Journal of Biomedical Informatics xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Methodological Review

An informatics framework for the standardized collection and analysis of medication data in networked research

Rachel L. Richesson

Duke University School of Nursing, 2007 Pearson Bldg, 307 Trent Drive, Durham, NC 27710, United States

ARTICLE INFO

Article history:
Received 27 August 2013
Accepted 3 January 2014
Available online xxxx

Keywords:
Medication terminological systems
Drug classification
Standards
Data coordinating centers
Data coding
Data analysis

ABSTRACT

Medication exposure is an important variable in virtually all clinical research, yet there is great variation in how the data are collected, coded, and analyzed. Coding and classification systems for medication data are heterogeneous in structure, and there is little guidance for implementing them, especially in large research networks and multi-site trials. Current practices for handling medication data in clinical trials have emerged from the requirements and limitations of paper-based data collection, but there are now many electronic tools to enable the collection and analysis of medication data. This paper reviews approaches to coding medication data in multi-site research contexts, and proposes a framework for the classification, reporting, and analysis of medication data. The framework can be used to develop tools for classifying medications in coded data sets to support context appropriate, explicit, and reproducible data analyses by researchers and secondary users in virtually all clinical research domains.

1. Introduction

Standardized coding and classification of medication data enable comparability within and across research studies and can bridge clinical and research networks and applications. However, there is currently no explicit standards-based process for organizing granular medication codes (e.g., "*acetaminophen*", "*Tylenol*", or "*acetaminophen 325 mg oral tablet*") into classes defined by various drug properties (e.g., "*antipyretic*" or "*prostaglandin receptor antagonists*"). Such a process is badly needed to ensure efficient, reproducible classification and analysis.

While property-based drug classification seems straightforward, there is ambiguity in the definitions of classes in coded data sets (e.g., "*Does the class 'antibiotics' in this analysis include topical antibiotics?*"), new indications ("*Is Donnatol™ included as a seizure medication in this analysis?*"), and new products ("*Is the recently approved drug Ravicti™ included in the group of urea cycle disorder medications in this data set?*"). Comprehensive reference terminology valid across the spectrum of clinical domains can ensure that the class memberships used in categorical analyses are explicit and easily reproducible, enabling comparability of different data sets. In this paper, we examine current approaches, propose a framework for standard medication coding and classification systems, and identify needed research and tools.

2. Background

2.1. Medication coding systems

Current systems for coding medication data include granular medication entities (e.g., the Unique Ingredient Identifiers [UNII] for active ingredients and chemical substances, and the National Drug Codes [NDC] for packaged products maintained by the U.S. FDA) with varying levels of adherence to sound coding management principles [1–3]. Other U.S. coding systems include the NDC Directory [4] and RxNorm [5,6], both of which include route of administration, in addition to the information on active ingredients, trade and generic names, drug strength, dosage form, and package size information found in the NDC codes. RxNorm has been shown to have near perfect coverage of ambulatory e-prescriptions [7] and it is being increasingly used in a variety of applications [8]. RxNorm does not include codes for dietary supplements such as mineral and herbal preparations, although there is no designated standard in this area for all research studies [9]. (It should be noted that the WHO Herbal Dictionary is used in many industry-sponsored trials [10].) In general, less specificity is required for coding of medications for research than is for clinical purposes (which require identifiers for branded products to support prescribing and pharmacy management activities).

2.2. Medication classification systems

RxNorm is linked to source vocabularies in the National Library of Medicine's (NLM) Unified Medical Language System (UMLS),

E-mail address: rachel.richesson@dm.duke.edu

1532-0464/$ - see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2014.01.002

including the National Drug File Reference Terminology (NDF-RT). NDF-RT was created by the Department of Veterans Affairs (VA) to organize drugs in the VA formulary, and it uses a logic-based model to group drug products into classes based on chemical structure, mechanism of action, physiological effect, drug–disease relationship describing therapeutic intent, pharmacokinetics describing the mechanisms of absorption and distribution of an administered drug in a body (e.g., hepatic metabolism), and legacy VA-NDF classes for pharmaceutical preparations (e.g., non-opioid analgesic). NDF-RT is the standard for the U.S. FDA's Structured Product Label (SPL) initiative [11].

Other reference terminologies and classifications that have formal systems for organizing and classifying medications include the Anatomical Therapeutic Chemical (ATC) Classification System maintained by the WHO Collaborating Centre for Drug Statistics Methodology and used widely in Europe [12]. Proprietary medication systems obviously use a formal class organization, presumably based on the American Hospital Formulary Service (AHFS Drug Information) [13,14], but they are not available for research or comparison. The fairly new Chebi ontology of chemical compounds of biological interest also contains medication compounds and information related to their molecular and biological function and interactions [15–17]. Existing reference terminologies and classifications, however, support different uses and user groups, and the identification of the most appropriate classification system is driven by the research question and the business case.

