Inter-translation of Biomedical Coding Schemes Using UMLS (Extended Abstract)

Jeffery L. Painter and **Kristopher M. Kleiner**
Logic and Cognitive Science Initiative
North Carolina State University
Raleigh, North Carolina

Gary H. Merrill
Drug Development Sciences
GlaxoSmithKline Research and Development
Research Triangle Park, North Carolina

Abstract

We report the results of our work in using the Unified Medical Language System (UMLS) [1] to apply biomedical ontologies to practical problems faced by epidemiologists in extracting study cohorts from large disparate observational data bases.

Introduction

In the areas of drug discovery, pharmaceutical development, epidemiology, and drug safety, the representation of data in large data bases is most often expressed in terms of a number of different "coding schemes". These include ICD-9, Read Codes, MedDRA, CPT, and others. [2]

The data sources themselves are also disparate with regard to their format and content; and comparison and analysis of data across such disparate sources requires some way of translating among the coding scheme representations (or "normalizing" the references) so that references to the "same disease", "same condition", "same procedure", or "same drug" may be identified across the sources. In addition, ideally, a drug safety scientist or epidemiologist should be provided with some automated or semi-automated mechanism for determining what codes from one scheme are related to codes from another in either direct or indirect ways.

In the domain of epidemiology, even what initially appears to be a simple task of data extraction can be a daunting problem because of the disparity of the coding schemes and data involved. An epidemiologist may be required to extract a cohort of data for a study on "thrombocytopenia" from a set of data bases. But how does the epidemiologist determine the correct set of codes to use in extracting the data? In ICD-9 one set of codes associated with thrombocytopenia falls in the 287.x group (287.1, 287.3, 287.4, and 287.5), but not all codes in the 287.x group pertain to thrombocytopenia. And thrombocytopenia also falls under other ICD-9 categories (446.6 and 776.1) as well. In MedDRA, thrombocytopenia is associated with at least 17 different (though sometimes related) categories at different levels of its hierarchy. And how are these to be related in a reliable way to the appropriate ICD-9 codes, Read codes, and (even more challenging) CPT codes?

The practical importance of this problem is illustrated by the fact that as the final version of this abstract was being prepared, we were contacted by an epidemiologist with just such a difficulty. He was faced with extracting data from data bases in which medical conditions were coded in the ICD-10 and Read schemes, but he had been presented with ICD-9 codes as the basis for extracting the data.

These problems set the context for both our need to relate disparate biomedical ontologies and for the methods we use in meeting this challenge.

UMLS, Concepts, and Relations

The UMLS *Metathesaurus* is a set of "sources" which can be thought of as vocabularies, dictionaries, taxonomies, or ontologies. We take a purely ontological stance and in the case of what are more commonly referred to as "coding schemes", we view these as ontologies of medical conditions, procedures, etc. (which we refer to generically as *categories*) with a set of associated terms having these categories as their denotata or extensions.

UMLS imposes a more abstract structure across coding schemes through the use of classifiers known as *concepts* and *relations*. An often misguided notion is that the *concept* itself has a name (i.e., what epidemiologists or informaticians would call a "verbatim string") which denotes, connotes, or otherwise represents the concept. Rather, the UMLS-imposed *concept* – which can be identified solely by a CUI (concept unique identifier) – should be thought of as a purely abstract entity, having as its extension a (possibly empty) set of *realizations* in each source, and with each realization identified by an AUI (atomic unique identifier). An AUI is characterized by many attributes, including the code

[1] UMLS is a project of the (US) National Library of Medicine.

[2] By 'ICD-9' we mean to refer to ICD-9-CM, the International Classification of Diseases, 9th Revision, Clinical Modification, which is maintained jointly by the National Center for Health Statistics and the Health Care Financing Agency. ICD-10 is copyrighted by the World Health Organization, and ICD-10-CM is developed by the National Center for Health Statistcis. MedDRA® (Medical Dictionary for Regulatory Activities) is a registered trademark of the International Federation of Pharmaceutical Manufacturers Association. The Clinical Terms Version 3 (Read Codes)© are maintained by the (UK) National Health Service Information Authority. *Current Procedural Terminology* (CPT®) is copyright the American Medical Association.

associated with it in its particular coding scheme (source) and a verbatim string that would commonly be used by medical professionals.

Understanding how relationships available within the Metathesaurus can be utilized is key to creating a robust and meaningful system. For our purposes, we have classified the most useful (binary) relations among objects into categories based on their strengths. In *direct strong relations*, a single CUI is realized by two AUIs in the same source; in *indirect strong relations*, the CUI is realized by AUIs in two different sources; and in *indirect weak relations*, there is a potentially complex AUI(realization)/CUI(concept) chain between the AUI relata. This classification of relations allows epidemiologists to assess the likely significance of connections among codes within and across schemes.

