


# SCIENTIFIC REPORTS



OPEN

## Development and validation of a multivariate predictive model for rheumatoid arthritis mortality using a machine learning approach

José M. Lezcano-Valverde<sup>1</sup>, Fernando Salazar<sup>2</sup>, Leticia León<sup>1</sup>, Esther Toledano<sup>1</sup>, Juan A. Jover<sup>1</sup>, Benjamín Fernández-Gutiérrez<sup>1</sup>, Eduardo Soudah<sup>2</sup>, Isidoro González-Álvaro<sup>3</sup>, Lydia Abasolo<sup>1</sup> & Luis Rodríguez-Rodríguez<sup>1</sup> 

We developed and independently validated a rheumatoid arthritis (RA) mortality prediction model using the machine learning method Random Survival Forests (RSF). Two independent cohorts from Madrid (Spain) were used: the Hospital Clínico San Carlos RA Cohort (HCSC-RAC; training; 1,461 patients), and the Hospital Universitario de La Princesa Early Arthritis Register Longitudinal study (PEARL; validation; 280 patients). Demographic and clinical-related variables collected during the first two years after disease diagnosis were used. 148 and 21 patients from HCSC-RAC and PEARL died during a median follow-up time of 4.3 and 5.0 years, respectively. Age at diagnosis, median erythrocyte sedimentation rate, and number of hospital admissions showed the higher predictive capacity. Prediction errors in the training and validation cohorts were 0.187 and 0.233, respectively. A survival tree identified five mortality risk groups using the predicted ensemble mortality. After 1 and 7 years of follow-up, time-dependent specificity and sensitivity in the validation cohort were 0.79–0.80 and 0.43–0.48, respectively, using the cut-off value dividing the two lower risk categories. Calibration curves showed overestimation of the mortality risk in the validation cohort. In conclusion, we were able to develop a clinical prediction model for RA mortality using RSF, providing evidence for further work on external validation.

Rheumatoid arthritis (RA) is a chronic, systemic, inflammatory disease, characterized by inflammatory arthritis and localized destruction of bone, cartilage, and periarticular structures. This condition is associated with an increased mortality risk and a reduced life expectancy of about 3 to 10 years compared with the general population<sup>1–5</sup>.

Several socio-demographic and clinical-related factors with a significant impact in RA mortality have been identified<sup>4–11</sup>, mostly through the use of traditional survival techniques, such as the Cox proportional hazards (CPH) model<sup>12</sup>. However, these models have several limitations, including their reliance on restrictive assumptions, such as proportional hazards, being often parametric, therefore having to model nonlinear effects and interactions, which increases the risk of over-fitting and diminishes the statistical power of the model<sup>13,14</sup>, and lacking reliability in the presence of high rates of censoring<sup>15</sup>.

In order to overcome these limitations and to improve the predictive performance, machine learning methods/models have been developed. These methods are able to “learn” from experience (data) and create predictive and prognostic models with high accuracy, reliability, and efficiency<sup>16</sup>.

Among the several machine learning methods that have been developed, random survival forest (RSF) has been proposed as an alternative approach to traditional survival methods<sup>13</sup>. RSF is a non-parametric method that generates multiple decision trees using bootstrap samples from the original data. Based on the majority votes of the individual decision trees, it is able to predict the outcome of interest. When the primary outcome is survival

<sup>1</sup>Rheumatology Department, Hospital Clínico San Carlos, and IdISSC, Madrid, Spain. <sup>2</sup>International Centre for Numerical Methods in Engineering (CIMNE), Madrid, Spain. <sup>3</sup>Rheumatology Department, Hospital Clínico Universitario de La Princesa, and IIS-IP, Madrid, Spain. José M Lezcano-Valverde and Fernando Salazar contributed equally to this work. Lydia Abasolo and Luis Rodríguez-Rodríguez jointly supervised this work. Correspondence and requests for materials should be addressed to L.R.-R. (email: [lrrodriguez@salud.madrid.org](mailto:lrrodriguez@salud.madrid.org))

(time to event), RSF produces a cumulative hazard function (CHF) from each individual decision tree that are averaged in an ensemble CHF. RSF has been used for the analysis of right-censored survival data in several human diseases, such as cancer and cardiovascular diseases<sup>17,18</sup>.

The objective of our study was to develop and validate, both internally and externally, a RSF prediction model of mortality in RA patients based on demographic and clinical-related variables collected during the first two years after disease diagnosis.

## Results

**Cohort description.** 1,461 patients from the Hospital Clínico San Carlos RA cohort (HCSC-RAC) and 280 RA patients from the Hospital Universitario de La Princesa Early Arthritis Register Longitudinal (PEARL) study were included in this study. The former is a day-to-day clinical practice cohort that includes subjects that have received a clinical diagnosis of RA by their usual rheumatologist<sup>5</sup>. The latter is an early arthritis cohort including RA and undifferentiated arthritis patients<sup>19</sup>. Demographic and clinical characteristics are shown in Table 1. Median follow-up time starting two years after RA diagnosis was 4.3 years [interquartile range (IQR): 2.0–6.8; range: 1 day–10.7 years] for the HCSC-RAC, and 5.0 (2.1–8.1; range: 3 days–11.3 years) for the PEARL. During follow-up, 148 (10.1%; time of observation of 6,707.6 person-years), and 21 (7.5%; time of observation of 1,441.4 person-years) patients died, resulting in a mortality rate of 22.1 [18.8 to 25.9], and 14.6 [9.5–22.3] events per 1,000 patients-year, respectively.

The variables presence of anti citrullinated peptide antibodies presence (ACPA) and median Health Assessment Questionnaire (HAQ) value in the first 2 years after RA diagnosis were excluded from the analysis due to their high proportion of missing data in the HCSC-RAC (Table 1).

**Model Development.** First, we assessed that the number of trees included in the models were enough to obtain the lowest possible prediction error rate for that model, a measure of its discrimination ability. As showed online in Supplementary Figs S1 and S2, the higher the number of trees, the lower the prediction error. Furthermore, the prediction error stabilized above 200 trees, approximately, regardless the splitting rule use to construct the model [log-rank ( $M_{LR}$ ) or log-rank score ( $M_{LRS}$ )].

The parameters used for the development of the  $M_{LR}$  and  $M_{LRS}$  are shown in Table 2. We present the results of the model using log-rank as splitting rule ( $M_{LR}$ ), since it exhibited the lowest prediction error, and therefore, the highest discrimination ability. In addition, the integrated Brier score (IBS; a measure of the model's accuracy) for the overall follow-up was also lower when the log-rank splitting rule was used (Table 2 and Supplementary Figure S3).

Next, we assessed the classification of the variables according to their predictive ability (Table 3), in order to select those variables to be included in the final model. The most important predictor variable was the patient's age at the time of RA diagnosis, followed by the median erythrocyte sedimentation rate (ESR) during the first 2 years after RA diagnosis, the number of hospital admissions, the calendar year of RA diagnosis, and being Spaniard. The variables presence of rheumatoid factor (FR), use of any biological therapies during the first 2 years after RA diagnosis, elapsed time from RA symptoms onset to diagnosis, and gender also showed some predictive capacity, although considerably lower. Because all the variables showed a positive variable importance (VIMP), none was excluded from the final model, and therefore the  $M_{LR}$  was our final model.

