ORIGINAL ARTICLE

AASLD
AMERICAN ASSOCIATION FOR
THE STUDY OF LIVER DISEASES

OPEN

MELD, MELD 3.0, versus Child score to predict mortality after acute variceal hemorrhage: A multicenter US cohort

Adam Buckholz[1] | Rochelle Wong[2] | Michael P. Curry[3] | Gyorgy Baffy[4] |
Eric Chak[5] | Tarun Rustagi[6] | Arpan Mohanty[7] | Brett E. Fortune[8]

[1]Division of Gastroenterology and Hepatology, Weill Cornell Medical Center, New York, New York, USA

[2]Department of Internal Medicine, Weill Cornell Medical Center, New York, New York, USA

[3]Division of Gastroenterology/Liver Center, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA

[4]Department of Medicine, VA Boston Healthcare System, Boston, Massachusetts, USA

[5]Division of Gastroenterology and Hepatology, University of California Davis School of Medicine, Sacramento, California, USA

[6]Interventional Endoscopy Services, California Pacific Medical Center, San Francisco, California, USA

[7]Section of Gastroenterology, Chobanian & Avedisian School of Medicine, Boston, Massachusetts, USA

[8]Division of Hepatology, Montefiore Einstein Medical Center, Bronx, New York, USA

Correspondence
Brett E. Fortune, 111 East 210th Street 2nd Floor, Rosenthal SE, Bronx, NY 10467, USA.
Email: bfortune@montefiore.org

Abstract

Background: Acute variceal hemorrhage is a major decompensating event in patients with cirrhosis and is associated with high 6-week mortality risk. Many prognostic models based on clinical and laboratory parameters have been developed to risk stratify patients on index bleeding presentation, including those based on the Model for End-Stage Liver Disease (MELD) and Child-Turcotte-Pugh (CTP). However, consensus on model performance remains unclear.

Methods: Using a large US multicenter cohort of hospitalized patients with cirrhosis who presented with acute variceal hemorrhage, this study evaluates, recalibrates, and compares liver severity index-based models, including the more recent MELD 3.0 model, to investigate their predictive performance on 6-week mortality. Models were also recalibrated and externally validated using additional external centers.

Results: All recalibrated MELD-based and CTP-based models had excellent discrimination to identify patients at higher risk for 6-week mortality on initial presentation. The recalibrated CTP score model maintained the best calibration and performance within the validation cohort. Patients with low CTP scores (Class A, score 5–6) were strongly associated with < 5% mortality, while high CTP score (Class C, score > 9) were associated with > 20% mortality.

Abbreviations: AUROC, area under the receiver operating curve; AVH, acute variceal hemorrhage; BIDMC, Beth Israel Deaconess Medical Center; BMC, Boston Medical Center; BMI, body mass index; CTP, Child-Turcotte-Pugh score; CC, Child-Turcotte-Pugh Class; aCTP, adjusted Child-Turcotte-Pugh based on previously published adjustments by Fortune et al; rCTP, recalibrated Child-Turcotte-Pugh score; rCTP CC, recalibrated Child-Turcotte-Pugh Class; HL, Hosmer–Lemeshow; INR, international normalized ration; MELD, Model for End-Stage Liver Disease; aMELD, adjusted Model for End-Stage Liver Disease based on previously published adjustments by Reverter et al; rMELD, recalibrated MELD score; rMELD3, recalibrated MELD 3.0; UC Davis, University of California, Davis; UNM, University of New Mexico; WCMC, Weill Cornell Medical Center.

Adam Buckholz and Rochelle Wong denotes co-first authorship with equal contribution to the manuscript.

Arpan Mohanty and Brett E. Fortune denotes co-senior authorship with equal contribution to the manuscript.

Supplemental Digital Content is available for this article. Direct URL citations are provided in the HTML and PDF versions of this article on the journal's website, www.hepcommjournal.com.

- -

Conclusion: Use of liver severity index-based models accurately predict 6-week mortality risk for patients admitted to the hospital with acute variceal hemorrhage and supports the utilization of these models in future clinical trials as well as their use in clinical practice.

INTRODUCTION

Cirrhosis is a common cause of morbidity and mortality, ranking as the 11th leading cause of death worldwide and responsible for 51,000 deaths per year in the United States.[1] Acute variceal hemorrhage (AVH) is a decompensating clinical event in cirrhosis and is associated with 15%–25% 6-week mortality.[2–4] Several factors have been associated with poor outcomes in AVH including liver disease severity, captured in indices such as the Model for End-Stage Liver Disease (MELD) score[5,6] and Child-Turcotte-Pugh (CTP) score and its components,[7–9] active bleeding on endoscopy,[6] transfusion of packed red blood cell,[6] renal failure,[10,11] and a HVPG.[8,12] Using these indicators to identify high-risk patients may allow for more timely intervention to reduce the risk of rebleeding and mortality, as evidenced by the use of CTP score for pre-emptive TIPS creation for AVH.[13]

As the use of HVPG for prognostication and risk stratification of patients with AVH is limited in general practice, many models based on clinical and laboratory parameters have been developed. The MELD and CTP scores are clinically relevant, used most frequently, and have been shown to independently predict 6-week mortality in AVH.[5,14] However, there is no consensus on which method is superior at prognostication and risk stratification. Reverter et al[15] observed that recalibration of a MELD-based model had better predictive performance for 6-week mortality in patients with AVH, while Fortune et al[16] suggested a CTP-based model worked better. Furthermore, MELD 3.0 has since been proposed using the published MELD formula with upper and lower bounds of albumin, sodium, and creatinine to reduce liver transplant waitlist mortality.[17] To our knowledge, no studies have compared all 3 models or validated these findings using a large multisite US cohort.

Thus, this study aims to evaluate and compare the capabilities of MELD-based and CTP score-based models in predicting 6-week mortality in AVH, so as to identify those at highest risk.

