34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

Predictive modeling of cardiovascular complications in incident hemodialysis patients

J. Ion Titapiccolo, *Student Member, IEEE*, M. Ferrario, *Member, IEEE*, C. Barbieri, D. Marcelli, F. Mari, E. Gatti, S. Cerutti, *Fellow, IEEE*, P. Smyth, *Member, IEEE*, M. G. Signorini, *Member, IEEE*

Abstract— **The administration of hemodialysis (HD) treatment leads to the continuous collection of a vast quantity of medical data. Many variables related to the patient health status, to the treatment, and to dialyzer settings can be recorded and stored at each treatment session. In this study a dataset of 42 variables and 1526 patients extracted from the Fresenius Medical Care database EuCliD was used to develop and apply a random forest predictive model for the prediction of cardiovascular events in the first year of HD treatment. A ridge-lasso logistic regression algorithm was then applied to the subset of variables mostly involved in the prediction model to get insights in the mechanisms underlying the incidence of cardiovascular complications in this high risk population of patients.**

I. Introduction

Technological advancements in the form of computer-based patient record software and personal computer hardware are making the collection of and access to health care data more manageable. However, few tools exist to evaluate and analyze clinical data after they have been collected and stored. Very large quantities of data are generated through the health care process. Evaluation of stored data can lead to the discovery of trends and patterns hidden within the data that could significantly enhance our understanding of disease progression and management [1][2]. Techniques are needed to search large quantities of clinical data for these patterns and relationships.

End stage renal disease (ESRD) patients need to be treated with dialysis treatment at least three times per week. The risk of death in patients with ESRD is high despite advances in the dialysis care. Cardiovascular deaths occurring among dialysis patients are approximately 30 times higher than in the general population [3], so chronic renal failure (CRF) has recently been defined as a 'vasculopathic state' [4]. The understanding and proper management of the determinants of cardiovascular disease have therefore become a major focus of nephrology care. The pathogenesis of cardiovascular damage in CRF patients is far more complex than in the general population, since the risk factors include those identified in the general population and additional risk factors typical of CRF. The epidemiological picture of the actual end-stage renal disease (ESRD) population shows a patient population with a growing proportion of "elderly individuals" and a high incidence of co-morbidities (diabetes, hypertension, congestive heart failure, multiple organ failure...). This is mainly due to the general increase in the number of patients admitted to renal replacement therapy (RRT) [5]. Risk factors in the general population include smoking, hypertension, diabetes mellitus, physical inactivity, inflammation-related factors. Their prevalence in the ESRD population is high because of the progressive aging of dialysis patients, the metabolic derangement caused by renal failure and because of the aetiology of the underlying renal disease. In addition, hemodynamic and metabolic risk factors, peculiar to CRF, further enhance the cardiovascular risk. These include hemodynamic overload due to plasma volume expansion and arterio-venous fistula, anaemia, hyperparathyroidism, electrolyte imbalance and increased oxidant factors.

When dialysis therapy is administered in hemodialysis (HD) clinics a large amount of treatment and patient data can be collected. Thus, HD databases represent a potentially very promising application of medical machine learning. Several methods have been developed to assess the burden of comorbidity conditions and to predict outcomes in dialysis patients, reflecting the increased risk compared with the general population [6]. Despite these efforts no conclusive results have been obtained in the prediction of patient condition course. This is mainly due to the fact that complex phenomena are involved in the pathophysiology condition of HD patients and in HD treatment outcome.

The high prevalence of cardio-vascular diseases (CVD) at the start of RRT suggests that the mechanisms leading to cardiovascular impairment have been operating early in the pre-dialysis phase of chronic renal disease. Cardiovascular state at the beginning of RRT strongly influences patients' outcome and it needs to be taken into account in the estimate of hemodialysis cardiovascular risk [7].

The aim of the present study is the development of "machine learning" methods to stratify incident HD patients, i.e. ESRD patient starting HD treatment for the first time in their life, with respect to the risk of cardiovascular and life-

*Research supported by Fresenius Medical Care Deutschland GmbH.

