REVIEW ARTICLE OPEN ACCESS

Retrospective Assessment of Endometriosis Pain Over the Life Course: A Reliability Study Within the ComPaRe-Endometriosis Cohort

Solène Gouesbet[1] | Sarah Lambert[1] | Hélène Amazouz[1] | Zélia Breton[1,2] | Viet-Thi Tran[3,4] | Stacey Missmer[5,6] | Marina Kvaskoff[1]

[1]Université Paris-Saclay, UVSQ, Inserm, Gustave Roussy, CESP, Villejuif, France | [2]Lyv Healthcare, Nantes, France | [3]Center for Clinical Epidemiology, Hôtel-Dieu Hospital, Assistance Publique-Hôpitaux de Paris (AP-HP), Paris, France | [4]Université de Paris, CRESS, INSERM, INRA, Paris, France | [5]Michigan State University, Grand Rapids, Michigan, USA | [6]Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA

Correspondence: Marina Kvaskoff (marina.kvaskoff@inserm.fr)

Received: 29 November 2023 | **Revised:** 10 March 2025 | **Accepted:** 1 May 2025

Funding: ComPaRe was funded by the Assistance Publique-Hôpitaux de Paris (APHP), Université Paris-Cité. The ComPaRe-Endometriosis cohort received funding from patient societies and initiatives (ENDOmind, EndoFrance, Les joyeux boulomanes Farledois, EndoAction, *La Belle et L'Endo, Les Chroniques Endométriques*) and was funded by grants from the *Fondation Apicil* and the *Fondation de France* (*00132975/WB-2022-44780*). S.G. has received a PhD scholarship from the French Ministry of Research and a grant from the Crédit Agricole Ile-de-France Mécénat. The funders had no role in the design, interpretation, or decision to publish the study.

Keywords: endometriosis | numeric rating scale | pain | reliability study | retrospective data | WERF-EPHect

ABSTRACT

Background: Endometriosis may manifest through various pain symptoms, such as dysmenorrhea, dyspareunia, dyschezia, dysuria and abdominal pain. While retrospective evaluation of these pain symptoms is less expensive and time-consuming compared to a prospective evaluation, there is potential for recall bias, and the reliability of such data needs to be assessed. We aimed to evaluate the reliability of questions on past endometriosis-related pain.

Methods: We conducted a reliability study within ComPaRe-Endometriosis, an ongoing prospective e-cohort including patients with endometriosis. We assessed past endometriosis-related pain over a lifetime using the WERF-EPHect Patient Questionnaire—Standard (EPQ-S). Participants rated the worst intensity of dysmenorrhea, dyspareunia, dyschezia, dysuria and abdominal pain that they experienced at ≤ 15, 16–20, 21–30, 31–40 and > 40 years using a numeric-rating scale (NRS). We asked the same questions about 1 year later and measured the agreement between participant responses by calculating intraclass correlation coefficients (ICC) (continuous NRS level) and weighted kappa coefficients (κ_w) (pain intensity categories).

Results: A total of 1752 participants completed both surveys. The global reliability was close to the 'good' and 'substantial' thresholds for dysmenorrhea ($ICC = 0.74$; $\kappa_w = 0.57$) and dyspareunia ($ICC = 0.72$; $\kappa_w = 0.57$), 'moderate' and close to the 'substantial' threshold for dysuria ($ICC = 0.68$; $\kappa_w = 0.59$), and 'moderate' for dyschezia ($ICC = 0.62$; $\kappa_w = 0.54$) and abdominal pain ($ICC = 0.58$; $\kappa_w = 0.49$).

Conclusions: In this population, questions on worst pain intensity over the life course showed moderate-to-good reliability depending on the type of pain, with higher reliability when pain level was considered as a continuous variable.

Significance Statement: While prospective measures are the most robust approach in epidemiological research, longitudinal data with pain recorded since childhood or adolescence are scarce. This study shows that the worst level of pelvic and abdominal pain over the life course are reliably reported by endometriosis patients after a 1-year interval. These findings suggest that retrospective pain assessment may reliably be used to assess trajectories of pain over the life course in order to gain insights into the progression of pain-related conditions such as endometriosis.

1 | Introduction

Endometriosis is a chronic inflammatory disease in which endometrium-like tissue develops outside of the uterus (Zondervan et al. 2018). It may manifest through several pain symptoms, including dysmenorrhea, dyspareunia, dyschezia, dysuria and abdominal pain. Patients may experience several of these symptoms in various patterns and at different intensity levels, which may evolve over time (Becker et al. 2021; Comptour et al. 2020). Measuring pain level is thus of particular importance to study endometriosis, particularly to assess treatment response or disease progression. While prospective measurements represent the highest methodological standard, very few data are available to observe endometriosis pain intensity over the life course starting from adolescence. In this context, using retrospective assessment of pain, although carrying the potential for recall bias, may constitute a useful tentative approach.

To facilitate large-scale collaborations on endometriosis research worldwide, the World Endometriosis Research Foundation (WERF) developed guidelines and tools to standardise endometriosis data collection based on an international consensus, the WERF Endometriosis Phenome and Biobank Harmonisation Project (WERF-EPHect) available on the website https://www.ephect.org/. One of these instruments, the WERF-EPHect Patient Questionnaire—Standard (EPQ-S), includes questions that retrospectively assess endometriosis-related pain over five age ranges across the life course using an 11-point numeric rating scale (NRS), from 0 (no pain) to 10 (worst pain imaginable) (Vitonis et al. 2014). Previous studies have shown that the NRS is a recommended tool for assessing the intensity of menstrual pain in the research context (Larroy 2002; Williamson and Hoggart 2005). However, the reliability of this tool is uncertain for retrospective assessment (Coughlin 1990; Talari and Goyal 2020).

