General Section
Research Paper

OPEN

PAIN REPORTS

Counting your chickens before they hatch: improvements in an untreated chronic pain population, beyond regression to the mean and the placebo effect

Monica Sean[a,b,c], Alexia Coulombe-Lévêque[a,d,e], William Nadeau[a], Anne-Catherine Charest[a], Marylie Martel[a,c], Guillaume Léonard[a,d,e], Pascal Tétreault[a,b,c,f,*]

Abstract

Introduction: Isolating the effect of an intervention from the natural course and fluctuations of a condition is a challenge in any clinical trial, particularly in the field of pain. Regression to the mean (RTM) may explain some of these observed fluctuations.

Objectives: In this paper, we describe and quantify the natural trajectory of questionnaire scores over time, based on initial scores.

Methods: Twenty-seven untreated chronic low back pain patients and 25 healthy controls took part in this observational study, wherein they were asked to complete an array of questionnaires commonly used in pain studies during each of 3 visits (V1, V2, V3) at the 2-month interval. Scores at V1 were classified into 3 subgroups (extremely high, normal, and extremely low), based on z-scores. The average delta ($\Delta = V2 - V1$) was calculated for each subgroup, for each questionnaire, to describe the evolution of scores over time based on initial scores. This analysis was repeated with the data for V2 and V3.

Results: Our results show that high initial scores were widely followed by more average scores, while low initial scores tended to be followed by similar (low) scores.

Conclusion: These trajectories cannot be attributable to RTM alone because of their asymmetry, nor to the placebo effect as they occurred in the absence of any intervention. However, they could be the result of an Effect of Care, wherein participants had meaningful improvements simply from taking part in a study. The improvement observed in patients with high initial scores should be carefully taken into account when interpreting results from clinical trials.

Keywords: Chronic pain, Regression to the mean, Effect of care, Pain questionnaires

1. Introduction

Pain is a highly subjective and variable phenomenon; as such, it is famously difficult to measure accurately—even more so when it comes to measuring *changes* in pain levels.[31] Indeed, pain levels fluctuate naturally, as they are affected by a wide array of biopsychosocial factors, such as sleep, mood, expectations, and beliefs, themselves fluctuating and difficult to measure.[3,14,15,16,19,21,27] Clinical trials do their best to quantify the effects of their interventions using the most valid and reliable questionnaires at their disposal.[6,9,28] Unfortunately, it remains difficult to isolate the effect of an intervention from the natural course and fluctuations of the condition.[14,19,26] This is further complicated by a relatively well-known phenomenon: regression to the mean (RTM).[8,17,18,23,30,32]

Sponsorships or competing interests that may be relevant to content are disclosed at the end of this article.

[a] Faculty of Medicine and Health Sciences, Université de Sherbrooke, Sherbrooke, QC, Canada, [b] Department of Anesthesiology, Sherbrooke, QC, Canada, [c] Centre de Recherche du CHUS, Sherbrooke, QC, Canada, [d] School of Rehabilitation, Sherbrooke, QC, Canada, [e] Research Centre on Aging, Sherbrooke, QC, Canada, [f] Department of Nuclear Medicine and Radiobiology, Sherbrooke, QC, Canada

**Corresponding author. Address: Faculty of Medicine and Health Sciences, Université de Sherbrooke, 3001 12e Ave Nord, J1H 5N4, Sherbrooke, Québec, Canada. Tel.: 819-821-8000 (ext: 74502). E-mail address: Pascal.Tetreault@USherbrooke.ca (P. Tétreault).*

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.painrpts.com).

PR9 9 (2024) e1157

http://dx.doi.org/10.1097/PR9.0000000000001157

Regression to the mean is not another biopsychosocial factor that *influences* pain levels: it is a *statistical concept* that can—in part—describe, explain, and predict those fluctuations. Regression to the mean is based on probability distributions and states that extreme scores are likely to be followed by less extremes scores that are closer to the individual's own sampling mean.[8] In chronic pain studies, multiple questionnaires are often used to assess pain and specific biopsychosocial factors, which all have their intrinsic variability.[2,7,25] It is therefore possible that two questionnaires will show different RTM for the same subject over the same time period. In other words, a questionnaire measuring a comparatively more fluctuating factor will be more susceptible to RTM.

Regression to the mean is often mentioned in the discussion section of clinical trial reports as a possible alternative explanation for observed changes in outcomes over time. However, RTM has rarely been the primary focus of investigation in a chronic pain population. This is unfortunate, seeing as the results obtained from such investigations could prove useful on many levels. For example, a better understanding of RTM could help guide the choice of outcome measures in clinical trials (favoring those less susceptible to RTM), and improve result interpretation by helping researchers to differentiate changes in outcome measures resulting from treatment effect vs RTM. The study design suitable for such RTM assessment requires that patients with chronic pain be assessed using a large array of validated, commonly used questionnaires, at different time points (at least twice, ideally more), with no concomitant intervention taking place outside of usual care. Our team had such an observational study taking place to assess changes in brain structure and functional activity over time in patients with chronic low back pain (CLBP) and healthy controls (HC). We were therefore able to conduct the analysis presented below as part of that study.

The objectives of this analysis were (1) to describe and quantify the natural trajectory of questionnaire scores over time, based on initial scores, with a subgoal of determining whether the observed fluctuations were compatible with RTM, and (2) to evaluate and compare the stability of each questionnaire over time, in 27 untreated CLBP and 25 healthy controls.

2. Methods

The study was conducted in accordance with the Declaration of Helsinki. All participants provided written informed consent for their participation into the study. Ethics approval was granted from the institutional review board of the Centre intégré universitaire de santé et de services sociaux de l'Estrie—Centre hospitalier universitaire de Sherbrooke (Sherbrooke, Quebec, Canada; approval #2021-3861). The trial has been registered on Open Science Framework (OSF), under the name "Pilot project on brain and lower back imaging of chronic pain" (https://doi.org/10.17605/OSF.IO/P2Z6Y).

2.1. Participants

Twenty-seven patients with CLBP and 25 HC aged 18 to 75 years old took part in this study (convenience sampling). Healthy controls were matched with CLBP patients for sex and age.