J. Ion Titapiccolo, M. Ferrario, M. G. Signorini and S. Cerutti are with Politecnico di Milano, Department of Bioengineering, P.zza Leonardo da Vinci 32, 20133 Milano, Italy (e-mail: jasmine.ion@mail.polimi.it, manuela.ferrario@biomed.polimi.it, mariagabriella.signorini@polimi.it, sergio.cerutti@polimi.it).

C. Barbieri, D. Marcelli, F. Mari and E. Gatti are with Fresenius Medical Care, E Kroenerstrasse 1, 61352 Bad Homburg.

P. Smyth is with University of California, Irvine, CA 92697-3435, Department of Computer Science, Center for Machine Learning and Intelligent Systems.

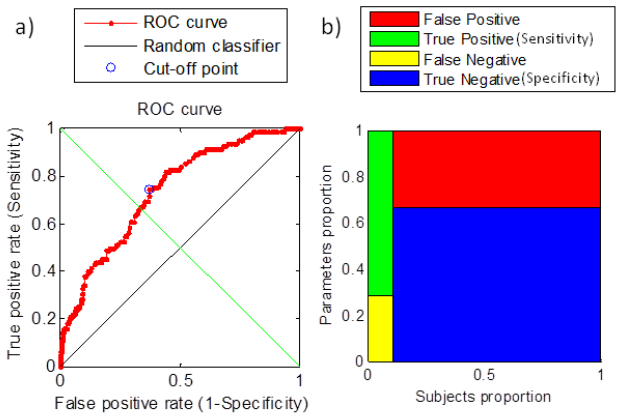

Fig. 1. a) ROC curve obtained on the test data with the random forest model. AUC value is 74.2%. b) Proportional representation of positive (yellow and green areas) and negative (blue and red areas) cases and their classification at best cut-off point: sensitivity is 71.6% and specificity is 66.7%. This is an intuitive representation of the confusion matrix.

threatening events. This work aims at assessing the impact of physiological and treatment variables trends and values on short-term mortality (1 year of follow-up) and/or incidence of cardiovascular complications.

II. Materials and Methods

A. Database description

Data used in the present study have been extracted from EuCliD, a clinical database designed by Fresenius Medical Care to monitor the key aspects of haemodialysis treatment. In particular EuCliD is a potentially useful database because it contains all the hemodialysis treatment-related information of patients treated in Fresenius Medical Care clinics. Specifically, this is a follow-up study of incident dialysis patients treated in Fresenius Medical Care clinics in Spain. Data about patients, data about each HD session, and monthly blood tests collected during the first year of hemodialysis treatment, have been extracted. Time series of treatment and blood test variables were obtained. The first challenge of the present work was to transform the temporal series database into a "static" database. The chosen approach was to extract features such as mean values and trends from the first 6 months of temporal series to predict patient course in terms of cardiovascular events (classified as "diseases of the circulatory system" in the ICD-10 coding, excluding cerebrovascular and veins and lymphatic vessels diseases) in the following six months. 166 patients had cardiovascular events during the time period of interest. 1360 patients had neither events nor kidney transplantation during the first year of HD and were included in the control group.

B. Outcomes and variables of interest

The primary outcome of prediction is the occurrence of one or more of the following cardiovascular events: a) all cause mortality, b) the incidence of cardiovascular comorbidity, c) cardiovascular hospitalization during the second semester of hemodialysis treatment. Thus the primary interest was to predict the occurrence of cardiovascular events in the next months, i.e. a binary classification. Treatment variables included in the prediction model were: mean values of blood pressure and heart rate measurement before and after each HD session (HD pre and HD post), weight loss trend during the first three months, mean value of fluid removal at each HD session, HD modality (hemodialysis or hemodiafiltration, HDF), dialyzer blood flow mean value (values in the first two month were not considered due to adjustments in the treatment strategy), number of hypotension events in the first 6 months of treatment. Mean values of some blood test variables concentration. The age of patients at HD initiation date and categorical variables about comorbidities such as diabetes, heart disease, angina, peripheral vascular disease, non-mortal cardiovascular events in the first semester of treatment were also included. Overall this resulted in a data set of 42 variables. The complete list of variables is shown in Tab.1. In the obtained dataset there were no missing values for the treatment variables and a missing value percentage between 5% and 15% for blood test variables. Missing values were substituted in the data set by the mean value of the corresponding variable.

