HEPATOLOGY

HEPATOLOGY, VOL. 71, NO. 2, 2020

Charting the Path Forward for Risk Prediction in Liver Transplant for Hepatocellular Carcinoma: International Validation of HALTHCC Among 4,089 Patients

Daniel J. Firl,[1*] Kazunari Sasaki,[1] Vatche G. Agopian,[2] Andre Gorgen ,[3] Shoko Kimura,[4] Wethit Dumronggittigule,[2] John C. McVey,[1] Samuele Iesari,[5] Gianluca Mennini,[6] Alessandro Vitale,[7] Armin Finkenstedt,[8] Simona Onali,[9] Maria Hoppe-Lotichius,[10] Giovanni Vennarecci,[11] Tommaso M. Manzia,[12] Daniele Nicolini,[13] Alfonso W. Avolio,[14] Salvatore Agnes,[14] Marco Vivarelli,[13] Giuseppe Tisone,[12] Giuseppe M. Ettorre,[11] Gerd Otto,[10] Emmanuel Tsochatzis ,[9] Massimo Rossi,[6] Andre Viveiros,[8] Umberto Cillo,[7] James F. Markmann,[4] Toru Ikegami,[15] Toshimi Kaido ,[16] Quirino Lai,[5,6] Gonzalo Sapisochin,[3] Jan Lerut,[5] on behalf of the European Hepatocellular Cancer Liver Transplant Study Group, and Federico N. Aucejo[1]

Prognosticating outcomes in liver transplant (LT) for hepatocellular carcinoma (HCC) continues to challenge the field. Although Milan Criteria (MC) generalized the practice of LT for HCC and improved outcomes, its predictive character has degraded with increasing candidate and oncological heterogeneity. We sought to validate and recalibrate a previously developed, preoperatively calculated, continuous risk score, the Hazard Associated with Liver Transplantation for Hepatocellular Carcinoma (HALTHCC), in an international cohort. From 2002 to 2014, 4,089 patients (both MC in and out [25.2%]) across 16 centers in North America, Europe, and Asia were included. A continuous risk score using pre-LT levels of alpha-fetoprotein, Model for End-Stage Liver Disease Sodium score, and tumor burden score was recalibrated among a randomly selected cohort (n = 1,021) and validated in the remainder (n = 3,068). This study demonstrated significant heterogeneity by site and year, reflecting practice trends over the last decade. On explant pathology, both vascular invasion (VI) and poorly differentiated component (PDC) increased with increasing HALTHCC score. The lowest-risk patients (HALTHCC 0-5) had lower rates of VI and PDC than the highest-risk patients (HALTHCC > 35) (VI, 7.7%[1.2-14.2] vs. 70.6% [48.3-92.9] and PDC:4.6% [0.1%-9.8%] vs. 47.1% [22.6-71.5]; $P < 0.0001$ for both). This trend was robust to MC status. This international study was used to adjust the coefficients in the HALTHCC score. Before recalibration, HALTHCC had the greatest discriminatory ability for overall survival (OS; C-index = 0.61) compared to all previously reported scores. Following recalibration, the prognostic utility increased for both recurrence (C-index = 0.71) and OS (C-index = 0.63). *Conclusion:* This large international trial validated and refined the role for the continuous risk metric, HALTHCC, in establishing pre-LT risk among candidates with HCC worldwide. Prospective trials introducing HALTHCC into clinical practice are warranted. (HEPATOLOGY 2020;71:569-582).

Hepatocellular carcinoma (HCC) is the fifth-most common cancer by global incidence.[1,2] For most patients, liver transplantation (LT) offers the only reasonable chance at cure given that it replaces the carcinogenic background liver and leaves the diseased liver capsule and vasculature

Abbreviations: AFP, alpha-fetoprotein; CI, confidence interval; HALTHCC, Hazard Associated with Liver Transplantation for Hepatocellular Carcinoma; KM, Kaplan-Meier; LRT, locoregional therapy; LT, liver transplant; MC, Milan Criteria; MELD, Model for End-Stage Liver Disease; MELD-Na, Model of End Stage Liver Disease Sodium score; MoRAL, model of recurrence after liver transplantation; oHALTHCC, original HALTHCC score; OS, overall survival; PDC, poorly differentiated component; PDT, poorly differentiated tumor; rHALTHCC, recalibrated HALTHCC score; TBS, tumor burden score; VI, vascular invasion.

undisturbed.[3,4] Although LT is preferred, limited donor organs and poorer outcomes from cancer recurrence limited its application until the groundbreaking Milan Criteria (MC:;a single-tumor ≤5 cm or up to 3 tumors, all <3 cm).[3-5] In the subsequent two decades, MC has allowed transplant centers around the world to treat HCC patients with similar, acceptable outcomes.[2,6-8] However, several groups have suggested that MC is too stringent.[8-10] With increasing penetration of downstaging locoregional therapy (LRT), the predictive quality of MC has decreased.[9,11] This can be attributed to tumor morphology alone not reflecting the biology of HCC; however, it is also worth noting LRT was uncommon when MC was developed.[12]

Many factors have been suggested to improve patient selection. Unfortunately, many models incorporated genetic parameters/serum assays not routinely collected in clinical practice[6,7,13-15] or pathological factors which cannot be used for selection.[8,16,17] Most studies presented dichotomous criteria; the patient is in or out. Although simple and convenient, modeling risk in binary terms discards variance, which could be associated with outcomes and results in poorer discrimination.[18] Such models need to be reconstructed *de novo* with population or clinical practice shifts. The Hazard Associated with Liver Transplantation for Hepatocellular Carcinoma (HALTHCC) model sought to remedy these strategic missteps by forming a continuous spectrum of risk upon which a patient can be judged relative to others.[9] Its value is calculated by common clinical variables which were determined in a single center to be relevant for post-LT survival and validated among the entire national experience of the United States.[9] It was also validated for post-LT recurrence, and its longitudinal measurement was associated with dropout in a multicenter study.[11]

Received December 16, 2018; accepted June 17, 2019.

Additional Supporting Information may be found at onlinelibrary.wiley.com/doi/10.1002/hep.30838/suppinfo.

K.S. received an American Society of Transplant Surgeons Fellowship related to this work. The corresponding author had full access to all of the data and the final responsibility to submit for publication.

**Current address: Department of Surgery, Duke University Hospital, Durham, NC.*

View this article online at wileyonlinelibrary.com.

DOI 10.1002/hep.30838

Potential conflict of interest: Dr. Cillo advises for and received grants from Novartis. He advises for Sanofi and received grants from Johnson & Johnson, Astellas, and Pfizer. Dr. Markmann advises for and owns stock in Egenesis-Qihan Bio.

ARTICLE INFORMATION:

From the [1]Department of General Surgery and Cleveland Clinic Lerner College of Medicine, Digestive Disease and Surgery Institute, Cleveland Clinic, Cleveland, OH; [2]Dumont-UCLA Transplant and Liver Cancer Center, Department of Surgery, Ronald Reagan UCLA Medical Center and David Geffen School of Medicine at UCLA, Los Angeles, CA; [3]Department of Abdominal Transplant and HPB Surgical Oncology, University Health Network and University of Toronto, Toronto, Ontario, Canada; [4]Transplant Center, Massachusetts General Hospital and Harvard Medical School, Boston, MA; [5]Starzl Unit of Abdominal Transplantation, St. Luc University Hospital, Université Catholique Louvain, Brussels, Belgium; [6]Department of General Surgery and Organ Transplantation, Umberto I Hospital, Sapienza University, Rome, Italy; [7]Department of Surgery, Oncology and Gastroenterology, University of Padua, Padua, Italy; [8]Department of Medicine I, Medical University Innsbruck, Innsbruck, Austria; [9]UCL Institute for Liver and Digestive Health and Royal Free Sherlock Liver Centre, Royal Free Hospital and UCL, London, United Kingdom; [10]Department of Transplantation and Hepatobiliary Surgery, University of Mainz, Mainz, Germany; [11]Division of General Surgery and Liver Transplantation, San Camillo Hospital, Rome, Italy; [12]Department of Transplant Surgery, Polyclinic Tor Vergata Foundation, Tor Vergata University, Rome, Italy; [13]Unit of Hepatobiliary Surgery and Transplantation, Azienda Ospedaliero-Universitaria Ospedali Riuniti, Torrette Ancona, Italy; [14]Liver Unit, Department of Surgery, Agostino Gemelli Hospital, Catholic University, Rome, Italy; [15]Kyoto University Hospital, Kyoto, Japan; [16]Kyushu University Hospital, Kyushu, Japan.

