Clinical Kidney Journal, 2021, vol. 14, no. 5, 1388–1395

doi: 10.1093/ckj/sfaa126
Advance Access Publication Date: 11 August 2020
Original Article

ORIGINAL ARTICLE

Predicting mortality in hemodialysis patients using machine learning analysis

Victoria Garcia-Montemayor[1], Alejandro Martin-Malo[1,2,3], Carlo Barbieri[4], Francesco Bellocchio[4], Sagrario Soriano[1], Victoria Pendon-Ruiz de Mier[1,2], Ignacio R. Molina[1], Pedro Aljama[1,2] and Mariano Rodriguez[1,2,3]

[1]Department of Nephrology, Reina Sofia University Hospital, Cordoba, Spain, [2]Maimonides Biomedical Research Institute of Cordoba (IMIBIC), Reina Sofia University Hospital, University of Cordoba, Spain, [3]RETICs-REDinREN (National Institute of Health Carlos III), Madrid, Spain and [4]Fresenius Medical Care Italia, Vaiano Cremasco, Cremona, Italy

Correspondence to: Dr. Alejandro Martin-Malo; E-mail: amartinma@senefro.org

ABSTRACT

Background. Besides the classic logistic regression analysis, non-parametric methods based on machine learning techniques such as random forest are presently used to generate predictive models. The aim of this study was to evaluate random forest mortality prediction models in haemodialysis patients.

Methods. Data were acquired from incident haemodialysis patients between 1995 and 2015. Prediction of mortality at 6 months, 1 year and 2 years of haemodialysis was calculated using random forest and the accuracy was compared with logistic regression. Baseline data were constructed with the information obtained during the initial period of regular haemodialysis. Aiming to increase accuracy concerning baseline information of each patient, the period of time used to collect data was set at 30, 60 and 90 days after the first haemodialysis session.

Results. There were 1571 incident haemodialysis patients included. The mean age was 62.3 years and the average Charlson comorbidity index was 5.99. The mortality prediction models obtained by random forest appear to be adequate in terms of accuracy [area under the curve (AUC) 0.68–0.73] and superior to logistic regression models (ΔAUC 0.007–0.046). Results indicate that both random forest and logistic regression develop mortality prediction models using different variables.

Conclusions. Random forest is an adequate method, and superior to logistic regression, to generate mortality prediction models in haemodialysis patients.

Keywords: haemodialysis, machine learning, mortality, predictive models, random forest

Received: 25.4.2020; Editorial decision: 18.5.2020

INTRODUCTION

The number of haemodialysis (HD) patients increases progressively in those >75 years of age [1]. It is a health challenge with high economic cost and mortality. Risks factors of mortality in HD patients are very different from other populations. It is important to stratify HD patients according to risk and develop, as early as possible, appropriate strategies aimed at optimizing survival.

Classic statistical analysis has identified variables that predict mortality in HD patients and the results have been rather uniform among the different publications [2–11]. Risk prediction models (or 'risk scores') are designed to predict the probability of an adverse outcome, such as death, from different types of variables—demographic, clinical and others. Classic methods of survival analysis, such as Cox proportional hazards regression and logistic regression, rely on the assumption that the relationship between variables and outcomes is linear. This assumption is very useful to generate simple and intelligible models in which the numerical value of a coefficient represents the contribution of that variable to the overall risk [12]. More recently, machine learning methods [13], such as random forest, have been proposed as more advanced valid procedures to predict outcome if there is enough data available to perform the analysis. These new methods of analysis may identify variables, not previously recognized, that can improve prediction of mortality. Studies in patients have proven the usefulness of random forest regression models in identifying variables with high predictive power [14–16], to estimating individualized treatment effects, by investigating the performance of random forest of interaction trees via extensive numerical experiments [17]. Analyses based on random forest have been used to quantify the association between parameters of chronic kidney disease–mineral and bone disorder (CKD-MBD) in HD patients [18]. Similar methods have been used to estimate the individual effect of a treatment, based on observational data [19]. It has been reported that Cox regression was inferior to random forest in developing prognostic models of lung adenocarcinoma [20].

The objective of the present study was to compare mortality prediction models in HD patients obtained by conventional logistic regression analysis and by random forest. The accuracy of the models was obtained by comparing the prediction obtained with each method with the actual mortality.

MATERIALS AND METHODS

Patients

This study analyses data collected by the Nephrology Department Database (Reina Sofia University Hospital) from 1995 to 2015. There were 2219 patients from seven HD facilities and 1571 patients fulfilled the criteria to be included in the study. Since 1995, all patients starting maintenance HD have been informed that their identity will not be revealed to third parties. All patients included in this study were >18 years of age and signed an informed consent form allowing the use of clinical records and laboratory data for analysis aimed at improving clinical practice, providing that their identity will not be disclosed. The study was performed in accordance with relevant guidelines and regulations and was submitted and subsequently approved by the Institutional Ethics Committee of the Reina Sofia University Hospital.