C. Machine learning methods

A random forest is a classifier consisting of a collection of tree-structured classifiers. Each tree casts a unit vote for the most popular class at each input [8]. A forest of trees is very difficult to interpret in terms of possible underlying mechanisms. In some applications, analysis of medical experiments for example, it is critical to understand the interaction of variables that is providing the predictive accuracy. A start on this problem is made by using internal out-of-bag estimates, and verification by reruns using only selected variables.

Suppose there are M input variables. M additional random forests were constructed and in each forest the values of the m^{th} variable are not included in the database. The out-of-bag data are run down the corresponding tree and the misclassification rate given on testing set is saved. This is repeated for m = 1, 2, ... , M and the M misclassification rates are compared with misclassification rate obtained using all the variables obtaining a measure of the importance of each variable in the model.

In this study, a random forest composed by 500 trees was implemented and Gini impurity index criterion was used to split the nodes. Then scores of variables importance were obtained and those with score higher than 5% were selected. These variables were used to run a lasso logistic regression to better understand their correlation with the output labels. The lasso-ridge algorithm is a popular statistical method for regression that uses a penalty term to achieve a sparse solution: only variables significantly involved in the regression model have non-null coefficients [9]. 8 fold cross-

validation was used to select the best penalization term value for the regression model.

III. Results

The random forest model was applied to the dataset randomly selecting half of the patients to be included in the training set and using the remaining part as testing set. A bootstrapping approach was used on the training set to improve the sensitivity of the model. Cardiovascular patients were randomly oversampled with replacement obtaining balanced classes.

For each patient in the database the mean value of the prediction of all the trees in the random forest was computed to estimate the outcome probability. The obtained values are not the actual probabilities because of bootstrapping but can be considered un-calibrated probabilities. The model was trained and tested and the ROC curve for testing was computed: an area under curve value of 74.2% was obtained (Fig. 1). A cut-off point for best sensitivity and specificity gave a misclassification rate equal to 32.8%, sensitivity 71.6% and specificity 66.7%.

Only the most influencing variables in the model were standardized to have zero mean and unit variance and the lasso-ridge logistic regression algorithm was applied to deeply understand the underlying mechanisms involved in the cardiovascular outcome. It is indeed important to understand if values of the variables higher or lower than the mean lead to an increase or to a decrease in the outcome probability. The same training and testing set were used. Testing was done to assess the accuracy of the model: an area under the ROC significantly higher than 50% was obtained (66.3%). Measures of variable importance in the random forest model and coefficient values obtained with logistic regression are reported in Table 1. Variables in the table are ordered by the "importance scores" in the random forest model.

IV. Discussion

The choice of a random forest approach for the prediction of short-term cardiovascular events in incident hemodialysis patients is due to its simplicity, accuracy and relatively robustness to outliers and noise. Moreover it gives useful internal estimates of error, strength, correlation and variable importance. Looking at the ROC obtained with such a model (Fig. 1) it can be noted that the initial slope of the curve is high. Thus it suggests that the model is able to classify to high precision patients belonging to the cardiovascular events group. Patients receiving high probability values are mostly those that really have a cardiovascular event in the next few months. Indeed the model has a good sensitivity, higher than 70%, despite the low number of patients in the cardiovascular event group.

It is of physiological interest to investigate the variables, which mostly affect the probability vector i.e. the outcome of the random forest. Random forest models can capture linear and also non-linear dependencies between variables and

