Psychometric evaluation of the pruritus numeric rating scale in patients with primary biliary cholangitis with moderate to severe pruritus

Abstract number 843 THU-277 (Immune-mediated and cholestatic disease: Clinical aspects)

Copies of this poster obtained through QR (Quick Response) and/or text key codes are for personal use only and may not be reproduced without written permission of the authors.

Andreas E. Kremer[1], David Jones[2], Anne M. Skalicky[3], Ariane Kawata[3], Milena Anatchkova[3], Adam Smith[4], Caroline Burk[5], Marvin Rock[5], Chong Kim[5], Cynthia Levy[6]

[1]Department of Gastroenterology and Hepatology, University of Zürich, University Hospital Zürich, Zurich, Switzerland; [2]Faculty of Medical Science, Newcastle University, Newcastle, United Kingdom; [3]Patient-Centered Research, Evidera, Inc., Wilmington, NC, United States; [4]Patient-Centered Research, Evidera, Inc., Hammersmith, London, United Kingdom; [5]Global Value and Access, Global Health Economics and Outcomes Research (GHEOR), Gilead Sciences Inc., Foster City, CA, USA; [6]Schiff Center for Liver Diseases, Division of Digestive Health and Liver Diseases, University of Miami Miller School of Medicine, Miami, FL, United States

Conclusions

- The PNRS is a valid, reliable, and responsive PROM for detecting antipruritic treatment benefits in patients with PBC and moderate to severe pruritus.

Plain language summary

- We compared the Pruritus Numeric Rating Scale (PNRS) to similar questionnaires to see if it is a good measure of itch severity for patients with Primary Biliary Cholangitis (PBC).
- Patients' scores on the PNRS were consistent over time and consistent with their scores on similar questionnaires.
- These results mean that the PNRS is a good measure of itch severity in PBC patients and can be used in clinical trials.

Abbreviations: 5-D, 5-domain itch scale; ICC, intraclass correlation coefficient; LS, least squares; PBC, primary biliary cholangitis; PBC-40, primary biliary cholangitis-40; PGI-C, Patient Global Impression of Change; PGI-S, Patient Global Impression of Severity; PNRS, Pruritus Numeric Rating Scale; PROM, patient-reported outcome measure; SE, standard error.

References: 1. Tanaka A, et al. Primary biliary cholangitis. *Lancet*. 2024; 404(10457):1053-1066. **2.** Trivella J, et al. Primary biliary cholangitis: Epidemiology, prognosis, and treatment. *Hepatol Commun*. 2023; 7(6):e0179. **3.** Kremer AE, et al. Seladelpar improved measures of pruritus, sleep, and fatigue and decreased serum bile acids in patients with primary biliary cholangitis. Liver Int. 2022 Jan;42(1):112-123. **4.** Phan NQ, et al. Assessment of pruritus intensity: Prospective study on validity and reliability of the visual analogue scale, numerical rating scale and verbal rating scale in 471 patients with chronic pruritus. *Acta Derm Venereol*. 2012; 92(5):502-7. **5.** Currie, BM, et al. PCR123 Validation of the worst itch numerical rating scale (WI-NRS) and related patient-reported outcomes (PROs) to assess severity and impact of cholestatic pruritus in primary biliary cholangitis (PBC): An observational study. *Value in Health*. 2024; 27(12):S529-S530. **6.** Elman S, et al. The 5-D itch scale: a new measure of pruritus. *Br J Dermatol*. 2010 162(3):587-93. **7.** Jacoby A, 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):1622-9. **8.** National Institute of Mental Health (NIMH). Patient Global Impressions scale - Change, Improvement, Severity (PGI-C, PGI-I, PGI-S). 2025. URL: https://eprovide.mapi-trust.org/instruments/patient-global-impressions-scale-change-improvement-severity **9.** Food and Drug Administration. Guidance for industry. Patient-reported outcome measures: Use in medical product development to support labeling claims. *U.S. Department of Health and Human Services*; 2009. **10.** Food and Drug Administration. Patient-focused drug development: Incorporating clinical outcome assessments into endpoints for regulatory decision-making. Guidance for industry, Food and Drug Administration staff, and other stakeholders. *U.S. Department of Health and Human Services*; 2023. **11.** Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420-428. **12.** Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika*. 1951; 16:297–334. **13.** Revicki D, Hays RD, Cella D, et al. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol*. 2008;61(2):102-109. **14.** Fleiss JL. 1986. The design and analysis of clinical experiments. New York: Wiley. Reliability of measurement. **15.** Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*. 1994; 6(4):284–29. **16.** Mathieson GJ. We need to talk about reliability: making better use of test-retest studies for study design and interpretation. *PeerJ*. 2019; 7:e6918.

Acknowledgments: Medical writing was provided by Maddy L. Dyer, PhD, (PPD clinical research business of Thermo Fisher Scientific) in accordance with Good Publication Practice guidelines and was funded by Gilead Sciences, Inc.