While the first two types of relations are easily found, the weak relations prove to be more challenging and possibly more useful. Although the indirect strong technique does allow for the transcending of sources, it is limited in its capacity to transcend disparate types of coding schemes. An example of this is found when we wish to discover codes for medical *conditions* (ICD-9) that are associated with medical *procedures* (CPT).

Procedures and Conditions

Developing a strategy of mapping between procedure and condition concepts has been a major challenge in our project. Certain coding schemes are similar in both content and organizing structure, which makes for simple translation. However, this is not always the case. We illustrate the difficulties that arise when translating between two dissimilar coding schemes by considering the case of ICD-9 and CPT.

ICD-9 codes represent medical *conditions* (e.g. infectious and parasitic diseases, injury and poisoning, etc.), while CPT codes represent medical *procedures* (e.g. x-ray, CT scan, appendectomy, etc.). At first it may seem possible to relate medical conditions to procedures by assuming that a form of the verbatim string of the condition ("appendicitis") will appear as a part of the verbatim string of the procedure ("appendectomy"). This assumption does not hold in most cases, and so the question of how to relate the two coding schemes remains.

Additionally, coding schemes often differ in their hierarchical structures. ICD-9 has a "source asserted" hierarchy (i.e., one that is part of ICD-9 itself), while CPT lacks such an intrinsic hierarchy. As a consequence, UMLS imposes a hierarchy on CPT, and this is represented in the MTHCH [3] source.

This dissimilarity in structure provides further difficulties in the inter-translation of coding schemes. Certainly it seems clear that many medical procedures are associated with some medical condition or other (such as the association of a biopsy with cancer), but identifying meaningful associations of this sort is non-trivial. First, it is impossible to use hierarchical relations (such as parent, child, or sibling) to establish such relations by ascending in one of these ontologies to a high "conceptual level" via the CUI hierarchy and then descending to concept (CUI) realizations in the alternate coding scheme. UMLS was not designed to meet this particular goal. UMLS itself only provides semantically oriented mappings of objects in and between ontologies. And consequently it is infeasible to start with a CPT (procedure) category, ascend to a higher level, and then "project" the concept at that level "downward" onto an ICD-9 (condition) category without broadening the concepts involved to the point of meaninglessness. Another way of describing this situation is to make the (on reflection, rather obvious) observation that a medical *condition* concept cannot be realized by a medical *procedure*.

As a result, the higher-level categories in ICD-9 are difficult at best to associate with CPT categories. We explored methods that would take advantage of the relations which provide a link between coding schemes. Our hope was to make use of other relationships provided by UMLS to map between procedures and conditions. UMLS provides a number of non-hierarchical relationships, but these are not uniformly or predictably distributed across its sources. Though we found non-hierarchical UMLS relations to be useful in mapping among "domain homogeneous" ontologies, they proved to be wholly insufficient for reliably establishing relationships among heterogeneous ontologies such as ICD-9 and CPT; there are no direct relations of any kind between ICD-9 and CPT, and only a limited number of indirect relations between AUIs from the two sources sharing the same CUI.

Recalling that our primary purpose in relating procedure to condition codes is to provide a knowledgeable user with suggestions for meaningful associations, we have decided on a more empirically oriented approach to discovering such relations. Our current research direction is to investigate the use of very large observational data bases of patient medical histories in this regard. By using data mining or machine learning techniques on such information sources we should be able to infer common relations among procedures and conditions. These empirically inferred relations may then be used in our application to present the user with a set of meaningful procedure/condition suggestions from which to choose.

Methodology in a Practical Context

We have described our general approach to using UMLS for the purpose of "mapping" among biomedical coding schemes in the domains of medical conditions, signs, symptoms, and procedures. While our methodology is of some interest in the areas of knowledge representation and knowledge engineering, our primary goal is to use this methodology to achieve beneficial practical results in the areas of drug discovery and drug safety. Accordingly, our methods are being used as the basis of a highly-interactive user-oriented application called *CodeSlinger* that will enable epidemiologists to explore relations among medical condition and procedure concepts, and to develop and manage "sets" of such concepts and their related codes. *CodeSlinger* itself will be the subject of future publications.

[3] By 'MTHCH' we mean the *Metathesaurus Hierarchical CPT Terms*, which is maintained by the National Library of Medicine.