Supplementary materials and methods

Statistics

TRIPOD criteria were followed in our study and statistical analyses were performed with SPSS 22.0 (IBM Corp, Armonk, NY, USA) and STATA/IC 12.1 (StataCorp, College Station, TX, USA). Generation of the prognostic signatures was performed in the MDA cohort using regression Cox analysis by steps, where the contribution of every potential predictor to the model performance was assessed manually by the researcher. Thus, we eliminated redundant variables one by one according not only to their significance order (from lower to higher significance), but also avoiding the loss of more than 10% of the initial magnitude of the model Chi square. This method was repeated until we obtained 20 different candidate prognostic models (all of them clinically meaningful), stratifying by stage. For each model, we calculated the prognostic index (PI) for each patient, as the sum of the products of the B coefficients for each variable (X, Y, Z…) and the H-Score value: (H-Score X · Coefficient B X) + (H-Score Y · Coefficient B Y) + (H-Score Z · Coefficient B Z) + ... In order to measure the discrimination of the models we used the Harrell´s Concordance coefficient (C-index) by using a macro downloaded from this website: http://www-01.ibm.com/support/docview.wss?uid=swg21478383 (May 2015). In addition, the models were evaluated by Kaplan-Meier curves (log rank test); stratifying the PI by the median into two groups of risk (low and high). The final model was selected according to the principle of parsimony (the simplest explanation model –i.e. less number of possible variables) and a high C-index coefficient. The C-index, proposed by Harrell, estimates the probability of concordance between predicted and observed responses. It can be used to quantify the predictive discrimination of any quantitative predictive method, whether the response is continuous, ordinal, or binary. A value of 0.5 indicates no predictive discrimination and a value of 1.0

indicates perfect separation of patients with different outcomes. The prognostic model was internally validated to quantify any optimism in the predictive performance through a shrinkage penalization strategy. We used the bootstrapping method to generate 100 bootstrap samples. We applied the model and calculated the 100 PIs and the 100 C-indexes in the bootstrap samples (observed C-index). The PI obtained in each bootstrap sample was applied to the original cohort and a new C-index was then calculated (validated C-index). We penalized the observed C-index subtracting the validated C-index values and we calculated the average, named as Shrinkage. The difference between the original C-index and the shrinkage yielded the adjusted C-index. The proportional hazards assumption was examined by testing interactions between the co-variables of the final model and time. Univariate and multivariate Cox proportional hazards analyses including clinical and pathological variables were used to assess the prognostic role of the molecular model (PI). Only variables with $P<0.25$ in the univariate analysis were included in the multivariate analysis. The external validation of the prognostic models was performed in the second cohort (CIBERES-CUN). We calculated the C-index and the survival curves with the Kaplan-Meier method methods, which differences were compared using log-rank test as previously described. Clinical utility of our models was tested by comparing the likelihood ratio of the stage itself to that after the addition of the molecular model (PI) through a bivariate Cox analysis in the MDA cohort. Moreover, we developed a new combined prognostic model (CPI) by adding the pathological stage and the molecular data following this formula: (B Coefficient PI x PI) + B Coefficient Stage; depending on the stage of each patient. We calculated the C-index coefficient of the new model and calibrated it with survival curves as previously described. The combined model was validated in CIBERES-CUN cohort. A predictive study was performed with patients from the MDA cohort, depending on whether they received or not adjuvant therapy.

We separated stage I-IIA patients into two groups according the CPI median. Then we analyzed the survival differences between patients treated or not treated with chemotherapy using the Kaplan-Meier method and log rank test.