ADDRESS CORRESPONDENCE AND REPRINT REQUESTS TO:

Kazunari Sasaki, M.D.
Department of General Surgery
Digestive Disease and Surgery Institute, Cleveland Clinic
9500 Euclid Avenue
Cleveland, OH 44195
E-mail: sasakik@ccf.org
Tel.: +1-216-444-2222

In setting potential LT candidates on a spectrum of risk, population characteristics can continuously change without the model being so stiff as to stop being useful. Furthermore, maintaining the equation as a set of beta coefficients is a flexible framework familiar to the transplant community in the form of the Model for End-Stage Liver Disease (MELD) equation and allows continual improvement in accuracy without reinvention of the concept. Although HALTHCC is a valuable lens through which to observe candidate risk, its utility among international centers is unknown. The aim of this study was 3-fold: to create a broad and heterogeneous cohort representative of worldwide LT for HCC; to test the hypothesis that HALTHCC will maintain its predictive character for post-LT outcomes in this cohort; and to test the feasibility and efficacy of recalibrating the beta coefficients for HALTHCC as has been done for MELD in the past.

Patients and Methods

PATIENT DATA AND COMPETITOR MODELS

This multicenter trial was conducted using data from prospectively collected transplant records of patients undergoing LT for HCC from 2002 to 2014. The inclusion criteria were intentionally broad to generate a heterogeneous sample: adult recipients of LT (whole organ or partial) with a primary or secondary diagnosis of HCC. Patients were included regardless of whether they had access to preoperative LRT. Across 16 sites and eight countries, 4,089 patients were registered. Patient selection, management, and follow-up were determined by center practice and local/regional/national law. Data sharing was restricted to variables from common clinical practice to construct the investigated risk models. Competitor models were included based on literature review and available variables. Search terms, search strategy, and model inclusion are described in Supporting Information Table S1. This study was approved by the institutional review boards (IRBs) at participating institutes, written informed consent was obtained at each center while data was maintained centrally at the Cleveland Clinic (IRB #17-772).

STUDY DESIGN AND STATISTICAL ANALYSIS

The TRIPOD-IV framework was followed in reporting this study.[19] Two lines of investigation were conducted:

1. The original HALTHCC score (oHALTHCC) was examined with respect to explant pathology and then on post-LT outcomes. For pathology, focus was placed on vascular invasion (VI) and poorly differentiated component (PDC) on explant analysis because these are most strongly associated with post-LT recurrence.[12,20] In the same oHALTHCC analysis, post-LT outcomes were analyzed to assess discriminatory ability in the broad international cohort.
2. To test the utility of recalibration, the study population was randomized into a training set and validation set. The training set was used to find the optimal weighting of HALTHCC parameters, and then this recalibrated HALTHCC score (rHALTHCC) was examined in the validation set to assess discriminatory ability of all competitor models, including oHALTHCC.

HALTHCC is a continuous score based on preoperatively accessible clinical characteristics, which was reported on.[9] Original HALTHCC (oHALTHCC) is calculated by Equation 1:

$$\begin{aligned} \boldsymbol{oHALTHCC} = & (1.85 * \ln(AFP)) \\ & + (1.27 * \textit{tumor burden score}\ (TBS)) \\ & + (0.26 * MELD - Na). \end{aligned}$$

Overall survival (OS) was defined as the time interval from LT to either mortality or last follow-up (where censored). Recurrence was monitored using center-specific practice and reported as location and time from LT. HCC-related mortality was defined as mortality after recurrence (either intra- or extrahepatic spread). Cox proportional hazards models, ordinal logistic models, and the Kaplan-Meier (KM) method were used to assess associations with OS, recurrence, and HCC-related mortality. In order to maintain statistical validity and avoid overfitting, the cohort was divided into a training and validation set for rHALTHCC. Significant differences between the original cohort and this international cohort (Supporting Information Table S2) strengthen the

generalizability of any findings. Block randomization was performed in a 3:1 fashion (with each center serving as a block) to maintain similar representation of every center.[21] Recalibration was performed on the training cohort (n = 1,021) using ordinal logistic modeling of a composite endpoint (1 point for either recurrence or mortality and 2 points for recurrence leading to mortality) in an effort to better balance the competing interests in allocation policy.[22-24] Throughout the modeling, attention was paid to ensure an adequate ratio of events to explanatory variables (ratios maintained >10:1) and proportional hazards assumption (assessed by Schoenfeld global test; $P = 0.08$). Postestimation concordance and discrimination was applied to the validation cohort (n = 3,068) using Harrell's C-index, Akaike information criterion (AIC), net reclassification improvement (NRI), and integrated discrimination improvement (IDI).[25] Calibration between the oHALTHCC and rHALTHCC was compared by estimating predicted 1-, 3-, and 5-year survival and overall recurrence rates for 10 randomly selected cohorts among the 3,068 patients. Categorical variables are displayed as counts and percentages whereas continuous variables are reported as medians and interquartile range (25th-75th percentiles). All testing performed was two-sided and compared against a 5% alpha using STATA software (version 13; StataCorp LP, College Station, TX).

Results

POPULATION COMPOSITION

Population distribution and characteristics are visualized in Fig. 1 and Supporting Information Fig. S1 and summarized in Supporting Information Table S3. Significant differences were seen in recipient underlying liver disease and MELD-Na, Model of End Stage Liver Disease Sodium score (MELD-Na) at listing, tumor burden, and alpha-fetoprotein (AFP) at listing, waiting time, proportion of candidates receiving LRT while waiting, and tumor characteristics at LT (Fig. 1A). In addition to heterogeneity across world region, there were trends over time (Fig. 1 and Supporting Fig. S1). For example, increased waiting time in North America attributed to shifts in allocation policy have led to increased penetration of LRT and decreased the active tumor burden at transplantation. The Asian cohort changed dramatically over the study period; government reimbursement and focus on outcomes have decreased the tumor burden of candidates undergoing LT. The European cohort was more stable over time with only minor trends toward increasing penetration of LRT and some, likely related, decreases in tumor burden at transplantation.