Variables collected at the initiation of regular HD included age and comorbidities that generate the Charlson comorbidity index, including myocardial infarction, coronary heart disease, heart failure, stroke, peripheral vascular disease, dementia, chronic obstructive pulmonary disease (COPD), autoimmune disease, peptic ulcer, non-cirrhotic liver disease, liver cirrhosis, severe kidney disease, non-metastatic tumour disease, solid metastatic tumour, malignant haematological disease (leukaemia or lymphoma), acquired immune deficiency syndrome and diabetes mellitus. The analytical variables available during follow-up were haemoglobin, ferritin, transferrin saturation index (TSI), creatinine, albumin, C-reactive protein, phosphorus, calcium, potassium, alkaline phosphatase, magnesium, parathyroid hormone (PTH) and β2-microglobulin. Additional parameters evaluated were body mass index (BMI), residual diuresis, type of vascular access for HD and dose of dialysis expressed as Kt/V_{urea}.

Baseline data were constructed with the information obtained during the initial period of regular HD. It is important to know if the length of time collecting the data influences the results. Therefore, to be more accurate, the initial period of time used to collect data was set at 30, 60 and 90 days after the first HD session. Patients that did not survive the first 30 days on regular HD were not included. Likewise, patients with no data collected during the first 90 days after the first regular HD session were not included in the analysis. Values of variables were the mean of measurements obtained within the first 30, 60 or 90 days after the initiation of regular HD (this will be indicated as data at 30, 60 and 90 days). The missing values were handled with a single imputation approach considering the mean for continuous variables and the mode for categorical variables.

A flow chart of the patients analysed in this study is shown in Figure 1. Prediction of mortality was analysed at 6 months, 1 year and 2 years after the initiation of regular HD. Analysis was performed separately according to the time period available for the collection of baseline data (30, 60 and 90 days).

Prediction of mortality was calculated using two different regression models: logistic regression and random forest. Receiver operating characteristics (ROC) curves specify the sensitivity and specificity of these predictive models. The values of the area under the curve (AUC) are used to compare the predictive value obtained by logistic regression and random forest [21]. A *t*-test was performed to determine if the differences between the mean AUCs obtained by logistic regression and random forest were statistically significant.

Since the data set is relatively small, each AUC is computed considering a set of models configured on 30 different randomizations of the original data set. In each randomization, 70% of the patients were used to train the model (training set) and 30% of the patients were used to evaluate the accuracy (test set). The final accuracy was computed as the AUC, considering all the predictions on the test sets of the 30 randomizations.

Logistic regression analysis identifies independent variables that significantly influence the risk of death. The analysis provides a coefficient for each variable [the odds ratio (OR)] that represents the probability of an increase or decrease of death if the variable is modified by one unit (e.g. if the OR of age is 1.1, there is a 10% increase in the expected risk relative to a 1-year increase in age). In random forest analysis [15], the probability of death is computed as the average of the probabilities of a set of submodels (decision tree). Each submodel is a composition of if-then-else decision rules, which are derived considering a random subset of the data. In contrast to logistic regression, in random forest, there is not a linear coefficient for each variable; it is assumed that the effect of the variable could be much more complex. In random forest, the predictive value of a variable on

Table 1. Baseline comorbidities and biochemistry obtained during the first 30 days of dialysis

Baseline	Characteristics
Gender (male/female), *n* (%)	953 (61)/618 (39)
Age (years), mean ± standard deviation (SD)	62.33 ± 15.89
Comorbidities, *n* (%)	
Diabetes mellitus	482 (31)
Cardiac failure	319 (20)
COPD	144 (9)
Tumoral disease (non-metastatic)	131 (8)
Myocardial infarction	102 (6)
Hepatopathy (non-cirrhotic)	68 (4)
Stroke	9 (1)
Charlson comorbidity index (mean)	6
Biochemical parameters, mean ± SD	
Haemoglobin (g/dL)	10.08 ± 2.79
Ferritin (ng/mL)	290.1 ± 362.64
TSI (%)	18.73 ± 10.32
Creatinine (mg/dL)	7.3 ± 4.4
Albumin (g/dL)	3.54 ± 0.55
CRP (median) (mg/L)	8.8 (IQR: 19.5)
Calcium (mg/dL)	9.04 ± 3.88
Phosphorous (mg/dL)	5.04 ± 1.66
PTH (pg/mL)	288.35 ± 297.72
Alkaline phosphatase (UI/L)	124.88 ± 108.64
Potassium (mEq/L)	4.91 ± 0.89
Magnesium (mg/dL)	2.22 ± 0.45
β2-microglobulin (µg/L)	19.44 ± 8.61
Others	
BMI, mean ± SD	27.1 ± 5.41
Residual diuresis (mL), mean ± SD	631.73 ± 730.6
Vascular access (catheter), *n* (%)	830 (53)

mortality is given not only by the influence of that specific variable, but also by the effect of other dependent variables (covariates) that in a non-linear manner may also affect mortality. This peculiarity gives to the model a higher degree of freedom to capture complex relationships between input variables and outcome. For example, age >80 years may increase the mortality risk for a patient with haemoglobin <10 g/dL but may have less effect (or no effect) for patients with haemoglobin between 10 and 12 g/dL. So the effect of each variable cannot be isolated and measured as it is in logistic regression. Thus the evaluation of the effect of each variable on the probability of the event is difficult. In general, a clear statistical description of the effect of the variables on the prediction of the event, as in logistic regression (magnitude of the coefficient and P-value), is not computable and only an indirect qualitative measure can be obtained. For this reason, in the present study, the method of analysing the change in the AUC value was used when the effect of each specified variable is turned off. This is the method used to compare random forest and logistic regression.

