ESC European Society of Cardiology

European Journal of Heart Failure (2021) **23**, 2035–2044
doi:10.1002/ejhf.2352

RESEARCH ARTICLE

Head-to-head comparison of contemporary heart failure risk scores

**Pau Codina[1,2], Josep Lupón[1,2,3], Andrea Borrellas[1], Giosafat Spitaleri[1], Germán Cediel[1,2], Mar Domingo[1], Joanne Simpson[4], Wayne C. Levy[5], Evelyn Santiago-Vacas[1,3], Elisabet Zamora[1,2,3], David Buchaca[6], Isaac Subirana[7], Javier Santesmases[1,2], Crisanto Diez-Quevedo[1], Maria I. Troya[1], Maria Boldo[1], Salvador Altmir[1], Nuria Alonso[1], Beatriz González[1], Carmen Rivas[1], Julio Nuñez[3,8,9], John McMurray[4], and Antoni Bayes-Genis[1,2,3]*

[1]Heart Failure Clinic and Cardiology Service, University Hospital Germans Trias i Pujol, Badalona, Spain; [2]Department of Medicine, Universitat Autonoma de Barcelona, Barcelona, Spain; [3]CIBERCV, Instituto de Salud Carlos III, Madrid, Spain; [4]British Heart Foundation Cardiovascular Research Centre, University of Glasgow, Glasgow, UK; [5]UW Medicine Heart Institute, University of Washington, Seattle, WA, USA; [6]Barcelona Supercomputing Center, Barcelona, Spain; [7]Hospital del Mar Medical Research Institute (IMIM), Barcelona, Spain; [8]Cardiology Department. Hospital Clínico Universitario, INCLIVA Valencia, Valencia, Spain; and [9]Departament of Medicine, Universidad de Valencia, Valencia, Spain

Received 2 March 2021; revised 14 September 2021; accepted 20 September 2021; online publish-ahead-of-print 1 October 2021

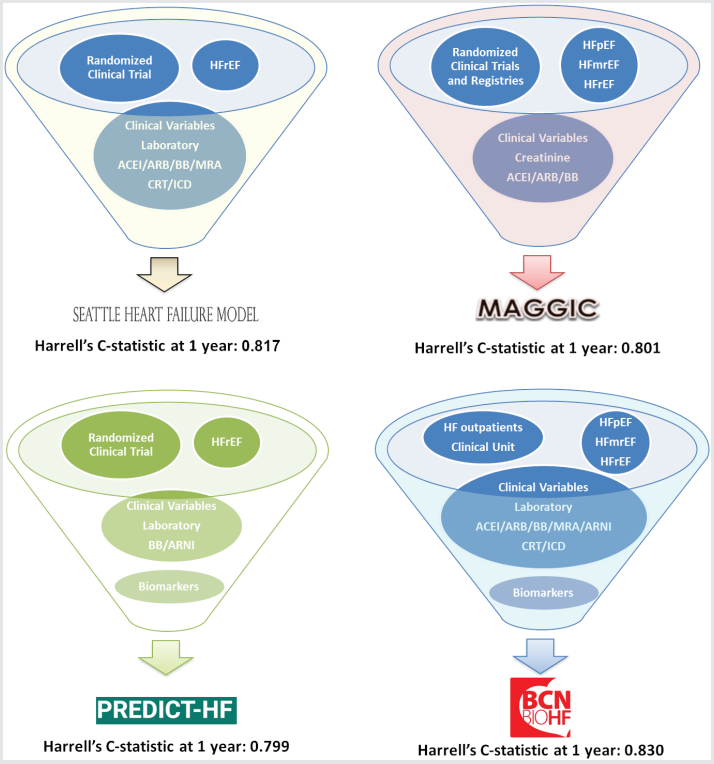
Aims Several heart failure (HF) web-based risk scores are currently used in clinical practice. Currently, we lack head-to-head comparison of the accuracy of risk scores. This study aimed to assess correlation and mortality prediction performance of Meta-Analysis Global Group in Chronic Heart Failure (MAGGIC-HF) risk score, which includes clinical variables + medications; Seattle Heart Failure Model (SHFM), which includes clinical variables + treatments + analytes; PARADIGM Risk of Events and Death in the Contemporary Treatment of Heart Failure (PREDICT-HF) and Barcelona Bio-Heart Failure (BCN-Bio-HF) risk calculator, which also include biomarkers, like N-terminal pro B-type natriuretic peptide (NT-proBNP).

Methods and results A total of 1166 consecutive patients with HF from different aetiologies that had NT-proBNP measurement at first visit were included. Discrimination for all-cause mortality was compared by Harrell's C-statistic from 1 to 5 years, when possible. Calibration was assessed by calibration plots and Hosmer–Lemeshow test and global performance by Nagelkerke's R^2. Correlation between scores was assessed by Spearman rank test. Correlation between the scores was relatively poor (rho value from 0.66 to 0.79). Discrimination analyses showed better results for 1-year mortality than for longer follow-up (SHFM 0.817, MAGGIC-HF 0.801, PREDICT-HF 0.799, BCN-Bio-HF 0.830). MAGGIC-HF showed the best calibration, BCN-Bio-HF overestimated risk while SHFM and PREDICT-HF underestimated it. BCN-Bio-HF provided the best discrimination and overall performance at every time-point.

Conclusions None of the contemporary risk scores examined showed a clear superiority over the rest. BCN-Bio-HF calculator provided the best discrimination and overall performance with overestimation of risk. MAGGIC-HF showed the best calibration, and SHFM and PREDICT-HF tended to underestimate risk. Regular updating and recalibration of online web calculators seems necessary to improve their accuracy as HF management evolves at unprecedented pace.

*Corresponding author. Heart Institute, Hospital Universitari Germans Trias i Pujol, Department of Medicine, UAB, Carretera del Canyet s/n, 08916 Badalona, Spain. Tel: +34 934 978915, Fax: +34 934 978939, Email: abayesgenis@gmail.com

Graphical Abstract



Head-to-head comparison of contemporary heart failure risk scores. ACEI, angiotensin-converting enzyme inhibitor; ARB, angiotensin II receptor blocker; ARNI, angiotensin receptor–neprilysin inhibitor; BB, beta-blocker; CRT, cardiac resynchronization therapy; ICD, implantable cardioverter-defibrillator; HFmrEF, heart failure with mildly reduced ejection fraction; HFpEF, heart failure with preserved ejection fraction; HFrEF, heart failure with reduced ejection fraction; MRA, mineralocorticoid receptor antagonist.