EXPLANT PATHOLOGY

The relationship between risk scores and pathological features is demonstrated in Fig. 2. Among MC-in patients, VI and PDC on explant pathology were present in 19.2% and 11.1% of patients; MC-out status increased the incidences to 38.3% and 18.4%, respectively ($P < 0.0001$). The Metroticket 2.0 paradigm has introduced three groups of candidates among which LT should have comparable outcomes.[26] There was an increase in both VI and PDC from Group 1 (AFP < 200 and Up-to-7; VI, 19.0% [95% confidence interval {CI},17.3-20.6] and PDT [poorly differentiated tumor], 11.0% [95% CI, 9.7-12.3]) to Groups 2 (AFP, 200-400 and Up-to-5; VI, 40.7% [95% CI, 27.5-54.0] and PDT, 22.2% [11.0-33.4]) and 3 (AFP, 400-1,000 and Up-to-4; VI, 32.1% [95% CI, 14.5-49.8] and PDT, 14.3% [95% CI, 10.8-27.5]). Interestingly, patients not indicated to receive LT, here termed Group 4 (AFP > 1,000 or sum of lesion size and number greater than allowed with a given AFP for Groups 1-3) had similar rates of VI (45.9% [95% CI, 41.4-50.3]) and PDC (21.3% [95% CI, 17.6-24.9]; Fig. 2A). Another alternative to HALTHCC is the model of recurrence after liver transplantation (MoRAL) score. There was poor corroboration between increasing MoRAL score and explant pathology (Fig. 2B). HALTHCC had a clear and robust association between estimated pre-LT risk and explant pathology; both VI and PDC increased with increasing HALTHCC score (Fig. 2C). The lowest-risk patients (HALTHCC 0-5) had much lower rates of VI and PDC than the highest-risk patients (HALTHCC > 35; VI, 7.7% [95% CI, 1.2-14.2] vs. 70.6% [95% CI, 48.3-92.9] and PDC, 4.6% [95% CI, 0.1-9.8] vs. 47.1% [95% CI, 22.6-71.5]; $P < 0.0001$ for both). This trend was robust to MC status, indicating that HALTHCC may be a useful tool to estimate risk of poor pathological features using only preoperatively accessible variables.

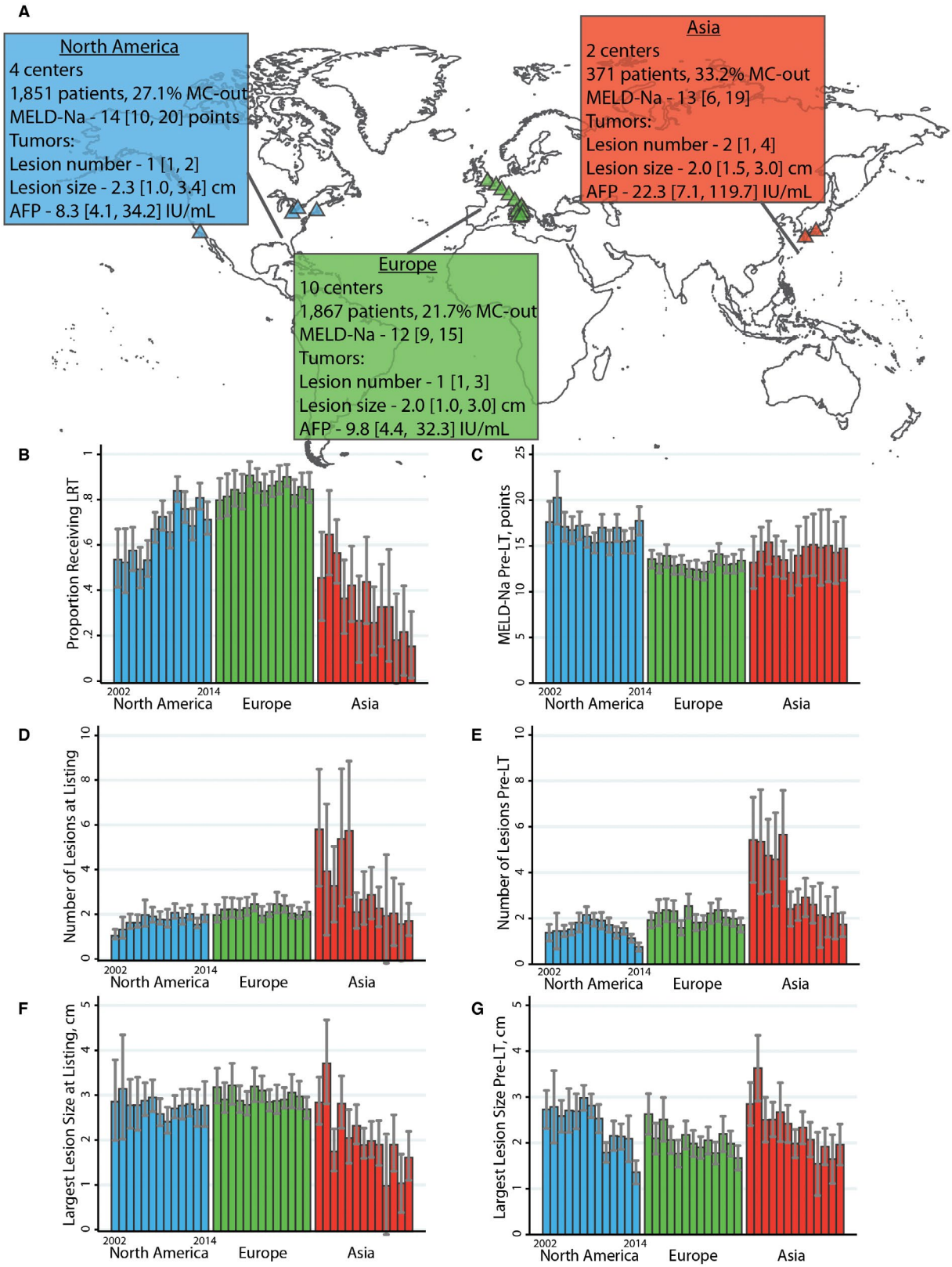
A
North America
4 centers
1,851 patients, 27.1% MC-out
MELD-Na - 14 [10, 20] points
Tumors:
Lesion number - 1 [1, 2]
Lesion size - 2.3 [1.0, 3.4] cm
AFP - 8.3 [4.1, 34.2] IU/mL
Asia
2 centers
371 patients, 33.2% MC-out
MELD-Na - 13 [6, 19]
Tumors:
Lesion number - 2 [1, 4]
Lesion size - 2.0 [1.5, 3.0] cm
AFP - 22.3 [7.1, 119.7] IU/mL
Europe
10 centers
1,867 patients, 21.7% MC-out
MELD-Na - 12 [9, 15]
Tumors:
Lesion number - 1 [1, 3]
Lesion size - 2.0 [1.0, 3.0] cm
AFP - 9.8 [4.4, 32.3] IU/mL
B
Proportion Receiving LRT
C
MELD-Na Pre-LT, points
D
Number of Lesions at Listing
E
Number of Lesions Pre-LT
F
Largest Lesion Size at Listing, cm
G
Largest Lesion Size Pre-LT, cm
2002
2014
North America
Europe
Asia

FIG. 1. Trends in worldwide LT for HCC. (A) Cartoon map of the world overlaid with the characteristics of LT for HCC in each of the three major regions in this study. (B) Proportion of LT recipients undergoing LRT before LT by year of LT and center region. (C) MELD-Na measured just before LT by year of LT and center region. (D) Number of viable lesions on radiographic evaluation at time of candidate listing, by year of LT and center region. (E) Number of viable lesions on radiographic evaluation at last pre-LT evaluation, by year of LT and center region. (F) Largest viable lesion size on radiographic evaluation at time of candidate listing, by year of LT and center region. (G) Largest viable lesion size on radiographic evaluation at last pre-LT evaluation, by year of LT and center region. (A) Boxed descriptive statistics are median [interquartile range]. (B-G) Bars represent means with 95% CIs.

MODELING POST-LT OUTCOMES

oHALTHCC was tested for association with post-LT outcomes (Fig. 3A,B). This analysis showed generally promising validation of the U.S. national experience; lower-risk patients had higher survival and lower recurrence. However, in this cohort, the very-low-risk patients (oHALTHCC < 5) and the penultimate high-risk group (oHALTHCC 30-35) were overlapping with nearby risk cohorts; identifying clear areas where risk estimation could be improved. oHALTHCC's discrimination was consistent with the U.S. national experience for recurrence (C-index, 0.69), OS (0.62), and HCC-related death (0.72).

