Prognostic scores for sorafenib-treated HCC patients:

A new application for the HAP Score

J Edeline*[1], J-F Blanc[2], B Campillo-Gimenez[1], Y-T Ma[3], J King[4], O Faluyi[5], J Mathurin[2], S Ghazi[3], D H Palmer*[5], T Meyer*[4, 6]

[1-] Oncology/Centre Eugène Marquis, Rennes, France

[2-] Hepatology/CHU Hôpital Saint André, Bordeaux, France

[3-] Oncology/University Hospitals Birmingham NHS Foundation Trust, United Kingdom

[4-] Oncology/Royal Free Hospital, London, United Kingdom

[5-] Department of Molecular and Clinical Cancer Medicine, University of Liverpool, and The Clatterbridge Cancer Centre NHS Foundation Trust, United Kingdom

[6] UCL Cancer Institute, University College London, London, United Kingdom

*These authors contributed equally

Corresponding Author:

Tim Meyer,

UCL Cancer Institute, University College London,

72 Huntley Street, London WC1E 6BT, UK

Tel: +44 0207 679 6731; Fax: +44 0203 108 2025

e-mail: t.meyer@ucl.ac.uk

Keywords: Hepatocellular carcinoma; sorafenib; prognosis; HAP score; BCLC; ALBI

Electronic work count: 2 745 words in the main text

2 tables, 1 figure, 3 supplementary tables and 1 supplementary figure

Abstract:

Background: No prognostic classification is currently used for patients treated with systemic therapies for Hepatocellular Carcinoma (HCC).

Methods: We retrospectively analyzed data from patients treated with sorafenib for HCC from five centers in France and in the United Kingdom. The training set comprised data from two centers and the validation set from three. Variables independently associated with Overall Survival (OS) in the training set were used to build the SAP (Sorafenib Advanced HCC Prognosis) score. The score was tested in the validation set, then compared with other prognostication systems.

Results: The training set and validation set included 370 and 468 patients respectively. In the training set, variables independently associated with OS in multivariable analysis were: performance status >0, AFP >400ng/mL, tumor size >7cm, bilirubin >17µmol/L and albumin <36g/L. The SAP score was built giving one point to each abnormal variable, and 3 classes were constructed. The SAP score was significantly associated with OS in the training set, with median OS of 14.9 months for SAP A, 7.2 months for SAP B and 2.5 months for SAP C (P<0.001). In the validation set, the SAP score was significantly associated with OS, and showed greater discriminative abilities than BCLC and ALBI. However, the HAP score showed greater discriminative abilities than the SAP score.

Conclusion: In European patients treated with sorafenib, the HAP was the most discriminant prognostic score and may facilitate stratification in trials and inform clinical decision making.

Abstract word count: 237

Keywords: Molecular Targeted Therapy; liver neoplasms; prognosis; sorafenib; drug therapy

Acknowledgments:

Financial support: JE was partly funded by the Fondation de France for this work. TM is part-funded by the NIHR UCLH Biomedical Research Centre

Introduction

Hepatocellular carcinoma (HCC) is the second leading cause of cancer death worldwide [1]. Many therapeutic interventions are available for HCC, ranging from curative (liver transplantation, resection, ablative treatments) to palliative options (transarterial treatments, systemic treatment) [2,3]. Sorafenib remains the current standard of care for advanced disease [4], with regorafenib recently shown active after first-line sorafenib [5].

Many prognostic classifications have been proposed for HCC [2,6–10], and some studies have compared the different scores [11,12]. However, most of the classifications addressed the prognosis across a wide range of clinical settings, and some are directly linked to a recommended treatment strategy according to stage. To address the prognosis with more accuracy within treatment groups, alternative prognostic scores including HAP and ART were developed in the setting of transarterial (chemo) embolization (TA(C)E) to enable selection of patients for therapy at baseline and after first treatment [13–17]. However, only a few studies have specifically addressed prognostic stratification in the advanced stage [18–20], and their application has been limited in clinical practice.

Developing a clinically relevant prognostic score for patients treated with systemic therapies might support more robust clinical decisions and enable better stratification within clinical trials. We therefore aimed to develop and validate a new score, focusing on patients treated with sorafenib, with the objective to keep the score simple, reproducible and objective to allow direct translation into clinical practice.

