main goal was to harvest historical NDCs. In contrast, the specific contribution of this work is twofold. First we investigate the enrichment of a reference set of NDC not only with historical NDCs, but also with NDCs from other drug information sources. More importantly, we provide a comprehensive evaluation of the impact of using a richer set of NDCs by measuring the results on several large sets of NDCs from various sources and time periods.

Methods

Our investigation of approaches to supporting the analysis of historical medication datasets with RxNorm can be summarized as follows. First we describe two approaches to enriching RxNorm with additional NDCs:

1. Collect curated NDCs from earlier versions of RxNorm,
2. Collect NDCs from all drug vocabularies in the latest release of RxNorm (curated or not by RxNorm).

We then evaluate these two sets of NDCs in their coverage of large datasets of NDCs collected from various sources and time periods.

Enriching RxNorm with additional NDCs

To enrich RxNorm with additional NDCs, we need to acquire NDCs from some source and to associate each NDC with a valid RxCUI in the current version of RxNorm.

Collect curated NDCs from earlier versions of RxNorm. As mentioned above, the RxNorm dataset is restricted to currently valid NDCs and does not contain historical information regarding legacy NDCs. To create an NDC

history, we retrieved the monthly releases of the RxNorm dataset starting in July 2007 through the March 2015 release to track all the curated NDCs active at any time during this time period. For each NDC, the start and end times of the period of activity were recorded, as well as the concept identifier in RxNorm with which it was associated. For some NDCs, the RxNorm concept identifier originally associated with the NDC became obsolete and was remapped. For example, NDC 00002036303 (Darvocet-N tablets) was originally linked to RxCUI = 687241, but that concept was later mapped to 849692. We refer to this set of NDCs and related information extracted from historical RxNorm versions as the History data.

Collect NDCs from all the drug vocabularies integrated in RxNorm. There exist in the RxNorm dataset NDCs which are provided by drug vocabularies integrated in RxNorm, but not curated by RxNorm. These NDCs are not part of the active set of NDCs and are not retrieved by the RxNorm API. The reasons for these NDCs not being in the active set could be that they deal with products, such as needles or syringes or that they represent experimental or unapproved drugs by the FDA. Moreover, these NDCs may come from proprietary sources and may not be publicly available. In this investigation, we looked at the March 2015 release of the RxNorm dataset and harvested all the NDCs from all the drug vocabularies. We refer to this set of NDCs and related information extracted as the All Sources data.

Associating NDCs with active RxCUIs

To determine if an NDC is active, we used the RxNorm API method *findRxcuiById*. For active NDCs, the RxNorm concept identifier (RxCUI) is identified. For obsolete NDCs (NDCs that were once active but are no longer active), the last active time period is identified along with the last known RxCUI from the historical data. We use the RxNorm API to determine if the RxCUI associated with the NDC is active, has been remapped, or is inactive. For example, the obsolete NDC 0018202833 (Bacitracin Ointment) was last mapped to RXCUI=308509 but that RXCUI was later remapped to 1366116. NDCs which are not contained in historical data but are contained in All Sources (which we call "Alien" NDCs) have an RxCUI associated with them, and we use the RxNorm API to determine if that RxCUI represents an active RxNorm drug product.

Based on the presence of an NDC in the various sources, we determine the status of an NDC in the following way.

- ***Active***. The NDC is currently recognized by RxNorm as an active drug code (i.e., is part of the curated NDCs from the latest release of RxNorm). All active NDCs are associated with an active RxCUI by definition, and can be found by using the RxNorm API. All active NDCs are also contained in both the History data and the All Sources data.
- ***Obsolete***. The NDC is no longer active, but was in the past (i.e., was part of the curated NDCs from some earlier version of RxNorm). Some obsolete NDCs can be associated with an active RxCUI (e.g., if the RxCUI to which they were originally associated is still active or can be remapped to an active RxCUI). These NDCs are found in the History data.
- ***Alien***. The NDC is not recognized by RxNorm as an active drug code, nor has it been in the past, but it is currently contained in at least one drug vocabulary. This indicates the possibility of an out of scope drug product. Most of these NDCs are not associated with an active RxNorm concept. These NDCs can be found in the All Sources data (and not in the History data).
- ***Unknown***. The NDC is not found in either the History data or All Sources data.

