Using National Drug Codes and drug knowledge bases to organize prescription records from multiple sources

LINAS SIMONAITIS AND CLEMENT J. MCDONALD

Purpose. The utility of National Drug Codes (NDCs) and drug knowledge bases (DKBs) in the organization of prescription records from multiple sources was studied.

Methods. The master files of most pharmacy systems include NDCs and local codes to identify the products they dispense. We obtained a large sample of prescription records from seven different sources. These records carried a national product code or a local code that could be translated into a national product code via their formulary master. We obtained mapping tables from five DKBs. We measured the degree to which the DKB mapping tables covered the national product codes carried in or associated with the sample of prescription records.

Results. Considering the total prescription volume, DKBs covered 93.0–99.8% of the product codes from three outpatient sources and 77.4–97.0% of the product codes from four inpatient sources. Among the inpatient sources, invented codes explained 36–94% of the noncoverage. Outpatient pharmacy sources rarely invented codes, which comprised only 0.11–0.21% of their total prescription volume, compared with inpatient pharmacy sources for which invented codes comprised 1.7–7.4% of their prescription volume. The distribution of prescribed products was highly skewed, with 1.4–4.4% of codes accounting for 50% of the message volume and 10.7–34.5% accounting for 90% of the message volume.

Conclusion. DKBs cover the product codes used by outpatient sources sufficiently well to permit automatic mapping. Changes in policies and standards could increase coverage of product codes used by inpatient sources.

Index terms: Codes; Databases; Hospitals; National Drug Code; Pharmacy, institutional, hospital; Records

Am J Health-Syst Pharm. 2009; 66:1743-53

Clinicians need an accurate list of their patients' medications to avoid prescribing errors and provide optimal care.[1] Over time, a patient's medications will be prescribed by many different providers and dispensed by many different pharmacists, and medication records tend to be scattered. Consequently, clinicians today must gather a medication history, including both active and inactive medications, directly from their patients. Hospitals must do the same as part of the medication reconciliation process required by the Joint Commission.[2] These processes are both time intensive and error prone and beg to be automated.[3]

Almost all inpatient and outpatient pharmacies use computers to process and fill prescriptions or drug orders. In theory, health care

LINAS SIMONAITIS, M.D., is Research Scientist, Regenstrief Institute, Indianapolis, IN, and Assistant Professor, School of Medicine, Indiana University, Indianapolis. CLEMENT J. MCDONALD, M.D., is Director, Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, MD; at the time this research was conducted, he was Director Emeritus, Regenstrief Institute, and Regenstrief Professor Emeritus of Medical Informatics, School of Medicine, Indiana University, Indianapolis.

Address correspondence to Dr. Simonaitis at Regenstrief Institute, 410 West 10th Street, Suite 2000, Indianapolis, IN 46202 (lsimonaitis@regenstrief.org).

Anne Belsito, John Clifford, and Larry Lemmon are acknowledged for their assistance in data processing. Sandy Poremba is acknowledged for her assistance in manuscript preparation. The following drug knowledge base vendors are gratefully acknowledged for allowing us to examine their products at no cost: Cerner Multum, Inc.; First DataBank, Inc.; Thomson Healthcare, Inc.; and Wolters Kluwer Health, Inc. The hospitals of Indianapolis are also gratefully acknowledged for providing their formularies and deidentified prescription messages.

This work was performed at the Regenstrief Institute and supported by grant T15 LM07117 from the National Library of Medicine; by National Cancer Institute grant U01 CA91343, a Cooperative Agreement for the Shared Pathology Informatics Network; by Biomedical Information Science and Technology Initiative grant P20 GM66402 from the National Institutes of Health; and by a grant from the Indiana Twenty-First Century Research and Technology Fund for proposal ID 510040784.

The authors have declared no potential conflicts of interest.

DOI 10.2146/ajhp080221

providers should be able to obtain a complete record of all of a patient's medications by pulling the prescription records from all of the pharmacies that have served their patient. This idea inspired a consortium of pharmacy benefit management companies to create RxHub, which merged with Surescripts in 2008, in order to coordinate the aggregation of pharmacy dispensing records into a unified medication history.[4] The same idea motivated the creation of Regional Health Information Organizations (RHIOs), which aggregate clinical information of many kinds from regional sources.[5]

Inpatient and outpatient pharmacy systems employ levels of standardization that could enable the delivery of computerized prescription records to care providers. Most community pharmacies use the National Council for Prescription Drug Programs (NCPDP) Telecommunication Standard,[6] and most hospital pharmacies use Health Level Seven (HL7) pharmacy-order messages.[7] We believe that most pharmacy systems either do or could include a universal product identifier (i.e., National Drug Code [NDC])[8] in these prescription messages because they carry this identifier in their master formulary for most of the medications they dispense.

However, this universal product identifier cannot directly enable decision support or performance measurements and cannot be used to organize medication profiles and flow charts. NDCs were designed for inventory management and reimbursement. Each product labeler assigns its own NDC for every product it markets; thus, every distinct combination of brand name, dosage form, strength, and package size gets many different NDCs. As a result, products that clinicians might consider as a single medication are represented by many different NDCs. For example, amoxicillin 500-mg capsules have at least 227 distinct NDCs, and there is nothing intrinsic to these codes that ties them together. Clinically equivalent NDCs should be mapped to a higher-level code that identifies the clinically relevant concept—the generic drug, dosage form, and strength.[9-11] Following common usage, we call this the "clinical drug code."

Commercial drug knowledge base (DKB) vendors provide the clinical drug codes needed for clinical use and provide tables for mapping NDCs to their clinical drug codes. These commercial DKBs have been adopted by pharmacies and clinical care systems to assist the prescribing and dispensing processes. The National Library of Medicine (NLM) also provides a public-use DKB called RxNorm, which contains a table for mapping NDCs to their "semantic clinical drug code."[12] So, in theory, hospitals and office practices could automatically capture these prescription records from all relevant pharmacy systems, map the NDCs to their respective clinical drug codes from one DKB, and file the prescription records from all prescription sources under the appropriate clinical drug within their medical records system.

