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 evaluate the use of NDC historical data from past RxNorm versions as well as NDC data from other (non-curated) drug sources in RxNorm to increase the number of NDCs that can be mapped of active RxNorm concepts. Methods: We created a historical active NDC data set from version of RxNorm 2005 to present. We also created a data set comprising all data sources of NDCs in the RxNorm data set current release. We evaluated these sets against four sets of NDCs obtained from various sources. Results: The mapping of NDCs to active RxNorm concepts increased 15-20%, and recognition of the NDCs increased over 20% to 97% total recognition. Conclusion: Adding historical data and NDCs from other sources significantly increases NDC recognition and mapping to active RxNorm concepts and will be added as a service in the RxNorm API.

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 findRx cuiById, 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 findRx cuiById 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. To our knowledge, there exists no reference set of legacy NDCs to support the analysis of historical medication datasets containing such 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, distributor or repackager). 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

<table>
<thead>
<tr>
<th>Code</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>5486810481</td>
<td>5-4-1</td>
</tr>
<tr>
<td>5511147679</td>
<td>5-3-2</td>
</tr>
<tr>
<td>0179011170</td>
<td>4-4-2</td>
</tr>
</tbody>
</table>

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:

<table>
<thead>
<tr>
<th>Code</th>
<th>Derivative</th>
</tr>
</thead>
<tbody>
<tr>
<td>5486810481</td>
<td>54868104801</td>
</tr>
<tr>
<td>5511147679</td>
<td>55111047679</td>
</tr>
<tr>
<td>0179011170</td>
<td>00179011170</td>
</tr>
</tbody>
</table>

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 contains NDC codes in 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 result in a fixed length 5-4-2 configuration without the dashes which are normally contained in the 10 digit NDC. Each version of the data set, updated monthly, contains NDC data which reflects the current state of availability of drug products. The active NDCs in RxNorm are derived from two terminologies – DailyMed and First Data Bank. New NDCs are added monthly, and others that become obsolete are removed. Historical data regarding the NDCs is not available in the RxNorm data set, so obsolete NDCs may be removed altogether.

The RxNorm API [12] provides functionality to access the RxNorm data set, 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 zero padded derivative.

Related work. Hanna et al. developed a historical NDC data set to use in the Drug Ontology. 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.

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 NDCs from earlier versions of RxNorm,
2. Collect NDCs from drug vocabularies integrated in RxNorm (but not curated 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 NDCs from earlier versions of RxNorm. As mentioned above, the RxNorm data set 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 data set starting in July 2007 through the March 2015 release to track all the NDCs active at any time during this time period. For each NDC, the start and end times were calculated 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 assigned 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 drug vocabularies integrated in RxNorm. There exist in the RxNorm data set 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 drug products such as needles or syringes or that they represent experimental or unapproved drugs by the FDA. Moreover, these NDCs may not be publicly available. In this investigation, we looked at the March 2015 release of the RxNorm data set and harvested all the NDCs from all the drug sources. 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. 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. 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 data sets, 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 data set came from random sample of Medicare Part D patients who enrolled in 2009. The set contains 27,191 unique NDCs administered to these patients in 2011, as well as the frequency of prescription for each NDC.

- Insurance NDCs – this data set came from a large private health insurance group and corresponds to NDCs collected during the period Jan. 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 findRxuciByld 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 data set 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, carotene)
- Cold medicine (e.g., pseudoephedrine, methorphan, menthol)
- Other kinds of over-the-counter drugs (e.g., fluordride, glycerin, ointment)
- Potential prescription drugs (e.g., furosemide, insulin, nidefipine)

**Results**

*Enriching RxNorm with additional NDCs*

**Collect 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 drug vocabularies integrated in RxNorm. The NDCs from the All Sources date 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 data sets. The numbers in parenthesis indicate the number of NDCs that are contained in active RxNorm concepts.
There are 623,323 unique NDCs contained in the two sets, of which 424,772 (68%) are contained in active RxNorm concepts. The History data contains 71% of the total NDCs and 98% of the total NDCs which are contained in active RxNorm concepts. The All Sources data contains 97% of the total NDCs and 97% of the total NDCs which are contained in active RxNorm concepts.

**Evaluation**

**Medicare NDCs**

The results returned by the NDC status function for the Medicare NDC data set are shown below (Table 1). There were 25,220 NDCs which were linked to RxNorm concepts, including 3939 NDCs that are not RxNorm NDCs. The Medicare set had 78% active NDCs and 93% of NDCs had active RxNorm concepts. 98% of the NDCs were found, leaving only 2% unknown.

