

EASL-AASLD Delphi consensus statement on surrogate endpoints and real-world evidence in primary biliary cholangitis

European Association for the Study of the Liver^{*}, American Association for the Study of Liver Diseases^{*,†}

Summary

Drug development in primary biliary cholangitis (PBC) has led to the conditional approval of three second-line therapies, yet the transition to full regulatory approval remains challenging due to (1) the reliance on liver biochemistry and non-invasive measures of liver fibrosis as biomarkers of clinical outcomes, which are not accepted by regulatory authorities as validated efficacy endpoints; (2) the limited integration of real-world data and real-world evidence to complement clinical trials; and (3) the absence of harmonised and standardised use of patient-reported outcomes (PROs). To address these limitations, the European Association for the Study of the Liver (EASL) and the American Association for the Study of Liver Diseases (AASLD) commissioned an international, multi-stakeholder consensus panel to conduct a modified Delphi process to review current evidence, identify unmet research needs, and create a framework aligned with regulatory expectations to guide ongoing and future research in PBC. A total of 62 panellists, including clinicians, methodologists, regulators, industry partners, and patient representatives, were divided into three working groups and participated in regular online meetings, one in-person conference, and two online Delphi voting rounds. Response rates across the two voting sessions were 88% and 60%, respectively. The Delphi process resulted in agreement on 16 statements and 42 recommendations across the three thematic domains. Collectively, these consensus recommendations provide a comprehensive and pragmatic framework to support and expand the existing regulatory pathway for the development and approval of new therapies in PBC.

© 2026 European Association for the Study of the Liver and American Association for the Study of Liver Diseases. Published by Elsevier BV on behalf of European Association for the Study of the Liver and by Wolters Kluwer Health on behalf of American Association for the Study of Liver Diseases. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

Introduction

Primary biliary cholangitis (PBC) is a chronic cholestatic liver disease that can progress to cirrhosis and is marked by debilitating symptoms that substantially impact quality of life.^{1–3} The approval of ursodeoxycholic acid (UDCA) for PBC by the US FDA in 1997 marked the beginning of a shift in disease management, as routine treatment began to delay disease progression.^{4,5} It nevertheless took many years for confirmation and consensus on the benefit of UDCA for patients.^{6–8} However, while UDCA provides benefit across the disease spectrum, a substantial proportion of patients do not achieve an adequate biochemical response and remain at risk of disease progression and serious complications.^{6,9,10}

Furthermore, treatment with UDCA has no significant impact on symptoms. For those who do not respond to first-line treatment with UDCA or cannot tolerate it, add-on second-line therapies are available including obeticholic acid (OCA, Ocaliva[®]) and more recently the peroxisome proliferator-activated

receptor (PPAR) agonists seladelpar (Livdelzi[®]) and elafibranor (Iqirvo[®]). These drugs received conditional approval from the FDA and EMA based on improvements in biochemical markers, mainly alkaline phosphatase (ALP) and total bilirubin.^{11–17} These biomarkers have demonstrated consistent associations with clinical outcomes across multiple large-scale international cohorts comprising thousands of patients, with no evidence to date contradicting their prognostic value. However, the supporting data are derived predominantly from retrospective analyses of tertiary referral populations. Consequently, regulatory agencies consider ALP and bilirubin as reasonably likely to predict clinical benefit but not yet validated surrogate endpoints; thus, confirmatory trials are required for full marketing authorisation.^{18–20} Notably, this approach does not take into consideration the impact a drug may have on symptoms: the ability to demonstrate a change in how the patient “functions” or “feels” can also lead to regulatory drug approval but is not currently explored due to existing limitations in our ability to

^{*} Corresponding authors. Addresses: E-mail address: easloffice@easloffice.eu European Association for the Study of the Liver. The EASL Building – Home of Hepatology, 7 rue Daubin, CH 1203, Geneva, Switzerland. Tel.: +41 (0) 22 807 03 60 or E-mail address: aasld@aasld.org American Association for the Study of Liver Diseases, 1001 North Fairfax Street 4th floor Alexandria, Virginia 22314, United States. Tel.: 703-299-9766. This paper was jointly developed by Journal of Hepatology, Hepatology and jointly published by Elsevier BV and Wolters Kluwer Health. The articles are identical except for minor stylistic and spelling differences in keeping with each journal's style. Either citation can be used when citing this article.

[†] Full list of collaborators provided in the [Appendix](#).
<https://doi.org/10.1016/j.jhep.2026.03.038>



implement and interpret results of data involving patient-reported outcomes (PROs). Alongside these conditionally approved agents, bezafibrate, a pan-PPAR agonist which is not currently licensed for the treatment of PBC, is widely used off label as add-on therapy, supported by both clinical trial data and real-world evidence (RWE) of safety and biochemical efficacy.^{21–23} Its widespread use in routine practice, despite the absence of formal regulatory approval, reflects the limitations of the current licensing framework in PBC. Similarly, fenofibrate, a PPAR- α agonist, is also used off label as add-on therapy, especially in countries where bezafibrate is not available (e.g. the United States).

Designing a confirmatory randomised controlled trial (RCT) in PBC presents significant challenges. Once a therapy receives conditional approval and becomes commercially available, recruitment and retention become increasingly difficult: patients and physicians will be reluctant to accept the placebo randomisation when effective treatments are accessible, particularly given the multi-year duration of confirmatory trials. Functional unblinding often occurs given that ALP and total bilirubin, used as endpoints in clinical trials, are also routinely tested in clinical practice, and this may motivate participants to crossover to commercially available drugs. Confirmatory trials preferentially enrol patients with advanced disease to maximise event rates, yet these patients face greater short-term risk from placebo exposure and may respond less well to therapy, increasing the likelihood of false negative results.

The challenges of the current framework for drug approval in PBC are exemplified by the COBALT study, a phase IV RCT, designed at the request of the regulatory agencies to confirm the effectiveness of OCA.²⁴ While a detailed analysis of the COBALT trial is beyond the scope of this document, it is important to note that the study was discontinued prematurely due to recruitment and retention difficulties, functional unblinding and placebo crossover to commercially available OCA or off-label therapies. Although real-world data (RWD) suggested OCA might have improved survival,^{25,26} COBALT failed to show significant benefits in primary outcomes. Concurrently, post-marketing safety signals, including serious liver injury in patients with and without cirrhosis, raised concerns about the overall benefit-risk profile of OCA.^{27–29} As a result, OCA was withdrawn from the EU and US markets at the end of November 2024 and September 2025, respectively.^{30,31} There is great concern that the ongoing confirmatory trials of seladelpar and elafibranor may ultimately mirror the outcome of COBALT, and more broadly, that future therapies targeting PBC and other rare liver diseases may struggle to advance beyond conditional approval.

In response to these setbacks, the European Association for the Study of the Liver (EASL) and the American Association for the Study of Liver Diseases (AASLD), with the collaboration of the Asociación Latinoamericana para el Estudio del Hígado (ALEH) and the Society on Liver Disease in Africa (SOLDA), co-hosted a multi-stakeholder conference in May 2025 to establish recommendations on three domains (1) the role of liver biochemistry and non-invasive measures of liver fibrosis as surrogates of clinical outcomes in PBC RCTs and real-world studies; (2) the use of RWD and RWE to support the assessment of safety and effectiveness of new therapies in PBC RCTs; and (3) the harmonisation and implementation of PROs for quality of life, pruritus and fatigue in PBC RCTs and real-world studies. The statements and recommendations

resulting from this Delphi panel aim to provide a new framework to support future drug approval in PBC.

Methods

Expert panel generation

A 12-person Steering Committee nominated a multi-stakeholder consensus panel. It comprised 62 members, including 36 expert clinicians, 4 EMA, 14 industry, and 8 patient representatives (Table S1). Representatives from the US FDA were invited to participate in the consensus process. However, due to travel limitations and other federal participation restrictions in place at the time, they were unable to attend. For broad representation, and generalisability, panelists included experts affiliated with major hepatology societies worldwide (AASLD, EASL, ALEH and the Asian Pacific Association for the Study of the Liver [APASL]). The Steering Committee identified three main topics and created three dedicated working groups, one for each of the consensus thematic domains. The consensus process used a modified Delphi method^{32–34} to integrate existing evidence with input from all stakeholder groups. Ideally, this model could provide a framework useful for implementation in other rare autoimmune liver diseases.

Literature review and evidence grading

Utilising the PICO format [P - Patient, Population, or Problem; I - Intervention, Prognostic Factor, or Exposure; C - Comparison or Intervention (if appropriate); O - Outcome] whenever possible, the panel formulated eight key questions (Tables S2 and S3A–C). To address them, an extensive literature search was conducted between March 18th and 26th, 2025, in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. An updated literature search was conducted on November 30th, 2025, targeting new publications since the first search date. The following databases were queried: MEDLINE, Embase, PubMed, Google Scholar, Web of Science, [ClinicalTrials.gov](https://www.clinicaltrials.gov), and the Cochrane Database of Systematic Reviews. To ensure comprehensive coverage, supplementary manual searches were performed in the bibliographies of relevant systematic literature reviews, international society guidelines, and abstract books from major international conferences in addition to grey literature sources. No restrictions were applied regarding language or country of publication. The screening and selection process was facilitated using the Covidence software.³⁵ All searches were led by the same core team [AM, MS and WW, see acknowledgements] to ensure methodological consistency and uniform application of inclusion criteria across groups. Each working group also conducted an independent literature search tailored to its specific PICO questions. Detailed search strategies and PRISMA flow diagrams illustrating the study selection process for each working group are available in Figs. S1–S3; Table S4A–C. The quality of evidence was graded using the Oxford Centre for Evidence-Based Medicine (OCEBM) criteria (Tables S5 and S6).

Development of recommendations and statements

All working groups convened approximately weekly for online discussions to review the selected literature and jointly draft

recommendations and statements. Recommendations are actionable guidance based on available evidence and designed to help end-users make informed decisions on how to undertake specific actions.³⁶ The grade of the recommendations was defined according to the OCEBM criteria and is reflected in the wording of the recommendations. Statements provide factual information or clarification based on evidence and expert opinion. The drafted recommendations and statements were subsequently discussed in person by the full consensus panel and then submitted to the Delphi panel for formal review and voting.

Delphi process and analysis

All Delphi panel members participated in at least one working group. Responses for each recommendation were recorded using a three-point Likert-type scale reflecting the level of agreement (*i.e.* agree, neutral, disagree). In addition, free-text fields were provided to allow panellists to comment or propose edits, in accordance with established Delphi methodology.^{33,34} The Delphi voting rounds data were subjected to both quantitative and qualitative analyses. Quantitative analysis involved generating response frequencies into a three-point Likert scale (agree, neutral, disagree) to evaluate the extent of consensus for each statement. Neutral responses were excluded from the denominator when calculating agreement proportions. Qualitative analysis of the accompanying free-text comments was independently conducted by chairs of each working group and discussed in follow-up online meetings to guide the revision of recommendations and statements.

Results

Delphi panel characteristics

The Delphi panel comprised 53 PBC experts, patient representatives, industry and regulatory members from 18 countries. Representatives from the EMA took part in the discussions solely in an observer/advisory capacity and did not participate in the Delphi voting rounds or act in a regulatory role. Industry participation was deliberately included to reflect their substantial contribution to therapeutic development in PBC. To ensure balanced representation, each industry organisation was limited to a single vote regardless of the number of representatives participating in the discussion. Detailed characteristics of the Delphi panel are provided in Table 1.

Response rates and panel participation

The first Delphi voting round was conducted between August 14 and September 10, 2025. Panel members voted on 16 statements and 42 recommendations distributed across three thematic domains. A total of 44 experts participated, corresponding to an 80% response rate. The first voting round generated 349 written comments, which were thematically reviewed and discussed by the working groups. Consensus was predefined as $\geq 75\%$ agreement among respondents. Statements achieving 50–75% agreement and/or $>10\%$ disagreement were earmarked for revision and resubmission. Following the first round, a virtual meeting was held including at least one steering committee chair from each working group and the consensus chairs ($n = 6$ participants) to refine the content of statements requiring further review. The second voting round was conducted

Table 1. Delphi panel characteristics.

Characteristic	n (%)
Total panel members	53
Expert clinicians	34 (64%)
Industry representatives	8 (15%)
Patient representatives	8 (15%)
Regulatory representatives*	3 (6%)
Sex	
Men	32 (60%)
Women	21 (40%)
Geographic region	
Europe	29 (55%)
North America	18 (33%)
Asia-Pacific	3 (6%)
South America	2 (4%)
Africa	1 (2%)
Countries represented	18
Delphi voting participation	
Round 1 respondents	44 (88%) [†]
Round 2 respondents	30 (60%) [†]

*Regulatory representatives contributed to discussions but did not participate in Delphi voting.

[†]Denominator excludes regulatory representatives; each industry organisation permitted one vote regardless of number of individuals involved.

between 19 and 22 September 2025 for four recommendations and one statement that met revision criteria. Thirty panel members participated (response rate 68%), and neutral responses were excluded from agreement calculations. The revised statements were subsequently incorporated into the consensus document, which underwent final review and approval by the EASL and AASLD Governing Boards. Upon completion of the Delphi process, the finalised statements and recommendations were organised according to the three working group domains. The following sections present the consensus statements and recommendations, supporting rationale, and future research priorities within each domain.

1. Surrogate endpoints in PBC clinical trials

Context: The primary endpoint in phase III PBC RCTs should be a clinically meaningful outcome, defined as either a clinical event relevant to the patient or an endpoint that directly measures how a patient feels, functions or survives.³⁷ However, the slow rate of disease progression in PBC limits the feasibility of using hard clinical outcomes such as hepatic decompensation, liver transplantation or death as primary endpoints in phase III RCTs. Surrogate endpoints are therefore needed to support drug approval. However, to date, no biomarker in PBC has met the evidentiary standard necessary for recognition as a validated surrogate endpoint by regulatory agencies.^{38–41} Although there are no standardised regulatory criteria defining a formal validation pathway, a combination of strong mechanistic rationale and evidence at the RCT level is expected. Nonetheless, some biomarkers have been accepted as surrogate endpoints in the absence of these criteria, particularly where they lie directly in the disease pathway (*e.g.* serum creatinine and glomerular filtration in chronic kidney disease) or represent the disease process itself (*e.g.* sustained virological response in hepatitis C). In PBC, RWE continues to show an association between some biomarkers and clinical outcomes, supporting their use as endpoints in clinical trials. The purpose of this section is to critically evaluate the available evidence, define acceptable thresholds for a pragmatic use of biomarkers as endpoints in RCTs, and outline a research agenda to move toward formal surrogate endpoint validation.

1.1. What is the role of liver biochemistry as a biomarker in PBC clinical trials?

Among liver biochemical parameters, ALP and total bilirubin are the most extensively studied. ALP levels are strongly related to disease pathogenesis^{42–45} and sustained reductions in ALP and total bilirubin are consistently associated with lower risk of hepatic decompensation, LT, and death in real-world retrospective cohorts. These associations have been shown in untreated natural history cohorts, in patients treated with UDCA, and more recently in those receiving second-line therapies.^{6,10,46–64} For this reason, regulatory agencies recognise ALP and total bilirubin as reasonably likely predictors of clinical outcomes, supporting conditional approval through accelerated pathways, although they have not yet been formally validated as surrogate endpoints.^{65–67} To date, RWE on the association between biochemical markers and clinical outcomes under second-line therapy is derived from studies of OCA and bezafibrate. Data for more recently approved agents, seladelpar and elafibranor, are not yet available owing to their shorter time in clinical use.

A systematic literature review identified 69 studies published between 1995 and 2025 assessing the role of biochemical markers in predicting clinically meaningful outcomes in patients with PBC. The median sample size was 540 patients (range 63–4,845), with follow-up ranging from 2.5 to 10 years. Most studies included naïve or UDCA-treated patients, but more recent studies also included patients on second-line therapies, OCA and bezafibrate.⁶⁸ Definitions of clinically meaningful outcomes varied, although the most common was a composite of liver transplantation and death (all-cause or liver-related).