However, the success of such an effort will depend on the degree to which the drug identifiers used in pharmacy messages are carried in DKB mapping tables. If DKBs do not include NDCs in common use or if pharmacy systems use locally invented product codes in these messages, this automated mapping process will fail.

We obtained mapping tables from five DKBs to assess the degree to which DKB mapping tables can translate the codes contained in pharmacy messages. We also obtained a large sample of production NCPDP and HL7 pharmacy messages and their associated master formulary files. We then assessed the degree to which the codes in the prescription records and formularies appear in the DKB mapping tables and attempted to determine reasons for any exclusions.

Methods

Coding systems. An NDC is a 10-digit code consisting of three parts delimited by dashes: (1) the labeler segment assigned by the Food and Drug Administration (FDA) to the distributor, manufacturer, or repackager of the product, (2) the product segment, which identifies a specific drug product (e.g., Zocor 20-mg tablets), and (3) the package segment, which distinguishes different package sizes produced by one labeler (e.g., bottle of 100 tablets). FDA does not control assignment of the entire code, only the first segment. This assignment process is akin to the assignment of Internet addresses: a root code is assigned to an organization, and the organization assigns more specific codes by adding digits to that root.

The NDC was introduced in 1972 as a 10-character code with a 4-4-2 configuration to identify the labeler, product, and package segments, respectively. Later, FDA expanded the labeler segment to 5 digits, with two configurations (5-4-1 and 5-3-2). All of these configurations used dashes as delimiters to distinguish the three segments. Today, most users convert the historic codes to an 11-digit format. In this format, the first 5 digits represent the labeler segment, the next 4 digits represent the product, and the last 2 digits, the package. To convert an older NDC to this newer 5-4-2 configuration, one must add a leading zero either to a 3-digit product segment or to a 1-digit package segment. The 11-digit format (with no dashes), is the only format permitted in NCPDP messages, and is mandated by the Health Insurance Portability and Accountability Act of 1996 (HIPAA) regulations for all HIPAA transactions.[13]

The primary focus of this study was the industry-accepted product

codes found in NCPDP messages, HL7 messages, pharmacy master files, and DKB mapping tables. These industry codes are used widely by pharmacies to identify prescribed products, allowing for the aggregation of prescription records from many sources. There is no accepted name for these codes. Many refer to them casually as NDCs. However, the codes are actually a collection of (1) registered NDCs (i.e., codes for medications, including therapeutic biological products), that have been officially registered in FDA's NDC directory, (2) semiofficial NDCs (i.e., codes for medications that have been properly assigned according to FDA rules but not yet registered in FDA's database), and (3) device and supply codes.

The two types of codes used for identifying devices and supplies are the National Health Related Items Code (NHRIC) system and the Universal Product Code (UPC) system. The NHRIC system was developed in the 1970s by FDA's Center for Devices and Radiological Health.[14] FDA set aside a block of 4-digit labeler codes for the NHRIC system, designed not to overlap with the labeler codes for the NDC system. Although NHRICs consist of two segments and 10 digits, they can be converted to a string of 11 digits, in accordance with NCPDP standards, in a way analogous to that used for NDCs.

UPCs derive from an even more complex system, beyond the authority of FDA. These codes are coordinated by the global standards development organization GS1 and its U.S. member organization GS1 US (formerly known as the Uniform Code Council).[15] Different versions of these codes exist, with different lengths. However, the most common type of UPC in the United States is the 10-digit pattern found on bar codes (typically accompanied by an 11th digit on the left to indicate product type and a 12th modulus check digit on the right). This 10-digit UPC is often converted to an 11-digit string by adding an additional zero between its two 5-digit segments.

In the DKBs examined in this study, the UPCs and NHRICs are stored in the same database columns as NDCs. However, commercial knowledge bases do provide some clues to the different origin of these codes, such as an additional indicator field.

All of these categories of codes (registered NDCs, semiofficial NDCs, NHRICs, and UPCs) are delivered in the same slot as HL7 and NCPDP messages, and they all appear in the same slot of pharmacy master files. Operationally, these codes are defined by the DKB vendors, who actively gather these codes and information about the products these codes represent. For convenience purposes, in this article we will call them "NDCs," using quotation marks to remind the reader that we are referring to the above defined universe of NDC-like codes. We also include industry-assigned supply codes in this rubric for the sake of simplicity. When speaking of NDCs officially registered with FDA, we will not use quotation marks and will always precede NDC with the word "official."

Another population of prescribed codes flows in messages in the same place as "NDCs." We will call them "invented codes," because they are invented locally by pharmacies and other organizations without FDA-assigned labeler codes and are not part of any national coding system. Invented codes may appear in the "NDC" slots of local formularies and electronic prescription messages where "NDCs" are usually found. Such codes are sometimes, but not always, easily distinguishable from "NDCs" by their format.

FDA does provide an enumeration of the official NDCs in its database. But FDA's enumeration does not include many of the "NDCs" found in electronic messages. Importantly, each DKB provider has its own enumeration of "NDCs" embodied in its mapping table. But each of these is also incomplete. Indeed, no complete enumeration of assigned "NDCs" exists. Because of this, we had to develop the operational definition of invented codes described below.

Sources of prescription data. *Inpatient samples.* Sample inpatient data were obtained from the five Indianapolis hospital systems participating in the Indiana Network for Patient Care (INPC).[16] Each of the five provided its inpatient master formulary file, which defines orderable medications. Four also provided HL7 messages from their inpatient pharmacy system. One of these provided HL7 "detailed financial transaction" messages that used "NDCs" to identify the prescribed medication. The other three hospital systems provided HL7 "pharmacy/treatment encoded order" (RDE) messages—one using "NDCs" and the other two using local service codes to identify the prescribed entity. In the latter two cases, we mapped these service codes to the "NDCs" in the respective hospital's formulary master file. Two of the hospitals provided a one-week sample and two provided a one-month sample of pharmacy prescription messages.