Keywords Heart failure • Mortality • Risk models • Risk prediction

Introduction

Scales for predicting death in patients with heart failure (HF) are continuously evolving, and risk prediction is a cornerstone of HF management. Over the last three decades, several web-based scores have been devised to aid clinicians in assessing patient prognosis, and ultimately, improving the appropriateness and timing of disease-modifying treatments.

Among contemporary risk scores, the Seattle Heart Failure Model (SHFM), published in 2006, was the pioneer, and it has been extensively validated.[1] In the last decade of the 20th century, the SHFM was derived using the data previously collected in the Prospective Randomized Amlodipine Survival Evaluation (PRAISE-1) cohort ($n = 1125$), and thereafter validated using five other cohorts ($n = 9942$) with predominantly reduced left ventricular ejection fraction (LVEF). PRAISE-1 was a randomized trial of amlodipine vs. placebo among patients in the United States and Canada with LVEF < 30% and New York Heart Association (NYHA) functional class IIIB to IV. The SHFM incorporates clinical variables, medical treatment (excluding sacubitril/valsartan), and conventional laboratory analytes (e.g. haemoglobin, lymphocyte count, uric acid, total cholesterol, and sodium). Remarkably, the relative effects of HF medications could not be obtained from the derivation cohort; instead, the benefits were estimated from published trials or meta-analyses. Model discriminant ability was determined by the 1-year receiver operating characteristic area under the curve (AUC). The 1-year AUC for PRAISE-1, the derivation cohort, was 0.725 [95% confidence interval (CI) 0.69–0.76]. The SHFM online calculator was updated in 2013 with a modification of the baseline survival function to provide a more accurate estimate of 2–5 year survival.

The Meta-Analysis Global Group in Chronic Heart Failure (MAGGIC-HF) score was published in 2012. The MAGGIC-HF score was derived around the turn of the 20th century, based

on a large, heterogeneous cohort using individual data on 39 372 patients with HF, both reduced and preserved LVEF, from 30 cohort studies, six of which were clinical trials. Thirteen independent variables were identified to predict all-cause mortality at 1 and 3 years,[2] including age, LVEF, NYHA class, serum creatinine, diabetes, systolic blood pressure, body mass index, HF duration, current smoking, chronic obstructive pulmonary disease (COPD), male sex, and treatment with beta-blockers, angiotensin-converting enzyme inhibitors (ACEI), and angiotensin II receptor blockers (ARBs). However, the MAGGIC-HF score does not incorporate mineralocorticoid receptor antagonists (MRAs), angiotensin receptor–neprilysin inhibitors (ARNIs), or biomarkers. Discrimination of the model in the derivation cohort was not available. However, the score was validated in 51 043 patients from the national Swedish Heart Failure Registry (SwedeHF), obtaining a C-index of 0.741 at 3 years.[3]

The PARADIGM Risk of Events and Death in the Contemporary Treatment of Heart Failure (PREDICT-HF) was developed using data (including natriuretic peptides) from the Prospective Comparison of ARNI With ACEI to Determine Impact on Global Mortality and Morbidity in Heart Failure (PARADIGM-HF) trial patient cohort ($n = 8399$). Patients were eligible at screening if they had NYHA class II–IV, LVEF ≤35% and had an elevated natriuretic peptide level. The model was validated using the Aliskiren Trial to Minimize Outcomes in Patients with Heart Failure Trial (ATMOSPHERE) study and the SwedeHF.[4] The cohorts were treated with contemporary evidence-based treatment and were geographically representative, even though it did not incorporate MRAs, cardiac resynchronization therapy (CRT) or implantable cardioverter-defibrillators (ICD). The Harrell's C-statistic was used to assess the discriminative ability of the model. For all-cause death, the C-statistic for 1 and 2 years was 0.71 (95% CI 0.69–0.74) and 0.70 (95% CI 0.67–0.72), respectively.

The Barcelona Bio-Heart Failure (BCN-Bio-HF) risk calculator was derived from consecutive outpatients ($n = 864$). The principal referral criterion was HF according to the European Society of Cardiology (ESC) guidelines irrespective of aetiology with at least one HF hospitalization or LVEF <40%. The first version of the calculator was published in 2014. It included clinical variables, medications, conventional laboratory analytes (sodium, estimated glomerular filtration rate), and the following biomarkers: N-terminal pro B-type natriuretic peptide (NT-proBNP), high-sensitivity troponin T (hs-TnT), and interleukin-1 receptor-like-1 (known as ST2). The BCN-Bio-HF calculator can operate with none, one, two, or all three biomarkers.[5] The BCN-Bio-HF was externally validated with the PROTECT cohort (Boston)[6] and updated in 2018 by incorporating the use of ARNIs, CRT, and ICD, being this version validated with the PARADIGM-HF cohort.[7] The C-statistic taking into account time to-event up to 5 years with the model containing NT-proBNP was 0.774, while AUC were 0.826 for 1-year, 0.809 for 2-year, 0.823 for 3-year and 0.838 for 5-year mortality, when considering death as a binary event. Of the three biomarkers, only NT-proBNP is routinely available in most HF clinics, and it is the only one we included in the current analyses.

Despite consistent evidence that has linked NT-proBNP to an increased risk of all-cause mortality in patients with HF,[8,9] this biomarker was not included in MAGGIC-HF or SHFM. In the present study, we aimed to perform head-to-head comparisons of the MAGGIC-HF, SHFM, PREDICT-HF and BCN-Bio-HF, by determining the correlation between them and comparing discrimination capacity for mortality risk prediction.

Methods

Study population and follow-up

All consecutive ambulatory patients with HF from different aetiologies that were admitted to a structured multidisciplinary HF clinic at a university hospital between July 2010 and December 2018 were eligible. The patients that participated previously in the derivation cohort (May 2006–July 2010) were not included in the present study. To be included the patients had to have undergone a measurement of NT-proBNP at the first visit. Patients were referred to the HF clinic mostly by the cardiology or internal medicine departments and to a lesser extent by the emergency or other hospital departments. The criteria for referral to the HF clinic were HF according to the ESC definition, with at least one hospitalization and/or reduced systolic function, as described previously.[10,11] For follow-up, all patients regularly visited the HF clinic according to their clinical needs, and they were treated according to a unified protocol.

Follow-up visits comprised a minimum of quarterly visits with a nurse, one visit with a physician (cardiologist, internist, or family physician) every 6 months, and optional visits with specialists in geriatrics, psychiatry, rehabilitation, endocrinology, or nephrology.