The effect on survival of these main variables was displayed with partial survival plots (Supplementary Figure S4) representing the predicted mortality rate for a given variable, after adjusting for all other variables. Older age at RA diagnosis, higher number of hospital admissions, higher median value of ESR during the first 2 years after RA diagnosis, and higher elapsed time from symptoms onset to diagnosis were associated with higher predicted mortality. Conversely, a more recent calendar year of RA diagnosis and the use of biological therapies were associated with lower mortality. In addition, we observed that the effect of the continuous variables in survival was not linear.

**Model Validation.** After we developed our model using the HCSC-RAC, we used the data from the PEARL study to externally validate its performance. Using the  $M_{LR}$ , we observed a prediction error in the validation cohort of 0.233. Comparing with the training cohort, we observed an increase in the prediction error, and therefore a worsening of the discrimination ability our model.

Finally, we performed a survival tree analysis using the individual predicted ensemble mortality from the HCSC-RAC to identify different mortality risk groups. The predicted ensemble mortality is the mean CHF estimated by the  $M_{LR}$  for each subject, and it was used as a measure of each patient estimated mortality risk. The cut-off values defining the risk groups are shown online in the Supplementary Table S1 and were applied to both cohorts. The mortality rate for each risk group and their comparison between groups, performed using a CPH model, both for the HCSC-RAC and for the PEARL study, are showed online in Supplementary Tables S2, and S3, respectively.

In order to reduce the number of groups and to maximize the differences among them, we decided to combine the three with intermediate risk, resulting in three final groups with low, intermediate, and high mortality risk (see Figs 1, 2, and Supplementary Table S4). In new CPH models, we observed that the intermediate and the high risk groups were significantly associated with higher mortality risk compared with the low risk, both in the training and in the validation cohorts.

As an example of how to apply the final model we developed and validated, we plotted the predicted survival curves of two different fictitious RA patients with different demographic and clinical characteristics (Supplementary Table S5 and Supplementary Figure S5). The predicted ensemble mortality were 1.1 and 12.5 for Patient A and B, respectively.

Variables	HCSC-RAC		PEARL	
	n = 1,461	Missing data, n (%)	n = 280	Missing data, n (%)
Women, n (%)	1,105 (75.6)	0	223 (79.6)	0
Age of RA diagnosis, median (IQR)	58.6 (45.2–72.0)	0	54.9 (45.3–67.6)	0
Elapsed time from RA symptoms onset to diagnosis, in years, median (IQR)	0.7 (0.3 to 3.5)	180 (12.3)	0.5 (0.3–0.7)	0
Presence of Rheumatoid Factor, n (%)	885 (61.5)	23 (1.6)	181 (64.6)	0
Presence of ACPA, n (%)	465 (44.9)	425 (29.1)	234 (83.9)	1 (0.4)
Nationality, n (%):		0		0
Spanish	1,160 (79.4)	—	232 (82.7)	—
South/Centre America, Caribbean	237 (16.2)	—	37 (13.2)	—
Other	64 (4.4)	—	11 (3.9)	—
Year of RA diagnosis, n (%):		0		0
2001–2005	614 (42.0)	—	107 (38.2)	—
2006–2011	847 (58.0)	—	128 (45.7)	—
2012–2014	0		45 (16.1)	—
Median HAQ in the first 2 years after RA diagnosis, median (IQR)	0.50 (0.19–1.10)	376 (25.7)	0.63 (0.25–1.00)	1 (0.4)
Median ESR in the first 2 years after RA diagnosis, median (IQR)	23 (14 to 36.5)	248 (17.0)	20 (13–30)	1 (0.4)
Any biological therapy in the first 2 years after RA diagnosis, n (%)	89 (6.1)	0	28 (10.0)	0
Hospital admissions in the first 2 years after RA diagnosis, n (%)		0		13 (4.6)
0	1,258 (86.1)	—	230 (86.1)	—
1	144 (9.9)	—	28 (10.5)	—
2	39 (2.7)	—	5 (1.87)	—
3	16 (1.1)	—	2 (0.8)	—
≥4	4 (0.28)	—	2 (0.8)	—
Inclusion period, calendar years	2001–2011	—	2001–2014	—

**Table 1.** Demographic and clinical characteristics of the rheumatoid arthritis patients from the “Hospital Clínico San Carlos - Rheumatoid Arthritis Cohort” and from the “Hospital Universitario de La Princesa Early Arthritis Register Longitudinal” with more than 2 years of follow-up after disease diagnosis. ACPA: Anti-citrullinated peptides antibodies, ESR: Erythrocyte sedimentation rate, HAQ: Health assessment questionnaire, IQR: Interquartile range, RA: Rheumatoid Arthritis.

**Sensitivity, specificity, and model calibration.** The time-dependent sensitivity and specificity was estimated for particular time points (1, 2, 5 and 7 years of follow-up, starting two years after RA diagnosis). Supplementary Figures S6 to S9 online show the relationship between sensitivity/specificity and the predicted individual ensemble mortality for that particular time point, both for the HCSC-RAC and the PEARL study. As expected, the higher the value of the predictive ensemble mortality, the higher the specificity and the lower the sensitivity. When compared both cohorts and regardless the time point and the predicted individual ensemble mortality, a greater specificity and a lower sensitivity was observed in the PEARL study. In the same way, similar results were observed for each of the cut-off points estimated using the survival tree analysis (online Supplementary Table S6).

Calibration curves for particular time points of follow-up (starting two years after RA diagnosis) are showed online in Supplementary Figures S10 and S11. For the HCSC-RAC, at 2 years of follow-up, the  $M_{LR}$  tended to underestimate the mortality risk of those with lower risk. Conversely, in those groups with higher risk, the model tended to overestimate the risk. For 5 and 7 years of follow-up, the model overestimate the risk for those groups with lower risk, and underestimate the mortality risk of those with higher risk. In the patients for the PEARL study, the  $M_{LR}$  tended to overestimate the mortality risk.

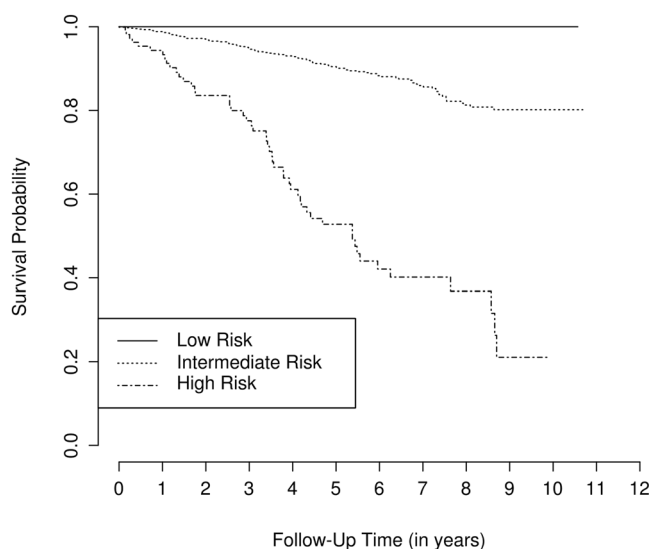
**Sensitivity Analysis.** We developed two new models in the training cohort, constructed with the log-rank splitting rule ( $M_{LR}$ ): a) an “expanded model” ( $M_{LRexp}$ ), including all available variables, even those previously excluded due to their high proportion of missing values, and b) a “reduced model” ( $M_{LRred}$ ), including only those variables with greater predictive capacity. Both  $M_{LRexp}$  and  $M_{LRred}$  were compared with the final model, in terms of prediction error, IBS, and VIMP. We observed that our final model ( $M_{LR}$ ) had both a prediction error and an IBS between the expanded  $M_{LR}$  ( $M_{LRexp}$ ) and the reduced  $M_{LR}$  ( $M_{LRred}$ ; see Supplementary Table S7 online). Regarding the predictive capacity of the variables included in the  $M_{LRexp}$  (see Supplementary Table S8 online), both ACPA presence and median HAQ had a negative value, and their presence did not alter the ranking of the variables with higher predictive capacity. Regarding the  $M_{LRred}$ , the same ranking as in  $M_{LR}$  was observed.

