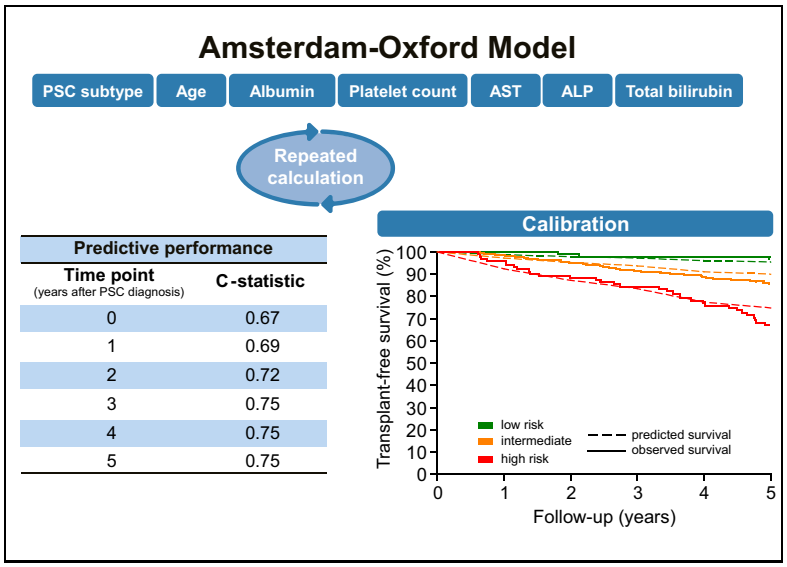
Research Article
Cholestasis and Autoimmune Diseases

JOURNAL
OF HEPATOLOGY

Validation, clinical utility and limitations of the Amsterdam-Oxford model for primary sclerosing cholangitis

Graphical abstract



Highlights

- Reliable estimates of survival are pivotal to optimize clinical management of patients with PSC.
- The Amsterdam-Oxford model (AOM) was recently introduced to estimate survival for patients with PSC at diagnosis.
- The AOM has adequate discriminatory performance and good predictive accuracy at PSC diagnosis.
- It maintains this performance and accuracy at other time points during follow-up.

Authors

Jorn C. Goet, Annarosa Floreani, Xavier Verhelst, ..., Adriaan J. van der Meer, Henk R. van Buuren, Bettina E. Hansen

Correspondence

j.c.goet@gmail.com
(J.C. Goet)

Lay summary

In our study we assessed whether the Amsterdam-Oxford model (AOM) is able to correctly estimate the risk of liver transplantation or death in patients with primary sclerosing cholangitis (PSC). This model uses 7 objective and readily available variables to estimate prognosis for individual patients at the time of PSC diagnosis. The AOM may aid in patient counselling and timing of diagnostic procedures or therapeutic interventions for complications of liver disease. We confirm that the model works well at PSC diagnosis, but also when the AOM is recalculated at different timepoints during follow-up, greatly improving the applicability of the model in clinical practice and for individual patients.

http://dx.doi.org/10.1016/j.jhep.2019.06.012
© 2019 European Association for the Study of the Liver. Published by Elsevier B.V. All rights reserved.

Research Article
Cholestasis and Autoimmune Diseases

JOURNAL OF HEPATOLOGY

Validation, clinical utility and limitations of the Amsterdam-Oxford model for primary sclerosing cholangitis

Jorn C. Goet[1,*], Annarosa Floreani[2], Xavier Verhelst[3], Nora Cazzagon[2], Lisa Perini[2], Willem J. Lammers[1], Annemarie C. de Vries[1], Adriaan J. van der Meer[1], Henk R. van Buuren[1], Bettina E. Hansen[1,4,5]

[1]*Department of Gastroenterology and Hepatology, Erasmus University Medical Center, Rotterdam, The Netherlands;* [2]*Department of Surgery, Oncology and Gastroenterology, University of Padua, Padua, Italy;* [3]*Department of Gastroenterology and Hepatology, Ghent University Hospital, Belgium;* [4]*Toronto Centre for Liver Disease, Toronto General Hospital, University of Toronto, Toronto, Canada;* [5]*Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Canada*

See Editorial, pages 867–870

Background & Aims: Recently the Amsterdam-Oxford model (AOM) was introduced as a prognostic model to assess the risk of death and/or liver transplantation (LT) in primary sclerosing cholangitis (PSC). We aimed to validate and assess the utility of the AOM.

Methods: Clinical and laboratory data were collected from the time of PSC diagnosis until the last visit or time of LT or death. The AOM was calculated at yearly intervals following PSC diagnosis. Discriminatory performance was assessed by calculation of the C-statistic and prediction accuracy by comparing the predicted survival with the observed survival in Kaplan-Meier estimates. A grid search was performed to identify the most discriminatory AOM threshold.

Results: A total of 534 patients with PSC and a mean (SD) age of 39.2 (13.1) years were included. The diagnosis was large duct PSC in 466 (87%), PSC with features of autoimmune hepatitis in 52 (10%) and small-duct PSC in 16 (3%). During the median (IQR) follow-up of 7.8 (4.0–12.6) years, 167 patients underwent LT and 65 died. The median LT-free survival was 13.2 (11.8–14.7) years. The C-statistic of the AOM ranged from 0.67 at baseline to 0.75 at 5 years of follow-up. The difference between the predicted and observed survival ranged from −1.6% at 1 year to + 3.9% at 5 years of follow-up. Patients that developed AOM scores >2.0 were at significant risk of LT or death (time-dependent hazard ratio 4.09; 95% CI 2.99–5.61).

Conclusions: In this large cohort of patients with PSC, the AOM showed an adequate discriminative performance and good prediction accuracy at PSC diagnosis and during follow-up. This study further validates the AOM as a valuable risk stratification tool in PSC and extends its utility.

Lay summary: In our study we assessed whether the Amsterdam-Oxford model (AOM) is able to correctly estimate the risk of liver transplantation or death in patients with primary sclerosing cholangitis (PSC). This model uses 7 objective and readily available variables to estimate prognosis for individual patients at the time of PSC diagnosis. The AOM may aid in patient counselling and timing of diagnostic procedures or therapeutic interventions for complications of liver disease. We confirm that the model works well at PSC diagnosis, but also when the AOM is recalculated at different timepoints during follow-up, greatly improving the applicability of the model in clinical practice and for individual patients.

© 2019 European Association for the Study of the Liver. Published by Elsevier B.V. All rights reserved.