The description of this study follows the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) Guidelines [22]. Therefore the completed TRIPOD checklist is provided as Supplementary Table S1.

RESULTS

The study includes a total of 1571 incident HD patients collected throughout a period of up to 20 years who fulfilled the inclusion criteria (Figure 1).

Of the total cohort, 61% were male and the mean age at the initiation of dialysis was 62.3 years. The mean Charlson comorbidity index score was 5.99. Table 1 shows baseline comorbidities and biochemistry results obtained during the first 30 days of dialysis.

Results from random forest analysis are shown in Figure 2 and from logistic regression analysis in Figure 3. For both approaches, the eight most important variables are reported. In Figures 2A-I and 3A–I, the dashed line represents the value of the AUC of the mortality prediction ROC curve obtained by both models. Each dot shows the AUC value obtained if the effect of the specified variable is turned off; this is achieved by randomly changing the values of the variable in the test set; therefore, the larger the decrease in the AUC value, the greater impact of the variable in the prediction of mortality.

Table 2 shows a comparison of the AUCs obtained by random forest and logistic regression for different baseline collection periods (30, 60 and 90 days) and prediction of mortality at 6 months, 1 year and 2 years.

Prediction of mortality of HD patients by random forest and logistic regression

Prediction of mortality after 6 months on HD. Separate analyses were performed using the baseline data collected at 30, 60 and 90 days. Results from the random forest analysis are shown in Figure 2A–C and from logistic regression in Figure 3A–C.

Using baseline data at 30 days (Figure 2A), the AUC obtained by the random forest mortality prediction model is 0.70. The most influential variable is serum creatinine; the absence of serum creatinine reduced the AUC value from 0.70 to 0.68. The next two most influential variables are Kt/V_{urea} and BMI. Using baseline data at 60 days (Figure 2B), the AUC is 0.68 and the three variables with the most predictive power are haemoglobin, calcium and potassium. With baseline data at 90 days, the AUC is 0.72 (Figure 2C) and Kt/V_{urea} becomes the variable with the greatest predictive power.

The results from logistic regression analysis are shown in Figure 3. Using the 30-day baseline data, the AUC is 0.69 and serum creatinine had the highest predictive power, followed by vascular access and Kt/V_{urea}. Using the 60-day data, diuresis becomes the variable with the highest predictive power, followed by vascular access and haematologic disease. Using data collected at 90 days, diuresis becomes the variable with the highest predictive power.

Prediction of mortality after 1 year on HD. Results of random forest analysis are shown in Figure 2D–F. Using the baseline data collected at 30 days (Figure 2D), the AUC by random forest is 0.73. The most influential variable is BMI, followed by serum albumin and serum creatinine. Using the 60-day baseline data (Figure 2E), the AUC is 0.73 and the variables with the most predictive power are BMI, serum albumin and Kt/V_{urea}. With 90-day data, the AUC is 0.73 (Figure 2F) and serum albumin becomes the variable with the most predictive power, followed by BMI and haemoglobin.

The results of logistic regression analysis are shown in Figure 3D–F. Considering the 30-day baseline data, the AUC is 0.71. The variables with the highest predictive power are age, diuresis and serum creatinine. With the 60-day data, the AUC is 0.71 and the variables with the most predictive power are age, diuresis and haemoglobin. With the 90-day data, the AUC is 0.72 and the variables with the most predictive power are age, albumin and haemoglobin.

Table 2. Comparisons of AUCs obtained by random forest and logistic regression

Prediction of mortality				AUC					
				Random forest		Logistic regression			
Prediction pPeriod	Period (days) after first HD for baseline data collection	Number of patients	Deaths	AUC (%)	95% CI	AUC (%)	95% CI	Difference in AUC (RF − LR) (%)	P-value
6 months	30	1456	80	70.14	67.95–72.33	69.01	66.8–71.21	1.13	0.32
	60	1432	56	67.55	64.88–70.22	66.84	64.15–69.52	0.71	0.61
	90	1419	43	71.75	68.84–74.65	67.15	64.18–70.13	4.60	0.18
1 year	30	1336	166	73.31	71.8–74.82	71.16	69.62–72.7	2.15	0.01*
	60	1312	142	73.19	71.56–74.81	71.22	69.57–72.87	1.97	0.02*
	90	1299	129	72.82	71.12–74.52	71.94	70.22–73.65	0.88	0.32
2 years	30	1244	271	72.59	71.37–73.81	68.73	67.47–69.99	3.86	<0.001
	60	1220	247	72.42	71.14–73.7	68.64	67.33–69.96	3.78	<0.001
	90	1207	234	72.06	70.75–73.37	69.78	68.45–71.12	2.28	<0.001

For each mortality prediction period (6 months, 1 year and 2 years), analysis was based on baseline variable values obtained during a minimum number of days after the first HD session: 30 days, 60 days and 90 days.