While retrospective measures are less expensive and time-consuming than those collected prospectively, their reliability needs to be assessed given the potential for recall bias (Munnangi and Boktor 2022). Previous studies evaluating the reliability of retrospective pain assessments suggested that people are able to remember and reliably assess the severity of their pain in a general context (Brauer et al. 2003; Jamison et al. 2006). The only study that assessed the reliability of retrospective evaluation of endometriosis-related pain concluded that women with endometriosis generally remembered their past pelvic pain accurately (Nunnink and Meana 2007). However, these studies have focused on pain that occurred from a few days to a few months prior (30 days for the study on endometriosis; Nunnink and Meana 2007), and the reliability of retrospective measures may vary considerably depending on the duration of the evaluated period. In comparison, the WERF-EPHect questionnaire assesses past pain over long periods experienced throughout life in five different age groups, from under 15 to above 40 years.

Our aim was to determine the reliability of retrospective assessment of past endometriosis-related pain across the life course using the WERF-EPHect EPQ-S questionnaire.

2 | Materials and Methods

2.1 | Study Sample

ComPaRe-Endometriosis is a sub-cohort of ComPaRe (*Community of Patients for Research*) a participatory research platform initiated in 2017 (Gouesbet et al. 2023). Briefly, ComPaRe is an ongoing prospective e-cohort of over 55,000 chronic disease patients volunteering to help advance research on chronic diseases (Tran and Ravaud 2020). Participants are French-speaking adults aged 18 or older reporting at least one chronic illness. Patients register on the online platform http://compare.aphp.fr and regularly reply to self-administered questionnaires after providing electronic consent. The cohort was approved by the Institutional Review Board of the Hôtel-Dieu Hospital in Paris (IRB: 0008367) and the French National Commission for Data Protection and Privacy (CNIL: 916397). ComPaRe-Endometriosis consists of patients who reported endometriosis and/or adenomyosis. After the initial launch of ComPaRe-Endometriosis in 2018, about 6000 women were active participants of the cohort in July 2022 (date of data extraction for analysis).

2.2 | Data Collection

All ComPaRe participants reply to a baseline questionnaire on their health and socio-demographic factors. Subsequent monthly questionnaires collect information on various factors including employment status, lifestyle and several patient-reported outcome (PROMs) and patient-reported experience measures (PREMs). ComPaRe-Endometriosis participants additionally reply to endometriosis-related questionnaires collecting information on factors such as medical path to diagnosis, self-reported stage and type of disease, surgical history and pain. We also used part of the WERF-EPHect EPQ-5 questionnaire to collect past data on five types of pain: dysmenorrhea, abdominal pain, dyspareunia, dyschezia and dysuria (Vitonis et al. 2014). More specifically, this part of the questionnaire collects data on the intensity of symptoms for these types of pain across the life course, assessing the worst pain levels on a NRS from 0 (no pain) to 10 (worst pain imaginable) for the 5 pain types over 5 age ranges (≤ 15, 16–20, 21–30, 31–40 and > 40 years of age). Each participant responded for the age ranges that applied to them. For instance, a 35-year-old participant responded to the first 4 age ranges (≤ 15, 16–20, 21–30 and 31–40). All individuals could assess their pain for at least the first 2 age ranges (i.e., ≤ 15, 16–20) because the cohort only included adults over 18 years. At each period, participants could report that the question did not apply to them (e.g., no sexual intercourse over the period so no possibility to assess dyspareunia, no menstruation so no possibility to assess dysmenorrhea) or indicate that they did not remember their pain level. These questionnaires were first sent to participants in March 2021 and then again in March 2022 to the participants who had answered the first survey in order to measure the reliability of data between the two surveys. However, participants could answer each survey until the date of data extraction (July 2022). To ensure a 1-year gap, we calculated the median and 25th/75th percentiles of the intervals between questionnaire responses: median 12 months (P25: 10 months/P75: 13 months). In addition, we used sociodemographic data (i.e., age, education level) and self-reported data on disease characteristics (i.e., stage, type) to describe our study population and to investigate reliability by subgroup (i.e., financial situation, education level, anxiety, depression, current level of pain and age). Financial situation was assessed every 2 years using a 6-level scale (from 'in debt' to 'very comfortable') evaluating participants' perceived financial situation. This instrument was proposed by the French national institute for statistical and economic studies and is largely used in epidemiological studies in France (Kranklader and Schreiber 2015). Anxiety and depression were assessed annually using the GAD-7 and PHQ-9 questionnaires, respectively (Kroenke et al. 2001; Spitzer et al. 2006). Current level of the 5 pain types was assessed annually using an 11-point (NRS) scale, indicating the worst level of pain experienced in the past 3 months (12 months for dyspareunia). We used the most recent data for financial situation, anxiety and depression and the data closest to March 2022 (i.e., date of the second survey) for current level of pain.

2.3 | Statistical Analysis

We measured reliability in two ways. First, we assessed the agreement of an 11-point (NRS) scale values between the 2021 and 2022 surveys by calculating intraclass correlation coefficients (ICC) using a 2-way random effects model with an agreement coefficient (McGraw and Wong 1996; Shrout and Fleiss 1979). ICC values were considered to be poor (<0.50), moderate (0.50–0.75), good (0.75–0.90), or excellent reliability (>0.90) (Koo and Li 2016). Second, to determine whether grouping NRS values into broader categories would lead to better reliability, we assessed the agreement of NRS values between surveys by calculating weighted kappa coefficients (κ_w). For this, we repeated the analyses by grouping values into 5 categories of pain intensity using a 5-point scale: no pain (0), mild (1–3), moderate (4–5), severe (6–7), or very severe (8–10). Kappa values were considered to be poor (<0.20), slight (0.21–0.40), moderate (0.41–0.60), substantial (0.61–0.80), or almost perfect agreement (>0.80) (Landis and Koch 1977).