TABLE I
VARIABLE RANDOM FOREST SCORES AND LASSO LOGISTIC REGRESSION COEFFICIENT VALUES

Variable	*Random Forest Score (%)*	*Coeff. Value*	*Variable*	*Random Forest Score(%)*	*Coeff. Value*
Mean sodium dialysate concentration (mEq/l)	12.423	0.224	*Mean Pulse pressure post HD (mmHg)*	5.361	0.293
Angina	11.842	0.414	*Number of hypotension events*	5.177	0.172
Mean C-reactive protein - blood test param (mg/dl)	10.955	0.284	*Presence of non mortal cardiovascular events*	4.955	-
Mean Diastolic pressure (post HD) (mmHg)	10.369	-0.060	*Heart disease*	4.872	-
Mean calcium - blood test param (mg/dl)	10.238	0.128	*Mean PTH value - blood test param (ng/l)*	4.702	-
Mean potassium - blood test param (mEq/l)	9.976	0	*Mean Diastolic pressure (pre HD) (mmHg)*	4.474	-
Mean bicarbonate dialysate concentration (mEq/l)	9.784	-0.011	*Diabetes*	4.210	-
Mean Delta pulse pressure (HD post - HD pre) (mmHg)	8.272	0	*Modality (0=HDF, 1=HD)*	4.209	-
Mean calcium phosphate - blood test param (mg/dl)	7.617	-0.063	*Mean total fluid lost per HD session (ml)*	3.683	-
Mean Systolic pressure (post HD) (mmHg)	7.529	0	*Weight percentage loss in six months (%)*	3.640	-
Mean haemoglobin - blood test param (g/dl)	7.494	-0.025	*Mean Systolic pressure (pre HD) (mmHg)*	3.410	-
Mean dializer blood flow (ml/min)	7.250	0.170	*Mean Delta diastolic(HD post-HD pre (mmHg)*	3.198	-
Mean Delta systolic (HD post - HD pre)	6.451	0.069	*Mean creatinine (pre HD) - blood test param (mg/dl)*	2.505	-
Mean Delta weight (HD post - HD pre)	6.375	0.060	*Mean haematocrit - blood test param (%)*	0.950	-
Peripheral vascular disease	6.353	0.064	*Mean sodium - blood test param (mEq/l)*	0.605	-
Mean totalprotein content - blood test param (g/dl)	6.306	0	*Mean albumin percentage-blood test param(%)*	0.054	-
Mean dialysis urea (pre HD) - blood test param (mEq/l)	6.260	-0.234	*Mean heart rate (post HD (bpm)*	-0.241	-
Mean Pulse pressure (pre HD) (mmHg)	6.091	0	*Dialysate temperature (°C)*	-1.195	-
Mean Delta heart rate (HD post - HD pre) (bmp)	6.006	0.017	*Mean phosphate - blood test param (mg/dl)*	-2.308	-
Mean dialysis urea (post HD) - blood test param (mEq/l)	5.913	0.264	*Age (years)*	-2.948	-
Mean albumin content - blood test param (g/dl)	5.367	-0.225	*Mean heart rate (pre HD) (bpm)*	-3.976	-

output. The logistic regression model was used as a starting point to get insights into the major linear relationships between selected variables and the output. Variables improving the random forest classification rate more than 5% were selected.

The measures of variable importance (the *score %* in Table 1) showed that variables more involved in the classification process included:

- sodium concentration in the dialysate;
- presence of angina and peripheral vascular disease;
- C-reactive protein;
- pulse pressure, systolic and diastolic pressure values measured after the treatment;
- albumin, haemoglobin and urea;
- blood flow in the dialyzer.

Looking at both the variable importance score and the corresponding logistic regression coefficient can help to understand the potentiality of the used predictive modeling techniques.

Sodium dialysate concentration, the variable with the highest importance score, has a positive regression coefficient. It means that, in the logistic model, positive normalized values of the variable lead to an increase in the probability of belonging to the cardiovascular event group. This is in agreement with [10]: clinics which predominantly use a dialysate sodium of 140 mmol/l (instead of 136 mmol/l) have increased inter-dialytic weight gains, with more difficult blood pressure control, and a greater percentage of patients requiring anti-hypertensive medication. This can lead to an increased cardiovascular risk. Moreover, the number of hypotension events was found to increase the risk of cardiovascular events as already shown [11]. In fact patients with cardiovascular problems (i.e. heart insufficiency) are more prone to intradialytic hypotension. Then, patients prone to hypotension are more likely to have prescribed a higher dialysate sodium concentration because it improves treatment tolerance to ultrafiltration.