RECALIBRATION OF HALTHCC

The ordinal logistic modeling used to recalibrate HALTHCC is detailed in Table 1. This confirmed the highly statistically significant nature of the original components of HALTHCC and also provided evidence that regional variation (especially Asia) was contributing to deviations from risk-outcomes relationships being observed in the U.S. experience. Together, these estimates were used to generate Equation 2:

$$\begin{aligned} \textbf{\textit{rHALTHCC}} = &(2.31 * \ln(\textit{AFP})) \\ &+ (1.33 * (\textit{TBS})) \\ &+ (0.25 * \textit{MELD} - \textit{Na}) - (5.57 * \textit{Asia}). \end{aligned}$$

The validation cohort was thusly estimated and then KM estimates of survival and recurrence compared to the oHALTHCC equation (Fig. 3C,D). This analysis provided a greatly improved range of outcomes across all levels of risk. The 5-year post-LT OS ranged from 82.4% in the lowest-risk group (rHALTHCC < 5; n = 145 in 3,068) to 32.4% in the highest-risk group (rHALTHCC > 35; n = 24 in 3,068). A similar trend was observed with 5-year post-LT recurrence, which ranged from 8.6% (rHALTHCC < 5; n = 145 in 3,068) to 70.0% (rHALTHCC > 35; n = 24 in 3,068). The robustness with which rHALTHCC stratified risk of post-LT events was much greater than all competitor scores (Supporting Fig. S2).

DISCRIMINATORY ABILITY AND ACCURACY OF HALTHCC

Following recalibration of HALTHCC in the training set, comparison of discrimination for post-LT mortality, recurrence, and HCC-related mortality was conducted among competitors. Compared to its closest competitor, rHALTHCC was a statistically superior discriminator as measured by Harrell's C (0.63 [95% CI, 0.61-0.65] vs. Metroticket 2.0 (0.57 [95% CI, 0.55-0.58]; $P < 0.0001$), NRI (0.330 [95% CI, 0.261-0.391] vs. Metroticket 2.0, 0.260 [95% CI, 0.201-0.315]), IDI (95% CI, 0.043 [0.031-0.057] vs. Metroticket 2.0, 0.016 [95% CI, 0.009-0.025]), and AIC (14,332 vs Metroticket 2.0, 14,401), for post-LT survival (Supporting Table S4). This difference persisted for recurrence (0.71 [95% CI, 0.68-0.74] vs. Metroticket 2.0, 0.65 [95% CI, 0.62-0.67]; $P < 0.0001$; Supporting Table S5) and HCC-related mortality (0.74 [95% CI, 0.71-0.77] vs. Metroticket 2.0, 0.66 [95% CI, 0.63-0.69]; $P < 0.0001$; Supporting Table S6). Last, estimates of 1-, 3-, and 5-year post-LT survival and overall recurrence rates derived from oHALTHCC and rHALTHCC were compared to observed values of these post-LT outcomes to demonstrate calibration of the estimating equations underlying their discriminatory ability (Supporting Fig. S3). It is clear to observe visually how the linear regressions of the randomly selected clusters improve in their estimation of the observed values moving from A-B, C-D, etc., which confirms the model's improvement.

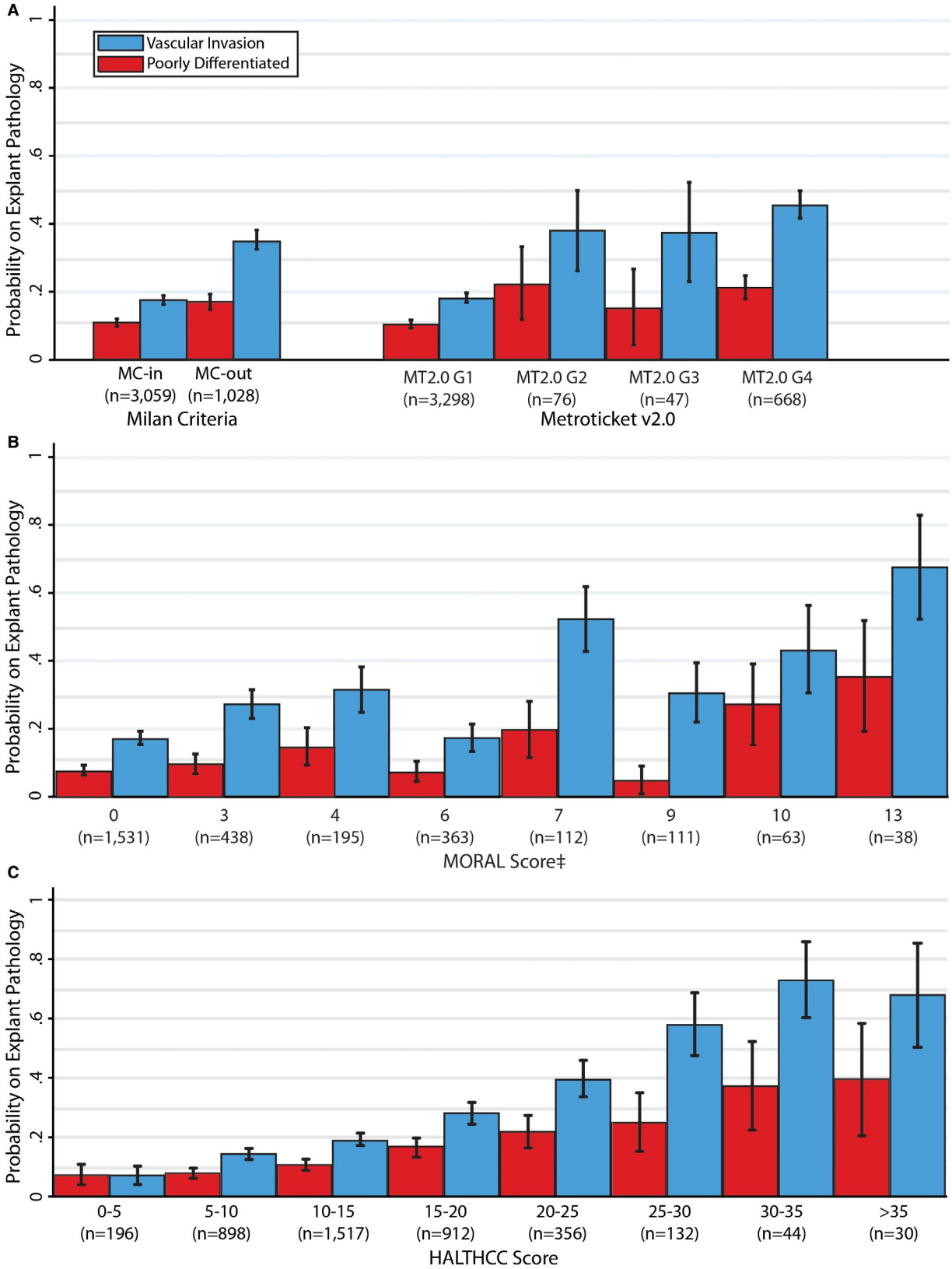
A
Vascular Invasion
Poorly Differentiated
Probability on Explant Pathology
0
.2
.4
.6
.8
1
MC-in (n=3,059)
MC-out (n=1,028)
Milan Criteria
MT2.0 G1 (n=3,298)
MT2.0 G2 (n=76)
MT2.0 G3 (n=47)
MT2.0 G4 (n=668)
Metroticket v2.0
B
Probability on Explant Pathology
0 (n=1,531)
3 (n=438)
4 (n=195)
6 (n=363)
7 (n=112)
9 (n=111)
10 (n=63)
13 (n=38)
MORAL Score‡
C
Probability on Explant Pathology
0-5 (n=196)
5-10 (n=898)
10-15 (n=1,517)
15-20 (n=912)
20-25 (n=356)
25-30 (n=132)
30-35 (n=44)
>35 (n=30)
HALTHCC Score

FIG. 2. Utility of preoperatively assessable risk metrics as measured by explant pathological features. (A) Demonstrates probability of finding vascular invasion or poor tumor differentiation of explant pathology for MC (in vs. out) and Metroticket 2.0 groupings (indicated LT patients [Groups 1-3] vs. not indicated [Group 4]). (B) Explant pathology rates for VI and poor tumor differentiation grading by MoRAL score. (C) Explant pathology rates of VI and poor tumor differentiation grading by HALTHCC score. Bars demonstrate means with 95% CIs for estimates. Abbreviations: G, group; MT2.0, Metroticket version 2.0.