We first calculated a unique ICC and κ_w for each type of pain regardless of age ranges (i.e., including all responses obtained for each type of pain). Next, we considered age ranges to assess the consistency of responses. We examined the ICC and κ_w for each pain type in the age group corresponding to the participant's current age, as well as in the previous age group relative to the woman's current age (age group—1), and so on for earlier age groups. This analysis aimed to examine whether the consistency of responses varied with the time elapsed since the onset of pain. To complement this, we performed descriptive analyses to explore alternative ways to observe the reliability of the data (e.g., mean levels of pain, standard deviations and point differences between the means). Also, we performed subgroup analyses by calculating ICCs and κ_w values to evaluate a potential influence of the following factors on data reliability: financial situation, education level, anxiety, depression and current level of pain (Coughlin 1990; Dillon and Pizzagalli 2018; Glazier and Alden 2017; Linden et al. 1993; Previtali et al. 2022). For each factor, scores were collapsed into two categories (e.g., 'none/moderate' and 'severe/very-severe' for the level of pain corresponding to scores 0–5 and 6–10, respectively). The Fisher-Z test was applied to compare ICC and κ_w values between these two categories. This test was also used in a sensitivity analysis to determine whether the reliability between the two timepoints in the age range 31–40 (with the highest number of participants) was greater than in other age ranges with fewer participants. Through this analysis, we also aimed to observe the reliability of responses in younger age groups, as it is possible that younger women are more likely to recall more accurately their pain intensity, even over a period of 1 year, and we wanted to test this potential bias. All analyses were performed using SAS 9.4 and R version 4.1.1, and the irr package (Gamer et al. 2019).

3 | Results

3.1 | Participant Characteristics

A total of 1752 participants replied to both the March 2021 and the March 2022 surveys on past pain and were included in this study. The mean age of participants was 35.7 years (SD = 8.0) (Table 1). Participants generally had a high education level (58.5% had a Bachelor's degree or more) but a financial situation perceived as bad (71.0% reported being in debt, in financial difficulty, a bit tight financially, or just at financial balance), although most participants were employed (72.0%). Most reported a diagnosis of endometriosis alone (63.6%), while only 5.5% reported a diagnosis of adenomyosis alone, and 30.8% reported a diagnosis of both diseases. Among participants who reported their endometriosis type (76.3%), 46.8% reported deep endometriosis, 34.1% reported ovarian endometrioma, and 19.2% reported superficial peritoneal endometriosis, including both exclusive and non-exclusive types. About one-third of participants (30.8%) reported their endometriosis stage, with fairly equal frequencies for stages I, II and III (approximately 14% each), and 58.3% for stage IV. Finally, approximately half of participants reported a history of endometriosis surgery (54.0%).

3.2 | Global Reliability

The global reliability (i.e., including all answers obtained for each type of pain) was very close to the 'good' and 'substantial' thresholds for dysmenorrhea (ICC = 0.74; $\kappa_w = 0.57$) and dyspareunia (ICC = 0.72; $\kappa_w = 0.57$) (Table 2). For dysuria, the κ_w value approached the 'substantial' threshold, while the ICC value was 'moderate' (ICC = 0.68; $\kappa_w = 0.59$). Agreement values were 'moderate' for dyschezia (ICC = 0.62; $\kappa_w = 0.54$) and abdominal pain (ICC = 0.58; $\kappa_w = 0.49$). Agreement levels obtained with the 5-point scale of pain intensity were 'slight' to 'moderate',

TABLE 1 | Characteristics of study participants, ComPaRe-Endometriosis cohort ($N=1752$).

Characteristics			Value
Age (years)			
Mean (SD)			35.7 (8.0)
Range			18–72
			n (%)
< 21			12 (0.7)
21–30			44 (2.5)
31–40			1245 (71.1)
> 40			451 (25.7)
Education level			*n* (%)
≤ High school graduate			391 (22.3)
Associate degree			322 (18.4)
≥ Bachelor's degree			1025 (58.5)
Unknown			14 (0.8)
Perception of financial situation			*n* (%)
In debt			40 (2.4)
It's difficult			175 (10.3)
A bit tight, need to be careful with finances			455 (26.9)
In financial balance			531 (31.4)
Fairly comfortable financially			404 (23.9)
Very comfortable financially			86 (5.1)
Missing			61
Professional status			*n* (%)
Employed			1261 (72.0)
Unemployed			172 (9.8)
Student			199 (11.4)
Long-term disability			54 (3.1)
Houseperson			27 (1.5)
Retired			8 (0.5)
Other			31 (1.8)
Self-reported diagnosis of endometriosis/adenomyosis			*n* (%)
Only endometriosis			1115 (63.6)
Only adenomyosis			97 (5.5)
Both			540 (30.8)
Self-reported type of endometriosis[a]			*n* (%)
SPE	OMA	DE	
X			172 (13.6)
	X		262 (20.7)
		X	492 (39.0)
X	X	X	32 (2.5)
X	X		66 (5.2)
X		X	43 (3.4)
	X	X	196 (15.5)
313	556	763	
(19.2)	(34.1)	(46.8)	
Don't know			392
Self-reported stage of endometriosis			*n* (%)
I			78 (14.5)
II			72 (13.4)
III			75 (13.9)
IV			314 (58.3)
Don't know			1212
Missing			4
History of surgery for endometriosis			
No surgery			797 (46.1)
At least one surgery			933 (53.9)
Missing			22
Number of comorbidities[b]			*n* (%)
0			765 (43.7)
1			433 (24.7)
2			231 (13.2)
≥ 3			323 (18.4)

(Continues)

TABLE 1 | (Continued)

Abbreviations: DE, deep endometriosis; OMA, ovarian endometrioma; SD, standard deviation; SPE, superficial peritoneal endometriosis.

[a] Excluding 97 participants with adenomyosis only ($N=1655$ patients).

[b] Comorbidities include all chronic diseases self-reported by the participant (defined as illness requiring care for at least six months) other than endometriosis and/or adenomyosis.

and all lower than those obtained with the 11-point (NRS) scale ($\kappa_w = 0.38$ vs. 0.57 for dysmenorrhea and dyspareunia, $\kappa_w = 0.49$ vs. 0.59 for dysuria, $\kappa_w = 0.45$ vs. 0.54 for dyschezia and $\kappa_w = 0.39$ vs. 0.49 for abdominal pain).

3.3 | Reliability According to the Time Period Since Pain Occurred

This section presents the results of the reliability levels obtained based on the time interval between the current age group and the age group for which reliability is assessed (current age group minus 1, 2, 3, or 4 age groups). For each type of pain, these results are presented in Table 3.

