Expert Systems with Applications 40 (2013) 4679–4686

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Artificial intelligence models to stratify cardiovascular risk in incident hemodialysis patients

Jasmine Ion Titapiccolo [a,*], Manuela Ferrario [a], Sergio Cerutti [a], Carlo Barbieri [b], Flavio Mari [b], Emanuele Gatti [b,c], Maria Gabriella Signorini [a]

[a] Politecnico di Milano, P.zza Leonardo da Vinci 32, 20133 Milano, Italy
[b] Fresenius Medical Care, E Kroenerstrasse 1, 61352 Bad Homburg, Germany
[c] Center for Biomedical Technology at the Danube, University of Krems, Donau Universität Krems, Karl Dorrek Str. 30, 3500 Krems, Austria

ARTICLE INFO

Keywords:
Random forest
Logistic regression
Lasso
Cardiovascular risk
Cardiovascular disease
Hemodialysis

ABSTRACT

End stage renal disease condition increases the risk of cardiovascular disease. The mortality rates among hemodialysis patients are 20% higher than the general population, thus in recent years the preservation of the cardiovascular system has become a major point of focus for nephrology care in patients. Cardiovascular events jeopardize the life of a dialysis patient and must therefore be prevented. The aim of this study is to develop forecast models that can predict the cardiovascular outcome of incident hemodialysis (HD) patients. Data relating to the treatment methods and the physiological condition of patients was collected during the first 18 months of renal replacement therapy and then used to predict the insurgence of cardiovascular events within a 6-month time window. Information regarding 4246 incident hemodialysis patients was collected. A Lasso logistic regression model and a random forest model were developed and used for predictive comparison. Every forecast model was tested on 20% of the data and a 5-fold cross validation approach was used to validate the random forest model. Random forest showed higher performance with AUC of the ROC curve and sensitivity higher than 70% in both the temporal windows models, proving that random forests are able to exploit non-linear patterns retrieved in the feature space. Out of bag estimates of variable importance and regression coefficients were used to gain insight into the models implemented. We found out that malnutrition and an inflammatory condition strongly influence cardiovascular outcome in incident HD patients. Indeed the most important variables in the model were blood test variables such as the total protein content, percentage value of albumin, total protein content, creatinine and C reactive protein. Age of patients and weight loss in the first six months of renal replacement therapy were also highly involved in the prediction. A greater understanding of the mechanisms involved in the insurgence of cardiovascular events in dialysis patients can ensure physicians to intervene in the appropriate manner when a high-risk cardiovascular condition is identified.

1. Introduction

Cardiovascular (CV) morbidity and mortality are very high incidence conditions in patients affected by chronic kidney disease (CKD). In fact, cardiovascular events are the leading cause of death among end stage renal disease (ESRD) patients, i.e. at the last stage of CKD when glomerular filtration is life-threatening (Wagner et al., 2011). A high number of ESRD patients die prematurely during the initial phase of renal replacement therapy (Herzog Asinger, & Berger, 2011). As a matter of fact, the mortality rate among hemodialysis (HD) patients exceeds that of the general population by 20% per year and an even higher mortality rate has been reported within the first year of initiation of HD treatment (Bradbury et al., 2007).

Since cardiovascular disease and kidney disease are closely related (Luke, 1998), cardiovascular risk factors among ESRD patients include those identified in the general population and the additional risk factors related with chronic renal failure (CRF). The risk factors include: progressive ageing of dialysis patients plus a high incidence of co-morbidities, such as diabetes, hypertension,

Abbreviations: CV, cardiovascular; CKD, chronic kidney disease; ESRD, end stage renal disease; HD, hemodialysis; CRF, chronic renal failure; RF, random forest; RRT, renal replacement therapy; TW, time window; CVE, cardiovascular event; PTH, parathyroid hormone; OOB, out of bag; VI, variable importance.

* Corresponding author. Address: p.zza Leonardo da Vinci 32, 20133 Milano, Italy. Tel.: +39 02 2399 3322; fax: +39 02 2399 3311.

E-mail addresses: jasmine.ion@mail.polimi.it (J. Ion Titapiccolo), manuela.ferrario@biomed.polimi.it (M. Ferrario), sergio.cerutti@polimi.it (S. Cerutti), carlo.barbieri@fmc-ag.com (C. Barbieri), flavio.mari@fmc-ag.com (F. Mari), emanuele.gatti@fmc-ag.com (E. Gatti), mariagabriella.signorini@polimi.it (M.G. Signorini).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.02.005

congestive heart failure, multiple organ failure and metabolic derangement due to renal injury. Moreover the risk of cardiovascular disease appearance is further increased by hemodynamic and metabolic risk factors, including hemodynamic overload due to plasma volume expansion, blood pressure overload, the presence of arterial-venous fistulae, anemia, hyperparathyroidism, electrolyte imbalance and increased oxidant factors (Locatelli et al., 2000). Consequently, in recent years, preservation of the cardiovascular system has become a major focal point in nephrology care as sudden cardiovascular events jeopardize the life of ESRD patients and must be prevented.

