External Validation of the United Kingdom-Primary Biliary Cholangitis Risk Scores of Patients With Primary Biliary Cholangitis Treated With Ursodeoxycholic Acid

Angela C. Cheung,[1] Aliya F. Gulamhusein,[2] Brian D. Juran,[1] Erik M. Schlicht,[1] Bryan M. McCauley,[3] Mariza de Andrade,[3] Elizabeth J. Atkinson,[3] and Konstantinos N. Lazaridis[1]

The United Kingdom-Primary Biliary Cholangitis (UK-PBC) risk scores are a set of prognostic models that estimate the risk of end-stage liver disease in patients with PBC at 5-, 10- and 15-year intervals. They have not been externally validated outside the United Kingdom. In this retrospective, external validation study, data were abstracted from outpatient charts and discrimination and calibration of the UK-PBC risk scores were assessed. A total of 464 patients with PBC treated with ursodeoxycholic acid were included. The median diagnosis age was 52.4 years, and 88% were female patients. The cumulative incidence of events was 6%, 9%, and 15% at 5, 10, and 15 years, respectively. Concordance (c-statistic) was 0.88, 0.85, and 0.84 using the 5-, 10- and 15-year risk scores, respectively, which was slightly lower than values observed in the United Kingdom validation cohort. Using the 5-year risk score, more events were observed than predicted (25 versus 16.8; $P = 0.046$); using the 10-year risk score, there was no difference between the observed and predicted number of events (35 versus 44.9; $P = 0.14$); conversely, using the 15-year risk score, fewer events were observed than predicted (46 versus 67.5; $P = 0.009$). Limiting evaluation by the 15-year UK-PBC risk score to those with >10 years of follow-up demonstrated no difference between observed and predicted events. Using the 5-year risk score, patients within the highest quartile had statistically significant worse event-free survival compared to the rest of the cohort: 82% versus 98% at 5 years, 73% versus 97% at 10 years, and 58% versus 93% at 15 years. *Conclusion:* In patients assessed at a North American tertiary medical center, the UK-PBC risk score had excellent discrimination and was reasonably calibrated both in the short and long term. (*Hepatology Communications* 2018;2:676-682)

Primary biliary cholangitis (PBC) is a chronic, autoimmune, cholestatic liver disease that is characterized by the destruction of small intrahepatic bile ducts.[1] Due to its extended natural history, one of the longstanding challenges in the care of patients with PBC has been the ability to prognosticate transplant-free survival.[2] The advent of ursodeoxycholic acid (UDCA) as a treatment for PBC led to development of different binary response criteria (e.g., Barcelona, Paris I, Rotterdam, Paris II, Toronto),[3] all predicated on responses after 1 to 2 years of UDCA treatment. More recently, the aspartate aminotransferase to platelet ratio was developed to predict outcome independent of UDCA, and continuous prognostic risk scores were developed by the Global PBC group (GLOBE score) and the United Kingdom-PBC group (UK-PBC risk scores).[4,5] The UK-PBC risk scores predict the probability of death, liver transplant, or severe hyperbilirubinemia (e.g., bilirubin >5.8 mg/dL) at 5, 10, and 15 years following 1 year of UDCA

Abbreviations: CI, confidence interval; GLOBE, Global primary biliary cholangitis group; IQR, interquartile range; MCPGE, Mayo Clinic PBC Genetic Epidemiology; MELD, Model for End-Stage Liver Disease; PBC, primary biliary cholangitis; UDCA, ursodeoxycholic acid; UK-PBC, United Kingdom-primary biliary cholangitis.

Received December 11, 2017; accepted March 6, 2018.
Additional Supporting Information may be found at onlinelibrary.wiley.com/doi/10.1002/hep4.1186/full.
Supported by the National Institute of Diabetes and Digestive and Kidney Diseases grant number DK80670 (to K.N.L.).

treatment.[4] These 5-, 10-, and 15-year UK-PBC risk scores were derived (n = 1,916) and validated (n = 1,249) from patient data collected from a research network of 155 centers across the United Kingdom.[4]

Given that the UK-PBC risk score models were developed across a variety of transplant and nontransplant centers in the United Kingdom, our aim was to externally validate the UK-PBC risk scores in a North American cohort of patients with PBC seen in a tertiary medical center. We evaluated the UK-PBC risk scores by first assessing discrimination and calibration and then performing a sensitivity analysis to determine the stability of these characteristics. In post hoc analysis, we also evaluated the clinical characteristics and UK-PBC risk scores of patients in the highest quartile of scores.