Introduction

Primary sclerosing cholangitis (PSC) is a chronic, variably progressive cholestatic liver disease characterized by inflammation of the intrahepatic and extrahepatic bile ducts, sclerosis and destruction of the biliary tract.[1–4] This leads to chronic cholestasis, biliary fibrosis and (decompensated) cirrhosis, which may eventually culminate into liver failure requiring liver transplantation; the only potential curative treatment for PSC.[2,3] Following a PSC diagnosis a median transplant-free survival of 13 years has been reported in studies from tertiary referral centres, although this may be longer in a population-based setting.[5]

One of the major challenges in the management of PSC is the lack of therapies that halt disease progression. Despite the biochemical improvement reported with ursodeoxycholic acid (UDCA) treatment in PSC, a survival benefit has never been reported.[6–11] Another challenge concerns reliable estimation of prognosis in PSC, largely because of the heterogeneity in clinical course progression and the variety of outcomes ranging from end-stage liver disease to development of hepatobiliary and colorectal malignancies.[12–14] In this setting, risk prediction models that quantify the risk of future events for individual patients with PSC are of critical importance for patient counselling, timely diagnostic procedures and subsequent therapeutic interventions for disease-related complications. Also, reliable risk stratification is important for the selection of patients in future drug development trials.

The Mayo risk score (MRS) is the most frequently used score to assess the short-term (4-year) mortality risk of patients with

Keywords: Primary sclerosing cholangitis; Cholestasis; Autoimmune liver disease; Risk stratification, prognostic modelling.
Received 19 November 2018; received in revised form 20 May 2019; accepted 19 June 2019; available online 3 July 2019

* Corresponding author. Address: Erasmus MC, University Medical Center, Department of Gastroenterology and Hepatology, Postbus 2040, 3000 CA Rotterdam, Internal Postal Address NA606, Visiting Address: Room 624, Dr. Molewaterplein 40, 3015 GD Rotterdam, Rotterdam, The Netherlands.
E-mail address: j.c.goet@gmail.com (J.C. Goet).

PSC. However, this score was mainly derived from a cohort of patients with end-stage disease in a liver transplant centre. This may limit its applicability in early stages of disease.[15] Recently the Amsterdam-Oxford model (AOM) was introduced; a prognostic model developed in a population-based cohort to predict the long-term risk of PSC-related death and/or liver transplantation.[16] The AOM incorporates PSC subtype, age at PSC diagnosis, aspartate aminotransferase (AST), alkaline phosphatase (ALP), total bilirubin, albumin and platelet count. This score, based on these 7 readily available variables, showed an adequate discriminative power and satisfactory calibration in a derivation and validation cohort. However, further validation of this population-based model is necessary to justify its application in other cohorts and centres and to extend its use at other timepoints during follow-up. In addition, the performance of this newly developed score has not been compared to the MRS. Therefore, we aimed to further validate the AOM and assess its utility in a large cohort of patients with PSC from 3 tertiary centres in Europe. A secondary aim was to compare the performance of the AOM with that of the MRS.

Patients and methods

Population and study design

This retrospective cohort study included patients with PSC from 3 tertiary centres in Europe: University of Padua, Italy, Ghent University Hospital, Belgium, and Erasmus University Medical Center, Rotterdam, The Netherlands. Data were collected from 1984 up to June 2016 for University of Padua, from 1977 up to June 2016 for the Rotterdam University Medical Center, and from 1993 up to June 2018 for Ghent University Hospital. Complete follow-up was defined as liver transplantation or death or clinical follow-up beyond 1 January 2016 for Padua and Rotterdam, and beyond 1 January 2018 for Ghent. Patients who were diagnosed with PSC at age ≥18 years and in accordance with the European Association for the Study of the Liver guidelines were included.[17] Patients with less than 6 months follow-up, with or without an event, were excluded to ensure exclusion of patients that were referred because of liver failure and that were consequently diagnosed with PSC in the process of being waitlisted for LT. In addition, patients were excluded if the date of diagnosis was unknown or in the case of concomitant liver disease. Clinical and laboratory data were collected from the start of follow-up until the last visit or clinical event at 6-monthly or 1-yearly intervals according to the intervals between visits at each centre. Biochemical parameters collected included AST, prothrombin time, international normalized ratio, alanine aminotransferase, ALP, gamma-glutamyl transferase, total bilirubin, albumin and platelets. Clinical data included sex, age, date of PSC diagnosis, liver histology, UDCA treatment, concomitant inflammatory bowel disease (IBD), last follow-up date, or the date of clinical outcomes. Patients with a diagnosis of IBD within the first year following a diagnosis of PSC were considered to have IBD at baseline. The primary endpoint of the current study was a combined endpoint of liver transplantation or death. For patients with a diagnosis of untreatable cholangiocarcinoma (CCA) and unknown clinical outcome we considered the last follow-up as the date of death.

This study was conducted in accordance with the protocol and the principles of the Declaration of Helsinki. The protocol was approved by the Institutional Research Board of the corresponding centre, and at each participating centre, in accordance with local regulations.

Statistical analyses

Normally distributed data are presented as mean ± SD and skewed distributed data as median and interquartile range. Statistical analyses were performed with IBM SPSS Statistics 22.0 (SPSS Inc, Chicago, IL) and SAS 9.4 (SAS institute Inc., Cary, NC) (Supplementary CTAT Table). To account for missing values SAS (SAS Proc MI, MCMC method) was used to generate 10 imputed datasets of laboratory results at yearly timepoints for up to 5 years following PSC diagnosis. Missing data were considered to be missing at random. Rubin's rules were used for estimation of the parameters and the standard error.[18–20] The imputation model included baseline variables that were potentially predictive for outcomes in PSC (*e.g.* year of diagnosis, age) as well as the outcomes themselves. Only continuous biochemical variables were imputed. All analyses were performed in the original database as well as in the imputed dataset.

The start of follow-up was set at PSC diagnosis defined by the first pathological imaging result (magnetic resonance cholangiography or endoscopic retrograde cholangiography) or liver biopsy. The AOM score was calculated at yearly intervals, starting at diagnosis and continuing for up to 5 years (laboratory values used ± 3 months around time point of calculation). The AOM score was calculated using the following formula: AOM score = 0.323*PSC subtype (1 = large duct PSC; 0 = small-duct PSC) + 0.018*age at diagnosis – 2.485*log(albumin*lower limit of normal [LLN]) + 2.451*abs(log[platelets-0.5] + 0.347*log[AST* upper limit of normal (ULN)] + 0.393*log[ALP*ULN]) + 0.337*log [total bilirubin*ULN]). For calculation of AOM values between 1 year and 5 years following a PSC diagnosis we used the actual age at the time of laboratory assessment instead of the age at diagnosis. The association of the AOM scores with the primary endpoint was assessed in Cox-regression analyses at diagnosis and at each year up to 5 years thereafter. Discriminatory performance of the model was assessed at various timepoints by calculation of the C-statistic. As a next validation step, we estimated the hazard ratio of the AOM score for the combined endpoint, to assess the fit of the model (*i.e.* whether a model overestimates or underestimates risk).[21] In case the log hazard ratio is equal to 1 the model has a perfect fit. A lower value indicates the model underestimates risk, while higher values suggest an overestimation of risk. In addition, the fit/potential misspecification of the AOM in our cohort at PSC diagnosis was assessed by running a Cox-regression analysis including the separate variables comprising the AOM as well as the AOM score itself.[22] In this regression model the coefficient score of the AOM was constrained to equal 1 (*i.e.* offsetting the score of the AOM).[22] If the β values of the separate variables of the model are not significantly different from 0 in these analyses, the AOM gives a perfect fit. If on the other hand a β is significantly different from 0, there is misclassification and the AOM can be improved by adjusting the βs of these variables. Finally, prediction accuracy (*i.e.* calibration) was assessed by comparing the predicted versus the observed Kaplan-Meier survival curves to assess calibration.[23] To assess prediction accuracy across different AOM score intervals, we divided patients into 3 risk groups based on their AOM score, using threshold points at the 20th and 80th percentiles.