During the baseline visit, patients provided written consent for the use of their clinical data for research purposes. Demographic, clinical, echocardiographic, and analytical data were recorded in a specific database (REGI-UNIC). Data that were not routinely recorded in that database were obtained by reviewing the electronic patient health records. We excluded patients that lacked data on more than three of the variables (five for PREDICT-HF) in any of the risk estimation tools.

In the present study we used the updated version of the BCN-Bio-HF calculator, which incorporated three new clinical variables (duration of HF in months, number of HF-related hospitalizations in the preceding year, and diabetes mellitus) and four new treatments (MRA, ARNI, CRT, and ICD) to the original variables. In the updated version CRT, ICD, diabetes and MRA became significant predictors. Beta values for ARNI were derived from the PARADIGM-HF trial.

Outcomes

Fatal events were identified by reviewing the patient health records from hospital wards, emergency room, and general practitioners, or by contacting their relatives. Data were verified with the databases of the Catalan and Spanish Health Systems, and also, with the Spanish National Death Index. In only a very small number of patients we needed to contact with their relatives and in those cases it was to precise the exact date and cause of death (only when in the medical records the exact date and cause of death were not reported). Annual all-cause death was the main endpoint for comparing the predictive abilities of the different risk calculators. Follow-up was closed on the 30th September 2020. Observed overall mortality at 1, 3, and 5 years

for all models was calculated using the Kaplan–Meier estimate of survival.

The study was performed in compliance with the laws that protect personal data, in accordance with the international guidelines on clinical investigations from the World Medical Association's Declaration of Helsinki. The local ethics committee approved the study.

NT-proBNP levels were determined in plasma samples with an immuno-electrochemiluminescence assay on the Modular Analytics E 170 instrument (Roche Diagnostics). This assay had <0.001% cross-reactivity with bioactive B-type natriuretic peptide. In the studies included in this report, the assay had inter-run coefficients of variation ranging from 0.9% to 5.5%.

Statistical analysis

Categorical variables are expressed as absolute numbers and percentages. Continuous variables are expressed as the mean ± standard deviation or the median and interquartile range (Q1 to Q3), according to normal or non-normal data distributions. Normal distributions were assessed with normal Q–Q plots. Comparisons between groups were performed with the chi-square and Fisher's test for categorical variables, and the Student's *t*-test or Mann–Whitney U test for continuous variables, as appropriate.

Missing values were treated by imputing either the default values of the calculators or the median values.

The discrimination abilities of the different prediction tools were compared with Harrell's C-index for 1-, 2-, 3-, 4- and 5-year all-cause mortality. The Harrell's C-index is a measure of the goodness of fit for models that produce risk scores in survival analyses, where data may be censored; thus, Harrell's C-index takes into account the time to event. It measures the ability of the model to discriminate between patients that will and patients that will not experience the event. AUCs considering death as a binary event at every time-point were also assessed in a sensitivity analysis. Calibration was assessed comparing the predicted and observed survival with the Kaplan–Meier estimate of survival. The D'Agostino–Nam version of the Hosmer–Lemeshow test using deciles of risk at all time-points of the study was used, and calibration plots comparing observed vs. expected events based on estimated risk by each calculator also by deciles, with the incorporation of LOWESS curves, which allow for the assessment of calibration at individual level, were plotted. Finally, overall performance of the tools was evaluated with the Royston modification of Nagelkerke's R^2 statistic for the same time-points at which mortality was recorded. Nagelkerke's R^2 statistic summarizes the proportion of variance in the dependent variable associated with the independent variables. Larger R^2 values indicated that more of the variation could be explained by the model, and the maximum R^2 value was 1.

Several sensitivity analysis were performed: (i) the performance of the four calculators was assessed in patients with reduced LVEF; (ii) the performance of the different risk prediction tools was assessed applying the median of our cohort to all missing variables; (iii) a sub-analysis in 421 patients without missing values (except allopurinol for SHFM) was performed for SHFM, MAGGIC-HF and BCN-Bio-HF; (iv) the performance of the BCN-Bio-HF model with NT-proBNP was compared with the model without NT-proBNP; (v) a sub-analysis of the BCN-Bio-HF in the last 413 patients in which ST2 and hs-TnT were available was performed, comparing its results with those of the model containing only NT-proBNP (up to 4 years); and (vi) an analysis of the BCN Bio-HF with the three biomarkers after recalculation of beta-coefficients in this more contemporary cohort.

Statistical analyses were performed with STATA V.15.1 software (StataCorp., College Station, TX, USA) and R software (A Language and Environment for Statistical Computing), distributed by the R Core Team (2017; R 9 Foundation for Statistical Computing, Vienna, Austria). A two-sided $P < 0.05$ was considered significant.

Results

Out of 1267 consecutive patients admitted to the HF clinic during the inclusion period, 22 lacked an NT-proBNP measurement, 30 were excluded because they had more than three missing values in MAGGIC, SHFM or BCN-Bio-HF and 49 because they had more than five missing values in PREDICT-HF. Our final cohort included 1166 patients (online supplementary *Figure S1*). *Table 1* provides the baseline demographic, clinical, biochemical, and echocardiographic characteristics and treatments of the studied cohort. Online supplementary *Table S1* compares included vs. excluded patients.

The included patients were predominantly men, aged 65.9 ± 13.4 years, with reduced LVEF (37.7 ± 14.7%), and they were mostly classified as NYHA class II (71.8%). Ischaemic heart disease was the most prevalent aetiology (40.7%), followed by dilated cardiomyopathy (19.5%). Contemporary HF treatments were optimized according to international guidelines.[12] A comparison was performed between surviving and non-surviving groups at 1 year (*Table 1*).

Online supplementary *Table S2* shows the number of missing values for each calculator and default or median values used for management of such missing values.

Follow-up was extended up to 5 years; the mean follow-up for surviving patients was 4.2 ± 1.1 years, and the median was 5 years. We recorded 358 deaths. Mortality rates at 1, 2, 3, 4 and 5 years were 9.5%, 16.4%, 21.9%, 30.6% and 35.8%, respectively.

Online supplementary *Figure S2* shows scatter plots between the risks of death at 1 year estimated by the different calculators.. Correlation between the scores was relatively poor (rho value from 0.66 to 0.79).