Outpatient samples. Outpatient prescription sample data were obtained from three sources: (1) a six-month collection of NCPDP messages from a large standalone outpatient pharmacy system associated with one of the Indianapolis hospitals, (2) a six-month collection of HL7 "pharmacy/treatment dispense" (RDS) outpatient prescription records delivered to INPC by RxHub, and (3) 24 different outpatient formularies for health insurance companies from across the country provided by RxHub. The "NDCs" from the 24 formularies were aggregated into one database and treated as a single source of "NDCs."

Archival outpatient sample. All of the sources used represent real content from production systems. With the data listed above, the problems of mapping the identifiers in current prescription records to a clinical drug code could be assessed. The problems an organization would face if it wanted to extract long-term medication histories from archival sources for care or research were also assessed. A list of all the "NDCs" contained in a 12-year period in the Indiana State Medicaid prescription database were obtained for this purpose.

Of these 13 samples, 4 were obtained in the second half of 2005 and 9 in the first half of 2006; the smallest of these samples contained more than 40,000 prescriptions, and the largest contained 60 million.

DKB mapping tables. Mapping tables were obtained from each of four commercial DKBs: (1) First DataBank National Drug Data File Plus (First DataBank, Inc., San Bruno, CA), (2) Medi-Span Master Drug Data Base (Wolters Kluwer Health, Inc., Conshohocken, PA), (3) Multum Lexicon (Cerner Multum, Inc., Denver, CO), and (4) Thomson Micromedex Red Book (Thomson Corporation, Greenwood Village, CO). These four DKBs are used widely in pharmacy information systems and are listed as knowledge base options in the NCPDP database indicator field.[17] Each of these vendors kindly provided us with their core database content at no cost for this study. Each DKB supplier included its mappings from over 100,000 current "NDCs" to its own clinical drug codes. We also obtained the RxNorm knowledge base, which the NLM makes freely available to the public.

Unlike these five DKBs, the FDA NDC directory currently does not include a clinical drug coding system and is not directly comparable to the other five. We included the NDC directory (obtained from FDA's website)[8] in some of the analyses to assess its coverage of "NDCs" in common use. Table 1 provides information about the five DKBs and the FDA directory used in this study.

Prescribed supplies. Pharmacies fill prescriptions for supplies (e.g., insulin syringes, glucose test strips, gauze sponges) as well as for medications. In pharmacy messages, the codes for these supplies are treated in the same way as the codes for medications. In NCPDP messages, these prescribed supplies are identified by a UPC or NHRIC. The UPC or NHRIC is recorded in the same slot used to record "NDCs." An additional field in the NCPDP message names the coding system used, and a field is available in HL7 to do the same. FDA has reserved a separate block of labeler segments for NHRICs, so these will not overlap with official NDCs. Neither the official NDC table nor RxNorm carry such supply codes.

Within hospital-delivered HL7 messages, the prescribed product identifier can be an "NDC" or a local "service code." Each hospital pharmacy has a master formulary, which includes a record for each service code that usually also carries the associated "NDC." Thus, if pharmacies send only local service codes in their messages, these codes can usually be translated into "NDCs," NHRICs, or UPCs via their formulary table.

Preprocessing of source prescriptions. In the case of NCPDP messages, the "NDC" was extracted from the item number field of the drug segment. In the case of HL7 RDE and RDS messages, the "NDC" was extracted from the first component of field 2 in the RXE (encoding) and RXD (dispensing) segments,

Table 1.
Drug Knowledge Bases Evaluated

Drug Knowledge Base (Manufacturer)	Contact Information	Version Evaluated	No. Distinct NDCs[a]	No. Distinct Clinical Drugs[b]
NDC directory (Food and Drug Administration)	drls@cder.fda.gov 301-210-2840	Jan 2006	122,117	Not available
Medi-Span Master Drug Data Base (Wolters Kluwer Health)	medispan-support @wolterskluwer.com 800-388-3884	Feb 2006	126,042	9,836
Multum Lexicon (Cerner Multum)	info@multum.com 800-968-5886	Feb 2006	113,221	8,383
National Drug Data File Plus (First DataBank)	inquiry@firstdatabank.com 800-633-3453	Feb 2006	117,025	14,405
Micromedex Red Book (Thomson Corporation)	mdx.info@thomson.com 800-525-9083	Feb 2006	181,113	13,801
RxNorm (National Library of Medicine)	rxnorminfo@nlm.nih.gov 888-346-3656	Feb 2006	232,111	8,082

[a]NDC = National Drug Code.
[b]Numbers are not directly comparable. Please see the appendix for details of the calculation.

respectively. In the case of the HL7 detailed financial transaction messages, the "NDC" was extracted from the first component of field 9 of the financial transaction (FT1) segment. All "NDCs" were converted to the 11-digit format before any matching was performed.

All but one of the prescription sources supplied a large sample of prescription messages and their master formulary table. The one inpatient service that provided us only with a master formulary table did not currently send HL7 messages. We used the same preprocessing method to analyze the contents of messages and master formulary files.

A sample of messages from each source was examined by hand. During this examination, one of our message sources was found to have included a null value in the HL7 field that carries the prescribed entity identification (ID) number. The messages with null ID numbers carried clinical data (mostly creatinine clearance values as free text), a misuse of the HL7, which is intended to carry information about prescriptions. We excluded these messages from all calculations because they were easy to distinguish and did not represent prescriptions.

