9-9-2021

Dynamic Risk Prediction of Response to Ursodeoxycholic Acid Among Patients with Primary Biliary Cholangitis in the USA

Jia Li

Mei Lu

Yueren Zhou

Christopher L. Bowlus

Keith Lindor

See next page for additional authors

Authors

Jia Li, Mei Lu, Yueren Zhou, Christopher L. Bowlus, Keith Lindor, Carla Rodriguez-Watson, Robert J. Romanelli, Irina V. Haller, Heather Anderson, Jeffrey J. VanWormer, Joseph A. Boscarino, Mark A. Schmidt, Yihe G. Daida, Amandeep Sahota, Jennifer Vincent, Kuan-Han Hank Wu, Sheri Trudeau, Loralee B. Rupp, Christina Melkonian, and Stuart C. Gordon

Digestive Diseases and Sciences
https://doi.org/10.1007/s10620-021-07219-4

ORIGINAL ARTICLE

Dynamic Risk Prediction of Response to Ursodeoxycholic Acid Among Patients with Primary Biliary Cholangitis in the USA

Jia Li[2] · Mei Lu[2] · Yueren Zhou[2] · Christopher L. Bowlus[3] · Keith Lindor[4] · Carla Rodriguez-Watson[5,6] · Robert J. Romanelli[7] · Irina V. Haller[8] · Heather Anderson[9] · Jeffrey J. VanWormer[10] · Joseph A. Boscarino[11] · Mark A. Schmidt[12] · Yihe G. Daida[13] · Amandeep Sahota[14] · Jennifer Vincent[15] · Kuan-Han Hank Wu[2] · Sheri Trudeau[2] · Loralee B. Rupp[16] · Christina Melkonian[2] · Stuart C. Gordon[1] · For the FOLD Investigators

Received: 19 February 2021 / Accepted: 5 August 2021

Abstract

Background Ursodeoxycholic acid (UDCA) remains the first-line therapy for primary biliary cholangitis (PBC); however, inadequate treatment response (ITR) is common. The UK-PBC Consortium developed the modified UDCA Response Score (m-URS) to predict ITR (using alkaline phosphatase [ALP] > 1.67 times the upper limit of normal [*ULN]) at 12 months post-UDCA initiation). Using data from the US-based Fibrotic Liver Disease Consortium, we assessed the m-URS in our multi-racial cohort. We then used a dynamic modeling approach to improve prediction accuracy.

Methods Using data collected at the time of UDCA initiation, we assessed the m-URS using the original formula; then, by calibrating coefficients to our data, we also assessed whether it remained accurate when using Paris II criteria for ITR. Next, we developed and validated a dynamic risk prediction model that included post-UDCA initiation laboratory data.

Results Among 1578 patients (13% men; 8% African American, 9% Asian American/American Indian/Pacific Islander; 25% Hispanic), the rate of ITR was 27% using ALP > 1.67*ULN and 45% using Paris II criteria. M-URS accuracy was "very good" (AUROC = 0.87, sensitivity = 0.62, and specificity = 0.82) for ALP > 1.67*ULN and "moderate" (AUROC = 0.74, sensitivity = 0.57, and specificity = 0.70) for Paris II. Our dynamic model significantly improved accuracy for both definitions of ITR (ALP > 1.67*ULN: AUROC = 0.91; Paris II: AUROC = 0.81); specificity approached 100%. Roughly 9% of patients in our cohort were at the highest risk of ITR.

Conclusions Early identification of patients who will not respond to UDCA treatment using a dynamic prediction model based on longitudinal, repeated risk factor measurements may facilitate earlier introduction of adjuvant treatment.

Keywords Primary biliary cirrhosis · Alkaline phosphatase · Paris II

Introduction

Ursodeoxycholic acid (UDCA) is the primary first-line treatment for primary biliary cholangitis (PBC), but an estimated 4 in 10 PBC patients fail to respond adequately to therapy[1]. Inadequate treatment response (ITR)—generally defined as a lack of improvement in liver laboratory parameters at 12 months after treatment initiation[2]—leaves patients at risk of progressive liver disease, including decompensated cirrhosis and liver transplant. However, use of this 12-month timeline means that patients who will ultimately require second-line therapies must wait at least a year prior to their initiation. Early and more robust identification of patients unlikely to benefit from UDCA would allow clinicians to target introduction of adjuvant or alternative therapies earlier in the disease process, when successful treatment might mitigate liver disease progression.

Several studies have identified factors that predict likelihood of ITR, including younger age at diagnosis, longer time between diagnosis and treatment initiation, as well as laboratory parameters such as alkaline phosphatase (ALP)[3,4]. A recent predictive model—the UDCA Response Score (URS)—developed by the UK-PBC Consortium demonstrated "good" accuracy (area under the receiver operator

Dr. Li accepts full responsibility for the conduct of the study.

✉ Jia Li
jli4@hfhs.org

Extended author information available on the last page of the article

Published online: 09 September 2021

characteristic curve [AUROC] = 0.83) when ITR was defined as ALP > 1.67 times the upper limit of normal (*ULN) at 12 months after UDCA initiation.[3] However, this cohort was limited to PBC patients with completed laboratory assessments at two time points (time of PBC diagnosis and time of UDCA initiation), which is not always feasible in "real world" patients. Accuracy of a version of this model adapted to use variables collected at a single time point—the modified URS (m-URS)—was slightly reduced (AUROC = 0.81 95% CI 0.77–0.84). Another group developed a model predicting UDCA response, using a cohort of PBC patients from Beijing, China; model accuracy was similar to the UK-PBC model (AUROC = 0.80, 95% CI: 0.68–0.89).[4]

The Fibrotic Liver Disease (FOLD) Consortium recently showed that, in contrast to the perception of PBC as a disease of White women in late middle age, significant proportions of this US-based cohort were Hispanic, African American, or Asian American/American Indian/Pacific Islanders (ASINPI). Given racial differences in PBC prevalence and benefits of UDCA treatment[5–7], it is not clear that the above-mentioned European and Chinese results are generalizable to a US cohort. Likewise, previous studies have documented geographic differences in PBC prevalence, suggesting possible differences in exposure to some yet-undefined risk factor(s) that could influence response to UDCA[8]. Therefore, we had three goals for this analysis: First, we sought to validate the m-URS[3] among a racially and geographically diverse US cohort. Second, we also evaluated whether the m-URS could be improved with additional calibration (adjustment of coefficients). Finally, we used machine-learning methods to explore a new predictive model (optimized based on accuracy, parsimony, and duration from treatment initiation) that would increase predictive ability and accurately classify patients at the highest risk of ITR. This resulted in the development of a dynamic ITR risk prediction model using longitudinal measurements of post-baseline biomarkers that included of laboratory measures in the first months after UDCA treatment initiation.