Discussion

In a large and heterogeneous worldwide cohort, this validation study confirmed superior discriminatory characteristics of HALTHCC for explant pathology and post-LT outcomes. Recalibration of the score improved its prognostic utility. Although MC was critical to widespread success in LT for HCC, its utility has decreased over time attributed to changes in clinical practice. Many models were developed to replace MC and several did possess improved prognostic discrimination[6-8,16,20,27-29]; however, most fell by the wayside because of poor generalizability. Those models were developed in single institutes where the patient population and treatment strategy are homogeneous. As a result, they failed to show similar excellence in other cohorts. Consequently, other groups made their own model in their cohorts, and our field has been trapped in a loop. HALTHCC was developed to defeat this loop and form the basis of a new allocation system to improve prognostic utility and fair and reasonable organ allocation between HCC and non-HCC patients. This validation study used a heterogeneous, worldwide cohort with the hypothesis: If HALTHCC is truly measuring the disease HCC, it should work in any environment. Statistical analyses showed that there were extensive differences of tumor/treatment characteristics in 4,089 worldwide patients. The original HALTHCC score's performance was consistent with past studies as a discriminator of recurrence (C-index, 0.69), OS (0.62), and HCC-related mortality (0.72). Recalibration of HALTHCC was conducted and improved statistical discrimination over oHALTHCC as measured by C-index ($P < 0.001$ for recurrence, $P = 0.051$ for survival, and $P = 0.003$ for HCC-related mortality) and also visual discrimination using KM and calibration curves (Fig. 3 and Supporting Fig. S3).

Although original HALT-HCC was validated in the U.S. national experience, it was unknown how this clinical score would perform in an entirely new population with high heterogeneity with most patients outside the United States (Supporting Table S2). This study cohort consisted of populations with different treatment strategies and tumor characteristics (Fig. 1). Although it is known that high heterogeneity decreases performance of prediction models,[30] the oHALTHCC score maintained a high level of discrimination, especially for recurrence. Moreover, HALTHCC showed a significant association with explant pathology, such as vascular invasion and poorly differentiated component (Fig. 2). MC demonstrated a single-level increase in rates of both VI and PDC consistent with its binary nature. However, when compared against all other criteria, increased HALTHCC score was most closely and reliably related to both VI and PDC. This is without any explicit modeling of pathology done at any phase in HALTHCC's development. These results suggest that this continuous risk score is measuring tumor features underlying risk of poor outcomes, not statistical static.

One might entertain the critique that a continuous score's form is too complicated and that our purpose seems more about chasing statistical excellence rather than usability. However, the transplant community is already familiar with the concept of continuous scoring; demonstrated by MELD/MELD-Na scores, those risk metrics were similarly statistically derived and optimized. Moreover, it is well known that the alternative, categorical, or ordinal scoring systems discard accuracy and their cut points promote controversy instead of easing decision making.[18] After HALTHCC's development, a close competitor, the Metroticket 2.0 model, was introduced. This model is derived from a statistical basis and estimates risk from a continuous formulation, which is then binned into risk categories based on risk planes. Although we applaud the derivation, its application to bins is overly simplistic. The benefit realized by switching from a categorical logic to a continuous risk score is significant and encourages the community to embrace statistically valid and efficient methodologies. Organ allocation is a national issue; therefore, even small improvements in predictive utility could affect and

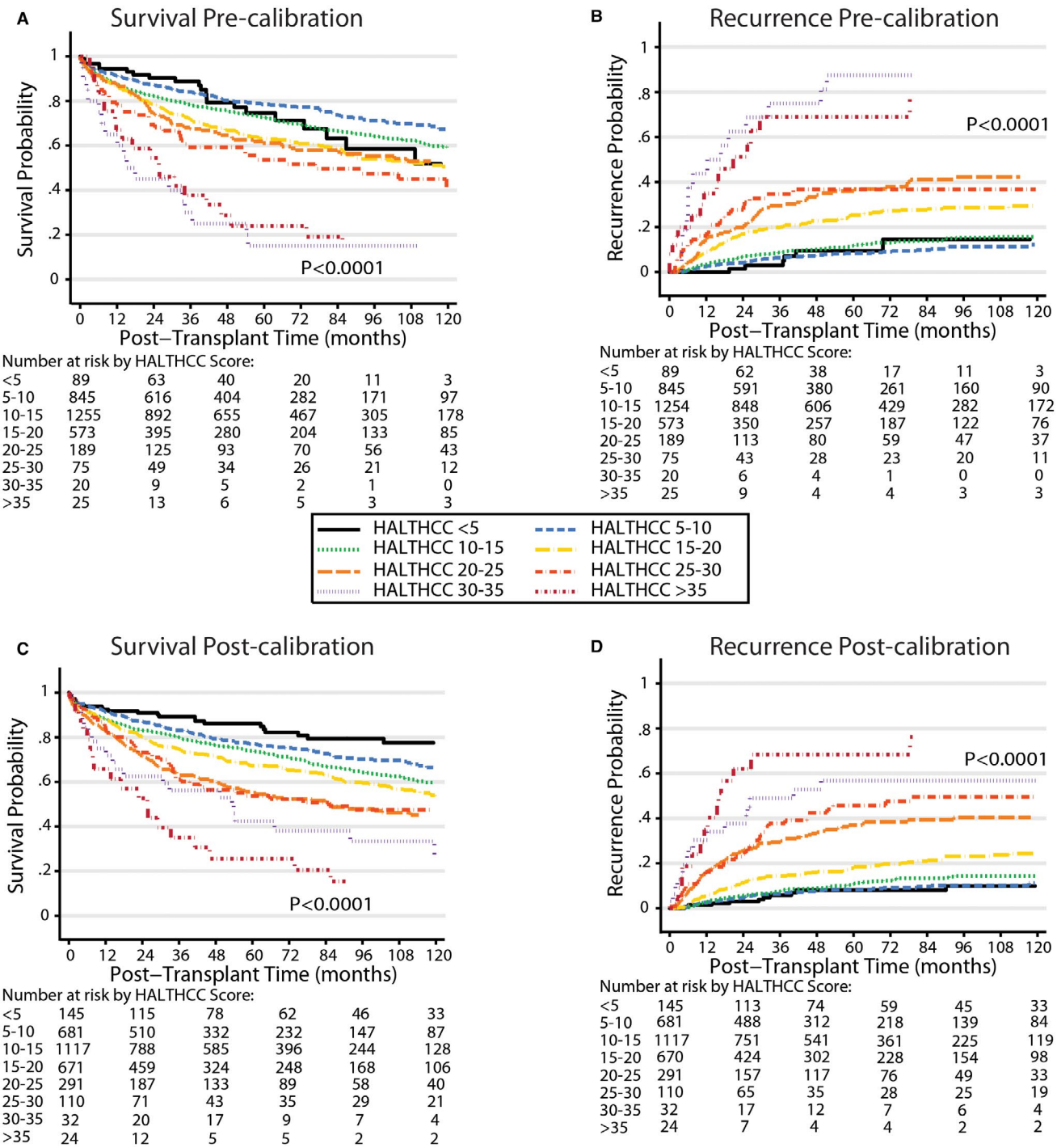

FIG. 3. HALTHCC score is associated with OS and recurrence and improved after recalibration. (A) KM survival estimates for OS by HALTHCC score before any model recalibration. (B) KM estimates for recurrence by HALTHCC score before any model recalibration. (C) KM survival estimates for OS following HALTHCC recalibration using the training cohort. (D) KM estimates for recurrence following HALTHCC recalibration using the training cohort. Only validation cohort patients are included in the analysis of (A-D); the training cohort for recalibration was excluded to avoid overfitting.

