Patients Undergoing Shoulder Stabilization Procedures Do Not Accurately Recall Their Preoperative Symptoms at Short- to Midterm Follow-up

Danielle Hope,*[†] MBBS, Jacqui French,[‡] BPhys, Tania Pizzari,[§] PhD,
Greg Hoy,[‡] FRACS (Orth), and Shane Barwood,[‡] FRACS (Orth)
Investigation performed at Melbourne Orthopaedic Group, Melbourne, Australia

Background: A patient's ability to recall symptoms is poor in some elderly populations, but we considered that the recall of younger patients may be more accurate. The accuracy of recall in younger patients after surgery has not been reported to date.

Purpose: To assess younger patients' abilities to recall their preoperative symptoms after having undergone shoulder stabilization surgery. We used 2 disease-specific, patient-reported outcome measures (PROMs)—the Western Ontario Shoulder Instability Index (WOSI) and the Melbourne Instability Shoulder Score (MISS)—at a period of up to 2 years postoperatively.

Study Design: Cohort study (diagnosis); Level of evidence, 2.

Methods: Participants (N = 119) were stratified into 2 groups: early recall (at 6-8 months postoperatively; n = 58) and late recall (at 9-24 months postoperatively; n = 61). All patients completed the PROMs with instructions to recall preoperative function. The mean and absolute differences between the preoperative scores and recalled scores for each PROM were compared using paired t tests. Correlations between the actual and recalled scores of the subsections for each PROM were calculated using an intraclass correlation coefficient (ICC). The number of individuals who recalled within the minimal detectable change (MDC) of each PROM was calculated.

Results: Comparison between the means of the actual and recalled preoperative scores for both groups did not demonstrate significant differences (early recall differences, MISS 1.05 and WOSI –38.64; late recall differences, MISS –0.25 and WOSI –24.02). Evaluation of the absolute difference, however, revealed a significant difference between actual and recalled scores for both the late and early groups (early recall absolute differences, MISS 12.26 and WOSI 216.71; late recall absolute differences, MISS 12.84 and WOSI 290.08). Average absolute differences were above the MDC scores of both PROMs at both time points. Subsections of each PROM demonstrated weak to moderate correlations between actual and recalled scores (ICC range, 0.17-0.61). Total scores for the PROMs reached moderate agreement between actual and recalled scores.

Conclusion: Individual recall after shoulder instability surgery was not accurate. However, the mean recalled PROM scores of each group were not significantly different from the actual scores collected preoperatively, and recall did not deteriorate significantly over 2 years. This suggests that recall of the individual, even in this younger group, cannot be considered accurate for research purposes.

Keywords: patient recall; patient-reported outcome measure; shoulder instability; surgery; Western Ontario Shoulder Instability Score

Patient recall is used commonly to assess the outcome of treatment in the clinical setting. In the research setting, it also appears common to rely on recall to obtain patient-reported outcome measures (PROMs) after treatment.[13,26] The accuracy of patient recall, however, has been questioned, and the relevance and accuracy of research using this technique are controversial.[11]

When evaluating intervention efficacy from a research perspective, minimizing the reliance on patient recall and administering pre- and postintervention outcome measures is considered the gold standard.[2] Such methods require an increase in time, planning, and expense for the collection of data at different time points and may not be possible with trauma. Modifications of existing PROMs and the introduction of new measures might also affect consistent data collection at each assessment time point after an intervention, particularly when collected for registry data over a long period of time.