Prediction of mortality after 2 years on HD. The results of the random forest analysis are shown in Figure 2G–I. With baseline data at 30 days (Figure 2G), the AUC is 0.73 and the most influential variable is BMI, followed by age and serum albumin. Using the 60-day data (Figure 2H), the AUC is 0.72 and the variables with the most predictive power are the same as with 30-day data. With 90-day data, the AUC is 0.72 (Figure 2I) and the variables with significant predictive power are the same as with 30- and 60-day data.

Figure 3G-I shows the results of logistic regression analysis. Considering data collected at 30, 60 and 90 days, the AUCs are 0.69, 0.69 and 0.70, respectively. Age, BMI and Kt/V_{urea} are the variables with significant predictive value for data at 30 days (Figure 3G). Age, diuresis and BMI are the variables with significant predictive value for data at 60 days (Figure 3H). Age, serum albumin and BMI are the variables with significant predictive value for data at 90 days (Figure 3I).

Comparison of results obtained with logistic regression and random forest

Using the baseline data collected at 30 days, the AUCs predicting mortality at 6 months, 1 year and 2 years obtained with random forest are greater than with logistic regression (Table 2). The difference in AUC in favour of random forest is greater in the prediction of 2-year mortality (3.86%, $P < 0.001$).

Similar results were obtained for data collected at 60 days. The AUC by random forest is significantly greater than that of logistic regression (3.78%, $P < 0.001$). The AUC by random forest is also superior to logistic regression for the 90-day data collection period at 2 years (2.28%, $P < 0.001$). In all the prediction periods, the benefit of random forest tends to increase with the number of events (deaths).

DISCUSSION

The aim of the present study was to compare models to predict mortality in HD patients based on conventional logistic regression and random forest analysis [23]. The mortality of incident HD patients at 6 months, 1 year and 2 years was analysed using baseline data collected during the 30, 60 and 90 days after the initiation of regular HD. The main finding was that the mortality prediction models obtained with the random forest method were more accurate than with logistic regression. The superiority of random forest versus logistic regression was greater in the prediction of mortality at 2 years. Furthermore, the variables that each method identified as predictors of mortality were not always the same.

Logistic regression showed that the most determinant variables predicting mortality were age, type of vascular access (the fistula being protective of mortality) and also the residual diuresis volume. These results agree with studies such as the Dialysis Outcomes and Practice Patterns Study [24], which identified age, vascular access, albumin and other comorbidities as variables that may influence mortality. Other studies [25] have shown that mortality at 6 months is determined by age, dementia, peripheral vascular disease and low serum albumin. The England registry [7], which included 5447 patients on HD and peritoneal dialysis (PD) followed for 3 years, showed that mortality was associated with the following variables: advanced age, being white, diabetes as the cause of end-stage renal disease, HD (versus PD), vascular disease, smoking, haemoglobin and serum values of albumin, creatinine, calcium and phosphorus.

According to random forest analysis, the main variables predicting mortality in HD at 6 months are serum albumin, Kt/V_{urea} and haemoglobin. Variables predicting mortality at 1–2 years were BMI, age, serum albumin and Kt/V_{urea}.

Interestingly, random forest identifies serum albumin more often as one of the most important predictive variables and logistic regression identifies residual diuresis and vascular access as predictors of mortality.

In our study, the ability to predict mortality by both tests was compared by analysing the AUC of the ROC curves obtained by both methods (Table 2). The superiority of random forest versus logistic regression in predicting mortality was statistically significant at years 1 and 2, but it was more evident in the prediction of mortality at year 2. At 6 months, random forest was not significantly superior to logistic regression, probably due to the small number of events (deaths). Furthermore, logistic regression has the worst prediction accuracy (67.2%) and random forest has the best (73.3%).

There are studies that demonstrate the superiority of random forest if the data include variables that have an influence

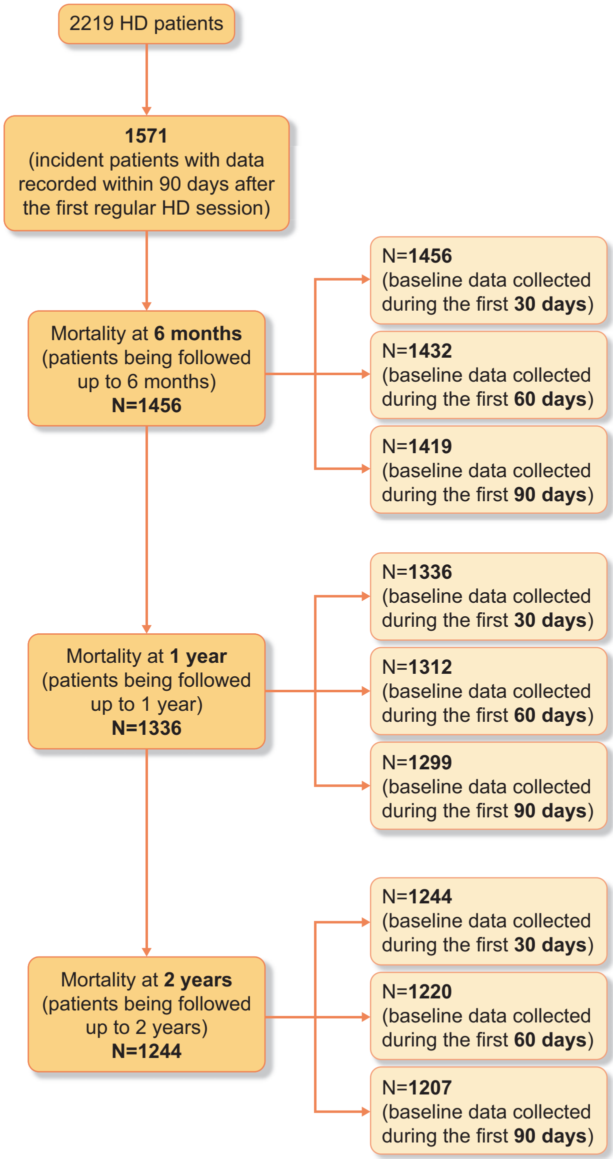
2219 HD patients
1571
(incident patients with data recorded within 90 days after the first regular HD session)
Mortality at 6 months
(patients being followed up to 6 months)
N=1456
N=1456
(baseline data collected during the first 30 days)
N=1432
(baseline data collected during the first 60 days)
N=1419
(baseline data collected during the first 90 days)
Mortality at 1 year
(patients being followed up to 1 year)
N=1336
N=1336
(baseline data collected during the first 30 days)
N=1312
(baseline data collected during the first 60 days)
N=1299
(baseline data collected during the first 90 days)
Mortality at 2 years
(patients being followed up to 2 years)
N=1244
N=1244
(baseline data collected during the first 30 days)
N=1220
(baseline data collected during the first 60 days)
N=1207
(baseline data collected during the first 90 days)