TABLE 1. Ordinal Logistic Model for Recalibrating HALTHCC Using Training Set

	Coefficient	Standard Error	Wald Test	Odds Ratio (95% CI)	*P* Value*
MELD-NA, per point	0.0247	0.0088	2.82	1.025 (1.008-1.043)	0.005
TBS, prepoint	0.1332	0.0233	5.71	1.143 (1.091-1.196)	<0.0001
AFP, per ln(IU/mL)	0.2308	0.0357	6.46	1.260 (1.175-1.351)	<0.0001
World region					
North America	Base case	—	—	—	—
Europe	0.1817	0.141	1.29	1.199 (0.910-1.581)	0.198
Asia	–0.5571	0.261	–2.13	0.573 (0.343-0.955)	0.033

*Based on likelihood test adjusted for the other factors in the final model.

contribute to large changes in organ utilization. We believed that balancing statistical accuracy with a minimally complex model would increase the number of LT for HCC without adversely affecting outcomes. In fact, there is evidence that priority allocation using HALTHCC would allow fine-tuning of "acceptable outcomes" and serve as a simple metric for transplant center comparison across regions.[11] Furthermore, a continuous risk score is advantageous compared to a dichotomous model given that the continuous score can be recalibrated easily when necessary, as was the case with the MELD and MELD-Na scores previously.

In this study, recalibration of HALTHCC was performed as planned at the conception of HALTHCC and this statistically significantly increased its discriminatory ability.[9] Following randomization of the study population into similarly composed cohorts (Table 2), the training set underwent ordinal logistic regression using a weighted composite endpoint. The estimation among the randomized cohort demonstrated statistical significance for all the factors in oHALTHCC. Supporting Table S7 outlines the composition of the validation cohort stratified by rHALTHCC score and MC status. If we were to reallocate the livers in this study based on composite risk, we can provide priority to nearly 74.6% of MC-out candidates without post-LT outcomes suffering. Furthermore, we can reallocate ~4-5% of MC-in recipients with higher composite risk. This approach would have the net effect of greater equity and lower post-LT recurrence rates. However, care must be taken given that these estimates are all derived post hoc in a cohort influenced by MC. A consensus meeting is necessary to finalize the form and composition of predictive models our community would like to explore before embarking on prospective trials. For example, we previously demonstrated the value of including MELD-Na in HALTHCC to improve prediction of OS; however, it has relatively less impact on modeling recurrence.[9,11] Clearly, recurrence is of prime interest in LT for HCC; however, whereas 3-year recurrence runs 7%-10%, transplant-related mortality at 3 years is 15%-20%[26]; candidate selection is about more than tumor size, and number and transplantation is not benign. As a community, we need to decide the most critical outcomes of interest for future prospective studies and then optimize our risk measure. A single system from presentation through LRT and into the transplant period is preferred given that it would simplify a complex clinical decision tree, but also improve outcomes.

Moving away from waiting-time–based standard exception points with individualized applications to a mathematic transformation of risk and benefit has many advantages. In Fig. 4, 6 example patients are outlined to demonstrate three proposed models of allocation through which to explore competing concepts in allocation related to HCC. It is important to note that these draft models are simply examples to discuss possible strategies in allocation; a final or "best model" will require advanced simulation studies to estimate the impact between non-HCC and HCC cohort dropout, transplant volumes, and post-LT outcomes. In the first draft model, post-LT survival is maximized (model alpha). This model prioritizes long-term survival after LT alone; patients who have lowest HALTHCC score before organ allocation wait the least time, have the least disease burden, and have the highest OS rates and lowest recurrence. However, one can argue that many of these low-risk patients would go for many months after LRT without disease recurrence or a new lesion; in contrast, many patients who could have benefited from timely transplantation may have progressed.

TABLE 2. Population Characteristics

	Training Cohort (n = 1,021)	Validation Cohort (n = 3,068)	*P* Value
Transplant year	2008 [2005, 2012]	2008 [2005, 2012]	0.920
World region			0.740
North America	460 (45.1)	1,391 (45.3)	
Europe	468 (45.8)	1,399 (45.6)	
Asia	93 (9.1)	278 (9.1)	
Sex, male	689 (67.5)	2,133 (69.5)	0.222
Age, years	58 [52, 63]	58 [52, 63]	0.360
Underlying liver disease*			0.667
HCV cirrhosis	560 (54.8)	1,683 (54.9)	
HBV cirrhosis	192 (18.8)	501 (16.3)	
Laennic's cirrhosis	203 (19.9)	623 (20.3)	
NASH cirrhosis	50 (4.9)	189 (6.2)	
Other	77 (7.5)	248 (8.1)	
Listing characteristics			
Laboratory MELD-Na, points	12.0 [9.0, 16.0]	12.0 [9.0, 16.0]	0.884
Tumor no.	1.0 [1.0, 3.0]	1.0 [1.0, 3.0]	0.655
Tumor size, cm	2.6 [1.9, 3.6]	2.5 [1.8, 3.6]	0.477
TBS, points	3.4 [2.5, 4.8]	3.4 [2.4, 4.6]	0.445
AFP, IU,mL	10.0 [5.0, 37.9]	10.0 [4.7, 39.3]	0.553
Patients meeting MC	746 (73.1)	2,173 (71.0)	0.786
HALTHCC, points	12.6 [10.0, 15.8]	12.6 [10.3, 15.9]	0.468
Waiting time, days	157 [60, 318]	149 [60, 315]	0.557
Pre-LT characteristics			
Laboratory MELD-Na, points	13.0 [9.0, 18.0]	13.0 [9.0, 17.6]	0.914
Tumor no.	1.0 [1.0, 2.0]	1.0 [1.0, 3.0]	0.218
Tumor size, cm	2.1 [1.1, 3.0]	2.1 [1.0, 3.0]	0.536
TBS, points	3.06 [1.4, 4.2]	3.05 [1.8, 4.2]	0.562
AFP, IU,mL	10.0 [4.5, 35.0]	9.8 [4.2, 37.4]	0.448
NLR	2.7 [1.7, 4.3]	2.7 [1.7, 4.4]	0.249
Patients meeting MC	752 (73.7)	2,308 (75.2)	0.303
HALTHCC, points	12.0 [9.0, 15.6]	12.2 [9.4, 15.6]	0.170
Follow-up time, days	1,369 [529, 2700]	1,445 [606, 2753]	0.378

Continuous variables: median [interquartile range].
*May not sum to unity because of overlapping etiologies.
Abbreviations: HBV, hepatitis B virus; HCV, hepatitis C virus; NASH, nonalcoholic steatohepatitis; NLR, neutrophil-lymphocyte ratio; SRTR, Scientific Registry of Transplant Recipients.