The ability to recall accurately is influenced by multiple factors,[7,14,20] and it is not surprising that the results of studies regarding the accuracy of patient recall of preoperative pain and function have reported variable results.[8,13,15,19,23,27] Several authors have reported that recall was inaccurate in patients who had undergone arthroplasty,[13,15,19] although early recall between 6 and 12 weeks seemed to be more accurate in this group.[8,17,27] These patients tend to be elderly patients with chronic conditions, and their recall ability appears to deteriorate with time. Recall has been reported to be accurate for up to 2 years in patients with nontraumatic hand and elbow conditions using the shortened version of the Disabilities of the Arm, Shoulder and Hand questionnaire, yet Lowe et al[15] reported that recall assessed using the American Shoulder and Elbow Surgeons scoring system in patients undergoing shoulder arthroplasty was accurate only at the 6-week postoperative assessment. They reported that function could be accurately recalled for 12 months, but pain was recalled for only 6 weeks. General recall is reported to be better in the young compared with the elderly,[5,21,22] although the specific ability of patients in the younger age group, typically involved in the sporting community, to recall dysfunction from injury has not been reported to date.

Shoulder instability is a condition that commonly occurs in younger patients undertaking various sporting activities, and it does not typically cause chronic and debilitating disability apart from the discrete episodes when the shoulder dislocates. Validated and disease-specific PROMs are available, including the Western Ontario Shoulder Instability Index (WOSI)[10] and the Melbourne Instability Shoulder Score (MISS).[25] Surgical procedures for shoulder instability provide reliable outcomes for patients, and typically the improvement in PROM scores after treatment is clinically significant.[1,9,12] It is feasible that this patient group could demonstrate accurate recall considering their younger average age and the fact that they experience relatively infrequent but significant episodic events. It is also feasible that their longevity of recall could be better than that of older patient populations.

If it can be shown that younger patients have accurate recall of preoperative dysfunction, even after surgical recovery, then investigators who conduct studies of shoulder instability surgery may be able to collect preoperative PROMs retrospectively with a degree of confidence. This study therefore aimed to assess the ability of patients undergoing surgery for shoulder instability to recall their preoperative dysfunction using the WOSI and MISS in the postoperative period for up to 2 years. We hypothesized that this group might be able to accurately recall the nature of their preoperative dysfunction and that recall would more likely be accurate at the 6-month time point and would decline with time following the intervention.

METHODS

All patients who had undergone surgery for shoulder instability, including primary shoulder labral repair, capsular plication, or Latarjet procedure between September 2013 and May 2016 by a single high-volume, subspecialized shoulder surgeon (S.B.) were considered for entry into the study. The technical aspects of the procedures did not change during this period and consisted typically of a 3-portal arthroscopic soft tissue reconstruction in the lateral position using an average of 4 glenoid anchors and suture knots or an open Latarjet reconstruction using the Arthrex Latarjet system. All participants had previously completed the validated and disease-specific MISS and WOSI PROMs at their initial assessment prior to their surgery. The participants were not told that they would be required to recall this information.

The WOSI has 21 items, completed on a visual analog scale representing 4 domains—physical symptoms, sport/recreation/work, lifestyle, and emotions—and is scored out of a total of 2100. The MISS has 24 items completed on a Likert scale representing 4 domains—pain, instability, function, and occupation and sport—and is scored out of a total of 100. The minimum detectable change (MDC) has been reported to be 5.5 points for the MISS and 10% for the WOSI.[10,25]

Patients were excluded from the study if their shoulder instability surgery was identified as a revision procedure or if an episode of shoulder redislocation occurred after the procedure and during the time points of the study. Patients who failed to complete the PROMs were also excluded. All participants provided written informed consent, and the study was approved by The Avenue Human Research Ethics Committee.

The enrolled participants were stratified into 2 groups based on time since surgery: the early recall group (6-8 months) and the late recall group (9-24 months). Participants completed the recall PROM at only 1 of the 2 time intervals to avoid potential bias by repetition. Participants were instructed to complete the form according to their recollection of preoperative pain and function, not their current postoperative state. The participants in the early recall group completed the recall PROM unaided after their final clinical review, and those in the late recall group completed their recall PROM online after being contacted by one of the investigators (J.F.).