FIGURE 1: Cohort selection flow chart. The comparison of mortality models was performed on nine different cohorts that are represented in the dashed boxes. On the left, the number of patients included to evaluate the prediction of mortality at 6 months, 1 year and 2 years. For each mortality prediction period there were three separate analyses according to the minimal period after the first HD session used for the collection of baseline data (input variables: 30, 60 or 90 days).

on each other [14, 15]. The use of random forest was instrumental in establishing an association among parameters of mineral metabolism in HD patients [18]. The identification of variables that predict outcome has allowed the construction of models and formulating algorithms with a specific objective, such as anaemia treatment with erythropoiesis-stimulating agents [26]. Other studies have shown the utility of random forest interaction trees on predictive covariates and the estimation of individualized treatment effects [17].

One limitation of our study is that the data come from a single centre and the values of mortality may not be extrapolated to other populations; however, comparison of different methods of analysis should be valid. Another limitation is that information on tobacco use and treatment with erythropoietin, intravenous iron and vitamin D was not included. Data collection was rigorous and the number of patients and follow-up periods were adequate and therefore the results obtained should be applicable to this type of population. Prospective studies are necessary to confirm our results. Analysis based on the artificial intelligence approach generally requires a large amount of data, which is now expedited by advances in computer processing power, relatively cheap digital storage and a