METHODS

Study population

In this national multicenter retrospective cohort study, adult patients who were hospitalized at 5 academic medical centers between 2012 and 2019 and had endoscopically proven AVH were included. Cirrhosis was diagnosed clinically with supporting laboratory and imaging characteristics such as nodular texture of liver and/or features of portal hypertension. Patients with nonvariceal bleeding, prior history of liver transplant, prior history of portosystemic shunt (TIPS or surgical shunt), or noncirrhotic portal hypertension were excluded. Mortality due to hemorrhage before endoscopic evaluation to ascertain bleeding source was excluded. As recent variceal bleeding is a strong predictor of rebleeding,[8] patients with an episode of portal hypertensive bleeding within 1 year of the current presentation were excluded. Patients with advanced HCC were also excluded given their high expected non–bleeding-related 6-week mortality. Liver transplant within 6 weeks of presentation was considered a competing risk, and these patients were excluded. All research was conducted in accordance with both the Declarations of Helsinki and Istanbul and all research was approved by the appropriate institutional review board at each participating site.

Data collection and therapeutic interventions

Clinical data were extracted by review of the electronic medical record. Variables collected included patient demographics, comorbidities, initial vital signs, admission labs, imaging reports, endoscopic interventions and reports, and hospitalization outcomes. Features of cirrhosis complications, including previous portal hypertensive bleeding, known PVT, ascites, spontaneous bacterial peritonitis, HE, and hepatorenal syndrome, were also collected. Severity of liver disease on index admission was captured by their MELD and CTP scores.

Therapeutic interventions on AVH presentation, including standard-of-care vasoactive drug administration with octreotide and antibiotic prophylaxis, were collected. Type of endoscopic intervention, such as banding, sclerotherapy, both or none, and need for repeat endoscopy, rescue TIPS, or balloon tamponade were included. Time of admission was considered time zero, and outcomes were followed until death or 1 year, as available. In accordance with the Baveno VI consensus guidelines,[18] the primary clinical outcome was mortality at 6 weeks. Additional outcomes included

24-hour rebleeding, 5-day rebleeding and mortality, and length of hospital stay.

Training and validation cohorts

Patients in this study were included from 5 academic centers across the country: the training cohort was derived from Weill Cornell Medical Center in New York, and the validation cohort comprised patients from 4 other institutions: University of California, Davis, University of New Mexico, Boston Medical Center, and Beth Israel Deaconess Medical Center. All serve as tertiary care referral centers for their local communities.

Selecting prognostic models

To maximize clinical relevance, we evaluated the discrimination of 4 clinically relevant, familiar models for 6-week mortality in AVH:

(1) MELD.
(2) CTP score (as a continuous variable).
(3) CTP Class (CC) with CTP score as a categorical variable where CC A: CTP score 5–6, CC B: CTP score 7–9, and CC C: CTP score 10–15.
(4) MELD 3.0, a recently derived prognostic model incorporating upper and lower bounds for albumin, sodium, and creatinine.[19]

We evaluated the initial calibration of MELD and previously published recalibrated model adjustments by Reverter et al[15] and Fortune et al[16] for MELD (henceforth aMELD) and CTP (aCTP), respectively. Both were assessed in our training and validation cohorts. Discrimination describes a model's ability to correctly rank likelihood of a binary outcome, that is, 6-week mortality. We tested discrimination by assessing area under the receiver operating curve to plot true positive and true negative rates. Calibration is the agreement between the predicted ("estimated") and observed ("true") numerical risk of the outcome. Predicted probability of 6-week mortality from MELD was calculated using the originally published equation $S(t) = S_0(t)^{\exp(R-1.127)}$ where $S(t)$ is the probability of survival, R is the risk score calculated from bilirubin, creatinine in the individual patient, and INR values and $S_0(t)$ is a previously published temporal constant (0.827) representing the underlying survival probability for an average patient in the original study. Predicted mortality is $1-S(t)$.[15] Predicted probability of 6-week mortality from aMELD was calculated as (logit = −5.312 + 2.07 × MELD).[15] Predicted probability of 6-week mortality for aCTP was calculated as logit adjustment (logit = −5.5125 + 0.4865 × CTP score).[16] No quantitative prediction of mortality is provided by the original CTP

score and class, so initial calibration could not be assessed. We likewise evaluated MELD 3.0, a recently derived prediction model incorporating albumin and gender into MELD; the previously published formula was used,[17] incorporating upper and lower bounds for albumin, sodium, and creatinine. Because there is no published temporal constant at this time for standard CTP/CC or MELD 3.0 at 6 weeks, initial calibration was not assessed.

Methods for assessing model predictive performance

We evaluated the performance of all chosen models (CTP score and Class, MELD, MELD 3.0, and aMELD and aCTP score) were evaluated for prediction of the primary outcome, 6-week mortality, according to 2 key statistical concepts: discrimination and calibration. Initial model calibration was assessed in those models with prior calibrations for predicted mortality (excluding standard CTP score/class). The lack of calibration was adjudicated by the Hosmer–Lemeshow (HL) goodness of fit test with the data split into quintiles; a $p < 0.05$ was considered significantly miscalibrated, while a higher p-value represents smaller deviation of observed from predicted mortality, or better calibration. Overall model performance was likewise assessed with the Brier score which includes components of both discrimination and calibration, scaled from 0 to 1 where a lower score represents lower difference in predicted versus observed outcomes and overall improved performance. Models with higher HL goodness of fit p values and lower Brier score were considered to have better model performance.[19,20]

Recalibration of models

All models (CTP/CC, MELD, MELD 3.0, aCTP, and aMELD) were miscalibrated across the full cohort. Therefore, we performed logistic recalibration of each model within the training cohort. Recalibration is a method of updating risk models for dichotomous outcomes such as 6-week mortality, by adjusting the model intercept and/or overall slope. The recalibrated models were then internally validated using bootstrapping with replacement in 200 random samples. For prognostic models to assist in clinical practice, it is important to externally validate in data collected in different settings. A model with good discrimination for 6-week mortality may be miscalibrated in another setting if it overestimates or underestimates the expected risk of the outcome. The recalibrated CTP score and CTP class (rCTP, rCC), MELD (rMELD), and MELD 3.0 (rMELD3) were then externally validated in our multicenter validation cohort (Boston Medical Center, Beth Israel Deaconess Medical Center, University of New Mexico, and University of California, Davis). Calibration

was not assessed at each individual site within the validation cohort given small sample size and low total number of deaths within 6 weeks at each site. All data analysis was performed using Stata version 17.0 (StataCorp LLC College Station, TX).