Methods:

Patients:

We retrospectively collected data from patients treated with sorafenib for HCC from five centers in the United Kingdom and France. All consecutive patients treated with sorafenib for HCC were entered in the databases. Relevant authorizations were obtained from institutional and ethical review boards for use of the data. Data was acquired under an ethically approved protocol (REC reference 12/LO/1088). Data collected at initiation of sorafenib included age, gender, cause of underlying liver disease, previous treatment for HCC, presence of extrahepatic spread, presence of macrovascular invasion, performance status, alpha-fetoprotein, prothrombin time or International Normalized Ratio (INR), albumin, bilirubin, ascites and encephalopathy (as coded by centers for the Child-Pugh classification), BCLC classification and Child-Pugh score.

Statistical analysis:

Statistical analyses and graphs were performed on R statistical software version 3.1.1 (2014-07-10). Overall survival (OS) was calculated from the start of treatment with Sorafenib to death, survival curves were estimated with the Kaplan-Meier approach and compared with Log-Rank tests. A p value < 0.05 was considered as statistically significant for all analyses.

Data from two centers were considered as a training set, and data from three other centers as a validation set.

With our training dataset, a Cox-regression multivariable model was fitted on the OS to select which variables were used to calculate the new score. Continuous variables were

dichotomized using the Receiver-Operating Curve (ROC) procedure. The score was then calculated attributing 1 point to the worst class of each binary and independently significant covariate and grouping patients into a 3-class system, named Sorafenib Advanced HCC Prognostic (SAP) score.

Then, we used the validation dataset to estimate performance metrics and to compare the SAP score with several existing prognostic scores that we were able to calculate from our database: BCLC, ALBI and HAP [2,8,13]. The HAP score is calculated giving 1 point when the following conditions are met: Albumin < 36 g/dl, AFP > 400 ng/ml, bilirubin > 17 µmol/l or maximum tumour diameter >7 cm, then grouping in 4 classes: HAP A if 0 point, HAP B if 1 point, HAP C if 2 points, HAP D if >2 points. We used goodness of fit parameters of each Cox model fitting OS to approximate homogeneity of each scoring system: a higher log-likelihood ratio (LR) and a lower Akaike Information Criterion (AIC) indicates a better goodness of fit and so a higher homogeneity. Relative likelihoods between models in order to represent the probability that the score with the higher AIC minimizes information loss as effectively as the score with the lower AIC, and is interpreted as a *P* value for the comparison of both AIC. In opposition with AIC, LR depends on the number of classes assessed, we analyzed the original scores and also reduced the HAP score to a 3-class system by combining HAP groups A and B. Monotonicity and discriminatory ability of each scoring system were assessed using the Cox OS regression models with the Harrel's C statistics as discriminative indicator and the linear predictor to compute Area Under the receiver operating Curve (AUC) at 6, 12, and 24 for each scoring system. These discriminative performance metrics were presented with 95% confidence intervals computed with a 2-thousand resampling bootstrap procedure.

Results:

Patient characteristics

From February 2003 to August 2014, 370 patients were treated with sorafenib and included in the training set (all consecutive patients), and 468 patients were included in the validation set (all consecutive patients with data available for calculation of the SAP score). The characteristics of the cohort are presented in Table 1. The median follow-up was 29.0 months [95% Confidence Interval (CI): 21.4-36.6], with 358 patients (76.5%) dead at last follow-up and 43.7 months [95% CI: 38.8-48.6] with 330 (89.2%) dead at last follow-up, in the training and validation sets, respectively.

Construction of the SAP score in the training cohort

Results from the Cox-regression model for OS are presented in Supplementary Table S1. Variables associated with OS in univariable analysis were albumin, bilirubin, previous treatment for HCC, performance status (PS), tumor size, alpha-fetoprotein (AFP) and macrovascular invasion. We performed ROC analysis to dichotomize continuous variables. Results of the ROC analysis are shown in Supplementary Figure 1. The following thresholds were chosen, based on their clinical relevance, their previous use in other classifications, the adequate power of discrimination in the ROC analysis and the fact that each split the population almost by half: >7cm for size, <36g/L for albumin, >17µmol/L for bilirubin and >400ng/mL for AFP. The chosen thresholds were associated with OS in univariable Cox-regression analysis (Supplementary Table S1).