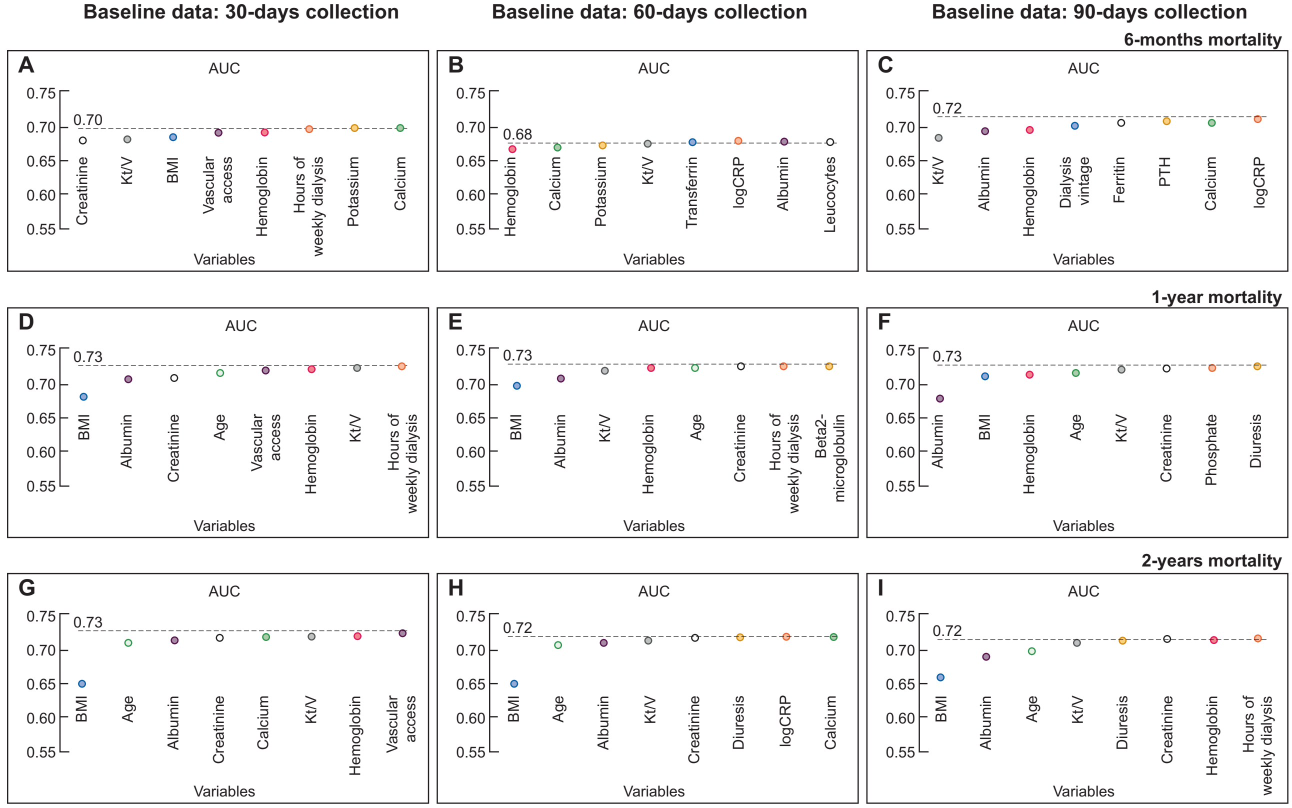

FIGURE 2: Prediction of mortality by random forest analysis. The dashed line represents the value of the AUC of the mortality prediction ROC curve obtained by the random forest regression model. Each dot shows the influence of each variable on the AUC value that is obtained if the effect of the specified variable is turned off. This is achieved by randomly changing the values of the variable in the test set. Each graph represents the result of mortality prediction (at 6 months, 1 year and 2 years) and for each mortality prediction period there were three separate analyses according to the minimal period of days after the first HD session used for the collection of the baseline data (input variables 30, 60 and 90 days). **(A)** Prediction of mortality at 6 months and a 30-day period after the first HD for baseline data collection. **(B)** Prediction of mortality at 6 months and a 60-day period after the first HD for baseline data collection. **(C)** Prediction of mortality at 6 months and a 90-day period after the first HD for baseline data collection. **(D)** Prediction of mortality at 1 year and a 30-day period after the first HD for baseline data collection. **(E)** Prediction of mortality at 1 year and a 60-day period after the first HD for baseline data collection. **(F)** Prediction of mortality at 1 year and a 90-day period after the first HD for baseline data collection. **(G)** Prediction of mortality at 2 years and a 30-day period after the first HD for baseline data collection. **(H)** Prediction of mortality at 2 years and a 60-day period after the first HD for baseline data collection. **(I)** Prediction of mortality at 2 years and a 90-day period after the first HD for baseline data collection.

flood of available digital data. The inherent requirement for large-scale, high-quality, well-structured data might ultimately limit the areas in which artificial intelligence can bring benefits to healthcare [13].

Not all published studies have found the same variables as predictors of mortality. In the present study, the analysis of the same data by two different methods did not identify the same variables as predictors of mortality. The diversity of methods may not allow uniformity of results.

The ability to interpret the logistic regression variable coefficients (ORs) is an important advantage of this method, but this oversimplification may limit the accuracy of the prediction with respect to random forest. Therefore one must be inclined to recognize that the variables identified by random forest are more reasonable than logistic regression. This last statement is not easy to assume after years of logistic regression analysis. Nevertheless, the analysis of the predictive value of single variables should be interpreted cautiously. In random forest, the predictive value of a variable on mortality is given not only by that specific variable, but also by the effect of other dependent variables (covariates) that in a non-linear manner may also affect mortality. It is easy to compare the predictive values of variables obtained by logistic regression and random forest.

Cox regression is another popular method for survival analysis and mortality prediction. Since with Cox the patients lost to follow-up are generally considered to build the model, we preferred comparing random forest with logistic regression so as to have the two models trained on exactly the same set of patients. Furthermore, different from random forest and logistic regression, Cox regression cannot be considered a binary classifier.

Additional studies will be needed to determine if one method like random forest is more accurate and useful than other more classical methods. The authors of the present work have no opinion as to whether one method of analysis is better than another; however, the information obtained clearly indicates that the applicability of artificial intelligence in determining mortality in HD patients is more than acceptable.

In conclusion, random forest regression analysis is an alternative valid method to identify variables and generate models that are useful to predict mortality in HD patients. The incorporation of statistical methods based on artificial intelligence holds promise for substantially improving healthcare delivery. It is predicted that familiarity with these methods for analysing big data will be a fundamental requirement for the next generation of physicians [27]. They will become important actors in

