Editorial

JOURNAL OF HEPATOLOGY

Risk stratification in primary sclerosing cholangitis: It's time to move on from replicating imperfection and break the glass ceiling

Palak J. Trivedi[1,2,3,4,⁎]

[1]National Institute for Health Research (NIHR) Birmingham Biomedical Research Centre, Centre for Liver and Gastroenterology Research, University of Birmingham, UK; [2]Institute of Immunology and Immunotherapy, University of Birmingham, UK; [3]Institute of Applied Health Research, University of Birmingham, UK; [4]Liver Unit, University Hospitals Birmingham NHS Foundation Trust Queen Elizabeth, UK

See Article, pages 992–999

Primary sclerosing cholangitis (PSC) represents the greatest unmet need in modern hepatology, given its ill-defined aetiology, critical absence of medical therapy, and the fact that liver transplantation remains the only life-saving intervention for patients. Although rare, PSC now accounts for 10–15% of all liver transplant activity in European liver transplant programmes, and is now the lead indication for transplantation in Nordic countries.[1,2] However, rates of progression vary, and accurately predicting the disease course is of relevance to clinical practice and interventional trial design.[3] Patient expectations are also rising, with a feeling that doctors must be able to tell them if they are at risk, if so in what way, and with a reasonable degree of confidence.[4]

To this effect, several attempts to construct 'the ideal' prognostic model have been made, each attributing different weights to clinical, laboratory, cholangiographic or histological variables.[3] Historic algorithms such as the Mayo PSC risk score (MRS),[5,6] derive mostly from tertiary referral centres, and predominantly liver transplant units. Although widely applied, these scores lose predictive accuracy beyond 4–5 years from the point of application. Moreover, the era in which they were conceived predates the modern management of variceal bleeding. In a similar vein, cholangiography-based systems rely on biliary imaging by endoscopic retrograde cholangiopancreatography, which is not standard of care for monitoring in PSC.[7,8]

More recently, intensive efforts have been conducted at a multicentre level to model the disease's natural history using contemporary patient cohort data.[9–11] Akin to older prognostic scores before it, the Amsterdam Oxford model (AOM) is composed of mainly laboratory parameters, each demonstrating stratification properties in their own right. Of note, only 7 variables were chosen during derivation of the AOM (out of a total 13 which showed predictive utility). In the original study, the authors state that this was to limit the number of covariates present in the model, to ease use in clinical practice, whilst yielding an overall model concordance (C)-statistic that was no more than 10% below the optimal reading possible. In turn, the way in which covariates were selected for the model was based on the rank of their individual C-statistic values. This approach is somewhat questionable, and less sensitive than selecting covariates based on likelihood, risk, or indeed their individual calibration accuracy.[12]

Within the final AOM, the chosen covariates were disease phenotype (large duct *vs.* small duct PSC), serum aminotransferases and alkaline phosphatase (ALP), together with biomarkers of more advanced liver disease such as bilirubin, albumin and total platelet count.[10] The landmark study from which the model derives, presents up to 30 years of patient follow-up data. Predicted event rates according to the AOM closely mirrored actuarial survival estimated by the Kaplan-Meier method (calibration accuracy). However, the chosen endpoint of PSC-related death included colorectal cancer mortality (in addition to liver transplantation), which is contentious given that all covariates in the model relate to hepatobiliary disease. Discriminant utility of the AOM was fair, as evidenced by a C-statistic of 0.68 and accompanying wide confidence intervals (95% CI 0.51–0.85). C-statistics were similar when the model was applied at diagnosis, and annually up to 3 years thereafter. No sub-stratification according to disease stage or variant clinical phenotypes was conducted.

In the current issue of *Journal of Hepatology*, Goet *et al.* present results of a very impressive prediction model study, using external patient data to validate the AOM in PSC.[13] The studied cohort comprised 534 patients, the vast majority having classical large duct disease (87%) and receiving ursodeoxycholic acid therapy (92%). Notable differences to the original AOM study, were the fact that patients were all diagnosed at 1 of 3 liver transplant units (as opposed to a predominant population-based cohort), with the selected endpoint being liver transplantation and all-cause mortality.[10,13]

The paper replicates several key findings but, most importantly, provides transparency surrounding the utility and limitations of the model. Firstly, discriminant performance was near identical in this validation exercise as in the original AOM study (C-statistic: 0.67 at baseline; 95% CI 0.64–0.70), but it was seen

Keywords: Amsterdam Oxford PSC score; Biomarker; Clinical trial; Endpoint; PREsTo; Prognostic model; UK PSC risk score.