TABLE 2 | Global agreement in numeric rating scale values between the March 2021 and March 2022 surveys for five types of pain, ComPaRe-Endometriosis cohort ($N = 1752$).

Type of pain	Dysmenorrhea	Dyspareunia	Dysuria	Dyschezia	Abdominal pain
Answers (n)[a]	5388[a]	3250[a]	4939[a]	5545[a]	5358[a]
ICC [95% CI]	0.74 [0.73; 0.75]	0.72 [0.70; 0.74]	0.68 [0.66; 0.69]	0.62 [0.60; 0.64]	0.58 [0.56; 0.60]
Weighted kappa (11-point scale)	0.57 [0.56; 0.59]	0.57 [0.55; 0.59]	0.59 [0.56; 0.62]	0.54 [0.52; 0.56]	0.49 [0.48; 0.51]
Weighted kappa (5-point scale)	0.38 [0.37; 0.41]	0.38 [0.36; 0.41]	0.49 [0.46; 0.53]	0.45 [0.43; 0.47]	0.39 [0.37; 0.41]

Abbreviations: CI, confidence interval; ICC, intraclass correlation coefficient.
[a] Global agreement was calculated by pooling all the data obtained from all age ranges.

For dysmenorrhea, ICC values indicated ‘moderate’ to ‘good’ agreement levels across time periods, while κ_w values indicated ‘moderate’ to ‘substantial’ agreement (ICCs = 0.66 to 0.77; $\kappa_w = 0.50$ to 0.61). A ‘moderate’ agreement was found for ‘participant's current age range’, ‘age range minus 1’ and ‘age range minus 4’. A ‘moderate’ to ‘good’ agreement was observed for ‘age range minus 2’. A ‘good’/‘substantial’ agreement was found for ‘age range minus 3’. Some participants did not assess dysmenorrhea for one or more age range(s) due to the absence of menstruation (menopause, period-suppressive treatment…). Our results showed that 0.4% to 89.0% of the study population did not assess dysmenorrhea in the second survey for this reason (Table S1).

For dyspareunia, ICC values ranged from ‘moderate’ to ‘excellent’ agreement levels across time periods, while κ_w values indicated ‘moderate’ to ‘almost-perfect’ agreement (ICCs = 0.63 to 0.99; $\kappa_w = 0.46$ to 0.93). However, the ‘excellent’/‘almost-perfect’ agreement was based on 5 answers only for the ‘age range minus 4’ (very few participants had sexual intercourse before age 16 [Table S1]), which may not be sufficient for kappa and ICC analyses (Bujang and Baharum 2017; Zou 2012). A ‘moderate’ agreement was found for ‘participant's current age range’ and ‘age range minus 1’. Agreement was very close to the ‘good’ and ‘substantial’ thresholds for ‘age range minus 2’ and ‘age range minus 3’. Some participants did not assess dyspareunia for one or more age range(s) due to the absence of intercourse. Our results showed that 1.1% to 86.9% of the study population did not assess dyspareunia in the second survey for this reason (Table S1).

For dysuria, ICC values ranged from ‘moderate’ to ‘good’ agreement levels across time periods, while κ_w values indicated ‘moderate’ to ‘substantial’ agreement (ICCs = 0.57 to 0.75; $\kappa_w = 0.46$ to 0.63). A ‘good’/‘substantial’ agreement was found for ‘age range minus 4’, while all other categories showed ‘moderate’ agreement.

For dyschezia, ICC values indicated ‘poor’ to ‘moderate’ agreement levels across time periods, while κ_w values indicated ‘slight’ to ‘moderate’ agreement (ICCs = 0.44 to 0.60; $\kappa_w = 0.38$ to 0.51). A ‘poor’/‘slight’ agreement was found for the ‘age range minus 4’, whereas ‘moderate’ agreement was observed for all other categories.

For abdominal pain, ICC and κ_w values showed ‘moderate’ agreement levels across time periods (ICCs = 0.52 to 0.57; κ_w = from 0.45 to 0.49).

3.4 | Supplementary Analyses

In supplementary analyses, we calculated the mean values of the 11-point (NRS) scale for each type of pain and age group. They appeared very similar between the two surveys, with a point difference of only 0 to 0.3 on the NRS scale (Table S2). The point difference distribution showed that 54% to 100% of the study population fell within a range of plus or minus 1 NRS point, and at least 68% of the participants fell within a range of plus or minus 2 NRS points (Table S3).

In stratified analyses, we found several differences in reliability according to various factors (Table S4). *Current level of pain*—Participants experiencing severe or very severe dysuria across all age ranges combined showed more reliable responses than those with absent or moderate dysuria (ICC of 0.75 vs. 0.59, $p < 0.001$; κ_w of 0.64 vs. 0.50, $p = 0.004$). Similar results in ICC were also noted for dysmenorrhea and abdominal pain, although with smaller differences. *Depression*—Participants with moderate or severe depression provided more reliable responses for their abdominal pain than those with no or mild depression (ICC of 0.56 vs. 0.48, $p = 0.006$; κ_w of 0.48 vs. 0.40, $p = 0.01$). *Anxiety*—When grouping all types of pain, participants with no or mild anxiety were slightly more likely to report the same category of pain level compared with those with moderate-to-severe anxiety (κ_w of 0.65 vs. 0.63, $p = 0.04$). However, no significant differences were identified for any specific type of pain, nor for continuous pain levels. *Financial situation*—Considering all types of pain combined, participants perceiving their financial situation as favourable tended to report their pain level slightly more reliably between surveys (ICC of 0.76 vs. 0.74, $p = 0.001$; κ_w of 0.66 vs. 0.63, $p < 0.001$). A similar difference was found for dyschezia, and for dysmenorrhea when pain level was considered as a continuous variable. *Education level*—Participants with an education level above a high school diploma tended to report their pain levels slightly more reliably than those with a lower education level, considering all types of pain combined (ICC of 0.75 vs. 0.73, $p = 0.005$; κ_w of 0.65 vs. 0.63, $p = 0.03$).