Second, the model prioritizing oncological risk in a fashion similar to the "sickest first" policy (model gamma): In this model, high HALTHCC patients have the highest priority because they presumably have the most to benefit from immediate and early transplantation. However, this model performs poorly in metrics of organ utilization and post-LT population recurrence rates. It could be argued that many moderate-risk patients with lower priority may not have progressed had they been transplanted early, whereas this model may inadvertently prioritize transplantation of patients with almost certain distant microscopic disease. Last, the model balancing the middle group of patients, those with near mean risk best served by timely transplantation (model beta): These patients are clearly at risk of dropout, but also have a reasonable chance at cure from timely transplant. In model beta, patients closest to the mean historical risk at transplant achieve the highest priority by simply transforming their mathematical distance from the mean using absolute values. Patients with very high risk at presentation may naturally drop out, but may be able to get sufficient priority to access LT if they have an appropriate response to LRT as measured by HALTHCC.[11] Patients with very low risk may go many months or even years after LRT without disease recurrence; however, with the field effect of cirrhosis in place, it is likely that another lesion will, at some point, crop up; in that case, their HALTHCC score will increase and their priority for transplantation will increase. This balanced approach provides a compromise between sickest first and survival conscious or utility concepts. The beauty of all of these models compared to something like MC or other binary systems (including the complex criteria currently required to meet exceptional status under United Network for Organ Sharing [UNOS]) is that there is no cutoff at which point patients with only marginally increased risk suffer from total ineligibility (e.g., one tumor sized 5.2 cm vs. one tumor sized 4.8 cm or AFP level of 1,005 vs. 995). Last, these models can be easily adjusted in two future oriented ways:

1. They can have the constant term adjusted by UNOS region to help negotiate disparities in rates of transplantation by MELD/priority score.
2. They can have the multipliers modified to adjust the risk tolerance of the society if recurrence rates are too high or there is too much dropout.

These models serve as the most basic and first in a series of studies our group hopes to utilize to estimate the impact advanced, HALTHCC-based allocation models can have on our society.

Despite its size, this large study has some limitations to consider. First, it is a retrospective study;

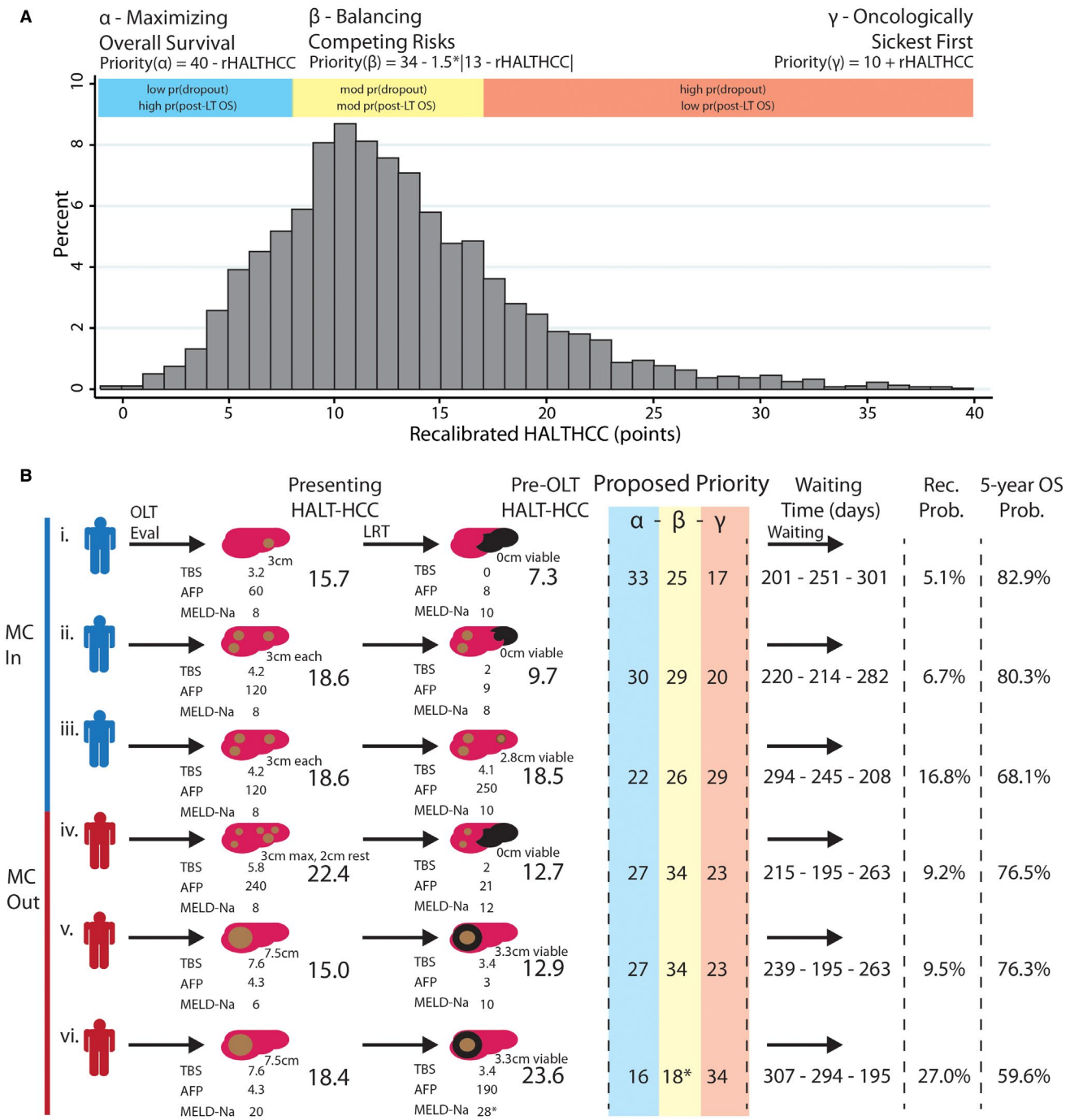

FIG. 4. Example patients to demonstrate utility of HALTHCC-based allocation. Six example patients illustrate the ease and utility of longitudinal assessment with HALTHCC of HCC patients undergoing evaluation for LT. The proposed models vary in emphasis from purely postoperative survival to a sickest first and finally balanced approach. Furthermore, a simple adjustment for regional competitiveness or to balance risk of dropout can be conducted through modification of the constant terms. Abbreviation: OLT, orthotopic liver transplantation.

therefore, we cannot definitively infer any causality related to HALTHCC score. Also, because of time constraints, we were not able to collect detailed longitudinal data on tumor size, serum tumor markers, exact times of multiple LRT, nor the response to LRT. Methodologically, assumptions need to be

made to adjust for regional differences in outcomes. One can either allow the individual coefficients to vary for each region (i.e., estimate a new equation per region) or allow an omnibus adjustment at the end, assuming biological determinants, like AFP, tumor burden score (TBS), and MELD-Na, will have similar orders of influence by region. In scenario 1 above, we assume knowledge of the mechanism by which outcomes vary between different locations in the world. In reality, it could be population genetics, donor types, or other unspecified practice patterns which most significantly account for the difference in outcomes for a given HALTHCC score (unadjusted by region). In contrast, in scenario 2, we acknowledge that outcomes for similar patient characteristics may differ around the world. However, that does not mean that they are not comparable, nor that the difference is not estimable. In fact, the difference is estimable; we just do not know the mechanism through which that difference acts. Our assumption in acting through scenario 2 is that AFP, TBS, and MELD-NA have a relatively similar contribution (at least in a similar order of magnitude) to risk to poorer outcome across location. However, overall risk of a given outcome is not proportional strictly to those biological factors. Given that the mechanism of risk modification is unknown, we chose the method of omnibus adjustment of risk for a given location. In fact, this method is the most conservative from both a clinical reasoning standpoint, but also statistically given that we introduce far fewer additional assumptions and tests and thus reduce opportunity for bias. Furthermore, a sensitivity analysis was performed, demonstrating that the coefficients for AFP, TBS, and MELD-Na were similar between the Western and Eastern cohorts (Supporting Tables S8 and S9). Estimating the baseline hazard function between each region separately demonstrated that proportionality was not violated, but that the baseline hazard is simply lower in Asia; thus, our estimates of the impact of AFP, TBS, and MELD-Na are valid (Supporting Fig. S4A,B). Next we calculated model utility on the Western data set alone only, which only marginally decreased the utility estimates by Harrell's C (Supporting Table S10). Additionally, calibration of HALTHCC-derived OS and recurrence predictions compared to observations were not systematically biased by region whereas its nearest competitor model, Metroticket 2.0, consistently overestimated the Asian cohorts risk (Supporting Fig. S4C,D) Finally, because we did not have a complete data set of all previously reported serum markers and all the individual patients' longitudinal courses, we were unable to draw comparisons on utility to several recent scoring systems, including the delta-HALTHCC and Time-Radiological-response-Alpha-fetoprotein-INflammation (TRAIN) scores.[7,11]

