Vrije Universiteit Brussel

VRIJE
UNIVERSITEIT
BRUSSEL

AFP score and metroticket 2.0 perform similarly and could be used in a "within-ALL" clinical decision tool
collaborative group for HCC and liver transplantation

AFP score and metroticket 2.0 perform similarly and could be used in a "within-ALL" clinical decision tool

Authors

Federico Piñero, Charlotte Costentin, Helena Degroote, Andrea NotarpaoloIlka FSF. Boin, Karim Boudjema, Cinzia Baccaro, Aline Chagas, Philippe Bachellier, Giuseppe Maria Ettorre, Jaime Poniachik, Fabrice Muscari, Fabrizio Dibenedetto, Sergio Hoyos Duque, Ephrem Salame, Umberto Cillo, Sebastián Marciano, Claire Vanlemmens, Stefano Fagiuoli, Flair Carrilho, Daniel Cherqui, Patrizia Burra, Hans Van Vlierberghe, Quirino Lai, Marcelo Silva, Fernando Rubinstein, Christophe Duvoux

Correspondence

fpinerof@cas.austral.edu.ar (F. Piñero), christophe.duvoux@hmn.aphp.fr (C. Duvoux).

Graphical abstract
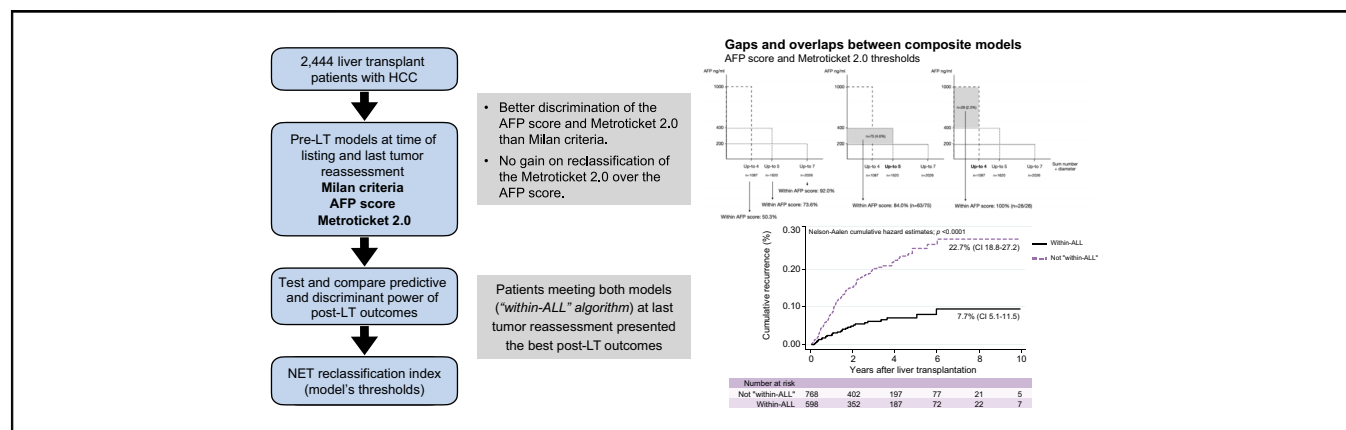
Highlights

- At listing but not last tumor reassessment, Metroticket 2.0 showed better discriminative ability than AFP score for HCC recurrence.

- Discriminative power using respective thresholds was similar between models, either at listing or last tumor reassessment.

- Gaps and overlaps were observed when stratifying recurrence risk according to proposed thresholds.

- Combining both models at listing and at last tumor reassessment in a *"within-ALL decision algorithm"* could optimize candidate selection.

Impact and implications

Composite models were recently proposed for the selection of liver transplant (LT) candidates among individuals with hepatocellular carcinoma (HCC). We found that both the AFP score and Metroticket 2.0 predicted post-LT HCC recurrence and survival better than Milan criteria; the Metroticket 2.0 did not result in better reclassification for transplant selection compared to the AFP score, with predictive gaps and overlaps between the two models; patients who met low-risk thresholds for both models had the lowest 5-year recurrence rate. We propose prospectively testing the combination of both models, to further optimize the LT selection process for candidates with HCC.

AFP score and metroticket 2.0 perform similarly and could be used in a "within-ALL" clinical decision tool

Federico Piñero,[1,*,†] Charlotte Costentin,[2,†] Helena Degroote,[3] Andrea Notarpaolo,[4] Ilka FSF. Boin,[5]

Karim Boudjema,[6] Cinzia Baccaro,[7] Aline Chagas,[8] Philippe Bachellier,[9] Giuseppe Maria Ettorre,[10]

Jaime Poniachik,[11] Fabrice Muscari,[12] Fabrizio Dibenedetto,[13] Sergio Hoyos Duque,[14] Ephrem Salame,[15]

Umberto Cillo,[16] Sebastián Marciano,[17] Claire Vanlemmens,[18] Stefano Fagiuoli,[19] Flair Carrilho,[8] Daniel Cherqui,[20]

Patrizia Burra,[21] Hans Van Vlierberghe,[3] Quirino Lai,[22] Marcelo Silva,[1] Fernando Rubinstein,[23]

Christophe Duvoux [24,*], for the collaborative group for HCC and liver transplantation, France, Latin America, Italy, Belgium

[1]Hospital Universitario Austral, School of Medicine, Austral University, Argentina and Latin American Liver Research Educational and Awareness Network (LALREAN), Argentina; [2]Grenoble Alpes University Institute for Advanced Biosciences, Research Center UGA/Inserm U 1209/CNRS 5309, Gastroenterology, hepatology and GI oncology department, Digidune, Grenoble Alpes University Hospital, La Tronche, France; [3]Department of Hepatology and Gastroenterology, Ghent University Hospital, Belgium; [4]Arcispedale Santa Maria Nuova, Reggio Emilia, Italy; [5]Hospital das Clinicas UNICAMP Campinas, Brazil; [6]Department of Hepatobiliary and Digestive Surgery, Pontchaillou Hospital Rennes 1 University, Rennes, France; [7]Lanciano's Hospital, Chieti, Rome, Italy; [8]Division of Clinical Gastroenterology and Hepatology, Hospital das Clínicas Department of Gastroenterology, School of Medicine, University of São Paulo, Brazil; [9]Digestive Surgery Unit, CHU Strasbourg, Strasbourg, France; [10]Ospedale San Camillo di Roma, Rome, Italy; [11]Hospital Clínico de la Universidad de Chile, Santiago, Chile; [12]Digestive Surgery and Transplant Unit, Hôpital Rangueil, Toulouse, France; [13]Hepato-Pancreato-Biliary Surgery and Liver Transplantation Unit, Department of General Surgery, University of Modena and Reggio Emilia, Italy; [14]Hospital Pablo Tobón Uribe y Grupo de Gastrohepatología de la Universidad de Antioquía, Medellín, Colombia; [15]Digestive Surgery Unit, CHU de Tours, Tours, France; [16]Hepatobiliary Surgery and Liver Transplant Unit, Padova University Hospital, Padova, Italy; [17]Hospital Italiano de Buenos Aires, Argentina; [18]Hepatology Unit, Hôpital Juan Minjoz, Besançon, France; [19]Gastroenterology, Hepatology and Transplantation, Papa Giovanni XXIII Hospital, Bergamo, Italy; [20]Hospital Paul Brousse-APHP, Villejuif, Paris, France; [21]Multivisceral Transplant Unit, Department of Surgery, Oncology and Gastroenterology, Padova University Hospital, Padova, Italy; [22]General Surgery and Organ Transplantation Unit, Sapienza University of Rome, Italy; [23]Institute for Clinical Effectiveness and Health Policy (IECS), Buenos Aires, Argentina; [24]Department of Hepatology, Medical Liver Transplant Unit, Hospital Henri Mondor AP-HP, University of Paris-Est Créteil (UPEC), France.