The procedure for determining the NDC status goes through the steps listed above in order and stops when the NDC is identified. Once the NDC is identified from one of the datasets, we use the procedure outlined above to find the active RxCUI (if it exists) associated with the NDC. In order to facilitate the evaluation, we created an application for looking up the NDCs and their status from a database.

Evaluation

In order to evaluate the practical benefit of using additional NDC sources to the analysis of medication datasets, we acquired several large datasets of NDCs collected from various sources and time periods. We performed a quantitative evaluation of the coverage each dataset. Additionally, we performed a qualitative analysis of the NDCs from one of these sources for which no mapping to RxNorm could be found.

Quantitative evaluation. To test the NDC status function, we used sets of NDCs from three distinct sources:

- ***Medicare NDCs***. This dataset came from a random sample of Medicare Part D patients who enrolled in 2009. The set contains 27,186 unique NDCs prescribed to these patients in 2011, as well as the frequency of prescription for each NDC.
- ***Private insurance NDCs***. This dataset came from a large private health insurance group and corresponds to NDCs collected during the period January 2010 to May 2014. The set contains 51,490 unique NDCs.
- ***API log file NDCs***. We took the NDCs specified in the RxNorm API calls to *findRxcuiById* for a single month (January 2015) from the API log files. This API call allows the user to find the RxNorm concept associated with the user specified NDC. We removed any input that was not in the 10-digit, 3-segment format or the 11-digit derivative format and converted all to the 11-digit derivative format. The resulting dataset contained 372,705 unique NDC entries.
- ***FDA NDC list***. A reference list of approved NDCs for drug products exists from the FDA. We downloaded this list for http://www.fda.gov/Drugs/InformationOnDrugs/ucm142438.htm in our investigation to check against the NDCs in the History and All Sources data. The FDA list contained 167,748 NDCs.

Qualitative evaluation. One of the authors (OB), a physician, reviewed all the NDCs from the Medicare dataset for which no mapping to an active RxNorm drug could be found. This evaluation was made possible by the fact that this source provided a generic drug name for each NDC. In practice, we performed a manual review assisted by the use of regular expressions to capture frequently occurring words, corresponding to five major categories:

- Supplies (e.g., needle, syringe, lancet)
- Vitamins and dietary supplements (e.g., hyoscyamine, ferrous sulfate, carotene)
- Cold medicine (e.g., pseudoephedrine, methorphan, menthol)
- Other kinds of over-the-counter drugs (e.g., fluoride, glycerin, ointment)
- Potential prescription drugs (e.g., furosemide, insulin, nifedipine)