In conclusion, this large international trial validated and refined the role for the continuous risk metric, HALTHCC. Its utility was confirmed among 4,089 patients undergoing LT for HCC, composing a heterogeneous data set, which strengthens generalizability. Although MC has served the transplant community well for two decades, it is time for providers in this field to have our practice "bridged" until whole-exome studies provide truly personalized guidance to HCC management. We proposed various models to inform allocation algorithms based on HALTHCC. Prospective trials introducing HALTHCC into observational frameworks to remove the bias laid out by two decades of practice under MC are warranted.

REFERENCES

1) El-Serag HB. Hepatocellular carcinoma. N Engl J Med 2011;365:1118-1127.
2) El-Serag HB, Kanwal F. Epidemiology of hepatocellular carcinoma in the United States: where are we? Where do we go? Hepatology 2014;60:1767-1775.
3) Iwatsuki S, Starzl TE, Sheahan DG, Yokoyama I, Demetris AJ, Todo S, et al. Hepatic resection versus transplantation for hepatocellular carcinoma. Ann Surg 1991;214:221-229.
4) Llovet JM, Fuster J, Bruix J. Intention-to-treat analysis of surgical treatment for early hepatocellular carcinoma: resection versus transplantation. Hepatology 1999;30:1434-1440.
5) Mazzaferro V, Regalia E, Doci R, Andreola S, Pulvirenti A, Bozzetti F, et al. Liver transplantation for the treatment of small hepatocellular carcinomas in patients with cirrhosis. N Engl J Med 1996;334:693-700.
6) Uchiyama H, Itoh S, Yoshizumi T, Ikegami T, Harimoto N, Soejima Y, et al. Living donor liver transplantation for hepatocellular carcinoma: results of prospective patient selection by Kyushu University Criteria in 7 years. HPB 2017;19:1082-1090.
7) Lai Q, Nicolini D, Inostroza Nunez M, Iesari S, Goffette P, Agostini A, et al. A novel prognostic index in patients with hepatocellular cancer waiting for liver transplantation: Time-Radiological-response-Alpha-fetoprotein-INflammation (TRAIN) Score. Ann Surg 2016;264:787-796.
8) Sapisochin G, Goldaracena N, Laurence JM, Dib M, Barbas A, Ghanekar A, et al. The extended Toronto criteria for liver transplantation in patients with hepatocellular carcinoma: a prospective validation study. Hepatology 2016;64:2077-2088.
9) **Sasaki K**, **Firl DJ**, Hashimoto K, Fujiki M, Diago-Uso T, Quintini C, et al. Development and validation of the HALTHCC score to predict mortality in liver transplant recipients with

hepatocellular carcinoma: a retrospective cohort analysis. Lancet Gastroenterol Hepatol 2017;2:595-603.
10) Yao F. Liver transplantation for hepatocellular carcinoma: Expansion of the tumor size limits does not adversely impact survival. Hepatology 2017;33:1394-1403.
11) Firl DJ, Kimura S, McVey J, Hashimoto K, Yeh H, Miller CM, et al. Reframing the approach to patients with hepatocellular carcinoma: longitudinal assessment with HALTHCC improves ablate and wait strategy. Hepatology 2018;68:1448-1458.
12) Silva MF, Sherman M. Criteria for liver transplantation for HCC: what should the limits be? J Hepatol 2011;55:1137-1147.
13) Hoshida Y, Villanueva A, Kobayashi M, Peix J, Chiang DY, Camargo A, et al. Gene expression in fixed tissues and outcome in hepatocellular carcinoma. N Engl J Med 2008;359:1995-2004.
14) Nault J, De Reyniès A, Villanueva A, Calderaro J, Rebouissou S, Couchy G, et al. A hepatocellular carcinoma 5-gene score associated with survival of patients after liver resection. Gastroenterology 2013;145:176-187.
15) Lee HW, Song GW, Lee SG, Kim JM, Joh JW, Han DH, et al. Patient selection by tumor markers in liver transplantation for advanced hepatocellular carcinoma. Liver Transpl 2018;24:1243-1251.
16) Marsh JW, Dvorchik I, Subotin M, Balan V, Rakela J, Popechitelev EP, et al. The prediction of risk of recurrence and time to recurrence of hepatocellular carcinoma after orthotopic liver transplantation: a pilot study. Hepatology 1997;26: 444-450.
17) Zhang M, Yin F, Chen B, Li B, Li YP, Yan LN, Wen TF. Mortality risk after liver transplantation in hepatocellular carcinoma recipients: a nonlinear predictive model. Surgery 2012;151:889-897.
18) Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 2006;25:127-141.
19) Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD Statement. Ann Intern Med 2015;162:55.
20) Agopian VG, Harlander-Locke M, Zarrinpar A, Kaldas FM, Farmer DG, Yersiz H, et al. A novel prognostic nomogram accurately predicts hepatocellular carcinoma recurrence after liver transplantation: analysis of 865 consecutive liver transplant recipients. J Am Coll Surg 2015;220:416-427.
21) Matts JP, Lachin JM. Properties of permuted-block randomization in clinical trials. Control Clin Trials 1988;9:327-344.
22) Armstrong PW, Westerhout CM, Van de Werf F, Califf RM, Welsh RC, Wilcox RG, Bakal JA. Refining clinical trial composite outcomes: an application to the Assessment of the Safety and Efficacy of a New Thrombolytic–3 (ASSENT-3) trial. Am Heart J 2011;161:848-854.
23) Rauch G, Jahn-Eimermacher A, Brannath W, Kieser M. Opportunities and challenges of combined effect measures based on prioritized outcomes. Stat Med 2014;33:1104-1120.
24) Ananth C. Regression models for ordinal responses: a review of methods and applications. Int J Epidemiol 1997;26:1323-1333.
25) Pencina MJ, D' Agostino RB, D' Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med 2008;27:157-172.
26) 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 following liver transplantation for hepatocellular carcinoma. Gastroenterology 2018;154:128-139.
27) Lai Q, Inostroza M, Rico Juri JM, Goffette P, Lerut J. Delta-slope of alpha-fetoprotein improves the ability to select liver transplant patients with hepatocellular cancer. HPB 2015;17: 1085-1095.
28) Halazun KJ, Najjar M, Abdelmessih RM, Samstein B, Griesemer AD, Guarrera JV, et al. Recurrence after liver transplantation for hepatocellular carcinoma: a new MORAL to the story. Ann Surg 2017;265:557-564.
29) Kaido T, Ogawa K, Mori A, Fujimoto Y, Ito T, Tomiyama K, et al. Usefulness of the Kyoto criteria as expanded selection criteria for liver transplantation for hepatocellular carcinoma. Surgery 2013;154:1053-1060.
30) Ioannidis JP, Lau J. Heterogeneity of the baseline risk within patient populations of clinical trials: a proposed evaluation algorithm. Am J Epidemiol 1998;148:1117-1126.

Author names in bold designate shared co-first authorship.

Supporting Information

Additional Supporting Information may be found at onlinelibrary.wiley.com/doi/10.1002/hep.30838/suppinfo.