Columbia University
Mailman School of Public Health

Automated ICF Coding Using Medical Language Processing

Project Summary

In order for the International Classification of Functioning, Disability and Health (ICF) to be widely adopted, it is critical that we explore methods to facilitate the inclusion of ICF information into standardized records. The proposed research addresses this task and represents a novel step to evaluate the function of the ICF for capturing and encoding data from patient records. The project evaluates the MedLEE NLP system in parsing relevant concepts, and coding in the ICF classification. Project goals are to evaluate 1) the ICF classifications' performance in coding concepts and, 2) the MedLEE NLP system in parsing relevant concepts, and coding in the ICF classification.

MedLEE, an extensible natural language and encoding system, has already been used to extract and structure information from several types of patient reports, including discharge summaries and radiology reports. It has allowed for the implementation of several automated clinical applications that depend on the availability of coded data. MedLEE's knowledge components consist of a lexicon database, grammar rules, a compositional mapping tool for multi-word phrases, and an encoding table to map clinical terms to controlled vocabularies. The system's processing components include a preprocessor, a parser, an encoder, an XML translator, and an error handler.

The initial phase of the project involved determining ICF coding issues, determining a set of frequently occurring ICF codes, and collecting patient records for training the natural language system. A rehabilitation expert provided a listing of 10 ICD-9 codes based on frequency of occurrence within rehabilitation records. There are no ICF codes available in the patient record, but these ICD-9 codes are linked to conditions that frequently require rehabilitation. These conditions also involve functional and structural impairments, and other limitations and environmental factors that are associated with ICF codes. For example, ICD-9 code 722.10 (herniated disk in back) occurs frequently, and is likely to be associated with ICF code **b280**, corresponding to **pain.** Similarly, a frequently occurring ICD-9 code of 716.9 (arthritis in knee) is likely to be associated with an ICF code of **b710**, corresponding to **mobility of joint** as well as **b280**, corresponding to **pain**.

The next phase consisted of manual coding, and analyzing coding issues. A random sample of 5 rehabilitation records and 5 discharge summaries that are associated with each of the 10 ICD-9 were selected. A rehabilitation expert then manually coded the sample reports using the ICF codes. The expert highlighted the information in the reports that corresponds to each code, and noted any coding difficulties, redundancies, or other issues or problems. The research team then reviewed the results of the manual coding to assess the completeness and adequacy of the ICF codes and associated qualifiers. This

initial evaluation was qualitative in nature. Some of the factors reviewed and analyzed included:

- Are the codes complete: determined based on the amount and types of information that were difficult to code or that could not be coded because an appropriate ICF code was not available.
- Can the same information be coded in multiple ways: analysis of the cause and frequency of coding redundancies because redundancies in codes are undesirable.
- Are the set of qualifiers complete and well defined: determined based on an analysis of qualifiers that could/could not be adequately assigned after the primary ICF code was determined.
- How direct is the association between the information in the record and the code: analysis of additional domain knowledge required to associate the ICF code and the actual language in the patient records.
- How complete is the patient record in relation to the information that is needed to adequately perform functional coding: Analysis of information that is directly in the patient record and information that can be inferred using domain knowledge.
- Does encoding of the rehabilitation record differ from encoding of the discharge summary; if the encoding does differ, what is the difference.

During the second half of year one, a training set of codes from the reports for the natural language processing system was determined. This training set was established based on the coding analysis performed in the first half of year one and is informing modifications to the NLP system's preprocessor, semantic categories, lexicon, grammar, and encoder. This is the phase of this project that is currently underway. Once the MedLEE system is trained to perform using ICF codes, we will perform a quantitative evaluation. We will measure the performance of MedLEE with respect to ICF coding; we will also measure the performance of physicians with respect to ICF coding. The reason we will evaluate physicians is to determine whether the application of ICF codes differ between physicians and rehabilitation experts. This is a possibility because physicians have a different focus than physical therapists: physicians have a disease-oriented view of patients whereas therapists have a functional orientation. The subjects in this quantitative evaluation, to be conducted in the final project phase are the computer (MedLEE), and human coders (physicians and rehabilitation experts). Two sets of measures will be obtained: one set will measure the ability of the subjects to determine the applicable function that was stated in the patient record; the second set will measure the ability of the subject to associate the appropriate qualifier to the ICF code.