We then performed multivariable Cox-regression analysis in the training set, testing the variables found positive in univariable analysis (Supplementary Table S1). Five factors were found to be independently associated with OS in multivariable analysis: size >7cm, albumin <36g/L, bilirubin >17µmol/L, PS >0 and AFP >400ng/mL, with Hazard Ratios between 1.34 and 1.82. Due to similar Hazard Ratios, we then attributed one point to each of these abnormal values, and constructed a 3-class score, named SAP: SAP A if score was 0 or 1, SAP B if score was 2 or 3 and SAP C if score was 4 or 5. The SAP score was significantly associated with OS in the training set (Supplementary Table S2 and Figure 1A, log-rank *P*<0.001).

Validation of SAP in the validation set, and comparison with alternative scores

As in the training set, the SAP score was significantly associated with OS in the validation set (Supplementary Table S2 and Figure 1B, log-rank *P*<0.001 overall, and every comparison between 2 classes *P*<0.001).

In the validation set we compared the performance of SAP with other scores: BCLC, HAP and ALBI. Of note, 409 (87%) patients were BCLC C as might be expected for this patient population, and only 55 and 4 patients were BCLC B and D respectively. For ALBI, there were only 26 patients (5.6%) in the grade 3 group. BCLC was prognostic (Supplementary Table S2 and Figure 1D, log-rank *P*<0.001 overall, and every comparison between 2 classes *P*<0.001) as was HAP (Supplementary Table S2 and Figure 1E and 1F, log-rank *P*<0.001 overall, p=0.030 for HAP A vs HAP B, p=0.001 for HAP B vs HAP C and *P*<0.001 for HAP C vs HAP D). ALBI was also prognostic, although comparison between grade 2 and 3 was not statistically significant, probably due to low numbers of grade 3 patients (Supplementary Table S2 and Figure 1C, log-rank p<0.001 overall, p<0.001 for ALBI grade 1 vs 2 but p=0.068 for ALBI grade 2 vs 3).

Similar results were seen when we focused on the population of ideal candidates for sorafenib treatment, namely Child-Pugh A patients with BCLC B or C stage (Supplementary Table S2).

Due to the lower number of patients in HAP A and B classes, and in order to compare 3-class scores, HAP A and HAP B were also grouped for further analysis (Table 2). The SAP score seemed superior to BCLC and ALBI, as illustrated by higher homogeneity (higher log likelihood) and by lower loss of information (lower AIC, with significant relative likelihood). Conversely, the SAP score appeared inferior to HAP, either when assessed as a 3-class score or as a 4-class score, according to all parameters tested. Overall, the 4-class HAP score had the best performance with the highest Log likelihood ratio and Harrell's C statistics, the lowest AIC criterion, and the highest AUC at all three survival points.

Finally, we focused on the worst prognostic class of each score in the validation set. Results are presented in Supplementary Table S3. The worst class of every classification was associated with low median OS and low 6- and 12-months survival. However, the number of patients in the worst class was very low for ALBI and BCLC, and the ability to predict the 6-months survival for a patient in the worst class as compared with the others classes was higher and significant for SAP and HAP only. The HAP D class represented almost a fifth of the population, and had a median OS of 3.8 months [95%CI: 3.4-4.2], and 6- and 12-months survival rates of 27% and 15%, respectively.

Discussion

We developed and validated a new prognostic score specific for HCC patients treated with systemic therapy, and more specifically the current first-line treatment, sorafenib. The aim was to construct a clinically-relevant score, easy to use by clinicians, and able to accurately predict the prognosis, that would support clinical decision making and aid stratification in clinical trials. The SAP score is easy to calculate from the characteristics readily available for treatment decision in HCC, namely patient's performance status (PS), liver function tests (albumin and bilirubin) and relevant tumor characteristics (tumor size and AFP). This new score showed better discriminatory abilities than BCLC and ALBI scores. However, the score appeared very similar to the HAP score (partly by design, but mainly due to the same variables being independently associated with OS in this population) but was inferior to the HAP score in terms of discriminatory abilities. The HAP score was originally constructed for patients treated with TA(C)E, but appeared in this study to have equivalent prognostic abilities in this population of more advanced cases.