The BCN-Bio-HF improved the discrimination of all-cause mortality risk, based on Harrell's C-index. Moreover, the BCN-Bio-HF showed the highest overall performance at every time-point that risk was estimated (*Table 2*). *Table 3* provides observed overall survival vs. the predicted by each model, while *Figure 1* shows calibration plots with LOWESS lines of expected and observed mortality at 1 year by risk deciles for every risk prediction tool and the results of the Hosmer–Lemeshow test; these figures at 2, 3 and 5 years are shown in online supplementary *Figures S3–S5*. The BCN-Bio-HF systematically overestimated risk whereas SHFM and PREDICT-HF underestimated the risk of death. MAGGIC-HF showed the best calibration, although also moderately overestimated risk of death. Finally, *Figure 2* shows survival curves based on quintiles of risk estimation by every tool.

In the sensitivity analyses, (i) the BCN-Bio-HF model remained with best performance in patients with reduced LVEF (online supplementary *Table S3*); (ii) the performance of the different risk prediction tools applying our cohorts' median to all missing variables tended to show similar general results, with improvement of

Table 1 Baseline population characteristics and comparison between alive and death groups at the end of the follow-up period

Characteristic	Total cohort (n = 1166)	Alive at end of follow-up (n = 808)	Dead at end of follow-up (n = 358)	P-value*
Age, years	65.9 ± 13.4	62.5 ± 13.2	73.7 ± 10.4	<0.001
Male sex	824 (70.7)	574 (71.0)	250 (69.8)	0.705
BMI, kg/m^2	26.9 [24.3–30.1]	27.1 [24.6–30.4]	26.5 [23.6–29.2]	0.011
Ischaemic aetiology	475 (40.7)	289 (35.8)	186 (51.2)	<0.001
Heart failure duration, months	6 [1–36]	5 [1–28]	12 [2–48]	0.006
Hypertension	771 (66.1)	487 (60.3)	284 (79.3)	<0.001
Diabetes	516 (44.3)	318 (39.4)	198 (55.3)	<0.001
Current smoker	208 (17.8)	175 (21.7)	33 (9.2)	<0.001
COPD	176 (15.1)	93 (11.5)	83 (23.2)	<0.001
Region				
Central Europe	12 (1.0)	8 (1.0)	4 (1.1)	0.871
Latin America	2 (0.2)	1 (0.1)	1 (0.3)	0.559
Race/ethnicity (Asian)	2 (0.2)	1 (0.1)	1 (0.3)	0.495
Systolic BP	128.6 ± 21.6	128.0 ± 20.6	130.2 ± 23.7	0.238
NYHA functional class				
I	103 (8.8)	99 (12.3)	4 (1.1)	<0.001
II	837 (71.8)	620 (76.7)	217 (60.6)	<0.001
III	223 (19.1)	89 (11.0)	134 (37.4)	<0.001
IV	3 (0.3)	0 (0)	3 (0.8)	0.006
Atrial fibrillation/flutter	265 (22.7)	162 (20.0)	103 (28.8)	0.001
LVEF, %	37.7 ± 14.7	37.0 ± 14.2	39.2 ± 15.6	<0.001
Blood tests				
Haemoglobin, g/dL	12.9 ± 1.9	13.3 ± 1.8	11.9 ± 1.7	<0.001
Lymphocytes, %	21.1 [16.0–27.1]	22.4 [17.2–28.3]	18.7 [13.0–24.0]	<0.001
Monocytes, %	8.4 [7.1–10.1]	8.4 [7.1–10.0]	8.5 [7.0–10.2]	0.329
Absolute neutrophils, /mm^3	4800 [3800–6200]	4800 [3700–6000]	4900 [3800–6600]	0.021
Sodium, mmol/L	137.6 ± 3.5	137.7 ± 3.2	137.5 ± 3.9	0.052
Potassium, mmol/L	4.3 ± 0.5	4.3 ± 0.5	4.3 ± 0.6	0.288
Chloride, mmol/L	N/A	N/A	N/A	
Uric acid, µmol/L	438 ± 139.7	427.4 ± 135.3	464.0 ± 147.1	<0.001
eGFR, mL/min/1.73 m^2	61.8 ± 28.5	68.0 ± 27.5	47.9 ± 25.5	<0.001
BUN, mmol/L	4.2 [3.1–6.2]	3.9 [2.9–5.4]	5.6 [3.9–8.6]	<0.001
Total cholesterol, mmol/L	4.11 ± 1.16	4.22 ± 1.19	3.86 ± 1.05	<0.001
LDL, mmol/L	2.47 ± 1.02	2.57 ± 1.05	2.19 ± 0.90	<0.001
Triglycerides, mmol/L	1.43 ± 0.76	1.45 ± 0.74	1.37 ± 0.81	0.137
AST, IU/L	23 [17–35]	24 [18–36]	21 [15–31]	0.031
Bilirubin, µmol/L	N/A	N/A	N/A	
Albumin, g/L	37.7 ± 5.5	38.0 ± 5.1	37.0 ± 6.3	<0.001
NT-proBNP, pg/mL	1700 [707–4238]	1310 [532–2875]	3580 [1510–7580]	<0.001
Treatments (at baseline)				
Beta-blocker	943 (80.9)	676 (83.7)	267 (74.6)	<0.001
ACEI/ARB/ARNI	799 (68.5)	596 (73.8)	203 (56.7)	<0.001
Loop diuretics				
Furosemide >40 mg/day	622 (53.3)	409 (50.6)	213 (59.5)	0.001
Furosemide ≤40 mg/day	544 (46.7)	399 (49.4)	145 (40.5)	0.001
Weight-adjusted diuretic dose, mg/kg	0.47 [0.18–0.59]	0.45 [0.16–0.57]	0.54 [0.28–0.63]	<0.001
MRA	555 (47.6)	401 (49.6)	154 (43.0)	0.08
CRT	58 (5.0)	44 (5.4)	14 (3.9)	0.165
ICD	105 (9.0)	86 (10.6)	19 (5.3)	0.002
Allopurinol	N/A	N/A	N/A	

Values are mean ± standard deviation, n (%), or median [interquartile range].
ACEI, angiotensin-converting enzyme inhibitor; ARB, angiotensin II receptor blocker; ARNI, angiotensin receptor–neprilysin inhibitor; AST, aspartate aminotransferase; BMI, body mass index; BP, blood pressure; BUN, blood urea nitrogen; COPD, chronic obstructive pulmonary disease; CRT, cardiac resynchronization therapy; eGFR, estimated glomerular filtration rate; ICD, implantable cardioverter-defibrillator; LDL, low-density lipoprotein; LVEF, left ventricular ejection fraction; MRA, mineralocorticoid receptor antagonist; N/A, not available; NT-proBNP, N-terminal pro B-type natriuretic peptide; NYHA, New York Heart Association.
*Based on Cox regression analysis.