Locally invented codes. The institutions in INPC sometime invent their own local prescribed entity codes and use them in the same fields as "NDCs." Many of the invented codes could be identified during initial inspection of the messages because their format was incompatible with the format of an "NDC." Invented codes will not be present in any DKBs and will not be mappable by automatic means. An operational definition for invented codes was developed to quantify their prevalence. A code was considered invented if it (1) had more than 11 or fewer than 10 digits, (2) contained alphabetic characters, (3) began with a sequence of 5 identical digits (e.g., 11111, 22222), or (4) began with the digits 991, as encountered at only one hospital.

The first two criteria were based on the format definitions of "NDCs," and the last two were based on an empirical review of the source databases. This working definition was tested by searching through the four commercial DKBs, which contained a combined total of 249,098 distinct "NDCs." Only two records were found that would have met the definition of invented codes.

Counting and calculations. For noninvented "NDCs," all unique codes in a given message stream were counted. For invented codes, unique combinations of prescribed entity name and code were counted, as a review of the product names associated with these codes showed that the same invented code could be used to identify many different products. To determine the total number of unique codes from a given source, the number of unique noninvented codes was added to the number of unique invented code–name combinations.

We had originally assumed that the first 9 digits of the "NDC" would be the appropriate level for comparison, because the last 2 digits of the "NDC" are supposed to represent only the package size (e.g., bottle of 100 capsules), which is irrelevant for most clinical purposes. We assessed the degree to which a given DKB covered the "NDCs" from each of our sources by using the leading 9 digits of the "NDC" and by using all 11 digits. We reported the 11-digit comparison only, as we found several hundred pairs of "NDCs" that differed only by the last 2 digits but represented quite different drugs. One such pair, selected at random, is 00686-0360-10 (ipecac syrup) and 00686-0360-67 (digoxin elixir); another such pair is 55289-0033-28 (ampicillin capsule) and 55289-0033-97 (prochlorperazine tablet). Apparently, labelers and manufacturers sometimes use the last 2 digits to distinguish ingredients rather than package size. Using 9 digits did not increase the coverage of any DKB by more than 1 or 2 percentage points. The appendix provides the details of how clinical drug codes in the DKBs were counted.

Deidentification and institutional review board approval. This study was conducted using nonpatient data from master formularies and deidentified portions of prescription messages. Approval for the study was obtained from Indiana University's institutional review board.

Results and discussion

Codes in the DKB tables. The number of unique "NDCs" contained in the mapping tables from DKB providers (the four commercial DKBs and NLM) ranged from 113,221 to 232,111. The number of distinct clinical drug codes in these same tables ranged from 8,082 to 14,405 (Table 1). The differences in the number of distinct drug codes are due to differences among the rules for distinguishing supplies from medications across DKBs, the granularity at which some drugs (e.g., multivitamins) are represented, and the inclusion of special content (e.g., allergy shots) in some DKBs and not in others. Therefore, the numbers are not directly comparable. Differences in these numbers did not predict success in covering the product codes we obtained from pharmacy sources.

Distribution of product codes among all messages. The frequency distribution of code use across the total message volume is highly skewed. Across all seven message sources, a very small percentage of the unique codes (1.4–4.4%) accounted for 50% of the total message volume, and 10.7–34.5% accounted for 90% of the message volume. The thin tail of this skewed distribution included large numbers of codes that occurred just once among 40,000–66 million prescriptions. Across six of the seven message sources, "NDCs" that occurred just once in the prescription sets made up 9–18% of the

unique codes but only 0.01–0.47% of the total message volume (i.e., 40,000–66 million prescriptions). For the seventh source, the single-instance codes comprised only 3% of the unique codes. The skewing of invented codes was even more severe than that of the noninvented codes.

DKB coverage of unique product codes found in messages and formularies. For completeness, the DKB coverage of unique product codes in formularies and prescription records across all sources and DKBs is reported in Table 2.

The DKBs covered less than 95% of the unique product codes in 12 of the 13 data sets listed in Table 2. The coverage was much better for the 13th data set—the RxHub prescription messages. Two DKBs covered nearly 98% and the other three DKBs covered at least 94% of the unique codes in the RxHub messages. Overall, DKB coverage of unique codes from inpatient sources was not as high as that from outpatient sources. DKBs failed to cover 8.1–23.7% of the unique codes in inpatient messages and 4.4–13.3% of those in inpatient formularies.

A DKB did not cover a product code for one of two reasons: (1) the unrecognized code was invented by a local institutional source and could not be known by the DKB, or (2) the unrecognized code is a valid "NDC" that was not included in the DKB.

Invented codes comprised a small percentage (≤1.6%) of the unique codes from all but one outpatient source (Table 3). One outpatient source invented 20.7% of its unique codes. Across the five inpatient formularies and four inpatient message sources, invented codes comprised 1.6–12.3% and 7.5–22.9% of the unique codes, respectively. By definition, DKBs will not carry invented codes in their mapping tables, so invented codes will always decrease the rate of coverage.

DKB coverage of the total volume of product codes. Coverage of unique codes is a misleading assessment of DKB coverage because the frequency distribution of product codes in messages is highly skewed. Using unique codes as the denominator weights codes that identify 1 prescription in 1 million the same as those that identify 20,000 prescriptions in 1 million. It is more appropriate to base the assessment on total prescription volume (i.e., counting the number of prescriptions covered by a DKB and dividing it by the total number of prescriptions from a given source).