Specific inclusion criteria for the CLBP were (1) low back pain (≥6 months); (2) average daily pain intensity of ≥3/10; (3) pain primarily localized in the lower back; and (4) no history of invasive treatment to manage their pain. Specific exclusion criteria for HC were (1) history of chronic pain; (2) recent (<3 months) acute pain; and (3) pain at the time of testing.

Exclusion criteria for the 2 populations included (1) neurological, cardiovascular, or pulmonary disorders; (2) comorbid pain syndrome; (3) history of back surgery; (4) use of opioids, antidepressants, anticonvulsants, or psychostimulants; (5) recent (<1 year) corticosteroid infiltration; (6) pregnancy; (7) inability to read or understand French; and (8) contraindication to magnetic resonance imaging (MRI).

2.2. Study design

The study had an observational longitudinal design. All participants attended 3 sessions (V1, V2, V3) at 2 months intervals where they completed several questionnaires (discussed in the present paper) and underwent brain and lumbar MRI (as part of the larger study, not discussed in this paper) at the Centre de recherche du CHUS.

2.3. Questionnaires

All questionnaires were completed online using the platform "Research Electronic Data Capture" (REDCap) and are presented in **Table 1**.

Participants with CLPB completed all questionnaires; HC completed only the Pain Catastrophizing Scale (PCS) and the State and Trait Anxiety Inventory (STAI/S-T) (two questionnaires applicable to a healthy population). To avoid fatigue caused by filling out multiple questionnaires, the Pain Disability Index (PDI) (for the CLBP participants) and the PCS (for all participants) were completed at home, one week before each visit.

2.4. Data processing and statistical analysis

2.4.1. Group attribution

To describe the behavior of "extreme" vs "normal" scores of participants, it was first necessary to establish a criterion to differentiate "extreme" and "normal" scores. This was done by transforming raw initial scores into studentized scores (ie, z-scores) for each questionnaire. Multiple z-score thresholds were tested, and a threshold of $|z| > 0.5$ was found to yield the most similar number of participants across the 3 subgroups ("extremely high," "normal," "extremely low"). As such, scores with $|z| > 0.5$ (ie, scores that were more than half a standard deviation above or below the group average) were considered "extreme," whereas scores with $|z| < 0.5$ (ie, scores within half a standard deviation of the group average) were considered "normal." An exploratory analysis with various thresholds revealed that, regardless of the threshold used, a similar pattern emerged from our results. All results obtained using the different thresholds tested ($|z| > 0.66$; $|z| > 0.8$ and $|z| > 1$) are included in supplementary materials, http://links.lww.com/PR9/A231.

Scores at V1 were thus classified as (1) extremely high, (2) normal, or (3) extremely low. This was done independently for each questionnaire such that a given participant could be in the "extremely high" subgroup for one questionnaire, but in the "normal" subgroup for another questionnaire. Next, the delta between V1 and V2 was calculated by subtracting the score at V1 from the score at V2 ($\Delta = V2 - V1$) such that a positive delta corresponds to a score increase (ie, worsening of the condition) and a negative delta corresponds to a score decrease (ie, improvement of the condition). The same analysis was conducted between V2 and V3: scores at V2 were again classified as (1) extremely high, (2) normal, or (3) extremely low, and the delta between V2 and V3 was calculated by subtracting the score at V2

Table 1

Questionnaires completed by patients with chronic low back pain and healthy controls

Questionnaires	Subscales	Description	Items	Scale	Total score	French validated version
Pain Catastrophizing Scale (PCS)[28]	—	Degree of catastrophic thoughts (helplessness, magnification, and rumination)	13	5-point Likert scale ("not at all" to "all the time")	0–52	Yes[10]
Pain Disability Index (PDI)[1]	—	Ability to perform daily activities (home, social, recreational, occupational, sexual, self-care, and life support activities)	7	Numerical rating scale (NRS) (0 = no disability to 10 = worst disability)	0–70	Yes[13]
Brief Pain Inventory (BPI) *short form*[6]	Pain severity (BPIs)	Intensity of pain (current, average, least and worst pain in the last 24 hours)	4	NRS (0 = no pain; 10 = worst pain imaginable)	0–40	Yes[6]
	Pain interference (BPIi)	Interference of pain with daily activities (sleeping, walking, mood, etc.)	7	NRS (0 = no interference; 10 = complete interference)	0–70	Yes[6]
PainDETECT (PD)[11]	PD	Presence of neuropathic pain components in patients with back pain, such as burning sensation and electric shocks.	9	7 items rated on a 6-point Likert Scale (0 = not at all, 5 = very strongly), 1 item based on pain behavior pattern score (−1, 0 or 1) & 1 item based on a radiation score (0 or 2)	0–38	Yes[11]
	PD Severity (PDs)	Intensity of pain (current, average in the past 4 wk, worst in the past 4 wk)	3	NRS (0 = no pain; 10 = worst pain imaginable)	0–10	Yes[11]
Pain Outcomes Questionnaire (POQ)[5]	—	Global function (eg mobility, vitality, affect, daily activities, and pain)	19	NRS (0 = less symptoms to 10 = more severe symptoms)	0–190	No (In-house translation)
State-Trait Anxiety Inventory (STAI-S/T)[27]	State anxiety (STAI-T)	Current anxiety	20	4-point Likert scale ("not at all" to "all the time")	20–80	Yes[12]
	State anxiety (STAI-S)	General anxiety	20	4-point Likert scale ("not at all" to "all the time")	20–80	Yes[12]
McGill Pain Questionnaire (MPQ) *short form*[22]	MPQ	Sensory-affective components of pain	15	4-point Likert scale ("no pain"; to "severe pain")	0–45	Yes[4]
	MPQ intensity (MPQi)	Pain intensity (previous week)	1	100-point visual analogue scale (left anchor: "no pain"; right anchor: "worst possible pain")	0–100	Yes[4]
Central Sensitization Inventory (CSI) *short form*[20]	—	Symptoms of central sensitization	25	5-point Likert scale (0 = "never"; 4 = "always")	0–100	Yes[24]

CLBP patients completed 8 questionnaires: (1) Pain Catastrophizing Scale (PCS), (2) Pain Disability Index (PDI), (3) Brief Pain Inventory (BPI), (4) Pain DETECT, (5) Pain Outcomes Questionnaire (POQ), (6) State-Trait Anxiety Inventory (STAI/S-T), (7) McGill Pain Questionnaire (MPQ), and (8) Central Sensitization Inventory (short form) (CSI). HC completed only the PCS and the STAI/S-T.

from the score at V3 (Δ = V3 − V2). Thus, for each questionnaire, 2 calculations were performed (V2 − V1 and V3 − V2).