Table 2 Performance of the different risk prediction tools for all-cause mortality

	1-year mortality		2-year mortality		3-year mortality		4-year mortality		5-year mortality	
	C-index	R^2	C-index	R^2	C-index	R^2	C-index	R^2	C-index	R^2
SHFM	0.817 (0.782–0.851)	0.45	0.765 (0.732–0.798)	0.42	0.753 (0.722–0.785)	0.44	0.740 (0.711–0.768)	0.41	0.742 (0.714–0.770)	0.41
MAGGIC-HF	0.801 (0.761–0.841)	0.55	–	–	0.758 (0.727–0.789)	0.47	–	–	–	–
PREDICT-HF	0.799 (0.761–0.837)	0.25	0.739 (0.704–0.773)	0.22	–	–	–	–	–	–
BCN-Bio-HF	0.830 (0.797–0.864)	0.66	0.788 (0.757–0.818)	0.56	0.782 (0.753–0.811)	0.53	0.772 (0.745–0.798)	0.46	0.771 (0.743–0.798)	0.44

BCN-Bio-HF, Barcelona Bio-Heart Failure risk calculator; C-index, Harrell's C-index for evaluating discrimination ability; MAGGIC-HF, Meta-Analysis Global Group in Chronic Heart Failure risk score; PREDICT-HF, PARADIGM Risk of Events and Death in the Contemporary Treatment of Heart Failure; R^2, Royston modification of Nagelkerke's R^2 statistic; SHFM, Seattle Heart Failure Model.
Statistical comparison (Harrell's C-index): At 1 year: *P*-value BCN-Bio-HF vs. SHFM 0.28; vs. MAGGIC = 0.054; vs. PREDICT-HF = 0.056/*P*-value SHFM vs. MAGGIC = 0.40; vs. PREDICT-HF = 0.30/*P*-value MAGGIC-HF vs. PREDICT-HF = 0.91. At 2 years: *P*-value BCN-Bio-HF vs. SHFM 0.045; vs. PREDICT-HF <0.001/*P*-value SHFM vs. PREDICT-HF = 0.082. At 3 years: *P*-value BCN-Bio-HF vs. SHFM 0.006; vs. MAGGIC-HF = 0.53/*P*-value SHFM vs. MAGGIC-HF = 0.69. At 4 years: *P*-value BCN-Bio-HF vs. SHFM 0.002. At 5 years: *P*-value BCN-Bio-HF vs. SHFM 0.009.

SHFM and PREDICT-HF (online supplementary *Tables S4* and *S5*); (iii) the sub-analysis in 421 patients without missing values showed somewhat better performance of the studied calculators (online supplementary *Tables S6* and *S7*); (iv) the BCN-Bio-HF model used in the present study containing NT-proBNP, compared with the model of the same tool without NT-proBNP, numerically increased C-statistic at every time-point and reached statistical significance at 1 and 4 years, while patients were significantly reclassified with the biomarker addition (online supplementary *Table S8*); (v) the BCN-Bio-HF model including also ST2 and hs-TnT performed in 413 patients provided minor discrimination improvement and less risk overestimation (online supplementary *Table S9*); (vi) the BCN-Bio-HF analysis after recalibration of beta-coefficients in this more contemporary cohort, showed better results, mainly on calibration (online *supp*lementary *Tables S10* and *S11* and online supplementary *Figures S6* and *S7*).

Discussion

This study compared the performances of the MAGGIC-HF, SHFM, PREDICT-HF and BCN-Bio-HF scores (the two latter included NT-proBNP) for predicting mortality in 1166 HF outpatients managed in a multidisciplinary HF clinic. None of the contemporary risk scores examined showed a clear superiority over the rest. The BCN-Bio-HF calculator, which included NT-proBNP, provided good discrimination and overall performance, but overestimated risk at all time-points, particularly for patients at high risk of death. In contrast, SHFM and PREDICT-HF tended to underestimate the risk of death. MAGGIC-HF showed good calibration with modest overestimation of risk. Our results support the routine use of natriuretic peptides in risk stratification tools for HF and eventually the inclusion of other biomarkers such as ST2 and hs-TnT to better refine risk of death.

Importantly, no patient involved in the development of the BCN-Bio-HF (derivation cohort) was included in the current study.

Risk prediction models are frequently used in HF to aid clinicians in assessing patient prognosis. Ultimately, they improve the appropriateness and timing of disease-modifying treatments. These tools are also valuable for comparing risk between populations from different studies. Over the last three decades, several multivariate risk models have been devised, though only some are available on the web and are currently used in clinical practice.

The widely used SHFM includes important prognostic variables; however, some of these variables are not readily available in routine clinical practice. Indeed, in some of the validation cohorts for the SHFM, up to 65% of uric acid levels and up to 100% of lymphocyte values were missing. Moreover, in the SHFM derivation cohort, only 3% of patients were taking an MRA, and none were prescribed a beta-blocker. Consequently, the relative effects of HF medications could not be analysed in the derivation cohort. Instead, the drug benefits were estimated from published trials or meta-analyses, which usually included selected patients. Therefore, the effects of these medications may actually differ in the routine management of patients with HF. The update of this calculator performed in 2013 undoubtably has improved its performance in more contemporary

Table 3 Observed vs. predicted overall survival

	Observed[a]	SHFM[b]	MAGGIC-HF[b]	PREDICT-HF[b]	BCN-Bio-HF[b]
1-year survival	90.5%	94.0%	85%	94.1%	80.6%
2-year survival	83.6%	85.5%	–	89.1%	66.4%
3-year survival	78.1%	83.2%	67.2%	–	55.5%
4-year survival	69.4%	78.1%	–	–	47.1%
5-year survival	64.2%	73.2%	–	–	39.3%

BCN-Bio-HF, Barcelona Bio-Heart Failure risk calculator; MAGGIC-HF, Meta-Analysis Global Group in Chronic Heart Failure risk score; PREDICT-HF, PARADIGM Risk of Events and Death in the Contemporary Treatment of Heart Failure; SHFM, Seattle Heart Failure Model.
[a]Observed average survival according to Kaplan–Meier survival curve.
[b]Average predicted survival.