Disclosures: This study was funded by Gilead Sciences, Inc. CB, MR, and CK are employed by Gilead Sciences. AEK is an advisor / consultant for Abbvie, Advanz, Alentis, AlphaSigma, Astellas, AstraZeneca, Attovia Therapeutics, Avior, Bayer, BMS, Böhringer-Ingelheim, CymaBay, Dr. Falk, Gilead, GSK, Guidepoint, Intercept, Ipsen, Mirum, MSD, Novo Nordisk, Rectify Pharma, Roche, and Takeda. AS, AMS and MA work for Evidera who received funding for conducting/presenting this data. CL holds research grants from Gilead, Intercept, and Roche. CL holds research grants from Calliditas, Cymabay, Escient, Gilead, GSK, Ipsen, Kowa, intercept, Mirum, Target RWE, and Zydus.

Correspondence: Andreas E. Kremer (andreas.kremer@usz.ch).

INTRODUCTION

- **Primary biliary cholangitis (PBC)** is a rare, chronic, autoimmune liver disease that is characterized by cholestasis and mainly affects women aged 40–70 years.[1,2]
- **Cholestatic pruritus** affects 70%–80% of patients with PBC and can be debilitating.[3]
- The **Pruritus Numeric Rating Scale (PNRS)** is a single-concept item patient-reported outcome measure (PROM) assessing the severity of worst itch over the past 24 hours. It is valid and reliable in various diseases.[4,5]
- The **psychometric properties** of the PNRS were evaluated in patients with PBC who experience cholestatic pruritus.[5] However, they have not previously been assessed within a clinical trial setting.
- This is important to support that the PNRS is **valid, reliable, and responsive** in this target population.
- Examining **clinically meaningful within-person changes** on the PNRS is also important to ensure that changes represent a meaningful improvement for patients (poster THU-277).

OBJECTIVE

- To evaluate the **psychometric properties** of the **PNRS** in patients with **PBC and moderate to severe pruritus** in the randomized, placebo-controlled, phase 3 RESPONSE trial of seladelpar (NCT04620733).

METHODS

Pruritus NRS

- Responses on the 11-point PNRS range from 0 (no itch) to 10 (worst imaginable itch). Patients completed the PNRS daily for 6 months from run-in (up to 2 weeks before day 1 [treatment initiation]) then for 7 consecutive days each month until month 12.

Patients

- The analysis included RESPONSE trial patients with moderate to severe pruritus, defined as scoring ≥4 on the PNRS at baseline.

METHODS

Psychometric evaluation

- The psychometric evaluation of the PNRS assessed **test-retest reliability, convergent validity, known-groups validity, and responsiveness** using pooled, blinded data up to month 6 from the active and placebo arms (**Table 1**).
- Other PROMs from RESPONSE were used to evaluate PNRS**:**
  - **5-Domain Itch Scale (5-D)**.[6] Patients completed the 5-D considering their experiences of pruritus during the past 2 weeks. The five domains are degree, duration, direction, disability, and distribution. The 5-D was collected bi-weekly from run-in until month 6 (week 26) then monthly until month 12 (treatment end).
  - **PBC-40**.[7] Patients completed the PBC-40 considering their experiences of pruritus during the past 4 weeks. The six domains are symptoms, itch, fatigue, cognition, social, and emotional**.** The PBC-40 was collected at run-in, day 1, month 1, and every 3 months until month 12.
  - **Patient Global Impression of Severity of pruritus (PGI-S)**.[8] Patients completed the PGI-S considering their experiences of pruritus during the past 7 days. The PGI-S was collected at run-in, day 1, month 1, and every 3 months until month 12.
  - **Patient Global Impression of Change of pruritus (PGI-C)**.[8] Patients completed the PGI-C considering the change they experienced in pruritus since the start of the study. The PGI-C was collected at month 1 and every 3 months until month 12.
- The psychometric evaluation was performed in accordance with US Food and Drug Administration guidance documents.[9,10]

RESULTS

Patients

- Of the 193 patients enrolled in the RESPONSE trial, **72** (seladelpar, n=49; placebo, n=23) were included in the psychometric evaluation who had moderate to severe pruritus at baseline (mean [range], 53.8 [32–75] years of age; 97.2% female).

Psychometric evaluation

- The PNRS had acceptable:
  - **Test-retest reliability:** mean difference of 1.1 between weekly mean PNRS scores at day 1 and week 4 (intraclass correlation coefficient [ICC]=0.48; r=0.63; p<0.001; n=34) demonstrates fair-to-moderate reliability.
  - **Convergent validity:** the strongest correlations were found between PNRS scores and 5-D Degree scores at day 1 (r=0.60), week 4 (r=0.50), week 12 (r=0.64), and week 26 (r=0.68) as predicted (**Table 2**). Convergent validity of the PNRS was also supported by moderate to strong correlations with 5-D itch total, 5-D duration, and PBC-40 itch scores. Correlations between PNRS scores and PBC-40 symptoms, fatigue, cognition, social, and emotional domain scores were weak to moderate.
  - **Known-groups validity:** weekly mean PNRS scores differentiated by 5-D degree (week 12: F=15.24, p<0.0001; week 26: F=15.28, p<0.0001; **Table 3**) and by PGI-S (week 12: F=30.75, p<0.0001; week 26: F=16.32, p<0.0001 **Table 4**). PNRS scores were higher in patients with higher disease severity (per 5-D degree and PGI-S).
  - **Responsiveness:** incremental improvements in weekly mean PNRS scores were found, with greater improvement in PGI-S scores (F=7.28, p<0.0001) and better PGI-C scores (6.15, p<0.0001) over time.