Model	Splitting rule	Minimum terminal node size, n	Terminal nodes, mean	Variables tried at each split, n	Prediction error, mean (SD)	1 year IBS, mean (SD)	2 years IBS, mean (SD)	5 years IBS, mean (SD)	7 years IBS, mean (SD)	Overall IBS, mean (SD)
M <sub>LR</sub>	Log-rank	3	131.7	3	0.187 (0.002)	0.003 (0.0001)	0.013 (0.0004)	0.070 (0.002)	0.128 (0.003)	0.150 (0.003)
M <sub>LRS</sub>	Log-rank score	3	228.04	3	0.209 (0.003)	0.003 (0.0001)	0.012 (0.001)	0.071 (0.002)	0.140 (0.004)	0.167 (0.004)

**Table 2.** Parameters and quality measures of two random survival forests models using the log-rank or the log-rank score splitting rules developed for the prediction of mortality in a cohort of rheumatoid arthritis patients (HCSC-RAC). IBS: Integrated Brier Score; SD: Standard deviation.

Variables	VIMP, mean (SD)	IR (%)
Age of RA diagnosis	0.110 (0.001)	100
Median ESR in the first 2 years after RA diagnosis	0.014 ( $9.8 \times 10^{-4}$ )	12.7
Hospital admissions in the first 2 years after RA diagnosis	0.012 ( $7.0 \times 10^{-4}$ )	10.5
Calendar year of RA diagnosis	0.009 ( $9.1 \times 10^{-4}$ )	8.4
Spaniard	0.005 ( $5.3 \times 10^{-4}$ )	4.5
Presence of Rheumatoid Factor	$6.1 \times 10^{-4}$ ( $5.7 \times 10^{-4}$ )	0.6
Any biological therapy in the first 2 years after RA diagnosis	$3.5 \times 10^{-4}$ ( $1.7 \times 10^{-4}$ )	0.3
Elapsed time from RA symptoms onset to diagnosis	$2.8 \times 10^{-4}$ ( $9.4 \times 10^{-4}$ )	0.2
Gender	$0.5 \times 10^{-4}$ ( $5.5 \times 10^{-4}$ )	0.1

**Table 3.** Variables included in the random survival forest M<sub>LR</sub> ranked based on their variable importance value (VIMP). ACPA: Anti-citrullinated peptides antibodies; ESR: Erythrocyte sedimentation rate; HAQ: Health assessment questionnaire; RA: Rheumatoid Arthritis; SD: standard deviation; VIMP: Variable importance.

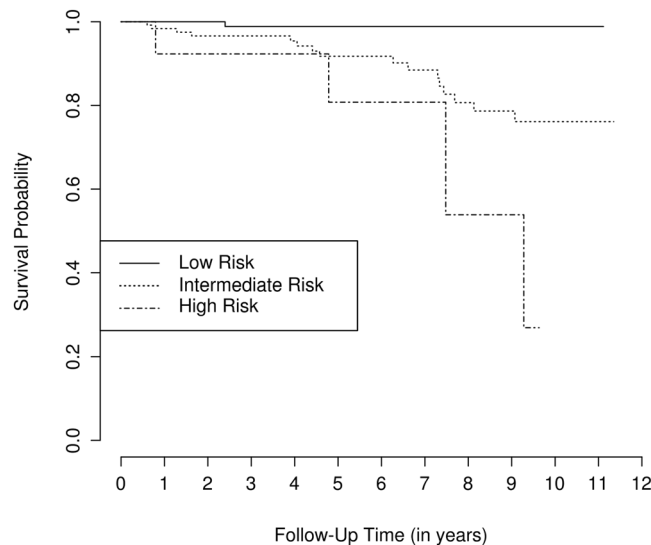


**Figure 1.** Kaplan Meier curves for the observed mortality of patients from the HCSC-RAC. Patient were grouped in mortality risk categories (low, intermediate, and high) according to a rheumatoid arthritis mortality random survival forest model using the log-rank splitting rule (M<sub>LR</sub>).

## Discussion

We have developed and externally validated a clinical predictive model for RA mortality using RSF, based on data collected during the first two years after disease diagnosis. To the best of our knowledge, this is the first time that this machine learning method is applied to analyze RA mortality. Furthermore, we identified several prognostic categories that were able to discriminate among subjects with different mortality risk in our validation cohort.

Several studies have used the RSF method to predict numerous disease outcomes, such as cancer mortality<sup>17</sup> (including glioblastoma<sup>20,21</sup>, leukaemia<sup>22–25</sup>, colon<sup>26</sup>, and thyroid<sup>27</sup> cancer), cancer recurrence<sup>28</sup>, survival of kidney graft<sup>15</sup>, development of Huntington disease<sup>29</sup>, bed occupancy in an intensive care unit<sup>30</sup>, or time to initiation of symptomatic therapy in early Parkinson's disease<sup>31</sup>. In rheumatology, RSF has been used to analyze the mortality risk in juvenile idiopathic inflammatory myopathies<sup>32</sup>, and in systemic lupus erythematosus<sup>33</sup>. Regarding the former, RSF was used to identify the most important variables of mortality in a cohort of 441 patients<sup>32</sup>. Later, those variables were included in a multivariate CPH model, with a prediction error of 23.4%. In addition, 3,839 SLE



**Figure 2.** Kaplan Meier curves for the observed mortality of patients from the PEARL. Patient were grouped in mortality risk categories (low, intermediate, and high) according to a rheumatoid arthritis mortality random survival forest model using the log-rank splitting rule ( $M_{LR}$ ).

patients were used to create a model to predict in-hospital mortality, and to identify the most important variables. The prediction error of the best model was 11.9%. Regarding RA, although two studies used RSF to generate propensity scores to analyze, with a CPH model, the influence of methotrexate<sup>34</sup> and corticosteroids<sup>35</sup> in mortality, this method have not been used to predict mortality. Conversely, CPH models have been used in several studies to analyze the role of demographic and clinical-related variables in the mortality risk of RA patients. Although these studies uncovered numerous associations, they did not characterize these variables in terms of their capacity to predict mortality. Most of the variables that our model identified as having high predicting capacity have been associated with a higher mortality risk in most studies, including older age at inclusion<sup>4,5,7,9</sup> and higher disease activity<sup>4,7,8</sup>. In turn, among those variables with low predictive capacity, association studies have observed either conflicting results or lack of association: use of biological therapy has been associated with lower mortality in some, but not all studies<sup>11,36,37</sup>, probably due to the issue of confounding by indication<sup>38</sup>. Regarding the elapsed time from RA symptoms to diagnosis, Naz *et al.*<sup>6</sup> observed lower mortality risk with greater elapsed time in a cohort of recent-onset inflammatory polyarthritis, although association was lost after adjustment.