Received 7 August 2019; accepted 18 August 2019

☆ DOI of original article: http://dx.doi.org/10.1016/j.jhep.2019.06.012.

⁎ Address: National Institute for Health Research (NIHR) Birmingham Biomedical Research Centre, Centre for Liver and Gastroenterology Research, University of Birmingham, Birmingham, UK.

E-mail address: p.j.trivedi@bham.ac.uk.

to improve when applied at 5 years following diagnosis (0.75; 95% CI 0.71–0.78). The authors then go on to show good calibration accuracy of the model, and actually quantify differences between observed versus expected clinical events for the overall cohort. However, Fig. S1 provides a major take-home message.[13] When testing the model over different time points, calibration accuracy was greatest for patients in the lower percentile risk groups (<20th); whereas the incidence of clinical events was underestimated in mid- and high-percentile scorers, particularly when applying the AOM at 3- and 5-years following the date of PSC diagnosis. A direct comparison of the AOM versus MRS was also provided, something that was lacking in the original AOM study. The MRS exhibited greater discriminatory value (C-statistic: 0.73, 95% CI 0.73–0.76 at baseline; 0.79, 95% CI 0.76–0.82 at 5 years), but overestimated the risk of future clinical events long-term. This trade-off between discrimination versus calibration accuracy is recognised in prognostic modelling, particularly if one 'assumes' that distribution of disease risk is uniform across patient populations, when it may not be.[12] Nevertheless, in the related cholestatic disorder primary biliary cholangitis, prognostic models demonstrate both high-level discrimination as well as calibration accuracy.[14–17] This difference may relate to the fact serum ALP exhibits wider intra-individual variability between time-points in PSC, which is likely to impact the performance of any ALP-based stratification system such as the AMS.[18] Notably, the current publication points toward significant limitations to serum ALP as a biomarker, as well as raising questions surrounding its utility as a surrogate endpoint in PSC clinical trials. Firstly, the discriminant value of ALP was marginal at best (C-statistic ranging