Table 2. Convergent validity results: Key 5-D and PBC-40 itch domains.

Assessment	N	Correlation with PNRS, p-value
5-D Itch Total		
Day 1	72	0.58, <0.0001
Week 4	70	0.54, <0.0001
Week 12	68	0.66, <0.0001
Week 26	65	0.60, <0.0001
5-D Degree		
Day 1	72	0.60, <0.0001
Week 4	70	0.50, <0.0001
Week 12	68	0.64, <0.0001
Week 26	65	0.68, <0.0001
5-D Duration		
Day 1	72	0.53, <0.0001
Week 4	70	0.54, <0.0001
Week 12	68	0.64, <0.0001
Week 26	65	0.55, <0.0001
PBC-40 Itch		
Day 1	72	0.59, <0.0001
Week 4	63	0.40, 0.0012
Week 12	65	0.49, <0.0001
Week 26	59	0.62, <0.0001

Table 3. Known-groups validity results: Pruritus NRS by 5-D degree.

	5-D Degree										
	Not present		Mild		Moderate		Severe		Unbearable		
PNRS	N	LS mean (SE)	N	LS mean (SE)	N	LS mean (SE)	N	LS mean (SE)	N	LS mean (SE)	Overall F* value	p-value
Week 12	2	0.75 (1.25)	25	2.64 (0.35)	29	4.58 (0.33)	11	6.47 (0.54)	0	-	15.24	<0.0001
Week 26	3	0.67 (0.93)	30	2.57 (0.29)	25	4.66 (0.32)	5	6.22 (0.72)	0	-	15.28	<0.0001

*the ANOVA F statistic compares the variances between two groups of variables is significant

Table 4. Known-groups validity results: Pruritus NRS by PGI-S.

	PGI-S									
	None		Mild		Moderate		Severe			
PNRS	N	LS mean (SE)	N	LS mean (SE)	N	LS mean (SE)	N	LS mean (SE)	Overall F* value	p-value
Week 12	5	0.92 (0.66)	31	2.88 (0.27)	22	5.92 (0.32)	5	6.63 (0.66)	30.75	<0.0001
Week 26	4	0.68 (0.80)	32	2.67 (0.28)	20	5.26 (0.36)	2	5.67 (1.13)	16.32	<0.0001

*the ANOVA F statistic compares the variances between two groups of variables is significant

Table 1. Psychometric assessments.

Assessment	Description	Analysis	Statistical test
Test-retest reliability	Scores on the PNRS are consistent over time.	Association between weekly average PNRS scores at day 1 and week 4 in "stable" patients (i.e., those with the same 5-D Degree score on day 1 and week 4). Values: 0.4–0.6 (fair), 0.6–0.8 (moderate), and >0.8 (substantial).[11]	Intraclass correlation coefficient, Pearson correlation, and paired t-test
Convergent (construct) validity	The PNRS correlates with other PROMs that measure related concepts.	Associations of weekly average PNRS scores with 5-D scores and PBC-40 scores at day 1, week 4, week 12, and week 26. Values: 0.1–0.3 (weak), 0.3–0.5 (moderate), and >0.5 (strong).[12]	Spearman correlation coefficient
Known-groups (construct) validity	The PNRS discriminates between groups known to differ.	Differences in weekly average PNRS scores between subgroups of 5-D Degree severity (levels of itch: not present=0, mild=1, moderate=2, severe=3, and unbearable=4) and PGI-S severity (levels of itch: none=0, mild=1, moderate=2, and severe=3) at day 1, week 4, week 12, and week 26.[13]	Analysis of variance, Least squares (LS) method and SE estimates difference between observed mean and the fitted mean value from the model
Responsiveness	The PNRS can detect change over time.	Differences in weekly average PNRS score changes from baseline to month 6 between subgroups of PGI-S change (from -3-point improvement to +3-point improvement) and PGI-C (from -3 [very much worse] to +3 [very much better]). Moderate or strong correlations (|r| ≥ 0.30) were predicted.[13]	Analysis of covariance

LIMITATIONS

- There were some small sample sizes in the subgroups of itch severity.
- Thresholds for test-reliability sources can vary and be interpreted differently. ICC values between 0.40 and 0.75 have been described as "good", and ICC values between 0.40 and 0.59 have been described as "fair".[14,15,16]
- The moderate test-retest reliability results could be due to measurement error, actual changes in the underlying "true" change value, the small sample size of "stable" patients, or the long duration between timepoints.