All studies assessing ALP or bilirubin at 12 months of UDCA or second-line therapy reported significant associations with clinical outcomes (multivariate analyses, $p < 0.05$)^{6,8,46–49,52,58–60,69–72} (Table S7A–C). In the largest international cohort ($n = 4,845$), Lammers *et al.* demonstrated that the risk of poor clinical outcomes increases with higher ALP (hazard ratio [HR] 2.49 for ALP $> 2.0 \times$ the upper limit of normal [ULN], $p < 0.001$) and total bilirubin (HR 5.06 for total bilirubin $> 1.0 \times$ ULN, $p < 0.001$).⁴⁶ These findings were independently confirmed in a large UK nationwide cohort ($n = 3,165$) from both secondary and tertiary centres. An additional fifteen studies examined composite prognostic measures incorporating ALP and/or total bilirubin at 12 months and consistently predicted poor clinical outcomes.^{6,9,49,51,60,73–80}

The robustness of the evidence is limited by a substantial overlap among study cohorts, particularly within follow-up analyses derived from the Global PBC cohort which has been used for eleven studies,^{46,47,51–53,58,59,71,78,81} and heterogeneity in outcome definitions. However, the consistency of findings in large international cohorts and the strong mechanistic rationale linking ALP and total bilirubin with cholestatic injury have supported their use as primary endpoints in PBC clinical trials for conditional/accelerated marketing authorisation.^{40,82}

Delphi panellists were asked to evaluate the use of liver biochemical markers as endpoints in PBC clinical trials. Agreement across statements was uniformly high, with $\geq 94\%$ consensus for all items. Specifically, there was complete agreement (100%) that ALP and total bilirubin levels after 12 months of UDCA therapy are associated with clinical

outcomes, and near-complete agreement (96–98%) for corresponding associations under second-line treatments. Similar strong consensus ($\geq 98\%$) supported the prognostic value of albumin and platelet count, while 94% of panellists agreed that evidence for AST (aspartate aminotransferase), ALT (alanine aminotransferase), and GGT (gamma-glutamyltransferase) remains insufficient (Table S7D–H). Overall, 90% endorsed the recommendation that complete biochemical response (normal ALP and bilirubin) or adequate biochemical response (ALP $< 1.5\text{--}1.67 \times$ ULN and normal bilirubin) should be considered ideal or acceptable primary endpoints, respectively, in phase III PBC RCTs of at least 12 months' duration, with eligibility defined by baseline biochemistry at screening (Fig. 1). This is supported by studies demonstrating that different thresholds of ALP ($1.5\text{--}1.67 \times$ ULN) and total bilirubin have consistently high discriminatory performance.⁴⁶ However, more stringent biochemical endpoints, *i.e.* complete response (ALP and total bilirubin normalisation), offer stronger prognostic value.^{52,58}

Statements

- 1.1.1. ALP levels ($<$ or $> 1 \times$ ULN, $1.5 \times$ ULN, $1.67 \times$ ULN or $2 \times$ ULN) after 12 months of UDCA are associated with clinical outcomes in patients with PBC (**LoE 2, 100% agreement**).
- 1.1.2. ALP levels ($<$ or $> 1 \times$ ULN, $1.5 \times$ ULN, $1.67 \times$ ULN or $2 \times$ ULN) after 12 months of second-line treatment with bezafibrate or OCA are associated with clinical outcomes in patients with PBC (**LoE 3, 98% agreement**).
- 1.1.3. Total bilirubin levels ($<$ or $> 0.6 \times$ ULN, $1 \times$ ULN or $2 \times$ ULN) after 12 months of UDCA are associated with clinical outcomes in patients with PBC (**LoE 2, 100% agreement**).
- 1.1.4. Total bilirubin levels ($<$ or $> 0.6 \times$ ULN or $1 \times$ ULN) after 12 months of second-line treatment with bezafibrate or OCA are associated with clinical outcomes in patients with PBC (**LoE 3, 96% agreement**).
- 1.1.5. There is insufficient evidence to determine the impact of AST, ALT and GGT on long-term clinical outcomes (**LoE 3, 94% agreement**).
- 1.1.6. Albumin levels ($<$ or $> 1 \times$ the lower limit of normal) after 6 to 12 months of UDCA are associated with clinical outcomes in patients with PBC (**LoE 2, 98% agreement**).
- 1.1.7. Circulating platelet count ($<$ or $> 1 \times$ the lower limit of normal) after 6 to 12 months of UDCA is associated with clinical outcomes in patients with PBC (**LoE 2, 100% agreement**).
- 1.1.8. Inadequate biochemical response after 6 to 12 months of UDCA – and at any time thereafter – is associated with increased risk of poor clinical outcomes in patients with PBC (**LoE 2, 98% agreement**).
- 1.1.9. Inadequate biochemical response after 12 months of second-line treatment with bezafibrate, fenofibrate or OCA is associated with increased risk of poor clinical outcomes in patients with PBC (**LoE 3, 96% agreement**).

Recommendation

1.1.1. In phase III PBC clinical trials (based on biochemical inclusion criteria) of at least 12 months' duration, a complete biochemical response (normal ALP and normal total bilirubin) or an adequate biochemical response (ALP $<1.5\text{--}1.67 \times \text{ULN}$ and normal total bilirubin) should be considered ideal or adequate endpoints, respectively (**LoE 2, strong recommendation, 90% agreement**).

1.2. Can non-invasive tests for liver fibrosis be considered surrogate endpoints for clinical outcomes in PBC?

In addition to liver biochemistry, histological progression could be used as a surrogate for clinically meaningful outcomes in PBC. However, repeated liver biopsies carry procedural risks, are poorly accepted by patients and are not part of routine clinical management.^{1,51} Moreover, while liver biopsy provides valuable information at a population level, its accuracy for tracking disease progression in individual patients remains imperfect mainly because of sampling variability and inter-observer heterogeneity. Non-invasive measures of liver fibrosis offer a more feasible alternative for use in clinical trials. Among these, liver stiffness measurement (LSM) by vibration-controlled transient elastography (VCTE) is the most extensively validated modality in PBC, with large retrospective cohorts demonstrating good performance in identifying advanced fibrosis and predicting clinical outcomes. As with liver biopsy, LSM has important limitations. In particular, LSM values can be influenced by factors other than fibrosis, including inflammation, cholestasis, operator technique, and patient habitus, and concerns remain regarding the reproducibility and consistency of individual measurements across centres and operators. Unlike histology, LSM does not allow post-test reassessment of tissue, which has been highlighted by regulatory agencies as a key challenge when considering its use as a surrogate endpoint at the individual patient level. Furthermore, much of the published evidence supporting the

prognostic value of LSM derives from multiple analyses of overlapping patient cohorts, while prospective data linking changes in LSM to hard clinical outcomes remain limited, representing an important unmet need in the field.

Despite these limitations, available data suggest that LSM provides clinically meaningful prognostic information. Specifically, LSM above 10 kPa at baseline and on treatment was significantly associated with poor clinical outcomes, independent of biochemical markers. Moreover, a $\geq 20\%$ increase in LSM over 1 year was associated with worse outcomes, whereas a $\geq 20\%$ decrease was associated with improved outcomes.^{54,79,83–86} The optimal timing of assessment remains uncertain. Although there is pressure to limit the double-blind period to 12 months to minimise placebo exposure, particularly if the endpoint is LSM stability (*i.e.* no increase), prospective data from the Bezurso trial suggest that a 24-month interval may be more appropriate, especially in patients with early-stage disease.²¹

Among non-invasive serum biomarkers of liver fibrosis, the enhanced liver fibrosis (ELF) score is the most studied, as it has been widely included in RCTs. The ELF score, based on serum hyaluronic acid, procollagen type III N-terminal peptide, and tissue inhibitor of metalloproteinase, correlates with fibrosis stage and has shown predictive value towards clinical outcomes in PBC at baseline; however, its prognostic value under treatment remains to be fully elucidated.^{87,88} Notably, despite its inclusion in trial protocols, the ELF score is not widely used in routine clinical practice. While non-proprietary non-invasive tests of liver fibrosis, such as APRI (AST-to-platelet ratio index) and FIB-4, have been evaluated in some observational studies, they have not been utilised as secondary or exploratory endpoints in phase III RCTs. Therefore, their role was not explored further in this consensus.

A systematic literature review identified 10 studies on LSM and two studies on the ELF score, evaluating the association between these tests and clinically meaningful outcomes in PBC, published between 1995 and 2025 (Table S7I and J). The median sample size was 426 patients (range: 102 to 3,284) with follow-up ranging from 15 months to 7.3 years. All studies included UDCA-treated patients. Definitions of clinically

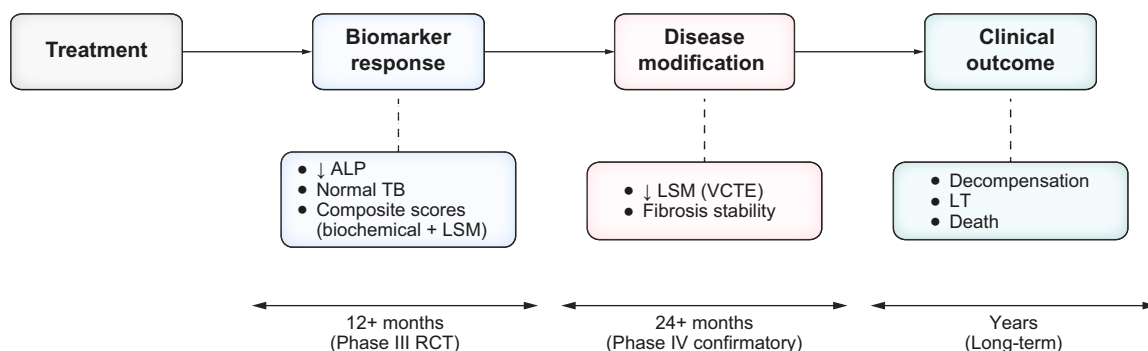


Fig. 1. Surrogate endpoint framework for PBC clinical trials. Schematic representation of the pathway from treatment initiation to clinical outcomes in PBC, illustrating the role of surrogate endpoints at different stages of disease assessment. Biomarker response, assessed through reductions in ALP, normalisation of TB, and composite scores, serves as an early endpoint measurable at ≥ 12 months in phase III RCTs. Disease modification, assessed through changes in LSM by VCTE and fibrosis stability, requires extended follow-up of ≥ 24 months in phase IV confirmatory trials. Clinical outcomes, including hepatic decompensation, LT, and death, represent the ultimate endpoints but require years of follow-up. Arrows indicate the temporal relationship between treatment effects and measurable outcomes. ALP, alkaline phosphatase; LSM, liver stiffness measurement; LT, liver transplantation; PBC, primary biliary cholangitis; RCT, randomised controlled trials; TB, total bilirubin; VCTE, vibration-controlled transient elastography.

meaningful outcomes varied, but most studies included a composite of liver decompensation, LT, and death (all-cause or liver-related).

Delphi panellists were asked their opinion about the use of LSM by VCTE and serum biomarkers of liver fibrosis as endpoints in PBC RCTs. Agreement was consistently high across all statements, with $\geq 96\%$ consensus on the prognostic value of baseline and on-treatment LSM, including its predictive ability independent of biochemical markers (98% agreement). Panellists also agreed (96%) that baseline ELF score is associated with clinical outcomes, although evidence for longitudinal ELF dynamics and their correlation with biochemical response was considered insufficient (98–100% agreement). In addition, 92% supported the use of LSM by VCTE as a potential endpoint in sufficiently long phase III RCTs, while 86% endorsed combining biochemical response with stable or decreased LSM as a composite endpoint. However, it should be noted that trials incorporating LSM as an endpoint may require longer durations than the 12-month minimum recommended for biochemical endpoints alone, as fibrosis-related changes are expected to occur more slowly, and the optimal timeframe may vary depending on disease stage and inclusion criteria.

Collectively, the panel agreed that strategies are needed to strengthen the predictive ability of current surrogate markers and support the marketing authorisation process for those treatments already under investigation. These could include the adoption of more stringent biochemical targets such as normalisation of ALP and total bilirubin as well as their combination with additional biomarkers into composite endpoints including non-invasive fibrosis measures to better reflect disease progression. In parallel, the panel identified the formal validation of liver biochemistry and non-invasive liver fibrosis markers as surrogate endpoints as a research priority for future therapeutic development.

Statements

- 1.2.1. Baseline and longitudinal on-treatment changes in LSM measured by VCTE are associated with clinical outcomes in patients with PBC (**LoE 2, 98% agreement**).
- 1.2.2. LSM by VCTE is associated with clinical outcomes in patients with PBC, independently of biochemical biomarkers (**LoE 2, 98% agreement**).
- 1.2.3. There is poor evidence regarding the optimal timeframe between subsequent LSM assessments (**LoE 3, 98% agreement**).
- 1.2.4. Baseline ELF score is associated with clinical outcomes in patients with PBC (**LoE 3, 96% agreement**).
- 1.2.5. There is insufficient data on the association of longitudinal ELF scores with clinical outcomes in PBC (**LoE 5, 100% agreement**).
- 1.2.6. There is insufficient data on the association between ELF and biochemical response to treatment in PBC (**LoE 5, 98% agreement**).

Recommendations

- 1.2.1. In phase III PBC clinical trials of appropriate duration, LSM measured by VCTE may be considered a potential surrogate of clinical outcomes (**LoE 3, weak recommendation, 92% agreement**).
- 1.2.2. In phase III PBC clinical trials of appropriate duration (minimum 24 months for early-stage disease), a biochemical response combined with decreased or stable LSM by VCTE should be considered as a composite endpoint (**LoE 2, strong recommendation, 86% agreement**).
- 1.2.3. Treatment failure may be assessed by a combined endpoint of inadequate biochemical response, LSM progression, and/or hepatic decompensation (**LoE 3, weak recommendation, 98% agreement**).

Research agenda

- a) Definition and standardisation of LSM-based endpoints, including thresholds, magnitude of meaningful changes, and optimal timing of assessment across different disease stages.
- b) Prospective evaluation of the ELF score under treatment for predicting clinical outcomes.
- c) Prospective assessment of transaminases and non-proprietary liver fibrosis markers as predictors of clinical outcomes.
- d) Establishment of a structured framework for drug development that integrates clinical safety and efficacy assessment with biomarker validation into a unified approval pathway.
- e) Conduct and publish subgroup analyses within long-term confirmatory trials such as COBALT,²⁴ stratifying by responder status, disease stage and treatment history.

2. Use of real-world data and real-world evidence in PBC

Context: While the uses of RWD/RWE in PBC can be multiple – including providing a deeper understanding of the natural history of the disease, identifying risk factors and potential surrogate endpoints, informing clinical trial design (power calculations, eligibility criteria and endpoints), conducting post marketing surveillance and elucidating various aspects of disease burden – the purpose of this section is to address the use of RWD/RWE to support causal inferences related to the safety and effectiveness of new therapies in PBC. These consensus statements are intended to support harmonised methodological standards and promote regulatory-grade research practices that can drive regulatory decisions. They are grounded in regulatory guidance, published experience in PBC, and discussions among all stakeholders.

2.1. What are the key design and reporting considerations for RWE studies intended to support regulatory decisions pertaining to confirmatory trials and post marketing trials in PBC?

RCTs remain the conventional standard for evaluating therapeutic efficacy, yet in PBC they are often constrained by slow disease progression, small patient populations, and ethical concerns regarding prolonged placebo exposure. These challenges have created a need to seek alternative pathways,

including the use of RWD to generate RWE capable of informing regulatory decisions.

