A Lexical Method for Assisted Extraction and Coding of ICD-10 Diagnoses from Free Text Patient Discharge Summaries

A. Blanquet, M.D., M.Sc., P. Zweigenbaum, Ph.D.


BACKGROUND
Codi ng diagnoses, procedures or other patient information in a controlled vocabulary includes two different subtasks: (i) finding which information to code for a given patient, and (ii) determining how to code a given piece of information. The objective of this work is to assist ICD-10 diagnosis encoding by automatically detecting possible codes based on a lexical analysis of the contents of patient discharge summaries.

MATERIAL AND METHODS
In a first step, our method identifies and adds XML tags to the “most relevant” phrases of the text, based on a dictionary of the words in the target controlled vocabulary (212 entries for a 382-term haematology thesaurus, and 11,064 entries for the 10,797-term French ICD). Attached to each word form are its lemma (root form) and a feature that characterizes its role in ICD terms: general-purpose word (GEN: “the”, “of”), domain-specific word (SPE: “acute”, “aspergillosis”), main word of a term (HEAD: “acute myeloblastic leukemia”). Relevant phrases are sequences of ICD-10 words matching the pattern

\[(\text{spe}\ (\text{gen}\ j\ \text{spe})\*)? \ \text{head}\ ((\text{spe}\ j\ \text{gen})\ *\ \text{spe})\ ?\ (*\ \text{means 0-n repetition and } ?\ \text{marks optionality})\]

Then the marked-up PDS is converted into an HTML page showing an index of the relevant phrases in the original PDS (see figure). The second step collects the possible codes for these phrases by lexically matching the head words and, if possible, other words of the candidate expressions, with the code terms, taking into account some lexical variants and morphologically related words. Codes are presented to the coder in separate HTML pull-down menus according to their head word and probable relevance.

RESULTS
This method has been fully implemented. Tested on a sample of 100 French patient discharge summaries in haematology against the diagnostic codes manually assigned by physicians in their routine work (we wish to thank Dr. N Taright for help and advice and Pr. JP Vernant for providing us PDSS), it yielded a near-null silence (sensitivity = 98.2%). The first step identified an average of 8 different head words per PDS. The average number of codes presented for each head word in the second step was 4.7 with the haematology thesaurus. We believe that as long as they can fit on one screen page, as was the case for nearly all our PDSS, this noise should be acceptable.

CONCLUSIONS
This method is organized around a simple but very useful notion of head word which makes it both less constraining than a full term-matching method and at the same time potentially more focused than a statistical partial matching method of the kind used in information retrieval. Its interface design is such that the corresponding precision, although low in absolute terms, enables a quick selection of the appropriate codes. Although simple and modest in its capabilities, we believe this method has the potential for providing valuable help as a coding assistant. Its generality over other departments and specialties now needs to be assessed.

References