RESULTS

Study cohort and outcomes

Our training set and validation set consisted of 125 patients and 229 patients, respectively. On admission, patients had a mean age of 56 ($\pm$ 12.9) years, body mass index of 27.7 ($\pm$ 5.65), and median CTP Score of 9 (interquartile range 4). Baseline characteristics were largely similar between the 2 groups (Table 1), although those in the validation cohort had a slightly higher admission MELD (17.9 vs. 16.2, $p = 0.02$). On endoscopy, 109 patients (31%) had active bleeding and 46 (12.9%) had failure to control bleeding at initial endoscopy or at 24 hours. Overall, 61 (17.2%) of participants died within 6 weeks, with similar mortality in both the training and validation cohorts (Table 2). Of these, 29 (47.5%) died within 5 days, 62% of whom had failure to control bleeding at initial endoscopy or at 24 hours. Pre-emptive TIPS was underutilized across the full cohort, with only 17/179 (9.5%) of those with CC B and active bleeding on endoscopy or CC C and score < 14

receiving a pre-emptive TIPS. In the training cohort, a clear explanation for foregoing TIPS was found in 33/62 potentially eligible patients who did not receive pre-emptive TIPS. Of these, 11 were deemed to have an unsafely high MELD, 11 had concurrent HE, 6 had anatomical contraindications such as thrombosis, and 5 had an early-stage liver cancer.

Discrimination and initial calibration of prognostic models in training cohort

We analyzed discrimination of MELD, MELD 3.0, and CTP score. Of note, recalibration (as in the aCTP and aMELD models) does not alter discrimination. MELD 3.0 had the highest discrimination for 6-week mortality after AVH, but all models fell well within statistical equivalence by 95% CI, with similar findings when considering the full cohort (n = 354) (Table 3). When evaluating calibration by HL goodness of fit, both MELD and aCTP were significantly miscalibrated at $p < 0.05$ (Figure 1). Both models significantly overestimated mortality overall, especially those at low MELD and CTP score. The aMELD was calibrated ($p = 0.83$) within the training cohort but overpredicted 6-week mortality at the highest MELD quintiles. By the Brier score, the aMELD (0.114) outperformed both MELD (0.126) and aCTP calibration (0.131) in the training cohort.

TABLE 1   Baseline characteristics on admission of included patients

	Training cohort (n = 125)	Validation cohort (n = 229)	Total (n = 354)	p (Train vs. Val)
Age, mean (SD)	58.5 (13.6)	54.7 (12.3)	56.0 (12.9)	0.01
Gender, M, n (%)	90 (72)	159 (69.4)	250 (70.4)	0.58
BMI (kg/m^2), mean (SD)	26.9 (5.9)	28.1 (5.4)	27.7 (5.7)	0.059
HCC, n (%)	25 (20)	30 (13.1)	55 (15.5)	0.09
PVT, n (%)	17 (13.6)	17 (7.4)	34 (9.6)	0.06
HE, n (%)	31 (24.8)	78 (34.1)	109 (30.7)	0.07
Ascites, n (%)	80 (64.0)	148 (64.6)	228 (64.2)	0.83
MELD, mean (SD)	16.2 (6.7)	17.9 (7.4)	17.4 (7.2)	0.02
CTP score, mean (SD)	8.8 (2.2)	8.9 (2.2)	8.9 (2.2)	0.76
CTP Class, n (%)				0.51
A	17 (13.6)	37 (16.2)	54 (15.2)	
B	64 (51.2)	102 (44.5)	166 (46.8)	
C	44 (35.2)	90 (39.3)	135 (38.0)	
Admission labs				
Systolic blood pressure, mean (SD)	115.9 (23.8)	116.8 (23.3)	116.5 (23.4)	0.72
Platelets, mean (SD)	134.9 (128.6)	120.7 (73.1)	125.7 (96.6)	0.18
Sodium, mean (SD)	136.2 (5.5)	137.3 (5.6)	136.9 (5.5)	0.07
INR, mean (SD)	1.50 (0.41)	1.68 (0.58)	1.62 (0.53)	0.003
Creatinine, mean (SD)	1.33 (1.3)	1.29 (1.6)	1.30 (1.5)	0.80
Bilirubin, mean (SD)	3.41 (4.7)	3.89 (6.0)	3.72 (5.6)	0.44
Albumin, mean (SD)	2.79 (0.60)	2.67 (0.71)	2.71 (0.67)	0.11

Abbreviations: BMI, body mass index; CTP, Child-Turcotte-Pugh; INR, international normalized ratio; M, male; MELD, Model for End-Stage Liver Disease

TABLE 2 Outcomes table in training and validation cohort

	Training cohort (n = 125), n (%)	Validation cohort (n = 229), n (%)	Total (n = 354), n (%)	p
6-wk mortality	21 (16.8)	40 (17.5)	61 (17.2)	0.85
5-d mortality	14 (11.2)	15 (6.6)	29 (8.2)	0.13
CTP B+ active bleed or CTP C	71 (56.8)	108 (47.2)	179 (50.6)	0.07
Received pre-emptive TIPS (if eligible)	5 (7.0)	12 (11.1)	17 (9.5)	0.44

Abbreviation: CTP, Child-Turcotte-Pugh.

