Journal of Biomedical Informatics 45 (2012) 613–625

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Cross-terminology mapping challenges: A demonstration using medication terminological systems

Himali Saitwal [a,1], David Qing [a,2], Stephen Jones [a,c], Elmer V. Bernstam [a,b], Christopher G. Chute [d], Todd R. Johnson [a,e,*]

[a] The University of Texas School of Biomedical Informatics at Houston, 7000 Fannin Suite 600, Houston, TX 77030, USA
[b] Department of Internal Medicine, The University of Texas, Health Science Center at Houston, 6431 Fannin, MSB 1.150, Houston, TX 77030, USA
[c] Department of Surgery, The Methodist Hospital Research Institute, 6550 Fannin, Smith Tower 1661A, Houston, TX 77030, USA
[d] Division of Biomedical Statistics and Informatics, Mayo Clinic, 200 First Street SW, Rochester, MN 55905, USA
[e] Division of Biomedical Informatics, Department of Biostatistics, College of Public Health, University of Kentucky, 725 Rose Street 230H, Lexington, KY 40536, USA

ARTICLE INFO

Article history:
Received 11 July 2011
Accepted 14 June 2012
Available online 28 June 2012

Keywords:
Medication terminological systems
Standards
Terminology mapping
Review of medication terminological systems

ABSTRACT

Standardized terminological systems for biomedical information have provided considerable benefits to biomedical applications and research. However, practical use of this information often requires mapping across terminological systems—a complex and time-consuming process. This paper demonstrates the complexity and challenges of mapping across terminological systems in the context of medication information. It provides a review of medication terminological systems and their linkages, then describes a case study in which we mapped proprietary medication codes from an electronic health record to SNOMED CT and the UMLS Metathesaurus. The goal was to create a polyhierarchical classification system for querying an i2b2 clinical data warehouse. We found that three methods were required to accurately map the majority of actively prescribed medications. Only 62.5% of source medication codes could be mapped automatically. The remaining codes were mapped using a combination of semi-automated string comparison with expert selection, and a completely manual approach. Compound drugs were especially difficult to map: only 7.5% could be mapped using the automatic method. General challenges to mapping across terminological systems include (1) the availability of up-to-date information to assess the suitability of a given terminological system for a particular use case, and to assess the quality and completeness of cross-terminology links; (2) the difficulty of correctly using complex, rapidly evolving, modern terminologies; (3) the time and effort required to complete and evaluate the mapping; (4) the need to address differences in granularity between the source and target terminologies; and (5) the need to continuously update the mapping as terminological systems evolve.

1. Introduction

Much progress in biomedical information systems and informatics research would not be possible without the wide availability and content coverage of standardized terminological and coding systems. However, because these systems were created at different times and for different purposes, their content coverage (and overlap in coverage) varies, as does their use in health information technology and research. As a result, practical use cases often require mapping or integrating information across several terminological systems. Despite ongoing efforts to link systems through the Unified Medical Language System (UMLS) Metathesaurus and in individual terminological systems, making practical use of these systems presents a number of challenges. Given the size and complexity of each terminological system, it is often difficult to know the quality and completeness of the content of each system, the quality of the mappings across systems, and how these mappings may alter semantics. Because these systems continue to evolve, ongoing use of any cross-terminology mappings must also include a plan to accommodate these changes, such as occurs when codes are given different meanings, removed from a system, or new codes are added.

In this paper we demonstrate these challenges in the context of a project to map medications from a commercial electronic health

* Corresponding author at: Division of Biomedical Informatics, Department of Biostatistics, College of Public Health, University of Kentucky, 725 Rose Street 230H, Lexington, KY 40536, USA. Fax: +1 859 257 6430.
E-mail addresses: hsaitwal@Apelon.com (H. Saitwal), pqing@mdanderson.org (D. Qing), sljones2@tmhs.org (S. Jones), Elmer.V.Bernstam@uth.tmc.edu (E.V. Bernstam), chute@mayo.edu (C.G. Chute), Todd.R.Johnson@uky.edu (T.R. Johnson).
[1] Present address: Apelon Inc., 100 Danbury Rd., Suite 202, Ridgefield, CT 06877, USA.
[2] Present address: Department of Pharmacy Informatics, The University of Texas MD Anderson Cancer Center, 1515 Holcombe Blvd., Houston, TX 77030, USA.

1532-0464/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2012.06.005

Table 1
Acronyms used throughout this paper.

Acronym	Explanation	Acronym	Explanation
AHRQ	Agency for Healthcare Research and Quality	NCI	National Cancer Institute
AMP	Actual Medicinal Product	NCIt	National Cancer Institute thesaurus
AUI	Atom Unique Identifier	NDC	National Drug Code
CMS	Centers for Medicare and Medicaid Services	NDDF plus	National Drug Data File Plus
CUI	Concept Unique Identifier	NDF – RT	National Drug File – Reference Terminology
DDI	Drug Description Identifier	NLM	National Library of Medicine
DoD	Department of Defense	PharmGKB	Pharmacogenomics Knowledge Base
EPA	Environmental Protection Agency	PTR	Path to Root
FDA	Food and Drug Administration	RXCUI	RXNorm Concept Unique Identifier
GO	Gene Ontology	SAB	Source ABbreviation
GPI	Generic Product Identifier	SNOMED CT	Systematized Nomenclature of Medicine – Clinical Terms
GPPC	Generic Product Packaging Code	STR	String or Text Label
HIPAA	Health Insurance Portability and Accountability Act	UMLS	Unified Medical Language System
LOINC	Logical Observation Identifiers Names and Codes	UNII	Unique Ingredient Identifier
MDDB	Master Drug Data Base	UPC	Universal Product Code
MedDRA	Medical Dictionary for Regulatory Activities	USP	United States Pharmacopeia
MeSH	Medical Subject Headings	VHA	Veterans Health Administration
MRCONSO	UMLS Metathesaurus Concept Names and Sources		

record to a drug classification system. The goal of the mapping was to allow researchers to retrieve patient records from an i2b2 clinical data warehouse (CDW) based on indications and/or classes of prescribed medications. For example, to answer a question such as "Which patients have been pharmacologically treated for depression?" Thus, our main goal was to map all drugs into a polyhierarchical classification system that a clinical researcher could use to form queries in i2b2. A secondary goal was to provide interoperability with information in our other informatics research projects, which were primarily based on UMLS Metathesaurus codes.

To provide context, we first give a brief overview of medication terminology systems, emphasizing the distribution of information across them and the linkages among them. Although we did our initial review of systems at the start of the study, the review we present here reflects the present state of these systems., We then present the case study of mapping from medications in a commercial electronic health record to SNOMED CT and the UMLS Metathesaurus (see Table 1 for a list and definitions of the acronyms used throughout this paper). Since this study was completed there have been many changes to medication terminological systems, such that if we were to repeat the mapping today, even using the same terminologies, we would obtain different results with respect to the completeness and quality of the mapping. However, the challenges that are demonstrated in this paper transcend the particular details of terminological systems and are therefore relevant to ongoing mapping efforts.

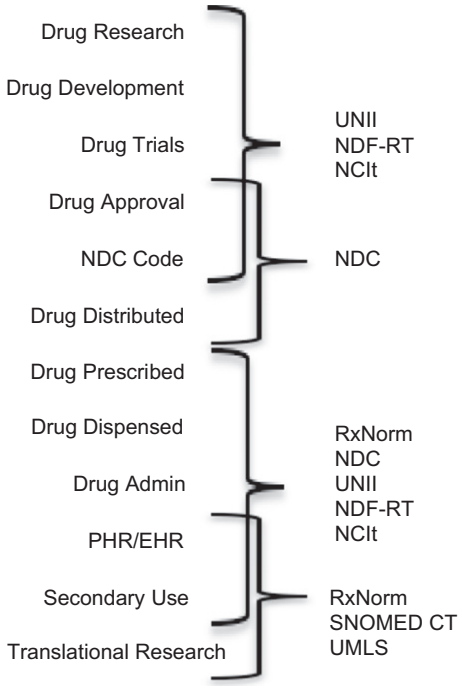

Fig. 1. Medication terminology standards and their uses. Modified from presentations of the Federal Medication Terminology standards [3].

2. Review of controlled terminologies for medication information

Useful medication information includes drug components such as trade name (if any), generic name, active ingredient(s), drug strength and unit of measure, dosage form, route of administration, chemical substance, drug class, mechanism of action, physiologic effect, manufacturer details, and package type and size. Fig. 1 gives an overview of common medication terminological systems and how they are often used. Fig. 2 is a Circos diagram [1] showing each of these systems and how they are linked through common codes. In the diagram, ribbons show the connections between each system. Ribbon color corresponds to the source system—the system that contains the code to the target system. For example, the ribbon linking RxNorm to MDDB has the same color as the RxNorm segment, meaning that RxNorm contains MDDB codes. In the remainder of this section, we describe these terminological systems and the linkages between them.