However, we also observed contradictory results regarding our observations and previous studies. Some variables with high predictive capacity in our model lacked association with mortality in previous studies, such as calendar time. Most studies have observed either no clear influence of this variable in the mortality rate<sup>4,39</sup>, or a small association with decreased risk<sup>40</sup>. In our study it was the fourth variable with the highest predictive capacity, even when considering that the effect of the other variables in the predictive capacity of calendar time in mortality was taken into account by the model. In the same way, variables with low predictive capacity in our model have showed a consistent association with mortality in most previous studies, including male gender<sup>4,6,9</sup>, or presence of RF<sup>4,6,8,9</sup>.

Regarding hospital admissions, its total number during follow-up has been independently associated with a higher mortality risk in our RA cohort<sup>2</sup>. In turn, the number of the rate of hospitalizations has been associated with increased disease activity, and reduced physical activity<sup>41,42</sup>. Therefore, we could consider this variable as a surrogate marker for RA severity. The fact that it showed predictive capacity even when another measure of disease activity (median ESR) was included in the model could mean that this variable conveys different aspects of the disease inflammatory burden, or other factors associated with mortality.

In daily practice, our prediction model could be implemented in an electronic health record. That way, the data needed for the model could be included automatically from the databases storing the information previously collected by physician or nurses during patient's visits, or from the laboratories. In addition, the information from the model would be displayed in the same application used by the physician during the outpatient visits, not needing to access a different one. Considering the heavy workload of outpatient clinics that would facilitate the visualization of the results and its incorporation in the physician's decision-making.

Our study has several limitations. Although we compared both splitting rules to produce the model with the lowest average prediction error, we used the default values for the rest of the parameters of the R package. In further studies, a fine tuning could increase the accuracy and precision of the models. In addition, despite we were able to develop a clinical model that showed accuracy and precision, we included a limited number of variables in our analysis. The inclusion of other variables, such as comorbidity, lifestyle habits, prescribed treatments, or other potentially relevant variables, and the introduction of interaction among them, could improve the model. Furthermore, another limitation is the high proportion of missing data, particularly for disease activity, disability, and presence of ACPA in the HCSC-RAC. We want to point out that this is an observational study, using retrospective information from a cohort set in real life clinical practice, meaning that the data used was collected

by physicians and nurses during day-to-day practice in an environment of heavy workload. Under those circumstances, higher rates of missing information are expected. In addition, there was no way to retrieve the missing information, as it was not collected (on paper or electronically) at the time of the patient's appointment, or it was the result of erroneous or unsolicited laboratory tests. In addition, in our sensitivity analysis we observed that including those variables with higher proportion of missing data after imputation did not increase the predicting capacity of our models or modified the ranking of the predictive capacity of those variables included in the final model.

Another limitation is that, because RSF models involve thousands of decision trees collectively yielding a prediction, it is not possible to make a straightforward presentation of the classification criteria. However, it is a deterministic system, as the same combination of demographic and clinical characteristics will make the same prediction.

It is important to consider that the determination of the VIMP tends to favour continuous variables over categorical<sup>43</sup>. This could explain the low predictive capacity attributed to gender. However, another dichotomic variable, such as being Spaniard, ranked among those with higher predictive capacity. Therefore we think that this issue was not the main responsible for the low predictive capacity of this variable.

Finally, the same predicted ensemble mortality cut-off value showed a lower sensitivity in our external validation cohort. Considering that our outcome is mortality, it is important to take into account that the cost of a decision that results in a false negative (to consider a patient with high mortality risk as having low risk) is not the same as the cost of a false positive (to consider a patient with low mortality risk as having high risk), and that the former it is likely more desirable. In order to determine a satisfactory cut-off value we will need to test out model in new RA patient cohorts.

This study also has important strengths. We have used an independent RA cohort for external validation. Although most studies only perform an internal validation in order to assess the goodness-of-fit of the developed models, we also tested our model in an independent cohort, observing only a small increase in the prediction error. Comparing with previous studies of development and validation of prediction models, in a recent review<sup>44</sup> the median C-index of samples used for external validation was 0.78 (IQR: 0.69–0.88), similar to what was observed in our study (C-index = 1 – prediction error = 0.77).

Regarding RA mortality prediction models, Provan *et al.*<sup>45</sup> observed a C-index of 0.91 in multivariate logistic regression model of 10 year mortality including age, gender, disease activity and the value of the N-terminal pro-brain natriuretic peptide. Although their model exhibited higher discrimination ability, we want to point out that no internal or external validation was performed, and therefore we have to be cautious with the interpretation of their results. Morisset *et al.*<sup>46</sup> developed a mortality model for patients with rheumatoid arthritis-associated interstitial lung disease. The cross-validated C-index of the final model (a multivariate CPH model) was 0.75, similar to our results. No external validation was performed by the authors.

It is important to point out that the number of events in that cohort was low (21 cases) and that some authors recommend to include at least 100–200 cases<sup>47, 48</sup>. Therefore, our model needs to be validated in different RA cohorts in order to properly evaluate its performance. In addition, it is important to point out that the same researchers that developed the prediction model carried out the external validation analysis, which could lead to bias<sup>49</sup>.

Another strength is that RSF methods do not rely on a restrictive assumption as traditional CPH models do, therefore requiring minimal data assumptions and automatically accounting for complex relationships among variables and with time<sup>13, 31, 50–52</sup>. Furthermore, RSF allows us to compare intuitively the predictive capacity among variables, adjusting for potential multiple interactions<sup>13</sup>. It also presents a reliable method for variable selection despite the presence of multicollinearity<sup>13</sup>, and it is able to reduce over-fitting due to the bootstrapping process used in the generation of decision trees. In line with this advantages over other methods, several studies have shown a higher generalization of the RSF results<sup>18, 53–56</sup>.

Finally, the partial plots have allowed us to visualize nonlinear relationships between variables and mortality, enabling us to identify cut-off point for these variables in further studies.

In conclusion, we have identified potentially modifiable mortality risk factors, which are mostly related to the inflammatory burden of RA. Therefore, a thorough control of inflammation during the early stages of disease could allow the patients to start out, after two years of RA diagnosis, with a lower mortality risk. In addition, we have developed a model that allows us, two years after RA diagnosis, to identify a subgroup of subjects with a higher mortality risk. Further studies need to be performed in order to assess if in this subgroup of patients a particular intervention can be implemented in order to reduce their risk.