Calibration assessment across full cohort

After confirming that the aMELD was well calibrated within the training cohort, its calibration was tested (along with aCTP and MELD) using the full cohort. Considering its prior validation in the older European cohort, if it remained well calibrated further analysis may have proven unnecessary. However, in the validation cohort ($p < 0.01$) and in the full cohort ($p < 0.01$), the aMELD was significantly miscalibrated (Supplemental Figure S1, http://links.lww.com/HC9/A510) and had an inferior Brier score to the original MELD in both the validation (0.145 vs. 0.135) and full cohorts (0.134 vs. 0.132). This was primarily due to significant over prediction of mortality at high MELD score. Similarly, aCTP and MELD remained miscalibrated in the full cohort.

Recalibration of MELD, MELD 3.0, and CTP score

Given the significant miscalibration of previously published models for 6-week mortality in AVH, as well as the similar discriminative performance of all models, we recalibrated all of MELD, MELD 3.0, and CTP score, henceforth rMELD, rMELD3, rCTP, using logistic calibration in the training cohort as described above. The recalibrated formulas are as follows [the models have been slightly adjusted]:

$$\text{MELD 3.0 logit} = -4.86 + 0.16 \times \text{MELD 3.0 score}.$$

$$\text{MELD logit} = -4.72 + 0.17 \times \text{MELD score}.$$

$$\text{CTP logit} = -6.47 + 0.51 \times \text{CTP score}.$$

Absolute predicted values for 6-week mortality can be found in Table 2. In the training cohort, all scores were well calibrated by HL goodness of fit (Figure 2A–C). Likewise, CTP was well calibrated when grouped into class (Figure 2D). After random bootstrapping, all models demonstrated low bias, suggesting strong internal validation.

Assessment of recalibrated model performance in validation cohort

When assessing calibration within the validation cohort, rMELD3 and rCTP/rCC remained technically calibrated, while rMELD was significantly miscalibrated in the validation cohort (Figure 3A–D). The rCTP score maintained the best calibration within the validation cohort ($p = 0.83$) followed by rCC ($p = 0.33$) and rMELD3 ($p = 0.26$), while even after recalibration MELD overpredicted mortality at high values ($p = 0.04$). By the Brier score, rCTP (1.26) and rCC (1.30) outperformed rMELD (1.47) and rMELD3 (1.39) in the validation cohort. Across the entire cohort (training + validation), rMELD3 and rCTP had superior performance relative to MELD, as well as the rMELD, aMELD, and aCTP (Table 4). A sensitivity analysis controlling for TIPS before 6 weeks did not change the conclusion that rCTP and rMELD3 were the best-calibrated models for 6-week mortality prediction. However, the aMELD was calibrated when excluding those receiving TIPS or transplant ($p = 0.21$).

## Ascertaining clinically relevant thresholds for 6-week mortality after AVH

The primary value of recalibration is to allow practitioners to accurately use common indices, unaltered in their calculation, to understand probable outcomes.

TABLE 3 Discrimination of MELD, MELD 3.0, and CTP score for 6-week mortality

	Training cohort AUROC [95% CI]	Validation cohort AUROC [95% CI]	Full cohort AUROC [95% CI]
CTP score	0.79 [0.70, 0.88]	0.75 [0.67, 0.83]	0.77 [0.71, 0.83]
MELD	0.80 [0.70, 0.91]	0.74 [0.66, 0.81]	0.76 [0.70, 0.82]
MELD 3.0	0.81 [0.71, 0.91]	0.74 [0.66, 0.81]	0.77 [0.71, 0.83]

Abbreviations: AUROC, area under the receiver operating curve; CTP, Child-Turcotte-Pugh; MELD, Model for End-Stage Liver Disease.
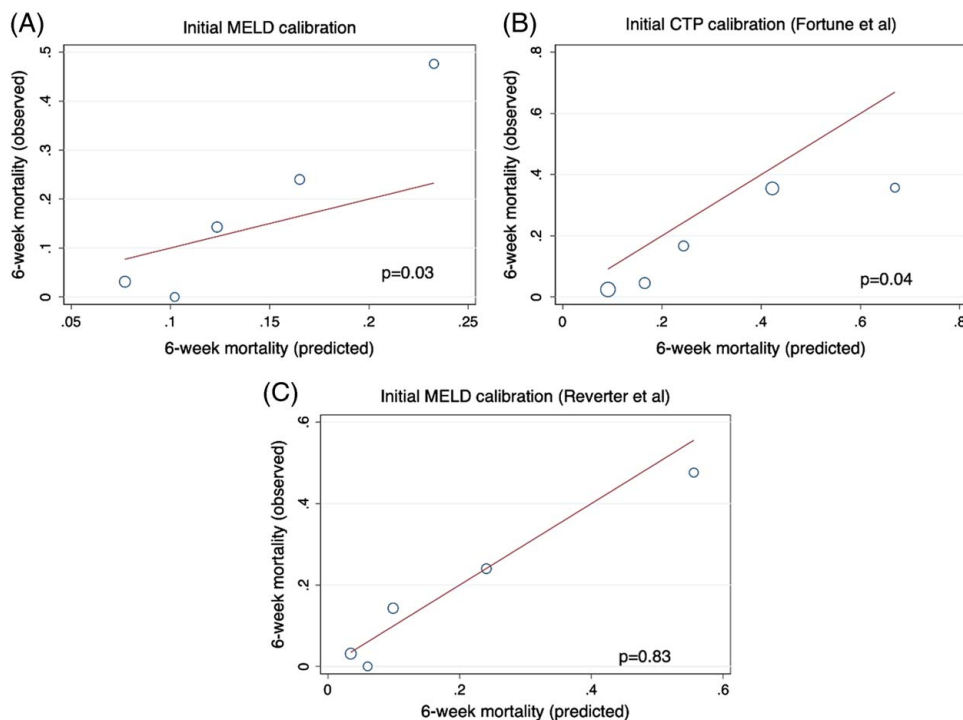
FIGURE 1 Initial assessment of calibration of MELD (A), Fortune CTP (B), and Reverter MELD (C) in the training cohort demonstrates significant miscalibration of both MELD and the previously published recalibration of CTP score. The previously published recalibration of MELD was calibrated in the training cohort. Abbreviations: CTP, Child-Turcotte-Pugh; MELD, Model for End-Stage Liver Disease.