A repeated linear model with a random intercept and slope per patient using an unstructured covariance matrix was performed to analyse the evolution of AOM scores over time in those with and without an endpoint at the end of

follow-up. To determine an AOM threshold with the highest power to discriminate patients achieving the primary endpoint from those not achieving the primary endpoint, we performed a grid search with calculation of C-statistic between an AOM score of 0.8 and 4.0 in steps of 0.1 at each year of follow-up. The optimal threshold was subsequently included in Cox-proportional hazards analyses in order to estimate the strength of association with the liver transplantation-free survival, as a baseline variable and time-dependent variable separately.

The value of ALP alone in making absolute risk predictions of transplant-free survival was assessed. Cox-proportional hazard regression analyses were used to assess the association between baseline log(ALP) and time to event. From this Cox model the baseline linear prediction equation of ALP (prognostic index), along with the baseline survival estimate $S_0(t)$, t = time, were derived. The prediction accuracy of ALP was assessed by comparing the observed versus the predicted transplant-free survival rate, for the total cohort as well as for different percentiles of risk (<20th, 20th-80th, >80th)." Finally, to compare the AOM with the MRS, the MRS was calculated with the formula: 0.0295*age in years + 0.5373*ln(total bilirubin in mg/dl) – 0.8389*serum albumin in g/dl + 0.5380*ln(AST in IU/L) + 1.2426*(points for variceal bleeding [0 = not present or 1 = present]).

Results

Baseline cohort characteristics

Data were obtained from 601 patients with PSC of whom 534 patients met the inclusion criteria. A total of 48 patients were excluded because we were unable to obtain a date of diagnosis and 19 patients had a follow-up <6 months (Fig. S1)." A total of 13,344 patient visits and a mean of 25 visits per patient were reported across the entire cohort.

The mean (SD) age was 39.2 (13.1), 66% were male, and 93% were UDCA-treated. The baseline patient characteristics are summarized in Table 1. The diagnosis was large duct PSC in 466 (87%), 52 (10%) had PSC with features of autoimmune hepatitis and 16 (3%) had small-duct PSC. The year of PSC diagnosis ranged from 1977 to 2017. At baseline, 268 (60%) patients had IBD: 77% had ulcerative colitis and 20% had Crohn's disease. In total, 427 (80%) patients had complete follow-up. During the median follow-up period of 7.8 years (interquartile range, 4.0–12.6 years) a total of 232 (43%) patients reached a clinical endpoint: liver transplantation was performed in 167 patients and 65 patients died. The transplant-free survival rates were 98.3% at 1 year, 84.4% at 5 years, and 65.9% at 10 years of follow-up, as shown in Fig. 1. The median transplant-free survival was 13.2 (11.8–14.7) years.

For the 65 patients who died in our cohort, the cause of death was variceal bleeding in 5 patients, spontaneous bacterial peritonitis in 4, hepatorenal syndrome in 2, liver failure (unspecified) in 6, hepatocellular carcinoma in 2, signet ring cell carcinoma metastasized to the ductus choledochus in 1, pancreatic cancer in 1, colorectal cancer in 4, lung cancer in 2, renal insufficiency in 1, sepsis in 5, and surgical complications in 3. For 4 patients we could not determine the cause of death. A total of 25 patients died from CCA. Six of these patients had untreatable CCA and unknown clinical outcome. For these patients we considered the last follow-up as the date of death.

Table 1. Baseline patient characteristics.

	Total cohort, n = 534
Age at diagnosis, y, mean (SD)	39.2 (13.1)
Male, n (%)	351 (65.7)
UDCA treated, n (%)	493 (92.3)
PSC type, n (%)	
Large duct PSC	466 (87.3)
PSC with features of AIH	52 (9.7)
Small–duct PSC	16 (2.6)
Year of diagnosis	2004 (1995–2009)
Year of diagnosis, range	1977–2017
IBD at baseline[a], n(%)	268 (60.2)
UC	206 (76.9)
CD	54 (20.1)
IBD–U	8 (3.0)
Indeterminate	0 (0)
Follow-up, years	7.8 (4.0–12.6)
Laboratory data at diagnosis[b]	
Serum total bilirubin × ULN	1.0 (0.52–2.30)
Serum ALP × ULN	1.99 (1.11–3.59)
Serum GGT × ULN	4.94 (2.11–9.94)
Serum AST × ULN	1.79 (1.08–3.00)
Serum ALT × ULN	2.13 (1.27–3.92)
Serum albumin × LLN	1.17 (1.03–1.30)
Serum platelets x10^3/mm^3	258 (195–332)
Prediction model scores at diagnosis	
Amsterdam–Oxford model score	1.70 (1.30–2.17)
Mayo risk score[c]	–0.41 (–1.21–0.56)

ALP, alkaline phosphatase; ALT, alanine aminotransferase; AST, aspartate aminotransferase; CD, Crohn's disease; GGT, gamma-glutamyl transferase; IBD, inflammatory bowel disease; IBD-U, IBD unclassified; LLN, lower limit of normal; UC, ulcerative colitis; UDCA, ursodeoxycholic acid; ULN, upper limit of normal.
Data presented as median (interquartile range) unless specified otherwise.
[a] Unknown for 50 patients (9.4%).
[b] Laboratory data presented here from imputed data. Missing values in our cohort ranged from 26% to 30%.
[c] Data available for 498 patients. For 36 patients data on variceal bleeding was not available.