Methods

The FOLD Consortium has been previously described.[5, 6] Briefly, FOLD comprises 11 geographically diverse health systems, representing four US Census Bureau-defined regions (Northeast, Midwest, Northwest, and South). FOLD follows the guidelines of the US Department of Health and Human Services for the protection of human subjects. The study protocol was approved by the Institutional Review Board of each participating site. The requirement for written informed consent was waived due to the observational and de-identified nature of the data. All authors had access to the study results, and reviewed and approved the final manuscript.

Patient Cohort

FOLD PBC patient identification methods have been previously described.[6] Our PBC cohort was identified using an automated Classification and Regression Tree (CART) algorithm. Patients were classified as having PBC if they met one of three conditions: 1) AMA-positive with at least one PBC diagnosis code; 2) AMA-positive with ALP > 150 (in the absence of a PBC diagnosis code); or 3) presence of two PBC diagnosis codes (in the absence of an AMA-positive result). Classification accuracy was excellent: AUROC = 93.4%, sensitivity 93.4%, specificity 86.5%, PPV 79.2%, and NPV 96.8%. Application of the CART algorithm to the preliminary patient pool identified 4,241 "true" PBC cases. Of these, 1645 had an AMA-positive test and one or more PBC diagnosis codes; 832 had an AMA-positive test and ALP > 150; and 1764 patients presented with two or more PBC diagnosis codes. These categories were mutually exclusive. All cases were confirmed with chart abstraction performed by trained medical abstractors; FOLD hepatologists provided adjudication of indeterminate cases.

PBC patients that initiated UDCA treatment between January 1, 2006 and December 31, 2016 and had at least 12 months of follow-up data after UDCA initiation were included in this analysis. We excluded patients with a history of UDCA treatment prior to 2006 and those without at least one of three laboratory results for calculation of Paris II criteria[2] (ALP, aspartate aminotransferase [AST], and total bilirubin) 12 months after treatment initiation. For validation of the m-URS, data were randomly divided into a training dataset and a validation dataset using a 2:1 ratio.

Covariates

"Index date" was defined as the date of UDCA treatment initiation. Variables collected at index date included patient demographics (gender, race/ethnicity, age) and the following laboratory parameters: bilirubin; ALP in relation to the upper limit of normal (ULN) defined by the assay used at each site; albumin (again in relation to the site-defined "normal"); ratio of AST to alanine aminotransferase (ALT) [9]; AST to Platelet Ratio Index (APRI); and Fibrosis 4 Index (FIB4; a biomarker comprising age, ALT, AST, and platelet counts). For the dynamic parameters, ALP, AST, ALT, bilirubin, and albumin were summarized using a median smoother in monthly intervals, up to 7 months from the index date.

Outcomes

We assessed the accuracy of prediction models using two established criteria: 1) Toronto criteria (ALP > 1.67*ULN at 12 months post-UDCA initiation), as used by the modified UDCA Response Score (m-URS); and 2) Paris II criteria[2] (any of the following: total bilirubin > 1 mg/dL; AST ≥ 1.5*ULN; or ALP ≥ 1.5*ULN, based on the first laboratory results ≥ 12 months after UDCA treatment initiation).

Statistical Analysis

In order to validate and recalibrate the m-URS, data were randomly divided into a training dataset and a validation dataset using a 2:1 ratio. The m-URS includes four variables measured at the time of UDCA treatment initiation (age; $1/\sqrt{\text{total bilirubin}}$; $_{\log}$ALT; and $_{\log}$ALP), calculated with a logit model; ITR is defined as ALP > 1.67*ULN at 12 months post-treatment initiation. We validated the model several ways: 1) by applying the m-URS variables/coefficients and main outcome (ALP > 1.67*ULN) directly to our validation data set; 2) by calibrating the coefficients of m-URS variables based on our training data using logistic regression, and testing these refined coefficients using the validation data. Model accuracy was assessed for ITR defined by both ALP > 1.67*ULN or Paris II criteria. Multiple imputation was used to address missing laboratory data at baseline.

Next, we sought to determine whether the inclusion of dynamic longitudinal laboratory measures obtained after treatment initiation improved prediction of ITR by performed a two-step analysis based on information derived from each patient's individual trajectory of laboratory markers. First, a linear regression analysis was performed with "time from treatment initiation" as the independent variable and repeated measurements of ALP as the dependent variable. This analysis was performed separately for each patient, using data from baseline up to 7 months post-index, in order to obtain subject-specific growth curves over this time interval. In addition to a linear growth curve, we also applied quadratic and piecewise functions; the final function for the selected growth curve was chosen based on goodness-of-fit. Similar strategies were used with other laboratory markers (AST, ALT, total bilirubin). Next, each patients' baseline characteristics and their patient-specific growth parameters (i.e., slope coefficient[s]) were entered into a logistic model to predict ITR at 12 months after treatment initiation, generating an individual "risk score." Dynamic risk prediction models were developed separately for the ALP > 1.67*ULN and Paris II criteria for ITR. Training data were used to build this dynamic risk prediction model.

Model selection was based on balancing model predictive ability and complexity. Model parameters such as number of covariates and number of time intervals to calculate growth parameters were evaluated by tenfold cross-validation on the training data. Predictive ability of the models was further assessed at baseline and each month post-treatment initiation on the validation data. Model prediction accuracy for both m-URS and our dynamic model was assessed by receiver operating characteristic curves (ROC) and the precision-recall curve (PRC) as an alternative metric, which is preferable when outcomes are imbalanced. Confidence intervals and comparisons between models for area under ROC (AUROC) and area under PRC (AUPRC) were estimated by 1,000 bootstraps. Other metrics such as sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were used to calculate a cutoff value for the risk score. The cutoff point for this score was derived by optimizing the PPV. This model was developed using imputed data.

We further conducted two sensitivity analyses. First, the model was tested on a subset of the validation data that excluded patients with missing baseline laboratory measurements. Second, the model was tested on the subset of patients in validation data who had data from at least one set of post-treatment laboratory measures.