Many prognostic scores have been developed for HCC [2,6–9,13]. However, some classifications are designed to guide the treatment strategies (as the BCLC and the Hong Kong Liver Cancer classifications), rather than to provide accurate prognostic information [2,9]. Others have used subjective criteria (such as the >50% liver involvement in the Okuda, CLIP and other classifications). The ALBI grade, which was recently developed, focusses only on liver function and in our study, despite significant prognostic abilities, the ALBI grade showed lower discriminative abilities as compared to SAP or HAP, emphasizing the need to incorporate some tumoral characteristics for optimal prognostication. Currently, there is no

consensus on the criteria required for stratification in clinical trials and the SAP or HAP scores provide a potential means to do this.

We chose a pragmatic, yet statistically robust approach in the development of the SAP score so that the score used clinically-relevant thresholds. The ROC analysis was strongly supportive of the 7-cm size threshold, but other thresholds could have been chosen for bilirubin, albumin or AFP, due to a relative plateau. However, we believe that such easy-to-remember thresholds based on limits of the normal range, already used in the HAP and Child-Pugh scores, will facilitate easier translation into clinical practice, while diminishing only moderately prognostic abilities. Linear predictors, even when translated into nomograms, often encounter difficulties for wide adoption into clinical practice. This is well illustrated by the ALBI score, which was developed as a linear predictor, but was translated into three grades for easier use in the clinic while maintaining advantage over the Child-Pugh score [8,21]. It might appear surprising that neither macrovascular invasion nor extra-hepatic spread were found prognostic in this study. One explanation that we can propose is that while in the overall population of HCC macrovascular invasion and extra-hepatic spread do have strong prognostic value, when we consider the more advanced population of patients treated with sorafenib, this value is less prominent. As regards to extra-hepatic spread, the prognostic value has frequently been studied grouped with macrovascular invasion. In this cohort of patients treated with sorafenib, extra-hepatic spread by itself was not a prognostic factor. Interestingly, it was also not prognostic in the validation cohort (data not shown). As regards to macrovascular invasion, the prognostic value disappeared in multivariable analysis, suggesting that other factors correlated with macrovascular invasion (AFP level, size of the lesion, frequent hepatic dysfunction…) might be more directly prognostic that macrovascular invasion itself.

At least three other prognostic scores were proposed in the setting of HCC treated with systemic therapies [18–20]. The NIACE score focused on the BCLC C population [18]. The score incorporates 5 variables: number of tumors ≥3, infiltrative tumor, AFP ≥200ng/mL, Child-Pugh B, and PS ≥1. The calculated score ranges from 0 to 7. However, the variable "infiltrative tumor" might be viewed as subjective, and no grouping in classes was proposed, thus limiting the applicability of the score. We were not able to test the NIACE score in our cohort because only a few of the databases had the "infiltrative tumor" criterion recorded. The score proposed by Choi et al still retains the subjective "CLIP-morphology", provides a linear score and is more challenging to implement in the clinic [19]. The Japan Red Cross score, despite interesting prognostic abilities, uses des-γ-carboxy prothrombin, which is not commonly available in western clinical practice, and also uses the subjective >50% liver involvement criterion [20]. We believe that these characteristics might explain the lack of adoption of such classifications in clinical practice, and that the limitations are overcome by the SAP and HAP scores. Moreover, the HKLC group developed an alternative system to the BCLC. We were not able to test this system in our series because it uses the variable extra-hepatic vascular invasion, while our databases did not distinguish the extent of macrovascular invasion. However, the HKLC system was developed more as a way to propose alternative treatment strategies rather than a better prognostic system.