Because overall performance for predicting 6-week mortality was highest in the rCTP and rMELD3, we sought to identify simple cutoffs which accurately grouped patients by 6-week mortality after AVH. Using nonparametric locally weighted scatterplot smoothing, we stratified participants across the full cohort (n = 354) and in each available cohort (Weill Cornell Medical Center, Boston Medical Center, Beth Israel Deaconess Medical Center, University of New Mexico, and University of California, Davis) and plotted observed and predicted mortality. We found that CC A (5–6) was strongly associated with <5% mortality, while CC C (>9) was associated with >20% mortality (Figure 4A–B). Similarly, a MELD 3.0 <12 was associated with <5% mortality and >20 was associated with >20% mortality (Figure 4C, D). At high MELD 3.0 scores (and similarly at high MELD score), there was significant skewing of the observed mortality relative to that seen between cohorts with CTP score.

DISCUSSION

Using a large contemporaneous multisite US cohort of hospitalized patients with cirrhosis presenting with AVH, we extensively investigated the prediction performance of the CTP, MELD, and MELD 3.0 scores to further refine our understanding of mortality risk after AVH. As expected, mortality increased among those with greater severity of liver disease severity, and MELD, MELD 3.0,

and CTP all had excellent discrimination to identify those at higher likelihood of death. However, after recalibration and external validation of these scores, we found that CTP score had the best prediction for 6-week mortality.

The performance of recalibrated CTP score was better than a recalibrated MELD 3.0, even though they had similar Brier scores. Unlike the recalibrated MELD 3.0, the predictive accuracy for CTP score was not attenuated at higher scores. In the study by Reverter et al,[15] the authors posited that MELD was the most appropriate predictive score, thus the one chosen for recalibration, based on higher discrimination than the D'Amico, Augustin and CTP scores; however, the increased discrimination did not reach statistical significance. Assessing model discrimination within only a single cohort before re-calibrating and validating only a single model (MELD) has the potential to introduce bias.[21,22] Moreover, temporal (data collected at different time) and external (new data in different location; by different investigators) validation of the model is important to determine its generalizability.[23] A significant strength of this study is the decision to recalibrate all models and assess performance more comprehensively and generally, especially as we found similar discrimination across all models. Of note, we also chose not to evaluate the D'Amico and Augustin scores due to their uncommon usage in clinical practice, thus limiting their value even after recalibration.
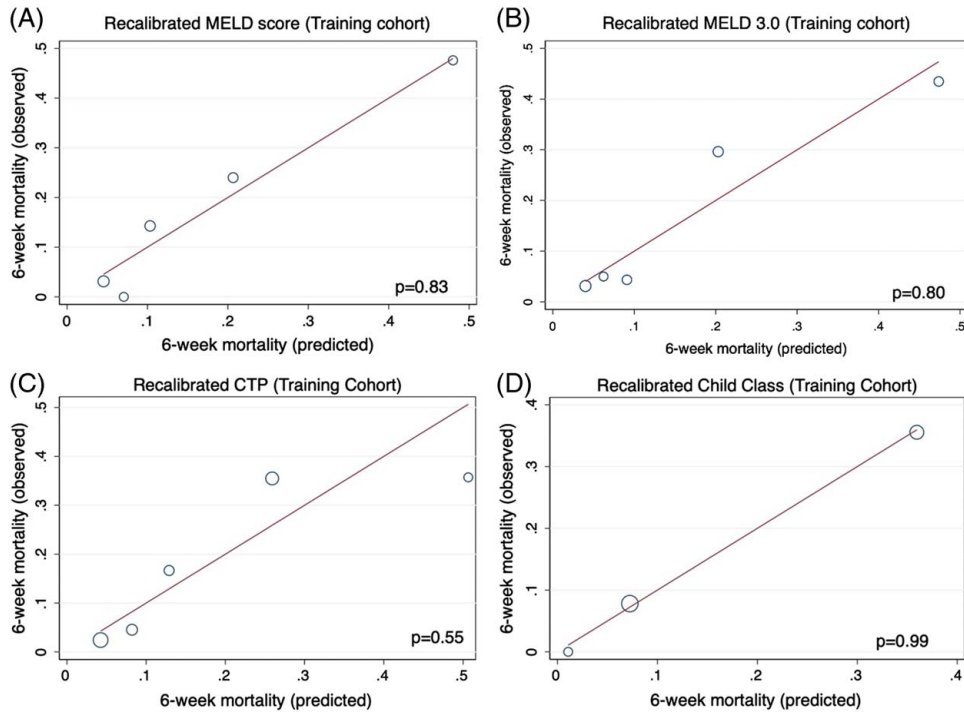
FIGURE 2 Recalibration of MELD (A), MELD 3.0 (B), CTP Score (C) and Child class (D) demonstrated improved performance within the training cohort. Abbreviations: CTP, Child-Turcotte-Pugh; MELD, Model for End-Stage Liver Disease.