Results

Table 1 displays the twelve baseline variables considered in the univariate analysis, including imputed values, by UDCA response status. We identified 1578 patients who initiated UDCA treatment during the study period: 8% were African American; 9% were ASINPI; 67% were White; and 25% were Hispanic of any race. Thirteen percent were men. A total of 419 (27%) demonstrated ITR at 12 months from index using ALP > 1.67*ULN, whereas 706 (45%) had ITR using Paris II criteria. Values were missing for 6–23% of baseline covariates; multiple imputation was used to estimate these missing values.

Validation of the Modified UDCA Response Score (m-URS)

Among a total of 1578 patients (1052 in the training data and 526 in the validation data), baseline measurements were available for 1270 patients. After imputation to address missing data, the AUROC of the m-URS based on baseline covariates was 0.89 (95% CI 0.86–0.92) and the AUPRC was 0.69 (95% CI 0.60–0.79); Table 2) in the validation dataset. Using the m-URS variables but allowing the coefficients to vary yielded similar predictive ability (see formula in Supplemental Table 1), the AUROC for the model for predicting ITR defined as ALP > 1.67*ULN was 0.89 (95% CI 0.86–0.92); the AUPRC was 0.69 (95% CI 0.61–0.78). Using Paris II criteria as the outcome of interest, application of the m-URS yielded an AUROC of 0.76 (95% CI 0.72–0.80),

Table 1 Univariate analysis of baseline variables for the full sample (A), for the full sample with unknown values imputed (B), and by response to UDCA treatment at 12 months after initiation of therapy (C and D) defined by alkaline phosphatase (< 1.67 times the upper limit of normal) at 12 months post-treatment initiation

Variable	Response	A. All (N = 1578)	B. All with imputation (N = 1578)	C. Responder (N = 872)	D. Non-responder (N = 706)	*P* value
US Census Region	Midwest	308 (20%)	308 (20%)	163 (19%)	145 (21%)	0.006
	Northeast	99 (6%)	99 (6%)	69 (8%)	30 (4%)	
	South	157 (10%)	157 (10%)	96 (11%)	61 (9%)	
	West	1014 (64%)	1014 (64%)	544 (62%)	470 (67%)	
Gender	Women	1371 (87%)	1371 (87%)	768 (88%)	603 (85%)	0.119
	Men	207 (13%)	207 (13%)	104 (12%)	103 (15%)	
Race	AS/IN/PI	137 (9%)	151 (10%)	88 (10%)	63 (9%)	0.382
	Black	126 (8%)	139 (9%)	70 (8%)	69 (10%)	
	White	1057 (67%)	1288 (82%)	714 (82%)	574 (81%)	
	Missing	258 (16%)				
Hispanic	Yes	392 (25%)	412 (26%)	211 (24%)	201 (28%)	0.055
	No	1088 (69%)	1166 (74%)	661 (76%)	505 (72%)	
	Missing	98 (6%)				
Age	≤40	123 (8%)	123 (8%)	59 (7%)	64 (9%)	0.047
	41−50	283 (18%)	283 (18%)	143 (16%)	140 (20%)	
	51−60	496 (31%)	496 (31%)	280 (32%)	216 (31%)	
	61−70	428 (27%)	428 (27%)	237 (27%)	191 (27%)	
	> 70	248 (16%)	248 (16%)	153 (18%)	95 (13%)	
Albumin	< LLN	328 (21%)	385 (24%)	150 (17%)	235 (33%)	< .001
	Normal	887 (56%)	1193 (76%)	722 (83%)	471 (67%)	
	Missing	363 (23%)				
Alkaline phosphatase	< ULN	280 (18%)	424 (27%)	119 (17%)	305 (35%)	< .001
	[1,1.5)*ULN	345 (22%)	345 (22%)	98 (14%)	247 (28%)	
	[1.5,2)*ULN	248 (16%)	248 (16%)	101 (14%)	147 (17%)	
	[2,3)*ULN	263 (17%)	263 (17%)	154 (22%)	109 (13%)	
	≥3*ULN	298 (19%)	298 (19%)	234 (33%)	64 (7%)	
	Missing	144 (9%)				
Bilirubin (mg/dL)	> 2.0	116 (7%)	127 (8%)	25 (3%)	102 (14%)	< .001
ALT	2.0 > 1.5	64 (4%)	103 (7%)	37 (4%)	66 (9%)	
	1.5 > 1.0	131 (8%)	144 (9%)	54 (6%)	90 (13%)	
	1.0 > 0.7	252 (16%)	317 (20%)	152 (17%)	165 (23%)	
	0.7 > 0.5	254 (16%)	283 (18%)	163 (19%)	120 (17%)	
	0.5 > 0.4	163 (10%)	192 (12%)	133 (15%)	59 (8%)	
	≤0.4	349 (22%)	412 (26%)	308 (35%)	104 (15%)	
	Missing	249 (16%)				
Platelets	Normal	1084 (69%)	1240 (79%)	743 (85%)	497 (70%)	< .001
	< LLN	311 (20%)	338 (21%)	129 (15%)	209 (30%)	
	Missing	183 (12%)				
AST/ALT ≥1.1	Yes	451 (29%)	464 (29%)	217 (25%)	247 (35%)	< .001
	No	919 (58%)	1114 (71%)	655 (75%)	459 (65%)	
	Missing	208 (13%)				
APRI Score	N	1213	1578	872	706	0.022
	Mean (SD)	1.8 (4.93)	1.7 (4.99)	1.5 (5.66)	2.0 (4.00)	
	Median	0.73	0.77	0.57	1.09	
FIB4 Index	N	1207	1578	872	706	0.029
	Mean (SD)	10.7 (36.92)	10.5 (37.11)	8.5 (27.49)	12.9 (46.22)	
	Median	1.95	2.06	1.75	2.62	

AS/IN/PI: Asian American, American Indian, Pacific Islander; ULN: upper limit of normal, as defined by the assay used at each site; LLN: lower limit of normal, as defined by the assay used at each site; AST/ALT: ratio of aspartate aminotransferase to alanine aminotransferase (≥ 1.1 indicates cirrhosis); APRI: Aspartate aminotransferase to Platelet Ratio Index; SD: standard deviation; FIB4: Fibrosis 4; UDCA: ursodeoxycholic acid

