**Supplement: Additional risk models and sensitivity analyses**

**Baseline covariates**

“Index date” was defined as the date of UDCA treatment initiation (for treated patients) or PBC diagnosis date (for untreated patients). An index date of 1/1/2006 was assigned to patients diagnosed prior to 2006 to permit inclusion of previously-diagnosed PBC patients the cohort. aseline covariates included: sex; ethnicity (Hispanic, non-Hispanic); race (White; African American; Asian American/ American Indian/ Pacific islander [ASINPI]); age category (≤40; 41–50; 51–60; 61–70; >70); health insurance type; Charlson-Deyo comorbidity score (calculated without liver disease); household income; total bilirubin (categorized as >2, 2>1.5, 1.5>1.0, 1.0>0.7, 0.7>4, ≤0.4 mg/dL); index date albumin and ALP (categorized in relation to normal ranges as defined at each site); and the ratio of aspartate aminotransferase (AST) to alanine aminotransferase (ALT) (ratios ≥1.1 have been shown to indicate advanced fibrosis/ cirrhosis among patients with PBC).(1) Missing data was categorized as “unknown”; the unknown category was included in all risk model analyses to reflect the real-world setting but omitted from figures for the sake of clarity. Patients who had received a liver transplant prior to index date were excluded from analyses of the composite outcome of liver transplant/ death.

**Models incorporating covariates measured at Year 1 post-index**

A similar set of parameters was collected for the first laboratory data available on or after 12 months post-index. We also noted whether patients have levels above or below the Paris-II criteria (2) at 12 months post-index (alkaline phosphatase [ALP] or AST ≥1.5 times the upper limit of normal [ULN, as defined by the assay used at each site] or total bilirubin >1 mg/dL). Missing data was categorized as “unknown.”

We generated two models for risk of mortality that added covariates collected at Year 1 of follow-up to those in the baseline model to determine whether the inclusion of these variables improved the accuracy of the model: 1) BL+P2: baseline variables plus the interaction between UDCA and Paris-II criteria; 2) BL+1: baseline variables plus a set of categorical variables—ALP (normal or >ULN, based on the assay used at each site), the ratio of AST/ALT, and total bilirubin. We also compared these Year 1 models to a model that included the Paris-II criteria alone. A similar analytical approach was used to study the risks and the impact of UDCA treatment on liver transplant/ death.

**Results of mortality risk models incorporating covariates measured at Year 1**

Predictive accuracy (AUROC) of all models is summarized in Table S6. AUROC for the BL+P2 model was “very good” at Year 2 post-index (0.89), and “good” at Year 5 (0.84). For the BL+1 model, the incorporation of the Year 1 variables improved the predictive accuracy of the model (AUROC=0.90 at Year 2 post-index; 0.85 at Year 5). The predictive accuracy of both of these models was better than a model using Paris-II criteria alone (P2; AUROC=0.84 at Year 2 post-index; 0.74 at Year 5).

**Models for impact of UDCA treatment on liver transplant/ death**

For this analysis, we excluded 111 patients who had received a liver transplant prior to their index date, leaving a sample of 4127. After IPTW, there were no significant differences in patient characteristics (including index year) between treatment groups. We generated two TXF models that added covariates measured at Year 1 after index: 1) TXF+P2, using baseline covariates plus the Paris-II criteria at Year 1; and 2) TXF+1: baseline covariates plus a set of covariates measured at Year 1 post-index.

**Results of liver transplant/ death risk models incorporating covariates measured at Year 1**

The TXF+P2 model was consistent with the TXF model; model accuracy (AUROC) was 0.85 at Year 2 post-index and 0.84 at Year 5. Likewise results from the TXF+1 model were similar to those from the TXF baseline model, but predictive accuracy improved at both Year 2 and Year 5 post-index (AUROC=0.86 and 0.84, respectively). A model using Paris-II criteria yielded AUROC of 0.79 and 0.73 (at Years 2 and 5 post-index, respectively).

**Selection of the optimal models to predict all-cause mortality and liver transplant/ death**

Predictive accuracy of all models, is presented in Table S6. All models reached “good” to “excellent” predictive ability (AUROC≥0.80). As expected, models incorporating Year 1 data improved AUROCs by 4–9% compared to models using baseline data. The model using values collected at baseline and at Year 1 post-index demonstrated the highest predictive accuracy for both Year 2 and Year 5 of follow-up. However, the baseline models were considerably more parsimonious and demonstrated reasonable predictive accuracy. Given the burden of additional observation for the models using covariates measured at Year 1, we believe the greatest utility lies in the baseline models, without much loss of accuracy.

1. 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 Int 2006;26:840-845.

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.