2.4.2. Standardization across questionnaires

To facilitate the comparison between questionnaires, all raw scores were reported on a scale from 0 to 100. Fluctuations larger than 10 percentage points were considered clinically meaningful, and fluctuations of 5 percentage points or less were considered random noise. Fluctuations between 5 and 10 percentage points, while of debatable clinical relevance, were still considered likely enough to denote an effect to warrant being reported. The use of such standardized thresholds, as opposed to the Minimal Detectable Change (MDC) specific to each questionnaire, was favored because it allowed for direct comparisons between questionnaires; the advantages and drawbacks of this methodological choice are highlighted in the Discussion.

2.4.3. Average delta scores

Once participants were divided into the 3 subgroups (based on their initial scores), average delta scores were calculated for each subgroup within each questionnaire. Delta scores were calculated by subtracting V1 scores from V2 scores (and V2 scores from V3 scores) such that a negative delta represents a decrease in score (which, for all outcome measures, corresponds to an improvement of the condition).

Delta scores were averaged within each subgroup, for each questionnaire, to yield a measure of the average evolution over time of each subgroup.

2.4.4. Fluctuation scores

In addition to *average* delta scores, "fluctuation scores" were calculated for each subgroup within each questionnaire by averaging the *absolute value* of delta scores for the given subgroup. This measure was particularly informative in cases where both large decreases and large increases scores had taken place: such decreases and increases would cancel each other out in the "average delta scores" and could lead us to conclude that scores remained roughly stable over time, when in fact large fluctuations (in opposite directions) had taken place.

3. Results

3.1. Participants and raw scores

Fifty-two participants (25 HC and 27 CLBP) were recruited in the study. Three CLBP participants dropped out after the first visit (unexpected pregnancy [n = 1], discomfort during MRI [n = 1], scheduling conflicts [n = 1]), and one dropped out after the second visit (move to a different city [n = 1]) such that 23 CLBP completed the entire study and were included in the analysis.

Table 2

Sociodemographic characteristics of the sample.

	CLBP (n = 23)	HC (n = 25)
Biological sex		
Women	11	15
Men	12	10
Age (average ± SD)	44 ± 15	40 ± 14
Etnicity		
Caucasian	16	24
Asiatic	1	1
Hispanic	4	NA
African	1	NA
Arabic	1	NA
Education level		
Primary school	NA	NA
High school	NA	NA
Apprenticeship	5	2
College	4	6
University	14	17
Annual income		
less than 20K	2	5
20K–35K	5	2
35K–50K	5	1
50K–65K	5	6
65K–80K	2	5
80K–100K	3	4
100K and more	1	2
Pain duration		
4 mo–5 mo	0	NA
6 mo–12 mo	5	NA
1–4 y	7	NA
5 y and more	11	NA

There were no dropouts among the HC. Sociodemographic characteristics of the sample are presented in **Table 2**. All participants complied with the instructions to avoid any treatment other than over-the-counter medication and their usual, non-invasive rehabilitation treatments. This allowed us to evaluate the natural course of the condition during the period of the study. Average scores for each questionnaire at each visit are presented in **Table 3a and b** for the 2 populations.

3.2. Average evolution over time as a function of initial score

The evolution of questionnaire scores for each subgroup over time is presented in **Table 4** (CLBP) and **Table 5** (HC). The same data are represented in 3 different ways: **Tables 4a and 5a** present the *average* delta within each subgroup, **Tables 4b and 5b** presents the *individual* delta scores of each member of the subgroup, and **Tables 4c and 5c** presents the average *absolute* delta within each subgroup. As a core example, PCS scores for the CLBP participants in the 'extremely high' subgroup at V1 (ie, participants with a z-score >0.5) showed, on average, a reduction of 22 percentage points at V2 (corresponding to a raw reduction of 12 points on the PCS scale) (**Table 4a**). *Individual* delta scores making up that average delta score are presented in **Table 4b**. Still on the PCS, in the "normal" CLBP subgroup, some participants had an *increase* in scores between V1 and V2, and some had a *decrease* in scores (**Table 4b**). This yielded an *average* delta score of only −4/100 (**Table 3a**), but a much larger average *absolute* delta score of 14/100 (**Table 4c**). To visually illustrate the grouping, we have plotted McGill Pain Questionnaire intensity (MPQi) average pain scores for each subgroup and each time point (**Fig. 1**).

3.3. Chronic low back pain sample

3.3.1. Average evolution of "extremely high" scores

Participants with an initial "extremely high" score at V1 tended to show a reduction in score at V2, and those with an "extremely high" score at V2 similarly tended to show a reduction in score at V3 (**Table 4a and b**, top rows). Indeed, the *average* deltas from V1 to V2 and from V2 to V3 were mostly negative in that subgroup, and most *individual* deltas were negative (**Table 4b**, top row).

This trend for high scores to be followed by lower scores is expected, as RTM predicts that extreme scores will be followed by more normal scores, which in the case of extremely *high* scores means that the subsequent score should be *lower*.

3.3.2. Average evolution of "normal" scores

The evolution of "normal" scorers is presented in the second rows of **Tables 4a–c**. *Average* scores tended to remain stable or to slightly decrease over time, and the visual representation of individual delta scores (**Table 4b**, middle row) reveals that participants in this subgroup tended to show an uneven split between increases and decreases in scores from one visit to the next, with a larger number of individual delta scores being negative.

Table 3

Average scores for each questionnaire during the 3 visits (V1, V2, and V3), for the chronic low back pain sample and the healthy controls sample.