Regulatory grade RWE is distinguished from exploratory observational research by its reliance on “fit-for-use” data sources – those that are relevant, reliable, and sufficiently detailed to support causal inference – and prespecified protocols with transparent data provenance and analytic reproducibility.^{89–92} This need is particularly acute in PBC, where therapies granted conditional approval based on surrogate endpoints deemed reasonably likely to predict outcomes require confirmatory evidence of long-term clinical benefit.⁹³ RWE therefore represents a pragmatic and ethically acceptable pathway to complement traditional trials.

Methodological advances have demonstrated that external control arms derived from RWD can supplement single-arm trials when conventional randomised designs are not feasible – a strategy already applied in rare disease contexts and increasingly considered for rare liver diseases.^{94,95} Key design elements – including comparator selection, temporality, and approaches to confounding adjustment – determine reliability, while frameworks such as target trial emulation and registry-based or hybrid prospective-retrospective designs provide structured methodologies to improve causal inference and reproducibility.⁹⁵

Regulatory agencies have recognised this evolving landscape. The FDA’s Real World Evidence Program, established under the 21st Century Cures Act, and EMA guidance explicitly support using RWE for evaluating new indications and fulfilling post-approval requirements, provided that data reliability and relevance are demonstrated.^{15,82,96–98} This approach has precedent: in oncology, drugs such as blinatumomab and avelumab received accelerated approval based on single-arm trials supplemented by RWD-derived external control arms.^{167,168} Both the FDA and EMA emphasise pre-specified study designs, judicious use of external control arms, rigorous data quality standards, and early engagement with regulators as essential.^{99,100}

Despite these advances, important limitations persist: RWD sources vary in representativeness and completeness; unmeasured confounding threatens validity; and cross-system heterogeneity adds complexity. These uncertainties underscore the need for careful design, transparent reporting, and formalised interaction with regulatory bodies when planning real-world studies intended to support drug approval or label expansion in PBC.⁹⁸

From the consensus discussions, two principles emerged. First, engagement with regulatory authorities must occur early and throughout the design and execution of real-world studies. Second, in situations where RCTs are not feasible, well-designed single-arm trials incorporating target trial emulation frameworks and rigorously defined RWD-based external controls can provide credible, regulatory-grade evidence of treatment benefit in PBC. These principles directly underpin the recommendations below.

Recommendations

2.1.1. Early and ongoing engagement with regulatory agencies is strongly recommended for studies aiming to support approval or label expansion (**LoE 5; strong recommendation, 100% agreement**).

2.1.2. When RCTs are not feasible, single-arm trial designs that incorporate external control arms derived from rigorously defined RWD should be considered acceptable to support regulatory-grade evidence of safety and effectiveness (**LoE 3; strong recommendation, 100% agreement**).

2.2. Which data sources should be used for RWD studies in PBC?

The choice of data sources is central to the credibility and regulatory utility of RWE in PBC. Regulatory-grade RWE depends on “fitness-for-use” – the degree to which data are relevant, reliable, and sufficiently complete to address the study objective.^{98,101}

Disease-specific registries provide the most granular and clinically meaningful information on treatment exposure, laboratory dynamics, and PROs. These registries – such as the Global PBC and UK-PBC cohorts – have generated validated prognostic and survival models^{47,48} and remain indispensable for endpoint validation. Retrospective and prospective observational cohorts further demonstrate the feasibility of structured multicentre data (national and international) collection with long-term follow-up.^{102–104} Together, these initiatives, summarised in Table S8, constitute the empirical backbone of PBC-focused RWE.

Electronic health records (EHRs) capture routine clinical practice, comorbidities, laboratory values, and prescribing patterns offering rich context for effectiveness and safety analyses.¹⁰⁵ Claims databases complement these strengths by providing population-level coverage, standardised coding, and detailed information on healthcare utilisation and cost. They are well suited to capturing hard clinical endpoints, such as liver transplantation or hospitalisation for cirrhosis complications, though they lack granularity with respect to laboratory and PRO data.^{106,107} Distributed data networks enable federated analyses across systems, enhancing representativeness while preserving privacy.^{98,101}

To ensure regulatory acceptability, endpoints and covariates must be harmonised across data types and defined *a priori*.^{82,96} Structured approaches to improve completeness – such as prospective capture of safety variables, documenting treatment decisions, linking registry and EHR data, and integrating PROs – are necessary to close current knowledge gaps and align with international data-governance standards prioritising quality over volume.^{108–111}

Building on this context, the panel recognised that no single data source is sufficient in isolation. While PBC benefits from mature registry infrastructures that have shaped risk stratification and endpoint development, registries alone may limit generalisability by reflecting tertiary-centre practice. EHR and claims databases are essential complements for capturing treatment patterns and safety at scale, despite limitations in laboratory and PRO availability. Distributed data networks offer promise for future integration, provided cross-system harmonisation and standardised coding are achieved. Optimal RWE generation requires thoughtful alignment of data sources with study aims, endpoint harmonisation, and proactive measures to strengthen data quality and interoperability. These shared conclusions form the basis of the recommendations that follow.

Recommendations

- 2.1.3. RWD studies in PBC should utilise appropriate data sources, which may include registries, EHRs, claims databases, and distributed data networks, selected based on the study objectives and fitness-for-use (**LoE 5; strong recommendation, 98% agreement**).
- 2.1.4. Endpoint definitions should be harmonised across data sources and aligned with regulatory standards (**LoE 5; strong recommendation, 100% agreement**).
- 2.1.5. The quality and utility of emerging and existing data sources intended to generate regulatory-grade evidence should be enhanced by structured approaches addressing key knowledge gaps (**LoE 5; strong recommendation, 100% agreement**).

2.3. How can we identify and control confounding and various biases in RWD specifically in PBC studies?

Bias and confounding are pervasive challenges in RWD studies and represent a primary barrier to generating regulatory-grade evidence. Unlike random error, bias introduces systematic distortion that can misrepresent causal relationships. Because regulatory decisions depend on valid causal inference, the identification, quantification, and mitigation of bias are central to RWE credibility in PBC.

PBC poses unique challenges in the identification and mitigation of bias. As a rare, slowly progressive disease, it is prone to confounding by indication, selection bias, and

missing data. Observational datasets may under- or over-represent patients with advanced or symptomatic disease depending on the data source, and treatment allocation is frequently influenced by baseline risk factors. Other threats include immortal time bias, exposure or outcome misclassification from coding errors, and attrition bias from incomplete follow-up.^{112–115} Because clinical events such as decompensation or transplantation are infrequent, even modest imbalances can yield over- or underestimated effects (Table 2).

Methodological frameworks such as target trial emulation, external control arms, and hybrid prospective-retrospective designs highlight the need for explicit adjustment of time-related biases and robust comparator definitions.^{116–122} Validation of automated case-finding algorithms against clinical reference standards – such as EHR-based detection of cirrhosis – is essential to ensure accurate population identification.^{103,109} Regulatory agencies stress that internal validation of RWD through benchmarking against registries, adjudicated outcomes, or manual chart review is indispensable.^{96,98}

Bias mitigation in data linkage represents an emerging frontier. Tokenisation strategies linking de-identified patients across datasets introduce the possibility of false matches (“collisions”) and missed matches (“splits”).^{110,123} Linkage methods must be prespecified, justified, and quantitatively evaluated.^{101,124}

Post hoc investigation of unanticipated or inconclusive findings is an essential component of scientific transparency. PBC registry analyses have demonstrated regional variation in laboratory practices and treatment uptake, emphasising the importance of contextualising observed effects. Experience from oncology RWD frameworks illustrates how systematic

Table 2. Biases relevant to real-world data studies in PBC and recommended mitigation strategies.

Type of Bias	Brief explanation	Statistical mitigation techniques	Examples in PBC
Selection bias	Study population may not represent the general PBC population	Population weighting; sensitivity analyses; broader data sources ^{91,92,98,114–118}	Registries over-represent engaged patients; ICD codes have suboptimal sensitivity/specificity
Confounding by indication/channeling bias	Sicker patients more likely to receive certain treatments	Propensity score matching/weighting; multivariable adjustment ^{91,92,118}	Patients with progressive disease more likely to initiate second-line therapy; access influenced by insurance coverage
Immortal time bias	Unaccounted time between cohort entry and treatment initiation	Time-dependent Cox models; proper time alignment; consistent index date definitions ^{91,109–111}	Using treatment initiation date as time zero creates survival bias; match treated patients with controls at equal time from cohort entry
Information bias	Inaccurate or inconsistent data collection leads to incorrect exposure or outcome assessment	Validation studies; standardised data collection protocols; algorithm validation against chart review ^{91,98,100,115}	Diagnostic algorithms require validation; transplant waitlisting is more accurate and proximal than transplantation itself as an outcome
Misclassification bias	Incorrect categorisation of exposure or outcome distorts associations	Validated algorithms; sensitivity analyses; independent adjudication committees ^{91,92,112–115,122}	Comparing ITT and as-treated analyses critical for understanding early discontinuation effects; adverse event detection challenging without chart review
Inconsistent follow-up timing	Uneven measurement schedules affect outcome detection	Modelling time-varying covariates; adjustment for follow-up intensity; active comparator design ^{118–122}	Patients with greater symptom burden have more frequent visits; treated patients have regular visits for prescription renewal, unlike untreated patients
Symptom/PRO documentation bias	Subjective symptoms may be under-reported or inconsistently recorded.	Structured symptom tools; standardised PRO instruments ^{91,92,109,110}	PROs may influence treatment initiation and discontinuation decisions; intractable pruritus may drive transplant listing independent of biochemical response
Missing data bias	Missing information, especially if non-random (MNAR), leads to biased estimates	Multiple imputation; sensitivity analyses for MNAR ¹¹	Both greater and lesser disease severity can increase missingness; prolonged gaps between visits common in asymptomatic stable disease

bias audits can refine effect estimates and improve future trial emulation.¹²⁵

The panel concluded that bias identification and mitigation must be prespecified, reproducible, and transparently reported. Internal validation through registry benchmarking or chart review should be considered mandatory when using EHR or claims data. Tokenisation must include quantitative error assessment. Transparency in exploring residual bias was viewed as a scientific obligation, especially in rare diseases where every dataset informs future research (Fig. 2).

Recommendations

- 2.1.6. All real-world studies in PBC should pre-specify anticipated biases (**LoE 5; strong recommendation, 100% agreement**).
- 2.1.7. Mitigation strategies should be pre-specified and implemented to control for biases in studies utilising RWD in PBC (**LoE 5; strong recommendation, 100% agreement**).
- 2.1.8. EHR and claims-based datasets should be internally validated – either through benchmarking against clinical reference standards or via manual chart review – to ensure accurate identification of the study population, exposures, outcomes, and covariates (**LoE 2; strong recommendation, 100% agreement**).
- 2.1.9. If tokenisation is used to link de-identified patients across datasets, the choice of token(s) should be justified, and risk of linkage errors (e.g. collisions and splits) should be quantified (**LoE 5; strong recommendation, 100% agreement**).
- 2.1.10. When RWD analyses lead to unanticipated or inconclusive results, *post hoc* investigations of bias sources should be encouraged and transparently reported to support future studies and regulatory decision-making (**LoE 5; strong recommendation, 100% agreement**).

Research agenda

- a) Adapt target trial emulation and external control arm designs for PBC, with clear strategies for comparator selection and confounding control.
- b) Strengthen registries, EHRs, claims, and distributed networks with standardised endpoints and integrated PRO collection.
- c) Benchmark EHR/claims data against clinical standards, and systematically identify, mitigate, and report key biases.
- d) Refine tokenisation approaches, quantify linkage errors, and promote transparent reporting for multi-source studies.
- e) Ensuring that the variability in clinical care and practice globally is included in the design and interpretation of RWD.

3. Patient-reported outcomes and symptoms assessment in PBC

Context: The impact of PBC on quality of life is multifaceted, encompassing symptoms such as fatigue and pruritus, as well as broader physical, emotional, and functional limitations. It is important to recognise that these symptoms can have a significant impact on patients' life experiences and quality of life and warrant treatment irrespective of interventions to prevent disease

progression. Thus, there is a need for harmonised approaches considering the patient's experience, to assess these outcomes in both clinical trials and real-world settings. These consensus statements aim to guide the selection and implementation of outcome measures that are patient-centred, psychometrically robust, and aligned with regulatory expectations. The routine inclusion of PROs in all clinical trials in PBC is essential – not only to evaluate the efficacy of symptom-targeted therapies but also to assess the broader impact, both beneficial and adverse, of treatments not primarily designed to address symptoms. The recommendations presented here are informed by current evidence, the lived experiences of patients, input from clinicians, and accumulated expertise in PBC research. The ultimate objective is to promote consistency and rigor in outcome measurement, enabling meaningful cross-study comparisons and supporting the development of regulatory-grade evidence in PBC.

3.1. What are the key considerations for PROs in PBC?

PROs are defined as direct patient-derived reports of symptoms, functioning, and well-being that quantify how disease and treatment affect daily life. In PBC, they capture domains that characterise disease burden – most prominently fatigue, pruritus, cognitive dysfunction, emotional distress, and social participation – none of which are reflected in biochemical or histological markers.^{126–128} These manifestations are core therapeutic targets, as many patients remain symptomatic despite biochemical response. PRO instruments may be disease-specific (e.g. PBC-40)¹²⁹ or generic (e.g. EuroQoL-5 Dimension [EQ-5D], Short Form Health Survey [SF-36], Patient-Reported Outcomes Measurement Information System 29 [PROMIS-29]),^{130–133} offering complementary perspectives on treatment benefit.

Validated PROs (*i.e.* instruments with demonstrated content validity, reliability, and responsiveness to change in the target population) are indispensable for evaluating outcomes that matter most to patients and for establishing therapeutic value.¹³⁴ Because symptoms such as fatigue and pruritus cannot be objectively quantified through laboratory or imaging markers, patient report is the sole feasible means of assessing efficacy when symptom relief is the therapeutic target.^{135,136} Regulatory agencies increasingly regard PROs as essential to demonstrating treatment benefits. The EMA emphasises that PROs complement clinical endpoints by providing direct insight into how patients feel and function.¹³⁶ The FDA's PRO Guidance states that validated PROs can support labelling claims and post-approval commitments.¹³⁷ Beyond regulatory requirements, PRO-based endpoints influence clinical practice guidelines, health technology assessments, and payer decisions.¹³⁴

The credibility of PRO data depends on rigorous implementation. First, the timing and frequency of assessments must align with disease course and expected treatment response. Symptoms such as pruritus may fluctuate over days to weeks. Measurements across clinically meaningful windows are therefore preferable to single time-point comparisons; however, the downsides of overly frequent assessment must be considered, including completion fatigue and recall of previous responses. Recall periods and scoring algorithms for the various instruments utilised in the study must be prespecified to minimise inconsistent interpretation.^{136,137} Second, interpreting PRO change requires establishing thresholds of minimal clinically important change (MCIC). Anchor-based methods – typically using global impression instruments such as Patient Global

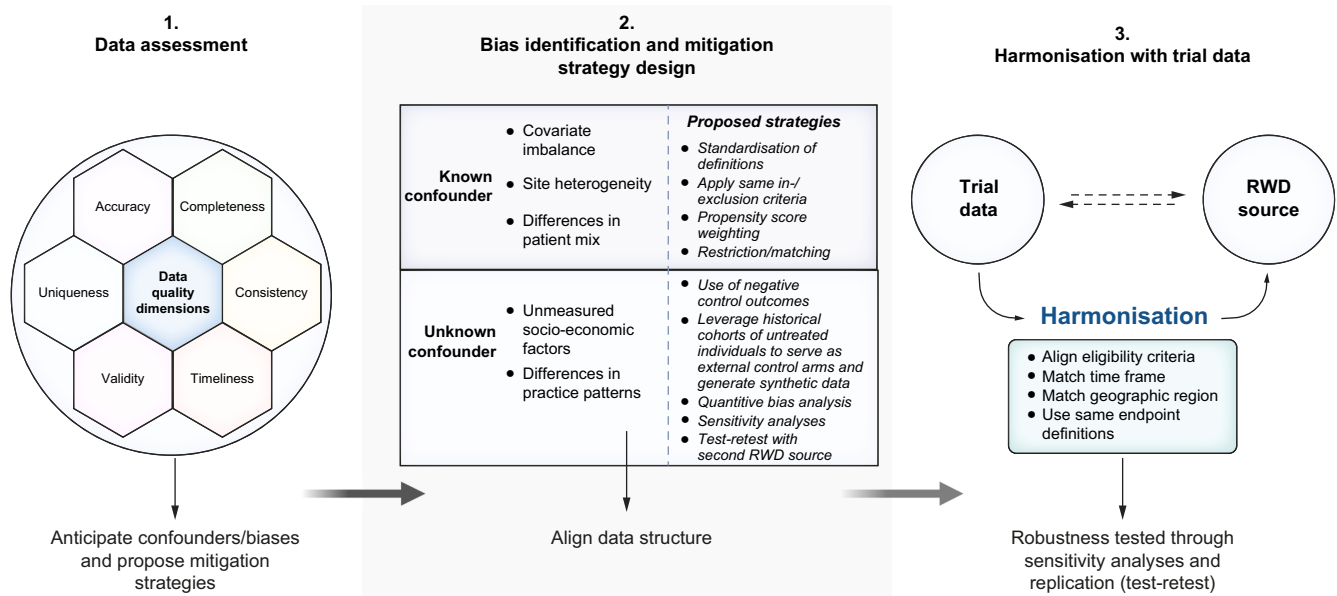


Fig. 2. Methodological framework for generating regulatory-grade real-world evidence in primary biliary cholangitis. Three-stage workflow for designing real-world studies intended to support regulatory decisions in primary biliary cholangitis. RWD, real-world data.