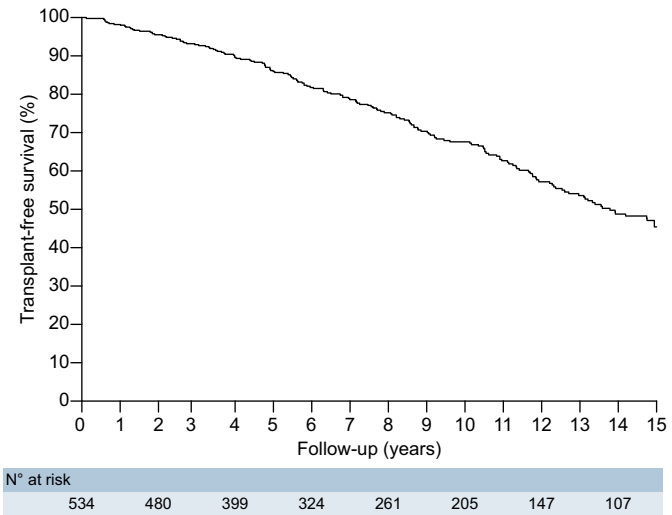


Fig. 1. Kaplan-Meier estimate of transplant-free survival.

Discriminatory performance of the Amsterdam-Oxford model and misspecification/fit

The overall discriminatory performance for death or liver transplantation of the AOM score at diagnosis calculated with C-statistic was 0.67 (95% CI 0.64–0.70) and ranged to 0.75 (95% CI 0.71–0.78) at 5 years following diagnosis (Table 2). The

Table 2. Discriminative performance and assessment of the fit of the Amsterdam-Oxford model calculated n years after diagnosis.

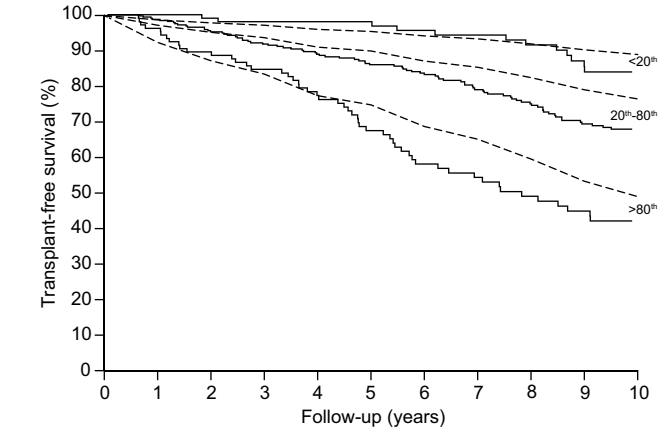
	Descriptives		Performance	Measure of fit
Year(s) after diagnosis	n	Median (IQR) AOM value	C–statistic (95% CI)	Hazard ratio (95% CI)[a]
0	534	1.70 (1.30–2.17)	0.6704 (0.6392–0.7015)	2.18 (1.77–2.68)
1	516	1.63 (1.24–2.17)	0.6927 (0.6615–0.7238)	2.84 (2.31–3.47)
2	480	1.67 (1.24–2.26)	0.7230 (0.6936–0.7524)	3.36 (2.75–4.11)
3	433	1.76 (1.29–2.41)	0.7484 (0.7195–0.7773)	3.70 (3.04–4.51)
4	399	1.76 (1.32–2.40)	0.7458 (0.7139–0.7777)	2.61 (2.21–3.09)
5	365	1.74 (1.31–2.44)	0.7461 (0.7114–0.7807)	2.94 (2.42–3.57)

[a] Hazard ratios with 95% CIs that include exp(^1) (=2.72) indicate good fit. Hazard ratios greater and not including 2.72 in the confidence interval indicate overestimation of risk by the model.

AOM had a good fit, by univariable Cox-regression analyses, at baseline and at most timepoints during follow-up, with a hazard ratio for clinical events ranging from 2.18 (95% CI 1.77–2.68) at diagnosis to 2.94 (95% CI 2.42–3.57) at 5 years of follow-up. Detailed assessment of the AOM fit/misspecification at PSC diagnosis indicated that only the βs for platelet count and age were significant predictors when offsetting the score of the AOM equal to 1. When adding these variables to the AOM score at baseline in the calculation of C-statistic, the C-statistic was 0.682 (95% CI 0.644–0.719).

Prediction accuracy (calibration) of the Amsterdam-Oxford model

For the total cohort, the difference between the calculated mean survival based on AOM values at PSC diagnosis and the observed Kaplan-Meier survival ranged from −1.6% (96.7% predicted *vs.* 98.3% observed) at 1 year to 3.9% at 5 years of follow-up (88.3% predicted *vs.* 84.4% observed). Fig. 2 shows further assessment of AOM calibration in different risk groups stratified by AOM score percentiles (<20th, 20-80th and >80th percentile). Similar prediction accuracy was observed when the score was recalculated at 1 year, 3 years and 5 years after diagnosis for the 5 years following calculation, with the most accurate predictions being made in the lower-percentile and mid-percentile groups of the AOM score (Fig. S2A-C). Underestimation of the risk of death or LT was greater in the highest percentile (Fig. S2-A-C).



N° at risk/events						
<20th	107/0	101/1	88/2	72/4	65/7	49/12
20th-80th	321/0	288/15	240/33	206/33	161/66	128/82
>80th	106/0	91/12	71/23	46/40	35/47	28/52

Fig. 2. Predicted versus observed liver transplant-free survival according to Amsterdam-Oxford model score percentiles. Fig. shows prediction accuracy (calibration) of the Amsterdam-Oxford model score up to 10 years of follow-up across different percentiles of the scores (divided into 3 groups based on 20th and 80th percentile) at diagnosis. Solid lines = actual observed transplant-free survival probabilities estimated by Kaplan-Meier analyses. Dashed lines = the predicted mean transplant-free survival probabilities as predicted by the Amsterdam-Oxford model.

Utility of the Amsterdam-Oxford model and evolution over time

Visualization of AOM scores over time by a repeated linear model revealed that patients who were alive at the end of follow-up had consistently low AOM scores during follow-up (Fig. 3). A subsequent grid search of AOM scores based on the C-statistic revealed that the most discriminatory threshold for death or liver transplantation ranged between 1.8 and 2.1, if calculated at baseline and during the first 5 years of follow-up (C-statistic 0.61–0.69; Fig. 4 and Fig. S3A-E). In Cox-regression analyses AOM scores above 2.0 were significantly associated with clinical events (hazard ratio ranging between 2.30 [95% CI 1.75–2.95] at diagnosis and 4.46 [95% CI 3.19–6.24] at 5 years following diagnosis, Fig. 4). At baseline, a total of 174 (32.6%) patients had AOM values above 2.0. During the first 5 years following PSC diagnosis an additional 8.4% within 1 year, 13.9% within 3 years, and 25.4% within 5 years developed AOM values above the threshold of 2.0. Patients that reached an AOM score