Over the past few decades special attention has been paid to the diagnosis and management of CV complications in CKD patients: with this aim personalized, optimized strategies for prevention have been proposed. However, the basic guidelines are limited and clinical intervention is often administered too late, thus proving ineffective in preventing life-threatening events in HD patients. Both the early signs of a CV event and the prognostic factors need further investigation in order to better understand the mechanisms underlying cardiovascular system impairment in CKD patients and prevent or reduce its occurrence. Understanding which treatment- and physiological- related variables are most involved in such phenomenon is also essential. That way more effective and focused treatment strategies can be devised to decrease the risk of cardiovascular disease in CKD patients.

Large quantities of data can be collected during renal replacement therapy administration: hemodialysis patients are treated three times per week in clinics and the data relating to their physiological condition, in addition to the treatment parameters, can be recorded. In fact, clinicians can easily access data from the dialysis machines and the monitoring devices (such as blood volume monitors, on-line clearance monitors and flow measurement equipment) and said data can be automatically stored in a clinical database.

Data mining methods have shown their helpfulness in discovering hidden information and patterns among recorded clinical data (Rosset, Perlich, Swirszcz, Melville, & Liu, 2010; Savage, 2012). By managing huge quantity of clinical data, the cardiovascular risk could be estimated more reliably in order to stratify patients and foster early intervention, which could prevent the occurrence of sudden life-threatening events (Chang, Wang, & Jiang, 2011; Eom, Kim, & Zhang, 2011).

The complexity of the underlying phenomena and the presence of strong non-linear relationships among the variables involved and the cardiovascular outcome suggest the use of a non-linear machine learning method, such as a random forest model (Sut & Simsek, 2011). Random forests were selected because they are able to identify non-linear patterns in the data and they are supposed to improve the predictive capability of commonly used linear methods.

The aim of this study is therefore to develop a random forest (RF) model for the short-term prediction of cardiovascular events in incident HD patients and verify its improvement with respect to a standard linear model. For this reason the RF model is compared with a standard linear model, the logistic regression.

2. Materials and methods

2.1. Study design

This is a retrospective study of incident hemodialysis patients, i.e. patients who started HD treatment for the first time. The HD treatments were performed in Fresenius Medical Care clinics. Data used in this study has been extracted from EuCliD database (Marcelli et al., 2001; Steil et al., 2004), an electronic database designed by Fresenius Medical Care to monitor the key parameters of dialysis treatment. In our study the data comes from patients in clinics located in Portugal and Spain. In particular, the data collected in Spain came from patients starting HD treatment between 2006 and 2009; the data collected in Portugal came from patients starting the treatment between 2006 and 2010.

The analysis of incident HD patients is of interest due to the high prevalence of cardiovascular disease occurrence and cardiovascular events reported during the beginning period of renal replacement therapy (RRT). Much literature suggests that the key processes leading to the impairment of the cardiovascular system need to be investigated during the first period of HD administration (Bradbury et al., 2007; Herzog et al., 2011).

Incident HD patients were followed for 18 months after the beginning of RRT in order to predict the occurrence of life-threatening events. The analyzed period of 18 months was split into 6-month intervals, i.e. into three successive time windows (TW): TW1 = months 1–6, TW2 = months 7–12, TW3 = months 13–18. Two different models were constructed: (1) the prediction model of events in TW2 given the data of TW1; (2) the prediction model of events in TW3 given the data of TW2.

Exclusion criteria were: age below 18 years, death during the observational period of unknown or non-cardiovascular origin, occurrence of renal transplantation, missing of data from HD treatment.

Patients were subdivided into two groups for binary classification: patients who experienced a cardiovascular events (CVEs) in the successive 6 months after the current observational time window (CVE group) and patients who did not experience any significant event in the successive 6 months (control group). The following conditions determined an event classification as cardiovascular: cardiovascular mortality, insurgence of new cardiovascular co-morbidity, or cardiovascular hospitalization.

The ICD-10 international coding system, adopted by the EuCliD system, enables identification of a CVE. All the diseases of the circulatory system – besides cerebrovascular, vein and lymphatic vessel diseases – were considered as preventable CV diseases.

Fig. 1 summarizes the details about patients used to construct the models as well as the details of the patients who dropped out of the project.

2.2. Variables and feature extraction

Hemodialysis patients are commonly treated three times per week on HD treatment. In Fresenius Medical Care clinics, at each HD session, physiological variables and treatment settings are stored in the EuCliD database. Moreover blood tests and medical examinations are generally performed once per month, to monitor the health status of patients: said records are also stored in the EuCliD database.

Our analysis selected, for each patient, the following set of variables extracted from the EuCliD database: the time series of the physiological variables measured at each HD session (blood pressure, heart rate, body weight - measured before and after the treatment); the time series of blood test variables (urea, potassium, sodium, calcium phosphate, phosphate, parathyroid hormone (PTH), calcium, haematocrit, haemoglobin, total proteins, albumin, creatinine, C-reactive protein); the dialysate concentrations (sodium, bicarbonate); the dialysate temperature; the dialyzer blood flow and total fluid lost at each HD session; the dialysis modality (hemodialysis or hemodiafiltration); the patient age and comorbidities at RRT starting point (diabetes, angina pectoris, heart disease, peripheral vascular disease); and the diagnosed comorbidities during the first 18 months of RRT (diabetes, angina pectoris, heart disease, peripheral vascular disease).