In the outpatient setting, all DKBs covered at least 93.0% of the total prescription volume across all sources (Table 4). In the case of hospital E's outpatient pharmacy, three DKBs covered at least 97.6% and another covered 99.6% of its total prescription volume. Invented codes accounted for only 0.21% of the total prescription volume from this pharmacy, com-

Table 2.
Percentage of Unique Codes Identified in Drug Knowledge Bases (DKBs)[a]

Data Source	FDA[b]	MDDB	MMSL	MMX	NDDF	RxNorm	Not Covered in Any DKB
Outpatient Sources							
Hospital E pharmacy messages (*n* = 2,834)	56.5	72.5	77.7	74.7	69.6	76.4	21.2
Medicaid archival records (*n* = 41,727)	45.5	75.8	80.3	80.9	74.5	79.3	2.4
RxHub							
Formularies (*n* = 117,151)	29.7	78.4	51.1	64.9	90.1	52.5	2.5
Messages (*n* = 7,838)	78.5	97.8	94.3	95.2	97.9	94.0	1.6
Inpatient Sources							
Hospital A							
Formulary (*n* = 3,598)	65.0	83.4	85.9	83.1	80.9	86.0	13.3
Messages (*n* = 1,641)	70.6	87.4	88.8	86.7	85.8	89.2	10.3
Hospital B							
Formulary (*n* = 2,317)	59.0	71.6	92.3	84.9	67.8	93.5	5.9
Messages (*n* = 1,716)	54.3	64.9	80.2	75.5	62.2	81.1	18.5
Hospital C							
Formulary (*n* = 3,820)	76.0	94.7	93.5	93.3	94.7	93.6	4.4
Messages (*n* = 1,206)	74.5	91.3	90.7	90.3	91.4	90.1	8.1
Hospital D							
Formulary (*n* = 2,905)	70.2	86.8	91.4	90.2	86.2	91.3	7.8
Messages (*n* = 3,439)	59.9	72.1	75.6	74.4	71.4	75.8	23.7
Hospital E formulary (*n* = 2,543)	73.5	90.8	91.1	90.1	90.8	90.4	7.2

[a]FDA = Food and Drug Administration, MDDB = Medi-Span Master Drug Data Base, MMSL = Multum Lexicon, MMX = Thomson Micromedex Red Book, NDDF = First DataBank National Drug Data File. Prescription records and formularies from the same institution are listed as separate data sources.
[b]Not literally a DKB but included for comparison purposes.

Table 3.
Summary of Distinct Identifiers by Data Source[a]

Data Source	No. Unique Codes	% Invented Codes	% Noninvented NDCs[b] Not Found in Any DKB[c]	% Noninvented NDCs Found in ≥1 DKB
Outpatient Sources				
Hospital E pharmacy messages	2,834	20.7	0.42	78.9
Medicaid archival records	41,727	0.01	2.4	97.6
RxHub				
Formularies	117,151	0.48	2.0	97.5
Messages	7,838	1.6	0.01	98.4
Inpatient Sources				
Hospital A				
Formulary	3,598	12.3	0.92	86.8
Messages	1,641	9.6	0.67	89.7
Hospital B				
Formulary	2,317	1.6	4.2	94.2
Messages	1,716	16.0	2.6	81.4
Hospital C				
Formulary	3,820	4.1	0.31	95.6
Messages	1,206	7.5	0.66	91.8
Hospital D				
Formulary	2,905	6.8	1.0	92.2
Messages	3,439	22.9	0.73	76.4
Hospital E formulary	2,543	6.2	0.94	92.9

[a]The first data column gives the number of distinct codes. The remaining three data columns give the percentage breakdown of these distinct codes.
[b]NDC = National Drug Code.
[c]DKB = drug knowledge base.

Table 4.
Percentage of Total Medication Records Covered in Drug Knowledge Base (DKB) by Data Source[a]

Data Source	FDA[b]	MDDB	MMSL	MMX	NDDF	RxNorm	Not Covered in Any DKB
Outpatient Sources							
Hospital E pharmacy messages (*n* = 464,404)	88.5	97.6	99.6	93.6	93.0	98.2	0.28
Medicaid archival records (*n* = 66,291,780)	81.0	98.7	98.7	98.2	97.3	98.7	0.17
RxHub messages (*n* = 304,855)	92.2	99.8	99.0	99.5	99.8	99.0	0.17
Inpatient Sources							
Hospital A messages (*n* = 40,855)	74.7	89.0	91.7	87.7	87.5	92.1	7.8
Hospital B messages (*n* = 149,675)	63.3	80.7	95.4	90.9	77.4	96.0	4.0
Hospital C messages (*n* = 40,622)	73.7	96.9	96.9	95.8	97.0	96.7	2.6
Hospital D messages (*n* = 396,689)	81.7	93.7	95.7	93.9	93.5	95.8	3.5

[a]FDA = Food and Drug Administration, MDDB = Medi-Span Master Drug Data Base, MMSL = Multum Lexicon, MMX = Thomson Micromedex Red Book, NDDF = First DataBank National Drug Data File.
[b]Not literally a DKB but included for comparison purposes.

pared with 20.7% of the unique codes in this data set (Table 5). In the case of the RxHub prescription messages, every DKB covered at least 99% of the total prescription volume.

In the inpatient setting, the coverage of the codes from any one source by any one DKB was also better when measured in terms of total prescription volume, from 77.4% to 97.0% (median, 93.7%) for the 20 inpatient source cells (Table 4). But the coverage did not reach the heights of the outpatient sources. The lower rate of coverage in the inpatient setting was partly due to the higher use of invented codes (1.3–7.4% of the total prescription volume) compared with the outpatient setting. For each data cell in Table 4, the number of invented codes was divided by the total number of noncovered codes. For three of the four inpatient sources, the invented codes explained

Table 5.
Results of Evaluation Based on Total Prescription Messages or Records for Each Data Source[a]

Data Source	Total No. Records	% Invented Codes	% Noninvented NDCs Not Found in Any DKB	% Noninvented NDCs Found in ≥1 DKB
Outpatient Sources				
Hospital E pharmacy messages	464,404	0.21	0.06	99.73
Medicaid archival records	66,291,780	0.11	0.06	99.83
RxHub messages	304,855	0.17	0.001	99.83
Inpatient Sources				
Hospital A messages	40,855	7.4	0.35	92.2
Hospital B messages	149,675	1.3	2.7	96.0
Hospital C messages	40,622	1.5	1.1	97.4
Hospital D messages	396,689	3.2	0.30	96.5

[a]The first data column gives the total number of prescription messages or records in each source. The remaining three data columns give the percentage breakdown of the message volume.