After developing the SAP score in the training cohort, we validated it in the validation set, and showed benefit over BCLC and ALBI. However, despite having been developed in another context, the HAP score showed better characteristics, with better homogeneity and higher discrimination. This could be explained by the exclusion of the only subjective parameter of the SAP score, PS. Indeed, the use of PS in the BCLC classification has been the subject of controversy [22,23]. Our analysis illustrates the fact that every score should be

tested in a separate validation cohort: while the inclusion of PS significantly improved the results in the training set, its use in the validation set diminished its performance compared to the HAP score. The HAP score thus offers the advantages of being totally objective, easy-to-use, and we would recommend its use in preference to the SAP score. The HAP D class defined a significant proportion (20.4%) of the population with a poor prognosis (median OS of 3.8 months, 6-months OS of 26% and 12-months OS of 15%), for which the benefit of sorafenib is questionable. Interestingly, the previously reported median survival of HAP D patients treated with TA(C)E was 3.6 months suggesting that HAP D defines a poor prognostic group regardless of therapy [13].

Our analysis has some limitations: First, the analysis was retrospective and should be validated in prospective studies, second, the results were obtained in a European population and require validation in other populations, and finally we were not able to compare it with the NIACE and HKLC classifications.

In conclusion, we have developed a new prognostic score for patients treated with systemic therapies, but shown that the HAP score is superior in this setting. After appropriate validation in confirmatory cohorts, we would recommend the following use of HAP score in the context of systemic therapies: HAP A/B vs C could be used for stratification in clinical trials, while the benefit/risk ratio of sorafenib in the HAP D population should be questioned in view of the very low survival of this population.

References:

[1] Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012: Global Cancer Statistics, 2012. CA Cancer J Clin 2015;65:87–108. doi:10.3322/caac.21262.

[2] EASL, EORTC, others. EASL–EORTC Clinical Practice Guidelines: management of hepatocellular carcinoma. J Hepatol 2012;56:908–943.

[3] Verslype C, Rosmorduc O, Rougier P, ESMO Guidelines Working Group. Hepatocellular carcinoma: ESMO-ESDO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol Off J Eur Soc Med Oncol 2012;23 Suppl 7:vii41-48. doi:10.1093/annonc/mds225.

[4] Di Marco V, De Vita F, Koskinas J, Semela D, Toniutto P, Verslype C. Sorafenib: from literature to clinical practice. Ann Oncol Off J Eur Soc Med Oncol 2013;24 Suppl 2:ii30-37. doi:10.1093/annonc/mdt055.

[5] Bruix J, Qin S, Merle P, Granito A, Huang Y-H, Bodoky G, et al. Regorafenib for patients with hepatocellular carcinoma who progressed on sorafenib treatment (RESORCE): a randomised, double-blind, placebo-controlled, phase 3 trial. Lancet Lond Engl 2017;389:56–66. doi:10.1016/S0140-6736(16)32453-9.

[6] A new prognostic system for hepatocellular carcinoma: a retrospective study of 435 patients: the Cancer of the Liver Italian Program (CLIP) investigators. Hepatol Baltim Md 1998;28:751–5. doi:10.1002/hep.510280322.

[7] Okuda K, Obata H, Nakajima Y, Ohtsuki T, Okazaki N, Ohnishi K. Prognosis of primary hepatocellular carcinoma. Hepatol Baltim Md 1984;4:3S–6S.

[8] Johnson PJ, Berhane S, Kagebayashi C, Satomura S, Teng M, Reeves HL, et al. Assessment of Liver Function in Patients With Hepatocellular Carcinoma: A New

Evidence-Based Approach--The ALBI Grade. J Clin Oncol 2015;33:550–8. doi:10.1200/JCO.2014.57.9151.

[9] Yau T, Tang VYF, Yao T-J, Fan S-T, Lo C-M, Poon RTP. Development of Hong Kong Liver Cancer staging system with treatment stratification for patients with hepatocellular carcinoma. Gastroenterology 2014;146:1691–1700.e3. doi:10.1053/j.gastro.2014.02.032.

[10] Chevret S, Trinchet JC, Mathieu D, Rached AA, Beaugrand M, Chastang C. A new prognostic classification for predicting survival in patients with hepatocellular carcinoma. Groupe d'Etude et de Traitement du Carcinome Hépatocellulaire. J Hepatol 1999;31:133–41.