2.3. Current medication coding and classification in clinical research

To date, paper workflows have dominated clinical research [18]. Medication data for clinical trials are typically collected as free text and coded later, usually by a medical coder employed by the trial sponsor; errors and uncertainties are coded after data collection as they are discovered [19]. The completeness of data collection depends in part on the method of elicitation used by the investigator (e.g., open ended questions, symptom checklists) [19]. In industry-sponsored trials, quality control checks are performed by a clinician or pharmacist; reported medications are often compared with reported events/diseases to verify the accuracy of the information.

There has been little empirical research on the quality of coding in clinical trials. A 2012 systematic review [19] found only one published study on the validity and inter-rater reliability of adverse event coding using Medical Dictionary for Regulatory Activities (MedDRA) [20]. That study found that 12% of sampled adverse events were coded differently by two coders, and 8% of the coding was declared medically inaccurate by experts. Similarly, White et al. looked at 204 post marketing surveillance events to examine the impact of different classification systems on the identification of adverse events [21]. When the same verbatim text was coded with terms from MedDRA and the WHO Adverse Reaction Terminology (WHO-ART), 32 coded pairs (16%) of events were rated as medically different by expert reviewers.

Richesson et al. found high completion rates using RxNorm for coding in two different multi-site studies, but they did not evaluate coding accuracy [22]. The designation of the WHO Drug Dictionary as a medication coding standard for regulated clinical trials may eliminate the need for comparative research on coding schemes in industry-sponsored trials, but observational studies do not have the same restrictions. As electronic health records (EHRs) become more widely adopted, there will be greater potential to capture medication data from clinical repositories; the Clinical Data Standards Interchange Consortium (CDISC) is driving these efforts as part of its Electronic Source Data Interchange (eSDI) Group [18]. The emergence of "big data" and increasing interest in observational research are creating a sizable constituency of prospective data users who will demand free and open data standards. Thus comparative studies of the fitness of various medication classifications, and ways to integrate them into various workflows, are urgently needed.

"Big data" necessitate specialized data centers with massive capacities for data collection, storage, exchange, aggregation, visualization, and analysis [23]. Data centers for research networks and multi-site trials can benefit greatly from the use of standards to support data collection and analysis [24], yet there is little published guidance on how to implement these in large research networks and multi-site trials. The framework we provide can guide the development of tools to support improvements and standardization of medication data and enable sharing of research data sets and meaningful interpretation, comparison, communication, and application of results.

3. Framework for handling medication data in research environments

The major issues related to the standardization of medication data are (1) selection of a standardized medication coding scheme; (2) development of a process for collecting and coding data, including systems and interfaces to support data collection, entry, or coding that are customized to study needs and workflows; (3) choice of a classification system to support reporting and analysis; and (4) development of methods and tools for inserting the classification into the data management and analysis workflow. Fig. 1 represents this framework, including where potential sources of error lie and points of opportunity for informatics tools and theories to reduce error and increase efficiency.

3.1. Selection of a standard coding system

Medication data must be coded for analysis, and data centers can benefit from the re-use of tools, personnel, and training materials by using the same medication system for all studies. Cimino's desiderata [2] has obvious relevance for selecting a structurally sound coding system, but other considerations (e.g., regulatory reporting requirements, sponsor requirements, costs, licensing issues, and assurance of ongoing maintenance) are often more critical to research networks.

In the U.S., RxNorm is the designated standard for representing medication brand names, clinical drug names, and allergies/adverse events for medication products [25]. RxNorm uses a relational model to name drugs at different levels of specificity, ranging from active ingredients to packaged products with trade name, dosage, formulary and packing information. This is an appealing feature for data centers supporting multiple studies with multiple sources of medication data. RxNorm includes content from the FDA (UNII and NDC codes) as well as codes from commercial medication knowledge vendors. Because a core objective of the U.S. Meaningful Use regulations is to increase the use of computer provider order entry [26], it is likely that future regulations will require providers to use RxNorm for primary medication order entry [8].

Several organizations have reported mapping local medication code lists and free text data to RxNorm [7,27–30]. Mapping approaches include syntactic, semantic, and hybrid methods [31]. Others have demonstrated the feasibility of using RxNorm as a primary coding source (for medical history) [31] and research [9,32–34]. Bennett has described a live clinical system implementing RxNorm for primary data capture [31].

3.2. Systems and processes for implementing a medication coding system

As shown in Fig. 1, medication coding systems can be applied before, at, or after (electronic or paper-based) data collection. Once

Please cite this article in press as: Richesson RL. An informatics framework for the standardized collection and analysis of medication data in networked research. J Biomed Inform (2014), http://dx.doi.org/10.1016/j.jbi.2014.01.002