Background & Aims: Two recently developed composite models, the alpha-fetoprotein (AFP) score and Metroticket 2.0, could be used to select patients with hepatocellular carcinoma (HCC) who are candidates for liver transplantation (LT). The aim of this study was to compare the predictive performance of both models and to evaluate the net risk reclassification of post-LT recurrence between them using each model's original thresholds.

Methods: This multicenter cohort study included 2,444 adult patients who underwent LT for HCC in 47 centers from Europe and Latin America. A competing risk regression analysis estimating sub-distribution hazard ratios (SHRs) and 95% CIs for recurrence was used (Fine and Gray method). Harrell's adapted c-statistics were estimated. The net reclassification index for recurrence was compared based on each model's original thresholds.

Results: During a median follow-up of 3.8 years, there were 310 recurrences and 496 competing events (20.3%). Both models predicted recurrence, HCC survival and survival better than Milan criteria (p <0.0001). At last tumor reassessment before LT, c-statistics did not significantly differ between the two composite models, either as original or threshold versions, for recurrence (0.72 vs. 0.68; p = 0.06), HCC survival, and overall survival after LT. We observed predictive gaps and overlaps between the model's thresholds, and no significant gain on reclassification. Patients meeting both models ("within-ALL") at last tumor reassessment presented the lowest 5-year cumulative incidence of HCC recurrence (7.7%; 95% CI 5.1-11.5) and higher 5-year post-LT survival (70.0%; 95% CI 64.9-74.6).

Conclusions: In this multicenter cohort, Metroticket 2.0 and the AFP score demonstrated a similar ability to predict HCC recurrence post-LT. The combination of these composite models might be a promising clinical approach.

ELSEVIER

EASL
The Home of Hepatology

Impact and implications: Composite models were recently proposed for the selection of liver transplant (LT) candidates among individuals with hepatocellular carcinoma (HCC). We found that both the AFP score and Metroticket 2.0 predicted post-LT HCC recurrence and survival better than Milan criteria; the Metroticket 2.0 did not result in better reclassification for transplant selection compared to the AFP score, with predictive gaps and overlaps between the two models; patients who met low-risk thresholds for both models had the lowest 5-year recurrence rate. We propose prospectively testing the combination of both models, to further optimize the LT selection process for candidates with HCC.

Introduction

For more than 20 years, listing of patients with hepatocellular carcinoma (HCC) for liver transplantation (LT) has been based on the Milan criteria.[1] However, recently, composite predictive models, including size and number of tumor nodules, and serum alpha-fetoprotein (AFP) levels, have been shown to outperform Milan criteria for the prediction of post-LT outcomes[1] and have been proposed as next-generation selection tools in recently published international guidelines.[2,3] These new composite models include the French AFP score, which was adopted nationally by the French organ sharing organization in 2013 after external assessment in a prospective cohort, and the Metroticket 2.0 model, published in 2018.[4,5]

The AFP score was developed to address HCC recurrence,[4] whereas Metroticket 2.0 was developed to capture HCC-specific survival through competing risk assessment.[5] External evaluation of the AFP model was performed in other European,[6] Latin American,[7] and Asian populations,[8] whereas Metroticket 2.0 was designed in an Italian multicenter derivation cohort and externally validated in a monocentric Chinese cohort of individuals with hepatitis B.[9,10] With these composite models now available, a major issue is comparing their respective performances, and the gain on risk reclassification using specific thresholds or cut-offs, before considering a change in local or regional organ sharing organization rules and adopting them for selection of LT candidates.

It has been suggested that Metroticket 2.0 has superior discriminative power compared to the AFP score for the evaluation of post-LT survival in patients transplanted for HCC.[5] However, this comparison was performed using Metroticket 2.0 as a multivariable model that included continuous variables,[11] but not a specific clinical threshold for LT selection. As a consequence, the net gain on clinical reclassification of risks when using LT selection criteria (within or beyond approach) was not assessed.[12,13] The aim of the present study was therefore to further compare the AFP score and Metroticket 2.0 in a large multinational multicenter cohort, to reassess their respective performance for post-LT HCC recurrence, HCC survival and overall survival, and to evaluate the net reclassification between models using each model's original thresholds.

Patients and methods

This was a multicenter, multinational cohort study of consecutive adult patients with HCC who underwent LT in 47 different centers from Europe and Latin America. For this purpose, four databases including patients transplanted with HCC from France, Italy and Belgium between 2000 and 2018 and from Latin America (Argentina, Uruguay, Chile, Brazil, Ecuador, Colombia and Mexico) between 2005 and 2018 were considered. These four regional databases were merged, harmonized, quality controlled and hosted on a central server, following the agreement of all participating centers. Data was reviewed center-by-center and region-by-region in a step-by-step process that took several months. This final database was named the "Western-Latin American HCC LT Consortium", approved by the Austral University committee (CIE 17-065) and registered as part of an open public registry (NCT03775863; www.clinicaltrials.gov). All procedures were followed in accordance with STROBE and REMARK guidelines,[14,15] and complied with ethical standards following the revised version of the Helsinki Declaration in 2008. Each investigator was subject to a confidentiality agreement.

We consecutively included adult patients with HCC who underwent a first LT and were followed after transplantation. The pre-transplant HCC diagnosis was made according to international guidelines.[2,16] We excluded patients if: 1) extrahepatic or macrovascular tumor invasion was observed during pre-transplant evaluation, 2) incidental HCC was found at explant pathology, 3) there were tumors other than HCC found in the explant, and 4) they were included in the French training cohort in which the AFP score was developed.[4]

Exposure variables at HCC diagnosis, at listing and during the waiting list period

Common exposure variables in all cohorts consisted of recipient characteristics, radiological tumor burden on pre-LT CT or MRI, including number and diameter of each HCC nodule, paired with AFP serum levels at HCC diagnosis, at time of listing and at last tumor reassessment during the waitlist period when available.

Patients were classified according to the Milan criteria, the AFP score and Metroticket 2.0, at listing, and then at last pre-transplant assessment when appropriate. The Milan criteria were the common standardized criteria used for patient selection in all centers but, according to local practices and allocation policies, transplantation for patients exceeding Milan was also considered and discussed at each transplant center on a case-by-case basis. The AFP score (0 to 9 points) was calculated depending on largest tumor diameter (≤3 cm = 0 points, 3-6 cm = 1 point, >6 cm = 4 points), number of HCC nodules (1-3 nodules = 0 points, ≥4 nodules = 2 points), and AFP levels ng/ml (≤100 = 0 points, 101-1000 = 2 points, and >1,000 = 3 points).[4] Metroticket 2.0 included the sum of the largest nodule diameter with the total number of HCC nodules, and the log10 AFP values.[5] These transplant criteria were compared as originally described and using their thresholds or cut-offs at time of listing, including within/beyond Milan,[1] AFP scores ≤2/>2 points,[4] and the three thresholds as originally proposed for Metroticket 2.0, including Up-to 7 plus AFP below 200 ng/ml, Up-to 5 plus AFP below 400 ng/ml and Up-to 4 plus AFP below 1,000 ng/ml.[5]

In addition, in patients receiving bridging therapies during the waiting list period, last available pre-transplant radiologic tumor assessment and AFP values following these procedures were also registered. Patients without any reassessment or those transplanted within 3 months of listing were not included in the

final pre-transplantation reassessment analysis. Tumor treatment and type of bridging therapies before transplantation were decided at each transplant center.