Tables 2a and 2b shows how medication information is distributed across various standard terminological systems at the time of this writing. These tables show the kinds of medication information that is included in each system, but not whether the coverage is complete. Note that because the UMLS Metathesaurus contains and links concepts from many source terminologies, its content coverage is the set of all concepts that are included from those terminologies.

The terms used to describe different types of terminological systems are rarely used consistently. In this paper, we use the terms and definitions proposed by de Keizer et al. [2], as shown in Table 3. Table 4 shows how these types apply to the openly available terminological systems described in this paper. We were not able to classify the proprietary systems described below, because the manufacturers provide insufficient information to adequately assess each type. In addition, some proprietary systems consist of a

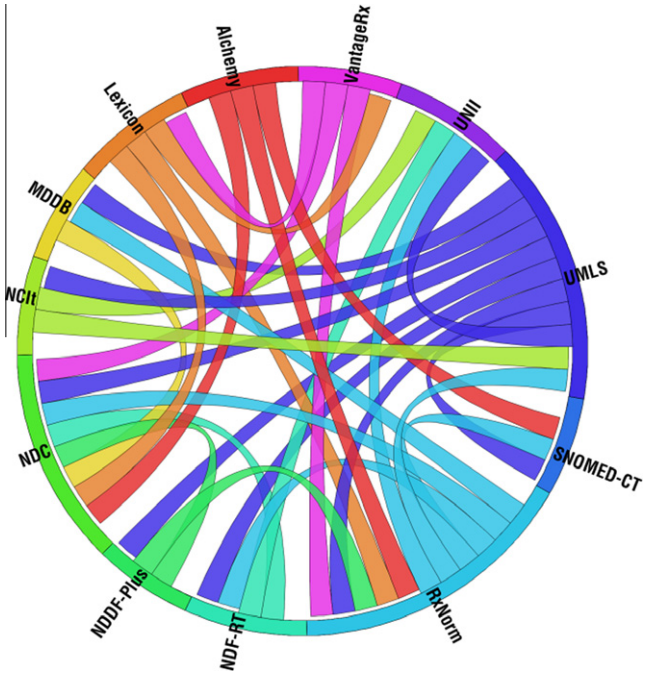

Fig. 2. A Circos diagram showing terminological systems containing medication information and connections between systems. Ribbon colors indicate the terminology containing the code to the target system. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

core package, along with a number of add-on packages and applications that extend their coverage. As a result, the type of terminological system may vary depending upon the particular set of purchased packages.

2.1. UNII (Unique Ingredient Identifier)

The UNII is a thesaurus that provides unique identifiers for substances in drugs, biologics, foods, and devices. Each concept has a preferred substance name and synonyms. There are 16,655 unique concepts and 67,715 synonyms, including preferred names [4]. UNII is maintained by the FDA and USP (United States Pharmacopeia), and is assigned by the FDA/USP Substance Registration System (SRS). Although the UNII does not contain mappings to other coding systems, it is used to identify ingredients in the NCIt [5], NDF-RT [6], and the UMLS Metathesaurus [7].

2.2. NDC (National Drug Code)

The NDC is a coding system that provides a universal, but not necessarily unique, product identifier for human drugs that is maintained and distributed by the FDA [8]. An NDC is a 10-digit, three-segment numeric string that identifies the labeler, product, and package size. The first segment, the labeler code, is assigned by the FDA (a labeler is any firm that manufactures or distributes the drug). The second segment, the product code, identifies specific drug strength, dosage form and formulation. The third segment, the package code, identifies package sizes and types. Both the product and package codes are assigned by the firm. Because an NDC is composed of three codes, it may be considered a simple nomenclature.

Because some firms have four digit labeler codes and some have five, the NDC can be in one of three configurations: 4–4–2, 5–3–2, or 5–4–1. A firm with a five digit code can choose between the 5–3–2 and 5–4–1 configurations.

Table 2a
Overview of medication information used for clinical and research purposes and the terminological systems that include it.

Information about Drug	Representative example	UNII	NDC	NCIt	RxNorm	NDF-RT	SNOMED CT	UMLS
Brand Name	Inderal capsules		✓	✓	✓	✓		✓
Generic Name	Propranolol Hydrochloride 60 MG extended release capsule		✓	✓	✓	✓	✓	✓
Active Ingredient	Propranolol Hydrochloride	✓	✓	✓	✓	✓	✓	✓
Drug Strength and Unit of Measure	60 mg		✓	✓	✓	✓	✓	✓
Dosage Form	Capsule (Extended Release)		✓	✓	✓	✓	✓	✓
Route of Administration	Oral		✓	✓	✓	✓	✓	✓
Chemical Substance	Propranolol Hcl	✓		✓	✓	✓	✓	✓
Drug Class	Antiarrhythmic drug			✓		✓	✓	✓
Mechanism of Action	Beta adrenergic antagonist			✓		✓	✓	✓
Physiologic Effect	Depression of sinus node function and AV node conduction			✓		✓	✓	✓
Manufacturer Details	Pharmacy Service Center		✓			✓		✓
Package Type and Size	PKG and 50		✓	✓	✓	✓		✓
Side effects	Dizziness, lightheadedness, drowsiness, tiredness, diarrhea, etc.			✓			✓	✓
Synonyms	2-Propanol-1-[(1-methylethyl)amino]-3-(1-naphthalenyloxy) Hydrochloride	✓			✓	✓		✓
Drug-Drug Interaction	For example – Quinidine increases the concentration of Propranolol and produces greater degrees of clinical beta-blockade and may cause postural hypotension					✓		✓
Drug-Allergy Interaction	Hypersensitivity reactions, pharyngitis and agranulocytosis; erythematous rash laryngospasm; respiratory distress, etc.						✓	✓
Prescription indication	Antihypertensive, antiarrhythmic, migraine			✓		✓		✓
Contraindications	cardiogenic shock, sinus bradycardia and greater than first-degree block, etc.					✓		✓

✓ – Terminology set contains the represented drug information, regardless of completeness.

Table 2b
Overview of medication information used for clinical and research purposes and the proprietary terminological systems that include it.

Information about drug	Representative example	MDDB (Wolters Kluwer Health)	NDDF Plus (First DataBank)	Lexicon (Cerner Multum)	Vantage Rx Database (Cerner Multum	Alchemy (Elsevier/Gold Standard)
Brand name	Inderal capsules	✓	✓	✓	✓	✓
Generic name	Propranolol Hydrochloride 60 MG extended release capsule	✓	✓	✓	✓	✓
Active ingredient	Propranolol Hydrochloride	✓	✓	✓	✓	✓
Drug strength and unit of measure	60 mg	✓	✓	✓	✓	✓
Dosage form	Capsule (Extended Release)	✓	✓	✓	✓	✓
Route of administration	Oral	✓	✓	✓	✓	✓
Chemical substance	Propranolol HCl	✓	✓	✓	✓	✓
Drug class	Antiarrhythmic drug		✓		✓	✓
Mechanism of action	Beta adrenergic antagonist		✓		✓	✓
Physiologic effect	Depression of sinus node function and AV node conduction		✓		✓	✓
Manufacturer details	Pharmacy Service Center	✓	✓		✓	✓
Package type and size	PKGand50	✓	✓		✓	✓
Side effects	Dizziness, lightheadedness, drowsiness, tiredness, diarrhea, etc.		✓		✓	
Synonyms	2-Propanol-1-[(1-methylethyl)amino]-3-(1-naphthalenyloxy) hydrochloride	✓	✓		✓	✓

Although the NDC is limited to 10 digits, the HIPAA standard requires an 11 digit NDC. To accommodate this requirement, the 10 digit NDC is padded with an extra character. To ensure that the original 10 digit code can be unambiguously recovered from the 11 digit code, the FDA recommends using an asterisk (*); however, some programs use a zero (0). The resulting HIPAA NDC is an 11 character, three-segment code with the configuration of 5–4–2. The NDC for two different concentrations of Propranolol oral solution are shown in Table 5.

A subset of registered drugs is also listed in the National Drug Code Directory, which connects an NDC with other related information, including trade name, active ingredient(s), dosage form, drug strength, unit of measure, route of administration, package size and type. Trade names are names assigned to drugs by manufacturers. Codes for this related information are unique to the NDC directory instead of being based on an existing standard, such as UNII.