36–94% of the DKBs noncoverage. For the fourth inpatient source, the invented codes accounted for a smaller, but still important, percentage of the noncoverage: 6–32%. The remaining percentage of noncoverage in a given comparison was due to gaps in "NDC" coverage within the DKB.

Interestingly, RxNorm provided the highest or second-highest coverage of most sources. FDA's published "official" NDC directory usually had the lowest coverage rate—a median of 17.4% below that with the highest coverage rate. The DKB with the highest coverage rate for a given source organization was sometimes the DKB used by that source organization. The DKB vendors get feedback from their customers about missing "NDCs" and add them to their database product. Over time, we would anticipate most vendors to provide good coverage of the "NDCs" that their customers encounter.

Why are codes invented? To determine why organizations invent codes, we hand reviewed records containing invented codes. We found that pharmacies invent codes to accommodate (1) locally compounded medications, (2) drugs used in randomized controlled trials, (3) special items (e.g., pig skin), and (4) pharmacy actions not associated with a dispensing event. Of course, official NDCs do not cover any of these uses. Pharmacies also invent codes to dispense a product that has not yet been registered in their master formulary file, even if that product already carries a valid official NDC assigned by the manufacturer.

At one institution, compounded dermatological products accounted for over 315 unique invented codes. At another, compounded i.v. admixtures and medications accounted for 113 invented codes. Total parenteral nutrition accounted for 4–8% of the invented codes at hospitals that used this code. One hospital invented 47 codes for as many clinical trials. Most invented a few codes for identifying pharmacy actions not associated with the dispensing of any product (e.g., remove fentanyl patch, compounding fee, read tuberculin skin test). Although the number of distinct codes in this category was small, the codes tended to be frequently used. None of the product codes categorized as "invented" by our operational definition appeared in any of the DKB mapping tables, giving credence to our definition.

Coverage of archival sources of prescription information. To assess the DKB coverage of archival prescription records, we examined one source of "old" prescription records: the Indiana Medicaid database, which contains records of 66 million dispensed prescriptions dated from 1994 to 2006 and includes 41,727 unique "NDCs." The standard releases of the DKBs covered 74.5–80.9% of these unique "NDCs" and 97.3–98.7% of the total volume of the product codes in the Medicaid database.

To serve the requirements of their pharmacy customers, commercial vendors include only active and recently inactivated "NDCs." However, most of them do keep the inactive "NDCs" in their internal databases. We obtained a custom database from one of the DKBs that included all the active and inactive "NDCs" from its internal database. This more complete set of "NDCs" covered 95.0% of unique Medicaid "NDCs" and 99.8% of the Medicaid prescription volume.

Reuse of "NDCs." Official NDCs can be reused (i.e., reassigned to a completely different product) five years after the supplier has reported their inactivation to FDA.[18] Some DKBs keep track of the discontinuation, reactivation, and reuse of these codes. Based on the content of one large DKB, labelers reuse codes infrequently in practice, with 0.4% of "NDCs" flagged as reused.

Coverage of supplies. All of the commercial DKBs contain commonly prescribed supply items (e.g., cotton balls, alcohol swabs, insulin syringes, home glucose test strips). A system using one of these commercial databases can convert product codes for supplies to a more general DKB code, analogous to that DKB's clinical drug code, and convert all prescribed items into a more clinically useful form. However, using the definition of supplies provided by one DKB, we found that only 2.1% of the unique codes and 1.8% of the prescription volume (from a combined data set consisting of all our sample prescriptions) represented supplies. At this time, RxNorm does not include general supplies in its database. This did not have much influence on RxNorm's coverage rate due to the relatively low prevalence of general supplies in prescriptions.

Summary and recommendations

The goal of this study was to assess the problems of converting prescribed entity codes, as delivered in HL7 and NCPDP pharmacy messages, into clinical drug codes as defined by DKB suppliers. This conversion is just one step in the process of aggregating prescription records from many different sources. But if the DKBs do not include the product codes delivered in pharmacy messages, this step will be the rate-limiting one.

In the case of outpatient pharmacy records, DKBs did cover the codes for prescribed entities well. The success with RxHub and Medicaid prescription coverage is notable because both sources comprised data from multiple pharmacy benefit managers and hundreds of pharmacies. The high outpatient coverage rate was because outpatient pharmacies rarely invent product codes (<0.21% of the message volume) and because DKBs tend to cover the commonly prescribed outpatient "NDCs" well. DKB coverage of inpatient medication messages was not as good as that for comparable outpatient messages, due in large part to the higher usage rates of invented codes, up to 7.4% of the total prescription volume at some hospitals.

Based on our calculations, a clinical system that received prescription messages from outpatient settings could automatically map more than 99% of them to a clinical drug code that would enable useful clinical displays, decision support, performance measurement, and research. For the few outpatient messages that could not be mapped automatically (<1%), the receiving system could store them under one "miscellaneous drug" code. The drug names of these few could still be displayed under that miscellaneous code for human viewing, but these nonmapped prescriptions could not be used for the analytic purposes described above. In the matter of a few minutes per week, receivers could choose to review the unmapped codes as they accumulate and manually map those that occur frequently.

This same set of strategies could also be applied to the inpatient setting. However, a greater proportion of the received messages would not be covered, so more orders (5–10%) would be coded as "miscellaneous drug" and proportionately more codes would have to be reviewed to reach the level of coverage available in the outpatient setting. However, nonmapped medications could still be displayed to care providers by name. It may not be as important to have inpatient medications fully coded for long-term care, because most of the drugs would be retrieved from outpatient sources.