Materials and Methods

STUDY POPULATION

Eligible subjects from the Mayo Clinic PBC Genetic Epidemiology (MCPGE) Registry followed between April 1987 and October 2017 at the Mayo Clinic were included in this study. The establishment of the MCPGE Registry has been previously described.[6] Subjects were included if their PBC diagnosis was based on internationally accepted criteria, i.e., elevated alkaline phosphatase and positive antimitochondrial antibody or in the case of antimitochondrial antibody-negative subjects, a liver biopsy with classic histologic findings of PBC.[7] Exclusion criteria included age <18 years at study entry, history of concomitant liver disease (e.g., autoimmune hepatitis, viral hepatitis, or nonalcoholic fatty liver disease), no UDCA treatment or unknown first UDCA treatment date, history of hepatocellular carcinoma prior to treatment with UDCA, occurrence of a defined endpoint within a year of starting treatment with UDCA, and lack of biochemical variables required to sufficiently calculate UK-PBC risk scores (see below for necessary variables). This study was approved by the Mayo Clinic Institutional Review Board.

STUDY ENTRY AND OUTCOMES

Values for alkaline phosphatase, aspartate aminotransferase/alanine aminotransferase, and bilirubin were taken at 12 months following initiation of UDCA (where 12-month values were missing, values up to 24 months following UDCA initiation were used); values for albumin and platelets were taken from the date of PBC diagnosis. The UK-PBC risk scores were calculated using an approach consistent with the original UK-PBC manuscript (personal communication with Dr. M. Carbone, Department of Health Sciences, School of Medicine and Surgery, University of Milano, Italy, marco.carbone@unimib.it.). Outcomes were defined using the same criteria as those defined in the derivation of the UK-PBC risk scores[4]: death from a liver-related cause (i.e., liver failure, variceal hemorrhage, or hepatocellular carcinoma), liver transplant or serum bilirubin >5.8 mg/dL. For patients with multiple events, only the first to occur was

View this article online at wileyonlinelibrary.com.
DOI 10.1002/hep4.1186

Potential conflict of interest: Nothing to report.

ARTICLE INFORMATION:

From the [1]Division of Gastroenterology and Hepatology, Mayo Clinic, Rochester, MN; [2]Division of Gastroenterology, Toronto Center for Liver Disease, University Health Network, Toronto, ONT, Canada; [3]Department of Health Sciences Research, Mayo Clinic, Rochester, MN.

ADDRESS CORRESPONDENCE AND REPRINT REQUESTS TO:

Konstantinos N. Lazaridis, M.D.
200 First Street SW
Rochester, MN 55905
E-mail: lazaridis.konstantinos@mayo.edu
Tel: +1-507-284-1006