Table 2 Assessment of the performance of the modified UDCA Response Score (m-URS) to predict inadequate treatment response (ITR) at 12 months post-treatment initiation in patients from the Fibrotic Liver Disease (FOLD) cohort. Missing baseline data were imputed by multiple imputation

ITR Criteria	Model	AUROC (95% CI)	AUPRC (95% CI)
ALP > 1.67*ULN	Original (full cohort)	0.87 (0.85,0.89)	0.71 (0.66,0.76)
	Original (validation)	0.89 (0.86, 0.92)	0.69 (0.60, 0.79)
	Recalibrated (validation)	0.89 (0.86, 0.92)	0.69 (0.61, 0.78)
Paris II	Original (full cohort)	0.77 (0.74, 0.79)	0.74 (0.70, 0.77)
	Original (validation)	0.76 (0.72,0.80)	0.74 (0.68, 0.79)
	Recalibrated (validation)	0.79 (0.76, 0.83)	0.75 (0.70,0.82)

ITR Inadequate treatment response, *AUROC* Area under the receiver operating curve, *AUPRC* Area under the precision-recall curve

and an AUPRC of 0.74 (95% CI 0.68–0.79). When the coefficients of m-URS were recalibrated, AUROC was 0.79 (95% CI 0.76–0.83), and AUPRC was 0.75 (95% CI 0.70–0.82).

We also performed a separate sensitivity analysis using m-URS variables/coefficients, but including an "unknown" category for covariates with missing (see Table 1 Column A). This demonstrated lower predictive ability (AUROC 0.72, 95% CI 0.68–0.77 using validation data).

Improvement of ITR Prediction Using Dynamic Risk Prediction

During model development, we observed a biphasic evolution of laboratory markers over time after treatment initiation. To address this, a piecewise linear spline model was fitted with "time from treatment" as the predictor and ALP as the outcome. This method was used to estimate changes in the longitudinal pattern of each laboratory marker, with one slope representing the change ≤ 2 months post-treatment initiation and the second slope presenting the change > 2 months of treatment initiation. The final logistic model included gender, age, and baseline measurements of ALP, AST, ALT, and bilirubin, and their growth parameters after treatment initiation. The formula for the dynamic predictive score of ITR for each patient is provided in Supplemental Table 1S.

In order to validate the dynamic prediction model at each follow-up time point, the growth curve estimates were updated monthly if additional ALP data became available; if no additional measurement of ALP was available, the slope derived from the previous measure was used. Similar strategies were used with other laboratory markers. Model predictive ability (Fig. 1a) was consistently better than the recalibrated m-URS over time; there were significantly improved (p value < 0.05) starting at month 4 (AUROC 0.91, 95% CI 0.88–0.94) compared to m-URS (0.89, 95% CI 0.86–0.92). Similarly, AUPRC of the dynamic model was significantly better than the m-URS beginning at month 4 (0.74, 95% CI 0.64–0.82, versus 0.69, 95% CI 0.60–0.78, respectively). Predictive accuracy of the dynamic model improved over time as more measurements became available. At month 7, AUROC was 0.91 (95% CI 0.89–0.94) and AUPRC was 0.77 (95% CI 0.68–0.85). Both were significantly higher than the m-URS (p values < 0.01). Figure 2a shows additional metrics at each time point based on the cut-points optimized by PPV. Patients with risk scores > 0.84 were considered to be at the highest risk of ITR. Using this cut-point, the model showed that roughly 9% of patients in our cohort were in the highest risk group for ITR. At baseline, the model classified 74 of the 525 available patients (14%) into the high-risk group, of whom 53 of these patients ultimately had ITR (ALP > 1.67*ULN), for a PPV of 71.6%. At 7 months post-treatment initiation, the model classified 52 patients (9.9%) as high risk, of whom 44 had ITR (ALP > 1.67*ULN); the PPV was significantly improved to 84.6%. Specificity was improved from 95% at baseline to 98% at month 4 and was consistent afterward; NPV remained stable at 83%.

For the first sensitivity analysis, we excluded patients in the validation cohort with missing baseline data (n = 414). At month 4, the AUROC of our dynamic risk prediction model became significantly higher than m-URS (Supplemental Fig. 1; 0.90; 95% CI: 0.87–0.94; $p = 0.034$); the AUPRC became significantly higher than the m-URS at month 3 (0.74, 95% CI 0.65–0.84; $p = 0.05$). Results were also consistent when we excluded patients without post-treatment laboratory data (n = 347; Supplemental Fig. 2).

Figures 3a and b illustrate how our dynamic modeling facilitates accurate and personalized prediction rules for ITR (ALP > 1.67*ULN at 12 months post-treatment initiation) using four select patients. For patient A, predicted risk for ITR was low at treatment initiation (ITR risk: 0.06, 95% CI 0.04–0.08) and generally remained low across the 6-month observation (ITR risk: 0.07; 95% CI 0.04–0.09). Patient B appeared to have only moderate risk of ITR at baseline (0.74, 95% CI 0.68–0.74), but estimated risk increased as more laboratory data became available. At 6 months post-treatment initiation, patient B passed into "high risk" threshold (0.84) for ITR (0.86; 95% CI 0.81–0.91). For patient