Endpoints and statistical analysis

The primary endpoint analyzed was post-LT HCC recurrence because it was considered the most important event affecting HCC-specific survival after transplantation.[5] Although there is still not a clear consensus on the optimal surveillance algorithm,[3] monitoring for post-transplant HCC recurrence consisted of CT or MRI and serum AFP assays once every 6 months in most centers.[17] Recurrence was assessed based on imaging criteria plus serum AFP or by biopsy. Secondary endpoints were HCC survival and overall post-LT survival. All patients were followed until death or last outpatient visit.

The cumulative incidence of recurrence was estimated in a competing risk framework, with death without HCC as a competing event, and the association of exposures with the risk of recurrence was analyzed using a Fine-Gray model, estimating sub-distribution hazard ratios (SHRs) and 95% CIs.[18] The proportional hazards assumption for competing risk regression was evaluated using the Grambsch-Thernau test. Each model's performance was compared including calibration and discrimination. Calibration was assessed comparing observed and predicted risk curves and discrimination with Harrell's adapted c-statistics for competing risk analysis was estimated for both, original (as originally proposed) and each model's threshold.[19] We used the *somersd* command to estimate the 95% CIs and compare the concordance statistics between models. First, we estimated the inverse hazard estimates, changing the coding for censored and uncensored lifetime observations. Afterwards, we compared the inverse sub-distribution hazards. The net reclassification index (NRI) was estimated to evaluate and quantify the agreement between "upward" and "downward" risk reclassifications and event status using each model's threshold.[12,13] In order to test independent variables associated with HCC survival, we again use the Fine and Gray method, considering non-HCC-related deaths as competing events.

For overall survival analysis after transplantation, Kaplan Meier survival curves were compared using the log-rank test (Mantel-Cox), and hazard ratios (HR) and 95% CIs were estimated using a Cox proportional regression model. The proportional hazards assumption was evaluated through graphical diagnostics and the Schoenfeld residual test. The performance of each model was evaluated in terms of calibration (observed *vs.* predicted curves) and discriminative power (Harrell's c-index).[20]

A planned sensitivity analysis according to pre-LT tumor reassessment, and LT periods of time was performed, estimating c-statistics for each model (LT periods 2000-2005, 2006-2011 and 2011-2018). Collected data were analyzed with StataBE v.17 and R software (Supplementary CTAT table).

Results

A total of 2,444 patients who underwent LT in 47 LT centers were included, 55.6% from Europe (n = 1,359) and 44.4% from Latin America (n = 1,085). The cohort was also categorized based on the period of transplantation: with 24.3% of transplants performed between 2000 and 2005 (n = 594), 40.0% between 2006 and 2011 (n = 978), and 35.7% between 2012 and 2018 (n = 872). Baseline patient and tumor characteristics at time of listing and at last pre-LT reassessment are shown in Table 1.

At time of listing, 80.9% of the patients (n = 1,978) were within the Milan criteria, 88.7% of the patients had an AFP score ≤2 points at listing (n = 2,163) (Table 1). According to each of the three Metroticket 2.0 thresholds,[5] 83.1% (n = 2,026) met the Up-to 7 plus AFP <200 ng/ml cut-off, 66.4% (n = 1,620) the Up-to 5 plus AFP <400 ng/ml cut-off and 44.5% (n = 1,087) the Up-to 4 plus AFP <1,000 ng/ml cut-off. Overall survival and recurrence rates at 1, 5 and 10 years of follow-up are shown in Fig. S1. Only five patients had missing AFP values at listing, in whom composite models could not be assessed. In 70.1% of the patients (n = 1,713), bridging therapies were received during the waiting list period, with a median time from last therapy to LT of 4.9 months (IQR 2.1-9.7). Overall, median follow-up in the entire cohort was 3.8 years (IQR 2.4-5.5 years).

Performance of the AFP score and metroticket 2.0 in evaluating post-transplant outcomes

At time of listing, patients within Milan but exceeding either the AFP score or the Metroticket 2.0 presented higher recurrence and lower post-transplant survival rates than patients within Milan but fulfilling the AFP score or the Metroticket 2.0 model. On the contrary, among patients exceeding Milan, both composite models identified a population with excellent outcomes after LT (Fig. S2A,B).

Similar results were observed when evaluating the effect and risk categorization for HCC recurrence in patients with tumor reassessment before transplantation (n = 1,377). No significant difference in the risk of recurrence was observed between patients exceeding Milan criteria compared to those within Milan criteria with AFP scores ≤2 (SHR 1.32; 95% CI 0.74-2.37; *p* = 0.34). Increasing recurrence risk was observed in patients either within (SHR 2.24; 95% CI 1.26-4.00; *p* = 0.006) or beyond (SHR 4.20; 95% CI 2.74-6.43; *p* <0.0001) the Milan criteria with AFP scores higher than 2 points. Similarly, patients exceeding Milan but within Metroticket 2.0 had a similar risk as patients within the Milan criteria (SHR 1.51; 95% CI 0.77-2.94; *p* = 0.22). On the other hand, patients within (SHR 4.43; 95% CI 2.90-6.77) or beyond (SHR 4.85; 95% CI 3.17-7.44) the Milan criteria but exceeding Metroticket 2.0 were at a higher risk of recurrence compared to those who met the Milan and Metroticket 2.0 criteria, respectively (*p* <0.0001). Time on the waiting list (as a continuous variable) and locoregional treatment were not independently associated with HCC recurrence.

Discrimination and net reclassification of risk of recurrence

A better discriminative power for risk of recurrence was observed for both composite models at time of listing when compared to the Milan criteria, either as original or threshold models (Table 2). There was not a significant change on each model's performance when including the effect of waiting list time and locoregional therapy.

Also, discrimination of HCC survival (Table S1) and overall survival (Table S2) was significantly better with each composite model.

When comparing both composite models at time of listing as originally proposed,[4,5] although Metroticket 2.0 showed a higher discriminative power for HCC recurrence compared to the AFP score, there were no significant differences when comparing each model's thresholds (Table 2). Also, there were not significant differences for HCC survival and overall survival.

Stratified by LT periods, there was not a significant difference in c-statistics between both composite models (Table S3). Similarly, at last tumor reassessment, c-statistics for both originally proposed models, the AFP score (0.68; 95% CI 0.63-0.73) and

Table 1. Patient and tumor characteristics.