Noncoverage can be attributable to gaps in the chosen DKB's coverage or to the delivery of invented codes by the source pharmacy. Receivers can mitigate noncoverage by selecting a DKB that covers the sources of most interest to them. We understand that DKBs tend to fill in "NDC" gaps that their customers experience, so the proportion of noncoverage due to DKB gaps is likely to be self-correcting. However, much of the noncoverage, especially in the inpatient setting, is due to invented codes. Changes in pharmacy processes and pharmacy message standards could reduce or eliminate most invented codes. Inpatient pharmacies should send the "NDC" for the main active ingredient in the HL7 message—as community pharmacies now do in their NCPDP message—or use the available HL7 mechanism consistently for delivering "NDCs" for all ingredients. HL7 and NCPDP could mitigate the problem with invented codes for randomized trials, rare supplies, and nondispensing actions by defining a field that identified these special purpose codes as such.

The practice of dispensing medications without registering them in the pharmacy's master formulary accounts for a large share of the invented codes. This practice should be discouraged, because pharmacy safety checks cannot be applied to drugs that have not been registered in the pharmacy's master formulary. Also, hospital pharmacy systems tend to load the "NDC" for a given clinical drug when they initially create a formulary entry. They rarely change this code, even when they get a new supply of the same clinical drug with a different "NDC." However, this practice will only have consequences when the initially stored "NDC" is retired from the DKB the systems use and could be corrected when that happens.

A better approach than mapping at the receiver site would be mapping at a central facility, such as an RHIO. It would be even better for prescription sources to identify the medications within their prescription message via a universal clinical drug code in addition to the "NDC." Such a solution would eliminate any need for mapping and speak to the 2008 resolution from the American

Society of Health-System Pharmacists to explore "the potential benefits of supplementing or modifying the National Drug Code with a coding system that can be effectively used across the medication-use continuum."[19]

In this context, the coverage of the "NDCs" by RxNorm, the public-use knowledge base, deserves comment. Considering all of the DKBs, the RxNorm mapping table provided the best or second-best coverage of "NDCs" from all but one of our prescription sources. RxNorm is nonproprietary and has been designated by the federal Consolidated Health Information committee as the standard for the clinical drug code in prescription messages.[20] It could supply the universal clinical drug codes that we need.

FDA's table of official NDCs covered a smaller proportion of the "NDCs" in our sample (range, 63.3–92.2%) compared with the DKBs. This relatively low coverage is because labelers generate new "NDCs" as they need them and put them into use, and the rest of the industry (including DKB vendors) adopt them before they get into the FDA database. Delays in labelers' submissions to FDA[21] and some delay in the entry of submissions into FDA's database account for most of the discrepancy. In August 2006, FDA proposed changes to its regulations to eliminate these discrepancies; provide a unique, chemical structure-based identifier for drug ingredients; and provide a rich machine-readable source of information about all drug products.[22] Some of these changes have already been adopted.[23] The comparisons in this report predate those changes, so it is likely that the coverage of FDA's data has improved since the time of this study.

Through collaboration with FDA, NLM is deploying this drug information and cross-linking FDA's identifiers with RxNorm.[24] The full set of FDA proposals, when fully implemented, will ensure that all drug products have an official NDC and will ease the process of mapping official NDCs to clinical drug codes.

Since much of our data came from Indianapolis hospitals, we cannot be certain that our results for DKB coverage and invented codes are generalizable to other parts of the country. But the experience does reach beyond one major city. The Medicaid database included records from the entire state of Indiana and many hundreds of pharmacies, as did the RxHub prescription sample. Nearly 1.4% of the RxHub prescriptions came from outside of Indiana (Majkowski K, RxHub, personal communication, 2006 Mar 27). Further, Indiana is in the middle range of drug use for all but one class of medications (macrolides),[25] so it may be representative of other states. Though the hospital pharmacy data came from hospitals in Indianapolis or collar counties, they represent a broad spectrum of information from five different hospitals and pharmacy systems. Nonetheless, this is the first report of its kind. More data are needed that represent other parts of the country to fully understand any possible variation.

The mapping of incoming codes to the more general clinical drug code is one step in the process of integrating prescription information from many outside sources in a medical record system. Prescriptions for the same patient from different sources may have different patient identifiers. The linking of these disparate identifiers is another required step but is beyond the scope of this article. However, this problem can be solved through standard linkage techniques if the prescription records come with enough patient registration information.[26] It has been solved by RxHub and by RHIOs, such as INPC, for their targeted scope.

Receiving systems also need machinery to aggregate the clinical drug to a higher level of categorization and to enable decision support and statistical analysis. For example, amoxicillin 250-mg capsules can be generalized to amoxicillin oral preparations or further to penicillins or even further to antibiotics. All of the commercial DKBs have built data structures and hierarchies that permit such aggregation. Indeed, they carry rich content, including the ingredients for each drug (important for allergy checking) and definitions of drug–drug, drug–test, drug–diagnosis, and drug–food interactions. Most also include human-readable information about the drug, designed for physicians, pharmacists, patients, or all three groups, depending on the source DKB. They also feature attributes that facilitate charging (e.g., average wholesale prices). We did not obtain or examine this information, so we cannot comment on these capabilities, except to say that they are invaluable for many drug-prescribing and clinical care purposes and that the DKB vendors differ in the specific tools and services they provide.

For research and historical purposes, many groups will want to incorporate archival prescription records into their medical record database, along with current medications. To include such older prescriptions, the DKB vendor must be asked to supply all of their old "NDCs" that are not part of their standard database release.

The results of this study suggest that a receiving system can automatically aggregate outpatient prescriptions from many sources and store them under the clinical drug code of an appropriate DKB. Because of the maturity of standards for codes and messages, medication records are the most ripe for electronic sharing. RHIOs, office practices, hospitals, nursing homes, payers, and researchers could all benefit from shared access to all of the medication records of their patients.

Conclusion

DKBs cover the product codes used by outpatient sources sufficiently well to permit automatic mapping. Changes in policies and standards could increase coverage of product codes used by inpatient sources.