*Address correspondence to Danielle Hope, MBBS, Melbourne Orthopaedic Group, 33 The Avenue, Windsor, Victoria, Australia 3181 (email: jfrench@mog.com.au).

†Registrar, Australasian College of Sports and Exercise Physicians, Melbourne, Australia.

‡Department of Shoulder Surgery, Melbourne Orthopaedic Group, Melbourne, Australia.

§Department of Rehabilitation, Nutrition and Sport, Latrobe University, Melbourne, Australia.

Statistical Analysis

Data were assessed for normality and descriptive statistics. Means, standard deviations, and the standard error of measurement (SEM) were calculated for the actual and recalled WOSI and MISS scores using IBM SPSS (version 24). The absolute difference was calculated using the numerical value between the actual score and the recalled score. The absolute value between 2 numbers provides the magnitude of the difference between the numbers but discards information about direction. Calculation of the absolute mean difference minimizes the effect of averaging difference scores and provides an indication of the true difference, irrespective of direction. Comparison of these means was calculated through the use of paired t tests.

A P value less than .05 was considered statistically significant. Bland-Altman limits of agreement comparing the differences between the 2 measures (actual and recalled) against the averages of the 2 measures were calculated and plotted.[3]

The correlation between actual and recalled total scores and subsection scores for each PROM were calculated by use of an intraclass correlation coefficient (ICC) (3,1). ICC values less than 0.36 were interpreted as weak reliability, 0.36 to 0.67 moderate reliability, 0.68 to 0.90 good reliability, and greater than 0.90 very strong reliability.[24] The number of individuals who recalled within the MDC was also analyzed.

To determine whether recall changes over time, an independent-groups t test was used to compare the absolute difference for the MISS and WOSI between the early recall and late recall groups.

RESULTS

In total, 180 patients (58 in the early recall group and 122 in the late recall group) were eligible and were approached to participate in the study. All 58 eligible patients in the early group completed the recall PROM at the final surgical consultation. Of the 122 patients contacted by phone or email for entry into the late recall group, 61 patients (50%) were excluded because of missing data. Of these 61 patients, 46 did not respond to any communication, 4 returned incomplete PROMs, and 11 consented to participate but did not return any data. As a result, a total of 119 participants were included in the study (Table 1). The mean period of recall was 7 months (range, 6-8 months) in the early group and 20 months (range, 9-24 months) in the late group. The majority of the procedures performed were arthroscopic labral repair surgeries, but 4 open Latarjet procedures were included in the early group (Table 1).

Comparison between the means of the actual and recalled PROM scores in the early and late groups were not significantly different (Tables 2 and 3). However, a comparison of the mean absolute difference between the actual and recalled PROM scores showed significant differences in both groups ($P < .01$). The mean absolute differences were greater than the MDC scores of both questionnaires in both groups. No difference was found in absolute recall between

TABLE 1
Participant Demographics

	Early Recall Group (n = 58)	Late Recall Group (n = 61)
Female, n (%)	6 (11)	11 (18)
Male, n (%)	52 (89)	50 (82)
Mean age (range), y	23 (14-58)	26 (14-49)
Surgery type, n		
Anterior labral repair	39	41
Posterior labral repair	6	10
Anterior and posterior labral repair	9	10
Latarjet	4	0

TABLE 2
Data for Early Recall Group (6-8 Months Postsurgery)[a]

Preoperative Scores	Mean	SD	SEM	P Value
MISS				
Actual	48.81	14.95	1.96	
Recalled	47.76	17.56	2.31	
Recall difference	1.05	15.07	1.98	.60
Absolute difference	12.26	8.68	1.14	<.01
WOSI				
Actual	1250.28	372.23	48.88	
Recalled	1288.91	379.05	49.77	
Recall difference	−38.64	306.28	40.22	.34
Absolute difference	216.71	218.04	28.63	<.01

[a]MISS, Melbourne Instability Shoulder Score; WOSI, Western Ontario Shoulder Instability Index.