Variable	Values
Age, years (±SD)	58 ± 8.1
Sex, male, n (%)	1,968 (80.5)
Median time on waiting list, (IQR), months	5.5 (2.5-10.5)
MELD, (±SD)*	12.2 ± 4.7
Non-cirrhotic liver, n (%)	21 (0.9)
Child Pugh A/B/C, n (%)	1,180 (48.3)/930 (38.0)/334 (13.7)
Etiology of liver disease, n (%)	
Hepatitis C	1,153 (47.2)
Hepatitis B	234 (9.6)
HCV-HBV co-infection	9 (0.4)
Alcohol	609 (24.9)
Other	439 (18.0)
Variables at time of listing*	
Number of HCC nodules, median (IQR)	1 (1-2)
1-3 nodules, n (%)	2,275 (93.1)
≥4 nodules, n (%)	169 (6.9)
Major nodule diameter (cm), median (IQR)	2.7 (2.0-3.5)
≤3 cm, n (%)	1,593 (65.2)
3-6 cm, n (%)	780 (31.9)
>6 cm, n (%)	70 (2.9)
AFP, ng/ml, median (IQR)	9.1 (4.2-33.9)
AFP ≤100 ng/ml, n (%)	2,089 (85.6)
AFP 101-1,000 ng/ml, n (%)	294 (12.0)
AFP >1,000 ng/ml, n (%)	57 (2.3)
Within Milan criteria, n (%)	1,978 (80.9)
AFP score, median (IQR)	0 (0-1)
≤2 points, n (%)	2,163 (88.7)
>2 points, n (%)	276 (11.3)
Metroticket 2.0 up-to 7 + AFP <200 ng/ml, n (%)	
Within	2,026 (83.1)
Beyond	413 (16.9)
Metroticket 2.0 up-to 5 + AFP <400 ng/ml, n (%)	
Within	1,620 (66.4)
Beyond	820 (33.6)
Metroticket 2.0 up-to 4 + AFP <1,000 ng/ml, n (%)	
Within	1,087 (44.5)
Beyond	1,354 (55.5)
Any locoregional treatment, n (%)	1,713 (70.1)
TACE	1,419
RFA	351
PEI	123
LR	107
Variables at last tumor reassessment*	**1,377 (56.0))**
Months from last tumor evaluation to LT, median (IQR)	2.4 (1.1-4.4)
Number of HCC nodules, median (IQR)	1 (1-2)
1-3 nodules, n (%)	1,303 (94.6)
≥4 nodules, n (%)	74 (5.4)
Major nodule diameter (cm), median (IQR)	2.6 (1.9-3.5)
≤3 cm, n (%)	910 (66.1)
3-6 cm, n (%)	441 (32.0)
>6 cm, n (%)	26 (1.9)
AFP (ng/ml), median (IQR)	9.0 (4.1-32.9)
≤100 ng/ml, n (%)	1,127 (83.2)
101-1,000 ng/ml, n (%)	175 (12.9)
>1,000 ng/ml, n (%)	53 (3.9)
Within Milan criteria, n (%)	1,165 (84.6)
AFP score, median (IQR)	0 (0-1)
≤2 points, n (%)	1,203 (87.9)
>2 points, n (%)	166 (12.1)
Metroticket 2.0 up-to 7 + AFP <200 ng/ml, n (%)	
Within	1,121 (82.5)
Beyond	237 (17.4)
Metroticket 2.0 up-to 5 + AFP <400 ng/ml, n (%)	
Within	920 (67.7)
Beyond	439 (32.3)
Metroticket 2.0 up-to 4 + AFP <1,000 ng/ml, n (%)	
Within	644 (47.1)

(continued on next page)

Table 1 (*continued*)

Variable	Values
Beyond	722 (52.9)
Explant pathology findings	
Up-to 7, n (%)	
Within	1,569 (64.2)
Beyond	875 (35.8)
Presence MVI, n (%)	618 (25.3)
Tumor differentiation, n (%)	
Nuclear grade I-II	1,756 (79.5)
Nuclear grade >II	452 (20.5)

AFP, alpha-fetoprotein; LR, liver resection; MELD, model for end-stage liver disease; MVI, microvascular invasion; PEI, percutaneous ethanol injection; RFA, radio-frequency ablation; SHR, sub-hazard ratio; TACE, trans-arterial chemoembolization.
* Laboratory MELD score before liver transplant.

Metroticket 2.0 (0.72; 95% CI 0.68-0.77), were significantly better than for the Milan criteria (0.56; 95% CI 0.52-0.60; p <0.0001) (Table 2). However, there was no significant difference when comparing the AFP score and Metroticket 2.0 as original models (p = 0.065).

Metroticket 2.0 did not lead to a significant reclassification of risk compared to the AFP score cut-off value at time of listing nor at last tumor reassessment (Table 3 and Table S4). Although Metroticket 2.0 gained on sensitivity, a decreasing specificity was observed according to each threshold.

Gaps and overlaps between both composite models

We further conducted a stratified analysis according to the AFP score and each Metroticket 2.0 threshold at time of listing. Gaps and overlaps between the two models were observed for each threshold (Fig. S3). First, according to the first Metroticket 2.0 threshold (sum up-to 7 plus AFP <200 ng/ml), in 8.7% of the population (n = 211) this Metroticket 2.0 threshold did not categorize the risk of recurrence in patients with AFP scores ≤2 or >2 points. Patients meeting this Metroticket 2.0 threshold with AFP scores >2 points presented similar 5-year recurrence rates as patients exceeding Metroticket 2.0 with AFP scores ≤2. The models clearly separated two distinct populations, those patients meeting both models showed the lowest cumulative incidence of HCC recurrence (14.2%; 95% CI 12.1-16.5) and the highest 5-year post-LT survival rate (67.9%; 95% CI 65.3-70.4) (Fig. 1A-B). Furthermore, according to the other two Metroticket 2.0 thresholds, again, higher recurrence rates were observed in patients exceeding one or the other corresponding threshold of each model, and the lowest cumulative recurrence rates were observed in those patients meeting both of each model's thresholds (Fig. 2A,B).

Similarly, at last pre-LT assessment, lower 5-year recurrence rates were observed in patients meeting both models criteria and, on the contrary, higher cumulative 5-year recurrence rates were observed in those patients exceeding both models for each Metroticket 2.0 threshold (Table 4).

"Within-ALL" clinical-decision algorithm

Taking into account the aforementioned gaps and overlaps, we considered using both composite models for selection of LT candidates at time of listing in a "within-ALL" clinical-decision algorithm (*Yes/No approach*), but stratified according to the Milan criteria. Patients meeting both composite models' thresholds either within or beyond the Milan criteria, showed the best post-LT outcomes with a lower risk of HCC recurrence (SHR of 0.28; 95% CI 0.22-0.36; p <0.0001), whereas patients exceeding both composite models, even meeting the Milan

Table 2. Discrimination power of HCC recurrence between Milan criteria, the AFP model and Metroticket v2.0 criteria at time of listing.

Models	Harrell's adapted c-index (95% CI)	p value vs. Milan as reference	p value vs. AFP model as reference
Original models at listing			
Milan criteria	0.59 (0.56-0.62)		−
AFP score*	0.66 (0.63-0.69)	<0.0001	−
Metroticket 2.0**	0.69 (0.66-0.72)	<0.0001	0.04
Threshold models at listing			
Milan criteria	0.59 (0.56-0.62)		−
AFP model ≤2 points	0.60 (0.58-0.63)	<0.0001	−
Up-to 7 + AFP <200 ng/ml	0.62 (0.59-0.65)	<0.0001	0.10
Up-to 5 + AFP <400 ng/ml	0.60 (0.57-0.63)	<0.0001	0.96
Up-to 4 + AFP <1,000 ng/ml	0.58 (0.56-0.61)	<0.0001	0.20
Original models at last reassessment			
Milan criteria	0.56 (0.52-0.60)		−
AFP score*	0.68 (0.63-0.73)	<0.0001	−
Metroticket 2.0**	0.72 (0.62-0.77)	<0.0001	0.065

AFP, alpha-fetoprotein; HCC, hepatocellular carcinoma.
* AFP as originally proposed by Duvoux *et al.*
** Metroticket 2.0 as originally reported by Mazzaferro *et al.* Level of significance, *p* <0.05.