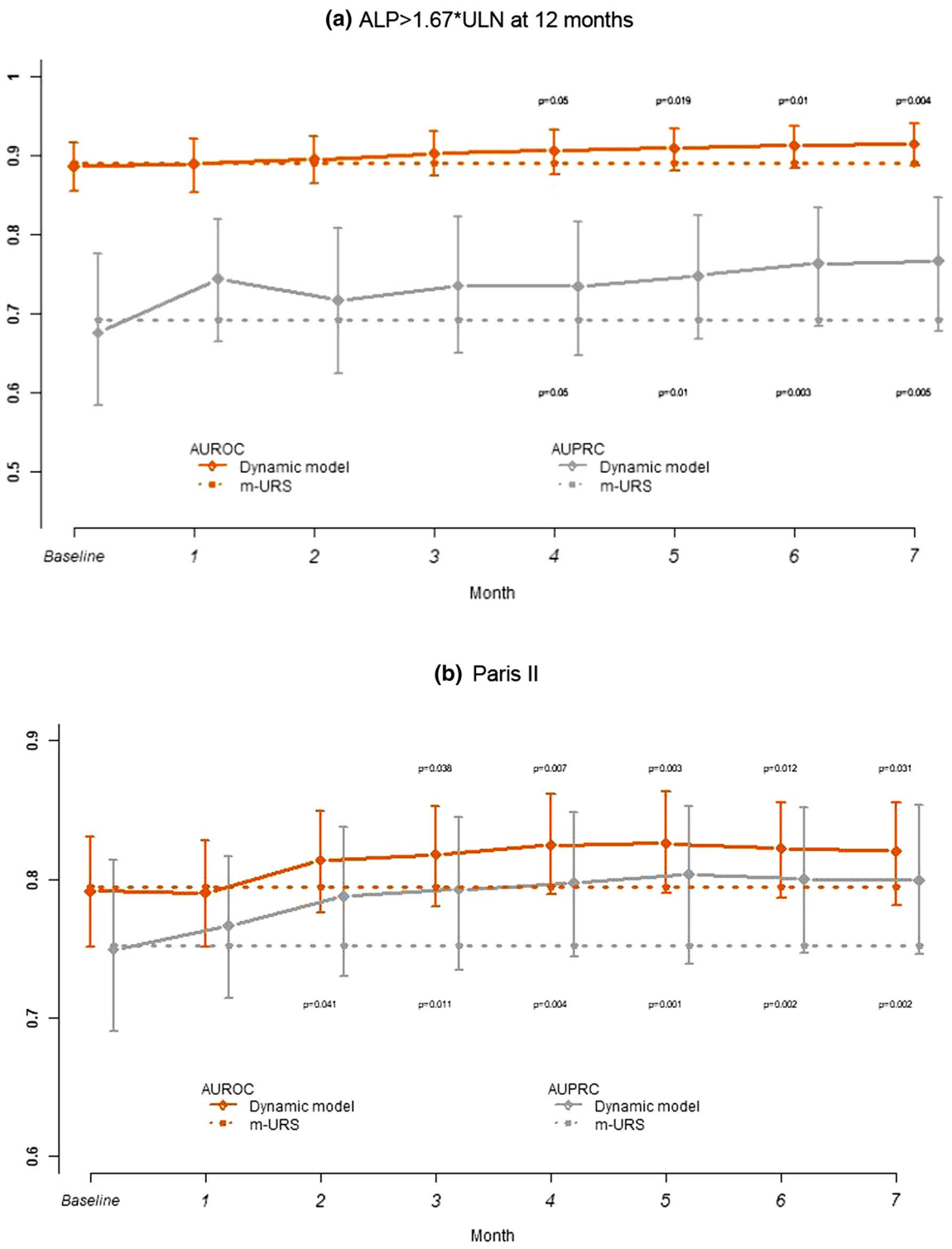

Fig. 1 Comparison of predictive ability between our dynamic model and the m-URS, using area under the ROC curves (AUROC) and area under the PR curves (AUPRC) based on data for up to 7 months after initiation of ursodeoxycholic acid treatment, using two criteria: **a** Alkaline phosphatase (ALP) > 1.67 times the upper limit of normal (ULN) and **b** Paris II criteria. Patients with missing baseline data were imputed ($n = 525$ for validation data)