Impression of Severity (PGI-S) or Change (PGI-C) – define the magnitude of improvement perceived as meaningful by patients.^{134,138,139} Disease-specific clinician-rated tools, including the Clinical Global Impression–Severity and Change for PBC (CGI-S-PBC and CGI-C-PBC),^{140,141} may serve as complementary anchors (*i.e.* external reference points used to interpret whether PRO score changes are clinically meaningful). The panel recognised that MCIC validation in PBC remains incomplete and highlighted the need for harmonised interpretation across trials. Where validated MCIC thresholds are not yet available for a PRO intended as a key endpoint, the panel emphasised that such thresholds should be developed and prespecified prior to study initiation to support regulatory acceptance of the findings. Third, consistency in data collection is important. PROs should be collected using the validated format (electronic or paper), with the same mode used across participants and time points to reduce measurement error and ensure comparability. PROs should be validated for use in the language in which they are administered.

Objective or physiological tools – such as actigraphy, orthostatic testing, or wearable sensors – can offer exploratory insights into activity, sleep, or autonomic symptoms but capture constructs distinct from patient-reported experience. For instance, actigraphy may underestimate fatigue in patients who maintain activity despite substantial symptom burden.^{142,143} These methods may complement but cannot replace PROs in regulatory or confirmatory settings until further validation occurs.

Finally, accurate interpretation requires accounting for comorbidities and overlapping mechanisms. Conditions such as thyroid disease, depression, or primary sleep disorders can confound the interpretation of fatigue as directly related to PBC, whilst autonomic dysfunction may independently contribute to fatigue.^{144–146} Measuring these variables and prespecifying analytical adjustments are essential to avoid misattribution of treatment effects.

Overall, the consensus panel concluded that PROs are foundational – not optional – endpoints in PBC research. Their

appropriate selection, standardised implementation, and transparent analysis are indispensable for generating patient-centred regulatory-grade evidence that truly reflects how individuals feel and function in response to therapy. The following principles apply to all patient-reported measures used to assess quality of life, fatigue and pruritus in people living with PBC. These general recommendations aim to promote rigorous and harmonised implementation of PROs across both clinical trials and real-world settings (Table 3).

Recommendations

- 3.1.1. PROs should be collected at multiple, prespecified time points that align with the trial design and the anticipated trajectory of treatment response (**LoE 2; strong recommendation, 98% agreement**).
- 3.1.2. Studies should clearly define the recall period and scoring approach for each instrument used. In clinical trials, this information should be specified *a priori* in the protocol or statistical analysis plan (**LoE 2; strong recommendation, 100% agreement**).
- 3.1.3. Thresholds for minimal clinically important change (MCIC) should be pre-established to support interpretation of patient-level and group-level treatment effects (**LoE 2; strong recommendation, 100% agreement**).
- 3.1.4. Global impression instruments, such as the Patient Global Impression of Severity (PGI-S) or Change (PGI-C), and when applicable disease-specific clinician-rated tools, like Clinical Global Impression–Severity and Change for PBC (CGI-S-PBC and CGI-C-PBC), may be used to support interpretation of PRO changes, and are particularly helpful when used as anchor items for establishing MCIC thresholds (**LoE 5; weak recommendation, 98% agreement**).

Table 3. Ideal characteristics of PRO instruments for use in PBC.

Characteristic	Description	Rationale
Validated in PBC population	Psychometric properties (reliability, validity, responsiveness) established in patients with PBC	Ensures instrument measures what it intends to measure in this specific disease context
Patient-derived content	Item content developed through qualitative input from patients with PBC	Captures domains and language relevant to patients' lived experience
Covers PBC-relevant domains	Addresses key symptom domains: fatigue, pruritus, cognitive function, emotional well-being, social participation	Reflects the multifaceted burden of PBC beyond biochemical markers
Established MCIC threshold	Minimal clinically important change defined using anchor-based methods	Enables interpretation of whether observed changes are meaningful to patients
Responsiveness to change	Demonstrated sensitivity to detect treatment effects over time	Essential for evaluating therapeutic benefit in clinical trials
Regulatory familiarity	Accepted or reviewed by FDA/EMA for use in clinical trials	Facilitates regulatory submissions and labelling claims
Appropriate recall period	Recall period aligned with symptom variability and study design	Minimises recall bias; captures fluctuating symptoms (e.g. pruritus) accurately
Feasible for repeated use	Low respondent burden; suitable for multiple administrations	Supports longitudinal assessment without excessive patient fatigue
Available in validated translations	Linguistically validated versions for multinational studies	Ensures cross-cultural equivalence in global trials
Compatible with electronic administration	Validated for ePRO platforms	Enables real-time data capture and integration with EHR/registries

ePRO, electronic patient-reported outcome; EHR, electronic health record; MCIC, minimal clinically important change; PBC, primary biliary cholangitis; PRO, patient-reported outcome.

3.1.5. PROs should ideally be administered using the format in which the tool was validated, either electronically or on paper. Regardless of the mode chosen, consistency across all participants and time points within a study is essential to preserve data integrity (**LoE 5; strong recommendation, 98% agreement**).

3.1.6. Timely and frequent data collection is recommended to minimise recall bias, particularly in symptoms known to fluctuate (e.g. pruritus) (**LoE 5; strong recommendation, 98% agreement**).

3.1.7. Objective tools (e.g. actigraphy, orthostatic testing) may offer valuable exploratory insights but measure constructs distinct from patients' lived experiences and may be considered complementary rather than substitutes for PROs (**LoE 5; weak recommendation, 98% agreement**).

3.1.8. Both comorbidities and symptom-modifying mechanisms should be considered for the accurate interpretation of PRO abnormalities (**LoE 5; strong recommendation, 92% agreement**).

3.2. How should global quality of life be assessed in PBC?

Global quality of life in PBC reflects the cumulative impact of symptoms, treatment effects, and psychosocial factors on patients' lived experience. It integrates domains such as fatigue, pruritus, cognitive function, and emotional and social well-being, providing comprehensive insight beyond biochemical or histologic endpoints.

The PBC-40 remains the most robustly validated and widely adopted tool for assessing quality of life in PBC.¹²⁹ Developed through qualitative interviews with patients, it integrates six domains – symptoms, fatigue, itch, cognitive, emotional, and social – and has demonstrated robust construct validity, responsiveness to therapeutic change, and reliability in both clinical trials and real-world studies.^{147–150} These characteristics make it the cornerstone instrument for evaluating quality of life in PBC.

Generic measures – EQ-5D,¹³⁰ SF-36,¹³¹ and PROMIS-29^{132,133} – are valuable for generating health-utility data, enabling cross-disease comparisons but lack coverage of hallmark PBC symptoms (i.e. fatigue and pruritus) and may underestimate disease burden when used in isolation.¹⁵¹ Both the FDA and EMA require that quality of life measures be conceptually sound, validated within the target population, and sensitive to change when quality of life is a treatment objective.^{136,137} The PBC-40 fulfils these expectations through its patient-derived content and disease specificity.

Global impression instruments, including PGI-S and PGI-C,^{138,139} as well as emerging clinician-rated tools, such as the CGI-S/C-PBC,^{140,141} may support determination of MCIC thresholds and strengthen interpretability of patient-reported effects (Table S9).

The panel endorsed a hierarchical approach: PBC-40 as the core disease-specific instrument, complemented by generic tools when economic or comparative analyses are required, and global impression measures providing anchor-based interpretability. Aggregated PBC-40 total scores should be contextualised by domain-specific results to capture domain-level benefits or residual burdens. This approach ensures that global quality of life measurement in PBC remains scientifically robust, interpretable, and aligned with regulatory expectations.

Recommendations

3.2.1. PBC-40 should be considered the primary disease-specific tool for assessing global quality of life in PBC (**LoE 2; strong recommendation, 96% agreement**).

3.2.2. EQ-5D, SF-36 and PROMIS-29 should not be used in isolation to assess quality of life in PBC (**LoE 5; strong recommendation, 100% agreement**).

3.2.3. Patient-reported global impression tools (e.g. PGI-S, PGI-C) may be used to support analyses aimed at establishing meaningful within-patient change thresholds (**LoE 5; weak recommendation, 100% agreement**).

3.3. How should pruritus and associated symptoms be assessed in PBC clinical studies?

Pruritus is among the most burdensome and treatment-refractory symptoms in PBC, profoundly impairing sleep, mood, and daily functioning. Its accurate measurement is critical both for evaluating therapeutic benefit and for supporting regulatory decision-making.^{152,153}

The Worst Itch Numerical Rating Scale (WI-NRS), including the nearly identical but separately validated and trademarked Itch Ro[®] score, is a 0-10 patient-reported scale that rates peak itch severity over the preceding 12 to 24 hours and has been validated in cholestatic liver diseases, including PBC.^{154,155} Owing to its clarity, sensitivity to change, and regulatory acceptance, it has become the preferred primary endpoint for pruritus trials. Both the FDA and EMA have accepted WI-NRS in pivotal trials of ileal bile acid transporter inhibitors and PPAR agonists^{66,67,156,157} (Table 4).

Mean change from baseline in WI-NRS scores between treatment and placebo groups is the current regulatory standard for primary analysis, though consensus on clinically meaningful between-group differences remains to be defined. At the individual level, a ≥ 3 -point reduction has emerged as the benchmark for clinically meaningful improvement, particularly among patients with moderate-to-severe itch. Studies using related instruments suggest smaller changes (≥ 2 points) may also be meaningful for some patients. Exploratory analyses incorporating multiple thresholds (≥ 2 , ≥ 3 , ≥ 4 points) are encouraged to capture interindividual variability and align with regulatory expectations.

Beyond intensity, pruritus requires multidimensional assessment. The 5D-Itch scale – evaluating duration, degree, direction, disability, and distribution – offers insight into functional and psychosocial consequences of pruritus.¹⁵⁸ The PBC-40 itch domain reflects patient-derived experiences within a validated PBC-specific framework.^{129,150} These instruments are best positioned as secondary or composite endpoints that contextualise WI-NRS changes and provide fuller assessment of therapeutic benefit.

Placebo responses in pruritus trials are substantial (20-40% mean improvement) and must be anticipated in study design and interpretation.^{1,2} Complementary endpoints, such as sleep interference scores (e.g. the Sleep Interference Numerical Rating Scale [Sleep-NRS] and the PROMIS Sleep Disturbance Scale), help quantify downstream effects of itch and capture aspects not reflected by intensity scores alone.¹⁵⁹ Objective tools like actigraphy provide exploratory data on nocturnal restlessness or activity but remain unvalidated as regulatory endpoints (Table S10).

The panel agreed that WI-NRS should remain the primary patient-reported endpoint for quantifying pruritus severity in PBC trials. However, multidimensional and complementary measures – including the 5-D Itch Scale, PBC-40 itch domain, and sleep-focused instruments – were viewed as essential for capturing the broader functional and psychosocial consequences of pruritus. These perspectives guided the recommendations below, which aim to harmonise the use of validated instruments, establish consistent thresholds for clinically meaningful change, and ensure that symptom assessment in PBC trials reflects both intensity and impact from the patient perspective.

Recommendations

- 3.3.1 The Worst Itch Numerical Rating Scale (WI-NRS) is recommended as the primary tool for assessing pruritus severity in clinical trials for PBC (**LoE 2; strong recommendation, 98% agreement**).
- 3.3.2 Mean change from baseline in WI-NRS scores between treatment and placebo groups over the study period is currently recommended as the primary endpoint in pruritus trials (**LoE 2; strong recommendation, 94% agreement**).
- 3.3.3 A 3-point reduction from baseline in the WI-NRS (0-10 scale) is recommended as the primary threshold for clinically meaningful within-patient improvement in PBC trials (**LoE 2; strong recommendation, 94% agreement**).
- 3.3.4 Global impression tools (e.g. PGI-S and PGI-C) may support interpretation of meaningful within-patient change (**LoE 5; weak recommendation, 96% agreement**).
- 3.3.5 Wearable technologies (e.g. actigraphy) may offer exploratory data to complement PROs but are not yet validated for use as endpoints in regulatory trials (**LoE 5; weak recommendation, 100% agreement**).
- 3.3.6 Visual analogue scales should be used in a supportive or real-world context (**LoE 5; strong recommendation, 98% agreement**).
- 3.3.7 Multidimensional measures that capture the broader impact of pruritus – such as the 5-D Itch Scale and the PBC-40 itch domain – should be considered secondary endpoints or components of composite endpoints (**LoE 5; strong recommendation, 96% agreement**).
- 3.3.8 Instruments targeting specific consequences of pruritus, such as sleep disruption (e.g. Sleep Interference NRS, PROMIS Sleep Disturbance), may be used as secondary endpoints in trials involving patients with moderate-to-severe pruritus (**LoE 5; weak recommendation, 96% agreement**).
- 3.3.9 While global quality-of-life measures (e.g. EQ-5D, SF-36) may be used for context when interpreting treatment impact, their low sensitivity to change in symptom-focused interventions limits their utility as primary or secondary endpoints in pruritus-specific trials (**LoE 5; weak recommendation, 100% agreement**).

3.4. How should fatigue be assessed in PBC, and what is the role of objective tools?

Fatigue is the most common and often the most debilitating symptom in PBC, affecting approximately 40-80% of patients.⁹³ Its severity does not consistently correlate with disease stage and may persist despite biochemical response.¹⁴⁶ The pathophysiology involves both central (altered neurotransmission, motivation) and peripheral (neuromuscular dysfunction, deconditioning) mechanisms,^{144,145} with contributions from autonomic dysfunction, sleep disturbance, and mood disorders. Currently licensed therapies do not consistently improve fatigue, making it a distinct therapeutic target requiring additional studies with validated PRO measures.⁹³

Table 4. Summary of validation status for patient-reported outcome instruments in PBC.