Furthermore, as expected, the presence of angina and peripheral vascular disease were found to be cardiovascular risk factors, having high importance score and positive coefficient values. C-reactive protein increases in response to inflammation status. The positive regression coefficient confirms the relationship between an inflammation status and an augmented mortality risk [12]. Furthermore patients with low albumin levels, low haemoglobin, low levels of predialysis urea and high values of post dialysis urea were found to be more subject to cardiovascular events. Indeed, nutritional status and interdialytic weight gain are important mortality risk factors for HD patients [13]. Furthermore, blood pressure measurements before and after the treatment were found to be in some way predictive of cardiovascular events. In particular, higher values of post HD pulse pressure (difference between systolic and diastolic pressure), lower values of post HD diastolic, and higher values of post HD systolic pressure (measured after the treatment), all increase the probability of cardiovascular events. This is in accordance with new findings on the importance of pulse pressure on the preservation of the cardiovascular system [14]. Since this is a multivariate analysis, the increased risk of a cardiovascular event is dependent on a combination of the values of all these different variables.

V. Conclusion

A random forest model was developed and applied to a database of incident hemodialysis patients to predict the incidence of cardiovascular events in the first year of treatment. To gain insights into the model the most important variables involved in the prediction were selected.

Logistic regression applied to these variables enabled to interpret the results from a clinical and physiological point of view. The application of machine learning models to larger HD datasets will permit to understand the mechanisms underlying cardiovascular events and to predict more accurately these events.

References

[1] N. Lavrac, "Selected techniques for data mining in medicine," *Artificial Intelligence in Medicine*, 16: 3-23, 1998.

[2] S. Rosset, C.Perlich, G. Swirszcz, P. Melville, Y. Liu, "Medical data mining: insights from winning two competitions," Data Min Knowl Disc 20: 439-469, 2010.

[3] F. Locatelli, D. Marcelli, F. Conte, M. D'Amico, L. Del Vecchio, A. Limido, F. Malberti and D. Spotti "Cardiovascular disease in chronic renal failure: the challenge continues" *Nephrol Dial Transplant* , vol. 15 [Suppl 5]: pp. 69–80, 2000.

[4] R.G. Luke, "Chronic renal failure - a vasculopathic state", *New Engl J Med* vol. 339, pp. 841–843, 1998.

[5] US Renal Data System. 1999 Annual Data Report. *Am J Kideny Dis*, 34 [Suppl 1], 1999.

[6] J M. Wagner, D. Ansell, D. M. Kent, J. L. Griffith, D. Naimark, C. Wanner and N. Tangri, "Predicting mortality in incident dialysis patients: an analysis of the United Kingdom renal registry," *Am J Kidney Dis*, vol. 57(6), pp. 894-902, 2011.

[7] P.S. Pafrey, R.N. Foley, J.D. Harnett, J.M. Kent, D.C. Murray, P.E. Barre, "The outcome and risk factors for left ventricular disorders in chronic emiauraemia," *Nephrol Dial Trasplant*, 11: 1277-1285, 1996.

[8] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[9] J. Friedman, T. Hastie, R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33(1), pp. 1-22, 2010.

[10] A. Davenport, "Audit of the effect of dialysate sodium concentration on inter-dialytic weight gains and blood pressure control in chronic haemodialysis patients," *Nephron Clin Pract*, vol. 104, pp. c120-c125, 2006.

[11] T. Shoji, Y. Tsubakihara, M. Fujii, E. Imai, "Hemodialysis-associated hypotension as an independent risk factor for two-year mortality in hemodialysis patients," *Kidney International*, 66:1212–1220, 2004.

[12] E. Ritz, "Intestinal-renal syndrome: mirage or reality?" *Blood Purification*, 31:70–76, 2011.

[13] J.M. Lopez-Gomez, M. Villaverde, R. Jofre, P. Rodriguez-Benitez, R. Perez-Garcia, "Interdialytic weight gain as a marker of blood pressure nutritional and survival in hemodialysis patients," *Kidney International*, 67(93): S63-S68, 2005.

[14] J.K. Inrig, U.D. Patel, R.D. Toto, D.N. Reddan, J. Himmelfarb, R.M. Lindsay, J. Stivelman, J.F. Winchester and L.A. Szczech "Decreased pulse pressure during hemodialysis is associated with improved 6-month outcomes" *Kidney Int*, 76: 1098–107, 2009.