## Methods

**Subjects.** We performed a retrospective longitudinal study, using two independent RA cohorts to train and to externally validate our predictive model. For the training part of our study, we used the HCSC-RAC (Madrid, Spain). A thorough description of the cohort, including inclusion and exclusion criteria, follow-up and clinical assessments can be found in the article of Abásolo *et al.*<sup>5</sup>. Briefly, this is a day-to-day clinical practice cohort that includes subjects that have received a clinical diagnosis of RA by their usual rheumatologist. In this cohort we have included those patients that a) are attending or have attended the rheumatology outpatient clinic of the *Hospital Clínico San Carlos* (Madrid, Spain), with at least two registered visits, b) have received any ICD9 and/or ICD10 codes for RA by their usual rheumatologist, at least in two consecutive visits, c) were 16 years old or older at symptoms onset, and d) RA diagnosis was established from January 1, 1994 to February 15, 2013. In the case that the patient also receives a diagnosis of other autoimmune disease (such as inflammatory bowel disease, psoriasis or psoriatic arthritis, systemic lupus erythematosus, scleroderma, juvenile idiopathic arthritis, or ankylosing spondylitis), either before being diagnosed of RA or after being included in the RA cohort, his/her clinical record

is reviewed by a rheumatologist (LA or LRR) that decides if the patient is included or excluded from the cohort, based on clinical, laboratory and treatment data. In addition to their routine clinical visits, patients included in the HCSC-RAC attend evaluation visits performed at baseline (when RA is diagnosed) and annually thereafter. In these visits, demographic, clinical and laboratory data is collected by a trained health professional evaluator. The present study was performed in a subset of the HCSC-RAC, selected based on a) the date of RA diagnosis and b) the length of follow-up. Because we wanted to minimize the missing information and to use data collected during the first two years after disease diagnosis, we included only those patients diagnosed in or after 2001, and those with at least 2 years of follow-up from the RA diagnosis.

For the external validation part of the study, we used the Hospital Universitario de La Princesa Early Arthritis Register Longitudinal (PEARL) study<sup>19</sup>. Briefly, this is an early arthritis cohort that includes patients diagnosed with RA<sup>57</sup> or chronic undifferentiated arthritis<sup>58</sup>. Patient a) attending or that have attended the rheumatology outpatient clinic of the *Hospital Universitario La Princesa* (Madrid, Spain), and b) with 1 or more swollen joints at presentation, for at least 4 weeks and symptoms for less than a year, are included in the PEARL study. Patients are excluded if they are diagnosed with other specific cause of arthritis at presentation or during follow-up (such as gouty arthritis, septic or viral arthritis, osteoarthritis, spondyloarthropathies, or connective tissue diseases). Patients started to be included in PEARL since 2000. Patients attend 5 structured visits (at baseline, 6, 12, 24 and 60 months) in which sociodemographic, clinical, laboratory, therapeutic, radiological data and biological samples are systematically collected by protocol. In order to get more reliable data, especially regarding joint counts, these visits are performed by two experienced rheumatologist, but there is no pre-established therapeutic protocol, so the decision on when and how to treat the patients relies on the responsible physicians from the department. In the present study, only patients diagnosed with RA were included.

Informed consent was obtained from all the patients. This study was conducted in accordance with the Declaration of Helsinki and Good Clinical Practices<sup>59</sup>, and study protocols were approved by the HCSC and Hospital Universitario de La Princesa Ethics Committee.

**Variables.** Our primary outcome was all cause mortality. For the HCSC-RAC, the date of death was obtained from the INDEF (*Índice Nacional de Defunciones*, Spanish for “National Mortality Index<sup>60</sup>), a national register depending on the Spanish Ministry of Health that records all deaths in the Spanish territory, regardless the nationality of the deceased (no cause of death is registered in this registry). The time of observation comprised the elapsed time between the date two years after the RA diagnosis and the date of patient’s death, or the date when mortality data was collected from the INDEF (September the 10, 2013).

For PEARL study, the date of death was obtained from HYGELA (the electronic clinical tool used at Hospital Universitario La Princesa) or, in case of patients’ loss of follow-up, from HORUS (an electronic health record that integrates information collected from primary care centres, outpatient clinics, and hospital admissions in the Madrid Region). Therefore, the time of observation comprised the elapsed time between the date two years after the RA diagnosis and the date of patient’s death, loss of follow-up, or January the 1st, 2017).

Regarding independent variables, we used demographic and clinical-related variables, including gender (dichotomic), age and calendar year at RA diagnosis (continuous), nationality (dichotomic: Spaniard, not Spaniard), RF presence (dichotomic: yes, no), ACPA (dichotomic: yes, no), use of any biological therapy during the first two years after disease diagnosis (dichotomic: yes, no), number of hospital admissions regardless the cause during the first two years after disease diagnosis (continuous), and median values of the HAQ<sup>61</sup> and of the ESR, performed at RA diagnosis, 1 and 2 years after (continuous)<sup>8</sup>. This information was obtained from a departmental electronic health record (Medi-LOG<sup>62</sup>), in the case of the HCSC-RAC, and from the PEARL database.

**Statistical analysis.** Continuous variables were described using median and IQR, or mean and standard deviation (SD), based on their distribution. Dichotomous and categorical variables were described using proportions.

Random Survival Forests were implemented using the R software package *randomForestSRC*<sup>63</sup>, developed by Ishwaran *et al.*<sup>13</sup>. Each run of RSF was performed based on 1,000 decision trees.

Regarding model training, the performance of the developed models was assessed using two measures: the prediction error<sup>64</sup> and the IBS<sup>65</sup>. The prediction error is a measure of the model’s discrimination ability. It is equal to  $1 - C\text{-index}$ <sup>13</sup>, which in turn is the probability that in two randomly selected pair of cases, the case with the shorter follow-up time has a worst predictive outcome<sup>66–68</sup>. The lower the prediction error (and therefore the higher the C-index), the better the model’s goodness of fit<sup>13</sup>. The IBS is a measure of the model’s accuracy, and it is calculated by squaring the differences between the patient primary outcome at a particular point in time (being alive or dead) and the predicted probability of this outcome at that time<sup>69</sup>. The lower the IBS, the better the model’s accuracy.

Those variables with more than 20% of missing data in the training cohort were excluded from the analysis (Table 1). For the rest of the variables, missing data was imputed using an iterative algorithm<sup>13</sup> supplied by the *randomForestSRC* package. As showed in other studies<sup>70</sup>, this algorithm appears to be reliable.

Internal validation was performed through bootstrapping, meaning that each of the decision trees that integrate the forest are created from a subset (in-bag data) of a bootstrapped sample from the original dataset (in our case the patients from the HCSC-RAC). The other subset (out-of-bag data) of the bootstrapped sample is used to calculate the prediction error<sup>71</sup>.

Two RSF models were constructed applying either the log-rank<sup>72,73</sup> ( $M_{LR}$ ) or the log-rank score<sup>74</sup> ( $M_{LRS}$ ) splitting rules. In order to obtain an unbiased measure of the models quality, we performed 100 iterations of each model, and we estimated the mean and SD of the prediction error. We selected the model with the lowest prediction error, and ranked the included variables according to their predictive capacity by an internal measure of variable importance (VIMP). The VIMP compares a variable’s predictive power to its power under randomness.

If the VIMP is close to zero, then the predictive capacity of that variable is lower, as the difference between the predictive power of the variable and its predictive power under randomness is small. Conversely, if the VIMP is large, then the predictive capacity of that variable is higher, as there is a greater difference between the predictive power of the variable and its predictive power under randomness. We calculated the mean (SD) VIMP based on 100 iterations of the model. In addition, we calculated for each variable their relative importance (RI), by dividing the mean VIMP score assigned to a specific variable by the mean VIMP score assigned to the first ranked variable. Based on the mean VIMP, we constructed a final model, excluding those variables with a negative VIMP<sup>75</sup>, and performed 100 iterations of the final model in order to estimate its prediction error, IBS, and VIMP of the included variables. We used partial plots to represent the effect on the predicted mortality of each variable included in the final model, after accounting for the average effects of the other variables.