Instrument	Domain(s) assessed	Validated in PBC	MCIC proposed in PBC	Recommended use in clinical studies	Notes
PBC-40	Global QoL, fatigue, itch, cognitive, emotional, social	Yes	No	Primary endpoint in trials	Disease-specific; patient-derived content
PBC-40 fatigue domain	Fatigue severity and impact	Yes	No	Secondary/co-primary	Core disease-specific fatigue measure
PBC-40 itch domain	Pruritus impact	Yes	No	Secondary	Underrepresents itch intensity
WI-NRS	Pruritus severity	Yes	Yes (≥ 3 -point reduction)	Primary endpoint (pruritus trials)	FDA/EMA accepted
5-D itch scale	Multidimensional itch (duration, degree, direction, disability, distribution)	No*	No	Secondary	Validated in chronic pruritus; used in PBC cohorts
VAS (itch)	Pruritus severity	No	No	Supportive/RWE	Simple but lacks standardisation
Sleep interference NRS	Sleep disturbance due to pruritus	No**	No	Secondary	Validated in pruritus populations
PROMIS fatigue SF-7a	General fatigue	No†	No	Complementary	Strong psychometrics; cross-disease comparability
Fatigue impact scale (FIS)	Multidimensional fatigue (cognitive, physical, social)	No*	No	Supportive	Validated in liver disease
EQ-5D	Health utility	No**	N/A	Economic modelling	Not disease-specific; HTA acceptance
SF-36	General health-related QoL	Yes	No	Secondary	Validated in liver disease; lacks PBC symptom specificity
PROMIS-29	Multiple domains (function, mood, sleep, pain, fatigue)	No**	No	Complementary	Comprehensive; enables utility derivation
PGI-S/PGI-C	Global impression of severity/change	No**	N/A (anchor function)	Anchor for MCIC	Supports interpretation; not stand-alone endpoint
CGI-S-PBC/CGI-C-PBC	Clinician-rated severity/change	No†	N/A (anchor function)	Anchor for MCIC	Disease-specific; undergoing validation
Epworth sleepiness scale (ESS)	Daytime somnolence	No	No	Supportive	Correlates with fatigue contributors
Orthostatic grading scale (OGS)	Dysautonomia symptoms	No	No	Supportive	Relevant to fatigue phenotyping
Actigraphy/wearables	Objective activity/sleep	No	No	Exploratory only	Not validated for regulatory endpoints

CGI-C-PBC, Clinical Global Impression of Change for PBC; CGI-S-PBC, Clinical Global Impression of Severity for PBC; EQ-5D, EuroQol 5-Dimensions; FIS, Fatigue Impact Scale; HTA, Health Technology Assessment; MCIC, Minimal Clinically Important Change; NRS, Numerical Rating Scale; OGS, Orthostatic Grading Scale; PBC-40, Primary Biliary Cholangitis-40 Questionnaire; PGI-C, Patient Global Impression of Change; PGI-S, Patient Global Impression of Severity; PROMIS, Patient-Reported Outcomes Measurement Information System; QoL, Quality of Life; RWE, Real-World Evidence; SF-36, Short Form Health Survey-36; VAS, Visual Analog Scale; WI-NRS, Worst Itch Numerical Rating Scale.

*Validated in other liver diseases or chronic pruritus populations.

**Used in PBC trials but not specifically validated in PBC.

†Undergoing validation in PBC.

The PBC-40 fatigue domain remains the most validated and widely applied tool for assessing fatigue in PBC.^{129,150} It captures both perceived severity and functional impact, with demonstrated responsiveness to change in prospective cohorts and therapeutic studies.^{148,160} The dynamic range for the PBC-40 fatigue domain score is 11 to 55, with a score ≥ 29 identifying clinically significant fatigue.¹⁶¹ Its integration within a broader PBC-specific quality of life framework establishes it as the primary disease-specific fatigue measure.

The PROMIS Fatigue Short Form 7a (PROMIS Fatigue SF-7a) complements the PBC-40 by offering strong psychometric properties, digital adaptability, availability in multiple languages, regulatory familiarity, and the ability to enable cross-disease comparisons.¹³³ Together, the PBC-40 fatigue domain and PROMIS SF-7a provide a balanced framework that is both disease-specific and generalisable for health technology and regulatory evaluation. Establishing validated thresholds for MCIC remains a priority, as interpretation of treatment effects is currently limited by the absence of consensus anchor-based thresholds in PBC-specific fatigue instruments.^{93,162} An important element of the patient experience in PBC is the balance between peripheral and central elements of fatigue, both of which are

common. Central fatigue symptoms can overlap with cognitive symptoms, as assessed using the PBC-40 cognitive symptom domain.

Other instruments can enrich interpretation or provide information regarding specific facets of fatigue. The Fatigue Impact Scale (FIS) and its modified forms have sub-domains for physical, cognitive, and psychosocial fatigue, although the relevance of these to objective peripheral and central fatigue and social isolation in PBC is not clear.¹⁶³ Global impression tools – including the PGI-S/PGI-C,¹⁶⁴ and the developing CG-S/C-PBC^{140,141} – provide complementary anchors for determining MCIC thresholds and contextualising patient-level responses but should remain secondary, supporting interpretation rather than replacing validated PROs as primary measures. Non-PBC-specific tools such as the Hospital Anxiety Depression Scale (HADS),¹⁶⁵ Epworth Sleepiness Scale (ESS),¹⁶⁶ and Orthostatic Grading Scale (OGS)¹⁴⁶ identify contributors to fatigue – mood disturbance, daytime somnolence, autonomic dysfunction – that may confound PRO interpretation. Objective tools including actigraphy and wearable sensors show promise for characterising sleep-wake patterns and physical activity but are not yet validated as regulatory endpoints (Table S11).

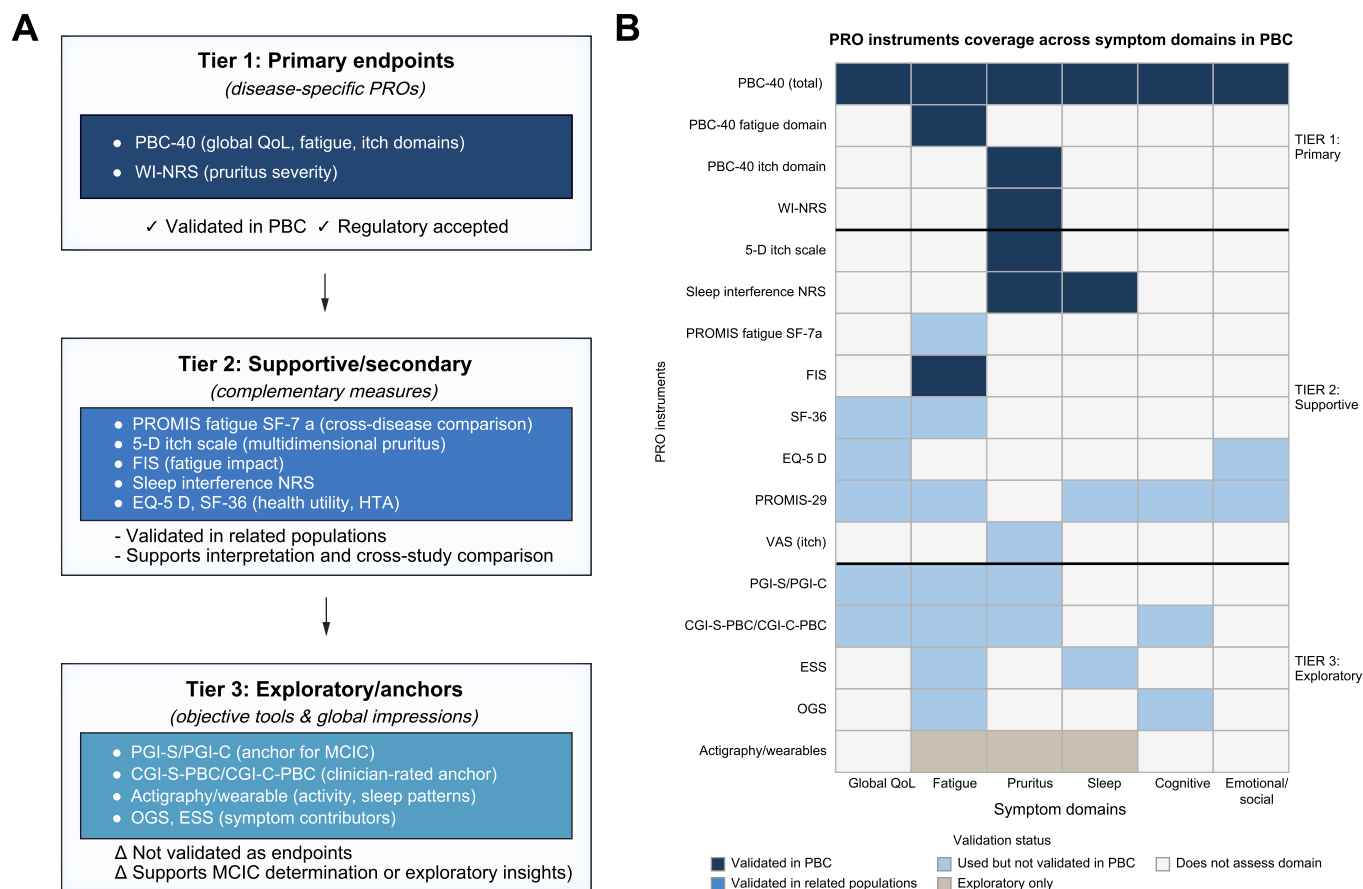


Fig. 3. Use of PROs in the assessment of quality of life in people living with PBC (A) Tiered framework for patient-reported outcome assessment in PBC. Hierarchical classification of PRO instruments recommended for use in PBC clinical trials and real-world studies. Tier 1 (primary endpoints): disease-specific PROs validated in PBC and accepted by regulatory agencies, including the PBC-40 questionnaire (assessing global quality of life, fatigue, and itch domains) and the WI-NRS for pruritus severity. Tier 2 (supportive/secondary): complementary measures validated in related populations that support interpretation and cross-study comparison, including PROMIS Fatigue SF-7a, 5-D itch scale, FIS, sleep interference NRS, and generic health utility instruments (EQ-5D, SF-36). Tier 3 (exploratory/anchors): objective tools and global impression instruments not yet validated as endpoints but useful for supporting MCIC determination, including PGI-S/PGI-C, CGI-S-PBC/CGI-C-PBC, actigraphy/wearables, OGS, and ESS. (B) PRO instrument coverage across symptom domains in PBC. Heatmap displaying the coverage and validation status of PRO instruments across key symptom domains in PBC. Instruments are organised vertically by tier (Tier 1: primary; Tier 2: supportive; Tier 3: exploratory) and mapped horizontally against six symptom domains: global QoL, fatigue, pruritus, sleep, cognitive function, and emotional/social well-being. Colour intensity indicates validation status: dark blue = validated in PBC; medium blue = validated in related populations; light blue = used but not validated in PBC; grey = exploratory only; white = does not assess domain. The PBC-40 (total) provides the most comprehensive coverage across domains with PBC-specific validation, while the WI-NRS offers robust pruritus-specific assessment. Supportive instruments provide complementary domain coverage, and exploratory tools offer additional insights for MCIC determination and symptom characterisation. 5-D Itch Scale, 5-Dimensional Itch Scale; CGI-C, Clinical Global Impression of Change; CGI-S, Clinical Global Impression of Severity; ESS, Epworth Sleepiness Scale; EQ-5D, EuroQol 5-Dimensions questionnaire; FIS, Fatigue Impact Scale; MCIC, minimal clinically important change; NRS, Numerical Rating Scale; OGS, Orthostatic Grading Scale; PBC, primary biliary cholangitis; PBC-40, primary biliary cholangitis 40-item questionnaire; PGI-C, Patient Global Impression of Change; PGI-S, Patient Global Impression of Severity; PROMIS, Patient-Reported Outcomes Measurement Information System; PRO, patient-reported outcome; QoL, quality of life; SF-36, Short Form-36 Health Survey; SF-7a, Short Form 7a (fatigue); WI-NRS, Worst Itch Numerical Rating Scale.

During the consensus discussions, the panel endorsed a tiered strategy for fatigue assessment in PBC: the PBC-40 fatigue domain as the core disease-specific instrument; PROMIS Fatigue SF-7a as a complementary measure ensuring regulatory comparability; and supportive scales such as FIS, PGI/CGI, HADS, ESS, and OGS to capture comorbid contributors and aid interpretation. Objective measures like actigraphy were considered promising adjuncts for exploratory endpoints. These deliberations guided the following recommendations (Fig. 3A,B).

Recommendations

- 3.4.1. Fatigue should be assessed using PROs that attempt to capture both central and peripheral components, recognising the multifactorial and individualised nature of fatigue in PBC (**LoE 2; strong recommendation, 100% agreement**).
- 3.4.2. The PBC-40 fatigue domain should be used as the primary disease-specific measure of fatigue in PBC (**LoE 2; strong recommendation, 96% agreement**).

- 3.4.3. The PROMIS Fatigue Short Form 7a (PSF-7a) should be included as a complementary fatigue measure (**LoE 5; strong recommendation, 98% agreement**).
- 3.4.4. The Fatigue Impact Scale (FIS) and its modified versions may be used as supportive tools (**LoE 5; weak recommendation, 100% agreement**).
- 3.4.5. Patient- and clinician-rated global impression tools (e.g. PGI-S, PGI-C, CGI-S/C-PBC) may support interpretation of meaningful within-patient change but should not replace PROs as the primary outcome measure (**LoE 5; strong recommendation, 100% agreement**).
- 3.4.6. Additional non-PBC-specific tools (e.g. Hospital Anxiety and Depression Scale [HADS], Epworth Sleepiness Scale [ESS] and Orthostatic Grading Scale [OGS]) may complement fatigue assessment (**LoE 5; weak recommendation, 98% agreement**).
- 3.4.7. Objective tools such as actigraphy or wearable sensors may be considered for exploratory or secondary endpoints in PBC (**LoE 5; weak recommendation, 98% agreement**).

Research agenda

- Establish robust thresholds for meaningful within-patient and between-group change in all applicable PROs.
- Align disease-specific and generic instruments (e.g. PBC-40, WI-NRS, PROMIS, PGI-S/PGI-C) to enable comparability across trials, registries, and regulatory submissions.
- Establish validated MCIC thresholds for PBC-40 fatigue domain and PROMIS Fatigue SF-7a using anchor-based methods.
- Test PRO instruments across diverse languages, cultures, and health systems, and confirm responsiveness to emerging therapies in both trial and real-world settings.

Abbreviations

AASLD, American Association for the Study of Liver Diseases; ALP, alkaline phosphatase; ALEH, Asociación Latinoamericana para el Estudio del Hígado; CGI-C, Clinical Global Impression-Change; CGI-C-PBC, Clinical Global Impression-Change for Primary Biliary Cholangitis; CGI-S, Clinical Global Impression-Severity; CGI-S-PBC, Clinical Global Impression-Severity for Primary Biliary Cholangitis EASL, European Association for the Study of the Liver; EHR, electronic health record; EQ-5D, EuroQoL 5-Dimension; ESS, Epworth Sleepiness Scale; FIS, Fatigue Impact Scale; HADS, Hospital Anxiety and Depression Scale; LoE, level of evidence; LSM, liver stiffness measurement; MCIC, minimal clinically important change; OCA, obeticholic acid; OCEBM, Oxford Centre for Evidence-Based Medicine; OGS, Orthostatic Grading Scale; PBC, primary biliary cholangitis; PBC-40, Primary Biliary Cholangitis-40 questionnaire; PGI-C, Patient Global Impression of Change; PGI-S, Patient Global Impression of Severity; PPAR, peroxisome proliferator-activated receptor; PRO, patient-reported outcome; PROMIS, Patient-Reported Outcomes Measurement Information System; PSF-7a, PROMIS Fatigue Short Form 7a; RCT, randomised controlled trial; RWD, real-world data; RWE, real-world evidence; SOLDA, Society on Liver Disease in Africa; UDCA, ursodeoxycholic acid; ULN, upper limit of normal; VCTE, vibration-controlled transient elastography; WI-NRS, Worst Itch Numerical Rating Scale.