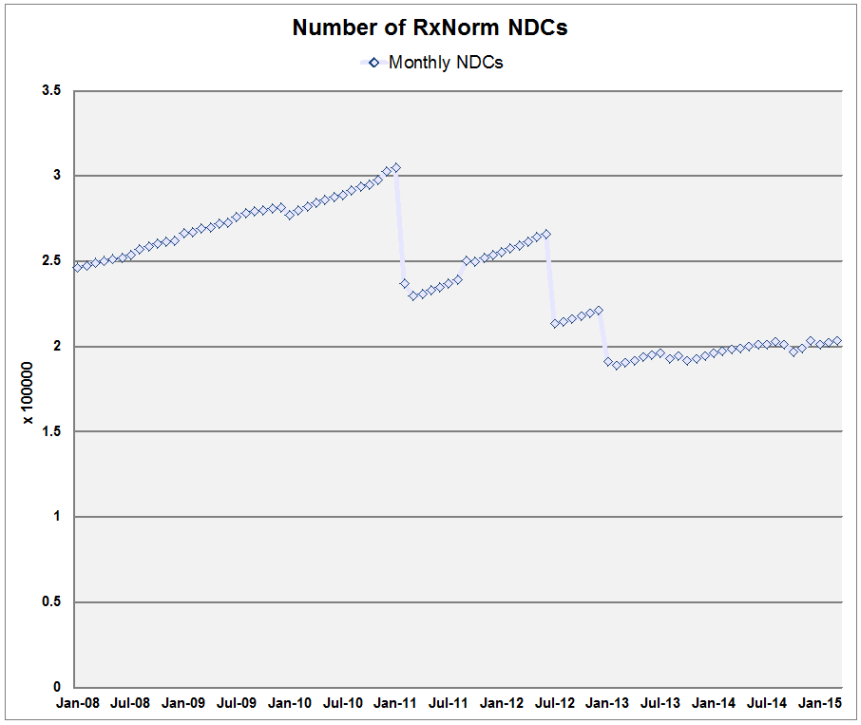

Figure 1. Active NDCs in RxNorm

Results

Enriching RxNorm with additional NDCs

Collect curated NDCs from earlier versions of RxNorm. The number of unique NDCs in the History data totaled 445,039. As shown in Figure 1, the number of active NDCs in each monthly version RxNorm fluctuates over time, reflecting not only addition of new drugs, but also curation efforts to eliminate obsolete or unreliable NDCs. In recent years, the number of NDCs curated by RxNorm is about 200,000. The History data includes 418,287 NDCs which can be linked to active RxNorm concepts, which includes 214,876 NDCs not already covered by (active) RxNorm NDCs.

Collect NDCs from all the drug vocabularies integrated in RxNorm. The NDCs from the All Sources data include 607,451 NDCs. There are 178,284 (alien) NDCs not covered by the History data. Of these alien NDCs, only 6,485 are linked to active RxNorm concepts.

Overlap between History NDCs and All Sources NDCs. Figure 2 shows the overlap of the NDCs between the two datasets. The numbers in parentheses indicate the number of NDCs that are linked to active RxNorm concepts. There are 623,323 unique NDCs contained in the extended set of NDCs composed of the union of the History and All Sources sets, of which 424,772 (68%) are linked to active RxNorm concepts. The History data contains 71% of the total NDCs and 98% of the total NDCs which are linked to active RxNorm concepts. The All Sources data contains 97% of the total NDCs and 97% of the total NDCs which are linked to active RxNorm concepts.

Figure 2. NDCs in History and All Sources
(Numbers in parentheses denote NDCs linked to active RxNorm concepts)

Evaluation

Medicare NDCs

The results returned by the NDC status function for the Medicare NDC dataset are shown below (Table 1). Of the 27,186 NDCs in this dataset, 26,528 (97.6%) could be recognized when using the extended set of NDCs ("active", "obsolete" or "alien" status), leaving only 658 (2.4%) unknown NDCs, compared to 22% unknown NDCs when only using the NDCs from the current release ("active" status). Moreover, 25,220 of all the NDCs could be linked to active RxNorm concepts, of which 24,924 (98.8%) were present in the History dataset ("active" or "obsolete" status).

Table 1. NDC status for the Medicare NDC dataset

NDC status	Total # of NDCs	# of NDCs linked to active concepts
Active	21281	21281
Obsolete	3806	**3643**
Alien	1441	**296**
Unknown	658	0
Total	**27186**	**25220**

Private insurance NDCs

The results returned by the NDC status routine for the private insurance NDC dataset are shown below (Table 2). Of the 51,490 NDCs in this dataset, 49,767 (96.7%) could be recognized when using the extended set of NDCs ("active", "obsolete" or "alien" status), leaving only 1723 (3.3%) unknown NDCs, compared to 33% unknown NDCs when only using the NDCs from the current release ("active" status). Moreover, 43,713 of all the NDCs could be linked to active RxNorm concepts, of which 43,359 (99.2%) were present in the History dataset ("active" or "obsolete" status).