Regarding external validation, the final model created with the patients from the HCSC-RAC was tested in a group of independent patients from the PEARL study, a different cohort of RA patients from a different centre, followed-up by different physicians, and collected and curated by different researchers. The observations from the PEARL cohort were dropped down the final RSF model and goodness-of-fit was assessed using the prediction error. Missing data was imputed using the iterative algorithm supplied by the *randomForestSRC* package<sup>13</sup>.

We also tested the performance of our final model by defining groups of patients based on their individual estimated mortality risks (i.e. groups with low, intermediate, and high risk) and then assessing and comparing the observed mortality among these groups. The out-of-bag mean cumulative hazard function<sup>22</sup> (referred to as ensemble mortality) estimated by the final RSF model was used as a measure of each patient's estimated mortality risk. For a particular patient, the ensemble mortality can be interpreted as the expected number of deaths in a cohort if all the patients had similar characteristics to those of this particular patient. The ensemble mortality cut-off values defining different risk groups were established using the data from the training cohort through a survival tree created with the R package *rpart* with default parameters<sup>76</sup>. These same cut-off values were used to define risk groups in the validation cohort. The survival probabilities along time for each mortality risk group were visually represented using Kaplan-Meier curves, and statistically compared with a bivariate Cox proportion hazard test, using the mortality risk group as a categorical variable and the lower risk category as reference. Differences in mortality risks among groups were expressed as hazard ratio (HR) and 95% confidence intervals (95% CI).

The relationship between time-dependent sensitivity/specificity and the ensemble mortality value for particular time points was estimated using the *survivalROC* R package<sup>77</sup>. Calibration was estimated using the *pec* R package<sup>78</sup>.

All analyses were performed by using R (version 3.3.2), the *randomForestSRC*, the *rpart*, the *timeROC*, and *survivalROC* the packages. A more detailed description of the statistical analysis can be found online in the Supplementary Methods).

**Sensitivity analysis.** Two new models were developed in the training cohort and compared with the final model, in terms of prediction error, IBS, and VIMP:

- a) A model including the variables from the final model and those excluded due to their high percentage of missing values (expanded model).
- b) A model including the variables from the final model except those with positive but low predictive capacity (RI < 1%; reduced model).

**Data availability.** The datasets analysed during the current study and the final Random Survival Forest Rheumatoid Arthritis Mortality Prediction Model are available from the corresponding author on reasonable request.

## References

1. Dadoun, S. *et al.* Mortality in rheumatoid arthritis over the last fifty years: systematic review and meta-analysis. *Joint. Bone. Spine* **80**, 29–33, doi:10.1016/j.jbspin.2012.02.005 (2013).
2. Myasoedova, E., Davis, J. M., Crowson, C. S. & Gabriel, S. E. Epidemiology of rheumatoid arthritis: Rheumatoid arthritis and mortality. *Current Rheumatology Reports* **12**, 379–385, doi:10.1007/s11926-010-0117-y (2010).
3. Widdifield, J. *et al.* Trends in excess mortality among patients with rheumatoid arthritis in ontario, Canada. *Arthritis Care Res. (Hoboken)*. **67**, 1047–53, doi:10.1002/acr.22553 (2015).
4. Radovits, B. J. *et al.* Excess mortality emerges after 10 years in an inception cohort of early rheumatoid arthritis. *Arthritis Care Res. (Hoboken)*. **62**, 362–70, doi:10.1002/acr.20105 (2010).
5. Abasolo, L. *et al.* Influence of demographic and clinical factors on the mortality rate of a rheumatoid arthritis cohort: A 20-year survival study. *Semin. Arthritis Rheum.* **45**, 533–8, doi:10.1016/j.semarthrit.2015.10.016 (2016).
6. Naz, S. M., Farragher, T. M., Bunn, D. K., Symmons, D. P. M. & Bruce, I. N. The influence of age at symptom onset and length of followup on mortality in patients with recent-onset inflammatory polyarthritis. *Arthritis Rheum.* **58**, 985–9, doi:10.1002/art.23402 (2008).
7. Pincus, T., Brooks, R. H. & Callahan, L. F. Prediction of long-term mortality in patients with rheumatoid arthritis according to simple questionnaire and joint count measures. *Ann. Intern. Med.* **120**, 26–34 (1994).
8. Wolfe, F., Michaud, K., Gefeller, O. & Choi, H. K. Predicting mortality in patients with rheumatoid arthritis. *Arthritis Rheum.* **48**, 1530–42, doi:10.1002/art.11024 (2003).
9. Turesson, C., O'Fallon, W. M., Crowson, C. S., Gabriel, S. E. & Matteson, E. L. Occurrence of extraarticular disease manifestations is associated with excess mortality in a community based cohort of patients with rheumatoid arthritis. *J. Rheumatol.* **29**, 62–7 (2002).
10. Book, C., Saxne, T. & Jacobsson, L. T. H. Prediction of mortality in rheumatoid arthritis based on disease activity markers. *J. Rheumatol.* **32**, 430–4 (2005).
11. Rodriguez-Rodriguez, L. *et al.* Treatment in rheumatoid arthritis and mortality risk in clinical practice: the role of biologic agents. *Clin. Exp. Rheumatol.* **34**, 1026–1032 (2016).
12. Kleinbaum, D. G. & Klein, M. *Survival Analysis. A Self-Learning Text*, Third Edition. (Springer New York, 2012).
13. Ishwaran, H., Kogalur, U. B., Blackstone, E. H. & Lauer, M. S. Random survival forests. *Ann. Appl. Stat* **2**, 841–860, doi:10.1214/08-AOAS169 (2008).