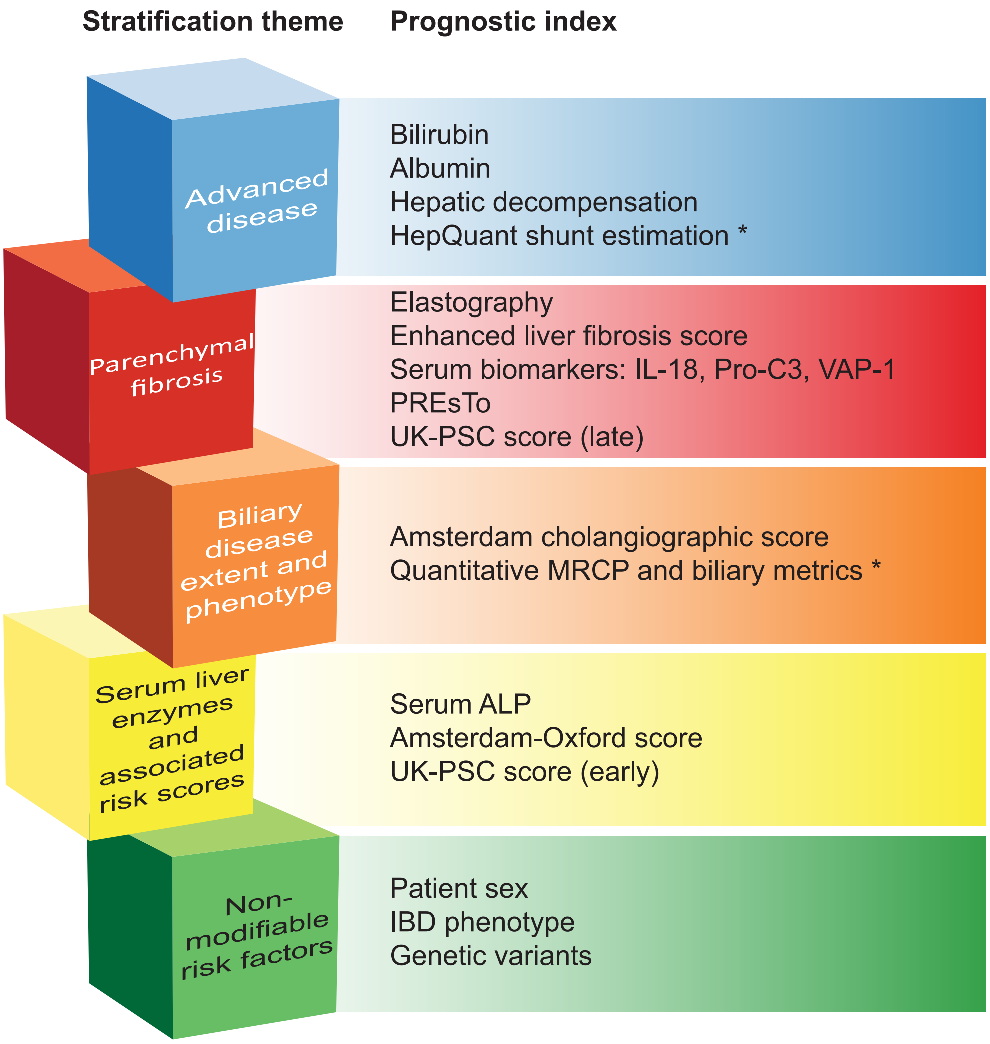
Stratification theme
Prognostic index
Advanced disease
Bilirubin
Albumin
Hepatic decompensation
HepQuant shunt estimation *
Parenchymal fibrosis
Elastography
Enhanced liver fibrosis score
Serum biomarkers: IL-18, Pro-C3, VAP-1
PREsTo
UK-PSC score (late)
Biliary disease extent and phenotype
Amsterdam cholangiographic score
Quantitative MRCP and biliary metrics *
Serum liver enzymes and associated risk scores
Serum ALP
Amsterdam-Oxford score
UK-PSC score (early)
Non-modifiable risk factors
Patient sex
IBD phenotype
Genetic variants

Fig. 1. Stratifying the stratifiers: a hypothetical approach to applying prognostic tools. In PSC, several predictive indices, biomarkers and prognostic models have been created, for which a hierarchical ranking based on disease stage is proposed. Whilst non-modifiable patient factors (green), such as sex and inflammatory bowel disease phenotype are proven in large observational cohorts, their predictive utility is seemingly lost during the validation of biochemical risk stratifiers (yellow), such as those presented in the current study. In a similar vein, cholangiographic methods have the potential to track biliary disease progression more readily (orange). However, their prognostic utility is likely attenuated once liver fibrosis has reached a certain stage, which is evident by the underestimation of clinical event rates in high-risk groups. At a certain point, surrogate biomarkers of fibrosis such as the enhanced liver fibrosis score (ELF^{TM}) and transient elastography, would become more meaningful (red). The critical challenge in PSC as a disease is identifying at what ELF score or elastography measure do earlier stratification tools become superseded? Regardless, the onset of advanced liver disease with features of hepatic decompensation, persistently elevated bilirubin, or hypoalbuminaemia, is more immediately linked to the development of clinical events (blue), following which the utility of any previous prognostic tool becomes moot. *Asterisks denote emerging stratification tools with potential, which have not yet been proven in PSC specifically. ALP, alkaline phosphatase; IBD, inflammatory bowel disease; MRCP, magnetic resonance cholangiopancreatography; PSC, primary sclerosing cholangitis.

between 0.52–0.63 during the first 5 years following diagnosis). Furthermore, assessment of ALP calibration revealed very large differences between observed and predicted survival rates.