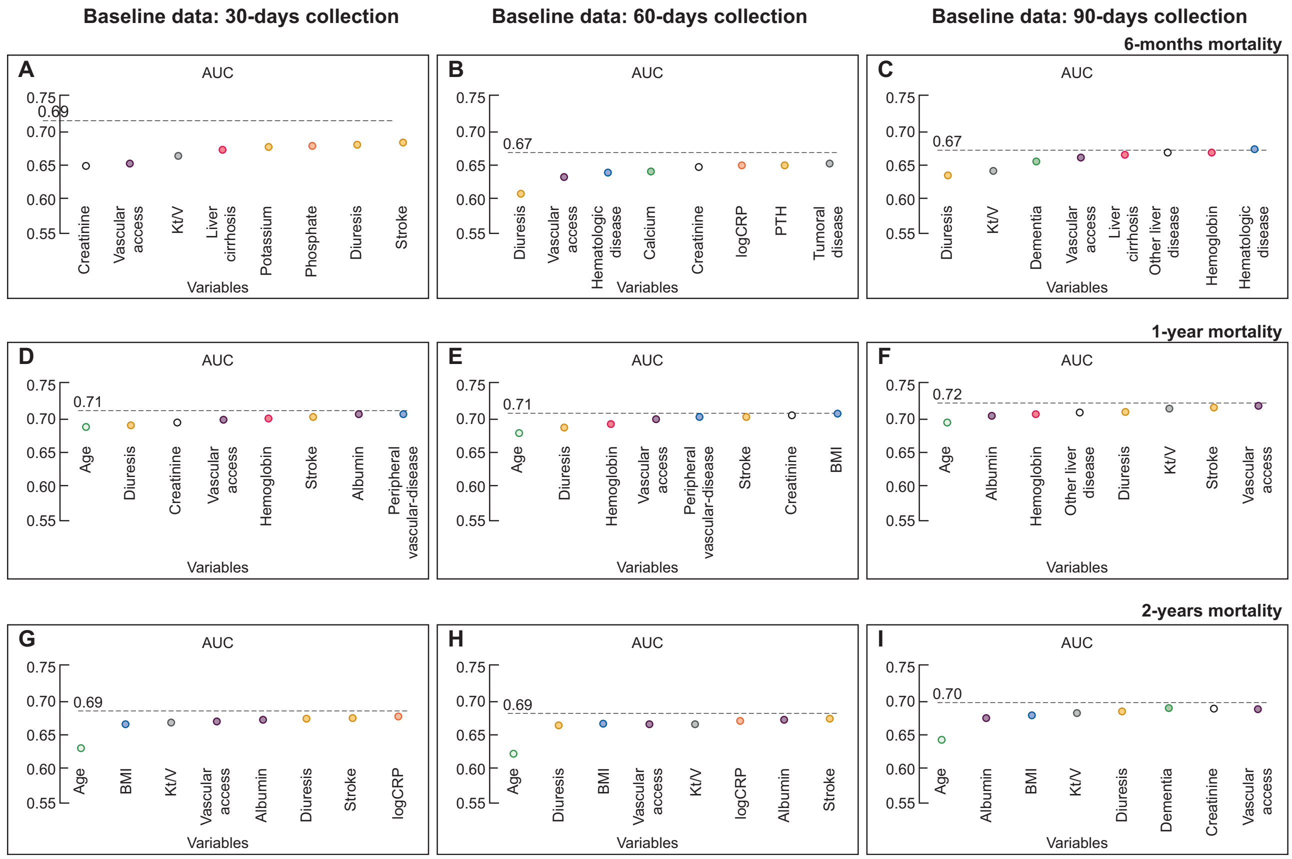

FIGURE 3: Influence of variables on mortality by logistic regression. The dashed line represents the value of the AUC of the mortality prediction ROC curve obtained by the logistic regression model. Each dot shows the AUC value obtained if the effect of the specified variable is removed. Each graph represents the result of mortality prediction at 6 months, 1 year and 2 years and for each mortality prediction period there were three separate analyses according to the minimal period of days after the first HD session used for the collection of baseline data (input variables 30, 60 and 90 days). **(A)** Prediction of mortality at 6 months and a 30-day period after the first HD for baseline data collection. **(B)** Prediction of mortality at 6 months and a 60-day period after the first HD for baseline data collection. **(C)** Prediction of mortality at 6 months and a 90-day period after the first HD for baseline data collection. **(D)** Prediction of mortality at 1 year and a 30-day period after the first HD for baseline data collection. **(E)** Prediction of mortality at 1 year and a 60-day period after the first HD for baseline data collection. **(F)** Prediction of mortality at 1 year and a 90-day period after the first HD for baseline data collection. **(G)** Prediction of mortality at 2 years and a 30-day period after the first HD for baseline data collection. **(H)** Prediction of mortality at 2 years and a 60-day period after the first HD for baseline data collection. **(I)** Prediction of mortality at 2 years and a 90-day period after the first HD for baseline data collection.

the therapeutic relationship and will need to be bound by the core ethical principles, such as beneficence and respect for patients, that have guided clinicians [28].

SUPPLEMENTARY DATA

Supplementary data are available at ckj online.

FUNDING

This study was supported by grants from the National Institute of Health Carlos III (FIS 17/01785, FIS 17/01010), RETICs Red Renal RD06/0016/0007, the Consejeria de Salud of Junta de Andalucia (PI-0311-2014), the REDinREN from the National Institute of Health Carlos III (RD16/0009/0034) and the European group EUTox and CKD-MBD group.

CONFLICT OF INTEREST STATEMENT

M.R. has received research grants from Amgen and lecture fees from the following companies: Amgen, Vifor Pharma, Kyowa and Sanofi. A.M.-M. has received lecture fees from Medtronic/Bellco, Vifor Pharma, Astellas and AstraZeneca in the last 3 years. C.B. and F.B. are employed by Fresenius Medical Care. There are no patents, products in development or marketed products to declare. All the other authors declare no potential conflicts of interest. The results presented in this article have not been published previously in whole or part, except in abstract format.