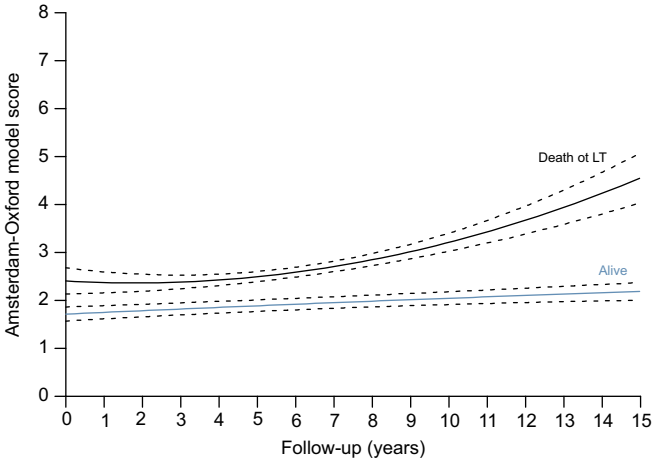
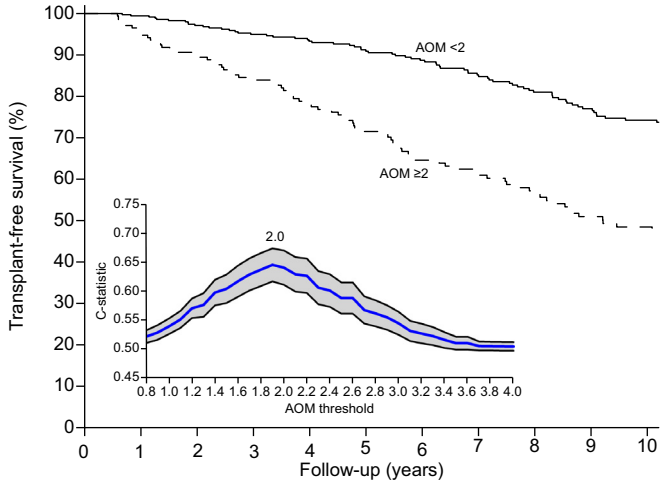


Fig. 3. Evolution of Amsterdam-Oxford model scores during follow-up stratified according to endpoint. Fig. shows Amsterdam-Oxford model scores over time in those with a clinical event (death or liver transplantation) and those without a clinical event at the end of follow-up from a repeated linear model with a random intercept and slope per patient in an unstructured covariance matrix. Solid line = predicted mean of the Amsterdam-Oxford model score; dashed lines = 95% confidence intervals. AOM, Amsterdam-Oxford model; LT, liver transplantation.



N° at risk											
	360	353	330	298	280	259	236	210	187	169	150
	174	163	150	135	119	106	90	83	74	63	55

Fig. 4. Kaplan-Meier estimate of transplant-free survival stratified according to an Amsterdam-Oxford model threshold of 2.0 at diagnosis. Kaplan-Meier estimate of transplant-free survival in patients with Amsterdam-Oxford model scores below 2.0 (solid line) and scores equal to or above 2.0 (dashed line) at diagnosis. Panel shows C-statistics for the grid search for a threshold in AOM score between 0.8 and 4.0 in steps of 0.1 at diagnosis. AOM, Amsterdam-Oxford model.

of 2.0 in the first 5 years of follow-up patients were at significant risk of death or liver transplantation (time-dependent HR 4.09; 95% CI 2.99–5.61).

ALP in isolation as a predictor of transplant-free survival

In the first 5 years following a PSC diagnosis, the C-statistic for ALP alone as a predictor of transplant-free survival ranged between 0.52 and 0.63 (Table S1). For the total cohort the difference between observed and predicted survival increased from −0.2% at 1 year after PSC diagnosis to −4.9% at 10 years. Further assessment of ALP calibration in different risk groups stratified according to percentiles (<20th, 20-80th and >80th percentile), revealed that the difference between observed and predicted survival between 1 year and 10 years following a PSC diagnosis ranged from −1.8% to −7.6% for the <20th percentile, from −0.5% to −2.9% for the 20th-80th percentile, and from −0.1% to −12.4% for the >20th percentile (Fig. S4).

Comparison of the Mayo risk score and Amsterdam-Oxford model

The MRS could be calculated for a sub-cohort of 498 patients at baseline. At diagnosis, a total of 311 (62.4%) of the patients had an MRS value below or equal to 0 (low-risk group); 161 (32.3%) had scores above 0 but less than 2 ('intermediate risk group') and 26 (5.3%) were considered at high risk of events (MRS greater than 2). The transplant-free survival rates were significantly different between the low, intermediate and high-risk group: 99.4%, 98.1% and 92.3% at 1 year, 95.5%, 88.4% and 65.4% at 3 years, and 91.5%, 77.3% and 47.4% at 5 years of follow-up (log-rank <0.001, Fig. 5). The discriminatory performance of the MRS calculated by C-statistic ranged from 0.73 (95% CI 0.73–0.76) at diagnosis to 0.79 (95% CI 0.76–0.82) at 5 years following PSC diagnosis. Direct comparison of discriminatory performance in patients for whom both the MRS and AOM score could be calculated at PSC diagnosis (n = 498) and

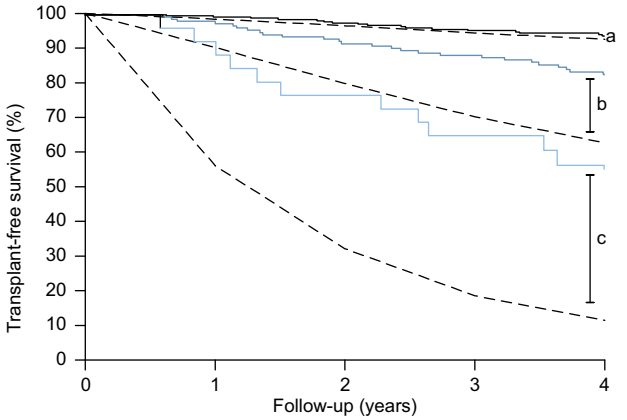


Fig. 5. Predicted versus observed liver transplant-free survival according to Mayo risk score risk group. Fig. shows prediction accuracy of the Mayo risk score (MRS) by comparing predicted survival based on the Mayo risk score and observed (actual) survival by Kaplan-Meier estimates at primary sclerosing cholangitis diagnosis according to Mayo risk score risk groups. Solid line: observed survival by Kaplan-Meier estimates. Dashed line: predicted survival by the Mayo risk score. [a]low-risk group (Mayo risk score value below or equal to 0); [b]intermediate risk group (scores above 0 but less than 2); [c]high risk group (Mayo risk score greater than 2).

1 year following diagnosis (n = 482) showed higher C-statistics for the MRS than the AOM score (0.73 *vs.* 0.68 at diagnosis and 0.75 *vs.* 0.70 at 1 year following diagnosis, respectively; Table 3).