One problem with the NDC, is that manufacturers may reuse a code 5 years after notifying the FDA that a code is inactive. Thus, the same NDC can represent more than one drug; a problem that is referred to as "semantic drift" [9]. A recent study, conducted by Simonaitis and McDonald, of a proprietary database that tracks changes to NDCs found that only 0.4% of NDCs were flagged as changed [10].

Simonaitis and McDonald also found several hundred codes in which manufacturers sometimes used the last two digits (the package code) to distinguish ingredients rather than package size, such as Ipecac Syrup (00686-0360-10) and Digoxin Elixir (00686-0360-67). In the same paper, they used prescription messages and formularies from five hospitals, one outpatient pharmacy, RxHub, and Medicaid archival records, to assess the percentage of NDCs in those datasets that are included in NDC, RxNorm, MDDB, Multum Lexicon, MMX (Thompson Micromedex Red Book), and the NDDF (First DataBank National Drug Data File). They found that NDC usually had the lowest coverage among the systems studied, whether measured as a percentage of unique codes or as a percentage of the total volume of codes in the prescriptions. In the latter case, NDC was a median of 17.4% points below the system with the best coverage. This low performance is due to the fact that medications with NDC codes are available on the market prior to being entered into the NDC database. Proprietary systems are often updated in a more timely manner due to customer feedback about new NDCs.

2.3. NCIt (National Cancer Institute thesaurus)

The NCIt is a cancer-centric thesaurus, vocabulary, taxonomy, ontology, and coding system that includes several concept hierarchies pertaining to medications [5,11,12]. It is maintained and distributed by the NCI Center for Bioinformatics caCORE [13,14]. The NCIt contains semantic relationships between genes, diseases, drugs and chemicals, anatomy, organisms, and proteins. In addition, the NCI metathesaurus connects terms from the NCIt to 76 different biomedical vocabularies such as MedDRA, NDF-RT, LOINC and GO. The NCIt also includes a substances concept hierarchy that maps to the corresponding UNII and UMLS Metathesaurus codes.

2.4. RxNorm

RxNorm is a thesaurus, taxonomy, ontology, nomenclature, and coding system that provides normalized names and concept codes (called the RxCUI) for clinical drugs and drug delivery devices [15]. Each drug is represented by its active ingredients (including trade names), drug strength and unit of measure, dosage form and route of administration. Each normalized name is one of several term types, such as Ingredient, Brand Name, Semantic Clinical Drug,

Table 3
Terminology types and their definitions.

Term	Definition
Terminology	List of terms referring to concepts
Thesaurus	Ordered terminology that includes synonyms
Vocabulary	Terminology or Thesaurus that includes concept definitions (formal or informal)
Nomenclature	System of terms with rules for combing the terms to define complex concepts
Taxonomy/classification	Organizes concepts into a hierarchy using "is-a" relationships. A classification uses the more general "is-member-of" relationship
Ontology	A specification of concepts, relations, and functions for a domain
Coding system	Any terminological system that uses codes for designating concepts

Table 4
Types of medication terminological systems.

Type	UNII	NDC	NCIt	RxNorm	NDF-RT	SNOMED CT	UMLS
Terminology	X	X	X	X	X	X	X
Thesaurus	X		X	X	X	X	X
Vocabulary			X		X	X	X
Nomenclature		X			X	X	
Taxonomy/classification			X	X	X	X	X
Ontology			X	X	X	X	X
Coding system	X	X	X	X	X	X	X

Table 5
An overview of assigning the eleven-digit NDC to drugs – labeler codes are assigned by FDA and Product and Package codes are assigned by manufacturing firms.

Digits 1–5	Labeler code (manufacturer)	Digit 6–9	Product code (drug name)	Digit 10–11	Package code (packet size)
00054	ROXANE LABORATORIES INC.	3727	Propranolol oral solution 20 mg/5 ml	63	Bottle size of 5 ml
00054	ROXANE LABORATORIES INC.	3730	Propranolol oral solution 40 mg/5 ml	63	Bottle size of 5 ml

etc. A Semantic Clinical Drug is an ingredient plus strength and dose form, as in *Fluoxetine 4 MG/ML Oral Solution* [16]. Normalized names are themselves composed of a number of elements, where each element is also a concept with its own term type. For instance "*Fluoxetine 4 MG/ML Oral Solution [Prozac]" with the RxCUI of 104850* is of term type Semantic Branded Drug, which is composed of terms for ingredient (Fluoxetine), strength and unit of measure (4 MG/ML), dose form (oral solution) and brand name (Prozac). RxNorm does not have separate terms for route of administration, but instead uses dose form, which combines route and drug form, as in "oral solution." Each of the component terms (other than strength and unit of measure) are examples of different term types with their own concept codes. Concepts and terms are connected within RxNorm by a number of relations, including constitutes, contains, dose_form_of, includes, ingredient_of, is_a, and tradename_of, and the corresponding inverse relations. These relationships mean that RxNorm also has the properties of a taxonomy and ontology.

Although RxNorm does not contain drug class information, a subset of drugs in RxNorm contain codes from the UMLS Metathesaurus, SNOMED CT and NDF-RT. All of which contain drug class information along with codes from other controlled vocabularies, including those in commercially available drug information sources. There is also a 1-to-many (1:M) mapping from RxNorm concepts to NDC's. A 1-to-1 (1:1) mapping is not possible because NDC's are specific to package size, whereas RxNorm codes are not. RxNorm also contains codes from several proprietary medication terminological systems such as NDDF Plus (First DataBank), Micromedex (Thomson Reuters), MDDB (Medi-Span), Alchemy (Gold Standard/Elsevier,), Lexicon (Cerner-Multum) and VantageRx Database (Cerner-Multum).

A recent evaluation [10] of RxNorm found that it could code all but one of 19,743 ambulatory e-prescriptions, resulting in a coverage rate of 99.995% coverage. The authors mapped from NDC codes in the e-prescriptions to RxNorm CUIs using three different methods applied sequentially: NDC to CUI mappings included in RxNorm (94.4% of the codes), a proprietary vendor supplied NDC to RxNorm mapping (an additional 4.4%), and manual search of RxNorm (for the remaining 1.2%). Similarly, in the study by Simonaitis and McDonald discussed above, RxNorm had the best or second best coverage of NDC codes. RxNorm's performance advantage over NDC derives from the fact that its NDCs are updated from multiple sources, including the Veterans Health Administration, the NDC Directory, the Centers for Medicare and Medicaid Services, Lexicon, and Gold Standard Alchemy [15]. Because of the importance of medication information and the maturity of RxNorm as a means of linking different terminological systems, the literature on the suitability of RxNorm for a variety of use cases is growing. However, a description of these studies is beyond the scope of this paper.

2.5. NDF-RT (National Drug File – Reference Terminology)

NDF-RT is produced by the VHA and distributed by NCI [6]. It includes information on drug characteristics, including drug ingredients, chemical substance, drug strength, unit of measure, dosage form, physiologic effect, mechanism of action, pharmacokinetics, and related diseases. NDF-RT contains UNII codes for generic ingredients, and codes for corresponding concepts in the NDC, UMLS Metathesaurus, MeSH, and RxNorm.

NDF-RT's drug classification assigns a single class to each drug. For instance, paroxetine HCl medications are classified under "antidepressants, other". However, NDF-RT provides a more comprehensive drug classification system through the relationships has_MoA (has mechanism of action) and has_PE (has pharmacologic effect). For instance, paroxetine HCl drug products include a has_MoA (has mechanism of action) relationship to the "serotonin reuptake inhibitors" concept. Both kinds of relations refer to

concepts that are organized in a classification hierarchy. For instance, ''serotonin reuptake inhibitors'' have an ''is-a'' relationship to ''serotonin transporter interactions'' which is in turn a ''neurotransmitter transporter interactions,'' and so on. Although MoA and PE provide technically more accurate drug classifications, the legacy drug classes tend to use more familiar clinical terms, which may make them better as an interface terminology.

In an analysis of the correspondence between NDF-RT (March 11, 2008 public inferred edition) and RxNorm (November 17, 2008 full release data), Pathak and Chute found that 54% of RxNorm drug products did not have a correspondence to NDF-RT, and that 45% of drug products in NDF-RT were missing from RxNorm [17]. They also found that drug products that have the same ingredient, but differ in other ways, such as dose or form, may be assigned to different drug classes [17]. At the time of the study, RxNorm and NDF-RT were linked only through the Veterans Health Administration unique identifiers (VUIDs). The RxNorm drug products that could not be linked to NDF-RT either were missing VUIDs or had VUIDs that were not in NDF-RT. However, beginning with the June 7, 2010 release of RxNorm, NDF-RT is now included as a source vocabulary and both RxNorm and NDF-RT are coordinated and released on the same date. However, the limitations of the legacy drug classification system remain, meaning that the mechanism of action and physiologic effect relations are the preferred methods for classifying drugs in NDF-RT. For a review of additional limitations of NDF-RT along with suggested remedies, see Pathak and Chute [18].