Table 1. NDC status for the Medicare NDC data set

<table>
<thead>
<tr>
<th>NDC status</th>
<th># of NDCs</th>
<th># active concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>21281</td>
<td>21281</td>
</tr>
<tr>
<td>Obsolete</td>
<td>3806</td>
<td>3643</td>
</tr>
<tr>
<td>Alien</td>
<td>1441</td>
<td>296</td>
</tr>
<tr>
<td>Unknown</td>
<td>658</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>27191</td>
<td>25220</td>
</tr>
</tbody>
</table>

**Insurance NDCs**

The results returned by the NDC status routine for the insurance NDC data set are shown below (Table 2). There were 43,713 NDCs which were linked to RxNorm concepts, including 9184 NDCs that are not RxNorm NDCs. The insurance set had 67% active NDCs and 85% of NDCs had active RxNorm concepts. 97% of the NDCs were found, leaving only 3% unknown.

Table 2. NDC status for the insurance NDC data set

<table>
<thead>
<tr>
<th>NDC status</th>
<th># of NDCs</th>
<th># active concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>34529</td>
<td>34529</td>
</tr>
<tr>
<td>Obsolete</td>
<td>9724</td>
<td>8830</td>
</tr>
<tr>
<td>Alien</td>
<td>5514</td>
<td>354</td>
</tr>
<tr>
<td>Unknown</td>
<td>1723</td>
<td>0</td>
</tr>
</tbody>
</table>
Total | 51490 | 43713

**API log file NDCs**

The results returned by the NDC status routine for the API log file NDC data set are shown below (Table 3). There were 222,393 NDCs which were linked to RxNorm concepts, including 85,219 NDCs that are not RxNorm NDCs. The API log file set had 37% active NDCs and 60% of NDCs had active RxNorm concepts. 80% of the NDCs were found, leaving 20% unknown. The high percent of unknown NDCs could be explained by the fact that the sources came from many users, whose intentions are unclear and whose sources may have included many more out of scope NDCs.

**Table 3.** NDC status for the API log file NDC data set

<table>
<thead>
<tr>
<th>NDC status</th>
<th># of NDCs</th>
<th># active concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>137174</td>
<td>137174</td>
</tr>
<tr>
<td>Obsolete</td>
<td>94476</td>
<td>81042</td>
</tr>
<tr>
<td>Alien</td>
<td>67580</td>
<td>4177</td>
</tr>
<tr>
<td>Unknown</td>
<td>73475</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>372705</td>
<td>222393</td>
</tr>
</tbody>
</table>

**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 product that are out of scope for RxNorm.

**Qualitative evaluation.** We reviewed a total of 1966 unique NDCs 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

<table>
<thead>
<tr>
<th>Category</th>
<th># of NDCs</th>
<th># prescriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supplies</td>
<td>910</td>
<td>258,218</td>
</tr>
<tr>
<td>Vitamins and dietary supplements</td>
<td>673</td>
<td>19,825</td>
</tr>
<tr>
<td>Cold medicine</td>
<td>97</td>
<td>7504</td>
</tr>
<tr>
<td>Other kinds of OTC drugs</td>
<td>162</td>
<td>784</td>
</tr>
<tr>
<td>Potential prescription drugs</td>
<td>124</td>
<td>27,328</td>
</tr>
<tr>
<td>Total</td>
<td>1,966</td>
<td>313,659</td>
</tr>
</tbody>
</table>
Discussion

Significance. 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. Use of all drug sources of NDCs in RxNorm greatly increased the recognition of many additional NDCs, though these Alien NDCs 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.

Application. We are encouraged by the large number of non-active NDCs in the History data that we can link to an active RxNorm drug, and additional recognition of NDCs in the All Sources data. We have started working on the development of a new service, similar to an existing RxCUI status, for NDCs. Given the large interest in mapping NDCs to RxCUIs as evidenced by the RxNorm API requests, this service will be of benefit to many users. We plan to release this service in the next few months as part of the RxNorm API.

Challenges

With development of the service, procedures need to be in place to handle the RxNorm monthly releases which result in changes to the active NDCs (additions of some, retirement of others).

Conclusion

The large percentage of recovered active concepts from obsolete NDCs in the three data sets is a positive indication that a service to identify obsolete NDCs and their active concepts from the past would be beneficial. An added benefit is the identification of active concepts for NDCs from other sources in the RxNorm data set.

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 the Children's Hospital Informatics Program in Boston, who provided the use case with the private insurance.

References