Data Mining validation model to predict future health care cost

Mónica Schpilberg[a], Vanina Taliercio[a], Daniel Vazquez Vargas[b] , Silvana Figar[a], Hernán Michelangelo[c], Daniel Luna[a], Fernán González Bernaldo de Quirós[d]

[a] *Department of Epidemiology and Statistics, Hospital Italiano de Buenos Aires, Argentina*
[b] *Faculty of Natural Sciences, University of Buenos Aires, Argentina*

Introduction

There is evidence that the use of a health care system is related to the health of its members (greater costs incurred poorer health status). An efficient manner to manage health care is through the identification of subgroups of patients with high-risk and resulting high costs, so as to optimize medical interventions. Clinical risk scores try to predict adverse medical outcomes based on clinical features and only a few scores are based on administrative data.

The Health Information System allows us to design interventions based on the knowledge of possible future scenarios about health status of the assisted population.

In our area, the Italian Hospital of Buenos Aires is a Health Maintenance Organization that currently assists 150,000 members whose information is systematized, standardized and stored in a Hospital Information system (HIS).

In conjunction with SAS Institute, we created an analytical model. In this paper we propose to validate the model created.

Materials and Methods

Retrospective cohort of Health Plan members of the HMO who have been active for at least two months of 2007 and continue assets at 31 December 2007. " Model validation was performed by comparing the predicted cases (as high-cost) applied the model to the dataset of 2007 against those who were truly spenders in 2008. We determined the sensitivity (S), the specificity (E), the number of true positives (TP), false negative (FN), positive predictive value (PPV) and negative predictive value (PNV). It expressed its confidence interval 95%. The probability of each member of incurring high costs were evaluated in the ROC curve of different breakpoints

Results

Over a total of 141,873 members in 2007, the model predicted that in 2008 would be a high cost of 1957 (1.38%) for presenting a value greater than $ 3785. Patients who were really expensive in 2008 were 11084 (7.81%).For a probability of high of 50%, the sensitivity of the model is 11% and the false negative rate of 88.5%,99% of specificity,64%PPV, 92%PNV. While the sensitivity is increased significantly without increasing the number of false negatives when the cutoff is set at a probability 0.10. Define the cutoff point is appropriate depends on the capacity for intervention in each setting. If such is considered as cutoff score to a 50% chance we will have per 18 afli well identified as high monthly spend only one will be a FP (positive LR 18.42 IC (16,83-20,16))

Discussion

Result of this study shows that data mining is a valid tool for a specific detection of patients who have high cost. Data mining applied in health offers models that allow us to recognize times on subgroups of people who speak in favour of their health. It should be noted, however, that some of these subgroups might be trivial to current knowledge. For example, if a finding of the model is a subgroup that have neoplasic disease, can hardly be any different from the classical intervention to alter the course of the disease. Fortunately, times are nodes (subsets) that bring together features that in everyday practice as you do not recognize risk and yes, these subgroups would offer the possibility of a new and usable knowledge. That is, become sources of new hypotheses to test

References

[1] Fetter RB, Shin Y, Freeman JL, Averill RF, and Thompson JD. Casemix definition by diagnosis-related groups. Med Care 1980: 18 (2 Suppl): 1-53.

[2] Zadeh LA. Is probability theory sufficient for dealing with uncertainty in AI: a negative view. In: Kanal LN and Lemmer JF, eds. Uncertainty in Artificial Intelligence. Amsterdam: Elsevier, 1986; pp. 103-16.