Predictive thresholds are important, as they can help risk stratify patients immediately on admission or index endoscopy with basic clinical and laboratory parameters. Based on the findings of this study, in those with low MELD 3.0 (< 12) or CTP Score (Class A), AVH is unlikely to result in significant 6-week mortality. Conversely, those with high MELD 3.0 (> 20) or CTP Score (Class C) have a high short-term risk of rebleeding and
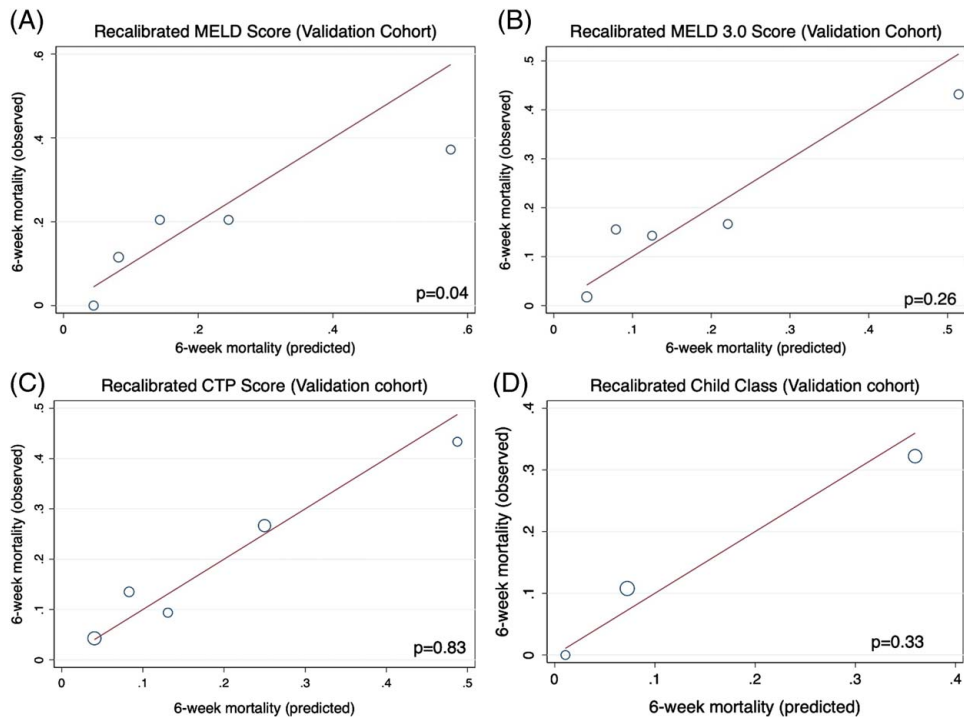
FIGURE 3 After recalibration of MELD (A), MELD 3.0 (B), CTP Score (C), and Child Class (D), all predictive models remained technically calibrated within the validation cohort, except for the recalibrated MELD, which was significantly miscalibrated in the validation cohort. Abbreviations: CTP, Child-Turcotte-Pugh; MELD, Model for End-Stage Liver Disease.

**TABLE 4** Overall comparison of model performance

	Brier score (training)	Brier score (validation)	Brier score (full cohort)	HL correlation (*p*)
Fortune CTP	0.131	0.140	0.137	< 0.001
Reverter MELD	0.114	0.145	0.135	0.006
Recalibrated MELD	0.112	0.145	0.134	0.02
MELD	0.125	0.135	0.132	0.003
Recalibrated MELD 3.0	0.114	0.139	0.130	0.72
Recalibrated CTP Class	0.118	0.130	0.126	0.46
Recalibrated CTP score	0.120	0.126	0.124	0.74

A lower Brier score and higher HL correlation is considered better calibrated. An HL correlation with $p < 0.05$ is considered significantly miscalibrated.
Abbreviations: CTP, Child-Turcotte-Pugh; HL, Hosmer–Lemeshow; MELD, Model for End-Stage Liver Disease.

death. Such thresholds do not rely on data from a hospital course including infection or other comorbidities, although certainly those with more advanced liver disease are likely to have additional concurrent confounders.

One limitation of our study is its retrospective nature, meaning we are unable to systematically control for all potential confounders which may impact the relative mortality in each site, thus affecting our assessment of model performance. We also are unable to ensure that all patients are diagnosed with cirrhosis by gold standard liver biopsy; however, it is common practice to use a clinical diagnosis of cirrhosis based on physical examination, imaging, and laboratory criteria in prospective trials of inpatients with end-stage liver disease.[24,25] One important observation from this study was the low rate of pre-emptive TIPS utilization among those who would qualify, based on previous recommendations.[26] Given that across the full cohort only 17 patients received a pre-emptive TIPS within eligibility criteria, our study was not powered to evaluate outcomes in this group relative to those who did not receive a TIPS. Several other studies have noted the low utilization (< 15%) of pre-emptive TIPS in clinical practice, which may be influenced by its
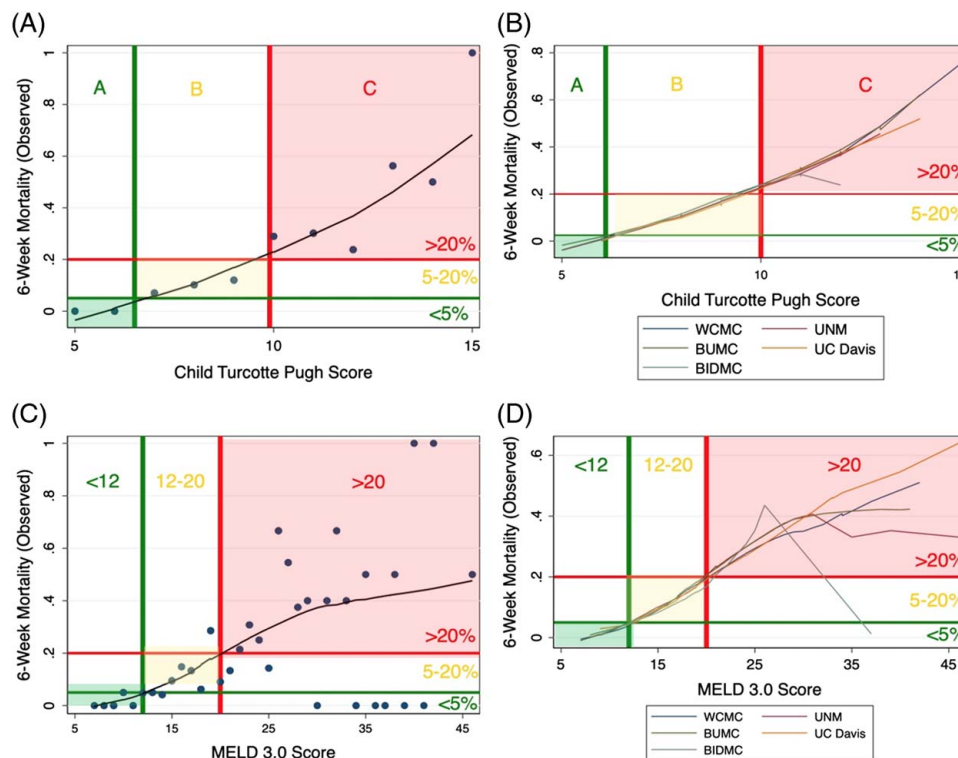
FIGURE 4 Locally weighted scatterplot smoothing smoothing and identification of thresholds for CTP Score in training cohort (A), CTP Score in validation cohort (B), MELD 3.0 Score in training cohort (C), and MELD 3.0 Score in validation cohort (D). Locally weighted scatterplot smoothing smoothing uses a robust weighting system to reduce the impact of outliers on a model, instead focusing on density. Plotted against observed outcomes, we have observed approximate mortality thresholds (denoted by the colored lines) which may be useful in clinical risk stratification. Abbreviations: BIDMC, Beth Israel Deaconess Medical Center; BUMC, Boston University Medical Center, MELD, Model for End-Stage Liver Disease; UC Davis, University of California, Davis; UNM, University of New Mexico; WCMC, Weill Cornell Medical Center