TABLE 3
Data for Late Recall Group (≥9 Months Postsurgery)[a]

Preoperative Scores	Mean	SD	SEM	P Value
MISS				
Actual	46.08	16.78	2.15	
Recalled	46.33	19.73	2.53	
Recall difference	−0.25	17.32	2.22	.91
Absolute difference	12.84	11.51	1.47	<.01
WOSI				
Actual	1241.00	400.01	51.22	
Recalled	1265.02	376.76	48.24	
Recall difference	−24.02	387.04	49.56	.63
Absolute difference	290.08	254.63	32.60	<.01

[a]MISS, Melbourne Instability Shoulder Score; WOSI, Western Ontario Shoulder Instability Index.

the early and late recall groups regarding the MISS ($P = .76$) or the WOSI ($P = .10$).

The ICCs of each of the subsections of the MISS and WOSI demonstrated weak to moderate agreements between the actual and recalled scores (Table 4). On the MISS, the consistency of recall was poorer for pain than for those in the other subsections in both the early and late groups. The recall of each subsection of the WOSI was

TABLE 4
Intraclass Correlation Coefficients for Recalled and
Actual Preoperative Scores for the Totals
and Subsections of the MISS and WOSI[a]

Totals and Subsections	Early Recall Group	Late Recall Group
MISS (n = 58)		
Total score	0.58	0.61
Pain	0.31	0.23
Instability	0.52	0.53
Function	0.47	0.52
Occupation and sport	0.51	0.44
WOSI (n = 58)		
Total score	0.67	0.61
Physical symptoms	0.20	0.48
Sport/recreation/work	0.51	0.45
Lifestyle	0.58	0.17
Emotions	0.27	0.51

[a]MISS, Melbourne Instability Shoulder Score; WOSI, Western Ontario Shoulder Instability Index.

variable, with no consistent pattern between time points. The ICC of the total WOSI and MISS scores reached moderate agreement between actual and recalled scores (Table 4).

A Bland-Altman plot analysis of the difference between recalled scores and the average of actual and recalled scores highlighted that recall did not appear to be influenced by preoperative clinical scores (Figure 1). In the early group, 63% of the recall scores were within the MDC of the WOSI and 26% were within the MDC of the MISS. In the late group, 51% of the recall scores were within the MDC of the WOSI and 31% were within the MDC of the MISS. The rate of individual participants who recalled within the MDC on both the MISS and the WOSI was 17% in the early group and 18% in the late group (Table 5).

DISCUSSION

This study used 2 PROMs to evaluate the accuracy of younger patients' recollection of preoperative dysfunction at different timepoints after shoulder instability surgery .

Our findings show that the mean recalled PROM scores of each group were not significantly different from the actual mean group scores collected preoperatively. This observation suggests that the use of retrospectively collected WOSI and MISS scores in this population in order to establish a group mean score is reasonable for up to 2 years postoperatively. This study also demonstrated that recall accuracy at the individual level was poor and that the accuracy of the results identified at the group level was attributable to equivalent numbers of errors in recall above and below the group mean. This finding suggests, therefore, that the recall ability of the individual even in this younger group cannot be considered accurate for research purposes. This significant variation in the absolute mean difference (irrespective of direction) highlights the limitation of comparing means between testing times and the potential for incorrect interpretation of results.

Previous studies that have accounted for the effect of averaging difference scores have also identified this phenomenon.[19,27] Where studies have evaluated only the mean difference in a cohort, the accuracy of recall appears acceptable.[17,23]

We did not find a difference in recall ability between the early and late groups. This differs from findings of several other studies[4,8,13,15,19]; however, those studies were completed in elderly patients undergoing arthroplasty. Lowe et al[15] studied 169 patients who had shoulder arthroplasty and found that although the patients appeared to recall function for up to 12 months, they could not recall their pain accurately for more than 6 weeks. No previously reported study included a population with a mean age similar to that of our study population. Stepan et al[23] evaluated a patient group with a mean age of 55 years and demonstrated that mean group recall was maintained over a 2-year period, similar to our finding in a patient group with an average age in the 20s.