2.6. SNOMED CT (Systematized Nomenclature of Medicine – Clinical Terms)

SNOMED CT is a thesaurus, nomenclature, taxonomy, ontology, and coding system of clinical concepts maintained by the International Health Terminology Standards Development Organization (IHTSDO) [19]. It has been designated as the U.S. standard for electronic health information exchange in the interoperability specifications produced by the Healthcare Information Technology Standards Panel and has also been adopted for use by the U.S. Federal Government, through the Consolidated Health Informatics (CHI) Initiative, for several clinical domains [20]. SNOMED CT organizes content into several hierarchies. Three of these hierarchies—pharmaceutical/biological product, substance and physical object – contain medication-related concepts and relationships. The pharmaceutical/biological product hierarchy provides information on different drug components including drug ingredients, drug strength, unit of measure, dosage form, route of administration, physiologic effect, mechanism of action, drug allergen and synonyms. The substance hierarchy provides information about the chemical substance. The physical object branch contains detailed concepts for commonly used appliances in the hospital and outpatient settings that may be prescribed along with medications [21] and sometimes appear as medications in clinical data. Many RxNorm concepts contain the SNOMED CT Concept ID.

SNOMED CT consistently ranks highly for content coverage [22,23]; however, researchers have noted many classification errors [24] For example, researchers found that SNOMED CT classified acetaminophen correctly as an ''Analgesic AND antipyretic product'' in the ''Pharmaceutical/biologic product'' hierarchy, but incorrectly in the Substance hierarchy, where it was classified under multiple parents, including an incorrect classification as a ''Para-aminophenol derivative anti-inflammatory agent''. In a comparison of NDF-RT and SNOMED CT, Mortensen and Bodenreider found very little consistency in how drugs were classified, in part because the pharmacologic classes differed across the two systems [25].

2.7. UMLS (Unified Medical Language System) Metathesaurus

The UMLS Metathesaurus combines many different thesauri, classifications, code sets, and lists of controlled terms used in patient records [7,26]. As a result, it contains codes to concepts in most of the medication-oriented terminologies (see Fig. 2), including some that are proprietary. The purpose of the UMLS Metathesaurus is to support the semantic interoperability of different terminological systems by linking terms and concepts with the same meaning. These terms are linked through a Concept Unique Identifier (CUI) in the UMLS Metathesaurus. For example, the concept Heart (CUI C0018787) includes the names and codes for heart from many different terminology systems, including SNOMED CT and MeSH.

In many cases the source terminologies are only partially represented. For instance, the UMLS Metathesaurus represents only NDF-RT concepts that are above the ''packaged products'' level. It also represents just over 12,000 clinical drugs from MDDB (Medi-Span), which contains over 50,000 codes. The UMLS Metathesaurus also includes codes from SNOMED CT and RxNorm.

2.8. MDDB (Master Drug Data Base)

MDDB is a coding system and ontology maintained by Medi-Span [27]. It provides descriptive drug information on trade name, generic name, drug class, active ingredient, drug components such as drug strength, unit of measure and dosage form, manufacturer and packaging details.

MDDB contains links to many standard medication terminological systems such as NDC, UPC, and RxNorm. It uses industry standard identifiers such as the NDC for all trade and generic drugs in the database. It also has proprietary drug identifiers such as Generic Product Identifier (GPI) and a Generic Product Packaging Code (GPPC). MDDB is linked to RxNorm through the GPPC.

The GPI code has 14 digits with seven subsets. The first 10 digits define therapeutic class code and the last four define route of administration, dosage form, drug strength and unit of measure [17]. A representive GPI code is shown in Table 6.

GPPC is an eight digit numeric code. The first five characters of the GPPC represent the generic ingredient code(s), drug strength code, unit of measure code, and dosage form code. These first five characters are also called the MDDB code, which is the code that is listed in RxNorm. The last three digits of the GPPC represent packaging size and type.

2.9. NDDF plus (National Drug Data File Plus)

NDDF plus is a coding system that provides descriptive drug information about trade names, generic names, drug class, drug components such as drug strength, unit of measure, dosage form, physiological effect, etc. It is a drug database produced by First DataBank (FDB). First DataBank uses unique drug identifiers from NDDF Plus that can be connected to RxNorm. The database can also be linked to any 11-digit NDC. The NDDF also includes knowledge bases that help physicians identify drug–drug interaction, drug–allergy interactions, drug dosing and administration details [28].

2.10. Lexicon

Lexicon is a drug database used by Cerner Multum, which include information about trade name, generic name, active ingredient, drug components such as drug strength and unit of measure, dosage form and route of administration. It is a subset of VantageRx Database. This is connected to proprietary system VantageRx Database and standard terminological systems RxNorm and NDC [29].

2.11. VantageRx Database

VantageRx Database is a coding system from Cerner Multum that includes proprietary drug identifiers for trade name drugs and generic drugs. It provides the drug information in the form of active ingredients, unit of measure and dosage form, route of administration, mechanism of action, packaging type and size, manufacturer information, and wholesale drug pricing. It also provides detailed information about Adverse Drug Events through drug-drug, drug-food, drug-disease and drug-allergy interactions. This proprietary system is connected to different coding systems such as RxNorm, NDC, J-codes and ICD-9-CM to enhance interoperability [30].

2.12. Alchemy

Alchemy is a drug database by Gold Standard/Elsevier. Alchemy is a coding system because it has proprietary codes for drugs, and taxonomy of drug classes. Alchemy provides trade names, generic names, drug strength, unit of measure, drug form, route of administration, physiologic effect, mechanism of action, and information about drug-drug and drug-allergy interactions. It also gives information about prescription indications, therapeutic intent, and off label uses. It is connected to standard medication terminological systems such as RxNorm and NDC. Alchemy uses a proprietary identifier to track product/package ID numbers. NDCs are attributes of the product, so changes to the NDC do not affect the structure of Alchemy. As described in the section regarding the NDC, manufacturers are free to reuse codes from obsolete products. Alchemy uses versioning to track these changes, ensuring clear product identification [31].

Each of the above commercial proprietary coding systems has its own, usually normalized, proprietary codes mapped to NDCs. Most, if not all, of the proprietary medication terminological systems available by subscription to vendors that sell and market medication prescribing systems have a coding system that includes: trade name, generic name, active ingredient(s), formulation and dose, route of administration, date of obsolescence, dose-range checking, prescription monographs for patients and health care providers, mechanism of action, drug class, drug–drug interactions, drug–allergy interactions, drug-disease interactions, indications, and even images of medication products and wholesale price information in one place. This is market validation that this information is relevant to clinicians at the point of care. However, these systems are only partially connected to more commonly used standards such as RxNorm and SNOMED CT. As a result, proprietary systems still require considerable mapping effort to integrate with other standards (either proprietary or open).

3. Case study: Mapping from Allscripts medications to SNOMED CT and the UMLS Metathesaurus

At the time of this case study, the clinical data warehouse (CDW) developed by the Center for Clinical and Translational Science at the University of Texas Health Science Center at Houston (UTHSCH) contained medical records on approximately 364,000 patients treated by clinicians using the Allscripts electronic health record (Allscripts, Chicago, IL). The data were received from the Allscripts outpatient medical record system used by UT Physicians; the UTHSCH clinical practice plan. The Center stored two forms of these data: one in its original SQL database and one in an i2b2 database [32]. One of the goals of the Center is to assist researchers by accurately retrieving patient records that match specific criteria. In most cases, these criteria are given in terms of an informal information need, such as "All adult patients who were diagnosed with generalized anxiety disorder and prescribed an anxiolytic" rather than in terms of the specific codes stored in the CDW. Center staff must then translate these informal queries into queries for stored codes—a time consuming and error-prone process. To improve this process we have begun to map codes in the CDW to concept codes from standard hierarchical terminological systems. These mappings offer three main benefits: (1) the concept hierarchies in standard ontologies allow us to automatically translate queries based on classes of concepts into the detailed codes required to retrieve records; (2) the additional information stored in standard ontologies makes it possible to conduct more knowledge-based searches, such as those based on the ingredients of a drug or possible side effects; and (3) mapping to standard ontologies allows interoperability with other databases, including data from other informatics research projects in the Center.