Table 2. NDC status for the private insurance NDC dataset

NDC status	Total # of NDCs	# of NDCs linked to active concepts
Active	34529	34529
Obsolete	9724	**8830**
Alien	5514	**354**
Unknown	1723	0
Total	**51490**	**43713**

API log file NDCs

The results returned by the NDC status routine for the API log file NDC dataset are shown below (Table 3). Of the 372,705 NDCs in this dataset, 299,230 (80.3%) could be recognized when using the extended set of NDCs ("active", "obsolete" or "alien" status), leaving "only" 73,475 (19.7%) unknown NDCs, compared to 63% unknown NDCs when only using the NDCs from the current release ("active" status). Moreover, 222,393 of all the NDCs could be linked to active RxNorm concepts, of which 218,216 (98.1%) were present in the History dataset ("active" or "obsolete" status). The high percentage of unknown NDCs in this dataset (compared to historical medication datasets) could be explained by the fact that these NDCs came from many users, whose intentions are unclear and whose sources may have included drugs and medical products that are out of scope for RxNorm. This dataset also has a much larger proportion of Alien NDCs.

Table 3. NDC status for the API log file NDC dataset

NDC status	Total # of NDCs	# of NDCs linked to active concepts
Active	137174	137174
Obsolete	94476	**81042**
Alien	67580	**4177**
Unknown	73475	0
Total	**372705**	**222393**

FDA NDCs

We compared the NDCs in the History and All Sources data to an external source, the FDA NDC list. When we examined the coverage of these NDCs in the FDA list, we found that 99.8% of the FDA NDCs were contained in either the History or All Sources data. On examination of the missing FDA NDCs, we found several were not valid NDCs (they contained letters) and several others we examined were for drug products that are out of scope for RxNorm.

Qualitative evaluation. We reviewed a total of 1966 unique NDCs (from the Medicare dataset) with no mapping to RxNorm, corresponding to 313,659 prescriptions. As shown in Table 4, the overwhelming majority (82%) of these prescriptions correspond to supplies. While there are quite a number of potential prescription drugs among the list, these NDCs represent a minute proportion of the entire Medicare dataset.

Table 4. Categorization of the NDCs with no mapping found to RxNorm

Category	# of NDCs	# prescriptions
Supplies	910	258,218
Vitamins and dietary supplements	673	19,825
Cold medicine	97	7504
Other kinds of OTC drugs	162	784
Potential prescription drugs	124	27,328
Total	1,966	313,659

Discussion

Findings. The results indicate there is much to be gained from the use of historical NDC data to allow obsolete NDCs to be linked to active RxNorm concepts. In the Medicare and private insurance datasets, 14-19% of the NDCs were obsolete, but 91-96% of these obsolete NDCs could be linked to an active RxNorm concept using the History data. In contrast, while the use of all sources of NDCs in RxNorm made it possible to recognize many additional NDCs, relatively few of these Alien NDCs were linked to active RxNorm concepts and the Alien data contributed very little in additional mapping to RxNorm concepts.

Comparison with the FDA list showed that the extensive coverage of the FDA NDCs eliminates the need for it to be included in our service that identifies NDCs.

Additionally, the qualitative analysis done on a set of NDCs without mappings to active RxNorm concepts indicate most of these NDCs correspond to entities other than drugs, mostly supplies.

Application. This investigation clearly demonstrates that the analysis of historical medication datasets can benefit from exploiting sources of NDCs beyond the active NDCs (i.e., the NDCs curated by RxNorm present in the latest release). Namely, we showed that a large number of obsolete NDCs in the History data can be linked to an active RxNorm drug. Moreover, additional recognition of NDCs is supported by the All Sources data.