Several studies have highlighted that the recall of pain seems to be exaggerated over time, and this has been considered a consequence of response shift.[13,14,16,18] Our findings suggest that the correlation of recalled pain was poor; however, we did not demonstrate a difference between actual and recalled pain scores. The ICCs for the subsections of the MISS and WOSI were either poor or moderate and appeared to change with time in some subsections quite differently from others. We were not able to identify the reasons for this in the current study. Since the ICCs of subsections were poor or moderate, analysis of recall of each subsection was not considered to be reliable. Further, the ICCs of the total MISS and WOSI scores are not clinically acceptable considering the significant differences found with univariate testing; only moderate ICCs were found for total WOSI scores in the early and late groups, and a wide spread of scores was seen on the Bland-Altman graphs.

Differences between recall within the MDC of each PROM were noted. The 95% limits of agreement in the Bland-Altman plot are wide, and it is not clinically acceptable that the scores lie within these. The MDC is a more relevant method of interpreting the accuracy of recall considering that change outside these limits can be considered true change in a patient's presentation.[6] The MDC was calculated individually for each PROM in its initial validation study, and although the absolute percentage of the MDC for each score may vary, each should identify a relevant clinical change. A noted difference is that the MDC of the WOSI (10%) appears wider than the MDC of the MISS (5.5%), and it was observed in this study that more participants were able to recall within the wider limits of the MDC of the WOSI. The most likely explanation is that the clinical change that each PROM considers significant is actually different despite both PROMs being regarded as valid for shoulder instability.[10,25] The differences in recall within the MDC of each PROM did not deteriorate with time.

Our study is subject to certain limitations. First, recall PROM scores were collected from only half of the eligible patients in the late group, which potentially introduces bias. We noted in this younger cohort that follow-up in the
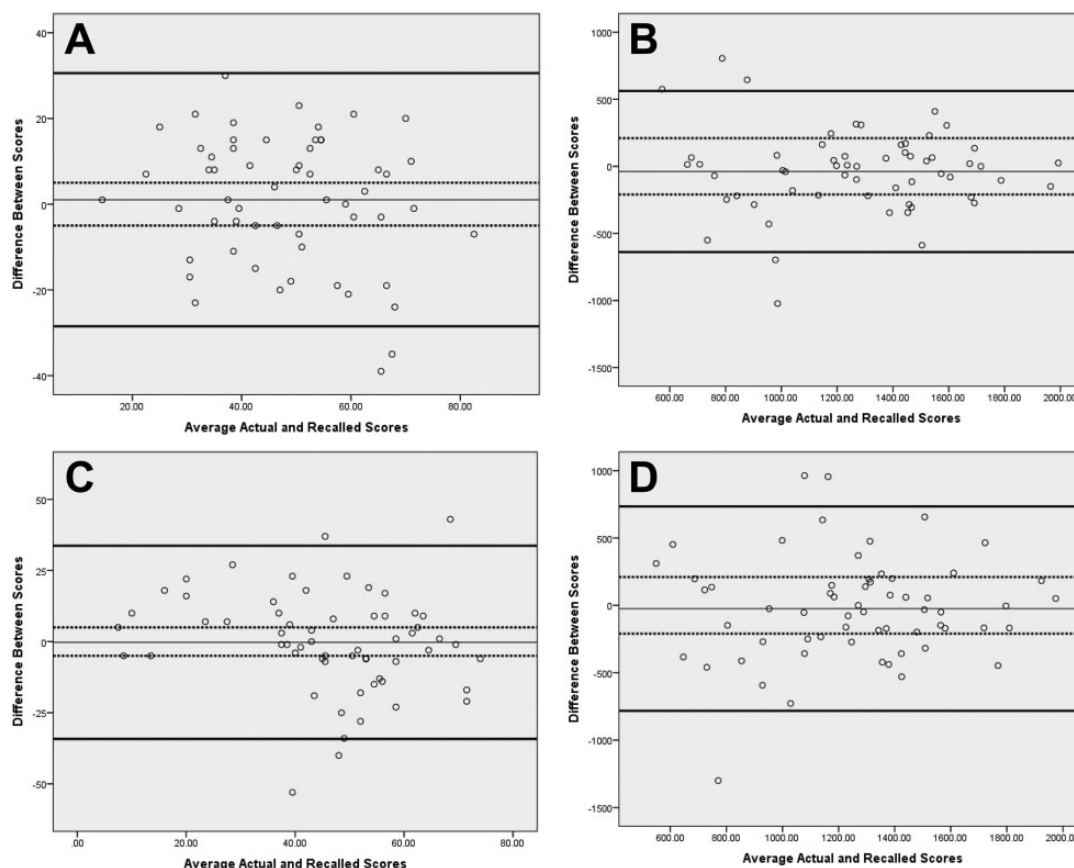
Figure 1. Bland-Altman plots showing the difference between recalled scores and the average of actual and recalled scores. Solid lines represent 95% limits of agreement; dotted lines represent the minimal detectable change score for the questionnaire. (A) Melbourne Instability Shoulder Scores (MISS) of the early recall group. (B) Western Ontario Shoulder Instability Index (WOSI) scores of the early recall group. (C) MISS of the late recall group. (D) WOSI of the late recall group.