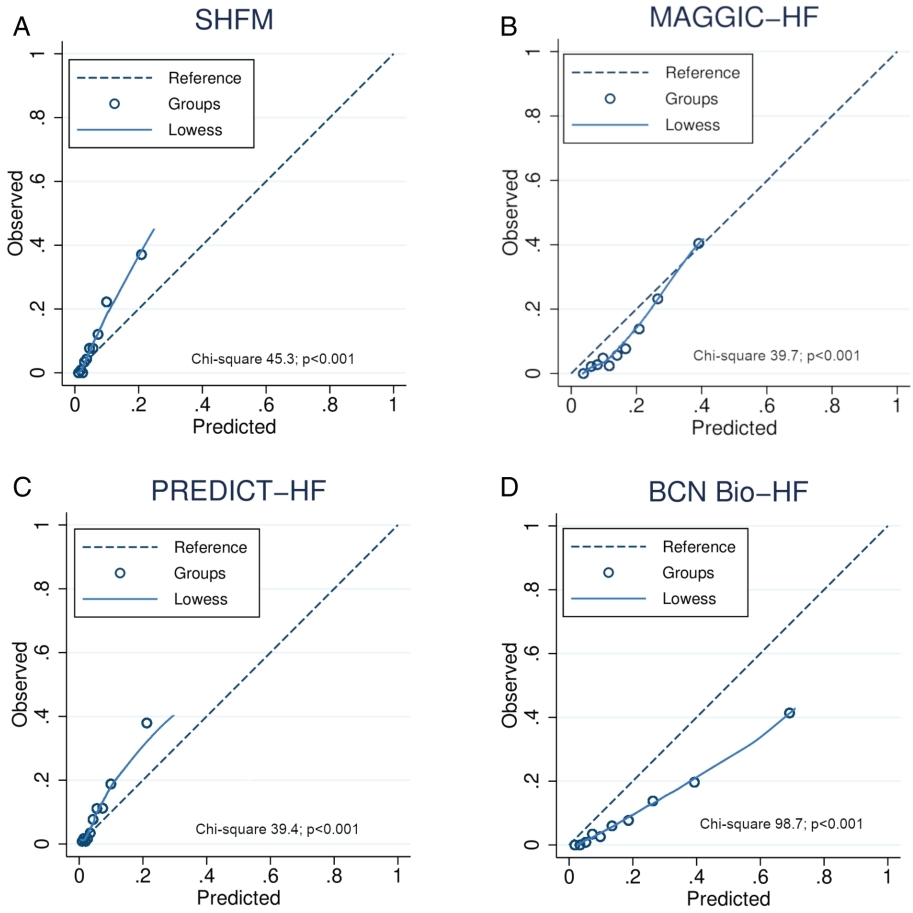


Figure 1 Calibration plots comparing observed vs. predicted mortality at 1 year: (*A*) Seattle Heart Failure Model (SHFM); (*B*) Meta-Analysis Global Group in Chronic Heart Failure (MAGGIC-HF) risk score; (*C*) PARADIGM Risk of Events and Death in the Contemporary Treatment of Heart Failure (PREDICT-HF); (*D*) Barcelona Bio-Heart Failure (BCN-Bio-HF) risk calculator. Y axis, observed mortality; X axis, expected mortality; dashed line represents best fitting curve; LOWESS smoother curve (blue line) allows assessing calibration at individual patient level; circles represent groups automatically created by the test.

cohorts (AUC at 1 year 0.725 in the derivation PRAISE-1 cohort, Harrell's C-statistic 0.817 in the present cohort).

The MAGGIC-HF risk score is easier to use than the SHFM. The MAGGIC-HF risk score was based on multivariable piece-wise Poisson regression methods with stepwise variable selection. The final model included 13 highly significant independent predictors of mortality.

The development of the MAGGIC-HF score was based on 39 372 patients from 30 studies. Notably, only 34% of those patients were treated with beta-blockers, and 21% were treated with