TABLE 1. CHARACTERISTICS OF THE MCPGE COHORT COMPARED TO THE UK COHORT

Parameters	MCPGE Cohort (n = 464)	UK Cohort (n = 1,916, Derivation)	UK Cohort (n = 1,249, Validation)
Age at diagnosis in years, median (IQR)	52.4 (45.2-59.9)	55.5 (48.5-62.7)	55.2 (47.9-62.8)
Female, no. (%)	409 (88.1)	1,707 (89.1)	1,140 (91.3)
OLT after diagnosis, no. (%)	20 (4.2)	155 (8.1)	105 (8.4)
Follow-up from diagnosis, median (IQR)	11.2 (6.1-17.8)	6.3 (3.2-10.7)	
Follow-up from 12 months following UDCA initiation, median (IQR)	9.7 (4.2-14.6)		
AMA positivity, no. (%)	404 (87.8)	1,667 (87.0)	1,070 (85.7)
ANA positivity, no. (%)	201 (43.3)	392 (20.5)	250 (20.1)
SMA positivity, no. (%)	43 (9.3)	111 (5.8)	91 (7.3)
Ascites at diagnosis, no. (%)	14 (3.0)	22 (1.1)	13 (1.0)
Laboratory values at UDCA initiation, median (IQR)			
ALP × ULN	3.1 (1.4-3.7)	1.9 (1.2-3.5)	2.1 (1.3-3.6)
ALT × ULN	2.6 (1.2-3.3)	1.4 (0.9-2.3)	1.4 (0.9-2.4)
AST × ULN	1.8 (1.1-2.7)	1.4 (0.9-2.3)	1.4 (0.9-2.4)
TB × ULN	0.6 (0.5-1.0)	0.5 (0.4-0.8)	0.5 (0.4-0.8)
ALB × LLN*	1.2 (1.1-1.6)	1.2 (1.1-1.3)	1.2 (1.1-1.3)
Plt × LLN*	1.6 (1.3-1.9)	1.8 (1.5-2.2)	1.8 (1.5-2.2)
Na × LLN	1.0 (1.0-1.0)	1.0 (1.0-1.1)	1.0 (1.0-1.0)
Cr × ULN	0.9 (0.8-1.1)	0.7 (0.6-0.8)	0.7 (0.6-0.8)
IgG × ULN	0.9 (0.8-1.1)	0.9 (0.7-1.1)	0.9 (0.7-1.1)
Laboratory values at 12 months following UDCA initiation, median (IQR)			
ALP × ULN*	1.3 (0.9-2.0)	1.2 (0.9-2.1)	1.3 (0.9-2.1)
ALT × ULN†	1.0 (0.6-1.6)	0.8 (0.6-1.3)	0.8 (0.6-1.3)
AST × ULN†	1.4 (0.9-2.5)	0.8 (0.6-1.3)	0.8 (0.6-1.3)
TB × ULN*	0.5 (0.4-0.8)	0.5 (0.4-0.7)	0.5 (0.4-0.7)

*Laboratory values are imputed to increase the number of usable samples for testing the model; †ALT and AST were not imputed but were imputed as the composite ALT/AST variable used in the model. The median (IQR) values reported here are not imputed.
Abbreviations: ALB, albumin; ALP, alkaline phosphatase; ALT, alanine aminotransferase; ANA, anti-nuclear antibody; AST, aspartate aminotransferase; Cr, creatinine; IgG, immunoglobulin G; LLN, lower limit of normal; Plt, platelets; Na, sodium; OLT, orthotopic liver transplantation; SMA, smooth muscle actin; TB, total bilirubin; ULN, upper limit of normal.

included in the analysis. Patients who did not reach an event were censored at the date of their most recent blood tests or date of non-liver related death. Patients for whom cause of death was unclear or unknown were assumed to have died from non-PBC-related causes in the initial analysis. To determine if this assumption would influence the discrimination or calibration of the UK-PBC risk scores, a subsequent sensitivity analysis was performed that assumed that these patients died a liver-related death.

STATISTICAL ANALYSIS

Descriptive statistics summarizing data are reported as medians, interquartile ranges (IQR), and proportions. Statistical comparisons of continuous and categorical variables were made using the Wilcoxon rank sum and chi-square tests, respectively. Kaplan-Meier curves were used to estimate the probability of being event-free over time and the event rates at specific points in time. Consistent with the original UK-PBC risk scores, laboratory values at least 12 months after UDCA treatment were used to calculate the risk scores. Multiple imputation was used to estimate any required missing values; imputation was run 5 times by chained equations available in the R package "mice," and results were reported using an average from the five data sets.[8]

Discrimination, the ability of a risk score to accurately rank subjects from low to high risk, was assessed using the concordance statistic (c-statistic). For a binary outcome, the c-statistic is equivalent to the area under the receiver operating characteristic curve. This method can be extended for use with the Cox model and was used in this analysis.[9] Calibration, the ability to accurately predict the absolute risk level, was assessed by comparing the observed and predicted number of events based on the person-years of observation. We computed 95% confidence intervals (CIs) assuming the observed number of events follow a Poisson distribution. To assess the consistency of calibration across the predicted risk distribution,