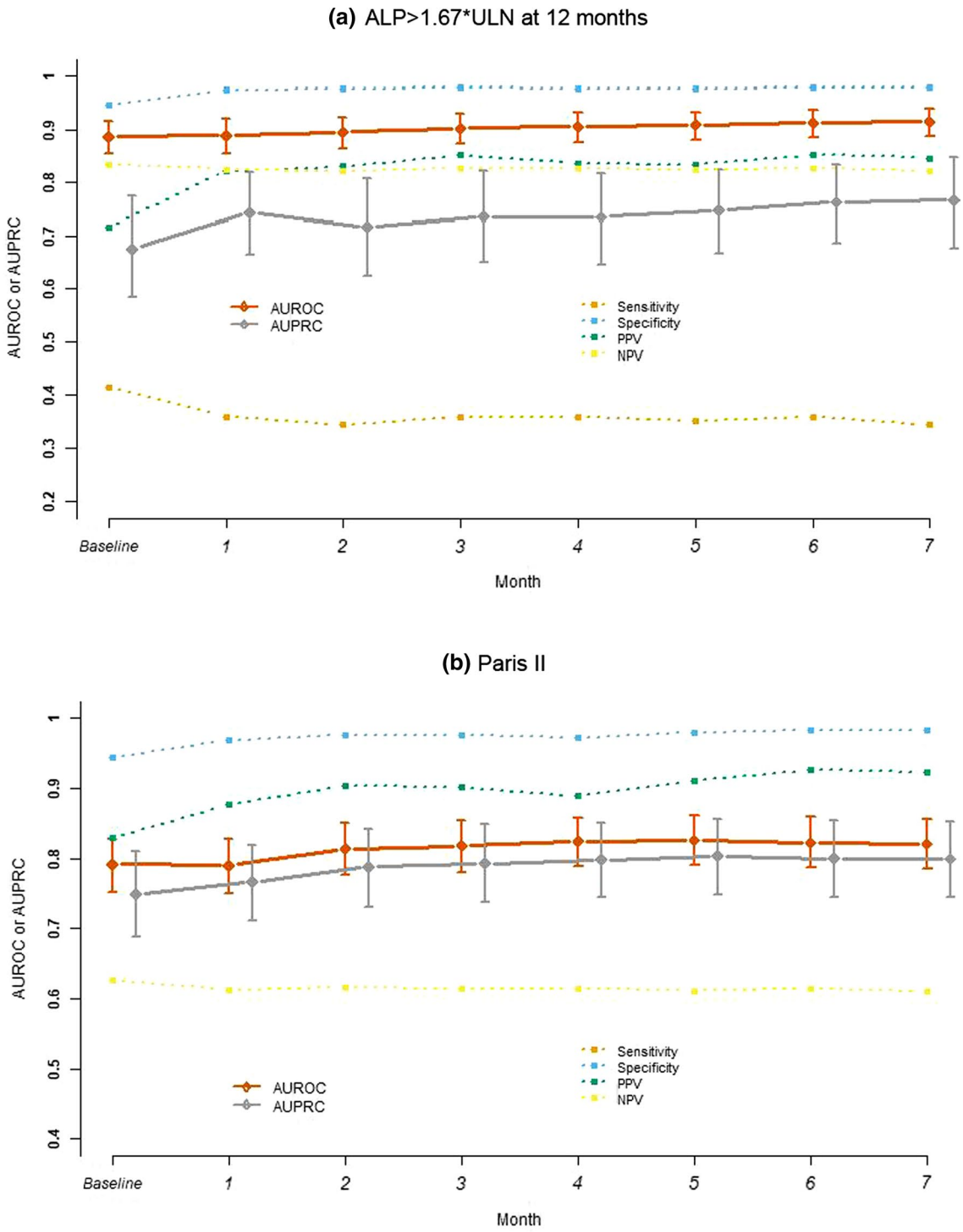

Fig. 2 Prediction metrics for predicting ITR to ursodeoxycholic acid at each time point after treatment initiation, using the dynamic risk prediction model, using **a**: alkaline phosphatase (ALP) > 1.67 times the upper limit of normal at 12 months after treatment initiation; and **b** Paris II criteria. Sensitivity, specificity, PPV, and NPV were calculated based on cutoff of 0.84. Patients with missing baseline data were imputed ($n = 525$ for validation data). *ITR* Inadequate treatment response, *AUROC* Area under the receiver operating curve, *AUPRC* Area under the precision-recall curve, *PPV* positive predictive value, *NPV* negative predictive value

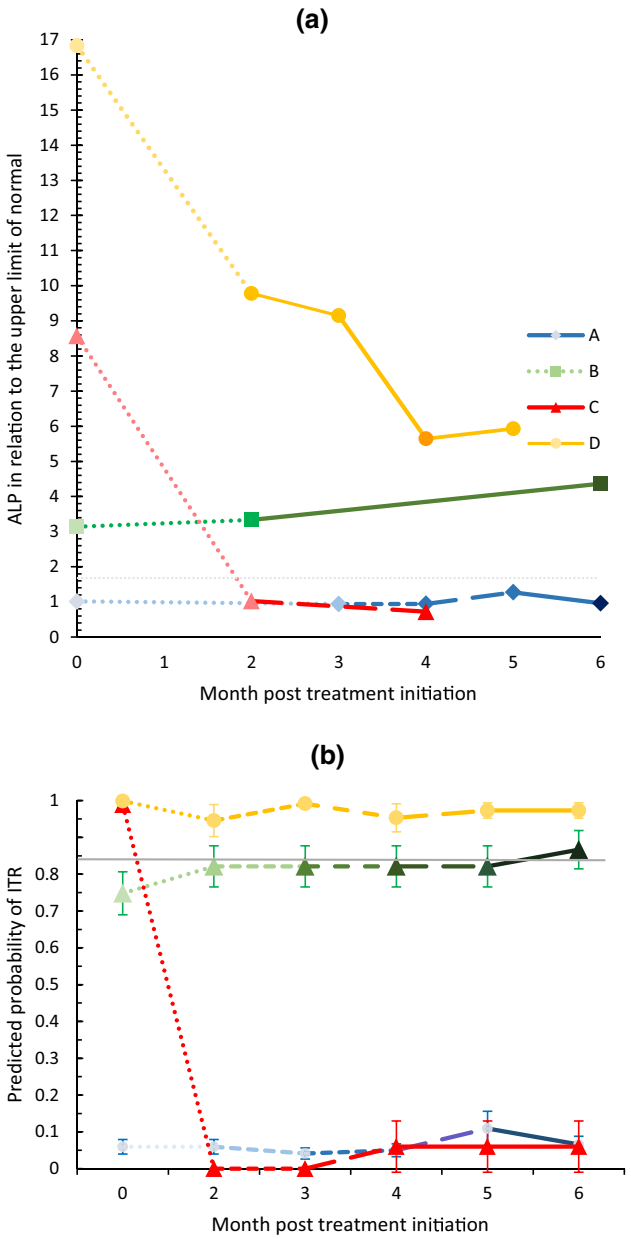

Fig. 3 Repeated measurement of alkaline phosphatase (ALP) data and dynamically updated predicted probabilities of ITR to UDCA treatment for four selected patients. **a** ALP measurements over time for four sample patients. **b** Corresponding predicted probabilities and 95% confidence intervals for ITR for the same sample patients

C, estimated risk of high ITR was high at baseline (predicted probability > 99%). However, as ALP levels decreased after 2 months, the predicted ITR risk also decreased; at 6 months, patient C's estimated risk of ITR dropped to 0.06 (95% CI: -0.01–0.13). In contrast, despite a similar precipitous decline in ALP in the first two months after treatment, patient D remained at high risk of ITR (≥ 95%) across the first 6 months post-treatment initiation.

Next, using the Paris II criteria to define ITR, we developed a dynamic model (Table S1) based on gender, age and repeated measures of ALP and bilirubin. AUROC and AUPRC at baseline were 0.79 (95% CI 0.75–0.83) and 0.75 (95% CI 0.69–0.81), respectively (Fig. 1b). The AUROC was significantly higher than the recalibrated m-URS starting in month 3, which was stable at 0.82 (95% CI: 0.79–0.86) beginning at 4 months post-treatment; the AUPRC was significantly higher than the value of m-URS from month 2 (0.79; 95% CI, 0.73–0.84). With a cut-point of 0.84, the model determined that roughly 30% of patients in our cohort would be considered to have the highest risk of ITR. PPV remained high for predicting ITR, ranging from 83.0% at baseline to 92.3% at 7 months post-treatment; corresponding specificities ranged from 94.4% to 98.3% (Fig. 2b).

Discussion

In a large sample of PBC patients from eleven US health systems, roughly half (45%) of UDCA-treated patients demonstrated ITR. These real-world results are consistent with rates of ITR (20–50%) observed in our own and other studies.[1] Accurate early identification of patients at greatest risk of ITR may allow clinicians to optimize monitoring and care. In response to the growing emphasis on personalized medicine—the "right treatment for the right person at the right time"[10]— we developed a personalized risk prediction model that can predict ITR in a subgroup of patients with high PPV and specificity within 4–6 months of treatment initiation, much sooner than the currently recommended observation time. Our model also showed that baseline values may simply not be sufficient to identify all likely responders or non-responders[10]. As shown in Fig. 3, both patients C and D started with ALP levels that suggested a similar likelihood of ITR; however, as data became available regarding the change in ALP over time, their risk of ITR diverged. This enhanced recognition of what essentially defines UDCA treatment futility could lead to more timely identification of high-risk patients; the excellent accuracy (AUROC = 0.91), PPV of 85% and near-perfect specificity (> 98%) should give clinicians confidence to explore additional treatment options for the subset of patients identified as non-responders.