TABLE 3 | Agreement in numeric rating scale values between the March 2021 and March 2022 surveys according to time period since pain occurred, ComPaRe-Endometriosis cohort ($N = 1752$).

	Participant's current age range	Participant's previous age range (age range −1)	Age range −2	Age range −3	Age range −4
Dysmenorrhea					
Answers (*n*)	1215	1554	1436	900	283
ICC [95% CI]	0.66 [0.63; 0.70]	0.67 [0.64; 0.69]	0.76 [0.73; 0.78]	0.77 [0.74; 0.80]	0.69 [0.62; 0.75]
Weighted kappa (11-point scale)	0.50 [0.47; 0.54]	0.50 [0.47; 0.53]	0.59 [0.56; 0.62]	0.61 [0.58; 0.64]	0.55 [0.48; 0.62]
Weighted kappa (5-point scale)	0.43 [0.39; 0.48]	0.36 [0.31; 0.40]	0.34 [0.30; 0.39]	0.29 [0.23; 0.35]	0.28 [0.17; 0.39]
Dyspareunia					
Answers (*n*)	1143	1196	712	188	5
ICC [95% CI]	0.63 [0.59; 0.66]	0.67 [0.64; 0.70]	0.72 [0.68; 0.75]	0.70 [0.61; 0.76]	0.99 [0.94; 1.00]
Weighted kappa (11-point scale)	0.46 [0.43; 0.49]	0.51 [0.48; 0.54]	0.60 [0.56; 0.64]	0.60 [0.51; 0.69]	0.93 [0.75; 1.00]
Weighted kappa (5-point scale)	0.58 [0.39; 0.77]	0.44 [0.39; 0.49]	0.48 [0.43; 0.54]	0.45 [0.39; 0.52]	0.23 [0.12; 0.34]
Dysuria					
Answers (*n*)	1235	1319	1230	899	306
ICC [95% CI]	0.68 [0.65; 0.71]	0.65 [0.61; 0.68]	0.57 [0.53; 0.61]	0.58 [0.54; 0.63]	0.75 [0.68; 0.78]
Weighted kappa (11-point scale)	0.58 [0.54; 0.63]	0.56 [0.51; 0.62]	0.48 [0.38; 0.58]	0.46 [0.26; 0.65]	0.63 [0.30; 0.95]
Weighted kappa (5-point scale)	0.47 [0.31; 0.63]	0.47 [0.34; 0.53]	0.48 [0.43; 0.54]	0.45 [0.39; 0.52]	0.58 [0.48; 0.69]
Dyschezia					
Answers (*n*)	1467	1586	1324	869	298
ICC [95% CI]	0.56 [0.53; 0.60]	0.60 [0.57; 0.63]	0.57 [0.54; 0.61]	0.50 [0.45; 0.55]	0.44 [0.35; 0.53]
Weighted kappa (11-point scale)	0.48 [0.44; 0.51]	0.51 [0.48; 0.54]	0.50 [0.45; 0.54]	0.44 [0.36; 0.52]	0.38 [0.23; 0.52]
Weighted kappa (5-point scale)	0.41 [0.35; 0.48]	0.43 [0.39; 0.48]	0.43 [0.39; 0.46]	0.38 [0.33; 0.42]	0.38 [0.30; 0.46]
Abdominal pain					
Answers (*n*)	1476	1524	1242	816	298
ICC [95% CI]	0.54 [0.51; 0.58]	0.53 [0.49; 0.57]	0.57 [0.53; 0.61]	0.55 [0.51; 0.60]	0.52 [0.44; 0.60]
Weighted kappa (11-point scale)	0.45 [0.41; 0.48]	0.45 [0.41; 0.48]	0.48 [0.45; 0.53]	0.49 [0.44; 0.55]	0.47 [0.37; 0.56]
Weighted kappa (5-point scale)	0.47 [0.42; 0.51]	0.40 [0.36; 0.44]	0.33 [0.30; 0.37]	0.34 [0.30; 0.38]	0.35 [0.27; 0.43]

Abbreviations: CI, confidence interval; ICC, intraclass correlation coefficient.

Higher ICC values were also found for dyschezia among participants with a higher education level (κ_w of 0.63 vs. 0.58, $p=0.02$).

The reliability of responses was assessed across different age groups. No significant differences in reliability were observed for dyspareunia and dysuria according to the Fisher-Z test (Table S5). However, for dysmenorrhea, reliability was statistically lower in participants aged >40 (ICC $=0.68$, $\kappa_w=0.53$) compared to the ≤ 30 and 31–40 age groups (ICC $=0.77$, $\kappa_w=0.59$ for both). A similar trend was noted for dyschezia, with the >40 age group showing lower reliability (ICC $=0.59$, $\kappa_w=0.51$) compared to the 31–40 group (ICC $=0.63$, $\kappa_w=0.54$). The ≤ 30 age group did not have a large enough sample to reach statistical significance, but the ICC and κ_w values were the same as those of the 31–40 group. For abdominal pain, the ≤ 30 age group demonstrated higher reliability (ICC $=0.69$, $\kappa_w=0.58$) compared to the 31–40 (ICC $=0.56$, $\kappa_w=0.48$) and >40 age groups (ICC $=0.53$, $\kappa_w=0.45$). These findings highlight that age may influence the reliability of pain reporting.

4 | Discussion

4.1 | Summary of Findings

In this study, we assessed the reliability of questions evaluating the worst intensity of endometriosis-related pain over the life course from the WERF-EPHect EPQ-S. The global reliability (i.e., including all answers obtained for each type of pain) was close to the 'good' and 'substantial' thresholds for dysmenorrhea (ICC $=0.74$; $\kappa_w=0.57$) and dyspareunia (ICC $=0.72$; $\kappa_w=0.57$), 'moderate' and close to the 'substantial' thresholds for dysuria (ICC $=0.68$; $\kappa_w=0.59$), and 'moderate' for dyschezia (ICC $=0.62$; $\kappa_w=0.54$) and abdominal pain (ICC $=0.58$; $\kappa_w=0.49$). When taking into account time between the considered age range for pain assessment and participants' age at recall, results were heterogeneous and did not follow a clear pattern based on the time elapsed. The most extreme levels of agreement were found in the 'age range minus 4' category, with the lowest level for dyschezia and the highest for dyspareunia, while dysuria also showed a high level of agreement. Interpreting these results requires caution, as the dyspareunia 'age range minus 4' group included small numbers, and the analysis included the score 0 (no pain), reported by a high proportion of women for dysuria (72.7% to 98.7%) and dyschezia (42.9% to 87.9%) (Table S2). High levels of agreement were also observed for dysmenorrhea and dyspareunia in the 'age range minus 2' and 'minus 3', without these limitations.