Financial support

The authors received no financial support to produce this manuscript.

- Evaluate the role of multidimensional tools (e.g. 5-D Itch Scale, FIS), global impression anchors, and exploratory objective measures (e.g. actigraphy) to capture domains not fully covered by current PROs.

Conclusion

Drug development in PBC faces significant, ethical, and operational challenges driven by slow disease progression, limited patient populations, and the absence of fully validated surrogate endpoints. Through a rigorous, multi-stakeholder modified Delphi process, this EASL-AASLD consensus provides a pragmatic framework to address these challenges. The panel reached strong agreement on the value of improvements in liver biochemistries in predicting clinical outcomes regardless of therapy and the complementary value of non-invasive fibrosis measures such as LSM. Importantly, the consensus highlights the need to define what additional data is required to move these biomarkers to validated surrogates for clinical trial efficacy endpoints.

In parallel, this work articulates clear principles for the generation of regulatory-grade real-world evidence, recognising its essential role in confirmatory effectiveness and safety assessments when RCTs are not feasible. Equally central is the systematic integration of validated PROs, ensuring that symptom burden, quality of life, and treatment impact are captured from the patient perspective.

Collectively, these recommendations offer a harmonised roadmap to support innovation, regulatory decision-making, and patient-centred care in PBC. Their implementation will require sustained collaboration among clinicians, patients, regulators, and industry, and should accelerate the development of safe, effective, and durable therapies for this rare but impactful disease. While tailored to PBC, the principles articulated herein may also inform endpoint selection, RWE generation, and PRO assessment across other rare cholestatic liver diseases facing similar barriers to therapeutic development.

Conflicts of interest

The authors declare no conflicts of interest pertaining to this manuscript.

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

Authors' contributions

Author contributions are reported according to the CRediT (Contributor Roles Taxonomy) framework. **Conceptualisation:** Christopher L. Bowlus, Marco Carbone, Christophe Corpechot, Claudia A. Couto, Bettina E. Hansen, Gideon Hirschfield, David E. Jones, Cynthia Levy, Ana Lleo, Marlyn J. Mayo, Alejandra Villamil, José Willemse. **Methodology:** Christopher L. Bowlus, Marco Carbone, Christophe Corpechot, Claudia A. Couto, Bettina E. Hansen, Gideon Hirschfield, David E. Jones, Cynthia Levy, Ana Lleo, Adrielly Martins, Marlyn J. Mayo, Miki Scaravaglio, Alejandra Villamil, José Willemse. **Data Curation:** Adrielly Martins, Miki Scaravaglio. **Visualisation:** Adrielly Martins, Miki Scaravaglio. **Writing – Original Draft Preparation:** Cynthia Levy, Ana Lleo, Adrielly Martins, Miki Scaravaglio. **Writing – Review & Editing:** Femi Adekunle, Andrey Belous, Therese Bittermann, Christopher L. Bowlus, Marco Carbone, Nora Cazzagon, Christophe Corpechot, Claudia A. Couto, Emma L. Culver, Sandrine Cure, Magnus Doverskog, Sven Francque, Michael W. Fried, Steve Gao, Bettina E. Hansen, Gideon Hirschfield, David E. Jones, Kris V. Kowdley, Andreas Kremer, Angela Leburgue, Cynthia Levy, Ana Lleo, Maria-Carlota Londono, Adrielly Martins, Marlyn J. Mayo, Megan McLaughlin, Adrian van der Meer, George F. Mells, Seema Meloni, Robert Michell-Thain, Aldo J. Montano-Loza, Paul

Nitschmann, Tiago Nunes, Deven Parmar, Vlad Ratzu, Carol Roberts, Stephen Rossi, Miki Scaravaglio, Francesca Scarpa, Christoph Schramm, Marwan Sleiman, Jan Span, C. Wendy Spearman, Mark Swain, Atsushi Tanaka, Ingo van Thiel, Douglas Thorburn, Palak J. Trivedi, Philip Troke, Alejandra Villamil, Martin Weltman, José Willemse, Gail Wright, Xiong Ma. **Project Administration:** Cynthia Levy, Ana Lleo, Adrielly Martins, Miki Scaravaglio.

Endorsing organisations

AASLD, American Association for the Study of Liver Diseases; ALEH, Asociación Latinoamericana para el Estudio del Hígado; EASL, European Association for the Study of the Liver; SOLDA, Society on Liver Disease in Africa.

Acknowledgements

Wenjing Wang, for the support in screening references and synthesising data in working group 2 systematic literature review.

Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jhep.2026.03.038>.

Collaborators

Chairs

Cynthia Levy: Division of Digestive Health and Liver Diseases, Schiff Center for Liver Diseases, University of Miami Miller School of Medicine, Miami, FL, USA
Ana Lleo: Department of Biomedical Sciences, Humanitas University; Division of Internal Medicine and Hepatology, Department of Gastroenterology, IRCCS Humanitas Research Hospital, Rozzano, Milan, Italy

Secretary

Adrielly Martins: Schiff Center for Liver Diseases, Division of Digestive Health and Liver Diseases, University of Miami Miller School of Medicine, Miami, FL, USA
Miki Scaravaglio: Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK

Steering committee

Christopher L. Bowlus: Division of Gastroenterology and Hepatology, University of California Davis School of Medicine, Sacramento, CA, USA
Marco Carbone: Department of Medicine and Surgery, University of Milano-Bicocca; Hepatology and Gastroenterology Unit, ASST Grande Ospedale Metropolitano Niguarda, Milan, Italy
Christophe Corpechot: Reference Center for Inflammatory Biliary Diseases and Autoimmune Hepatitis, Saint-Antoine Hospital, Assistance Publique – Hôpitaux de Paris (APHP); INSERM UMR_S938, Sorbonne University, Paris, France
Claudia A. Couto: Instituto Alfa de Gastroenterologia, Hospital das Clínicas, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil
Bettina E. Hansen: Department of Epidemiology and Biostatistics, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands; Institute of Health Policy, Management and Evaluation (IHPE), University of Toronto, Toronto, Canada
Gideon Hirschfield: Toronto Centre for Liver Disease, Toronto General Hospital, University Health Network; Division of Gastroenterology, University of Toronto, Toronto, Canada
David E. Jones: Translational and Clinical Research Institute, Newcastle University; Newcastle upon Tyne Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK
Marlyn J. Mayo: Division of Digestive and Liver Diseases, Department of Internal Medicine, University of Texas Southwestern Medical Center, Dallas, TX, USA
Alejandra Villamil: Liver Autoimmunity Unit, Hepatology Section, Hospital Italiano de Buenos Aires, Buenos Aires, Argentina
José Willemse: Patient Representative, ERDERA (European Rare Disease Research Alliance), The Netherlands

Working groups

Femi Adekunle: Advanz Pharma, London, UK
Andrey Belous: Kowa Pharma Europe, Germany
Therese Bittermann: Division of Gastroenterology and Hepatology, Penn Medicine, University of Pennsylvania, Philadelphia, PA, USA

Nora Cazzagon: Department of Surgery, Oncology and Gastroenterology, University of Padova; Padova University Hospital, Padova, Italy
Emma L. Culver: Translational Gastroenterology and Liver Unit, Oxford University Hospitals NHS Foundation Trust; NIHR Oxford Biomedical Research Centre, Oxford, UK
Sandrine Cure: Advanz Pharma, Ireland
Magnus Doverskog: Umeçrine Cognition AB, Solna, Sweden
Sven Francque: Department of Gastroenterology and Hepatology, Antwerp University Hospital; University of Antwerp, Antwerp, Belgium
Michael W. Fried: Division of Gastroenterology and Hepatology, University of North Carolina School of Medicine, Chapel Hill, NC, USA
Steve Gao: Gilead Sciences, Foster City, CA, USA
Kris V. Kowdley: Liver Institute Northwest, Seattle, WA, USA
Andreas Kremer: Department of Gastroenterology and Hepatology, University Hospital Zurich, University of Zurich, Zurich, Switzerland
Angela Leburgue: Patient Representative, ALBI France (Association pour la lutte contre les maladies Inflammatoires du foie et des Voies Biliaires), France
Maria-Carlota Londono: Liver Unit, Hospital Clínic Barcelona; FRCB-IDIBAPS; CIBERehd, Barcelona, Spain
Megan McLaughlin: GlaxoSmithKline, Brentford, UK
Adrian van der Meer: Department of Gastroenterology and Hepatology, Erasmus MC, University Medical Center Rotterdam, Rotterdam, the Netherlands
George F. Mells: Academic Department of Medical Genetics, University of Cambridge; Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK
Seema Meloni: Ipsen, Paris, France
Robert Michell-Thain: Patient Representative, PBC Foundation, Edinburgh, UK
Aldo J. Montano-Loza: Division of Gastroenterology and Hepatology, University of Alberta, Edmonton, AB, Canada
Paul Nitschmann: Intercept Pharmaceuticals, Morristown, NJ, USA
Tiago Nunes: Mirum Pharmaceuticals, Paris, France
Deven Parmar: Zydus Therapeutics, Bridgewater, NJ, USA
Vlad Ratzu: Sorbonne University; Hôpital Pitié-Salpêtrière, APHP, Paris, France
Carol Roberts: Patient Representative, PBCers Organization, USA
Stephen Rossi: Patient Representative, PSC Partners Seeking a Cure, Denver, CO, USA
Francesca Scarpa: Patient Representative, AMAF Italy (Associazione Malati Autoimmuni del Fegato), Milan, Italy
Christoph Schramm: Department of Medicine and Hamburg Center for Translational Immunology (HCTI), University Medical Center Hamburg-Eppendorf (UKE), Hamburg, Germany
Marwan Sleiman: Ipsen, Paris, France
Jan Span: Medicines Evaluation Board (CBG-MEB), Utrecht, the Netherlands
C. Wendy Spearman: Division of Hepatology, Department of Medicine, University of Cape Town, Cape Town, South Africa
Mark Swain: Division of Gastroenterology and Hepatology, Department of Medicine, University of Calgary, Calgary, AB, Canada
Atsushi Tanaka: Department of Medicine, Teikyo University School of Medicine, Tokyo, Japan
Ingo van Thiel: Patient Representative, Deutsche Leberhilfe e.V., Cologne, Germany
Douglas Thorburn: Sheila Sherlock Liver Centre, Royal Free Hospital NHS Foundation Trust; UCL Institute for Liver and Digestive Health, London, UK
Palak J. Trivedi: NIHR Birmingham Biomedical Research Centre; Centre for Liver and Gastroenterology Research, University of Birmingham, Birmingham, UK
Philip Troke: GlaxoSmithKline, Brentford, UK
Martin Weltman: Department of Gastroenterology and Hepatology, Nepean Hospital, University of Sydney, Sydney, NSW, Australia
Gail Wright: Patient Representative, PBC Society Canada, Toronto, Canada
Xiong Ma: Department of Gastroenterology, Renji Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China

References

Author names in bold designate shared co-first authorship

- [1] European Association for the Study of the Liver. EASL Clinical Practice Guidelines: the diagnosis and management of patients with primary biliary cholangitis. *J Hepatol* 2017;67(1):145–172.
- [2] Lindor KD, Bowlus CL, Boyer J, et al. Primary biliary cholangitis: 2018 practice guidance from the American association for the study of liver diseases. *Hepatology* 2019;69(1):394–419.
- [3] Tanaka A, Ma X, Takahashi A, et al. Primary biliary cholangitis. *Lancet* 2024;404(10457):1053–1066.