Given that there is no universally accepted method for defining inadequate response to UDCA treatment, we evaluated both the m-URS and our dynamic model against two of the most commonly used criteria. Our data showed that 27% of patients had ITR at 12 months from index using ALP > 1.67*ULN, whereas 45% had ITR using Paris II criteria. Our own recent study[7] showed that Paris II criteria are an independent prognostic marker for mortality among PBC patients regardless of UDCA treatment. Using data from our racially diverse cohort, we assessed the m-URS (a version

of the URS using baseline-only values) that was developed by the UK-PBC consortium using a European cohort. Accuracy of the m-URS in our racially diverse cohort was similar to that from the UK-PBC cohort (AUROC of 0.85, 95% CI 0.81–0.89) using ALP > 1.67*ULN to define ITR. The UK-PBC group also evaluated three other cutoffs (ALP > 1*ULN; ALP ≥ 1.5*ULN; and ALP ≥ 2*ULN); using these cutoffs, AUROC was reduced to 0.81–0.82 (95% CI 0.77–0.87). In light of this, we also assessed accuracy of the m-URS using Paris II cutoffs; predictive ability was only moderate (AUROC of 0.74).

An unavoidable limitation of any observational study is missing data, especially data collected in a real-world setting. For the validation of m-URS, we addressed this concern by performing multiple imputation[11]; analysis was performed on each imputed dataset, with each set of result combined into a single final result. This method addresses two common concerns regarding imputation of missing data: 1) that imputation based on a regression prediction may be too precise; 2) possible overestimation because observed values are used for the imputation. Although our model used parameters that should be readily available in the medical record of PBC patients under routine care, it is possible that values for some variables may be missing, reducing its predictive ability in these cases. In the dynamic risk prediction model, the parameter estimates were updated monthly if additional laboratory data became available; if no additional measurement was available, the slope derived from the previous measure was used. An advantage of this method is the flexibility with missing values and timing of measurements, as the model can still be applied to irregular measurement intervals, as commonly occurs in routine care settings. The method is also designed to be conceptually simple and practical for use by clinicians. Further sensitivity analysis using alternative imputation methods may be needed to test the robustness of the model. However, as the focus of this paper is to predict ITR rather than estimating a treatment effect, assessing different imputation methods is beyond the scope of this paper. Although our ability to make comparisons is limited by the relatively small sample size of our validation data, our predictive model does demonstrate significantly improved predictability of the dynamic risk prediction model starting from 3 to 4 months post-treatment initiation using both AUROC and AUPRC metrics. A larger validation study is needed to address ideal duration of follow-up measures and appropriate cutoffs.

Using data from a single time point (time of treatment initiation), the four-variable m-URS had "very good" predictive ability (AUROC 0.85–0.87) when validated against data from our multi-racial US cohort; PPV and specificity were roughly 80% and 60%, respectively. In contrast, our two-part dynamic predictive model that used both the values and trajectory of laboratory markers measured over time during the first seven months after treatment initiation reached "excellent" predictive ability (0.91), with very high PPV (85%) and specificity (98%)—which indicates that those patients identified by this model are unlikely to be misclassified as non-responders. Such a model could address a central issue pertaining to the current definition of UDCA response at 12 months post-treatment initiation—specifically, that those patients who might benefit the most from combination and second-line therapies must wait a year before such treatment is initiated. The high specificity (98%) of our model means it identifies only patients for whom UDCA treatment is futile. For these patients, an additional six months wait before initiating second-line treatments may be unnecessarily burdensome, particularly given variation in insurance status and volatility in generic UDCA pricing in the USA.

In conclusion, we used real-world data drawn from a racially and geographically diverse cohort of PBC patients in the USA to develop a dynamic prediction model for inadequate response to UDCA, which provided "excellent" predictive ability, as well as 98% specificity to identify patients at the highest risk of ITR within 6 months of UDCA treatment initiation. This method may allow clinicians to use routine care data to identify earlier those patients deemed ITR who will benefit from heightened monitoring and prompt initiation of adjuvant and/or second-line therapies.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10620-021-07219-4.

Author's contributions ML and SG contributed to study concept and design. ML, LR, IH, JV, RR, CR, MR, JB, MS, YD, AS, JV, CB, and KL contributed to acquisition of data. ML, HW, and JL contributed to analysis and interpretation of data. ML, LR, ST, and SG contributed to drafting of the manuscript. ML, LR, IH, JV, RR, CR, MR, JB, MS, YD, AS, JV, CB, KL, ST, CM, and SG contributed to critical revision of the manuscript for important intellectual content. ML and SG obtained funding. LR and CM contributed to administrative, technical, or material support. ML contributed to study supervision.

Financial support Intercept Pharmaceuticals Inc.

Declarations

Conflicts of interest Stuart C. Gordon receives grant/research support from AbbVie Pharmaceuticals, Conatus, CymaBay, Gilead Pharmaceuticals, Intercept Pharmaceuticals, and Merck. Mei Lu, Jia Li, Lora Rupp, Sheri Trudeau, Yueren Zhou, Christina Melkonian, Yihe G. Daida, Mark A. Schmidt, and Joseph A. Boscarino receive grant/research support from Gilead Pharmaceuticals. Carla V. Rodriguez owns stock in Gilead (< $5,000). Heather Anderson receives grant/research support from Intercept Pharmaceuticals. Jeffrey J. VanWormer receives grant/research support from Retrophin. Christopher L. Bowlus receives grant/research support from AbbVie Pharmaceuticals, Bristol-Myers-Squibb, Gilead Biosciences, Intercept Pharmaceuticals, Merck, Shire Pharmaceuticals, Takeda Pharmaceuticals, and has served as an advisor for Bristol-Myers-Squibb, Gilead Biosciences, Intercept Pharmaceuticals, and Takeda. Keith Lindor is a consultant/advisor for Biopharma and has served as an ad hoc advisor for HighTide, Takeda,

Shire, and Intercept Pharmaceuticals. He sits on a Data Safety Monitoring Board for Takeda. Robert J. Romanelli receives received grant/research support from Pfizer Inc. and Janssen Scientific Affairs. Irina V. Haller, Marsha A. Raebel, and Jennifer Vincent have no conflicts of interest to declare.