In terms of prediction accuracy (calibration) the MRS overestimated transplant-free survival with a difference between the calculated mean survival based on MRS scores and actual survival as observed by Kaplan-Meier estimates of 5.1% at 1 year, 6.9% at 2 years, 8.9% at 3 years, and 9.6% at 4 years of follow-up. Detailed analyses of prediction accuracy in different risk groups revealed that the difference between predicted and observed survival was most pronounced in the high-risk group (Fig. 5).

Discussion

This long-term study of a well-characterized large cohort of patients with PSC allowed us to assess the performance and utility of the AOM, a prognostic model to estimate survival for patients with PSC. We confirm that the AOM has adequate discriminatory performance and satisfactory prediction accuracy when applied at PSC diagnosis. In addition, we show that the performance and accurate 5-year prediction of the AOM score remained if recalculated at different timepoints during follow-up, thereby extending its utility in daily clinical practice. If dichotomized, an AOM score of 2.0 had the highest discriminative power to risk stratify patients during follow-up. In patients who remained alive without liver transplantation during the observation period in our study, the AOM stayed, on average, below this threshold. Although the MRS had a higher C-statistic than the AOM, the utility of this traditional prognostic tool for the individual patient may be limited due to the suboptimal prediction accuracy. The results confirm that the AOM is a valuable prognostic tool in patients with PSC.

While our study was performed in tertiary referral centres, the baseline characteristics in our cohort are similar to those presented in the original AOM derivation and validation study.[16] In this population-based cohort, including patients from general hospitals and academic centres without transplant

Table 3. Direct comparison of the discriminatory performance of the Amsterdam-Oxford model and Mayo Risk Score calculated n years after diagnosis.

Year(s) after diagnosis	n	Amsterdam–Oxford model C–statistic (95% CI)	Mayo Risk Score C–statistic (95% CI)
0	498	0.6816 (0.6482–0.7149)	0.7309 (0.7005–0.7613)
1	482	0.7008 (0.6685–0.7332)	0.7507 (0.7200–0.7813)
2	446	0.7334 (0.7027–0.7641)	0.7696 (0.7391–0.8001)
3	402	0.7589 (0.7291–0.7887)	0.7800 (0.7501–0.8099)
4	370	0.7592 (0.7257–0.7927)	0.7709 (0.7389–0.8029)
5	337	0.7566 (0.7199–0.7933)	0.7902 (0.7567–0.8237)

facilities, De Vries et al. reported 10-year transplant-free survival rates between 75–80% and a median survival of 22 years. In comparison, in our cohort the 10-year transplant-free survival was 66% and the median survival was 13 years. The lower event-rate in the AOM development study may explain the AOMs slight underestimation of clinical events beyond 5 years of follow-up in the total cohort as well as in the subgroup with intermediate AOM scores (20th-80th percentile). As expected, the difference between predicted and observed survival was most pronounced in the subgroup of patients with highest AOM scores (>80th percentile). Nonetheless, the C-statistic we found is comparable to that reported in the AOM development study (0.68).

In terms of discriminative performance, the MRS outperformed the AOM in our cohort, meaning that with the MRS more patients who experienced an event had a higher risk score and more patients without an event had a lower risk score than with the use of the AOM. However, prediction accuracy of the MRS was unsatisfactory as the score substantially overestimated the risk of liver transplantation or death. A potential explanation for the limited prediction accuracy may be the indirect clinical endpoint that was used in the development cohort of the MRS. Patients who underwent liver transplantation were considered to have died with time to death based on the expected survival in absence of transplantation. This is probably no longer accurate today. An alternative explanation may be the use of time of referral in the derivation of the MRS, which may be long after the date of diagnosis as used in our cohort. Also, the MRS was developed in specific expert centres in which it is likely that patients present with more advanced disease, especially at the time of referral. The higher event-rate in the MRS development cohort (the 5-year transplant-free survival was 65% in the MRS development cohort *vs.* 84% in our cohort) could thus contribute to the overestimation of events by the MRS in the current study, as well as in many centres managing PSC.

As the AOM and MRS have different prognostic qualities, it is difficult to determine which score should be used in clinical practice and how and when a score should be applied. From a clinical point of view, it is difficult to derive certainty from a single long-term estimate of risk, especially in PSC. Rather clinicians as well as patients are likely to want to re-evaluate the risk of adverse events during follow-up. As such, accurate short-term or intermediate-term calculations of risk may be better suited. We show the AOM can be used to make such repeated estimates for patients in different categories of risk. With the inclusion of variceal bleeding – a direct clinical complication of end-stage liver disease – the MRS has a high discriminative performance across different cohorts. With a current 30-day mortality of 15–20%, variceal bleeding is a strong predictor of death.[24,25] In our cohort, the MRS provides a more discriminative short-term mortality risk assessment as opposed to the AOM, albeit less accurate. This score may thus be more appropriate to estimate whether patients are at high risk of progressive disease necessitating LT on a group level. In daily practice, however, clinicians may prefer the accurate estimates of prognosis as provided by the AOM at various timepoints in order to optimize the management of individual patients.

Important to consider is that the AOM was developed in an early disease stage population and that it includes ALP. ALP is elevated early during the course of disease,[26,27] has often been used in drug development trials as a primary endpoint, and lower ALP levels have been correlated with a favourable course of disease.[28–33] In fact, a previous study in 366 patients with PSC showed that a more simplistic approach using ALP in isolation had a near identical c-statistic to that of the AOM, challenging the necessity of a complicated risk prediction model.[33] However, absolute predictions of transplant-free survival in our cohort using ALP in isolation, revealed that using ALP alone may grossly overestimate survival in different risk groups at diagnosis and when reapplied during follow-up. Moreover, the C-statistic varied between 0.52 and 0.63 during follow-up, which is clearly lower than that observed for the AOM. Therefore, the use of a prognostic score that includes more biochemical variables than ALP, especially when reapplied during follow-up and/or in a tertiary cohort, seems more appropriate.

A strength of our study is the inclusion of a well-characterized study population from multiple centres with complete follow-up in 80% of the patients. Second, we assessed a combined endpoint of LT and all-cause mortality rather than PSC-related mortality specifically. In clinical practice it is difficult to distinguish true liver-related or PSC-related death from other causes of death. In addition, one may argue that the separation between the two has limited relevance for patients and clinicians. Inherit to the retrospective nature of our study some laboratory values were missing. To overcome this problem multiple imputation techniques were used.[19] Importantly, analyses in our raw dataset revealed similar results (data not shown). While our results further validate and justify the clinical use of the AOM, the validation mainly pertains to Caucasian patients treated in tertiary centres. Further validation in a population-based setting, in particular in other ethnicities or countries, is thus warranted. Finally, the threshold of 2.0 found for the AOM should be interpreted with caution. The use of thresholds in clinical practice is widespread but has several limitations. Categorization results in loss of predictive information and thresholds are difficult to generalize to other populations.[34,35] This study shows that with dichotomous application of the AOM, patients may switch in risk category during follow-up. This is actually a limitation of dichotomous criteria/risk group thresholds in general. However, the AOM was not derived to be used as a dichotomous score but rather as a continuous measure of risk. The analyses in our manuscript therefore only visualize that 2 risk groups can be readily obtained by using the threshold of 2.0, akin to the thresholds used in the MRS.