4.2 | Comparison With Previous Work and Discussion of Findings

Only one study to date has investigated pain recall in endometriosis, finding that women generally recalled pain experienced in the previous 30 days accurately, although a different method was used to calculate agreement based on a ratio between recalled pain score and current pain score (Nunnink and Meana 2007). In that study ($n=100$ women), inaccuracies were skewed towards overestimation. Our data did not suggest over- or underestimation, except for abdominal pain, for which average scores were slightly higher in the second survey (0.1 to 0.3 points on the average pain scale). Nunnink and Meana's study also explored recall bias in relation to psychological well-being and current pain at the time of recall, finding no significant bias associated with psychological well-being. Conversely, our findings indicate that participants with lower anxiety levels provide more reliable responses. Additionally, we observed that participants experiencing moderate or severe depression reported abdominal pain with greater reliability. However, 3018 responses came from women with 'moderate to severe' depression versus 787 in the 'no or mild depression' category (Table S4), and this unequal distribution must be taken into account. Nunnink and Meana's study also identified differences in past pain levels according to current pain levels at the time of recall, observing that participants with lower levels of pain recalled past pain more accurately, similar to findings from previous studies conducted in different contexts or conditions (Bryant 1993; Previtali et al. 2022; Rasmussen et al. 2018; Smith and Safer 1993). In contrast, in our study, current pain at the time of recall impacted report of past pain in the opposite direction: participants who experienced severe pain for dysmenorrhea, abdominal pain and dysuria at the time of recall tended to provide more reliable responses than those experiencing lower pain levels then. This discrepancy may be explained by stark differences in study design: while these studies assessed past pain over a period of 30 days, our study compared two assessments of past pain occurring over the life course, and the two assessments were sent about 1 year apart, each potentially affected by the pain experienced at the time of assessment. This could suggest that current pain levels may impact recall of past pain differently according to the length of recall.

A study on fibromyalgia analysed pain recall over a longer period of time (up to 18 months) for momentary pain level, pain at its peak and pain at its lowest level (Van Liew et al. 2019). Pain was assessed at 5 time points using a 6-point pain scale. Since fibromyalgia also involves chronic pain, it is interesting to consider how this type of pain was analysed. The authors concluded that pain at its peak was the most stable of the painful experiences over time (ICC $=0.45$), which is the one we considered in our study. Additionally, this study was conducted with a relatively large female population ($n=572$), making it even more comparable to our study. To our knowledge, no previous study examined the reliability of measuring pain experienced many years ago. The present study is thus the first to address this topic.

We assessed agreement using weighted kappa coefficients and found that grouping pain levels into 5 categories was not relevant compared with the 11-item scale, which provided better reliability. Our observations are in line with those from Sim and Wright, which highlighted that grouping pain levels involves needless sacrifice of information from the original scale. This procedure implies a difficulty in the choice of categories while the results will largely depend on these choices (Sim and Wright 2005).

In our study, we found a better level of reliability for dysmenorrhea and dyspareunia, which may be due to their higher prevalences among patients (Kotowska et al. 2021; Schliep et al. 2015; Signorile et al. 2022). Also, some symptoms (e.g., abdominal pain) are common to many diseases, and it may be more difficult to remember less specific symptoms. Furthermore, although we initially anticipated the levels of agreement to be highest for the 'current age of participant' category due to the closer time proximity of assessed pain, this was not observed. This discrepancy might be due to the possibility that some participants may have experienced new pain or a different pain level during the interval between the two surveys. Notably, reliability was the highest for the longest interval between pain and survey response for dyspareunia and dysuria, which could be explained by the fact that these symptoms are less common under age 15, and also that most women had not yet had intercourse at this age.

While the WERF-EPHect EPQ-S questionnaire contains questions on both average and worst pain level, we elected to focus on the worst pain levels for several reasons. First, averaging involves remembering and summarising many experiences of pain over a period of time, and then aggregating and averaging these data into a single number (Broderick et al. 2006; Stone et al. 2004). This process can be challenging, particularly when assessing pain levels over a 10-year timeframe. Also, several studies have demonstrated that patients' memory is mainly influenced by the worst pain experienced during the period (Jensen et al. 2008; Redelmeier and Kahneman 1996; Stone et al. 2000). In addition, the aforementioned study on fibromyalgia, which analysed pain recall, found that peak pain was the most stable painful experience over time (Van Liew et al. 2019). Therefore, we considered that participants would more easily recall peak pain.

4.3 | Strengths and Limitations

Strengths of our study include the large sample size, with a high proportion of participants over 30, which provided many responses for several age ranges. Additionally, we used different reliability measures, allowing us to compare both exact pain levels and categories of pain levels. Moreover, we employed a standardised questionnaire used internationally in endometriosis studies (Vitonis et al. 2014), which will make our findings useful for future studies using this questionnaire. However, several limitations should be considered. Our study population is not representative of all endometriosis patients. Similar to other web-based cohorts, our results reflect a larger proportion of educated women (Andreeva et al. 2015; Kesse-Guyot et al. 2013). This population could also be particular in terms of disease severity, since among participants who self-reported endometriosis stage (around a third), over half reported stage IV. However, since half of the participants also reported that they had not undergone surgery, disease severity is probably lower than in clinical studies. Another limitation of our study lies in its entirely retrospective approach. All collected data were based on retrospective questions asked twice about past pain. This method differs from traditional reliability studies, which generally follow a prospective approach first, asking participants to report their pain level in real time, followed by a retrospective assessment to evaluate recall. Finally, it is essential to emphasise that retrospective assessments of pain trajectories cannot replace the value of prospective evaluations, which allow for a full control of recall bias. Further studies are thus needed to evaluate the evolution of pain levels among endometriosis patients prospectively, in order to describe accurate pain trajectories over time. Ideally, such evaluation should take place in adolescent prospective cohorts in order to explore the factors associated with the aggravation of pain symptoms over time.

