ABSTRACT

The main objective of this work is to develop machine learning models for the prediction of patient outcome in nephrology care as well as to validate and optimize the models with a feature selection approach. Cardiovascular events are a major cause of morbidity and mortality in hemodialysis (HD) patients and have an incidence of 20% in the first year of renal replacement therapy. Real data routinely collected during HD administration were extracted from the Fresenius Medical Care database EuCliD (39 independent variables) and used to develop a random forest predictive model to forecast cardiovascular events in the first year of HD treatment. Two feature selection methods were applied. Results of these models in an independent cohort of patients showed a significant predictive ability. The authors’ results were obtained with a random forest built on 6 variables only (AUC: 77.1% ± 2.9%; MCE: 31.6% ± 3.5%), identified by the variable importance out of bag (OOB) estimate.

Keywords: Decision Making, Feature Selection, Hemodialysis, Prediction of Cardiovascular Events, Random Forest

DOI: 10.4018/jkdb.2011100101
INTRODUCTION

During the last few decades the practice of recording electronic medical data has become routine, thus machine learning techniques are likely to play an increasing important role in clinical settings. Indeed computer-assisted analysis can be worthwhile to efficiently process the large quantity of recorded electronic clinical data and to extract useful information from it (Lavrac, 1998). Powerful techniques are needed to investigate patterns and relationships among medical variables and patient pathophysiological states. This can be especially useful in chronic diseases management where patho-physiological state of patients is steadily monitored. Appropriate management of chronic diseases aims at improving the quality of life by preventing or minimizing the effects of a disease, or chronic condition through integrative care. The care process of common chronic illnesses concerns first of all to the reduction of incidence life-threatening complications associated with the diseases. Different health professionals are involved in the process. In this scenario, the role of engineers and statisticians able to process medical data has to be reconsidered.

Physicians can be helped in the management of pathologies by the gained information, such as parameter alterations that jeopardize patient lives can be identified. Indeed, through these data processing methods, chronic diseases can be more deeply understood and a degeneration of the pathology can be more easily prevented intervening on the identified parameters. Moreover, a pattern in the data can be only identified by looking at a high amount of recordings acquired from patients at the same condition: an efficient search is only possible using appropriate data processing methods such as data mining techniques allowing to investigate large quantity of data (Rosset et al., 2010). Using a data mining approach together with the expertise and the ability of clinicians in interpreting the results, innovative and more effective treatment strategies can be devised.

The huge costs of health care, especially for the management of chronic kidney diseases, are continuously increasing. If it would be possible to identify some indicators helping to prevent sudden life-threatening events or to identify risk factors for the patients, the costs of chronic disease treatment would significantly decrease. Preventive medicine aims at identifying early signs of disease in order to improve the ability to operate before having a worsening in the pathology condition of the patient (Davis et al., 2010). Hospitalization of patients, worsening of the pathology or insurgence of life-threatening events could be prevented in this way (Savage, 2012). Prediction of events is one of the goals of machine learning. The application of machine learning techniques in preventive medicine can lead up to the identification of factors that anticipate the risky conditions which are unknown to the current clinical practice (Visweswaran et al., 2010). The patient care process and/or the pathology course could surely take advantage of that.

Chronic hemodialysis (HD) patients experience a very high mortality, which is about 20% per year. In particular chronic renal failure (CRF) was recently defined as a “vasculo-pathic state” (Ion Titapiccolo et al., 2012; Luke, 1998) since cardiovascular deaths among dialysis patients are approximately 30 times higher than in the general population. Thus the understanding of factors involved in the cardiovascular events incidence among these patients is right now a clinical target of nephrology care.

End stage renal disease (ESRD) patients need to be treated with dialysis treatment commonly three times per week to remove the excess of fluid and toxins from their body. When dialysis therapy is administered in HD clinics, a large amount of data related to the treatment and to the patient status can be collected. For this reason HD databases can pave the way for a potentially very helpful application of medical machine learning. Furthermore clinical experience may be sometimes insufficient to stratify patients according to mortality risk because of the complexity of the chronic disease.