Table 3. Net reclassification index considering competing events for risk reclassification of recurrence (events) and non-recurrence (non-events).

	Overall NRI	Z value	p value	Up events relative change in sensitivity (95% CI)	Up non-events relative change in specificity (95% CI)
AFP score *vs.*				29.0%	91.7%
Up-to 7 + AFP <200	-0.033	−1.52	0.064	−8.5%	+5.1%
Up-to 5 + AFP <400	-0.006	−0.21	0.41	−22.5%	+21.8%
Up-to 4 + AFP <1,000	0.023	0.59	0.72	−42.0%	+44.4%

All models assessed at listing. The NRI was estimated to evaluate the ability of these models to discriminate between events and non-events by quantifying the agreement between "upward" and "downward" risk reclassifications and event status.[12,13] Level of significance, *p* value <0.05.
AFP, alpha-fetoprotein; NRI, net reclassification index.

criteria, showed significantly higher risk of post-LT recurrence (Fig. 3A). In addition, survival rates were similar in patients meeting the "within-ALL" clinical-decision algorithm whether within or beyond the Milan criteria, and were significantly higher than in patients beyond the "within-ALL" clinical-decision algorithm (Fig. 3B). Also, at last tumor reassessment, patients meeting both composite models presented the lowest 5-year cumulative incidence of HCC recurrence (7.7%; 95% CI 5.1-11.5)

(SHR 0.33; 95% CI 0.22-0.49; p <0.0001) (Fig. 4) and higher 5-year post-LT survival rates (70.0%; 95% CI 64.9-74.6).

Discussion

In this multinational cohort study of transplanted patients with HCC, we compared the ability of two composite models, the AFP score and Metroticket 2.0, to predict post-LT outcomes. Three
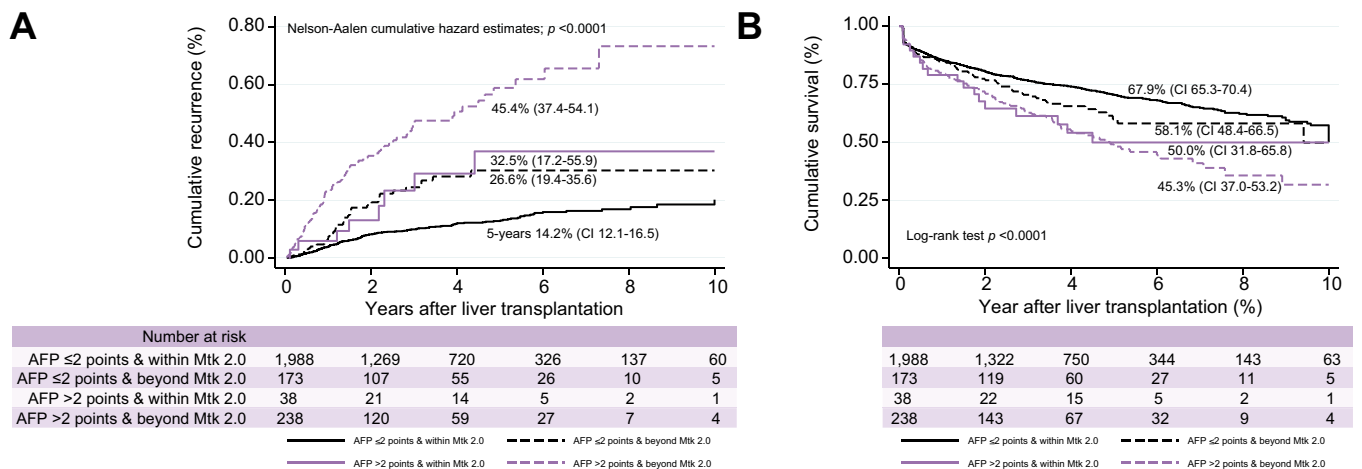
Fig. 1. Stratified analysis according to the AFP model and Metroticket 2.0. The first threshold for Metroticket 2.0 is shown here (sum up-to 7 plus AFP <200 ng/ml). (A) Nelson-Aelen cumulative hazard estimates. Corresponding cumulative post-LT recurrence at 5 years for patients meeting both model's thresholds, patients with AFP scores ≤2 and exceeding Metroticket 2.0 (SHR 2.2; 95% CI 1.51-3.13; p <0.0001), patients with AFP scores >2 and meeting Metroticket 2.0 (SHR 2.2; 95% CI 1.12-4.53; p = 0.02), and for patients with AFP scores >2 and exceeding Metroticket 2.0 (SHR 4.4; 95% CI 3.38-5.74; p <0.0001). (B) Corresponding 5-year cumulative survival for each stratum. Level of significance, p <0.05. AFP, alpha-fetoprotein; SHR, subdistribution hazard ratio; LT, liver transplantation.
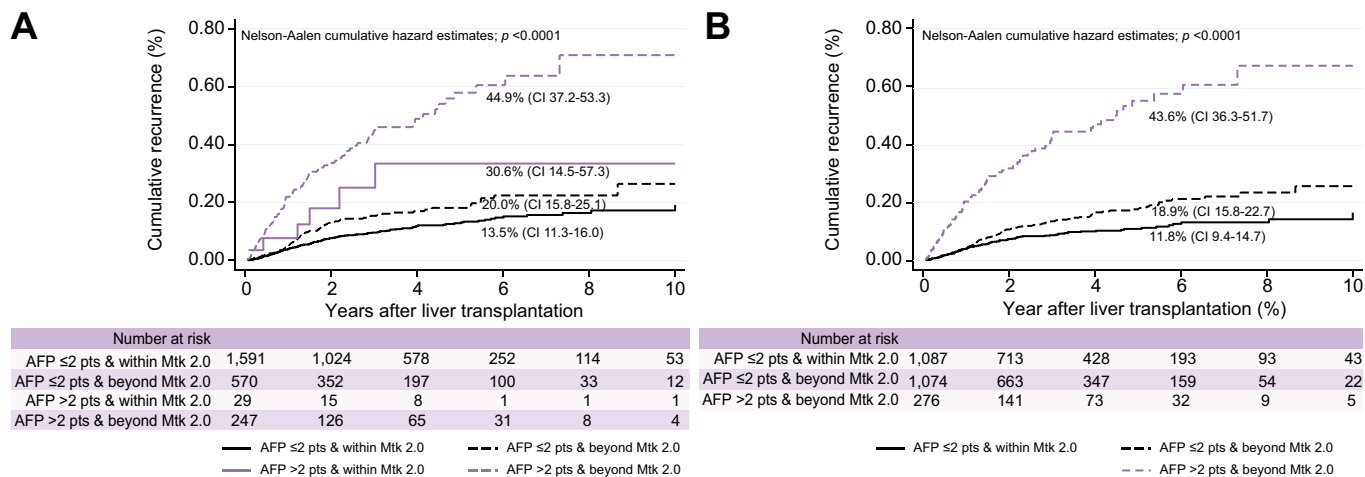
Fig. 2. Cumulative recurrence according to the AFP model and the other two thresholds of Metroticket 2.0. The other two thresholds for the Metroticket 2.0 are shown here. (A) Sum up-to 5 plus AFP <400 ng/ml: Nelson-Aelen cumulative hazard estimates. Corresponding cumulative post-LT recurrence at 5 years for patients meeting both model's thresholds, for patients with AFP scores ≤2 and exceeding Metroticket 2.0 (SHR 1.5; 95% CI 1.36-1.98; p = 0.004), for patients with AFP scores >2 and meeting Metroticket 2.0 (SHR 2.4; 95% CI 1.06-5.48; p = 0.04), and for patients with AFP scores >2 and exceeding Metroticket 2.0 (SHR 4.3; 95% CI 3.38-5.82; p <0.0001). (B) Sum up-to 4 plus AFP <1,000 ng/ml: Corresponding cumulative post-LT recurrence at 5 years for patients meeting both model's thresholds, for patients with AFP scores ≤2 and exceeding Metroticket 2.0 (SHR 1.5; 95% CI 1.17-2.01; p = 0.002), there were no patients with AFP scores >2 and meeting Metroticket 2.0, and for patients with AFP scores >2 and exceeding Metroticket 2.0 (SHR 4.7; 95% CI 3.48-6.30; p <0.0001). Level of significance, p <0.05. AFP, alpha-fetoprotein; SHR, subdistribution hazard ratio; LT, liver transplantation.