- [4] Harms MH, van Buuren HR, Corpechot C, et al. Ursodeoxycholic acid therapy and liver transplant-free survival in patients with primary biliary cholangitis. *J Hepatol* 2019;71(2):357–365.
- [5] Webb GJ, Rana A, Hodson J, et al. Twenty-year comparative analysis of patients with autoimmune liver diseases on transplant waitlists. *Clin Gastroenterol Hepatol* 2018;16(2):278–287.e7.
- [6] Parés A, Caballería L, Rodés J. Excellent long-term survival in patients with primary biliary cirrhosis and biochemical response to ursodeoxycholic Acid. *Gastroenterology* 2006;130(3):715–720.
- [7] Angulo P, Batts KP, Thorneau TM, et al. Long-term ursodeoxycholic acid delays histological progression in primary biliary cirrhosis. *Hepatology* 1999;29(3):644–647.
- [8] Poupon RE, Lindor KD, Cauch-Dudek K, et al. Combined analysis of randomized controlled trials of ursodeoxycholic acid in primary biliary cirrhosis. *Gastroenterology* 1997;113(3):884–890.
- [9] Corpechot C, Abenavoli L, Rabahi N, et al. Biochemical response to ursodeoxycholic acid and long-term prognosis in primary biliary cirrhosis. *Hepatology* 2008;48(3):871–877.
- [10] Kumagi T, Guindi M, Fischer SE, et al. Baseline ductopenia and treatment response predict long-term histological progression in primary biliary cirrhosis. *Am J Gastroenterol* 2010;105(10):2186–2194.
- [11] European Medicines Agency. Conditional marketing authorisation (glossary) 2024. Available from: <https://www.ema.europa.eu/en/glossary/conditional-marketing-authorisation>.
- [12] U.S. Food and Drug Administration. Approval letter – OCALIVA (obeticholic acid) tablets, NDA 207999, accelerated approval. In: Center for drug evaluation and research. U.S. Department of Health and Human Services; 2016.
- [13] European Medicines Agency. Summary of opinion (initial authorisation): Ocaliva (obeticholic acid). In: Committee for medicinal products for human use (CHMP); 2016.
- [14] U.S. Food and Drug Administration. Approval letter – livdelzi (seladelpar) capsules, NDA 217899, accelerated approval. In: Center for drug evaluation and research. U.S. Department of Health and Human Services; 2024.
- [15] European Medicines Agency. Summary of opinion (initial authorisation): seladelpar Gilead (seladelpar). In: Committee for medicinal products for human use (CHMP); 2024.
- [16] U.S. Food and Drug Administration. Approval letter – iqirvo (elafibranor) tablets, NDA 218860, accelerated approval. In: Center for drug evaluation and research. U.S. Department of Health and Human Services; 2024.
- [17] European Medicines Agency. Summary of opinion (initial authorisation): iqirvo (elafibranor). In: Committee for medicinal products for human use (CHMP); 2024.
- [18] U.S. Food and Drug Administration. Demonstrating substantial evidence of effectiveness for human drug and biological products: guidance for industry (draft guidance). In: Center for drug evaluation and research (CDER) CfBEaRC. U.S. Department of Health and Human Services; 2019.
- [19] U.S. Food and Drug Administration. Demonstrating substantial evidence of effectiveness with one adequate and well-controlled clinical investigation and confirmatory evidence: guidance for industry (draft guidance). In: Center for drug evaluation and research (CDER) CfBEaRC, oncology center of excellence (OCE). U.S. Department of Health and Human Services; 2023.
- [20] U.S. Food and Drug Administration. Accelerated approval and considerations for determining whether a confirmatory trial is underway: guidance for industry (draft guidance). In: Oncology center of excellence. U.S. Department of Health and Human Services; 2025.
- [21] Corpechot C, Chazouillères O, Rousseau A, et al. A placebo-controlled trial of bezafibrate in primary biliary cholangitis. *N Engl J Med* 2018;378(23):2171–2181.
- [22] de Vries E, Bolier R, Goet J, et al. Fibrates for itch (FITCH) in fibrosing cholangiopathies: a double-blind, randomized, placebo-controlled trial. *Gastroenterology* 2021;160(3):734–743.e6.
- [23] Honda A, Tanaka A, Kaneko T, et al. Bezafibrate improves GLOBE and UK-PBC scores and long-term outcomes in patients with primary biliary cholangitis. *Hepatology* 2019;70(6):2035–2046.
- [24] Kowdley KV, Hirschfield GM, Coombs C, et al. COBALT: a confirmatory trial of obeticholic acid in primary biliary cholangitis with placebo and external controls. *Am J Gastroenterol* 2025;120(2):390–400.
- [25] Murillo Perez CF, Fisher H, Hiu S, et al. Greater transplant-free survival in patients receiving obeticholic acid for primary biliary cholangitis in a clinical trial setting compared to real-world external controls. *Gastroenterology* 2022;163(6):1630–1642.e3.
- [26] Abbas N, Smith R, Lytvyak E, et al. Non-response to obeticholic acid is associated with heightened risks of developing clinical events in primary biliary cholangitis. *Aliment Pharmacol Ther* 2026;63(4):494–506.
- [27] U.S. Food and Drug Administration. FDA adds Boxed Warning to highlight correct dosing of Ocaliva (obeticholic acid) for patients with a rare chronic liver disease. U.S. Department of Health and Human Services; 2018. Available from: <https://www.fda.gov/drugs/drug-safety-and-availability/fda-adds-boxed-warning-highlight-correct-dosing-ocaliva-obeticholic-acid-patients-rare-chronic-liver>.
- [28] U.S. Food and Drug Administration. Due to risk of serious liver injury, FDA restricts use of Ocaliva (obeticholic acid) in primary biliary cholangitis (PBC) patients with advanced cirrhosis. U.S. Department of Health and Human Services; 2021. Available from: <https://www.fda.gov/drugs/drug-safety-and-availability/due-risk-serious-liver-injury-fda-restricts-use-ocaliva-obeticholic-acid-primary-biliary-cholangitis>.
- [29] U.S. Food and Drug Administration. Serious liver injury being observed in patients without cirrhosis taking Ocaliva (obeticholic acid) to treat primary biliary cholangitis. U.S. Department of Health and Human Services; 2024. Available from: <https://www.fda.gov/drugs/drug-safety-and-availability/serious-liver-injury-being-observed-patients-without-cirrhosis-taking-ocaliva-obeticholic-acid-treat>.
- [30] European Medicines Agency. Revocation of conditional marketing authorisation for Ocaliva (obeticholic acid). In: (CHMP) CfMPfHu; 2024.
- [31] Intercept Pharmaceuticals I. Intercept announces voluntary withdrawal of OCALIVA® for primary biliary cholangitis (PBC) from the US market; US clinical trials involving obeticholic acid placed on clinical hold. *GlobeNewswire*; 2025. Available from: <https://www.globenewswire.com/news-release/2025/09/11/3148535/23024/en/Intercept-Announces-Voluntary-Withdrawal-of-OCALIVA-for-Primary-Biliary-Cholangitis-PBC-from-the-US-Market-US-Clinical-Trials-Involving-Obeticholic-Acid-Placed-on-Clinical-Hold.html>.
- [32] Boulkedid R, Abdoul H, Loustau M, et al. Using and reporting the Delphi method for selecting healthcare quality indicators: a systematic review. *PLoS One* 2011;6(6):e20476.
- [33] Diamond IR, Grant RC, Feldman BM, et al. Defining consensus: a systematic review recommends methodologic criteria for reporting of Delphi studies. *J Clin Epidemiol* 2014;67(4):401–409.
- [34] Beiderbeck D, Frevel N, von der Gracht HA, et al. Preparing, conducting, and analyzing Delphi surveys: cross-disciplinary practices, new directions, and advancements. *MethodsX* 2021;8:101401.
- [35] Covidence systematic review software. Melbourne, Australia: Veritas Health Innovation.
- [36] National Institute for Health and Care Excellence. The guidelines manual. Chapter 9: developing and wording guideline recommendations. 2012. Available from: <https://www.nice.org.uk/process/pmg6/chapter/developing-and-wording-guideline-recommendations>.
- [37] Fleming TR, Powers JH. Biomarkers and surrogate endpoints in clinical trials. *Stat Med* 2012;31(25):2973–2984.
- [38] Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. *Stat Med* 1989;8(4):431–440.
- [39] Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? *Ann Intern Med* 1996;125(7):605–613.
- [40] U.S. Food and Drug Administration. Surrogate endpoint resources for drug and biologic development. U.S. Department of Health and Human Services; 2018. Available from: <https://www.fda.gov/drugs/development-resources/surrogate-endpoint-resources-drug-and-biologic-development>.
- [41] European Medicines Agency. Guideline on clinical trials in small populations. In: Committee for medicinal products for human use (CHMP); 2006.
- [42] Seetharam S, Sussman NL, Komoda T, et al. The mechanism of elevated alkaline phosphatase activity after bile duct ligation in the rat. *Hepatology* 1986;6(3):374–380.
- [43] Moss DW. Physicochemical and pathophysiological factors in the release of membrane-bound alkaline phosphatase from cells. *Clin Chim Acta* 1997;257(1):133–140.
- [44] Poupon R. Liver alkaline phosphatase: a missing link between cholestasis and biliary inflammation. *Hepatology* 2015;61(6):2080–2090.
- [45] Jones DEJ, Wetten A, Barron-Millar B, et al. The relationship between disease activity and UDCA response criteria in primary biliary cholangitis: a cohort study. *EBioMedicine* 2022;80:104068.
- [46] Lammers WJ, van Buuren HR, Hirschfield GM, 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(6):1338–1349.e5. quiz e15.
- [47] Lammers WJ, Hirschfield GM, Corpechot C, et al. Development and validation of a scoring system to predict outcomes of patients with primary biliary cirrhosis receiving ursodeoxycholic acid therapy. *Gastroenterology* 2015;149(7):1804–1812.e4.
 - [48] Carbone M, Sharp SJ, Flack S, 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 cholangitis. *Hepatology* 2016;63(3):930–950.
 - [49] Harms MH, Lammers WJ, Thorburn D, et al. Major hepatic complications in ursodeoxycholic acid-treated patients with primary biliary cholangitis: risk factors and time trends in incidence and outcome. *Am J Gastroenterol* 2018;113(2):254–264.
 - [50] Terziroli Beretta-Piccoli B, Stirnimann G, Cerny A, et al. Geoepidemiology of primary biliary cholangitis: lessons from Switzerland. *Clin Rev Allergy Immunol* 2018;54(2):295–306.
 - [51] Murillo Perez CF, Hirschfield GM, Corpechot C, et al. Fibrosis stage is an independent predictor of outcome in primary biliary cholangitis despite biochemical treatment response. *Aliment Pharmacol Ther* 2019;50(10):1127–1136.
 - [52] Murillo Perez CF, Harms MH, Lindor KD, et al. Goals of treatment for improved survival in primary biliary cholangitis: treatment target should be bilirubin within the normal range and normalization of alkaline phosphatase. *Am J Gastroenterol* 2020;115(7):1066–1074.
 - [53] Gerussi A, Bernasconi DP, O'Donnell SE, et al. Measurement of gamma glutamyl transferase to determine risk of liver transplantation or death in patients with primary biliary cholangitis. *Clin Gastroenterol Hepatol* 2021;19(8):1688–1697.e14.
 - [54] Osman KT, Maselli DB, Idilman IS, et al. Liver stiffness measured by either magnetic resonance or transient elastography is associated with liver fibrosis and is an independent predictor of outcomes among patients with primary biliary cholangitis. *J Clin Gastroenterol* 2021;55(5):449–457.
 - [55] Pinyopompanish K, Chadalavada P, Talal Sarmini M, et al. Simplified 6-month prediction scores for primary biliary cholangitis patients treated with ursodeoxycholic acid. *Eur J Gastroenterol Hepatol* 2022;34(4):411–416.
 - [56] Diaz-González Á, Rodríguez-Tajes S, Gómez MV, et al. THU482 - should we seek complete liver tests normalization in primary biliary cholangitis? Data from ColHai registry. *J Hepatol* 2022;77:S334.
 - [57] Takeuchi A, Matsumoto K, Komori A, et al. WED-293 - prognostic value of serum ALP levels during additional treatment in Japanese patients with primary biliary cholangitis treated with ursodeoxycholic acid and bezafibrate. *J Hepatol* 2023;78:S394–S395.
 - [58] Corpechot C, Lemoine S, Soret PA, et al. Adequate versus deep response to ursodeoxycholic acid in primary biliary cholangitis: to what extent and under what conditions is normal alkaline phosphatase level associated with complication-free survival gain? *Hepatology* 2024;79(1):39–48.
 - [59] Roberts SB, Choi WJ, Worobetz L, et al. Loss of biochemical response at any time worsens outcomes in UDCA-treated patients with primary biliary cholangitis. *JHEP Rep* 2024;6(10):101168.
 - [60] Kowdley KV, Victor 3rd DW, MacEwan JP, et al. Longitudinal relationship between elevated liver biochemical tests and negative clinical outcomes in primary biliary cholangitis: a population-based study. *Aliment Pharmacol Ther* 2025;61(11):1775–1784.
 - [61] de Vincentis A, Terracciani F, D'Amato D, et al. THU-272 Appraising improvement in prediction of clinical outcomes based on different response criteria and liver stiffness measurements in patients with primary biliary cholangitis treated with second-line therapies. *J Hepatol* 2025;82:S307–S308.
 - [62] Werner E, Weijsters G, van Hooff MC, et al. THU-296 Evaluation of response to Ursodeoxycholic acid (UDCA) in a nationwide population with primary biliary cholangitis. *J Hepatol* 2025;82:S316–S317.
 - [63] Hernández-Évole H, Salcedo M, Olivas I, et al. OS-086 Evaluation of the response criteria following second-line treatment with bezafibrate in patients with primary biliary cholangitis. *J Hepatol* 2025;82:S59–S60.
 - [64] Ronca V, de Vincentis A, Terracciani F, et al. OS-088 FIOCA study: comparative analysis of transplant-free survival in primary biliary cholangitis patients treated with obeticholic acid or fibrates. *J Hepatol* 2025;82:S60–S61.
 - [65] Nevens F, Lindor KD, Jones DE. Obeticholic acid in primary biliary cholangitis. *N Engl J Med* 2016;375(20):e41.
 - [66] Hirschfield GM, Bowlus CL, Mayo MJ, et al. A phase 3 trial of seladelpar in primary biliary cholangitis. *N Engl J Med* 2024;390(9):783–794.
 - [67] Kowdley KV, Bowlus CL, Levy C, et al. Efficacy and safety of elafibranor in primary biliary cholangitis. *N Engl J Med* 2024;390(9):795–805.
 - [68] Tanaka A, Hirohara J, Nakanuma Y, et al. Biochemical responses to bezafibrate improve long-term outcome in asymptomatic patients with primary biliary cirrhosis refractory to UDCA. *J Gastroenterol* 2015;50(6):675–682.
 - [69] Azemoto N, Kumagi T, Abe M, et al. Biochemical response to ursodeoxycholic acid predicts long-term outcome in Japanese patients with primary biliary cirrhosis. *Hepatol Res* 2011;41(4):310–317.
 - [70] Haldar D, Janmohamed A, Plant T, et al. Antibodies to gp210 and understanding risk in patients with primary biliary cholangitis. *Liver Int* 2021;41(3):535–544.
 - [71] de Veer RC, Harms MH, Corpechot C, et al. Liver transplant-free survival according to alkaline phosphatase and GLOBE score in patients with primary biliary cholangitis treated with ursodeoxycholic acid. *Aliment Pharmacol Ther* 2022;56(9):1408–1418.
 - [72] Wong YJ, Lam L, Soret PA, et al. Prognostic value of liver stiffness measurement vs. biochemical response in primary biliary cholangitis. *J Hepatol* 2026;84(2):275–283.
 - [73] Corpechot C, Chazouillères O, Poupon R. Early primary biliary cirrhosis: biochemical response to treatment and prediction of long-term outcome. *J Hepatol* 2011;55(6):1361–1367.
 - [74] Lammert C, Juran BD, Schlacht E, et al. Biochemical response to ursodeoxycholic acid predicts survival in a North American cohort of primary biliary cirrhosis patients. *J Gastroenterol* 2014;49(10):1414–1420.
 - [75] Yang F, Yang Y, Wang Q, et al. The risk predictive values of UK-PBC and GLOBE scoring system in Chinese patients with primary biliary cholangitis: the additional effect of anti-gp210. *Aliment Pharmacol Ther* 2017;45(5):733–743.
 - [76] Efe C, Taşçılar K, Henriksson I, et al. Validation of risk scoring systems in ursodeoxycholic acid-treated patients with primary biliary cholangitis. *Am J Gastroenterol* 2019;114(7):1101–1108.
 - [77] Pariente A, Chazouillères O, Causse X, et al. Management of primary biliary cholangitis: results from a large real-life observational study in France and Belgium. *Eur J Gastroenterol Hepatol* 2021;33(1S Suppl 1):e197–e205.
 - [78] de Veer RC, van Hooff MC, Corpechot C, et al. Ursodeoxycholic acid treatment-induced GLOBE score changes are associated with liver transplantation-free survival in patients with primary biliary cholangitis. *Am J Gastroenterol* 2023;118(7):1196–1203.
 - [79] Ding D, Hu Y, Jia G, et al. Low-risk individuals with primary biliary cholangitis and significant liver stiffness: prognosis and treatment. *Hepatol Int* 2025;19(3):673–681.
 - [80] Goet JC, Murillo Perez CF, Harms MH, et al. A comparison of prognostic scores (Mayo, UK-PBC, and GLOBE) in primary biliary cholangitis. *Am J Gastroenterol* 2021;116(7):1514–1522.
 - [81] Gatselis NK, Goet JC, Zachou K, et al. Factors associated with progression and outcomes of early stage primary biliary cholangitis. *Clin Gastroenterol Hepatol* 2020;18(3):684–692.e6.
 - [82] European Medicines Agency. Reflection paper on regulatory requirements for the development of medicinal products for primary biliary cholangitis (PBC) and primary sclerosing cholangitis (PSC). In: Committee for medicinal products for human use (CHMP); 2023. Amsterdam, The Netherlands.
 - [83] Corpechot C, Carrat F, Poujol-Robert A, et al. Noninvasive elastography-based assessment of liver fibrosis progression and prognosis in primary biliary cirrhosis. *Hepatology* 2012;56(1):198–208.
 - [84] Corpechot C, Carrat F, Gaouar F, et al. Liver stiffness measurement by vibration-controlled transient elastography improves outcome prediction in primary biliary cholangitis. *J Hepatol* 2022;77(6):1545–1553.
 - [85] Lam L, Soret PA, Lemoine S, Hansen B, Hirschfield G, Gulamhusein A, et al. Dynamics of liver stiffness measurement and clinical course of primary biliary cholangitis. *Clin Gastroenterol Hepatol* 2024;22(12):2432–2441.e2.
 - [86] Ding D, Guo G, Cui L, et al. Prognostic significance of liver stiffness in patients with primary biliary cholangitis: validation of Baveno VII criteria. *Hepatol Int* 2024;18(1):206–215.
 - [87] Mayo MJ, Parkes J, Adams-Huet B, et al. Prediction of clinical outcomes in primary biliary cirrhosis by serum enhanced liver fibrosis assay. *Hepatology* 2008;48(5):1549–1557.
 - [88] Fujinaga Y, Namisaki T, Takaya H, et al. Enhanced liver fibrosis score as a surrogate of liver-related complications and mortality in primary biliary cholangitis. *Medicine (Baltimore)* 2021;100(39):e27403.
 - [89] Dang A. Real-world evidence: a primer. *Pharmaceut Med* 2023;37(1):25–36.
 - [90] Hennessy S, Atsuta Y, Hill S, et al. Real-world data and real-world evidence in regulatory decision making: report summary from the council for