[11] Liu P-H, Hsu C-Y, Hsia C-Y, Lee Y-H, Su C-W, Huang Y-H, et al. Prognosis of hepatocellular carcinoma: Assessment of eleven staging systems. J Hepatol 2016;64:601–8. doi:10.1016/j.jhep.2015.10.029.

[12] Marrero JA, Fontana RJ, Barrat A, Askari F, Conjeevaram HS, Su GL, et al. Prognosis of hepatocellular carcinoma: comparison of 7 staging systems in an American cohort. Hepatol Baltim Md 2005;41:707–16. doi:10.1002/hep.20636.

[13] Kadalayil L, Benini R, Pallan L, O'Beirne J, Marelli L, Yu D, et al. A simple prognostic scoring system for patients receiving transarterial embolisation for hepatocellular cancer. Ann Oncol Off J Eur Soc Med Oncol ESMO 2013;24:2565–70. doi:10.1093/annonc/mdt247.

[14] Hucke F, Pinter M, Graziadei I, Bota S, Vogel W, Müller C, et al. How to STATE suitability and START transarterial chemoembolization in patients with intermediate stage hepatocellular carcinoma. J Hepatol 2014;61:1287–96. doi:10.1016/j.jhep.2014.07.002.

[15] Hucke F, Sieghart W, Pinter M, Graziadei I, Vogel W, Müller C, et al. The ART-strategy: sequential assessment of the ART score predicts outcome of patients with hepatocellular carcinoma re-treated with TACE. J Hepatol 2014;60:118–26. doi:10.1016/j.jhep.2013.08.022.

[16] Adhoute X, Penaranda G, Naude S, Raoul JL, Perrier H, Bayle O, et al. Retreatment with TACE: the ABCR SCORE, an aid to the decision-making process. J Hepatol 2015;62:855–62. doi:10.1016/j.jhep.2014.11.014.

[17] Bolondi L, Burroughs A, Dufour J-F, Galle PR, Mazzaferro V, Piscaglia F, et al. Heterogeneity of patients with intermediate (BCLC B) Hepatocellular Carcinoma: proposal for a subclassification to facilitate treatment decisions. Semin Liver Dis 2012;32:348–59. doi:10.1055/s-0032-1329906.

[18] Adhoute X, Pénaranda G, Raoul JL, Blanc JF, Edeline J, Conroy G, et al. Prognosis of advanced hepatocellular carcinoma: a new stratification of Barcelona Clinic Liver Cancer stage C: results from a French multicenter study. Eur J Gastroenterol Hepatol 2016;28:433–40. doi:10.1097/MEG.0000000000000558.

[19] Choi GH, Han S, Shim JH, Ryu M-H, Ryoo B-Y, Kang Y-K, et al. Prognostic Scoring Models for Patients Undergoing Sorafenib Treatment for Advanced Stage Hepatocellular Carcinoma in Real-Life Practice: Am J Clin Oncol 2014:1. doi:10.1097/COC.0000000000000132.

[20] Takeda H, Nishikawa H, Osaki Y, Tsuchiya K, Joko K, Ogawa C, et al. Proposal of Japan Red Cross score for sorafenib therapy in hepatocellular carcinoma. Hepatol Res Off J Jpn Soc Hepatol 2015;45:E130-140. doi:10.1111/hepr.12480.

[21] Edeline J, Blanc J-F, Johnson P, Campillo-Gimenez B, Ross P, Ma YT, et al. A multicenter comparison between Child Pugh and ALBI scores in patients treated with sorafenib for

Hepatocellular Carcinoma. Liver Int Off J Int Assoc Study Liver 2016. doi:10.1111/liv.13170.

[22] Hsu C-Y, Lee Y-H, Hsia C-Y, Huang Y-H, Su C-W, Lin H-C, et al. Performance status in patients with hepatocellular carcinoma: determinants, prognostic impact, and ability to improve the Barcelona Clinic Liver Cancer system. Hepatol Baltim Md 2013;57:112–9. doi:10.1002/hep.25950.