TABLE 2. DISTRIBUTION OF THE EVENTS USED IN THE COMPOSITE ENDPOINT

	MCPGE Cohort (n = 464)
Any first liver-related event, no.*	46
Death, no.	
Liver related	2 (0.4)
Not liver related†	5 (1.1)
Cause unknown‡	19 (4.1)
Liver transplant	20 (4.3)
Bilirubin ≥100 μmol/L	24 (5.2)
Event rates, no. (%)	
5 years	25 (6.0)
10 years	35 (8.9)
15 years	46 (15.2)

*Patients censored after first liver-related event; †these were not included in any analysis but are included here for reference only; ‡these were included in the sensitivity analysis where unknown cause of death was assumed to be due to PBC.

individual UK-PBC risk scores were calculated and then divided into quartiles. The observed median risk was then calculated for each group. The predicted risk was plotted against the observed risk for each group of patients to allow visual assessment of agreement, while Poisson regression was used to statistically assess agreement.[10] We also assessed the calibration of Barcelona, Paris-I, Rotterdam, Toronto, Paris-II, and GLOBE scores. Statistical significance was defined as a two-tailed value of $P < 0.05$, and the analysis was performed using R (version 3.3.1; R Core Team, Vienna, Austria).

Results

Medical records of 1,003 patients with PBC were reviewed, and 539 patients were excluded, primarily due to lack of laboratory values at specified time points (Supporting Fig. S1). Compared to those who were included, those excluded were more likely to need transplantation and had higher bilirubin at the time of diagnosis (Supporting Table S1). A total of 464 patients were included in the study. Characteristics for these patients along with the characteristics of the derivation and validation cohorts from the original UK-PBC risk scores study are provided in Table 1. Missing information is summarized in Supporting Table S2; 113 subjects had at least 1 laboratory value imputed for use in calculating the UK-PBC risk scores. Median follow-up was 11.2 years after diagnosis (IQR, 6.1-17.8 years) and 9.7 years after 1 year of UDCA treatment (IQR, 4.2-14.6). Forty-six patients (9.9%) in our cohort experienced an event by the end of follow-up (Table 2): 2 (0.4%) patients died a liver-related death, 20 (4.3%) patients received a liver transplant, and 24 (5.2%) patients had a rise in bilirubin >5.8 mg/dL (18 of whom subsequently received a transplant and 4 died during follow-up). The overall event-free survival rate in our cohort was 94.0% at 5 years, 91.1% at 10 years, and 84.8% at 15 years. By comparison, the event-free

