Application of Artificial Intelligence in Tumors Sizing Classification for Breast Cancer.

R. González Otal,[1] J. L. López Guerra,[1,2] C. L. Parra Calderón, [1,3] A. Martínez García,[1] V. Suarez Gironzini,[2] J. Peinado Serrano,[2] A. Moreno Conde [1] y M. J. Ortiz Gordillo [2]

[1] Group of technological innovation, University Hospital Virgen Del Rocío, Sevilla
[2]Department of Radiation Oncology, University Hospital Virgen Del Rocío, Sevilla
[3]Department of Information Technology, University Hospital Virgen Del Rocío, Sevilla

Abstract. The staging in breast cancer is one of the most important prognostic factors. However, the complex coding TNM criteria, which includes clinical and pathological components, the existence of different versions of TNM classification guides over time, and the variability of the source used to obtain data, makes the manual collection of TNM staging in free text be variable and imprecise. The aim of this project is to develop a tool based on artificial intelligence that allows the collection of tumor size (T) staging data for breast cancer automatically, reducing the variability. Our approach, based on two steps, starts with the detection and extraction of tumor's size characteristics in free text, using a simple natural language processor. Secondly, based on the data extracted, we applied different data mining algorithms for the T classification such as the J48 classifier tree, LADtree and NaiveBayes. Then, structured TNM reports for patients are created.

1 Introduction

Despite the fact that electronic health records are currently more and more common, not all the information contained can be used in automatic process. The main reason for this is because of the use of free text structured to abstract certain values in a table related to the health records. This information is sometimes written down altogether as a free text format in fields like Pathology diagnosis, or sometimes, somewhere else. In our project, we have used a simple natural language processor (NLP) which attends to indentify tumor characteristics from the diagnosis text, using these data extracted for making an accurate tumor size (T) status's classification based on different classification trees. After this, the classification results were compared with the T classification table from an expert, and also with the T classification table from a second year resident.

2 Materials and Methods

Our study contains two main activities. The first one is the use of a NLP for identifying T characteristics from different electronic health record sources. Factors like tumor type, tumor size, and some other attributes such us: 'is Invasive' were established. On having patient's tumor characteristics, this patient is T classified using different classification trees from the java data-mining open source software: WEKA, and a static classification algorithm, not based on data-mining. Weka has been used in many medical studies and applications. The process of comparison between different classifier trees allows us to know which is the most optimal classifier as well as the most accurate compared to an expert .The objective of our study is to show that this tool can save time and efforts, reduce errors with accuracy similar to any expert within this domain. In addition, the application is capable of updating classification models, if the classification criteria have changed from previous editions. For this study we used postoperative tumor size (pT, from the 7th edition TNM). This tool was tested on patients with non-metastatic breast cancer treated with conservative surgery and adjuvant radiation therapy at our Institution from January to April 2012 (N=68).

3 Results

The rates of coincidences between the resident's classification and the: (1) static algorithm, (2) J48, (3) LADtree, and (4) NaiveBayes were 86, 83, 82, and 77%, respectively. After the expert revision, the rates of coincidences were: 96, 93, 87, and 82%, respectively. There were only 3 errors when using the static algorithm, being 5, 9 and 12 when using J48, LADtree, and NaiveBayes algorithms, respectively. The reasons for the errors in the classification by the algorithms were mainly due to the lack of recognition of multifocal status and inflammatory disease. The resident´s classification errors were mostly due to the use of the clinical T stage when the pathology revealed a complete response after neoadyuvant chemotherapy.

4 Conclusions

The static algorithm had the highest percentage of correctly classified cases, although this algorithm does not allow updates when needed. J48 algorithm achieves a high percentage of correctly classified cases and also allows changes when needed, what is really important when guidelines are updated. This innovative system based on artificial intelligence automatically enables the classification of tumor size in breast cancer. This tool would help saving time in the data collection, preventing errors and improving tumor classification as well as the quality of the therapeutic decision.