In general, as models with a C-statistic >0.8 are considered good prognostic models, further optimization of the risk stratification in PSC remains warranted.[22] In PSC, utility of prediction models is hampered due to heterogeneity in disease progression and outcomes, as well as the lack of effective therapies. Still, reliable estimates of survival are important for patient counselling, optimization of follow-up regimens, and selection and timing of listing for LT. Once effective therapies become available, repositioning of prediction models in clinical management of patients with PSC may be necessary. Cox-regression analyses, with the variables comprising the AOM while offsetting the AOM score (*i.e.* keeping its value equal to 1), revealed a satisfactory fit at PSC diagnosis. Only the βs for age and platelet count were suboptimal for our cohort, but adjustments of these βs only yielded a minor increase in terms of C-statistic (0.67 *vs.* 0.68). Therefore, other approaches, with for example, the addition of measures of liver fibrosis could be of value. A potentially important addition to existing risk stratification models in PSC may be the inclusion of liver stiffness measurement (LSM). LSM reflects severity of fibrosis and absolute LSM values as well as longitudinal changes are strongly linked with clinical events in PSC.[36] Enhanced liver fibrosis (ELF™) score, a serological measure of liver fibrosis consisting of a combination of serum concentrations of hyaluronic acid, procollagen III peptide and tissue inhibitor of metalloproteinase 1, could provide another addition to existing risk prediction models. The ELF score has been shown to correlate well with LSM values and to have incremental prognostic utility alongside the Mayo risk score.[37] In order to better assess the additive value of such measures of liver fibrosis to existing risk prediction models, large prospective multicentre collaborations with extensive datasets are necessary. New statistical techniques could further aid in the derivation of more accurate models as well. A recently introduced model by Eaton et al., using a different statistical approach with machine learning, showed an impressive accuracy (C-statistic >0.9) for the prediction of endpoints.[38] This score is, however, limited by the prediction of hepatic decompensation rather than solid clinical endpoints such as LT, death or CCA. Furthermore, further validation of this score as well as the statistical technique should be awaited. Still, PSC remains a disease with a highly variable course that may not be easily captured by static prognostic scores.

In conclusion, we confirm the AOM has adequate discriminatory performance and good predictive accuracy for LT-free survival, both at PSC diagnosis and other timepoints during the course of disease. Hereby we demonstrate the validity and extended utility of the AOM as a prognostic tool in PSC.

Financial support

This investigator-initiated study was funded by the Foundation for Liver and Gastrointestinal Research (a not-for-profit foundation) in Rotterdam, the Netherlands. The supporting parties had no influence on the study design, data collection and analyses, writing of the manuscript, or on the decision to submit the manuscript for publication.

Conflicts of interest

Jorn C. Goet has nothing to declare. Annarosa Floreani has acted as advisor in the PBC Committee sponsored by Intercept. Xavier Verhelst served as a consultant for Intercept Pharmaceuticals, Gilead, MSD and Abbvie and received speaker's fees from Gilead, Bayer, MSD and Abbvie. Nora Cazzagon is consultant for Intercept Pharmaceuticals. Lisa Perini has nothing to declare. Willem J. Lammers is consultant for Intercept Pharmaceuticals. Annemarie C. de Vries received unrestricted grants from De Maag Lever Darm Stichting (MLDS) and Tramedico, and is consultant for Janssen Pharmaceutica, Takeda and Abbvie. Adriaan J. van der Meer received unrestricted grant from Gilead, speakers fee from Gilead, Zambon and AbbVie. Henk R. van Buuren received unrestricted grants from Intercept Pharmaceuticals. Bettina E. Hansen received unrestricted grants from and is consultant for Intercept Pharmaceuticals and is a consultant for Cymabay, Albireo and Janssen Pharma.

Please refer to the accompanying ICMJE disclosure forms for further details.

Authors' contributions

Jorn C. Goet and Bettina E. Hansen had full access to all data in the study and take responsibility for the integrity of the data and the accuracy of data analyses. Study concept and design: Jorn C. Goet, Annarosa Floreani, Xavier Verhelst, Nora Cazzagon, Lisa Perini, Willem J. Lammers, A.J. van der Meer, Annemarie C. de Vries, Henk R. van Buuren, Bettina E Hansen. Acquisition of data: Jorn C. Goet, Annarosa Floreani, Xavier Verhelst, Nora Cazzagon, Lisa Perini. Analysis and interpretation of data: Jorn C. Goet, Bettina E. Hansen and Henk R. van Buuren. Drafting of the manuscript: Jorn C. Goet, Henk R. van Buuren, Bettina E Hansen. Critical revision of the manuscript: Jorn C. Goet, A.J. van der Meer, Annarosa Floreani, Xavier Verhelst, Nora Cazzagon, Lisa Perini, Willem J. Lammers, Annemarie C. de Vries, Henk R. van Buuren, Bettina E Hansen. Statistical analysis: Jorn C. Goet, Bettina E. Hansen. Obtained funding: Bettina E. Hansen, Henk R. van Buuren. Study supervision: Jorn C. Goet, A.J. van der Meer, Henk R. van Buuren, Bettina E Hansen

Acknowledgements

The authors would like to express their gratitude to the students at the Department of Surgery, Oncology and Gastroenterology of the University of Padua for their assistance in data collection and translation of medical records for the current study, in specific: Nicola Perin, Alessandra Zago, Elena Salami, Giovanni Leardini, Chiara Mangini, Francesca Trentin and Filippo Simionato.

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhep.2019.06.012.

References

[1] Karlsen TH, Folseraas T, Thorburn D, Vesterhus M. Primary sclerosing cholangitis – a comprehensive review. J Hepatol 2017;67:1298–1323.
[2] Chapman RW, Arborgh BA, Rhodes JM, et al. Primary sclerosing cholangitis: a review of its clinical features, cholangiography, and hepatic histology. Gut 1980;21:870–877.
[3] Hirschfield GM, Karlsen TH, Lindor KD, Adams DH. Primary sclerosing cholangitis. Lancet 2013;382:1587–1599.
[4] Williamson KD, Chapman RW. Primary sclerosing cholangitis. Dig Dis 2014;32:438–445.
[5] Boonstra K, Beuers U, Ponsioen CY. Epidemiology of primary sclerosing cholangitis and primary biliary cirrhosis: a systematic review. J Hepatol 2012;56:1181–1188.