References

1. Aspden P, Wolcott JA, Bootman JL et al., eds. Preventing medication errors. Washington, DC: National Academies Press; 2006.
2. Joint Commission. 2007 Hospital and Critical Access Hospital National Patient Safety Goals webpage. www.jointcommission.org/PatientSafety/NationalPatientSafetyGoals/07_hap_cah_npsgs.htm (accessed 2008 May 1).
3. Poon EG, Blumenfeld B, Hamann C et al. Design and implementation of an application and associated services to support interdisciplinary medication reconciliation efforts at an integrated healthcare delivery network. *J Am Med Inform Assoc.* 2006; 13:581-92.
4. Surescripts. www.rxhub.com/thecompany (accessed 2009 Jul 17).
5. Halamka J, Aranow M, Ascenzo C et al. Health care IT collaboration in Massachusetts: the experience of creating regional connectivity. *J Am Med Inform Assoc.* 2005; 12:596-601.
6. National Council for Prescription Drug Programs. Telecommunication standard implementation guide, version 5.1. Scottsdale, AZ: 1999 Sep.
7. HL7 messaging standard, version 2.5. An application protocol for electronic data exchange in healthcare environments [standard]. Ann Arbor, MI: Health Level Seven; 2003.
8. Food and Drug Administration. National Drug Code directory webpage. www.fda.gov/cder/ndc (accessed 2008 May 1).
9. Cimino JJ, McNamara TJ, Meredith T et al. Evaluation of a proposed method for representing drug terminology. *Proc AMIA Symp.* 1999:47-51.
10. Nelson SJ, Brown SH, Erlbaum MS et al. A semantic normal form for clinical drugs in the UMLS: early experiences with the VANDF. *Proc AMIA Symp.* 2002:557-61.
11. Sperzel WD, Broverman CA, Kapusnik-Uner JE et al. The need for a concept-based medication vocabulary as an enabling infrastructure in health informatics. *Proc AMIA Symp.* 1998:865-9.
12. National Library of Medicine. RxNorm homepage. www.nlm.nih.gov/research/umls/rxnorm/index.html (accessed 2008 May 1).
13. Health insurance reform: standards for electronic transactions. Final rule. *Fed Reg.* 2000; 65:50312-72.
14. Food and Drug Administration. National Health Related Items Code webpage. www.fda.gov/cdrh/nhric/nhric.html (accessed 2008 May 1).
15. GS1 US. Bar codes and ecom webpage. http://barcodes.gs1us.org/dnn_bcec/About/WhatsinaBarcode/tabid/446/Default.aspx (accessed 2009 Jul 17).
16. McDonald CJ, Overhage JM, Barnes M et al. The Indiana Network for Patient Care: a working local health information infrastructure. *Health Aff.* 2005; 24:1214-20.
17. External code list. Scottsdale, AZ: National Council for Prescription Drug Programs; 2007 Jan.
18. Notification of registrant; drug establishment registration number and drug listing number (codified at 21 C.F.R. §207.35).
19. ASHP House of Delegates takes on pressing professional issues in Seattle. American Society of Health-System Pharmacists press release. Bethesda, MD; 2008 Jun 20.
20. National Committee on Vital and Health Statistics. Letter to secretary of the Department of Health and Human Services Tommy Thompson from chair of the National Committee on Vital and Health Statistics John Lumpkin. www.ncvhs.hhs.gov/031105lt3.pdf (accessed 2008 May 1).
21. Food and Drug Administration. National Drug Code directory. www.accessdata.fda.gov/scripts/cder/ndc/default.cfm (accessed 2009 Jul 17).
22. Requirements for Foreign and Domestic Establishment Registration and Listing for Human Drugs, Including Drugs that Are Regulated Under a Biologics License Application, and Animal Drugs [proposed rule]. Rockville, MD: Food and Drug Administration; 2006 Aug 23.
23. Schadow G. Assessing the impact of HL7/FDA Structured Product Label (SPL) content for medication knowledge management. *Proc AMIA Symp.* 2007:646-50.
24. National Library of Medicine website. DailyMed webpage. http://dailymed.nlm.nih.gov (accessed 2008 May 1).
25. Motheral BR, Cox ER, Mager D et al. Express Scripts prescription drug atlas, 2001: a study of geographic variation in the use of prescription drugs. St. Louis: Express Scripts, Inc.; 2002 Jan.
26. Grannis SJ, Overhage JM, McDonald CJ. Analysis of identifier performance using a deterministic linkage algorithm. *Proc AMIA Symp.* 2002:305-9.

Appendix—Methods for counting the unique clinical drugs in each drug knowledge base (DKB)

RxNorm

This DKB includes 29,734 unique identifiers of type "Semantic Clinical Drug." We removed 11,295 deprecated codes that were marked as "obsolete" in the "Suppress" field of the Concept (RXNCONSO.RRF) file and another 10,357 codes that were not linked to any "NDC" in the Attribute (RXNSAT.RRF) file. We obtained a count of 8,082 unique RxNorm "Semantic Clinical Drugs."

Medi-Span MDDB

This DKB includes 11,894 unique generic product identifiers (GPIs). We removed the 669 GPIs in the "Medical Devices" category and the 1,389 GPIs in the "Chemicals" category to obtain a total of the medications.

Thomson Micromedex Red Book

We began with all the 25-digit codes in the Ulti-Medex hierarchy. These codes define a polyhierarchy, so we focused on the last 10 digits of each code, which uniquely identifies the drug name, route, dose form, and strength, and squeezed out duplicates to obtain the count of unique clinical drug codes.

Multum Lexicon

We used the count of "main_multum_drug_codes"—a unique identifier for each combination of drug name, route, dose form, and strength in this DKB.

First DataBank NDDF Plus

We started with the 20,301 clinical formulation identifiers ("GCN_SEQNOs"). Then, we subtracted 2,165 identifiers with therapeutic class "Supply," 1,060 identifiers with therapeutic class "Bulk Chemicals," and 2,671 identifiers with a null therapeutic class.