The authors also stratified patients according to high- versus low-risk groups, following a grid search to identify the most discriminatory AOM cut-off point (<2.0 *vs.* ≥2.0). In so doing, they identify that 8.4%, 13.9% and 25.4% of patients who were identified as being low-risk at diagnosis actually move into a high-risk category when the model is applied at 1-, 3- and 5-years, respectively, indicating the progressive nature of PSC as a disease and a big caveat if applying the model to counsel patients in clinic.

A head-to-head comparison between the AOM and other contemporaneous models, such as the UK-PSC score and the PREsTo index, has yet to be conducted.[9,11] For a fair evaluation, the performance of each would need to be tested simultaneously in the same population. When assessing discriminant utility, factors such as censoring distribution and intra-predictor correlation also need to be considered, and study endpoints consistent to allow comparability. Of note, all existing models examine the cumulative incidence of events at specific time points; however, prognostic factors likely differ for endpoints developing in the first few years after diagnosis to those which manifest a decade later. Yet by current methodology, both early and late events are counted when calculating overall 10-year event-free survival. Perhaps a more precise method would be to identify predictors of clinical events occurring at specific intervals; for instance, from diagnosis up to 2 years, 2–5 years, and 5–10-years. A further drawback across all risk models, which has been identified by patient focus groups in the UK, is that they rely on the date of diagnosis being known and accurate. In reality, however, individuals may experience years from first presentation to the moment they are diagnosed. Therefore, patients request that dynamic biomarkers or prediction models be developed, which can be applied at any point and irrespective of the date they are told they have PSC.[4]

In any event, the development and validation of new risk scores represent a major advance for risk stratification in PSC. The question that follows is: "how and when to use them?" To a patient, it is more meaningful to know what the probability of a clinical event occurring is over a given time period. In this case the AOM is well placed, given the validation in calibration accuracy across 2 high quality studies.[10,13] However, it is important to re-calculate the score and update patients about their risk profile over sequential clinical visits, given that 25% of patients classified as low risk become high risk over 5 years. In turn, when stratifying patients toward clinical trials, scores with greater discriminant utility may be more appropriate.[9,11]

Efforts toward prognostic modelling continue to be refined, but there remains a need to differentiate variables associated with early-yet-rapidly progressive disease from that which is already advanced. Discriminant utility and accuracy of existing tools appear to have reached a ceiling, so perhaps a hierarchical approach to risk stratification is now needed, rather than repeated permutations and combinations of routinely available laboratory parameters to heterogeneous groups of patients. It is also plausible that different variables are relevant at distinct disease stages. For instance, measures of biliary disease involvement are likely to be more relevant early on in the disease process and help to predict longer-term outcomes. Reciprocally, tools that measure the extent of parenchymal disease are more likely to predict clinical events that develop more immediately (Fig. 1). The critical challenge in PSC lies in its heterogeneity, varying phenotypic presentations, and identifying the juncture at which early prognostic markers become redundant, and overridden by those directly linked to the burden of liver fibrosis.[19–20]

Financial support

PJT receives salary support from the NIHR Birmingham Biomedical Research Centre. This paper presents independent research supported by the NIHR Birmingham Biomedical Research Centre at the University Hospitals Birmingham NHS Foundation Trust and the University of Birmingham. The views expressed are those of the author and not necessarily those of the NHS, the NIHR or the Department of Health.

Conflict of interest

The authors declare no conflicts of interest that pertain to this work.

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

Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhep.2019.08.013.

References

Author names in bold designate shared co-first authorship

[1] Fosby B, Melum E, Bjøro K, Bennet W, Rasmussen A, Andersen IM, et al. Liver transplantation in the Nordic countries - An intention to treat and post-transplant analysis from The Nordic Liver Transplant Registry 1982–2013. Scand J Gastroenterol 2015;50:797–808. https://doi.org/10.3109/00365521.2015.1036359.

[2] Organ Donation and Transplantation Activity Report 2017 / 18. NHSBT 2018

[3] Trivedi PJ, Corpechot C, Pares A, Hirschfield GM. Risk stratification in autoimmune cholestatic liver diseases: opportunities for clinicians and trialists. Hepatology 2016;63:644–659. https://doi.org/10.1002/hep.28128.