(a)

CLBP	PCS (0–52)	MPQi (0–100)	BPIs (0–40)	MPQ (0–45)	PDs (0–10)	BPIi (0–70)	PDI (0–70)	STAI/S (20–80)	PD (0–38)	POQ (0–190)	CSI (0–100)	STAI/T (20–80)
V1	19 ± 12	56 ± 16	18 ± 5	14 ± 7	5 ± 1	18 ± 10	16 ± 9	33 ± 8	7 ± 4	44 ± 14	35 ± 11	35 ± 11
V2	16 ± 10	54 ± 20	18 ± 6	13 ± 7	5 ± 1	16 ± 10	14 ± 9	33 ± 9	9 ± 4	43 ± 18	33 ± 13	35 ± 12
V3	12 ± 11	46 ± 23	15 ± 7	10 ± 8	5 ± 2	9 ± 7	9 ± 7	32 ± 10	9 ± 5	37 ± 19	30 ± 14	35 ± 12

(b)

HC	PCS (0–52)	STAI/S (20–80)	STAI/T (20–80)
V1	6 ± 8	26 ± 6	29 ± 9
V2	6 ± 7	26 ± 6	29 ± 9
V3	5 ± 7	26 ± 8	29 ± 10

Average for CLBP sample is presented in Table 3a for CLBP and Table 3b for HC. Scores are reported as average ± standard deviation. The theoretical min and max scores for each questionnaire are reported in the title row.
BPIi, brief pain inventory—interference; BPIs, brief pain inventory—severity; CSI, central sensitization inventory; MPQ, McGill pain questionnaire; MPQi, McGill pain questionnaire intensity; PCS, pain catastrophizing scale; PD, Pain DETECT; PDs, Pain DETECT severity; PDI, pain disability index; POQ, pain outcomes questionnaire; STAI/S-T, state-trait anxiety inventory.

Table 4

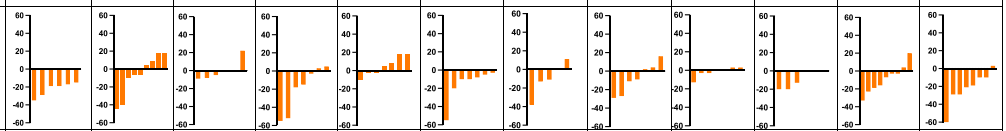
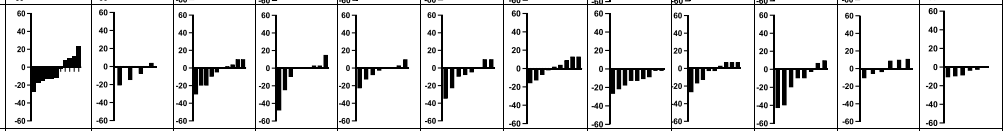
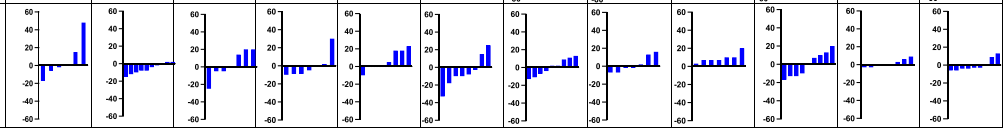
The evolution over time of participants with "extremely high," "extremely low," or "normal" initial scores at V1 and V2, in the chronic low back pain sample.

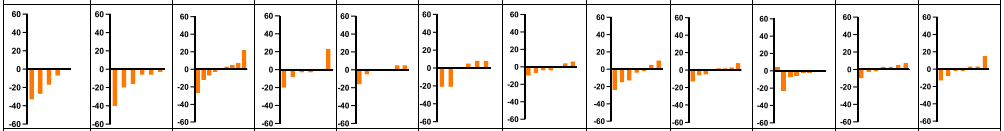
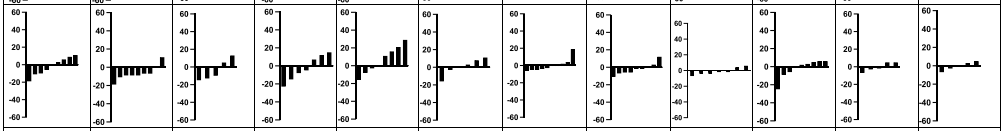
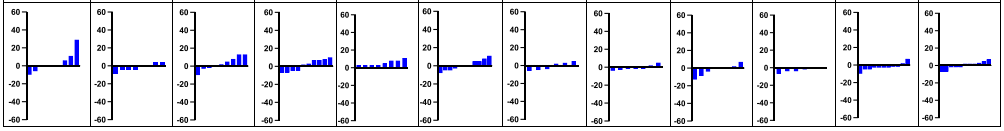
a)

	PCS (0-52)		MPQi (0-100)		BPIs (0-40)		MPQ (0-45)		PDs (0-10)		BPIi (0-70)	
CLBP	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2
Extremely high	**-22 (-12)**	-7 (-3)	0	**-19 (-19)**	5	**-16 (-6)**	-10 (-5)	-8 (-3)	-2	-9 (-1)	-9 (-6)	**-22 (-15)**
Normal	-4	-10 (-5)	-6 (-6)	-7 (-7)	-4	-8 (-3)	0	**-13 (-6)**	-4	**-14 (-1)**	1	-6 (-4)
Extremely low	6 (3)	-5	3	0	7 (3)	-5	0	2	9 (1)	0	1	0

	PDI (0-70)		STAI/S (20-80)		PD (0-38)		POQ (0-190)		CSI (0-100)		STAI/T (20-80)	
	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2
	-17 (-12)	**-15 (-11)**	-1	-2	-2	-4	-2	-6 (-12)	-1	-5	0	0
	-2	-7 (-5)	-4	-3	6 (2)	0	0	-2	-1	-2	0	0
	3	-1	3	1	5	1	-1	-1	-2	-3	-3	0

b)

CLBP	PCS		MPQi		BPIs		MPQ		PDs		BPIi	
	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2
Extremely high												
Normal												
Extremely low												

	PDI		STAI/S		PD		POQ		CSI		STAI/T	
	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2
												
												
												

c)