14. Radespiel-Tröger, M., Rabenstein, T., Schneider, H. T. & Lausen, B. Comparison of tree-based methods for prognostic stratification of survival data. *Artif. Intell. Med.* **28**, 323–341, doi:10.1016/S0933-3657(03)00060-5 (2003).
15. Hamidi, O., Poorolajal, J., Farhadian, M. & Tapak, L. Identifying important risk factors for survival in kidney graft failure patients using random survival forests. *Circ. Cardiovasc. Qual. Outcomes* **45**, 27–33, doi:10.1161/CIRCOUTCOMES.110.939371 (2016).
16. Churpek, M. M. *et al.* Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards. *Crit. Care Med.* **44**, 368–74, doi:10.1097/CCM.0000000000001571 (2016).
17. Yuan, Y. *et al.* Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nat. Biotechnol.* **32**, 644–652, doi:10.1038/nbt.2940 (2014).
18. Sloan, R. A. *et al.* A Fit-Fat Index for Predicting Incident Diabetes in Apparently Healthy Men: A Prospective Cohort Study. *PLoS One* **11**, e0157703, doi:10.1371/journal.pone.0157703 (2016).
19. González-Álvarez, I. *et al.* Interleukin 15 levels in serum may predict a severe disease course in patients with early arthritis. *PLoS One* **6**, e29492, doi:10.1371/journal.pone.0029492 (2011).
20. Ingrisch, M. *et al.* Radiomic Analysis Reveals Prognostic Information in T1-Weighted Baseline Magnetic Resonance Imaging in Patients With Glioblastoma. *Invest. Radiol.* **52**, 360–366, doi:10.1097/RLI.0000000000000349 (2017).
21. Jain, R. *et al.* Outcome prediction in patients with glioblastoma by using imaging, clinical, and genomic biomarkers: focus on the nonenhancing component of the tumor. *Radiology* **272**, 484–93, doi:10.1148/radiol.14131691 (2014).
22. Ruffalo, M. *et al.* Whole-exome sequencing enhances prognostic classification of myeloid malignancies. *J. Biomed. Inform.* **58**, 104–113, doi:10.1016/j.jbi.2015.10.003 (2015).
23. Wertheim, G. B. W. *et al.* Validation of DNA methylation to predict outcome in acute myeloid leukemia by use of xMELP. *Clin. Chem.* **61**, 249–258, doi:10.1373/clinchem.2014.229781 (2015).
24. Tremblay, C. S., Hoang, T. & Hoang, T. Early T cell differentiation: Lessons from T-cell acute lymphoblastic leukemia. *Prog. Mol. Biol. Transl. Sci.* **92**, 121–156, doi:10.1016/S1877-1173(10)92006-1 (2010).
25. Dal B, M. *et al.* CD49d prevails over the novel recurrent mutations as independent prognosticator of overall survival in chronic lymphocytic leukemia. *Leukemia*, doi:10.1038/leu.2016.88 (2016).
26. Manilich, E. A. *et al.* A novel data-driven prognostic model for staging of colorectal cancer. *J. Am. Coll. Surg.* **213**, 579–588, doi:10.1016/j.jamcollsurg.2011.08.006 (2011).
27. Banerjee, M., George, J., Song, E. Y., Roy, A. & Hryniuk, W. Tree-based model for breast cancer prognostication. *J. Clin. Oncol.* **22**, 2567–75, doi:10.1200/JCO.2004.11.141 (2004).
28. Gnep, K. *et al.* Haralick textural features on T2-weighted MRI are associated with biochemical recurrence following radiotherapy for peripheral zone prostate cancer. *Journal of Magnetic Resonance Imaging.* **45**, 103–117, doi:10.1002/jmri.25335 (2017).
29. Long, J. D. & Paulsen, J. S. Multivariate prediction of motor diagnosis in Huntington's disease: 12 years of PREDICT-HD. *Mov. Disord.* **30**, 1664–1672, doi:10.1002/mds.26364 (2015).
30. Ruysinck, J. *et al.* Random Survival Forests for Predicting the Bed Occupancy in the Intensive Care Unit. *Comput. Math. Methods Med.* **2016**, doi:10.1155/2016/7087053 (2016).
31. Simuni, T. *et al.* Predictors of time to initiation of symptomatic therapy in early Parkinson's disease. *Ann. Clin. Transl. Neurol.* **482–494** (2016).
32. Huber, A. M. *et al.* Early illness features associated with mortality in the juvenile idiopathic inflammatory myopathies. *Arthritis Care Res* **66**, 732–740, doi:10.1002/acr.22212 (2014).
33. Ward, M. M., Pajevic, S., Dreyfuss, J. & Malley, J. D. Short-term prediction of mortality in patients with systemic lupus erythematosus: classification of outcomes using random forests. *Arthritis Rheum.* **55**, 74–80, doi:10.1002/art.21695 (2006).
34. Wasko, M. C. M., Dasgupta, A., Hubert, H., Fries, J. F. & Ward, M. M. Propensity-adjusted association of methotrexate with overall survival in rheumatoid arthritis. *Arthritis Rheum.* **65**, 334–342, doi:10.1002/art.37723 (2013).
35. Chester Wasko, M., Dasgupta, A., Ilse Sears, G. & Fries, J. F. & Ward, M. M. Prednisone Use and Risk of Mortality in Patients With Rheumatoid Arthritis: Moderation by Use of Disease-Modifying Antirheumatic Drugs. *Arthritis Care Res. (Hoboken)* **68**, 706–710, doi:10.1002/acr.22722 (2016).
36. Carmona, L. *et al.* All-cause and cause-specific mortality in rheumatoid arthritis are not greater than expected when treated with tumour necrosis factor antagonists. *Ann. Rheum. Dis.* **66**, 880–5, doi:10.1136/ard.2006.067660 (2007).
37. Morgan, C. L. *et al.* Treatment of rheumatoid arthritis with etanercept with reference to disease-modifying anti-rheumatic drugs: long-term safety and survival using prospective, observational data. *Rheumatology (Oxford)* **53**, 186–94, doi:10.1093/rheumatology/ket333 (2014).
38. McMahon, A. D. & MacDonald, T. M. Design issues for drug epidemiology. *British Journal of Clinical Pharmacology* **50**, 419–425, doi:10.1046/j.1365-2125.2000.00289.x (2000).
39. Gonzalez, A. *et al.* The widening mortality gap between rheumatoid arthritis patients and the general population. *Arthritis Rheum.* **56**, 3583–7, doi:10.1002/art.22979 (2007).
40. Humphreys, J. H. *et al.* Mortality trends in patients with early rheumatoid arthritis over 20 years: results from the Norfolk Arthritis Register. *Arthritis Care Res. (Hoboken)* **66**, 1296–301, doi:10.1002/acr.22296 (2014).
41. Michet, C. J., Strobova, K., Achenbach, S., Crowson, C. S. & Matteson, E. L. Hospitalization rates and utilization among patients with rheumatoid arthritis: A population-based study from 1987 to 2012 in Olmsted County, Minnesota. *Mayo Clin. Proc.* **90**, 176–183, doi:10.1016/j.mayocp.2014.12.009 (2015).
42. Metsios, G. S. *et al.* Disease activity and low physical activity associate with number of hospital admissions and length of hospitalisation in patients with rheumatoid arthritis. *Arthritis Res. Ther.* **13**, R108, doi:10.1186/ar3390 (2011).
43. Ishwaran, H., Kogalur, U. B., Gorodeski, E. Z., Minn, A. J. & Lauer, M. S. High-Dimensional Variable Selection for Survival Data. *J. Am. Stat. Assoc.* **105**, 205–217, doi:10.1198/jasa.2009.tm08622 (2010).
44. Collins, G. S. *et al.* External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Med. Res. Methodol.* **14**, 40, doi:10.1186/1471-2288-14-40 (2014).
45. Provan, S., Angel, K., Semb, A. G., Atar, D. & Kvien, T. K. NT-proBNP predicts mortality in patients with rheumatoid arthritis: results from 10-year follow-up of the EURIDISS study. *Ann. Rheum. Dis.* **69**, 1946–50, doi:10.1136/ard.2009.127704 (2010).
46. Morisset, J. *et al.* The performance of the GAP model in patients with rheumatoid arthritis associated interstitial lung disease. *Respir. Med.* **127**, 51–56, doi:10.1016/j.rmed.2017.04.012 (2017).
47. Collins, G. S., Ogundimu, E. O. & Altman, D. G. Sample size considerations for the external validation of a multivariable prognostic model: A resampling study. *Stat. Med.* **35**, 214–226, doi:10.1002/sim.6787 (2016).
48. Vergouwe, Y., Steyerberg, E. W., Eijkemans, M. J. C. & Habbema, J. D. F. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J. Clin. Epidemiol.* **58**, 475–483, doi:10.1016/j.jclinepi.2004.06.017 (2005).
49. Ioannidis, J. P. A. Scientific inbreeding and same-team replication: Type D personality as an example. *J. Psychosom. Res.* **73**, 408–410, doi:10.1016/j.jpsychores.2012.09.014 (2012).
50. Hsieh, E., Gorodeski, E. Z., Blackstone, E. H., Ishwaran, H. & Lauer, M. S. Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. *Circ. Cardiovasc. Qual. Outcomes* **4**, 39–45, doi:10.1161/CIRCOUTCOMES.110.939371 (2011).
51. Datema, F. R. *et al.* Novel head and neck cancer survival analysis approach: Random survival forests versus cox proportional hazards regression. *Head Neck* **34**, 50–58, doi:10.1002/hed.21698 (2012).
52. Chen, X. & Ishwaran, H. Random forests for genomic data analysis. *Genomics* **99**, 323–329, doi:10.1016/j.ygeno.2012.04.003 (2012).