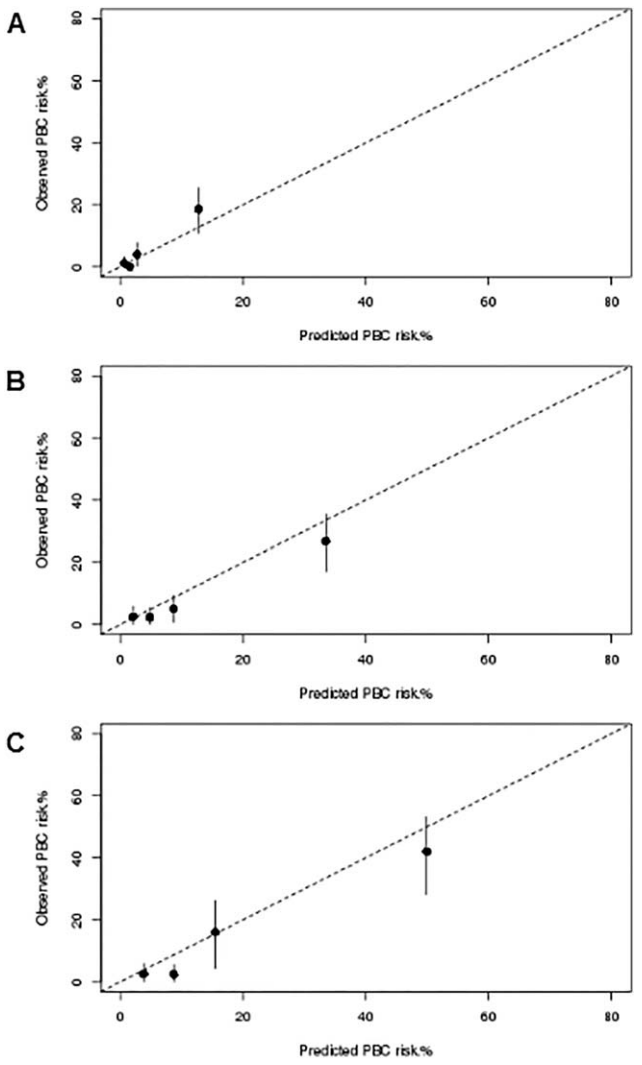

FIG. 1. Calibration plot comparing predicted and observed PBC risk. Data points represent MCPGE cohort patients grouped into four groups (quartiles) of predicted risk (obtained from the UK-PBC risk score for 5, 10, and 15 years). The observed 5-, 10-, and 15-year risk of events for each group of patients and confidence intervals were estimated from a Poisson regression model that included indicator covariates for each group. The risk is shown at (A) 5, (B) 10, and (C) 15 years. Note: The dots represent the mean predicted risk of the 4 groups vs the mean observed risk of the 4 groups. The bars represent a 95% confidence interval of the observed risk of the 4 groups.

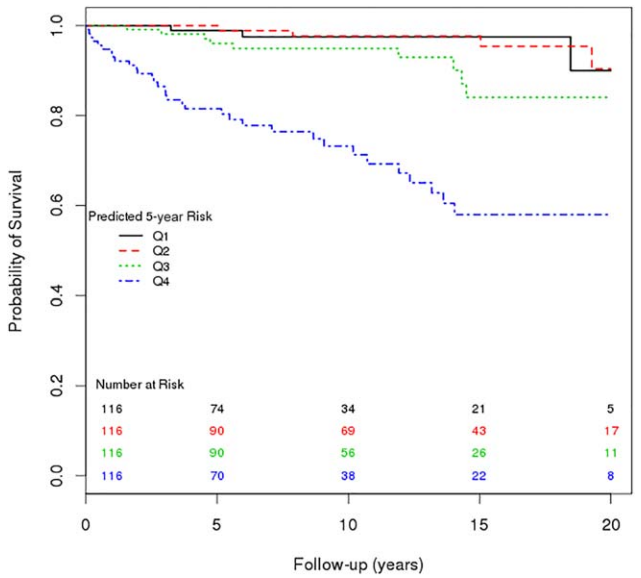

FIG. 2. Kaplan-Meier curve using all available follow-up after at least 12 months of UDCA treatment, with separate lines for MCPGE cohort patients grouped into four groups of predicted 5-year risk obtained from the UK-PBC risk score. Abbreviation: Q, quartile.

survival reported for the overall UK-PBC cohort was 96%, 89%, and 86% at 5, 10, and 15 years, respectively.

Discrimination of the UK-PBC risk scores in the MCPGE cohort was evaluated using the c-statistic and was 0.88 (95% CI, 0.76-0.99) for the 5-year risk score, 0.85 (95% CI, 0.75-0.94) for the 10-year risk score, and 0.84 (95% CI, 0.75-0.93) for the 15-year risk score. The ratio of observed to predicted number of events in the MCPGE cohort was 1.5 (25/16.8; $P = 0.046$), 0.8 (35/44.9; $P = 0.14$), and 0.7 (46/67.5; $P = 0.009$) for the 5-, 10-, and 15-year risk scores, respectively (Fig. 1A-C). To determine whether the UK-PBC risk scores have good performance after long-term follow-up, we used the 15-year UK-PBC risk score and limited patients to those with >10 years of follow-up; following this analysis, we observed 11 events and predicted 11.75 events ($P = 0.83$).

We then performed a sensitivity analysis where the model assumed that those with unknown cause of death died of liver-related causes. The resulting c-statistics were similar to the main model: 5 years 0.86 (0.76-0.97), 10 years 0.82 (0.73-0.90), and 15 years 0.81 (0.73-0.86). The ratio of observed to predicted number of events in the MCPGE cohort was 1.7 (29/17; $P < 0.001$), 1 (47/45; $P = 0.757$), and 1 (65/68; $P = 0.761$) for the 5-, 10-, and 15-year risk scores, respectively, indicating that the calibration was no worse than with the main version of the endpoint with the exception of the 15-year risk score.