CLBP	PCS		MPQi		BPIs		MPQ		PDs		BPIi	
	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2
Extremely high	22	17	8	22	9	16	15	14	4	9	14	23
Normal	14	12	11	12	7	13	9	13	10	18	9	6
Extremely low	15	6	13	9	9	15	8	7	9	11	3	5
Total (average of all participants)	16	11	11	14	8	14	10	11	8	13	9	12
Total (average of both deltas)	14		12		11		11		10		10	

	PDI		STAI/S		PD		POQ		CSI		STAI/T	
	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2
	17	15	9	10	5	11	5	10	5	6	5	7
	8	9	11	12	13	7	5	5	4	7	4	4
	7	4	6	6	5	5	4	3	5	3	4	3
	10	9	9	9	8	7	5	6	5	5	4	4
	9		9		7		5		5		4	

The analyses from V1 to V2 and from V2 from V3 were conducted independently such that a participant could be in the "extremely high" subgroup at V1 (and in the V1-V2 analysis) and in the "extremely low" subgroup at V2 (and in the V2-V3 analysis). Similarly, questionnaires were analysed independently such that a participant could be in the "extremely high" subgroup for one questionnaire and in the "normal" subgroup for another questionnaire. Scores are reported on 100, with raw scores in parenthesis when a significant change (>10%) occurred.

The same data are presented from 3 different angles: Tables 4a–c. Table 4a presents the average delta for each subgroup (changes larger than 10% are in bold); Table 4b presents individual deltas within each subgroup; and Table 4c presents the average of absolute deltas for each subgroup.

BPIi, brief pain inventory—interference; BPIs, brief pain inventory—severity; CSI, central sensitization inventory; MPQ, McGill pain questionnaire; MPQi, McGill pain questionnaire intensity; PCS, pain catastrophizing scale; PD, Pain DETECT; PDs, Pain DETECT severity; PDI, pain disability index; POQ, pain outcomes questionnaire; STAI/S-T, state-trait anxiety inventory.

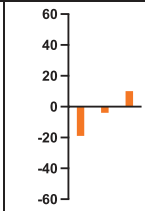
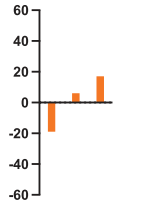
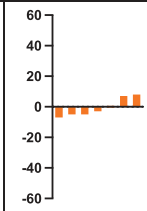
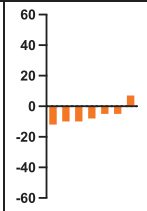
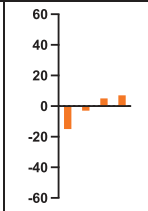
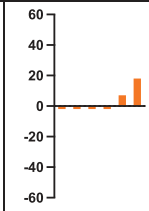
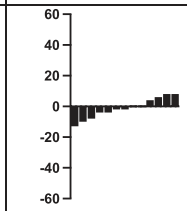
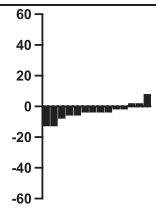
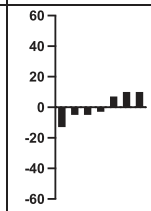
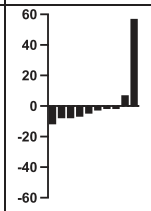
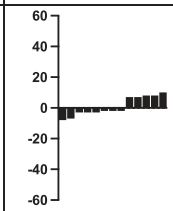
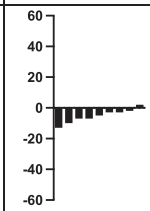
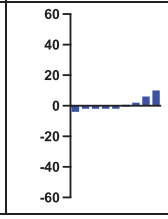
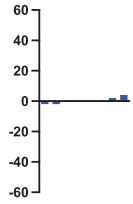
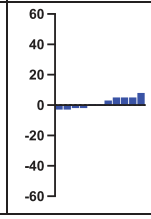
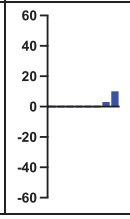
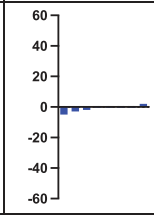
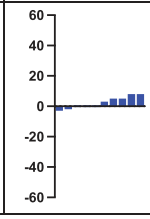
Table 5

The evolution over time of participants with "extremely high," "extremely low," or "normal" initial scores at V1 and V2, in the healthy controls sample.

a)

HC	PCS (0-52)		STAI/S (20-80)		STAI/T (20-80)	
	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2
Extremely high	-4	1	-1	-6 (-4)	-2	3
Normal	-1	-4	0	2	1	-5
Extremely low	1	0	2	2	-1	3

b)

HC	PCS		STAI/S		STAI/T	
	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2
Extremely high						
Normal						
Extremely low						

c)

HC	PCS		STAI/S		STAI/T	
	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2	Δ=V2-V1	Δ=V3-V2
Extremely high	11	14	5	8	8	5
Normal	5	5	8	11	5	6
Extremely low	3	1	3	2	1	4
Total (average of all participants)	5	5	5	7	4	5
Total (average of both deltas)	5		6		5	

The analyses from V1 to V2 and from V2 from V3 were conducted independently such that a participant could be in the "extremely high" subgroup at V1 (and in the V1-V2 analysis) and in the "extremely low" subgroup at V2 (and in the V2-V3 analysis). Similarly, questionnaires were analysed independently such that a participant could be in the "extremely high" subgroup for one questionnaire and in the "normal" subgroup for another questionnaire. Scores are reported on 100, with raw scores in parenthesis when a significant change (>10%) occurred.

The same data are presented from 3 different angles: Tables 5a–c. Table 5a presents the average delta for each subgroup; Table 5b presents individual deltas within each subgroup; and Table 5c presents the average of absolute deltas for each subgroup.

PCS, Pain Catastrophizing Scale; STAI/S-T, state-trait anxiety inventory.