53. Yosefian, I., Mosa Farkhani, E. & Baneshi, M. R. Application of Random Forest Survival Models to Increase Generalizability of Decision Trees: A Case Study in Acute Myocardial Infarction. *Comput. Math. Methods Med.* **2015**, doi:10.1155/2015/576413 (2015).
54. Bou-Hamd, I., Larocque, D. & Ben-Ameur, H. A review of survival trees. *Stat. Surv* **5**, 44–71, doi:10.1214/09-SS047 (2011).
55. Walschaerts, M., Leconte, E. & Besse, P. Stable variable selection for right censored data: comparison of methods. *arXiv Prepr. arXiv1203.4928*, 1–29 (2012).
56. Austin, P. C., Lee, D. S., Steyerberg, E. W. & Tu, J. V. Regression trees for predicting mortality in patients with cardiovascular disease: What improvement is achieved by using ensemble-based methods? *Biometrical J.* **54**, 657–673, doi:10.1002/bimj.201100251 (2012).
57. Arnett, F. C. *et al.* The american rheumatism association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis Rheum.* **31**, 315–324, doi:10.1002/art.1780310302 (1988).
58. Verpoort, K. N. *et al.* Undifferentiated arthritis - Disease course assessed in several inception cohorts. *Clinical and Experimental Rheumatology* **22**, (2004).
59. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA: the journal of the American Medical Association* **310**, 2191–4, doi:10.1001/jama.2013.281053 (2013).
60. Ministerio de Sanidad, S. S. e I. National Mortality Index. Available at: [http://www.msssi.gob.es/estadEstudios/estadisticas/estadisticas/estMinisterio/IND\\_TipoDifusion.htm](http://www.msssi.gob.es/estadEstudios/estadisticas/estadisticas/estMinisterio/IND_TipoDifusion.htm).
61. Ramey, D. R., Raynauld, J. P. & Fries, J. F. The health assessment questionnaire 1992: status and review. *Arthritis Care Res* **5**, 119–29 (1992).
62. Leon, L. *et al.* Health-related quality of life as a main determinant of access to rheumatologic care. *Rheumatol. Int.* **33**, 1797–1804, doi:10.1007/s00296-012-2599-6 (2013).
63. Ishwaran, H. & Kogalur, U. Random Forests for Survival, Regression and Classification (RF-SRC). Available at: <https://cran.r-project.org/package=randomForestSRC>. (Accessed: 15th December 2016) (2016).
64. Fontana, A. *et al.* Development of a metabolites risk score for one-year mortality risk prediction in pancreatic adenocarcinoma patients. *Oncotarget* **7**, 8968–8978, doi:10.18632/oncotarget.7108 (2016).
65. Gerds, T. A. & Schumacher, M. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical J.* **48**, 1029–1040, doi:10.1002/bimj.200610301 (2006).
66. Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L. & Rosati, R. A. Evaluating the yield of medical tests. *JAMA* **247**, 2543–6, doi:10.1001/jama.1982.03320430047030 (1982).
67. Heagerty, P. J. & Zheng, Y. Survival model predictive accuracy and ROC curves. *Biometrics* **61**, 92–105, doi:10.1111/j.0006-341X.2005.030814.x (2005).
68. Zweig, M. H. & Campbell, G. Receiver-operating characteristic (ROC) plots: A fundamental evaluation tool in clinical medicine. *Clin. Chem.* **39**, 561–577 (1993).
69. Mogensén, U. B., Ishwaran, H. & Gerds, T. A. Evaluating Random Forests for Survival Analysis Using Prediction Error Curves. *J. Stat. Softw.* **30**, 1–3, doi:10.1126/scisignal.2001449.Engineering (2009).
70. Diouf, M. *et al.* Prognostic value of health-related quality of life in patients with metastatic pancreatic adenocarcinoma: a random forest methodology. *Qual. Life Res.* **25**, 1713–1723, doi:10.1007/s11366-015-1198-x (2016).
71. Bamba, S. *et al.* Predicting Mucosal Healing in Crohn's Disease Using Practical Clinical Indices with Regard to the Location of Active Disease. *Hepatogastroenterology* **61**, 689–696, doi:10.1111/codi.13414 (2014).
72. Segal, M. R. Regression Trees for Censored Data. *Biometrics* **44**, 35–47 (1988).
73. Leblanc, M. & Crowley, J. Survival Trees by Goodness of Split. *J Am Stat Assoc* **88**, 457–467 (1993).
74. Hothorn, T. & Lausen, B. On the exact distribution of maximally selected rank statistics. *Comput. Stat. Data Anal.* **43**, 121–137, doi:10.1016/S0167-9473(02)00225-6 (2003).
75. Vistisen, D. *et al.* Prediction of first cardiovascular disease event in type 1 diabetes mellitus the steno type 1 risk engine. *Circulation* **133**, 1058–1066, doi:10.1161/CIRCULATIONAHA.115.018844 (2016).
76. Therneau, T. M., Atkinson, B. & Ripley, B. rpart: Recursive Partitioning. (2011).
77. Heagerty, P. J. & Saha-Chaudhuri, P. survivalROC: Time-dependent ROC curve estimation from censored survival data. Available at: <https://cran.r-project.org/web/packages/survivalROC/index.html>. (Accessed: 1st June 2017) (2013).
78. Gerds, T. A. pec: Prediction Error Curves for Risk Prediction Models in Survival Analysis. Available at: <https://cran.r-project.org/web/packages/pec/index.html>. (Accessed: 1st June 2017) (2017).

## Acknowledgements

We want to thank the patients for making this study possible. This work was supported by the grants RD16/0012/0004 (Instituto de Salud Carlos III), and PI14/01007 (Instituto de Salud Carlos III; awarded to LA, IGA, ES). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## Author Contributions

J.M.L.V., L.R.R., J.A.J., and B.F.G. designed this study and prepared the manuscript. E.T., L.L., I.G.A., L.A., and L.R.R. collected and analyzed the clinical data. J.M.L.V., E.S., L.R.R., and F.S. performed statistical analysis. All authors reviewed the manuscript.

## Additional Information

**Supplementary information** accompanies this paper at doi:10.1038/s41598-017-10558-w

**Competing Interests:** The authors declare that they have no competing interests.

**Publisher's note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017