The remainder of the post hoc analysis evaluating the highest risk quartile was thus performed on the main model (e.g., the model with the assumption that unknown deaths were not due to PBC). We sought to identify factors associated with liver-related events. This demonstrated that the quartile of patients with the highest 5-year UK-PBC risk scores had significantly worse event-free survival than the remainder of the cohort ($P < 0.001$) at 5 (82% versus 98%), 10 (73% versus 97%), and 15 years (58% versus 93%) (Fig. 2). Results were similar when the 10-year or 15-year UK-PBC risk scores were used. These patients were characterized by severe cholestasis, hepatitis, and portal hypertension (Supporting Table S3). Boxplots of the median scores and ranges of the UK-PBC risk scores for each quartile can be found in Fig. 3, which demonstrates that the patients in the highest risk quartile had a wide range of values ranging from 4%-68% (median 9%) for the 5-year score, 14%-98% (median

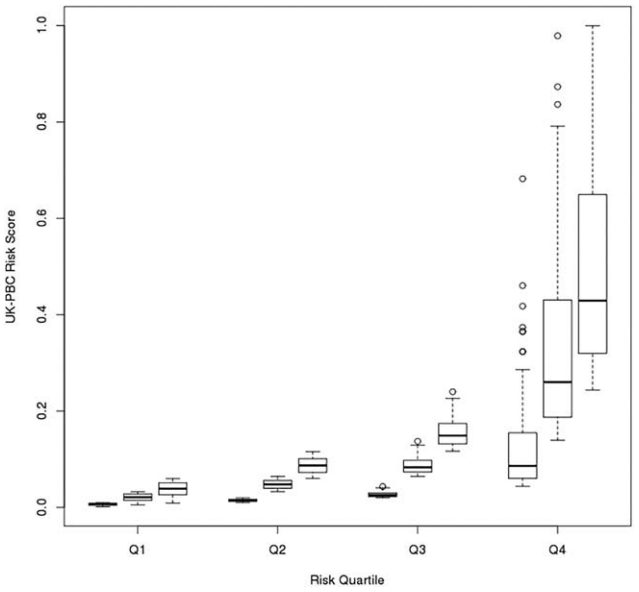

FIG. 3. Boxplots of predicted scores for MCPGE cohort patients. The boxplots of predicted 5-, 10-, and 15-year risk obtained from the UK-PBC risk scores and stratified by risk quartiles. For instance, the three boxes above Q4 indicate the fourth quartile for the 5-, 10-, and 15-year risk. Abbreviation: Q, quartile. Note: Horizontal lines within the boxplots represent the median UK-PBC Risk Scores, while the outer edges of each box represent the first and third quartiles. The ends of each dotted line represent values 1.5 times the IQR; the empty dots represent outliers.

26%) for the 10-year score, and 24%-100% (median 43%) for the 15-year score. The variability of values within each of the other quartiles was much smaller and was generally low. Lastly, the concordance of other known prognostic scores for PBC was lower than that of the UK-PBC risk scores (and the GLOBE score) (Supporting Table S4).

Discussion

One of the main challenges in the application of prognostic scores is their ability to perform adequately in patient populations outside which they were initially derived and validated. We have herein demonstrated that the UK-PBC risk scores have excellent short- and long-term performance in patients assessed at a North American tertiary medical center. Patients with the highest quartile in 5-year UK-PBC risk scores had an event-free survival of only 82%, 73%, and 58% at 5, 10, and 15 years compared to those with lower scores whose event-free survival ranged from 93%-98%. Further, the UK-PBC risk scores had excellent performance even for patients whose survival exceeded 10 years, demonstrating that it has both good short-term and long-term applicability.

The excellent discrimination of the UK-PBC risk scores indicates that subjects were well ranked according to their risk of a future event. Calibration (i.e., observed versus predicted events) demonstrated that events occurred early than predicted in the MCPGE cohort compared to the UK-PBC cohort, suggesting differences between the cohorts that skew the overall calibration of the PBC-risk scores. This likely reflects the fact that the MCPGE cohort has a slightly different case mix than the United Kingdom cohort, as demonstrated by the characteristics at the time of UDCA commencement as well as characteristics 12 months after UDCA treatment. The MCPGE cohort consisted of patients with a greater degree of cholestasis, hepatitis, portal hypertension, and worse renal function compared to the UK-PBC cohort.