5 | Conclusion

In this study, the reliability of evaluating the worst intensity of pain throughout the life course among women with endometriosis ranged from moderate to good depending on pain type, and reliability was higher when pain intensity was considered continuous. This suggests that, in the absence of prospective data, retrospective assessment of pain may reliably be used to assess the evolution of peak pain over the life course among women with endometriosis, and that the 11-point (NRS) scale should be used.

Author Contributions

M.K. conceived and designed the study. S.G. and S.L. performed the statistical analysis, and S.G. drafted the manuscript. All authors contributed to the interpretation of data discussed in the manuscript, critically revised the manuscript, and approved its final version.

Acknowledgements

We gratefully acknowledge the study participants for their continued involvement in this cohort and the patient organisations contributing to the success of this work (ENDOmind, EndoFrance, EndoAction, *La Belle et L'Endo*). We also thank Prof. Philippe Ravaud, principal investigator of ComPaRe, as well as the members of the ComPaRe team: Isabelle Pane, Elise Diard, Leslie Toko-Kamga, Carolina Riveros, Charline Garnier and Alain Brebion. We are also grateful to members of the ComPaRe-Endometriosis Scientific Committee: Nizar Aflak, Bruno Borghese, Juliette Gautier, Emilie Indersie, Delphine Lhuillery, Anne-Charlotte Millepied, Margaux Neve de Mevergnies, Erick Petit, Stéphane Ploteau, Horace Roman, Alice Romerio, Eric Sauvanet and Sylvain Tassy. In addition, we are grateful to the *Fondation Apicil* and the *Fondation de France* for their financial support of this work. S.G. is grateful to the French Ministry of Research and to the *Crédit Agricole Ile-de-France Mécénat* for their financial support during her PhD.

Ethics Statement

ComPaRe was approved by the Institutional Review Board of the Hôtel-Dieu Hospital, Paris (IRB: 0008367) and the French National Commission for Data Protection and Privacy (CNIL: 916397). This study's methods and procedures were conducted in accordance with the Declaration of Helsinki or comparable ethical standards. All participants provided written informed consent to participate and were all adults over the age of 18.

Consent

Patients who participated in the analyses of this study consent to publication.

Conflicts of Interest

Z.B.'s PhD was co-funded by the French National Association for Research and Technology (ANRT) and by Lyv Healthcare. The other authors have declared no conflicts of interest.

Data Availability Statement

All data collected for the study, including individual participant data and a data dictionary, are available for research purposes upon request to the corresponding author under the rules of the ComPaRe e-cohort (https://compare.aphp.fr).

References

Andreeva, V. A., B. Salanave, K. Castetbon, et al. 2015. "Comparison of the Sociodemographic Characteristics of the Large NutriNet-Santé e-Cohort With French Census Data: The Issue of Volunteer Bias Revisited." *Journal of Epidemiology and Community Health* 69: 893–898.

Becker, K., K. Heinemann, B. Imthurn, et al. 2021. "Real World Data on Symptomology and Diagnostic Approaches of 27,840 Women Living With Endometriosis." *Scientific Reports* 11: 20404.

Brauer, C., J. F. Thomsen, I. P. Loft, and S. Mikkelsen. 2003. "Can We Rely on Retrospective Pain Assessments?" *American Journal of Epidemiology* 157: 552–557.

Broderick, J. E., A. A. Stone, P. Calvanese, J. E. Schwartz, and D. C. Turk. 2006. "Recalled Pain Ratings: A Complex and Poorly Defined Task." *Journal of Pain* 7: 142–149.

Bryant, R. A. 1993. "Memory for Pain and Affect in Chronic Pain Patients." *Pain* 54: 347–351.

Bujang, M. A., and N. Baharum. 2017. "Guidelines of the Minimum Sample Size Requirements for Kappa Agreement Test." *Epidemiology, Biostatistics and Public Health* 14: e12267-1.

Comptour, A., C. Lambert, P. Chauvet, et al. 2020. "Long-Term Evolution of Quality of Life and Symptoms Following Surgical Treatment for Endometriosis: Different Trajectories for Which Patients?" *Journal of Clinical Medicine* 9: 2461.

Coughlin, S. S. 1990. "Recall Bias in Epidemiologic Studies." *Journal of Clinical Epidemiology* 43: 87–91.

Dillon, D. G., and D. A. Pizzagalli. 2018. "Mechanisms of Memory Disruption in Depression." *Trends in Neurosciences* 41: 137–149.

Gamer, M., J. Lemon, and I. F. P. Singh. 2019. "irr: Various Coefficients of Interrater Reliability and Agreement."

Glazier, B. L., and L. E. Alden. 2017. "Social Anxiety and Biased Recall of Positive Information: It's Not the Content, It's the Valence." *Behavior Therapy* 48: 533–543.

Gouesbet, S., M. Kvaskoff, C. Riveros, et al. 2023. "Patients' Perspectives on How to Improve Endometriosis Care: A Large Qualitative Study Within the ComPaRe-Endometriosis e-Cohort." *Journal of Women's Health* 32: 463–470.

Jamison, R. N., S. A. Raymond, E. A. Slawsby, G. J. McHugo, and J. C. Baird. 2006. "Pain Assessment in Patients With Low Back Pain: Comparison of Weekly Recall and Momentary Electronic Data." *Journal of Pain* 7: 192–199.

Jensen, M. P., J. Mardekian, M. Lakshminarayanan, and M. E. Boye. 2008. "Validity of 24-h Recall Ratings of Pain Severity: Biasing Effects of "Peak" and "End" Pain." *Pain* 137: 422–427.