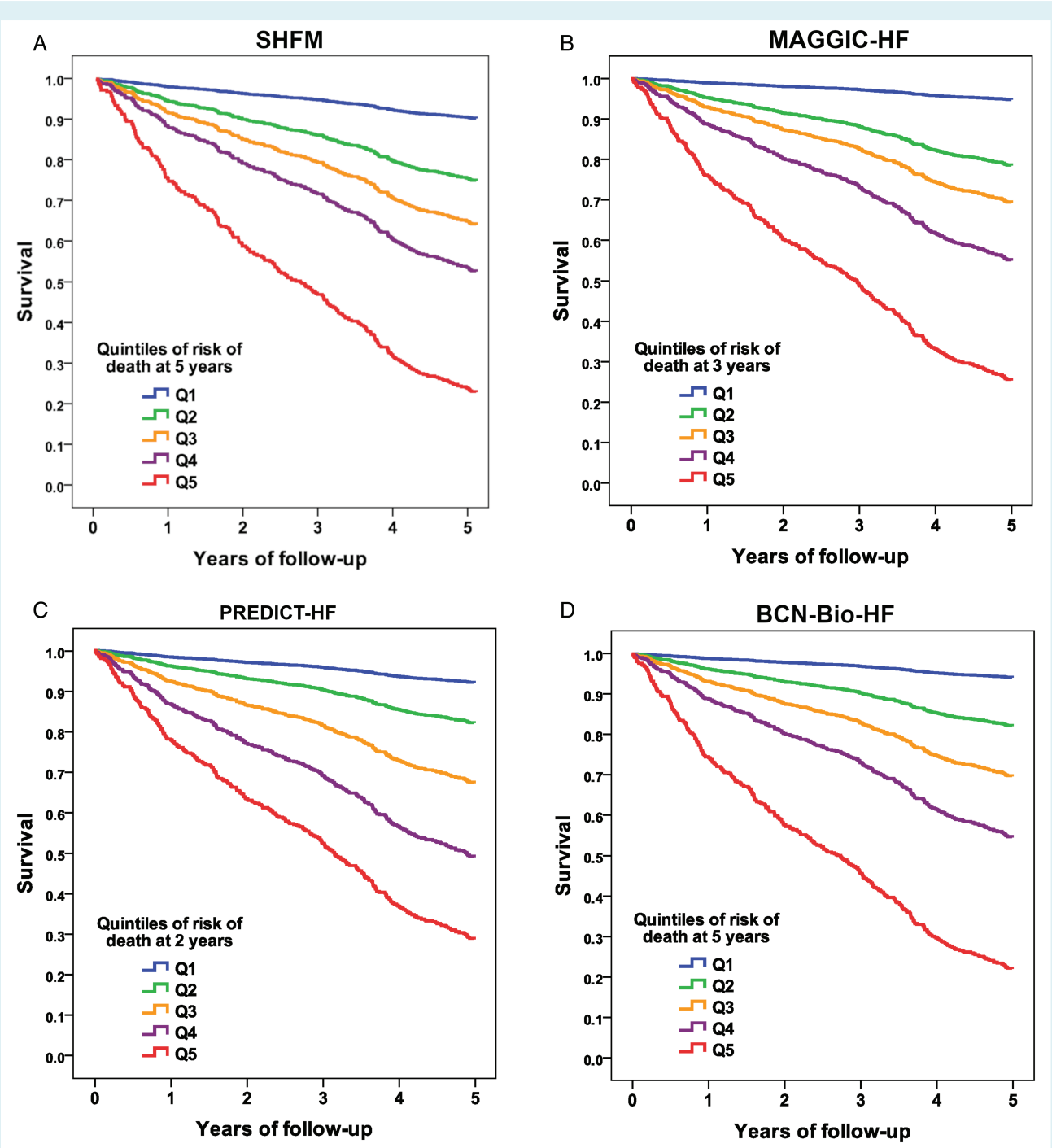


Figure 2 Survival curves based on quintiles of risk of death estimated by each calculator: (*A*) Seattle Heart Failure Model (SHFM); (*B*) Meta-Analysis Global Group in Chronic Heart Failure (MAGGIC-HF) risk score; (*C*) PARADIGM Risk of Events and Death in the Contemporary Treatment of Heart Failure (PREDICT-HF); (*D*) Barcelona Bio-Heart Failure (BCN-Bio-HF) risk calculator. Q1, lowest quintile of risk; Q5, highest quintile of risk.

MRAs. As mentioned, all the subjects included in the derivation cohorts for the MAGGIC-HF and SHFM scores were participating in clinical trials or registries; therefore, it is unknown how well they represented patients in routine clinical practice. Moreover, neither the MAGGIC-HF nor the SHFM score incorporated biomarkers.

The PREDICT-HF included natriuretic peptides and the cohorts used had a more extensive and complete collection of clinical and laboratory data. It was derived from the PARADIGM-HF cohort and patients were receiving contemporary levels of guideline-recommended therapies. It includes some variables

infrequent in routine clinical practice, such as chloride, bilirubin and albumin but, in contrast, it did not incorporate MRAs, CRT, or ICD. Despite a high proportion of missing values in these variables, in our study the model showed similar discrimination to the MAGGIC-HF and SHFM. However, the consistent absence of chloride and bilirubin could have carried that PREDICT-HF lost some discriminative ability. The online calculator provides the risk of cardiovascular death, but does not provide this for all-cause mortality.

In contrast, the BCN-Bio-HF was derived from a clinical HF cohort of patients managed in a multidisciplinary HF unit. It incorporates contemporary treatments, such as ARNI, and it can operate with or without biomarkers that are known to refine pathophysiological pathways in HF. Previous studies have revealed that the addition of NT-proBNP to the classic scores was beneficial, in terms of predicting mortality in patients with HF[13,14] but neither MAGGIC-HF nor SHFM added this biomarker into their online calculator (*Graphical Abstract*).

As decisions regarding patient management may be influenced by the magnitude of the patient's predicted risk, a model is most useful when it can significantly discriminate but also when is appropriately calibrated. In the present study, using the model of the BCN-Bio-HF calculator containing NT-proBNP only, this score showed the best discrimination capacity, but systematically overestimated the risk of the patients at all time-points; by contrast, the MAGGIC-HF score showed the best calibration, although also moderately overestimated the risk. On the other hand, the SHFM and the PREDICT-HF score underestimated risk at any time-point. The observation that the SHFM underestimated the risk in patients at high risk of HF-related mortality at 1 year was consistent with findings in previous studies.[15]

To our knowledge, this study was the first to compare these four web-based contemporary HF risk scores directly and comprehensively in a real-life prospective cohort of patients managed in a multidisciplinary HF clinic. The best discrimination at every time-point that risk was estimated in BCN-Bio-HF was likely due to the incorporation of NT-proBNP. Consequently, these data suggest that NT-proBNP should be included routinely in all risk-stratification tools for HF. The sensitivity analysis with such calculator including also ST2 and hs-TnT performed in 413 patients provided a slight improvement of discrimination and tended to lower overestimation of risk.

Finally, the significant improvement found in the sensitivity analysis of BCN-Bio-HF after its recalibration, together with the significant reduction in HF mortality trends observed in recent years[16] and the introduction of new mainstay drugs in the treatment of HF patients, suggest that recalibration and updating of all these online calculators should be regularly performed for improving accuracy of these tools in estimating risk of death.

Study limitations

Our study has some limitations. First, our sample comprised patients with general HF. Most patients had depressed ejection fractions and were treated at a multidisciplinary HF unit in a tertiary hospital. Additionally, most of the patients were referred from the cardiology department. Thus, our cohort comprised mostly relatively young men with HF of ischaemic aetiology. Consequently, our results might not be generalizable to a global HF population that may include patients with HF with preserved ejection fraction. Although patients with more than three missing values (five for PREDICT-HF) were excluded, we could not rule out the possibility that some bias occurred due to the missing variables. Our sample is limited, single centre, and included during a long time period. A more robust comparison of risk scores should ideally be undertaken in a larger multicentre contemporary patient population. All patients were derived from the same clinic as the original BCN-Bio-HF calculator, so we cannot discard a potential bias in the analysis. Finally, the BCN-Bio-HF model that incorporated hs-TnT and ST2 could not be used in the whole cohort, because the measurements of these two biomarkers at the first visit were not available for a substantial proportion of patients. Although the addition of only NT-proBNP did not improve significantly the performance of this tool in the derivation cohort, in the present study C-statistic numerically improved at any time-point, reaching statistical significance in several comparisons and reclassification measurements were statistically significant at any time-point. Indeed, external validation of this tool has only been performed with the model containing the three biomarkers. Yet, we could perform a sensitivity analysis with the three biomarkers that were available in 413 of the last included patients, which provided an improvement of the performance of the calculator.

Finally, sensitivity analyses – especially those with subgroup analysis – should be considered with caution both due the smaller number of patients and to potential selection bias.

Conclusions

This study illustrates the complexity of risk prediction in HF. Four comprehensive and quite contemporary online risk scores were compared head-to-head, and no one was found uncontrovertibly better than the rest. BCN-Bio-HF showed best discrimination and overall performance, and MAGGIC-HF showed best calibration. The four studied scores either overestimated (BCN-Bio-HF, MAGGIC-HF) or tended to underestimate (SHFM, PREDICT-HF) the risk of death. Biomarkers, in particular NT-proBNP, seem to provide value in risk stratification. Regular updating and calibration of online web calculators are recommended to keep them with value in the ever-changing landscape of HF management.

Supplementary Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Acknowledgements

We wish to thank the nurses in the heart failure unit for data collection and for their invaluable work in the unit.