Although the Allscripts database included a hierarchy of drug classes, each drug was restricted to a single drug class, meaning that queries for classes of drugs could produce incomplete results. In addition, these drug classes were not directly linked to any external terminology systems, so they did not meet our interoperability requirement. Thus, the first challenge was to select a target terminology for the mapping project. Evaluating possible target terminologies is difficult in part because formal evaluations lag terminologies by one or more years, and in part, because each use-case can present unique needs that may not be addressed by published evaluations. For example, a terminology might provide good content coverage, but may be too granular (or not granular enough) for a specific need, or may not be appropriate as an interface terminology.

We chose to map to SNOMED CT and the UMLS Metathesaurus for several reasons. The main motivation was that SNOMED CT uses a polyhierarchical drug classification system, meaning that a drug can be listed in one or more drug classes. Secondary considerations were that SNOMED CT along with the UMLS Metathesaurus provided access to a variety of additional knowledge sources; the Health Information Technology Standards Panel's recommendation of SNOMED CT as a US standard in their interoperability specification [33]; and its consistently high ranking on content coverage. We also considered NDF-RT, but chose not to use it because its drug classification system restricted drugs to a single class, and we were not sure that NDF-RT's alternative classification method, using mechanism of action, and pharmacologic effect would be familiar to clinical users. Although we mapped some of the drugs using RxNorm, we did not choose RxNorm as the primary target, because it lacked a classification hierarchy.

While we reviewed medication terminology systems, we simultaneously explored possible methods for mapping codes from Allscripts to the target system. Allscripts, prescribed medications were stored using a proprietary Drug Description Identifier (DDI) code that was also linked to database tables that provided dosage forms, drug strengths, units of measure, medication names, NDC, MDDB and GPI codes, as shown in Table 7. Although there are

Table 6
GPI code for an anti-depressant.

GPI	Coding	Example
58	Drug group	Antidepressants
58-20-	Drug class	Tricyclic agents
58-20-00-	Drug sub-class	–
58-20-00-60-	Drug name	Nortiptyline
58-20-00-60-10-	Drug name extension	Hydrochloride
58-20-00-60-10-01	Drug strength and unit of measure	10 mg
58-20-00-60-10-01-05	Dosage form	Capsule

approximately 50,000 unique DDI codes in Allscripts, the patient records in the CDW contained only 8500 unique DDIs. To simplify the mapping process, we limited our efforts to this subset of DDI codes. In the remainder of this paper, we discuss all results in the context of this subset.

Our goal was to map as many source medication codes as possible to concepts that captured drug (active ingredients), dosage form, and drug strength. If there was no exact match, we attempted to map to a concept that preserved as much of the drug, dosage form and drug strength as possible. For example, the closest match for "acetaminophen tablet 500 mg" might be "oral form of acetaminophen." We selected matches in the following order: (1) drug, dosage form and drug strength; (2) drug, dosage form (3) drug; (4) the most specific drug class. To prevent loss of granularity, when option 4 was required, we added the specific medication as a subclass of the selected SNOMED CT concept in our local version of the SNOMED CT class hierarchy.

At this point in the project, we expected that RxNorm's inclusion of both MDDB and SNOMED CT codes would cover a large percentage of the 8500 unique DDIs. However, as described below, we found that three different methods were required to obtain a complete mapping. We arrived at these three methods via an iterative process in which we identified the most direct method, carried out the mapping, determined which medications could not be mapped using that method, and then searched for an alternative method for the unmapped medications. Each of the methods increased the completeness of the mapping.

3.1. Method 1: Automatic Mapping

The Automatic Mapping method mapped each drug from Allscripts DDI codes to SNOMED CT and the UMLS Metathesaurus using the following steps: (1) map the DDI code to its corresponding MDDB code using tables in the Allscripts database; (2) map the MDDB code to an RxCUI and a SNOMED CT concept code using RxNorm's RXCONSO table; (3) map the SNOMED CT code to the corresponding UMLS Metathesaurus CUI and the Atom Unique Identifier (AUI) for the SNOMED CT concept code using the UMLS Metathesaurus' MRCONSO table; and (4) map the AUI for the SNOMED CT concept code to the corresponding SNOMED CT hierarchy tree code stored in the PTR (Path to Root) attribute of the UMLS Metathesaurus' MRHIER table. We used the 10/5/2009 full release of RxNorm provided through the UMLS website and the 4/6/2009 release of the UMLS (2009AA). We searched for MDDB codes in the RXCONSO database table, a table that includes the terms and identifiers from RxNorm's source vocabularies along with their corresponding RxCUI. This method was not able to map all drugs. First, approximately 5% of MDDB codes were not listed in RxNorm, and some of the MDDB codes that mapped to RxNorm did not have SNOMED CT codes listed in RxNorm. Second, even though SNOMED CT contained a large number of drugs, we accessed the SNOMED CT vocabulary through the UMLS Metathesaurus and not all drugs could be automatically mapped using this method, because 14% of SNOMED CT concept names that have different meanings in SNOMED CT were treated as synonyms in the UMLS Metathesaurus. [34]. In addition, RxNorm did not cover every drug listed in Allscripts, such as some compound drugs and nutritional supplements. Finally, some concepts listed under the medication list in Allscripts and prescribed by physicians, such as needles, syringes, and test kits, were listed in the physical object hierarchy in SNOMED CT and hence could not be mapped under the pharmaceutical/biological product or substance hierarchy. The physical objects (medical devices and appliances) were mapped manually (see method 3, below), and thus are included in our evaluation results.

3.2. Method 2: String based mapping using a semi-automated mapping tool

We developed a semi-automated tool that used string matching to determine and display potential matches between the Allscripts generic drug name and SNOMED CT drug names found in the UMLS Metathesaurus' MRCONSO table. A human expert could select a generic drug from the Allscripts drug list displayed on the left pane of the mapping tool. This highlighted the selected drug name on the left and then displayed the possible SNOMED CT drug names and SNOMED CT parents to the right. Once the expert selected the best match from the right pane, the tool showed the Allscripts drug name (for example, the trade name), and the SNOMED CT Path to Root (PTR) hierarchies at the bottom of the same window. To identify possible SNOMED CT matches, the tool matched generic drug names from Allscripts to generic drug names in RxNorm's RXCONSO table, then used the same table to map the generic names to corresponding RxCUIs and SNOMED CT concept codes. The remainder of the process was identical to steps 3–5 of the automated method described in Section 3.1. We used this tool as part of two slightly different sub-methods, described below.

3.3. Method 2a: String based method using Direct Map from generic names (Mapping Tool-DM)

This method was used to find the best SNOMED CT drug name for the generic name in Allscripts. There were 1-to-many (1:M) matches from an AllScripts drug name to SNOMED CT drug names, because the same drug often appeared with different strengths and dosage forms. Hence the expert had to determine the best match after reviewing the PTR hierarchies. In addition, for compound drugs, the expert needed to select all ingredients. Finally, some Allscripts drugs were identified only by trade name whereas all drugs in SNOMED CT were identified by generic name. Hence trade name drugs could not be mapped using this method.

3.3.1. Method 2b: String based method using Partially Automated Map (Mapping Tool-PAM)

This method is similar to Method 2a except that it maps trade names. Using RxNav (NLM's browser for RxNorm) [35] we mapped trade names to generic names and then used the mapping tool to map to SNOMED CT. In cases where it was not possible to get the generic name for a trade name using RxNav, we used other online sources, such as Drugs.com [36].

3.3.2. Method 3: Manual Mapping using SNOMED CT Browser (Manual Mapping)

In this method we used a SNOMED CT Browser maintained by Virginia–Maryland Regional College of Veterinary Medicine [37] to manually search for the Allscripts drug name. We narrowed down the results using the pharmaceutical/biological product hierarchy, then selected the best match, taking into consideration the drug ingredients, drug strength, unit of measure, dosage form and route of administration (if any). Whenever possible, we mapped each drug ingredient at the most complete level (ingredient, strength, unit of measure and dosage form) of the pharmaceutical/biological product hierarchy in order to include all possible parents and all possible nodes of a given parent for the given drug. One advantage of this method is that we could trace the entire PTR hierarchy to find the best match. However, this method is entirely manual, more time consuming and relies heavily on a human expert's knowledge of SNOMED CT. SNOMED CT codes mapped in this manner were also then mapped to corresponding UMLS Metathesaurus AUIs and CUIs.