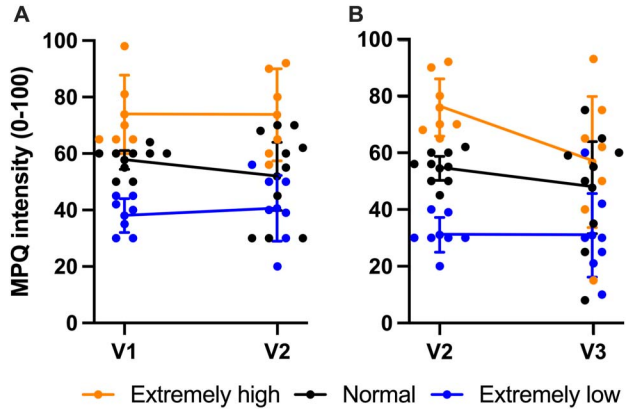


Figure 1. Average pain intensity scores from the McGill Pain Questionnaire (MPQi) for each subgroup, for the V1-V2 (A) and V2-V3 (B) analyses. For the V1-V2 analysis (A), subgroup allocation is based on scores at V1; for the V2-V3 analysis (B), subgroup allocation is based on scores at V2 such that a given participant can be allocated to different subgroups for the 2 analyses. Dots represent individual score. The error bars represent 95% confidence intervals (IC95).

Regression to the mean predicts that participants with a "normal" score will show little or no change from one session to the next, with an even distribution of increases and decreases cancelling each other out. This should translate in average delta scores being roughly equal to 0 (**Table 4a**) and, visually, in a roughly even and symmetrical split between individual increases and decreases (**Table 4b**). As this is not the pattern of results that we observed, our results suggest that RTM alone cannot account for the overall decrease in scores seen in some outcome measures (see discussion).

3.3.3. Average evolution of 'extremely low' scores

As can be seen in **Table 4a**, on average, participants with extremely low initial scores appear to remain stable from one visit to the next on most questionnaires. However, on roughly half of the questionnaires, these seemingly "stable" average deltas are a product of significant individual increases and decreases that roughly cancel each other out, as presented visually in **Table 4b** and quantified in **Table 4c**.

Regression to the mean predicts that extremely low scores will *increase* towards more "normal" scores on the subsequent measurement. Overall, there appears to have been a slight RTM effect on a few questionnaires, although most average deltas are close to 0, suggesting that no substantial RTM was at play—or that some other effect was at play that counteracted RTM (see discussion).

3.3.4. Analysis by questionnaire—average evolution

For the "extremely high" subgroup, the PCS, PDI, BPIi, BPIs, and MPQi all showed an average decrease larger than 15 percentage points for at least one time period (**Table 4a**), while the PD, CSI, and both subscales of the STAI showed no change on average from one visit to the next, for both time periods.

The "normal" subgroup was more stable overall, with clinically meaningful average fluctuations observed over a single time period only for the MPQ and PDs, and strong stability on the CSI, both subscales of the STAI, and the POQ (**Table 4a**).

The "extremely low" subgroup had the most stable scores of all, showing no clinically meaningful average change on any questionnaires (**Table 4a**).

Overall, the PCS shows the largest average change for all 3 subgroups, followed by the PDI, PDs, both subscales of the MPQ, and both subscales of the BPI. The questionnaires with the smallest average delta were the CSI and both subscales of the STAI (**Table 4a**).

3.3.5. Analysis by questionnaire—average absolute *fluctuation*

As mentioned previously, it is possible for a questionnaire to have an average delta of roughly 0 from one visit to the next, seemingly suggesting that all participants remained stable over time, while in fact large individual increases and decreases in scores have been taking place, cancelling each other out. For all 3 subgroups, the PCS shows the largest magnitude of fluctuation in scores (**Table 4c**), followed closely by the MPQi, BPIs, MPQ, PDs, and BPIi. The STAI-T was the most stable questionnaire.

3.4. Healthy controls

Healthy controls had much more stable scores overall compared with patients with CLBP (**Table 5**). Indeed, the *average* evolution over time was smaller than 5 percentage points for all 3 subgroups, on all questionnaires and at both time points (with one exception at 6 percentage points) (**Table 5a**). Visually, apart from a notable outlier on the STAI-S (a grad student who reported having a particularly stressful day), most individual fluctuations from one visit to the next were also negligible (**Table 5b**).

4. Discussion

The objectives of this analysis were (1) to describe and quantify the natural trajectory of questionnaire scores over time, based on initial scores, with a subgoal of determining whether the observed fluctuations were compatible with RTM, and (2) to evaluate and compare the stability of each questionnaire over time.

Our results show that the CLBP population had relatively large variations in outcome measures over time and that this effect varied across subgroups and across questionnaires. It bears repeating that these fluctuations were *observed in the absence of any experimental intervention*. Participants with high initial scores were overwhelmingly likely to show a decrease in score at the subsequent measurement, while participants with normal or extremely low scores were relatively more stable. In terms of questionnaires, the PCS showed the most variation in scores over time; both subscales of the MPQ and both subscales of the BPI as well as the PDs also showed meaningful variations. The most stable questionnaire overall was the STAI-T, followed by the CSI and POQ. Healthy controls, in contrast, showed very little variability. In this group, average deltas and average *absolute* deltas were similar - and very small - across all subgroups and questionnaires.

These observed fluctuations cannot be solely attributable to RTM. In the "extremely high" subgroup, the general decrease in score from one visit to the next is compatible with RTM. However, in the "normal" subgroup, the uneven split between score increases and decreases (skewed towards decreases, ie, improvements) is not compatible with RTM, which would predict roughly similar increases and decreases. Moreover, in the "extremely low" subgroup, the overall stability is also incompatible with RTM, which would predict a general increase in score.

Together, these results suggest the presence of a global effect responsible for a generalized decrease in scores (ie, clinical improvement) over time. There are 2 possible explanations. First, this effect could be the result of the attention and care received by the patients as part of their participation in the study; as such propose calling this effect "Effect of Care." Indeed, even if a participant is fully aware that they are not receiving any treatment (which therefore rules out a placebo effect, in its textbook definition[29]), simply having the chance to talk about their pain with understanding, thoughtful, and competent-looking research staff could contribute to improving their symptoms. In addition, the "seriousness" afforded by the inclusion of brain and lumbar MRI—a notably well-regarded and imposing modality—likely further increased the potency of Effect of Care in our study. It should be noted that this proposed Effect of Care is conceptually different than the Hawthorne effect, wherein participants of a study change their behavior when they know that they are being observed.