REFERENCES

1. Foote C, Woodward M, Jardine MJ. Scoring risk scores: considerations before incorporating clinical risk prediction tools into your practice. *Am J Kidney Dis* 2017; 69: 555–557
2. Kasza J, Wolfe R, McDonald SP *et al.* Dialysis modality, vascular access and mortality in end-stage kidney disease: a binational registry-based cohort study. *Nephrology (Carlton)* 2016; 21: 878–886.
3. Chen YM, Wang YC, Hwang SJ *et al.* Patterns of dialysis initiation affect outcomes of incident hemodialysis patients. *Nephron* 2016; 132: 33–42
4. Bradbury BD, Fissell RB, Albert JM *et al.* Predictors of early mortality among incident US hemodialysis patients in the Dialysis Outcomes and Practice Patterns Study (DOPPS). *Clin J Am Soc Nephrol* 2007; 2: 89–99
5. Canaud B, Tong L, Tentori F *et al.* Clinical practices and outcomes in elderly hemodialysis patients: results from the

Dialysis Outcomes and Practice Patterns Study (DOPPS). *Clin J Am Soc Nephrol* 2011; 6: 1651–1662

6. Foley RN, Parfrey PS, Hefferton D *et al*. Advance prediction of early death in patients starting maintenance dialysis. *Am J Kidney Dis* 1994; 23: 836–845
7. Wagner M, Ansell D, Kent DM *et al*. Predicting mortality in incident dialysis patients: an analysis of the United Kingdom Renal Registry. *Am J Kidney Dis* 2011; 57: 894–902
8. Chen JY, Tsai SH, Chuang PH *et al*. A comorbidity index for mortality prediction in Chinese patients with ESRD receiving hemodialysis. *Clin J Am Soc Nephrol* 2014; 9: 513–519
9. Couchoud CG, Beuscart J-BR, Aldigier J-C *et al*. Development of a risk stratification algorithm to improve patient-centered care and decision making for incident elderly patients with end-stage renal disease. *Kidney Int* 2015; 88: 1178–1186
10. Couchoud C, Labeeuw M, Moranne O *et al*. A clinical score to predict 6-months prognosis in elderly patients starting dialysis for end-stage renal disease. *Nephrol Dial Transplant* 2009; 24: 1553–1561.
11. Wick JP, Turin TC, Faris PD *et al*. A clinical risk prediction tool for 6-month mortality after dialysis initiation among older adults. *Am J Kidney Dis* 2017; 69: 568–575
12. Hsu JY, Roy JA, Xie D *et al*. Statistical Methods for Cohort Studies of CKD: survival analysis in the setting of competing risks. *Clin J Am Soc Nephrol* 2017; 12: 1181–1189
13. Artificial intelligence in health care: within touching distance. *Lancet* 2018; 390: 2739
14. Genuer R, Poggi J-M, Tuleau-Malot C. Variable selection using random forests. *Pattern Recognit Lett* 2010; 31: 2225–2236
15. Matsuki K, Kuperman V, Van Dyke JA. The random forests statistical technique: an examination of its value for the study of reading. *Sci Stud Read* 2016; 20: 20–33
16. Dankowski T, Ziegler A. Calibrating random forests for probability estimation. *Stat Med* 2016; 35: 3949–3960
17. Su X, Peña AT, Liu L *et al*. Random forests of interaction trees for estimating individualized treatment effects in randomized trials. *Stat Med* 2018; 37: 2547–2560
18. Rodriguez M, Salmeron MD, Martin-Malo A *et al*. A new data analysis system to quantify associations between biochemical parameters of chronic kidney disease-mineral bone disease. *PLoS One* 2016; 11: e0146801
19. Lu M, Sadiq S, Feaster DJ *et al*. Estimating individual treatment effect in observational data using random forest methods. *J Comput Graph Stat* 2018; 27: 209–219
20. Wang H, Shen L, Geng J *et al*. Prognostic value of cancer antigen -125 for lung adenocarcinoma patients with brain metastasis: a random survival forest prognostic model. *Sci Rep* 2018; 8: 5670
21. Roy J, Shou H, Xie D *et al*. Statistical methods for cohort studies of CKD: prediction modeling. *Clin J Am Soc Nephrol* 2017; 12: 1010–1017.
22. Collins G, Johannes B, Douglas G *et al*. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD Statement. *Ann Intern Med* 2015; 162: 55–65
23. Agarwal R, Sinha AD. Big data in nephrology—a time to rethink. *Nephrol Dial Transplant* 2018; 33: 1–3
24. Pisoni RL, Gillespie BW, Dickinson DM *et al*. The Dialysis Outcomes and Practice Patterns Study (DOPPS): design, data elements, and methodology. *Am J Kidney Dis* 2004; 44: 7–15
25. Chan KE, Maddux FW, Tolkoff-Rubin N *et al*. Early outcomes among those initiating chronic dialysis in the United States. *Clin J Am Soc Nephrol* 2011; 6: 2642–2649
26. Barbieri C, Molina M, Ponce P *et al*. An international observational study suggests that artificial intelligence for clinical decision support optimizes anemia management in hemodialysis patients. *Kidney Int* 2016; 90: 422–429
27. Couronné R, Probst P, Boulesteix AL. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics* 2018; 19: 270
28. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. *N Engl J Med* 2018; 378: 981–983