Table 4. Recurrence risk categorization in patients with tumor reassessment before transplantation.

Variables	Gaps and overlaps	5-year recurrence (95% CI)	SHR (95% CI); p values
AFP score ≤2 points, n = 1,203 (88.9%) Metroticket 2.0 up-to 7 + AFP <200 ng/ml			
Within, n (%)	1,101 (92.4)	10.9 (8.5-14.1)	0.21 (0.14-0.34); p <0.0001
Beyond, n (%)	91 (7.6)	35.2 (25.0-48.0)	
AFP score >2 points, n = 166 (12.1%) Metroticket 2.0 up-to 7 + AFP <200 ng/ml			
Within, n (%)	20 (12.1)	6.2 (0.9-36.8)	0.17 (0.02-1.28); p = 0.08
Beyond, n (%)	146 (87.9)	36.0 (27.0-46.8)	
AFP score ≤2 points, n = 1,203 (88.9%) Metroticket 2.0 up-to 5 + AFP <400 ng/ml			
Within, n (%)	902 (75.7)	10.7 (8.0-14.2)	0.46 (0.31-0.70); p <0.0001
Beyond, n (%)	290 (24.3)	20.2 (14.9-27.1)	
AFP score >2 points, n = 166 (12.1%) Metroticket 2.0 up-to 5 + AFP <400 ng/ml			
Within, n (%)	18 (10.9)	20.0 (4.9-62.6)	0.39 (0.1-1.58); p = 0.19
Beyond, n (%)	147 (89.1)	34.8 (26.0-45.6)	
AFP score ≤2 points, n = 1,203 (88.9%) Metroticket 2.0 up-to 4 + AFP <1,000 ng/ml			
Within, n (%)	644 (53.8)	9.9 (7.1-13.7)	0.66 (0.45-0.99); p = 0.05
Beyond, n (%)	553 (46.2)	17.0 (12.8-22.3)	
AFP score >2 points, n = 166 (12.1%) Metroticket 2.0 up-to 4 + AFP <1,000 ng/ml			
Within, n (%)	0 (-)	—	—
Beyond, n (%)	166 (100)	34.1 (27.8-53.3)	—

Level of significance, p value <0.05; Wald test.
Normal values: AFP 0.6-4.4 ng/ml.
AFP, alpha-fetoprotein; HCC, hepatocellular carcinoma; MVI, microvascular invasion; SHR, sub-hazard ratio.

major results can be drawn from this study. First, we confirmed that both models performed better for the prediction of HCC recurrence and survival than the Milan criteria. This finding suggests that the selection of LT candidates with HCC could be optimized by using composite models, as recently proposed by the European Association for the Study of the Liver.[2] Second, composite models performed similarly, in terms of intrinsic predictive values. Third, although the two models performed similarly, we observed gaps and overlaps in risk stratification. Consequently, using a clinical decision-approach combining each model's thresholds, we showed that patients meeting both models, especially at last reassessment, had the best post-LT outcomes.

Some technical issues should also be underlined when designing prognostic models.[21] Interestingly, the two models

A

B
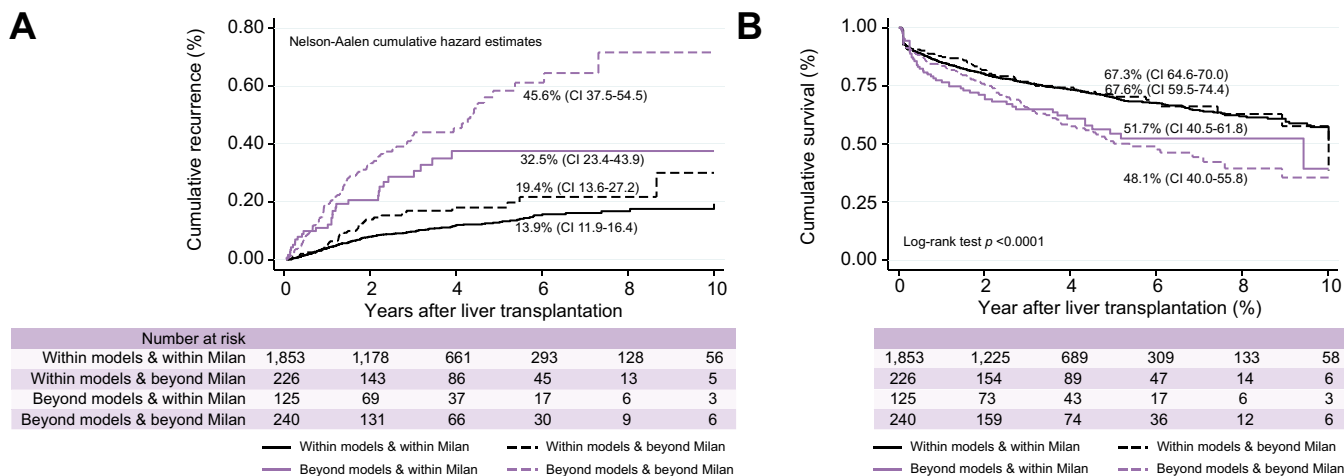
Fig. 3. Clinical use of the "within-ALL" decision algorithm. (A,B) Using both models, patients meeting or exceeding Milan criteria and within the AFP score and each of the Metroticket thresholds at time of listing showed lower recurrence rates (A) and better survival (B). Log-rank test, level of significance, p <0.05. AFP, alpha-fetoprotein.
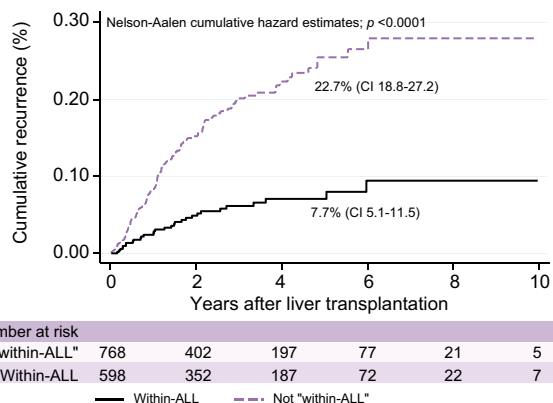
Fig. 4. Cumulative recurrence rates according to the "within-ALL" decision algorithm at last tumor reassessment. Nelson-Aelen cumulative hazard estimates at last tumor reassessment. Level of significance, p <0.05.