[6] Lindor KD. Ursodiol for primary sclerosing cholangitis. Mayo Primary Sclerosing Cholangitis-Ursodeoxycholic Acid Study Group. N Engl J Med 1997;336:691–695.
[7] Lindor KD, Kowdley KV, Luketic VA, Harrison ME, McCashland T, Befeler AS, et al. High-dose ursodeoxycholic acid for the treatment of primary sclerosing cholangitis. Hepatology 2009;50:808–814.
[8] Beuers U, Spengler U, Kruis W, Aydemir U, Wiebecke B, Heldwein W, et al. Ursodeoxycholic acid for treatment of primary sclerosing cholangitis: a placebo-controlled trial. Hepatology 1992;16:707–714.
[9] Stiehl A, Walker S, Stiehl L, Rudolph G, Hofmann WJ, Theilmann L. Effect of ursodeoxycholic acid on liver and bile duct disease in primary sclerosing cholangitis. A 3-year pilot study with a placebo-controlled study period. J Hepatol 1994;20:57–64.
[10] Chapman RW. Medical treatment of primary sclerosing cholangitis with ursodeoxycholic acid. Dig Liver Dis 2003;35:306–308.
[11] Olsson R, Boberg KM, de Muckadell OS, Lindgren S, Hultcrantz R, Folvik G, et al. High-dose ursodeoxycholic acid in primary sclerosing cholangitis: a 5-year multicenter, randomized, controlled study. Gastroenterology 2005;129:1464–1472.
[12] Bergquist A, Ekbom A, Olsson R, Kornfeldt D, Lööf L, Danielsson A, et al. Hepatic and extrahepatic malignancies in primary sclerosing cholangitis. J Hepatol 2002;36:321–327.
[13] Claessen MM, Vleggaar FP, Tytgat KM, Siersema PD, van Buuren HR. High lifetime risk of cancer in primary sclerosing cholangitis. J Hepatol 2009;50:158–164.
[14] Boonstra K, Weersma RK, van Erpecum KJ, Rauws EA, Spanier BW, Poen AC, et al. EpiPSCPBC Study Group. Population-based epidemiology, malignancy risk, and outcome of primary sclerosing cholangitis. Hepatology 2013;58:2045–2055.
[15] Kim WR, Therneau TM, Wiesner RH, Poterucha JJ, Benson JT, Malinchoc M, et al. A revised natural history model for primary sclerosing cholangitis. Mayo Clin Proc 2000;75:688–694.
[16] de Vries EM, Wang J, Williamson KD, Leeflang MM, Boonstra K, Weersma RK, et al. A novel prognostic model for transplant-free survival in primary sclerosing cholangitis. Gut 2017.
[17] European Association for the Study of the Liver. EASL Clinical Practice Guidelines: management of cholestatic liver diseases. J Hepatol 2009;51:237–267.
[18] Little RJA, Rubin DB. Statistical analysis with missing data. New York: John Wiley & Sons; 1987.
[19] Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ 2009;338 b2393.
[20] Rubin D. Multiple imputation aft er 18+ years. J Am Stat Assoc 1996;91:473–489.
[21] van Houwelingen HC. Validation, calibration, revision and combination of prognostic survival models. Stat Med 2000;19:3401–3415.
[22] Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. BMC Med Res Method 2013;13:33.
[23] Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology 2010;21:128–138.
[24] Graham DY, Smith JL. The course of patients after variceal hemorrhage. Gastroenterology 1981;80:800–809.
[25] D'Amico G, De Franchis R. Upper digestive bleeding in cirrhosis. Post-therapeutic outcome and prognostic indicators. Hepatology 2003;38:599–612.
[26] Brensilver HL, Kaplan MM. Significance of elevated liver alkaline phosphatase in serum. Gastroenterology 1975;68:1556–1562.
[27] Kaplan MM, Righetti A. Induction of rat liver alkaline phosphatase: the mechanism of the serum elevation in bile duct obstruction. J Clin Invest 1970;49:508–516.
[28] Stanich PP, Bjornsson E, Gossard AA, Enders F, Jorgensen R, Lindor KD. Alkaline phosphatase normalization is associated with better prognosis in primary sclerosing cholangitis. Dig Liver Dis 2011;43:309–313.
[29] Al Mamari S, Djordjevic J, Halliday JS, Chapman RW. Improvement of serum alkaline phosphatase to <1.5 upper limit of normal predicts better outcome and reduced risk of cholangiocarcinoma in primary sclerosing cholangitis. J Hepatol 2013;58:329–334.
[30] Lindstrom L, Hultcrantz R, Boberg KM, Friis-Liby I, Bergquist A. Association between reduced levels of alkaline phosphatase and survival times of patients with primary sclerosing cholangitis. Clin Gastroenterol Hepatol 2013;11:841–846.
[31] Ponsioen CY, Chapman RW, Chazouilleres O, Hirschfield GM, Karlsen TH, Lohse AW, et al. Surrogate endpoints for clinical trials in primary sclerosing cholangitis: Review and results from an International PSC Study Group consensus process. Hepatology 2016;63:1357–1367.
[32] Rupp C, Rossler A, Halibasic E, Sauer P, Weiss KH, Friedrich K, et al. Reduction in alkaline phosphatase is associated with longer survival in primary sclerosing cholangitis, independent of dominant stenosis. Aliment Pharmacol Ther 2014;40:1292–1301.
[33] de Vries EM, Wang J, Leeflang MM, Boonstra K, Weersma RK, Beuers UH, et al. Alkaline phosphatase at diagnosis of primary sclerosing cholangitis and one year later: evaluation of prognostic value. Liver Int 2016.
[34] Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. J Natl Cancer Inst 1994;86:829–835.
[35] Buettner P, Garbe C, Guggenmoos-Holzmann I. Problems in defining cutoff points of continuous prognostic factors: example of tumor thickness in primary cutaneous melanoma. J Clin Epidemiol 1997;50:1201–1210.
[36] Corpechot C, Gaouar F, El Naggar A, Kemgang A, Wendum D, Poupon R, et al. Baseline values and changes in liver stiffness measured by transient elastography are associated with severity of fibrosis and outcomes of patients with primary sclerosing cholangitis. Gastroenterology 2014;146:970–979, quiz e915–976.
[37] Vesterhus M, Hov JR, Holm A, Schrumpf E, Nygård S, Godang K, et al. Enhanced liver fibrosis score predicts transplant-free survival in primary sclerosing cholangitis. Hepatology 2015;62:188–197.
[38] Eaton JE, Vesterhus M, McCauley BM, Atkinson EJ, Schlicht EM, Juran BD, et al. Primary Sclerosing Cholangitis Risk Estimate Tool (PREsTo) predicts outcomes of the disease: a derivation and validation study using machine learning. Hepatology 2018.