Until recently, the RxNorm API only supported the mapping of active NDCs to RxNorm concepts. The findings of this investigation informed the design of a new API function, *getNDCStatus*, to support the mapping of historical NDCs to RxNorm. This function explores the History dataset and returns all RxNorm concepts ever associated with a given NDC, along with the period when this association was active. For example, the API indicates that the obsolete NDC 00364666854 was associated with the RxNorm concept 312656 (Promazine 50 MG/ML Injectable Solution) between June 2007 and January 2011. When an NDC is associated with several RxNorm concepts at different time periods, users can select the RxNorm concept corresponding to the date of prescription, if known. The function also indicates when an NDC is recognized as "alien" (i.e., is not curated by RxNorm). This API function was released in June 2015 and is available at https://rxnav.nlm.nih.gov/.

Conclusion

The large percentage of recovered active concepts from obsolete NDCs in the three datasets is a positive indication that a service to identify obsolete NDCs and their active concepts from the past will be beneficial for the analysis of historical medication datasets. An smaller added benefit is the identification of active concepts for NDCs from other sources in RxNorm.

Acknowledgments

This work was supported by the Intramural Research Program of the NIH, National Library of Medicine. We would like to thank Drs. Clem McDonald and Mallika Mundkur from the National Library of Medicine, who provided the Medicare use case. Similarly, our thanks go to Drs. Mark Homer and Ken Mandl from the Children's Hospital Informatics Program in Boston, who provided the use case with the private insurance.

References

1. Broverman C, Kapusnik-Uner J, Shalaby J, Sperzel D. A concept-based medication vocabulary: an essential requirement for pharmacy decision support. Pharm Pract Manag Q 1998;18(1):1-20.
2. Cai B, Katz L, Alexander CM, Williams-Herman D, Girman CJ. Characteristics of patients prescribed sitagliptin and other oral antihyperglycaemic agents in a large US claims database. Int J Clin Pract 2010;64(12):1601-8.
3. Chung CP, Rohan P, Krishnaswami S, McPheeters ML. A systematic review of validated methods for identifying patients with rheumatoid arthritis using administrative or claims data. Vaccine 2013;31 Suppl 10:K41-61.
4. Hoffman V, Xue F, Gardstein B, Skerry K, Critchlow CW, Enger C. Development and evaluation of an algorithm to identify users of Prolia((R)) during the early postmarketing period using health insurance claims data. Pharmacoepidemiol Drug Saf 2014;23(9):993-8.
5. Janzen ML, Dombrovskiy VY, Galinanes EL, Vogel TR. Clopidogrel and 1-year freedom from amputation after endovascular lower extremity revascularization in the medicare population. Vasc Endovascular Surg 2014;48(7-8):509-15.
6. Masseria C, Buikema AR, Liu F, Krishnarajah G. Mixing of diphtheria, tetanus and acellular pertussis (DTaP) vaccines in a population of children in managed care. Hum Vaccin Immunother 2015:0.
7. Simonaitis L, McDonald CJ. Using National Drug Codes and drug knowledge bases to organize prescription records from multiple sources. Am J Health Syst Pharm 2009;66(19):1743-53.
8. Strykowski J, Hadsall R, Sawchyn B, VanSickle S, Niznick D. Bar-code-assisted medication administration: a method for predicting repackaging resource needs. Am J Health Syst Pharm 2013;70(2):154-62.
9. Thorpe CT, Johnson H, Dopp AL, Thorpe JM, Ronk K, Everett CM, et al. Medication oversupply in patients with diabetes. Res Social Adm Pharm 2014
10. Wilner AN, Sharma BK, Thompson A, Soucy A, Krueger A. Diagnoses, procedures, drug utilization, comorbidities, and cost of health care for people with epilepsy in 2012. Epilepsy Behav 2014;41:83-90.
11. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. J Am Med Inform Assoc 2011;18(4):441-8.
12. RxNorm API: http://rxnav.nlm.nih.gov/
13. Hanna J, Joseph E, Brochhausen M, Hogan WR. Building a drug ontology based on RxNorm and other sources. J Biomed Semantics 2013;4(1):44.