slow implementation into clinical practice guidance statements.[27–29] We were unable to control for other patient factors which may have influenced whether they received a TIPS, such as consent to the procedure or amenable vasculature. As noted in the training cohort, many patients did not receive pre-emptive TIPS due to noted medical contraindications, but further investigation into practice quality is warranted in order to understand adoption of evidence-supported interventions. In addition, TIPS was not excluded in our primary analysis as a competing risk. Given the known impact of this intervention for reducing short-term mortality in portal hypertension, inability to perform a formal competing risk analysis as date of TIPS was unavailable is a weakness of this study. However, adjusting for these patients in a sensitivity analysis did not alter the overall conclusion or model calibration.

The reason for increased heterogeneity in the observed mortality at high MELD 3.0 among the various cohorts may be secondary to 2 factors. The first, which serves as a weakness in any comparison, is that the number of individuals with each MELD 3.0 score will be less than the number with each CTP score as the possible range of values is much larger in MELD. Therefore, with local weighting, 1 or 2 aberrant results (such as one patient with high MELD 3.0 who survived) is more likely to skew a locally weighted scatterplot smoothing curve. Second, which may be more clinically relevant, is that while CTP contains subjective variables, its composite factors are less variable in the short term from the overall health of the patient. In contrast, sodium and creatinine (both components of MELD 3.0 but not CTP) may have a significant departure from a patient's chronic baseline after a variceal bleed and on hospital admission. The creatinine used to calculate an admission MELD 3.0 may not accurately reflect the patient's underlying renal function. Furthermore, recent publication of using a combined score of MELD and peak serum lactate levels, MELD-lactate, highlights a need to additionally account for other acuity measures like organ hypoperfusion.[30] However, score-model–based simplicity must be considered to optimize utilization within clinical practice.

In summary, our study uniquely compared, recalibrated, and externally validated the 6-week mortality prediction capabilities of MELD, MELD 3.0, and CTP score-based models to help clinicians best identify those at highest risk of mortality in patients with cirrhosis presenting with AVH. Using one of the largest known US cohorts of patients with AVH, we found that CTP-based and MELD 3.0-based models best predicted 6-week mortality after AVH. These findings further support utilization of these liver disease score-based models in future therapeutic trials to treat acute variceal hemorrhage. By validating accurate risk stratification, we hope to aid in the design of pragmatic clinical trials to address tailoring risk-based therapeutic regimens in the management of acute variceal bleeding. For example, given the low risk of mortality, rebleeding and infection in CTP A or MELD 3.0 < 12 patients, trials can be performed to investigate outcomes without the use of antibiotics. Alternatively, more emphasis should be placed on escalated care for those in high-risk groups (CTP C or MELD 3.0 > 19) as standard therapy does not adequately reduce short-term mortality. Personalized regimens to treat AVH and recurrence are needed in future trials as the "one rule for all" management strategy requires modification.

CONFLICTS OF INTEREST
Michael P. Curry consults and received grants from Sonic Incytes. He consults for Mallinckrodt, Albireo and Alexion and advises Pfizer. Eric Chak received grants from Target RWE and GlaxoSmithKline. Tarun Rustagi consults for Boston Scientific. Arpan Mohanty received grants from Gilead and the Kinetix Group. Brett E. Fortune consults for Gore. The remaining authors have no conflicts to report.

ORCID
Adam Buckholz https://orcid.org/0000–0002–0068–353X
Rochelle Wong https://orcid.org/0000–0003–3326–3360
Gyorgy Baffy https://orcid.org/0000–0002–8334–0400
Arpan Mohanty https://orcid.org/0000–0002–8408–8075
Brett E. Fortune https://orcid.org/0000–0002–0646–467X

REFERENCES
1. FastStats – Chronic Liver Disease or Cirrhosis. Accessed January 9, 2023. https://www.cdc.gov/nchs/fastats/liver-disease.htm
2. Garcia-Tsao G, Sanyal AJ, Grace ND, Carey W. Practice Guidelines Committee of the American Association for the Study of Liver Diseases the Practice Parameters Committee of the American College of Gastroenterology. Prevention and management of gastroesophageal varices and variceal hemorrhage in cirrhosis. Hepatology. 2007;46:922–38.
3. Bosch J, Abraldes JG, Groszmann R. Current management of portal hypertension. J Hepatol. 2003;38:54–68.
4. Garcia-Tsao G, Abraldes JG, Berzigotti A, Bosch J. Portal hypertensive bleeding in cirrhosis: Risk stratification, diagnosis, and management: 2016 practice guidance by the American Association for the study of liver diseases. Hepatology. 2017;65:310–35.
5. Kamath PS, Wiesner RH, Malinchoc M, Kremers W, Therneau TM, Kosberg CL, et al. A model to predict survival in patients with end-stage liver disease. Hepatology. 2001;33:464–70.
6. Bambha K, Kim WR, Pedersen R, Bida JP, Kremers WK, Kamath PS. Predictors of early re-bleeding and mortality after acute variceal haemorrhage in patients with cirrhosis. Gut. 2008;57:814–20.
7. Augustin S, Muntaner L, Altamirano JT, González A, Saperas E, Dot J, et al. Predicting early mortality after acute variceal

hemorrhage based on classification and regression tree analysis. Clin Gastroenterol Hepatol. 2009;7:1347–54.