3.4. Manual review and correction

To verify the mappings, the expert (HS) used CliniClue [38] (a SNOMED CT browser) to display the concept and PTR hierarchy for each assigned SNOMED CT code, then evaluated the specific mapping according to the following criteria:

Completeness: The assigned SNOMED CT code represents all ingredients in the drug, including multiple ingredients for compound drugs.

Correctness: The assigned SNOMED CT code represents the same medication.

Accuracy (best-fit): If the exact match was not found, the assigned SNOMED CT code represents the best match. In this case the drug was mapped to its immediate parent node and classified as a Subclass of the parent. For example, there was no exact match for Triamcinolone Diacetate Micronized Powder in the Pharmaceutical/biological product branch of SNOMED CT, so it was mapped as a subclass of the respiratory form of triamcinolone. Since this effectively extends the SNOMED CT hierarchy to include the specific drug, there was no loss of granularity in the mapping.

If any of these criteria were not met, we documented the problem and attempted to correct it. Completeness, correctness, and accuracy were calculated for both the uncorrected and corrected mappings for each method.

3.5. Validation

A second human expert (SJ) reviewed a random sample of corrected drug mappings generated by each of the three methods. The sample size of each method was calculated using Yamane's simplified formula for proportions, using a 95% confidence interval and variability (P) of 0.5 (in our case agree or disagree) with 3% of precision [39]. For Set A (Automatic Map) a sample size of 20% was generated for review; for set B (String based method using the mapping tool) a sample size of 52% was selected for review, and for set C (Manual Mapping using SNOMED CT browser) a sample size of 41% was selected for review. The sample sizes varied for the three methods due to differences in their population sizes (the smaller the population, larger the percent sample sizes needed). Overall, 45% of the encoded terms were verified by a second human expert (SJ) using an online SNOMED CT Browser built and maintained by the Virginia-Maryland Regional College of Veterinary Medicine [37]. Differences were resolved by consensus. Inter-rater reliability was determined using the "joint probability of agreement method" [40].

4. Results

Fig. 3 gives an overview of the mapping of Allscripts drugs (coded with DDIs) to SNOMED CT using the three methods described above. Overall, of the 8447 DDIs, 8000 were mapped to RxNorm. The remaining 447 drugs (451 codes including overlaps) could not be mapped either because of their absence in RxNorm or (more commonly) because of the absence of a link from RxNorm to SNOMED CT. There were four Allscripts drug overlaps, which meant that those drugs were present in both sets of results (Mapped to RxNorm and Not Mapped to RxNorm). This arose because four DDIs had more than one MDDB code listed in Allscripts, as shown in Table 8. As a result, the same DDI code can have two or more different MDDB codes, some of which appear in RxNorm and others that do not. Table 8 shows an example in which the same DDI is associated with two different MDDB codes. Only the first MDDB code is listed in RxNorm. We later discovered that the second code (09569) was obsolete.

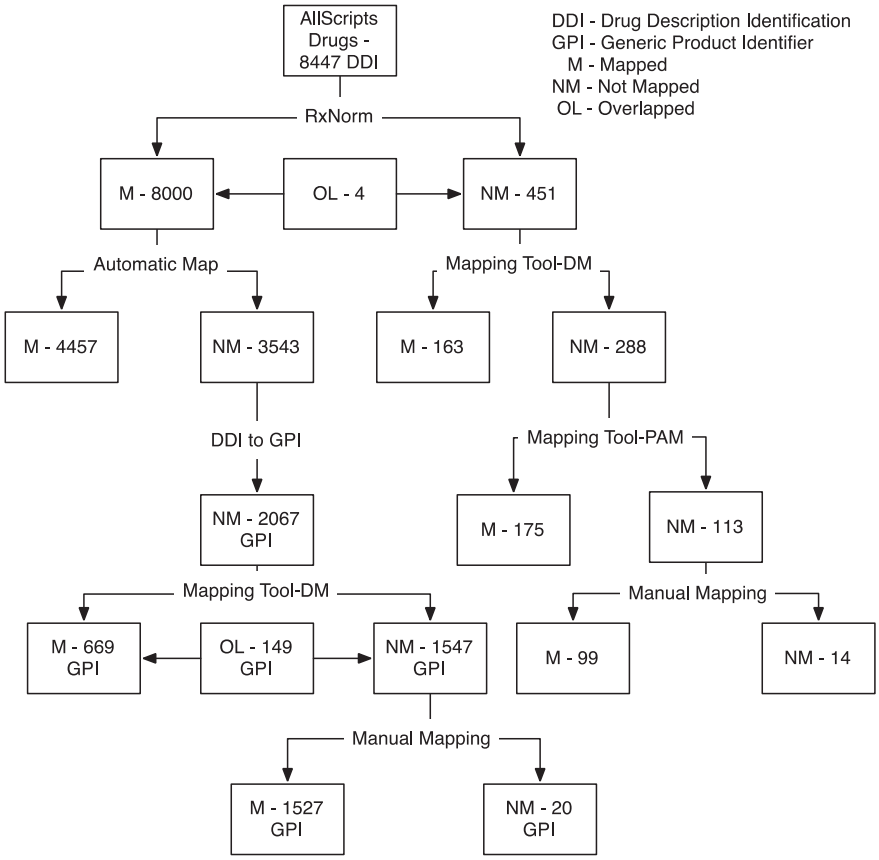

Fig. 3. Methods used for mapping Allscripts drugs to SNOMED CT.

Table 7
Representative example of links from DDI to other codes that include NDC, MDDB, GPI and drug components such as drug form, strength, unit of measure and drug display name.

DDI	Dosage form	Drug strength	Unit of measure	Display name	NDC	MDDB	GPI
284	TABS	500	MG	Acetaminophen 500 MG Tablet	68387021430	12340	64200
22567	TABS	500	MG	Tylenol Extra Strength 500 MG Tablet	54868333700	25999	64200

Of 8000 drugs, 4457 were mapped automatically using method 1: mapping from DDI, to MDDB, to RxNorm, then SNOMED CT. The majority of the 3543 drugs with DDIs that could not be mapped automatically were compound drugs containing more than one ingredient. To map these drugs, we mapped from their trade name-specific DDI codes to the corresponding generic ingredient-specific GPI codes—a many-to-1 (M:1) mapping. Since all drugs in SNOMED CT are represented by their generic name, this did not result in a loss of granularity. Thus 3543 DDIs were reduced to 2067 GPIs. Of the 451 drugs that could not be mapped due to the lack of MDDB codes in RxNorm or a link from RxNorm to SNOMED CT (see NM – 451 in Fig. 3), 163 were mapped using Mapping Tool-DM, and 175 were trade name drugs that were mapped using Mapping Tool-PAM. The remaining 99 were mapped using Manual Mapping. Fourteen drugs were not mapped due to their absence in SNOMED CT.

The branch of Fig. 3 labeled "DDI to GPI" gives an overview of the mapping of 2067 Allscripts drugs (with GPIs). First, we use Mapping Tool-DM to map drugs to SNOMED CT based on the generic name associated with each GPI. Of the 2067 GPI codes, only 669 could be mapped using this tool. The remainder did not directly match any names in SNOMED CT. We mapped the remaining 1547 GPI codes using Manual Mapping. Twenty drugs remained unmapped due to their absence in SNOMED CT.

Table 9 shows the distribution of the number of drugs mapped using different methods, drugs that could not be mapped, duplication loss from DDI to GPI (due to the many to one mapping), and the number of duplicate mappings throughout the mapping process. Once the mapping was completed, one author (HS) performed a manual review of the entire mapping to evaluate the accuracy, correctness, and (for compound drugs only) completeness of the mapping by the method used and to produce a corrected mapping. Prior to making corrections, the accuracy of each of the methods was: Automatic Mapping, 99.93%; Manual Mapping, 99.19%; Mapping Tool-PAM, 65%, Mapping Tool-DM, 51%. Likewise, correctness was: Automatic Mapping, 100%; Manual Mapping, 99.82%, Mapping Tool-PAM, 98.29%, and Mapping Tool-DM, 98.92%.

Table 10 gives an overview of the completeness of the mapping for compound drugs with ingredients ranging from two to six. Overall, of the 1516 compound drugs, 172 were incompletely mapped for an overall completeness of 88.65%. By method, completeness was: Automatic Mapping, 100%; Manual Mapping, 97.92%; Mapping Tool-DM, 48.42%; and Mapping Tool-PAM 75.86%.