[4] Walmsley M, Leburgue A, Thorburn D, Hirschfield G, Trivedi P. Identifying research priorities in primary sclerosing cholangitis: driving clinically meaningful change from the patients' perspective. J Hepatol 2019;70:e412–e413. https://doi.org/10.1016/S0618-8278(19)30812-6.

[5] Wiesner RH, Grambsch PM, Dickson ER, Ludwig J, MacCarty RL, Hunter EB, et al. Primary sclerosing cholangitis: natural history, prognostic factors and survival analysis. Hepatology 1989;10:430–436.

[6] 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. https://doi.org/10.4065/75.7.688.

[7] Tischendorf JJW, Hecker H, Krüger M, Manns MP, Meier PN. Characterization, outcome, and prognosis in 273 patients with primary sclerosing cholangitis: a single center study. Am J Gastroenterol 2007;102:107–114. https://doi.org/10.1111/j.1572-0241.2006.00872.x.

[8] Ponsioen CY, Vrouenraets SME, Prawirodirdjo W, Rajaram R, Rauws EAJ, Mulder CJJ, et al. Natural history of primary sclerosing cholangitis and prognostic value of cholangiography in a Dutch population. Gut 2002;51:562–566.

[9] Eaton JE, Vesterhus M, McCauley BM, Atkinson EJ, Schlicht EM, Juran BD, et al. Primary sclerosing cholangitis risk estimate tool (PREsTo) predicts outcomes in PSC: a derivation & validation study using machine learning. Hepatology 2018. https://doi.org/10.1002/hep.30085.

[10] **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 gutjnl-2016-313681. Gut 2017. https://doi.org/10.1136/gutjnl-2016-313681.

[11] Goode EC, Clark AB, Mells GM, Srivastava B, Spiess K, Gelson WTH, et al. Factors associated with outcomes of patients with primary sclerosing cholangitis and development and validation of a risk scoring system. Hepatology 2018. https://doi.org/10.1002/hep.30479.

[12] Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 2007;115:928–935. https://doi.org/10.1161/CIRCULATIONAHA. 106.672402.

[13] Goet JC, Floreani A, Verhelst X, Cazzagon N, Perini L, Lammers WJ, et al. Validation, clinical utility and limitations of the amsterdam-oxford model for primary sclerosing cholangitis. J Hepatol 2019;71:992–999. https://doi.org/10.1016/j.jhep.2019.06.012.

[14] Carbone M, Sharp SJ, Flack S, Paximadas D, Spiess K, Adgey C, et al. The UK-PBC risk scores: derivation and validation of a scoring system for long-term prediction of end-stage liver disease in primary biliary cirrhosis. Hepatology 2015. https://doi.org/10.1002/hep.28017.

[15] Cheung AC, Gulamhusein AF, Juran BD, Schlicht EM, McCauley BM, de Andrade M, et al. External validation of the United Kingdom-primary biliary cholangitis risk scores of patients with primary biliary cholangitis treated with ursodeoxycholic acid. Hepatol Commun 2018;2:676–682. https://doi.org/10.1002/hep4.1186.

[16] Lammers WJ, Hirschfield GM, Corpechot C, Nevens F, Lindor KD, Janssen HLA, et al. Development and validation of a scoring system to predict outcomes of patients with primary biliary cirrhosis receiving ursodeoxycholic acid therapy. Gastroenterology 2015. https://doi.org/10.1053/j.gastro. 2015.07.061.

[17] Lammers WJ, van Buuren HR, Hirschfield GM, Janssen HLA, Invernizzi P, Mason AL, et al. Levels of alkaline phosphatase and bilirubin are surrogate end points of outcomes of patients with primary biliary cirrhosis: an international follow-up study. Gastroenterology 2014;147:1338–1349. https://doi.org/10.1053/j.gastro. 2014.08.029.

[18] Trivedi P, Muir A, Levy C, Bowlus C, Manns MP, Lu X, et al. Prospective evaluation of serum alkaline phosphatase variability and prognostic utility in primary sclerosing cholangitis using controlled clinical trial data. J Hepatol 2019;70:e12–e13. https://doi.org/10.1016/S0618-8278(19)30022-2.

[19] 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, doi:10. 1053/j.gastro.2013.12.030.

[20] **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. https://doi.org/10.1002/hep.27825.