[23] Hsu C-Y, Liu P-H, Lee Y-H, Hsia C-Y, Huang Y-H, Chiou Y-Y, et al. Aggressive therapeutic strategies improve the survival of hepatocellular carcinoma patients with performance status 1 or 2: a propensity score analysis. Ann Surg Oncol 2015;22:1324–31. doi:10.1245/s10434-014-4151-2.

Table 1: Characteristics of the cohorts

Parameter	Training set (*N*=370)	Validation set (*N*=468 with data for SAP)
Gender: male	90.2%	84.8%
Median age (range)	67 (34-89)	67 (17-84)
Cirrhosis	73.9% (*N* =306)	81.4% (*N* =208)
Median albumin value, g/L (range) - Albumin < 36g/L	36 (19-54) 181 (49.7%) (*N* =364)	38 (18-61) 168 (35.9%)
Median bilirubin value, mcmol/L (range) - Bilirubin >17mcmol/L	17 (3-436) 169 (46.3%) (*N* =365)	15 (3-533) 206 (44.0%)
Child-Pugh class A	257 (75.8%) (*N* =339)	362 (77.4%)
Etiology of underlying liver disease (some patients might have several): - Alcohol - HBV - HCV - Metabolic Syndrome or Hemochromatosis - No known underlying liver disease	 206 (55.7%) 24 (6.5%) 72 (19.5%) 148 (40.0%) 66 (17.8%)	(n=365) 160 (43.8%) 38 (10.4%) 59 (15.3%) 103 (28.2%) 83 (22.7%)
Previous treatment for HCC	160 (52.3%) (*N* =306)	203 (43.4%) (*N* =370)
Performance status >0	161 (43.8%) (*N* =368)	308 (65.8%)

Median size of the largest liver lesion, mm (range) - Size >7cm	60 (0-250) 151 (40.8%)	67 (0-300) 200 (42.7%)
Extra-hepatic Spread	133 (44.4%) (*N* =360)	124 (26.5%)
Macrovascular invasion:	143 (39.8%) (*N* =359)	192 (41.0%)
Median AFP, ng/mL (range) - AFP >400ng/mL	121 (0-849,553) 150 (41.0%)	109 (1-690,000) 178 (38.0%)
BCLC stage: - A - B - C - D	(n=362) 4 (1.1%) 68 (18.8%) 263 (78.2%) 7 (1.7%)	 0 (0%) 55 (11.8%) 409 (87.4%) 4 (0.9%)

Table 2: Comparison between scores in the validation set. A higher log likelihood ratio indicated greater homogeneity, a higher Harrell's C statistics greater discriminative abilities, a lower Akaike Information Criterion (AIC) indicates lower loss of information.

	Log likelihood ratio	Harrell's C statistics	Akaike Information Criterion (AIC)	Relative likelihood of the score equivalence of AIC vs SAP	Relative likelihood of the score equivalence of AIC vs HAP 4 classes	AUC 6 months	AUC 12 months	AUC 24 months
SAP	77.5	0.640 [0.614-0.667]	3683		0.03	0.699 [0.655 - 0.741]	0.675 [0.628 - 0.721]	0.732 [0.669 - 0.789]
ALBI	25.7	0.595 [0.566-0.625]	3735	<0.001	<0.001	0.636 [0.590 - 0.681]	0.593 [0.545 - 0.643]	0.646 [0.570 - 0.719]
BCLC	25.8	0.550 [0.528-0.571]	3723	<0.001	<0.001	0.564 [0.535 - 0.593]	0.593 [0.556 - 0.632]	0.597 [0.533 - 0.664]

HAP 4 classes	86.2	0.653 [0.624-0.681]	3676	0.03		0.724 [0.674 - 0.768]	0.724 [0.669 - 0.775]	0.761 [0.701 - 0.817]
HAP 3 classes	81.6	0.642 [0.614-0.671]	3679	0.14	0.22	0.712 [0.662 - 0.760]	0.706 [0.654 - 0.754]	0.736 [0.679 - 0.787]

Figure Legends:

Supplementary Figure: Receiver Operating Characteristics curves used for the dichotomization of continuous variables.

Figure 1: OS according to SAP score in the training set (A) and in the validation set (B), and according to ALBI grade (C), BCLC (D), HAP as a 4-class (E) or 3-class (F) score in the validation set.