Funding

The NT-proBNP assays were partially provided by Roche Diagnostics. Roche Diagnostics had no role in the design of the study or the collection, management, analysis, or interpretation of the data. **Conflict of interest**: A.B.G. has received lecture honoraria from Abbott, AstraZeneca, Boehringer-Ingelheim, Novartis, Vifor, Roche Diagnostics and Critical Diagnostics. A.B.G. and J.L. report a relationship with Critical Diagnostics. All other authors have nothing to disclose.

References

1. Levy WC, Mozaffarian D, Linker DT, Sutradhar SC, Anker SD, Cropp AB, Anand I, Maggioni A, Burton P, Sullivan MD, Pitt B, Poole-Wilson PA, Mann DL, Packer M. The Seattle Heart Failure Model: prediction of survival in heart failure. *Circulation* 2006;**113**:1424–1433.
2. Pocock SJ, Ariti CA, McMurray JJ, Maggioni A, Køber L, Squire IB, Swedberg K, Dobson J, Poppe KK, Whalley GA, Doughty RN; Meta-Analysis Global Group in Chronic Heart Failure. Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies. *Eur Heart J* 2013;**34**:1404–1413.
3. Sartipy U, Dahlström U, Edner M, Lund LH. Predicting survival in heart failure: validation of the MAGGIC heart failure risk score in 51,043 patients from the Swedish Heart Failure Registry. *Eur J Heart Fail* 2014;**16**:173–179.
4. Simpson J, Jhund PS, Lund LH, Padmanabhan S, Claggett BL, Shen L, Petrie MC, Abraham WT, Desai AS, Dickstein K, Køber L, Packer M, Rouleau JL, Mueller-Velten G, Solomon SD, Swedberg K, Zile MR, McMurray JJV. Prognostic models derived in PARADIGM-HF and validated in ATMOSPHERE and the Swedish Heart Failure Registry to predict mortality and morbidity in chronic heart failure. *JAMA Cardiol* 2020;**5**:432–441.
5. Lupón J, de Antonio M, Vila J, Peñafiel J, Galán A, Zamora E, Urrutia A, Bayes-Genis A. Development of a novel heart failure risk tool: the Barcelona Bio-Heart Failure risk calculator (BCN Bio-HF calculator). *PLoS One* 2014;**9**:e85466.
6. Lupón J, Januzzi JL, de Antonio M, Vila J, Peñafilel J, Bayes-Genis A. Validation of the Barcelona Bio-Heart Failure risk calculator in a cohort from Boston. *Rev Esp Cardiol (Engl Ed)* 2015;**68**:80–81.
7. Lupón J, Simpson J, McMurray JJV, de Antonio M, Vila J, Subirana I, Barallat J, Moliner P, Domingo M, Zamora E, Bayes-Genis A. Barcelona Bio-HF Calculator Version 2.0: incorporation of angiotensin II receptor blocker neprilysin inhibitor (ARNI) and risk for heart failure hospitalization. *Eur J Heart Fail* 2018;**20**:938–940.
8. Hartmann F, Packer M, Coats AJ, Fowler MB, Krum H, Mohacsi P, Rouleau JL, Tendera M, Castaigne A, Anker SD, Amann-Zalan I, Hoersch S, Katus HA. Prognostic impact of plasma N-terminal pro-brain natriuretic peptide in severe chronic congestive heart failure: a substudy of the Carvedilol Prospective Randomized Cumulative Survival (COPERNICUS) trial. *Circulation* 2004;**110**:1780–1786.
9. Bettencourt P, Azevedo A, Pimenta J, Friões F, Ferreira S, Ferreira A. N-terminal-pro-brain natriuretic peptide predicts outcome after hospital discharge in heart failure patients. *Circulation* 2004;**110**:2168–2174.
10. Zamora E, Lupón J, Vila J, Urrutia A, de Antonio M, Sanz H, Grau M, Ara J, Bayés-Genís A. Estimated glomerular filtration rate and prognosis in heart failure: value of the Modification of Diet in Renal Disease Study-4, Chronic Kidney Disease Epidemiology Collaboration, and Cockroft-Gault formulas. *J Am Coll Cardiol* 2012;**59**:1709–1715.
11. Lupón J, Gavidia-Bovadilla G, Ferrer E, de Antonio M, Perera-Lluna A, López-Ayerbe J, Domingo M, Núñez J, Zamora E, Moliner P, Díaz-Ruata P, Santesmases J, Bayés-Genís A. Dynamic trajectories of left ventricular ejection fraction in heart failure. *J Am Coll Cardiol* 2018;**72**:591–601.
12. Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JGF, Coats AJS, Falk V, González-Juanatey JR, Harjola VP, Jankowska EA, Jessup M, Linde C, Nihoyannopoulos P, Parissis JT, Pieske B, Riley JP, Rosano GMC, Ruilope LM, Ruschitzka F, Rutten FH, van der Meer P; ESC Scientific Document Group. 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: the Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. *Eur J Heart Fail* 2016;**18**:891–975.
13. Khanam SS, Choi E, Son JW, Lee JW, Youn YJ, Yoon J, Lee SH, Kim JY, Ahn SG, Ahn MS, Kang SM, Baek SH, Jeon ES, Kim JJ, Cho MC, Chae SC, Oh BH, Choi DJ, Yoo BS. Validation of the MAGGIC (Meta-Analysis Global Group in Chronic Heart Failure) heart failure risk score and the effect of adding natriuretic peptide for predicting mortality after discharge in hospitalized patients with heart failure. *PLoS One* 2018;**13**:e0206380.
14. Arzilli C, Aimo A, Vergaro G, Ripoli A, Senni M, Emdin M, Passino C. N-terminal fraction of pro-B-type natriuretic peptide versus clinical risk scores for prognostic stratification in chronic systolic heart failure. *Eur J Prev Cardiol* 2018;**25**:889–895.
15. Sartipy U, Goda A, Yuzefpolskaya M, Mancini DM, Lund LH. Utility of the Seattle Heart Failure Model in patients with cardiac resynchronization therapy and implantable cardioverter defibrillator referred for heart transplantation. *Am Heart J* 2014;**168**:325–331.
16. Spitaleri G, Lupón J, Domingo M, Santiago-Vacas E, Codina P, Zamora E, Cediel G, Santesmases J, Diez-Quevedo C, Troya MI, Boldo M, Altmir S, Alonso N, González B, Núñez J, Bayes-Genis A. Mortality trends in an ambulatory multidisciplinary heart failure unit from 2001 to 2018. *Sci Rep* 2021;**11**:732.