Second, this global trend towards improvement could also be the result of a biased sample selection, wherein patients are more likely to volunteer for a study when their symptoms are worse than usual and less likely to volunteer when their symptoms are better than usual. Because RTM dictates that patients in a bad phase are likely to improve over time and patients in a good phase are likely to worsen, a sample biased towards patients in a bad phase would also yield RTM biased towards improvements.

4.1. Similarities and differences with test–retest

At first glance, this study presents superficial similarities with the well-known test–retest; however, it is important to point out that while test–retest studies generate a single overall score for a questionnaire, we conducted an analysis by subgroup. This allowed us to isolate and quantify differing degrees of variability *within* a questionnaire and to highlight directional trends *depending on the initial score*, providing more nuanced and precise results than a single overall score.

4.2. Biases and limitations

The most important limitation in this study is obviously the small sample size, especially as we further divided our sample into 3 subgroups. However, having 3 assessment time points allowed us to conduct 2 separate analyses (V1 to V2 and V2 to V3) which showed similar results. Furthermore, the objective of this study was not to precisely quantify specific effects, but rather to explore our data set and identify general trends and effects. Finally, the fact that a similar pattern was found regardless of the classification threshold used (see supplementary materials, http://links.lww.com/PR9/A231) lends further credibility to our findings.

Another potentially objectionable point was the decision to use an arbitrary threshold for fluctuations that are considered "noise" (≤5/100) vs "clinically meaningful" (>10/100), as opposed to using the established Minimal Detectable Change (MDC) or Minimal Clinically Important Difference (MCID) of each instrument. Standardized thresholds were chosen to facilitate comparisons between questionnaires, which would otherwise have been counterintuitive at best. This decision was again consistent with our objectives, which were to identify overall trends and not to quantify phenomena with a high degree of precision.

Last but not least, RTM is based on the score distribution within each individual—ie, a large RTM should be expected when a person's initial score is extreme relative to their own distribution mean. Because we did not have access to each participant's score distribution, we had to approximate their distribution mean using the group average.

4.3. Relevance for clinical trials

It is difficult to determine with certainty whether the variation in scores observed in this study is a manifestation of RTM, biased sample/biased RTM, Effect of Care, or some other effect. However, regardless of the underlying cause, being able to quantify this variability for specific questionnaires and specific subgroups has important clinical implications. Indeed, it will allow future researchers conducting clinical trials to compare their observed variations against our results, so that they can isolate and better estimate the "true" effect of their intervention. For example, a researcher might be thrilled to see a reduction of 10 points on the PCS following an experimental treatment. However, as shown in our study, such a decrease can easily be observed in the absence of any treatment.

Our results showed that CLBP patients with more severe symptoms at baseline will tend to show improvement at the subsequent measurement, even in the absence of an intervention—which could lead researchers to overestimate the effect of their intervention. Moreover, our results could suggest the presence of an Effect of Care, wherein patients generally show an improvement in symptoms simply by being part of a study.

Our results also provide a preliminary quantification of the variability in scores observed over time, in the absence of an intervention, in a CLBP population. This variability depends on initial score and is different across questionnaires. Our results can therefore be used to guide interpretation of results obtained in clinical trials.

Disclosures

The authors have no conflict of interest to declare.

Acknowledgments

The authors thank all participants for their collaboration in this study.

M.S. and A.C.L. are supported by a PhD scholarship from the CIHR. P.T. is supported by FRQS J1 salary award and Arthritis Society star career development award. G.L. is supported by FRQS senior salary award. P.T. has received financial grant for this study from the Quebec Bio-Imaging Network (QBIN) and Quebec Pain Research Network (QPRN).

Preprint on MedRxiv: https://doi.org/10.1101/2023.05.05.23289575 and a poster of this work was presented at the 13th congress of the European Pain Federation (EFIC)—poster.

Data availability: Our data are available from the corresponding author.

Supplemental digital content

Supplemental digital content associated with this article can be found online at http://links.lww.com/PR9/A231.

Article history:
Received 12 September 2023
Received in revised form 22 January 2024
Accepted 20 February 2024
Available online 26 April 2024

References

[1] Anagnostis C, Gatchel RJ, Mayer TG. The pain disability questionnaire: a new psychometrically sound measure for chronic musculoskeletal disorders. Spine 2004;29:2290–303.