Kesse-Guyot, E., V. Andreeva, K. Castetbon, et al. 2013. "Participant Profiles According to Recruitment Source in a Large Web-Based Prospective Study: Experience From the Nutrinet-Santé Study." *Journal of Medical Internet Research* 15: e2488.

Koo, T. K., and M. Y. Li. 2016. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." *Journal of Chiropractic Medicine* 15: 155–163.

Kotowska, M., J. Urbaniak, W. J. Falęcki, P. Łazarewicz, M. Masiak, and I. Szymusik. 2021. "Awareness of Endometriosis Symptoms—A Cross Sectional Survey Among Polish Women." *International Journal of Environmental Research and Public Health* 18: 9919.

Kranklader, É., and A. Schreiber. 2015. "Le Sentiment D'aisance Financière Des Ménages: Stable Au Fil Des Générations, Mais Fluctuant Au Cours de La Vie. Insee 69–86."

Kroenke, K., R. L. Spitzer, and J. B. Williams. 2001. "The PHQ-9: Validity of a Brief Depression Severity Measure." *Journal of General Internal Medicine* 16, no. 9: 606–613.

Landis, J. R., and G. G. Koch. 1977. "The Measurement of Observer Agreement for Categorical Data." *Biometrics* 33, no. 1: 159–174.

Larroy, C. 2002. "Comparing Visual-Analog and Numeric Scales for Assessing Menstrual Pain." *Behavioral Medicine* 27: 179–181.

Linden, M. V. D., C. Wyns, R. Bruyer, C. Ansay, and X. Seron. 1993. "Effect of Educational Level on Cued Recall in Young and Elderly Subjects." *Psychologica Belgica* 33, no. 1: 37–47.

McGraw, K. O., and S. P. Wong. 1996. "Forming Inferences About Some Intraclass Correlation Coefficients." *Psychological Methods* 1: 30–46.

Munnangi, S., and S. W. Boktor. 2022. "Epidemiology of Study Design." In *StatPearls*. StatPearls Publishing.

Nunnink, S., and M. Meana. 2007. "Remembering the Pain: Accuracy of Pain Recall in Endometriosis." *Journal of Psychosomatic Obstetrics & Gynecology* 28: 201–208.

Previtali, D., A. Boffa, A. Di Martino, L. Deabate, M. Delcogliano, and G. Filardo. 2022. "Recall Bias Affects Pain Assessment in Knee Osteoarthritis: A Pilot Study." *Cartilage* 13: 50–58.

Rasmussen, C. D. N., A. Holtermann, and M. B. Jørgensen. 2018. "Recall Bias in Low Back Pain Among Workers." *Spine (Phila Pa 1976)* 43: E727–E733.

Redelmeier, D. A., and D. Kahneman. 1996. "Patients' Memories of Painful Medical Treatments: Real-Time and Retrospective Evaluations of Two Minimally Invasive Procedures." *Pain* 66: 3–8.

Schliep, K. C., S. L. Mumford, C. M. Peterson, et al. 2015. "Pain Typology and Incident Endometriosis." *Human Reproduction* 30: 2427–2438.

Shrout, P. E., and J. L. Fleiss. 1979. "Intraclass Correlations: Uses in Assessing Rater Reliability." *Psychological Bulletin* 86: 420–428.

Signorile, P. G., M. Cassano, R. Viceconte, V. Marcattilj, and A. Baldi. 2022. "Endometriosis: A Retrospective Analysis of Clinical Data From a Cohort of 4,083 Patients, With Focus on Symptoms." *In Vivo* 36: 874–883.

Sim, J., and C. C. Wright. 2005. "The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements." *Physical Therapy* 85: 257–268.

Smith, W. B., and M. A. Safer. 1993. "Effects of Present Pain Level on Recall of Chronic Pain and Medication Use." *Pain* 55: 355–361.

Spitzer, R. L., K. Kroenke, J. B. W. Williams, and B. Löwe. 2006. "A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7." *Archives of Internal Medicine* 166, no. 10: 1092–1097.

Stone, A. A., J. E. Broderick, A. T. Kaell, P. A. DelesPaul, and L. E. Porter. 2000. "Does the Peak-End Phenomenon Observed in Laboratory Pain Studies Apply to Real-World Pain in Rheumatoid Arthritics?" *Journal of Pain* 1: 212–217.

Stone, A. A., J. E. Broderick, S. S. Shiffman, and J. E. Schwartz. 2004. "Understanding Recall of Weekly Pain From a Momentary Assessment Perspective: Absolute Agreement, Between- and Within-Person Consistency, and Judged Change in Weekly Pain." *Pain* 107: 61–69.

Talari, K., and M. Goyal. 2020. "Retrospective Studies - Utility and Caveats." *Journal of the Royal College of Physicians of Edinburgh* 50: 398–402.

Tran, V.-T., and P. Ravaud. 2020. "COllaborative Open Platform E-Cohorts for Research Acceleration in Trials and Epidemiology." *Journal of Clinical Epidemiology* 124: 139–148.

Van Liew, C., K. Standridge, G. Leon, and T. A. Cronan. 2019. "A Longitudinal Analysis of Pain Experience and Recall in Fibromyalgia." *International Journal of Rheumatic Diseases* 22: 497–506.

Vitonis, A. F., K. Vincent, N. Rahmioglu, et al. 2014. "World Endometriosis Research Foundation Endometriosis Phenome and Biobanking Harmonization Project: II. Clinical and Covariate Phenotype Data Collection in Endometriosis Research." *Fertility and Sterility* 102: 1223–1232.

Williamson, A., and B. Hoggart. 2005. "Pain: A Review of Three Commonly Used Pain Rating Scales." *Journal of Clinical Nursing* 14: 798–804.

Zondervan, K. T., C. M. Becker, K. Koga, S. A. Missmer, R. N. Taylor, and P. Viganò. 2018. "Endometriosis." *Nature Reviews Disease Primers* 4, no. 9: 1244–1256.

Zou, G. Y. 2012. "Sample Size Formulas for Estimating Intraclass Correlation Coefficients With Precision and Assurance." *Statistics in Medicine* 31: 3972–3981.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.