Table 11 gives an overview of the overall mapping results before and after manual review and corrections. Our goal was to map as many of the drugs as possible to concepts in the pharmaceutical/biological product branch. However, the 8447 actively prescribed "medications" from the CDW also included over-the-counter drugs, herbal medications, test kits, and hospital required materials such as I.V. tubing, needles, lancets, and syringes—some of which are not classified in the pharmaceutical/biological product branch. Therefore, the mapping included concepts from the substance and physical object branches of SNOMED CT. Of the 8447 entries, after corrections, 96% were mapped to the pharmaceutical/biological product branch, 2.06% to the substance branch, 1% to the physical object branch, and 0.48% were not mapped due to their absence in SNOMED CT.

Table 9
Quantitative overview of mapping process.

mapping method	No. of drugs	Final calculation	No. of drugs
Automatic Map	4457 (62.56%)	Total mapped	7124
Mapping Tool-DM	832 (11.68%)	DDI to GPI loss	+1476
Mapping Tool-PAM	175 (2.46%)	Duplicates	−153
Manual mapping	1626 (22.82%)	Final number	8447
Not mapped	34 (0.48%)		
Total	7124 (100%)		

The corrected mapping was verified independently by a second author (SJ). The inter-rater reliability was 98.30% for the Automatic Map Method, 96.02% for the String based method, and 84.51% for the Manual Map method. Inter-rater reliability for the combined mapping (aggregating across all three methods) was greater than 80%. The two evaluators (H.S. and S.J.) resolved all disagreements to produce a final mapping.

5. Discussion

The medication mapping case study presented here demonstrates the challenges of mapping across terminological systems. Only 62.56% of medications in our source vocabulary could be mapped automatically, using existing connections among terminological systems. An additional 14% were mapped using a semi-automated process based on matching medication names, but 23% had to be mapped manually. Only 34 of 7124 drugs could not be mapped due to missing concepts in the target terminological system.

One limitation is that we used fairly simple approaches to mapping medications that we could not map automatically. More advanced "ontology matching" algorithms may perform better than the ones we used. However, studies show that even these methods require significant expert intervention [41].

This study demonstrates several challenges to mapping across terminological systems. First, given the rapid evolution of terminological systems and the time required to evaluate and publish results, it is often difficult to find up-to-date information on the content coverage of existing systems or the linkages between systems. This can make it difficult to determine the suitability of a particular terminology for a specific use-case. Different dimensions

Table 8
Representative example of how overlaps occur in the mapping process resulting in a situation in which the same drug can be counted as both mapped and not mapped to RxNorm.

Drug no.	Allscripts DDI	Medi-Span MDDB	RxNorm RxCUI	Drug name
1	11612	29752	283880	Amylases 15,000 UNT/lipase 1200 UNT/protease 15,000 UNT Oral Capsule
2	11612	09569		Amylases 15,000 UNT/lipase 1200 UNT/protease 15,000 UNT Oral Capsule

Table 10
Comparison of the completeness of the mapping for compound drugs before and after correction based on number of ingredients. Columns marked CM contain the number of drugs that were completely mapped, whereas columns marked ICM show the number of drugs that had 1 to $n-1$ ingredients mapped, where n is the total number of ingredients listed at the top of the table. Cells containing multiple numbers show the number of drugs following, in parentheses, by the number of ingredients those drugs were mapped to.

	Ingredient No. Methods									
	2		3		4		5		6	
	CM	ICM	CM	ICM	CM	ICM	CM	ICM	CM	ICM
Automatic Mapping	549	0	87	0	8	0	–	–	–	–
Manual Mapping	322	9	164	1	32	1	16	0	4	0
Mapping Tool-DM	71	45	60	33 (2) 14 (1)	6	17 (3) 7 (2) 2 (1)	1	13 (3) 3 (2) 7 (1)	0	3 (3) 2 (2) 1 (1)
Mapping Tool-PAM	31	7	12	1 (2) 4 (1)	1	1	0	1	–	–
Total	973	61	303	53	47	28	17	24	4	6

Table 11
Comparison of data before and after correction.

Data analysis	Before correction (%)	After correction (%)
Completeness	91.16	99.73
Correctness	97.56	100
Accuracy	97.98	100
Substance/physical object	4.29	3.06
Not mapped	1.81	0.48

of terminological systems, such as content coverage, granularity, suitability as an interface terminology, or semantic interoperability with other systems, often require separate evaluations that may not be available at the time a project is initiated. In some cases, judging suitability might only be possible after first attempting and evaluating a mapping.

Second, as terminologies become more comprehensive, better linked, more frequently updated, and more computable, they also increase in complexity. This in turn, means that although they are ultimately more useful, they are harder to use correctly and it may be harder to judge their applicability to a particular problem.

Third, comprehensive cross-terminology mapping and evaluation projects may require significantly more time, effort, and terminology-specific expertise than can be predicted at the start of a project. The project described above required approximately one FTE over an entire year. However, our initial estimate, made after discovering the direct link from the source terminology to RxNorm was that it would take one FTE for just 3 months. Upon discovering that the existing linkages were not sufficient for a complete mapping, we had to explore several approaches before finally settling on the ones presented above.

Another challenge is the need to assess and address differences in granularity between the source and target terminologies, as well as any intermediate terminologies used in the mapping. For some terminological systems these differences may be relatively obvious. For instance, ICD-10 is obviously more granular than ICD-9. However, in more complex mapping projects, such as the one described above, the granularity may vary by concept. For instance, for a subset of medications, we were forced to extend the local instance of the SNOMED CT classification hierarchy with additional concepts to preserve granularity.

Further, all mappings must be maintained and updated as errors are found and corrected, and as the source and target terminologies change. For instance, during this project the source terminology was updated, forcing us to reconcile differences between the old and new versions and update our ongoing map. Changes to the source terminology often means that future data flowing into a data warehouse will be coded with the new version, whereas pre-existing data retains the old coding system. The difficulty of maintaining a mapping depends on whether the terminological systems follow best-practices (such as not reusing codes for new concepts) and on the mapping approach. For example, if the approach is fully automated, it may be possible to rerun the mapping algorithm, then computationally determine differences between the old and new mapping to determine what additional effort is needed. However, if a significant part of the mapping was done manually, maintenance may also require significant manual effort.

To address these challenges, we need to develop better methods for automatically linking concepts across systems and maintaining these links. Unfortunately, as with the case study presented here, creating new connections between systems often requires considerable manual effort [41]. Highly interactive partially-automated mapping tools that are directed by human experts to automate parts of the mapping process with specific expert input are a promising alternative to fully automated methods. For example, an expert might indicate that two concepts from two different standards are synonymous and then direct an Automatic Mapping tool to map only subconcepts of the two. Another approach, first explored in the Galen project, is to provide users with a convenient language for logically defining the terms and relations in an ontology using a common reference model, followed by the application of algorithms that can infer conceptual mappings and class/subclass relationships among terms from different source vocabularies [42]. A combination of these approaches is worth exploring, because satisfactory automated approaches that do not require significant human intervention may not be possible without significant artificial intelligence breakthroughs.

These challenges can also be addressed by setting terminology standards in clinical information systems. Such standards would decrease the number of mappings needed and allow terminology developers to focus on creating high quality links to and from a smaller set of systems. Stage 1 of the "meaningful use" rule adopted in August 2010, by the US Department of Health and Human Services, Centers for Medicare & Medicaid Services, includes a number of terminology standards in the certification criteria for electronic health records (see Federal Registry, 45 CFR Part 170.207). However, some of these standards (such as for medications) apply only to the exchange of health information, not to codes that are internal to a single system, whereas others (such as the use of SNOMED CT for problem lists) are required for internal use and information exchange. The standard for medications permits certified systems to exchange medication information using any one of the vocabularies included in RxNorm. Although these requirements fall short of a single, internal standard for the most important types of health information, the intent is to provide a "glide-path" to the adoption of single standards.

Finally, based on our experience with this project we offer several recommendations to others attempting similar cross-termi-

nology mappings. The first is to precisely define specific use-cases and then discuss those use-cases, as early as possible, with the developers of the related terminologies. The second is to carefully consider and plan for the dynamic nature of terminological systems. There are at least two issues to consider here. The first is that the results of papers that use or evaluate terminological systems may be partly (or completely) out of date or the specific evaluation needed to assess a terminology may not be available. The second is that any mapping will need to be updated on a regular basis, meaning that the choice of mapping methodology should consider and include a plan for regular updates. One possible approach for quickly determining the suitability of a terminology and possible mapping methods is to select a small random subset of the source terminology codes to map and evaluate. Another is to focus the mapping effort around only the most important codes, such as those that are most frequently used in the source data, or most frequently appear in queries. For example, in recent work at one of the author's institutions we choose to validate the ten most common diagnostic codes extracted from an inpatient medical record system, because they are also the most important concepts to researchers at the institution.