References

1. Carey EJ, Ali AH, Lindor KD. Primary biliary cirrhosis. *Lancet.* 2015;386:1565–1575. https://doi.org/10.1016/S0140-6736(15)00154-3.
2. Corpechot C, Chazouilleres O, Poupon R. Early primary biliary cirrhosis: biochemical response to treatment and prediction of long-term outcome. *J Hepatol.* 2011;55:1361–1367. https://doi.org/10.1016/j.jhep.2011.02.031.
3. Carbone M, Nardi A, Flack S et al. Pretreatment prediction of response to ursodeoxycholic acid in primary biliary cholangitis: development and validation of the UDCA Response Score. *Lancet Gastroenterol Hepatol.* 2018;3:626–634. https://doi.org/10.1016/S2468-1253(18)30163-8.
4. Cheung AC, Lammers WJ, Murillo Perez CF, et al. Effects of Age and Sex of Response to Ursodeoxycholic Acid and Transplant-free Survival in Patients With Primary Biliary Cholangitis. *Clin Gastroenterol Hepatol*. 2019; Doi:https://doi.org/10.1016/j.cgh.2018.12.028
5. Lu M, Zhou Y, Haller IV, et al. Increasing Prevalence of Primary Biliary Cholangitis and Reduced Mortality With Treatment. *Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association*. 2018;16:1342–1350 e1. Doi:https://doi.org/10.1016/j.cgh.2017.12.033
6. Lu M, Li J, Haller IV, et al. Factors Associated With Prevalence and Treatment of Primary Biliary Cholangitis in United States Health Systems. *Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association*. 2018;16:1333–1341 e6. https://doi.org/10.1016/j.cgh.2017.10.018
7. Gordon SC, Wu KH, Lindor K et al. Ursodeoxycholic Acid Treatment Preferentially Improves Overall Survival Among African Americans With Primary Biliary Cholangitis. *Am J Gastroenterol.* 2020;115:262–270. https://doi.org/10.14309/ajg.0000000000000512.
8. Tanaka A, Takikawa H. Geoepidemiology of primary sclerosing cholangitis: a critical review. *J Autoimmun.* 2013;46:35–40. https://doi.org/10.1016/j.jaut.2013.07.005.
9. Nyblom H, Bjornsson E, Simren M, Aldenborg F, Almer S, Olsson R. The AST/ALT ratio as an indicator of cirrhosis in patients with PBC. *Liver international : official journal of the International Association for the Study of the Liver.* 2006;26:840–845. https://doi.org/10.1111/j.1478-3231.2006.01304.x.
10. Ronca V, Gerussi A, Cristoferi L, Carbone M, Invernizzi P. Precision medicine in primary biliary cholangitis. *Journal of Digestive Diseases.* 2019;20:338–345. https://doi.org/10.1111/1751-2980.12787.
11. Rubin DB. *Multiple imputation for nonresponse in surveys*. Wiley series in probability and mathematical statistics Applied probability and statistics,. Wiley. 1987;xxix

Authors and Affiliations

Jia Li[2] · Mei Lu[2] · Yueren Zhou[2] · Christopher L. Bowlus[3] · Keith Lindor[4] · Carla Rodriguez-Watson[5,6] · Robert J. Romanelli[7] · Irina V. Haller[8] · Heather Anderson[9] · Jeffrey J. VanWormer[10] · Joseph A. Boscarino[11] · Mark A. Schmidt[12] · Yihe G. Daida[13] · Amandeep Sahota[14] · Jennifer Vincent[15] · Kuan-Han Hank Wu[2] · Sheri Trudeau[2] · Loralee B. Rupp[16] · Christina Melkonian[2] · Stuart C. Gordon[1] · For the FOLD Investigators

Mei Lu
mlu1@hfhs.org

Christopher L. Bowlus
clbowlus@ucdavis.edu

Keith Lindor
Keith.Lindor@asu.edu

Carla Rodriguez-Watson
crodriguezwatson@reaganudall.org

Robert J. Romanelli
Romanellir@pamfri.org

Irina V. Haller
IHaller@eirh.org

Heather Anderson
heather.anderson@ucdenver.edu

Jeffrey J. VanWormer
VanWormer.Jeffrey@marshfieldresearch.org

Joseph A. Boscarino
jaboscarino@geisinger.edu

Mark A. Schmidt
Mark.A.Schmidt@kpchr.org

Yihe G. Daida
yihe.g.daida@kp.org

Amandeep Sahota
Amandeep.Sahota@kp.org

Jennifer Vincent
Jennifer.Vincent@BSWHealth.org

Kuan-Han Hank Wu
wukh@umich.edu

Sheri Trudeau
strudea1@hfhs.org

Loralee B. Rupp
lrupp1@hfhs.org

Christina Melkonian
cmelkon1@hfhs.org

Stuart C. Gordon
sgordon3@hfhs.org

[1] Department of Gastroenterology and Hepatology, Henry Ford Health System, and Wayne State University School of Medicine, Detroit, MI, USA

[2] Department of Public Health Sciences, Henry Ford Health System, 3E One Ford Place, Detroit, MI 48202, USA

[3] University of California Davis School of Medicine, Sacramento, CA, USA

[4] College of Health Solutions, Arizona State University, Phoenix, AZ, USA

[5] Mid-Atlantic Permanente Research Institute, Kaiser Permanente Mid-Atlantic States, Rockville, MD, USA

[6] Innovation in Medical Evidence Development and Surveillance, The Reagan-Udall Foundation for the FDA, Washington, DC, USA

[7] Palo Alto Medical Foundation Research Institute, Palo Alto, CA, USA

[8] Essentia Institute of Rural Health, Essentia Health, Duluth, MN, USA

[9] Skaggs School of Pharmacy and Pharmaceutical Sciences, University of Colorado, Aurora, CO, USA

[10] Marshfield Clinic Research Institute, Marshfield, WI, USA

[11] Department of Epidemiology and Health Services Research, Geisinger Clinic, Danville, PA, USA

[12] Center for Health Research, Kaiser Permanente Northwest, Portland, OR, USA

[13] Center for Integrated Health Care Research, Kaiser Permanente Hawaii, Honolulu, HI, USA

[14] Department of Research and Evaluation, Kaiser Permanente Southern California, Los Angeles, CA, USA

[15] Baylor, Scott & White Research Institute, Temple, TX, USA

[16] Center for Health Policy and Health Services Research, Henry Ford Health System, Detroit, MI, USA