performed similarly to predict post-LT outcomes, although they were not designed in the same way. The AFP score was developed to address HCC recurrence,[4] whereas the Metroticket 2.0 was designed to capture HCC-specific survival.[5] Second, the Metroticket 2.0 included two continuous exposure variables, sum of number and tumor diameter, and AFP serum values, which in a mathematical model might lead to better predictive power.[11] Continuous data may lead to better discrimination compared to dummy or categorical variables, in which each stratum is "artificially" constructed and represents different risk classification. Third, Metroticket 2.0 was developed and validated in an HBV population, whereas the most frequent etiologies in this cohort were HCV and alcohol-related liver disease. Also, median AFP values, as well as median major nodule diameter and total number of tumor were similar between studies.[11]

However, in the end, we did not observe significantly better reclassification, that is to say, no significant clinical impact, regarding risk of HCC recurrence between these two composite models. Although Metroticket 2.0 was superior in terms of

sensitivity, it exhibited reduced specificity. Of note, both models showed discriminative power, with c-statistics of around 0.70. This underlines the need for further refinement of selection models.[21] Optimization of AFP thresholds and response to bridging therapy should probably be considered,[22] although we did not observe significant improvements in discrimination when this variable was included. Other models, such as the MORAL model,[23] the BALAD or GALAD score,[24] may perform adequately but have not been prospectively evaluated. Also, continuous scoring models, including the HALT-HCC,[25] showed similar discriminative power as these composite models but we could not directly compare them. While we wait for next-generation predictive tools based on molecular signatures, the available composite models offer a reasonable, reproducible, and user-friendly alternative as we move beyond the Milan criteria.

Using the alternative "within-ALL" decision process in this cohort, almost half of patients exceeding Milan could be included on the waiting list, while 6% of those within Milan could be excluded or granted better access to LT. Moreover, when considering changes over the waitlist period (including the effect of time and locoregional therapies), at last tumor evaluation prior to transplantation, meeting or exceeding this clinical decision point identified two populations at significantly different risks of recurrence. This shows how the clinical transplant decision should be reassessed over the waitlist period to further determine the risk of recurrence. The "within-ALL" strategy was proposed as clinical-decision-making algorithm, rather than a new predictive model. It aims to combine both composite model's thresholds to select the best candidates for LT who may be granted with MELD exception points, increasing their chances of receiving a transplant. Based on this critical finding, rather than opposing these two models, we propose considering both composite criteria to select patients with the best phenotype of HCC selection criteria. However, this depends not only on the pool of deceased donors, but also on tumor behavior over the waitlist period. This proposal will have to be confirmed by prospective cohort studies.

This present study has limitations common to observational studies, particularly on retrospective cohorts. First, cohorts from different regions of the world were merged. However, data

collection and outcome definitions were homogenous across cohorts, quality controlled, and there was not a differential assessment of the outcomes. Second, although we included cohorts in which the AFP score was externally validated, only one-third of the Latin American cohort was included in the prior external validation of the AFP score.[7] Other limitations include the relatively small number of patients with AFP scores >2 or exceeding the Metroticket 2.0 thresholds. We did not consider mRECIST or residual enhancement of treated tumors in order to avoid misinterpretation of necrotic or enhancing residual areas across centers. Unfortunately, based on the retrospective nature of this research, the independent effect of bridging therapies could not be completely analyzed. Finally, tumor reassessment was available in 50% of the study cohort, notably in patients transplanted before 2006, in part because of a short waiting time. We initially focused on data at time of listing because this time-point addresses the critical step in HCC candidate selection. It was therefore of major importance to reassess the discriminative power of the models in the subset of patients who have waited long enough to undergo tumor reassessment close to LT. In this last scenario, we again observed similar results.

In conclusion, in this multinational cohort study, both composite models shared similar predictive performances and showed better discrimination than the Milan criteria. However, gaps and overlaps were observed between each model when considering proposed thresholds, a limitation that can be counteracted by testing both models to guide the clinical decision-making process. Waiting for new biomarkers, we therefore suggest a clinical-decision algorithm, the "within-ALL" selection process; patients meeting both composite models achieve the best outcomes, broadening the scope for further optimization of candidate selection.

Authors' contributions

All the authors approved the final version of the manuscript. *Concept and study design:* Federico Piñero, Charlotte Costentin, Helena Degroote, Quirino Lai, Fernando Rubinstein and Christophe Duvoux. *Data collection: France:* Karim Boudjema, Philippe Bachellier, Filomena Conti, Olivier Scatton, Fabrice Muscari, Ephrem Salame, Pierre Henri Bernard, Claire Francoz, Francois Durand, Sébastien Dharancy, Marie-lorraine.Woehl, Claire Vanlemmens, Alexis Laurent, Sylvie Radenne, Jérôme Dumortier, Armand Abergel, Daniel Cherqui, Louise Barbier, Pauline Houssel-Debry, Georges Philippe Pageaux, Laurence Chiche, Victor Deledinghen, Jean Hardwigsen, J Gugenheim, M altieri, Marie Noelle Hilleret, Thomas Decaens, Christophe Duvoux. *Latin America:* Federico Piñero, Aline Chagas, Paulo Costa, Elaine Cristina de Ataide, Emilio Quiñones, Sergio Hoyos Duque, Sebastián Marciano, Margarita Anders, Adriana Varón, Alina Zerega, Jaime Poniachik, Alejandro Soza, Martín Padilla Machaca, Diego Arufe, Josemaría Menéndez, Rodrigo Zapata, Mario Vilatoba, Linda Muñoz, Ricardo Chong Menéndez, Martín Maraschio, Luis G Podestá, Lucas McCormack, Juan Mattera, Adrian Gadano, Ilka SF Fatima Boin, Jose Huygens Parente García, Flair Carrilho, Marcelo Silva. *Italy:* Andrea Notarpaolo, Giulia Magini, Lucia Miglioresi, Martina Gambato, Fabrizio Di Benedetto, Cecilia D'Ambrosio, Giuseppe Maria Ettorre, Alessandro Vitale, Patrizia Burra, Stefano Fagiuoli, Umberto Cillo, Michele Colledan, Domenico Pinelli, Paolo Magistri, Giovanni Vennarecci, Marco Colasanti, Valerio Giannelli, Adriano Pellicelli, Cizia Baccaro, Quirino Lai. *Belgium:* Helena Degroote, Hans Van Vlierberghe, Callebout Eduard, Iesari Samuele, Dekervel Jeroen, Schreiber Jonas, Pirenne Jacques, Verslype Chris, Ysebaert Dirk, Michielsen Peter, Lucidi Valerio, Moreno Christophe, Detry Olivier, Delwaide Jean, Troisi Roberto, Lerut Jan Paul. *Statistical analysis and revision:* Federico Piñero, Fernando Rubinstein, Quirino Lai. *Writing the article:* Federico Piñero, Charlotte Costentin, Helena Degroote, Quirino Lai and Christophe Duvoux. *Critical revision:* all authors.