8. Abraldes JG, Villanueva C, Bañares R, et al. Hepatic venous pressure gradient and prognosis in patients with acute variceal bleeding treated with pharmacologic and endoscopic therapy. J Hepatol. 2008;48:229–36.

9. D'Amico G, de Franchis R. Upper digestive bleeding in cirrhosis. Post-therapeutic outcome and prognostic indicators. Hepatology. 2003;38:599–612.

10. Olmo JAD, Peña A, Serra MA, Wassel AH, Benages A, Rodrigo JM. Predictors of morbidity and mortality after the first episode of upper gastrointestinal bleeding in liver cirrhosis. J Hepatol. 2000;32:19–24.

11. Hsieh YC, Lee KC, Chen PH, Su CW, Hou MC, Lin HC. Acute kidney injury predicts mortality in cirrhotic patients with gastric variceal bleeding. J Gastroenterol Hepatol. 2017;32:1859–66.

12. Moitinho E, Escorsell A, Bandi JC, Salmerón JM, García-Pagán JC, Rodés J, et al. Prognostic value of early measurements of portal pressure in acute variceal bleeding. Gastroenterology. 1999;117:626–31.

13. Garcia-Pagán JC, di Pascoli M, Caca K, Laleman W, Bureau C, Appenrodt B, et al. Use of early-TIPS for high-risk variceal bleeding: results of a post-RCT surveillance study. J Hepatol. 2013;58:45–50.

14. Durand F, Valla D. Assessment of the prognosis of cirrhosis: Child-Pugh versus MELD. J Hepatol. 2005;42(suppl 1):S100–7.

15. Reverter E, Tandon P, Augustin S, et al. A MELD-based model to determine risk of mortality among patients with acute variceal bleeding. Gastroenterology. 2014;146:412–19.e3.

16. Fortune BE, Garcia-Tsao G, Ciarleglio M, Deng Y, Fallon MB, Sigal S, et al. Child-Turcotte-Pugh Class is best at stratifying risk in variceal hemorrhage: analysis of a US Multi-Center Prospective Study. J Clin Gastroenterol. 2017;51:446.

17. Kim WR, Mannalithara A, Heimbach JK, Kamath PS, Asrani SK, Biggins SW, et al. MELD 3.0: the Model for End-Stage Liver Disease updated for the modern era. Gastroenterology. 2021; 161:1887–895.e4.

18. De Franchis R, Abraldes JG, Bajaj J, Berzigotti A, Bosch J, Burroughs AK, et al. Expanding consensus in portal hypertension: report of the Baveno VI Consensus Workshop: stratifying risk and individualizing care for portal hypertension. J Hepatol. 2015;63:743–52.

19. Brier WG. Verification of forecasts expressed in terms of probability. MWRv. 1950;78:1.

20. Hosmer DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression: Third Edition. Applied Logistic Regression: Third Edition. Published online August 29, 2013:1–510. doi:10.1002/9781118548387.

21. Samaga D, Hornung R, Braselmann H, Hess J, Zitzelsberger H, Belka C, et al. Single-center versus multi-center data sets for molecular prognostic modeling: a simulation study. Radiat Oncol. 2020;15:109.

22. Bafeta A, Dechartres A, Trinquart L, Yavchitz A, Boutron I, Ravaud P. Impact of single centre status on estimates of intervention effects in trials with continuous outcomes: meta-epidemiological study. BMJ. 2012;344:e813.

23. Prognosis and prognostic research: validating a prognostic model on JSTOR. Accessed July 11, 2023. https://www.jstor.org/stable/25671796

24. Wong F, Pappas SC, Curry MP, Reddy KR, Rubin RA, Porayko MK, et al. Terlipressin plus albumin for the treatment of type 1 hepatorenal syndrome. N Engl J Med. 2021;384:818–28.

25. Thursz MR, Richardson P, Allison M, Austin A, Bowers M, Day CP, et al. Prednisolone or pentoxifylline for alcoholic hepatitis. N Engl J Med. 2015;372:1619–28.

26. García-Pagán JC, Caca K, Bureau C, et al. Early use of TIPS in patients with cirrhosis and variceal bleeding. N Engl J Med. 2010;362:2370–9.

27. Hernández-Gea V, Procopet B, Giráldez Á, Amitrano L, Villanueva C, Thabut D, et al. Preemptive-TIPS improves outcome in high-risk variceal bleeding: an observational study. Hepatology. 2019;69:282–93.

28. Thabut D, Pauwels A, Carbonell N, Remy AJ, Nahon P, Causse X, et al. Cirrhotic patients with portal hypertension-related bleeding and an indication for early-TIPS: a large multicentre audit with real-life results. J Hepatol. 2017;68: 73–81.

29. De Franchis R, Bosch J, Garcia-Tsao G, Reiberger T, Ripoll C, Baveno VII Faculty. Baveno VII—renewing consensus in portal hypertension. J Hepatol. 2022;76:959–74.

30. Horvatits T, Mahmud N, Serper M, Seiz O, Reher D, Drolz A, et al. MELD-lactate predicts poor outcome in variceal bleeding in cirrhosis. Dig Dis Sci. 2022;68:1042–50.