Despite these differences, it should be noted that the MCPGE cohort included in this study consisted of a significant proportion of subjects with early PBC, owing to the fact that the cohort was developed for the purposes of genetic epidemiology studies in PBC.[6] This allowed for a case mix that included patients with early and advanced disease that might otherwise have included a disproportionate number of patients with advanced disease, given that Mayo Clinic is a center that specializes both in transplant and PBC.

Where there were missing values, this study used multiple imputation, which is an accepted approach to handle missing data. While a higher proportion of patients who were excluded had a liver transplant or significant hyperbilirubinemia, this is unsurprising given that the basis for exclusion for over 10% of these patients was the occurrence of an event within a year of starting UDCA.

Interestingly, despite including hyperbilirubinemia as an outcome, the derivation paper of the UK-PBC risk scores reported no patients with this event. By contrast, approximately 5% of the MCPGE cohort reached the threshold of hyperbilirubinemia prior to death or transplant. The majority of these patients (92%) received a liver transplant or died during follow-up. This may reflect differences in liver transplant eligibility where patients are given transplants prior to reaching the bilirubin threshold defined by the UK-PBC group. Indeed, since approximately 2007, the United Kingdom has used the United Kingdom Model for End-Stage Liver Disease rather than the Model for End-Stage Liver Disease (MELD), which is used in North America (now replaced by the Na-MELD).[11] While these scores use the same laboratory values and thus appear quite similar, each score was based on different risk models and was generated from different patient populations. Importantly, despite potential underlying differences between the Na-MELD and the United Kingdom Model for End-Stage Liver Disease, the UK-PBC risk scores maintain excellent performance.

In conclusion, this study demonstrates that the UK-PBC risk scores performed well in a cohort assessed at a tertiary medical center in North America, although there was a quartile of patients with earlier than expected events due to advanced disease.

REFERENCES

1) Nguyen DL, Juran BD, Lazaridis KN. Primary biliary cirrhosis. Best Pract Res Clin Gastroenterol 2010;24:647-654.
2) Pares A, Rodes J. Natural history of primary biliary cirrhosis. Clin Liver Dis 2003;7:779-794.
3) Pares A. Treatment of primary biliary cirrhosis: is there more to offer than ursodeoxycholic acid? Clinical Liver Disease 2014;3: 29-33.
4) Carbone M, Sharp SJ, Flack S, Paximadas D, Spiess K, Adgey C, et al.; UK-PBC Consortium. The UK-PBC risk scores: Derivation and validation of a scoring system for long-term prediction

of end-stage liver disease in primary biliary cholangitis. Hepatology 2016;63:930-950.
5) Lammers WJ, Hirschfield GM, Corpechot C, Nevens F, Lindor KD, Janssen HL, et al.; Global PBC Study Group. Development and validation of a scoring system to predict outcomes of patients with primary biliary cirrhosis receiving ursodeoxycholic acid therapy. Gastroenterology 2015;149:1804-1812. e1804.
6) Lazaridis KN, Juran BD, Boe GM, Slusser JP, de Andrade M, Homburger HA, et al. Increased prevalence of antimitochondrial antibodies in first-degree relatives of patients with primary biliary cirrhosis. Hepatology 2007;46:785-792.
7) Lindor KD, Gershwin ME, Poupon R, Kaplan M, Bergasa NV, Heathcote EJ; American Association for Study of Liver Diseases. Primary biliary cirrhosis. Hepatology 2009;50:291-308.
8) van Buuren S, Groothuis-Oudshoorn K. mice: multivariate imputation by chained equations in R. J Stat Software 2011;45:1-67.
9) Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361-387.
10) Crowson CS, Atkinson EJ, Therneau TM. Assessing calibration of prognostic risk scores. Stat Methods Med Res 2016;25:1692-1706.
11) Neuberger J, Gimson A, Davies M, Akyol M, O'Grady J, Burroughs A, et al.; Liver Advisory Group; UK Blood and Transplant. Selection of patients for liver transplantation and allocation of donated livers in the UK. Gut 2008;57:252-257.

Supporting Information

Additional Supporting Information may be found at onlinelibrary.wiley.com/doi/10.1002/hep4.1186/full.