Data availability statement

The data shown in this article are available from the corresponding authors upon a reasonable request.

References

[1] Mazzaferro VV, Regalia EE, Doci RR, Andreola SS, Pulvirenti AA, Bozzetti FF, et al. Liver transplantation for the treatment of small hepatocellular carcinomas in patients with cirrhosis. N Engl J Med 1996;334:693–699. https://doi.org/10.1056/NEJM199603143341104.

[2] Galle PR, Forner A, Llovet JM, Mazzaferro V, Piscaglia F, Raoul J-L, et al. EASL clinical practice guidelines: management of hepatocellular carcinoma. J Hepatol 2018;69:182–236. https://doi.org/10.1016/j.jhep.2018.03.019.

[3] Berenguer M, Burra P, Ghobrial M, Hibi T, Metselaar H, Sapisochin G, et al. Posttransplant management of recipients undergoing liver transplantation for hepatocellular carcinoma. Working group report from the ILTS transplant oncology consensus conference. Transplantation 2020;104:1143–1149. https://doi.org/10.1097/TP.0000000000003196.

[4] Duvoux C, Thoraval FR, Decaens T, Pessione F, Badran H, Piardi T, et al. Liver transplantation for hepatocellular carcinoma: a model including -fetoprotein improves the performance of milan criteria. Gastroenterology 2012;143:986–994.e3. https://doi.org/10.1053/j.gastro.2012.05.052.

[5] Mazzaferro V, Sposito C, Zhou J, Pinna AD, De Carlis L, Fan J, et al. Metroticket 2.0 model for analysis of competing risks of death after liver transplantation for hepatocellular carcinoma. Gastroenterology 2018;154:128–139. https://doi.org/10.1053/j.gastro.2017.09.025.

[6] Notarpaolo A, Layese R, Magistri P, Gambato M, Colledan M, Magini G, et al. Validation of the AFP model as a predictor of HCC recurrence in patients with viral hepatitis-related cirrhosis who had received a liver transplant for HCC. J Hepatol 2017;66:552–559. https://doi.org/10.1016/j.jhep.2016.10.038.

[7] Piñero F, Tisi Baña M, de Ataide EC, Hoyos Duque S, Marciano S, Varón A, et al. Liver Transplantation for Hepatocellular carcinoma: evaluation of the AFP model in a multicenter cohort from Latin America. Liver Int 2016;36:1657–1667. https://doi.org/10.1111/liv.13159.

[8] Rhu J, Kim JM, Choi GS, Kwon C-HD, Joh J-W. Validation of the α-fetoprotein model for hepatocellular carcinoma recurrence after transplantation in an asian population. Transplantation 2018;102:1316–1322. https://doi.org/10.1097/TP.0000000000002136.

[9] Lai Q, Vitale A, Halazun K, Iesari S, Viveiros A, Bhangui P, et al. Identification of an upper limit of tumor burden for downstaging in candidates with hepatocellular cancer waiting for liver transplantation: a west–east collaborative effort. Cancers 2020;12:452. https://doi.org/10.1080/01621459.1999.10474144.

[10] Cucchetti A, Serenari M, Sposito C, Di Sandro S, Mosconi C, Vicentin I, et al. Including mRECIST in the Metroticket 2.0 criteria improves prediction of hepatocellular carcinoma- related death after liver transplant. J Hepatol 2020;73:342–348. https://doi.org/10.1016/j.jhep.2020.03.018.

[11] Toll DB, Janssen KJM, Vergouwe Y, Moons KGM. Validation, updating and impact of clinical prediction rules: a review. J Clin Epidemiol 2008;61:1085–1094. https://doi.org/10.1016/j.jclinepi.2008.04.008.

[12] Kerr KF, Wang Z, Janes H, McClelland RL, Psaty BM, Pepe MS. Net reclassification indices for evaluating risk prediction instruments. Epidemiology 2014;25:114–121. https://doi.org/10.1097/EDE.0000000000000018.

[13] Pencina MJ, D'Agostino Sr RB, Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Statist Med 2010;30:11–21. https://doi.org/10.1002/sim.4780141909.

[14] Elm von E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet 2007;370:1453–1457. https://doi.org/10.1016/S0140-6736(07)61602-X.

[15] McShane LM, Altman DG, Sauerbrei W, Taube SE, Gion M, Clark GM, et al. Reporting recommendations for tumor marker prognostic studies (REMARK). JNCI J Natl Cancer Inst 2005;97:1180–1184. https://doi.org/10.1093/jnci/dji237.

[16] Heimbach JK, Kulik LM, Finn RS, Sirlin CB, Abecassis MM, Roberts LR, et al. AASLD guidelines for the treatment of hepatocellular carcinoma. Hepatology 2018;67:358–380. https://doi.org/10.1002/hep.29086/suppinfo.

[17] Kneteman N, Livraghi T, Madoff D, de Santibañez E, Kew M. Tools for monitoring patients with hepatocellular carcinoma on the waiting list and after liver transplantation. Liver Transplant 2011;17:S117–S127. https://doi.org/10.1002/lt.22334.

[18] Dignam JJ, Zhang Q, Kocherginsky M. The use and interpretation of competing risks regression models. Clin Cancer Res 2012;18:2301–2308. https://doi.org/10.1158/1078-0432.CCR-11-2097.

[19] Wolbers M, Blanche P, Koller MT, Witteman JCM, Gerds TA. Concordance for prognostic models with competing risks. Biostatistics 2014;15:526–539. https://doi.org/10.1111/j.1541-0420.2011.01671.x.

[20] Newson RB. Comparing the predictive powers of survival models using Harrell's C or Somers' D. Stata J 2010;10:339–358.

[21] Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. 2015. p. 1–10. https://doi.org/10.1186/s12916-014-0241-z.

[22] Cescon M, Cucchetti A, Ravaioli M, Pinna AD. Hepatocellular carcinoma locoregional therapies for patientsin the waiting list. Impact on transplantability and recurrence rate. J Hepatol 2013;58:609–618. https://doi.org/10.1016/j.jhep.2012.09.021.

[23] Halazun KJ, Najjar M, Abdelmessih RM, Samstein B, Griesemer AD, Guarrera JV, et al. Recurrence after liver transplantation for hepatocellular carcinoma. Ann Surg 2017;265:557–564. https://doi.org/10.1097/SLA.0000000000001966.

[24] Berhane S, Toyoda H, Tada T, Kumada T, Kagebayashi C, Satomura S, et al. Role of the galad and balad-2 serologic models in diagnosis of. CLinical Gastroenterol Hepatol 2016;14:875–886. https://doi.org/10.1016/j.cgh.2015.12.042.

[25] Firl DJ, Sasaki K, Agopian VG, Gorgen A, Kimura S, Dumronggittigule W, et al. Charting the path forward for risk prediction in liver transplant for hepatocellular carcinoma: international validation of HALTHCC among 4,089 patients. Hepatology 2019;71:569–582. https://doi.org/10.1093/oxfordjournals.aje.a009590.