6. Conclusion

Standardized biomedical terminologies are essential for making use of the growing amount of research, clinical and public health data. However, despite the increasing quality, scope, and cross-linkages, there remain several challenges to mapping across systems. These include (1) the availability of up-to-date information to assess the suitability of a given terminological system for a particular task, and to assess the quality and completeness of cross-terminology links; (2) the difficulty of correctly using complex, rapidly evolving, modern terminologies; (3) the time and effort required to complete and evaluate the mapping; (4) the need to address differences in granularity between the source and target terminologies; and (5) the need to continuously update the mapping as terminological systems continue to evolve.

Acknowledgments

This work was supported in part by National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through Grants 3UL1RR024148, UL1RR033173, UL1TR000117, HRSA Grant D1BRH20410, and Grants 10510592 and 90TR0002 under the Strategic Health IT Advanced Research Projects Program (SHARP) from the Office of the National Coordinator for Health Information Technology;. S.L. Jones was supported by a training fellowship from the Keck Center for Interdisciplinary Bioscience Training of the Gulf Coast Consortia (NLM Grant No. 5T15LM007093). We thank the anonymous reviewers and the guest editors of this special issue for their comments, suggestions, and tenacity.

References

[1] Krzywinski M, Schein J, Birol İ, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res 2009;19(9):1639–45.
[2] de Keizer NF, Abu-Hanna A, Zwetsloot-Schonk JH. Understanding terminological systems. I: Terminology and typology. Methods Inf Med 2000 Mar;39(1):16–21.
[3] Federal Medication Terminologies – National Cancer Institute. <http://www.cancer.gov/cancertopics/cancerlibrary/terminologyresources/fmt> (accessed 15.06.11).
[4] Office of the Commissioner, Center for Biologics Evaluation and Research. Substance Registration System – Unique Ingredient Identifier (UNII). <http://www.fda.gov/ForIndustry/DataStandards/SubstanceRegistrationSystem-UniqueIngredientIdentifierUNII/default.htm> (accessed 15.06.11).
[5] de Coronado S, Haber MW, Sioutos N, Tuttle MS, Wright LW. NCI Thesaurus: using science-based terminology to integrate cancer research results. In: Medinfo 2004: proceedings of the 11th world conference on medical informatics, San Francisco; 7–11 september 2004. p. 33.
[6] U.S. Department of Veterans Affairs, Veterans Health Administration. National Drug File – Reference Terminology (NDF-RT) Documentation. December 2010 Version. U.S. Department of Veterans Affairs, Veterans Health Administration; 2010.
[7] Lindberg DA, Humphreys BL, McCray AT. The unified medical language system. Methods Inf Med 1993;32(4):281.
[8] Center for Drug Evaluation and Research. Drug Approvals and Databases – National Drug Code Directory. <http://www.fda.gov/Drugs/InformationOnDrugs/ucm142438.htm> (accessed 15.06.11).
[9] Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med 1998;37(4):394–403.
[10] 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 Oct 1;66(19):1743–53.
[11] de Coronado S, Wright LW, Fragoso G, Haber MW, Hahn-Dantona EA, Hartel FW, et al. The NCI Thesaurus quality assurance life cycle. J Biomed Inform 2009 Jun;42(3):530–9.
[12] The National Cancer Institute's Thesaurus and Ontology; 8 March 2011. <http://www.websemanticsjournal.org/index.php/ps/article/view/27> (accessed 15.06.11).
[13] Covitz PA, Hartel F, Schaefer C, De Coronado S, Fragoso G, Sahni H, et al. CaCORE: a common infrastructure for cancer informatics. Bioinformatics 2003;19(18):2404.
[14] Cancer Common Ontologic Representation Environment (caCORE). <https://cabig.nci.nih.gov/tools/concepts/caCORE_overview> (accessed 15.06.11).
[15] Liu S, Ma W, Moore R, Ganesan V, Nelson S. RxNorm: prescription for electronic drug information exchange. IT Professional; 2005. p. 17–23.
[16] An Overview to RxNorm. <http://www.nlm.nih.gov/research/umls/rxnorm/overview.html> (accessed 16.10.11).
[17] Pathak J, Chute CG. Analyzing categorical information in two publicly available drug terminologies: RxNorm and NDF-RT. J Am Med Inf Assoc 2010;17(4):432.
[18] Pathak J, Chute CG. Further revamping VA's NDF-RT drug terminology for clinical research. J Am Med Inf Assoc; 17 March 2011. <http://jamia.bmj.com/content/early/2011/03/16/amiajnl-2011-000161.short> (accessed 31.03.11).
[19] Price C, Spackman K. SNOMED clinical terms. Br J Health Comp Inf Manage 2000;17(3):27–31.
[20] Insertion of SNOMED CT into the UMLS Metathesaurus: Explanatory Notes; 26 March 2004. <http://www.nlm.nih.gov/research/umls/Snomed/snomed_represented.html> (accessed 15.06.11).
[21] The International Health Terminology Standards Development Organisation. SNOMED clinical terms user guide. July 2009 international release. The International Health Terminology Standards Development Organisation; 2009.
[22] Elkin PL, Brown SH, Husser CS, Bauer BA, Wahner-Roedler D, Rosenbloom ST, et al. Evaluation of the content coverage of SNOMED CT: ability of SNOMED clinical terms to represent clinical problem lists. Mayo Clin Proc 2006 Jun 1;81(6):741–8.
[23] Fung KW, Hole WT, Nelson SJ, Srinivasan S, Powell T, Roth L. Integrating SNOMED CT into the UMLS: an exploration of different views of synonymy and quality of editing. J Am Med Inf Assoc 2005;12(4):486.
[24] Nadkarni Prakash M. Drug safety surveillance using de-identified EMR and claims data: issues and challenges. J Am Med Inf Assoc 2010;17(6):6671–4.
[25] Mortensen J, Bodenreider O. Comparing pharmacologic classes in NDF-RT and SNOMED CT. In: Proceedings of the fourth international symposium on semantic mining in biomedicine (SMBM 2010); 2010. p. 124.
[26] Fact Sheet: Unified Medical Language System; 23 March 2006. <http://www.nlm.nih.gov/pubs/factsheets/umls.html> (accessed 15.06.11).
[27] Wolters Kluwer Health – Medi-Span – Master Drug Data Base (MDDB®). <http://www.medi-span.com/master-drug-database.aspx> (accessed 09.07.11).
[28] NDDF Plus; 8 July 2010. <http://www.nlm.nih.gov/research/umls/sourcereleasedocs/2010AA/NDDF/> (accessed 08.07.10).
[29] Lexicon. <http://www.multum.com/Lexicon.htm> (accessed 15.06.11).
[30] VantageRx Database; 29 January 2011. <http://www.multum.com/VantageRxDB.htm> (accessed 29.01.11).
[31] Elsevier/Gold Standard – Advancing Healthcare through Medication Management Solutions. <http://www.goldstandard.com/index.html> (accessed 15.06.11).
[32] Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc 2010;17(2):124.
[33] HITSP Specifications Index Page. <http://wiki.hitsp.org/docs/#is>.
[34] Fung KW, Hole WT, Nelson SJ, Srinivasan S, Powell T, Roth L. Integrating SNOMED CT into the UMLS: an exploration of different views of synonymy and quality of editing. J Am Med Inf Assoc 2005 Aug;12(4):486–94.
[35] Bodenreider O, Nelson SJ. RxNav: a semantic navigation tool for clinical drugs. Medinfo; 2004. p. 1530.
[36] Drugs.com|Prescription drugs – information, interactions & side effects. <http://www.drugs.com/> (accessed 15.06.11).

[37] SNOMED CT Browser. <http://snomed.vetmed.vt.edu/sct/menu.cfm> (accessed 15.06.11).
[38] CliniClue; 2010. <http://www.cliniclue.com/> (accessed 07.07.10).
[39] Yamane T. Statistics: an introductory analysis. Harper & Row; 1964.
[40] Miller KJ, Vanni M. Inter-rater agreement measures and the refinement of metrics in the PLATO MT evaluation paradigm. In: Proc of the MT Summit X. Phuket, Thailand; 2005. p. 125–32.
[41] Shvaiko P, Euzenat J. Ten challenges for ontology matching. In: On the move to meaningful internet systems; 2008. p. 1164–82.
[42] Rector A, Rogers J. Ontological and practical issues in using a description logic to represent medical concept systems: experience from GALEN. Reasoning Web; 2006. p. 197–231.