- international organizations of medical Sciences (CIOMS) working group XIII. *Pharmacoepidemiol Drug Saf* 2025;34(3):e70117.
- [91] U.S. Food and Drug Administration. Real-world evidence 2025. Available from: <https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence>.
 - [92] European Medicines Agency. Real-world evidence provided by EMA: support for regulatory decision-making. In: Data analytics and methods task force R-WETT-R; 2024. Amsterdam, The Netherlands.
 - [93] Jones DEJ, Beuers U, Bonder A, et al. Primary biliary cholangitis drug evaluation and regulatory approval: where do we go from here? *Hepatology* 2024;80(5):1291–1300.
 - [94] Pradhan R, Abrahami D, Yin H, et al. Defining clinically relevant target populations using real-world data to guide the design of representative antidiabetic drug trials. *Clin Pharmacol Ther* 2021;109(5):1219–1223.
 - [95] Lugo-Palacios DG, Bidulka P, O'Neill S, et al. Going beyond randomised controlled trials to assess treatment effect heterogeneity across target populations. *Health Econ* 2025;34(1):85–104.
 - [96] U.S. Food and Drug Administration. Real-world data: assessing registries to support regulatory decision-making for drug and biological products. Guidance for industry. In: Center for drug evaluation and research; center for biologics evaluation and research; oncology center of excellence; 2023.
 - [97] European Medicines Agency. Data quality framework for EU medicines regulation. In: HMA-EMA joint big data steering group; 2023. Amsterdam, The Netherlands.
 - [98] European Medicines Agency. Data quality framework for EU medicines regulation: application to real-world data. In: Committee for medicinal products for human use (CHMP) MWP; 2024. Amsterdam, The Netherlands.
 - [99] U.S. Food and Drug Administration. Framework for FDA's real-world evidence Program. 2018.
 - [100] European Medicines Agency. Guideline on registry-based studies. 2021.
 - [101] U.S. Food and Drug Administration. Considerations for the use of real-world data and real-world evidence to support regulatory decision-making for drug and biological products. Guidance for industry. In: Center for drug evaluation and research; center for biologics evaluation and research; oncology center of excellence; 2023.
 - [102] Levy C, Bowlus CL, Carey E, et al. A real-world observational cohort of patients with primary biliary cholangitis: TARGET-primary biliary cholangitis study design and rationale. *Hepatol Commun* 2018;2(5):484–491.
 - [103] Lu M, Bowlus CL, Lindor K, et al. Validity of an automated algorithm to identify cirrhosis using electronic health records in patients with primary biliary cholangitis. *Clin Epidemiol* 2020;12:1261–1267.
 - [104] Murillo Perez CF, Ioannou S, Hassanally I, et al. Optimizing therapy in primary biliary cholangitis: alkaline phosphatase at six months identifies one-year non-responders and predicts survival. *Liver Int* 2023;43(7):1497–1506.
 - [105] Levy C, Buchanan-Pearl KA, MacEwan JP, et al. A nationwide study of primary biliary cholangitis prevalence, geographic distribution, and health care providers. *Hepatol Commun* 2025;9(5).
 - [106] Tyree PT, Lind BK, Lafferty WE. Challenges of using medical insurance claims data for utilization analysis. *Am J Med Qual* 2006;21(4):269–275.
 - [107] Ludwig RJ, Anson M, Zirpel H, et al. A comprehensive review of methodologies and application to use the real-world data and analytics platform TriNetX. *Front Pharmacol* 2025;16:1516126.
 - [108] Patrick DL, Burke LB, Powers JH, et al. Patient-reported outcomes to support medical product labeling claims: FDA perspective. *Value Health* 2007;10(Suppl 2):S125–S137.
 - [109] Schneeweiss S. A basic study design for expedited safety signal evaluation based on electronic healthcare data. *Pharmacoepidemiol Drug Saf* 2010;19(8):858–868.
 - [110] Tsai CH, Eghdam A, Davoody N, et al. Effects of electronic health record implementation and barriers to adoption and use: a scoping review and qualitative analysis of the content. *Life (Basel)* 2020;10(12).
 - [111] Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj* 2009;338:b2393.
 - [112] Brookhart MA, Wang PS, Solomon DH, et al. Instrumental variable analysis of secondary pharmacoepidemiologic data. *Epidemiology* 2006;17(4):373–374.
 - [113] Suissa S. Immortal time bias in observational studies of drug effects. *Pharmacoepidemiol Drug Saf* 2007;16(3):241–249.
 - [114] Lévesque LE, Hanley JA, Kezouh A, et al. Problem of immortal time bias in cohort studies: example using statins for preventing progression of diabetes. *BMJ* 2010;340:b5087.
 - [115] Jurek AM, Greenland S, Maldonado G, et al. Proper interpretation of non-differential misclassification effects: expectations vs observations. *Int J Epidemiol* 2005;34(3):680–687.
 - [116] Brookhart MA, Mayne TJ, Coombs C, et al. Hepatic real-world outcomes with obeticholic acid in primary biliary cholangitis (HEROES): a trial emulation study design. *Hepatology* 2025;81(6):1647–1659.
 - [117] Bittermann T, Schaubel DE. Obeticholic acid as second-line therapy for primary biliary cholangitis: does target trial emulation solve the issue? *Hepatology* 2025;81(6):1629–1631.
 - [118] Fewell Z, Wolfe F, Choi H, et al. Controlling for time-dependent confounding using marginal structural models. *Stata J* 2004;4:402–420.
 - [119] Glynn RJ, Schneeweiss S, Stürmer T. Indications for propensity scores and review of their use in pharmacoepidemiology. *Basic Clin Pharmacol Toxicol* 2006;98(3):253–259.
 - [120] Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res* 2011;46(3):399–424.
 - [121] Lash T, Fox M, Fink A. Applying quantitative bias analysis to epidemiologic data. 2009.
 - [122] Sedgwick P. Selection bias versus allocation bias. *BMJ* 2013;346:f3345–f.
 - [123] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. *Nature* 2023;620(7972):172–180.
 - [124] European Medicines Agency. Reflection paper on use of real-world data in non-interventional studies to generate real-world evidence for regulatory purposes. In: Committee for medicinal products for human use (CHMP) MWP; 2025.
 - [125] Rivera DR, Henk HJ, Garrett-Mayer E, et al. The friends of cancer research real-world data collaboration pilot 2.0: methodological recommendations from oncology case studies. *Clin Pharmacol Ther* 2022;111(1):283–292.
 - [126] Verma M, Desai AP. The role of patient-reported outcomes in a patient-centered care model for managing chronic liver diseases. *Clin Liver Dis (Hoboken)* 2024;23(1):e0222.
 - [127] Kim HP, Lieber SR, Rogers ME, et al. A systematic review of patient-reported outcomes in primary biliary cholangitis and primary sclerosing cholangitis. *Hepatol Commun* 2020;4(10):1502–1515.
 - [128] Rice S, Albani V, Minos D, et al. Effects of primary biliary cholangitis on quality of life and health care costs in the United Kingdom. *Clin Gastroenterol Hepatol* 2021;19(4):768–776.e10.
 - [129] Jacoby A, Rannard A, Buck D, et al. Development, validation, and evaluation of the PBC-40, a disease specific health related quality of life measure for primary biliary cirrhosis. *Gut* 2005;54(11):1622–1629.
 - [130] EuroQol—a new facility for the measurement of health-related quality of life. *Health Policy* 1990;16(3):199–208.
 - [131] Ware Jr JE, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care* 1992;30(6):473–483.
 - [132] DeWalt DA, Rothrock N, Yount S, et al. Evaluation of item candidates: the PROMIS qualitative item review. *Med Care* 2007;45(5 Suppl 1):S12–S21.
 - [133] Bode RK, Hahn EA, DeVellis R, et al. Measuring participation: the patient-reported outcomes measurement information system experience. *Arch Phys Med Rehabil* 2010;91(9 Suppl):S60–S65.
 - [134] Crossnohere NL, Schuster ALR, Bruckel J, et al. Patient-reported outcome measures add value as clinical trial endpoints. *Nat Med* 2025;31(11):3601–3604.
 - [135] Martin ML, Stassek L, Blum SI, et al. Development and adaptation of patient-reported outcome measures for patients who experience itch associated with primary biliary cholangitis. *J Patient Rep Outcomes* 2019;3(1):2.
 - [136] European Medicines Agency. Reflection paper on patient experience data. In: Committee for medicinal products for human use (CHMP); pharmacovigilance risk assessment committee (PRAC); 2025. Amsterdam, The Netherlands.
 - [137] U.S. Food and Drug Administration. Patient-reported outcome measures: use in medical product development to support labeling claims. In: Center for drug evaluation and research; center for biologics evaluation and research; center for devices and radiological health. Guidance for industry; 2009.
 - [138] Wyrwich KW, Bullinger M, Aaronson N, et al. Estimating clinically significant differences in quality of life outcomes. *Qual Life Res* 2005;14(2):285–295.
 - [139] Lenderking WR, Anatchkova M, Pokrzywinski R, et al. Measuring health-related quality of life in patients with rare disease. *J Patient Rep Outcomes* 2021;5(1):61.

- [140] Hussain N, Barnum S, Christie M, et al. THU-341 Implementation of a clinical global impression severity scale for primary biliary cirrhosis: results of a hepatologist focussed training programme. *J Hepatol* 2025;82:S336.
- [141] Hussain N, Smith R, Barnum S, et al. THU-340 Validation of the clinical global impression severity scale for primary biliary cholangitis: a clinical trials outcome tool. *J Hepatol* 2025;82:S335–S336.
- [142] Newton JL, Jones DE, Henderson E, et al. Fatigue in non-alcoholic fatty liver disease (NAFLD) is significant and associates with inactivity and excessive daytime sleepiness but not with liver disease severity or insulin resistance. *Gut* 2008;57(6):807–813.
- [143] Okifuji A, Hare BD. Nightly analyses of subjective and objective (actigraphy) measures of sleep in fibromyalgia syndrome: what accounts for the discrepancy? *Clin J Pain* 2011;27(4):289–296.
- [144] Newton JL, Bhala N, Burt J, et al. Characterisation of the associations and impact of symptoms in primary biliary cirrhosis using a disease specific quality of life measure. *J Hepatol* 2006;44(4):776–783.
- [145] Al-Harthi N, Kumagi T, Coltescu C, et al. The specificity of fatigue in primary biliary cirrhosis: evaluation of a large clinic practice. *Hepatology* 2010;52(2):562–570.
- [146] Phaw NA, Dyson JK, Mells G, et al. Understanding fatigue in primary biliary cholangitis. *Dig Dis Sci* 2021;66(7):2380–2386.
- [147] Jones DE, Sutcliffe K, Paiman J, et al. An integrated care pathway improves quality of life in Primary Biliary Cirrhosis. *QJM* 2008;101(7):535–543.
- [148] **Mells GF, Pells G**, Newton JL, et al. Impact of primary biliary cirrhosis on perceived quality of life: the UK-PBC national study. *Hepatology* 2013;58(1):273–283.
- [149] **Carbone M, Mells GF, Pells G**, et al. Sex and age are determinants of the clinical phenotype of primary biliary cirrhosis and response to ursodeoxycholic acid. *Gastroenterology* 2013;144(3):560–569.e7. quiz e13–4.
- [150] Dyson JK, Wilkinson N, Jopson L, et al. The inter-relationship of symptom severity and quality of life in 2055 patients with primary biliary cholangitis. *Aliment Pharmacol Ther* 2016;44(10):1039–1050.
- [151] Tapper EB, Lai JC. Time to embrace PROMIS-29 as the standard health-related quality of life instrument for patients with cirrhosis. *Hepatology* 2023;78(6):1688–1691.
- [152] Smith HT, Das S, Fettiplace J, et al. Pervasive role of pruritus in impaired quality of life in patients with primary biliary cholangitis: data from the GLIMMER study. *Hepatol Commun* 2025;9(3).
- [153] Mayo MJ, Carey E, Smith HT, et al. Impact of pruritus on quality of life and current treatment patterns in patients with primary biliary cholangitis. *Dig Dis Sci* 2023;68(3):995–1005.
- [154] Yosipovitch G, Reaney M, Mastey V, et al. Peak Pruritus Numerical Rating Scale: psychometric validation and responder definition for assessing itch in moderate-to-severe atopic dermatitis. *Br J Dermatol* 2019;181(4):761–769.
- [155] Gelhorn H, Currie BM, Bushnell DM, et al. Psychometric validation of the Worst Itch Numerical Rating Scale (WI-NRS) and other patient-reported outcome measures for assessing severity and impact of pruritus in patients with primary biliary cholangitis. *Orphanet J Rare Dis* 2025;20(1):390.
- [156] Kamath BM, Abetz-Webb L, Kennedy C, et al. Development of a novel tool to assess the impact of itching in pediatric cholestasis. *Patient* 2018;11(1):69–82.
- [157] Hirschfield GM, Bowlus CL, Jones DEJ, et al. Limerixibat in patients with primary biliary cholangitis and cholestatic pruritus (GLISTEN): a randomised, multicentre, double-blind, placebo-controlled, phase 3 trial. *Lancet Gastroenterol Hepatol* 2026;11(1):22–33.
- [158] Eiman S, Hynan LS, Gabriel V, et al. The 5-D itch scale: a new measure of pruritus. *Br J Dermatol* 2010;162(3):587–593.
- [159] Yu L, Buysse DJ, Germain A, et al. Development of short forms from the PROMIS™ sleep disturbance and Sleep-Related Impairment item banks. *Behav Sleep Med* 2011;10(1):6–24.
- [160] Untas A, Goffette C, Flahault C, Vioulac C, Chazouillères O, Soret PA, et al. PBC-HOPE: a randomized controlled trial of hypnosis and psycho-education in women with primary biliary cholangitis and fatigue. *Am J Gastroenterol* 2025;1–12.
- [161] Newton JL, Hudson M, Tachtatzis P, et al. Population prevalence and symptom associations of autonomic dysfunction in primary biliary cirrhosis. *Hepatology* 2007;45(6):1496–1505.
- [162] Koc OM, Toussaint AK, Untas A, et al. Fatigue in people with primary biliary cholangitis: a position paper from the European Reference Network for Rare Liver Diseases. *Lancet Gastroenterol Hepatol* 2026;11(1):71–86.
- [163] Fisk JD, Ritvo PG, Ross L, Haase DA, Marrie TJ, Schlech WF. Measuring the functional impact of fatigue: initial validation of the fatigue impact scale. *Clin Infect Dis* 1994;18(Suppl 1):S79–S83.
- [164] Guy W, National Institute of Mental Health. Psychopharmacology Research Branch, Early Clinical Drug Evaluation Program. ECDEU assessment manual for psychopharmacology. Rev. ed. Rockville, Md: U. S. Dept. of Health, Education, and Welfare, Public Health Service, Alcohol, Drug Abuse, and Mental Health Administration, National Institute of Mental Health, Psychopharmacology Research Branch, Division of Extramural Research Programs; 1976.
- [165] Zigmond AS, Snaith RP. The hospital anxiety and depression scale. *Acta Psychiatr Scand* 1983;67(6):361–370.
- [166] Johns MW. A new method for measuring daytime sleepiness: the Epworth sleepiness scale. *Sleep* 1991;14(6):540–545.
- [167] Subramaniam D, Anderson-Smiths C, Rubinstein R, et al. Framework for the use and likelihood of regulatory acceptance of single-arm trials. *Ther Innov Regul Sci*. 2024 Nov;58(6):1214–1232.
- [168] Arondekar B, Duh MS, Bhak RH, et al. Real-world evidence in support of oncology product registration: a systematic review of new drug application and biologics license application approvals from 2015–2020. *Clin Cancer Res*. 2022 Jan 1;28(1):27–35.

Keywords: Primary Biliary Cholangitis; Endpoints; Biomarkers; Real World Evidence; Real World Data; Patient Reported Outcome Measures; Delphi.

Received 30 March 2026; accepted 30 March 2026; Available online xxx