Creation of a Web-based ICPC encoder and a secondary codification variability Developing tools from a free text variability analysis of diagnostic descriptions records

Alejandro Lopez Osornio, Sebastian Garcia Marti, Laura Gambarte, Gustavo Staccia, Monica Schpillberg, Enrique Soriano, Paula Otero, Daniel Luna, Fernan Gonzalez Bernaldo de Quiros

Department of Medical Informatics, Hospital Italiano de Buenos Aires, Argentina

Abstract

In the management context the decision making process is based on epidemiological data from a population. This makes the data quality process of the clinical registry a critical task in modern health systems. We decided to analyze the variability in the secondary coding process of the problems originated by the assistant physicians with the purpose of measuring the quality of this process and develop an automated coding system based on the similarity of the previous text problems. We found that the variability index allows us to perform a quality analysis of a dataset in a quick and inexpensive manner.

Introduction

The most basic form of clinical registry is the manual recording of medical problems by the physician in an electronic medical record or in a specifically designed paper tool. For the posterior use of this data, these must be coded using a standardized vocabulary. This step solves the diversity problem of the medical vocabulary unifying them for a better analysis. The process by which the best code is assigned to a medical problem is known as “coding” and it has two basic modalities, one is called “primary” when the choice is made by the assistant physician and “secondary” when is made by another person without contact with the patient or the assistant physician. This second modality is more frequent.

Objective

To analyze the variability in the secondary coding process, with the purpose of measuring the quality of this process and develop an automated coding system, based on the similarity of the previous text problems..

Materials and Methods

The database used corresponds to the problem list generated during the ambulatory consults in the Hospital Italiano of Buenos Aires, which is a training university hospital in Buenos Aires, Argentina. Each problem text in the problems list represents the description of a diagnosis, symptom or procedure entered by a physician as a 50 characters string of free text.

The database contains 524,248 problems texts generated by the physicians during the medical encounters with their corresponding ICPC codes assigned by the coding team.

These 524,248 problems are contained in 104,354 groups of repeated texts; this means that the physicians wrote the texts in the same way for the descriptions of the same problems.After a normalization process we had 88,065 texts groups that contained inside all the 524,248 initial problems. In these 88,065 texts groups, 68,663 are composed by only one problem, meaning that the text inside this group is unique and was never repeated. In a modality as the one proposed by us, with the purpose of measuring the variability of the coding process, the texts appearing only once have no comparison pair and must be excluded from the analysis. The remaining 19,402 texts groups have more than one code assigned to them. Inside the groups not all the texts have the same code because of the variability inherent to the manual coding process. Our strategy is based in that there is only one correct code for the same text and the rest can be considered incorrect. We defined as correct the one assigned in more than 50% of the texts inside a group. The codes that are in a group where there is not a predominant code and those that are in a minor portion of those groups that do have a majority code are considered erroneous. Only are considered correct those codes that correspond to a text that appear more than once and belong to the majority portion of the groups that have a portion containing more than the 50% of the codes of its group.

Results

In the 19,402 text groups containing more than one text, 397,246 problems belong to a majority portion and are considered correct. Also in these groups there are 58,239 problems with a code that is different from the majority one and are considered incorrect. The variability index is the proportion that the supposedly correct codes represents from all the problems in the database excluding those texts that appear only once. In our dataset this index was 12.78% .

To apply this information to generate an automated coder we develop a thesaurus with the almost 15,000 text groups that contained a supposedly correct code, which relates the texts wrote by the physicians with the codes for each group. Also a web site was developed that allows the entrance of a text, and after applying to it the same normalization process and comparing it to the thesaurus, assigns an ICPC code automatically in the cases where a coincidence is found. Its access is free in http://www.hospitalitaliano.org.ar/encoder.php.

References

[1] Lopez Osornio A LD, Bernaldo de Quiros FG. Creación de un sistema para la codificación automática de una lista de problemas. In: SADIO, editor. 5to Simposio de Informática en Salud - 31 JAIIO; 2002; 2002 Septiembre 2002; Santa Fe, Argentina: SADIO; 2002.

[2] Letrilliart L, Viboud C, Boelle PY, Flahault A. Automatic coding of reasons for hospital referral from general medicine free-text reports. Proc AMIA Symp 2000:487-91.

[3] Chute CG, Elkin PL, Sherertz DD, Tuttle MS. Desiderata for a clinical terminology server. Proc AMIA Symp 1999:42-6.