[2] Ashar YK, Gordon A, Schubiner H, Uipi C, Knight K, Anderson Z, Carlisle J, Polisky L, Geuter S, Flood TF, Kragel PA, Dimidjian S, Lumley MA, Wager TD. Effect of pain reprocessing therapy vs placebo and usual care for patients with chronic back pain: a randomized clinical trial. JAMA Psychiatry 2022;79:13–23.
[3] Baird A, Sheffield D. The relationship between pain beliefs and physical and mental health outcome measures in chronic low back pain: direct and indirect effects. Healthcare 2016;4:58.
[4] Boureau F, Luu M, Doubrère JF. Comparative study of the validity of four French McGill Pain Questionnaire (MPQ) versions. PAIN 1992;50:59–65.
[5] Clark ME, Gironda RJ, Young RW. Development and validation of the pain outcomes questionnaire-VA. J Rehabil Res Dev 2003;40:381–95.
[6] Cleeland CS, Ryan KM. Pain assessment: global use of the brief pain inventory. Ann Acad Med Singapore 1994;23:129–38.
[7] Darnall BD, Roy A, Chen AL, Ziadni MS, Keane RT, You DS, Slater K, Poupore-King H, Mackey I, Kao M-C, Cook KF, Lorig K, Zhang D, Hong J, Tian L, Mackey SC. Comparison of a single-session pain management skills intervention with a single-session health education intervention and 8 sessions of cognitive behavioral therapy in adults with chronic low back pain: a randomized clinical trial. JAMA Netw Open 2021;4:e2113401.
[8] Davis CE. The effect of regression to the mean in epidemiologic and clinical studies. Am J Epidemiol 1976;104:493–8.
[9] Dworkin RH, Turk DC, Peirce-Sandner S, Burke LB, Farrar JT, Gilron I, Jensen MP, Katz NP, Raja SN, Rappaport BA, Rowbotham MC, Backonja M-M, Baron R, Bellamy N, Bhagwagar Z, Costello A, Cowan P, Fang WC, Hertz S, Jay GW, Junor R, Kerns RD, Kerwin R, Kopecky EA, Lissin D, Malamut R, Markman JD, McDermott MP, Munera C, Porter L, Rauschkolb C, Rice ASC, Sampaio C, Skljarevski V, Sommerville K, Stacey BR, Steigerwald I, Tobias J, Trentacosti AM, Wasan AD, Wells GA, Williams J, Witter J, Ziegler D. Considerations for improving assay sensitivity in chronic pain clinical trials: IMMPACT recommendations. PAIN 2012;153:1148–58.
[10] French D, Noël M, Vigneau F, French J, Cyr C, Evans R. L'Échelle de dramatisation face à la douleur PCS-CF: adaptation canadienne en langue française de l'échelle «Pain Catastrophizing Scale». Can J Behav Sci 2005;37:181–92.
[11] Freynhagen R, Baron R, Gockel U, Tölle TR. Pain DETECT: a new screening questionnaire to identify neuropathic components in patients with back pain. Curr Med Res Opin 2006;22:1911–20.
[12] Gauthier J, Bouchard S. Adaptation canadienne-française de la forme révisée du State–Trait Anxiety Inventory de Spielberger. Can J Behav Sci 1993;25:559–78.
[13] Gauthier N, Thibault P, Adams H, Sullivan MJ. Validation of a French-Canadian version of the pain disability Index. Pain Res Manag 2008;13:327–33.
[14] Gillving M, Demant D, Lund K, Holbech JV, Otto M, Vase L, Jensen TS, Bach FW, Finnerup NB, Sindrup SH. Factors with impact on magnitude of the placebo response in randomized, controlled, cross-over trials in peripheral neuropathic pain. PAIN 2020;161:2731–6.
[15] Haack M, Simpson N, Sethna N, Kaur S, Mullington J. Sleep deficiency and chronic pain: potential underlying mechanisms and clinical implications. Neuropsychopharmacology 2020;45:205–16.
[16] Henderson LA, Di Pietro F, Youssef AM, Lee S, Tam S, Akhter R, Mills EP, Murray GM, Peck CC, Macey PM. Effect of expectation on pain processing: a psychophysics and functional MRI analysis. Front Neurosci 2020;14:6.
[17] Johnson WD, George VT. Effect of regression to the mean in the presence of within-subject variability. Stat Med 1991;10:1295–302.
[18] Joshi S, Nuckols T, Escarce J, Huckfeldt P, Popescu I, Sood N. Regression to the mean in the medicare hospital readmissions reduction program. JAMA Intern Med 2019;179:1167–73.
[19] Kamerman PR, Vollert J. Greater baseline pain inclusion criteria in clinical trials increase regression to the mean effect: a modelling study. PAIN 2022;163:e748–58.
[20] Mayer TG, Neblett R, Cohen H, Howard KJ, Choi YH, Williams MJ, Perez Y, Gatchel RJ. The development and psychometric validation of the central sensitization inventory. Pain Pract 2012;12:276–85.
[21] McWilliams LA, Cox BJ, Enns MW. Mood and anxiety disorders associated with chronic pain: an examination in a nationally representative sample. PAIN 2003;106:127–33.
[22] Melzack R. The short-form McGill pain questionnaire. PAIN 1987;30: 191–7.
[23] Morton V, Torgerson DJ. Effect of regression to the mean on decision making in health care. BMJ 2003;326:1083–4.
[24] Pitance L, Piraux E, Lannoy B, Meeus M, Berquin A, Eeckhout C, Dethier V, Robertson J, Meeus M, Roussel N. Cross cultural adaptation, reliability and validity of the French version of the central sensitization inventory. Man Ther 2016;25:e83–4.
[25] Sandhu HK, Booth K, Furlan AD, Shaw J, Carnes D, Taylor SJC, Abraham C, Alleyne S, Balasubramanian S, Betteley L, Haywood KL, Iglesias-Urrutia CP, Krishnan S, Lall R, Manca A, Mistry D, Newton S, Noyes J, Nichols V, Padfield E, Rahman A, Seers K, Tang NKY, Tysall C, Eldabe S, Underwood M. Reducing opioid use for chronic pain with a group-based intervention: a randomized clinical trial. JAMA 2023;329:1745–56.
[26] Smith SM, Dworkin RH, Turk DC, McDermott M, Eccleston C, Farrar JT, Rowbotham MC, Bhagwagar Z, Burke LB, Cowan P, Ellenberg SS, Evans SR, Freeman RL, Garrison LP, Iyengar S, Jadad A, Jensen MP, Junor R, Kamp C, Katz NP, Kesslak JP, Kopecky EA, Lissin D, Markman JD, Mease PJ, O'Connor AB, Patel KV, Raja SN, Sampaio C, Schoenfeld D, Singh J, Steigerwald I, Strand V, Tive LA, Tobias J, Wasan AD, Wilson HD. Interpretation of chronic pain clinical trial outcomes: IMMPACT recommended considerations. PAIN 2020;161:2446–61.
[27] Spielberger CD, Gorsuch RL, Lushene R, Vagg PR, Jacobs GA. Manual for the state-trait anxiety inventory. Palo Alto, CA: Consulting Psychologists Press, 1983.
[28] Sullivan MJL, Bishop SR, Pivik J. The pain catastrophizing scale: development and validation. Psychol Assess 1995;7:524–32.
[29] Turner JA, Deyo RA, Loeser JD, Von Korff M, Fordyce WE. The importance of placebo effects in pain treatment and research. JAMA 1994;271:1609–14.
[30] Whitney CW, Von Korff M. Regression to the mean in treated versus untreated chronic pain. PAIN 1992;50:281–5.
[31] Wideman TH, Edwards RR, Walton DM, Martel MO, Hudon A, Seminowicz DA. The multimodal assessment model of pain: a novel framework for further integrating the subjective pain experience within research and practice. Clin J Pain 2019;35:212–21.
[32] Yu R, Chen L. The need to control for regression to the mean in social psychology studies. Front Psychol 2014;5:1574.