TABLE 5
Number of Patients Who Scored Within the MDC[a]

	Within the MDC
Early recall group (n = 58)	
MISS	15 (26)
WOSI	37 (63)
Both	10 (17)
Late recall group (n = 61)	
MISS	19 (31)
WOSI	31 (51)
Both	11 (18)

[a]Values are expressed as n (%). MDC, minimum detectable change; MISS, Melbourne Instability Shoulder Score; WOSI, Western Ontario Shoulder Instability Index.

early group was not difficult, as these patients reliably returned for postoperative appointments and the recall PROMs were completed at the conclusion of this appointment. However, we also noted in the late group that a high proportion of eligible participants were difficult to reach for follow-up owing to factors such as changes in place of residence and contact details. We also had multiple patients who agreed to participate but then failed to return their recall PROM despite multiple attempts to contact them.

Second, an increase in sample size may identify the effect of time since surgery on recall, but the sample size was adequate to show differences between actual and recalled scores. We specifically did not increase our sample size by asking participants to complete the recall PROMs in both the early and the late groups in order to reduce the influence of recall bias, and several similar studies have had equivalent sample sizes.[8,16,19,23] A larger cohort may have permitted the evaluation of the influence of baseline characteristics on recall accuracy, such as age, sex, sport, or number of instability episodes.

Although we identified poor individual patient recall for 2 validated and specific shoulder instability questionnaires, the reliability of other shoulder PROMs in this patient cohort is unknown.

Finally, we were not able to evaluate recall compared with actual outcome PROM score because we only asked the participants to complete recall PROM scores at the specified timepoint. Again, we did this to limit the number of times each participant was asked to complete the PROMs to reduce the time constraints on each participant and reduce potential bias.

CONCLUSION

The results of this study did not support our hypothesis that individual patients with shoulder instability have accurate recall between 6 months and 2 years after surgery. The study does suggest that the mean group recall in this population of younger patients is accurate and does not deteriorate over 2 years. This finding has implications for the use of retrospective data in assessment of shoulder instability after treatment and suggests that we cannot assume that individual patient recall is accurate enough in the research environment.

REFERENCES

1. Aboalata M, Plath JE, Seppel G, Juretzko J, Vogt S, Imhoff AB. Results of arthroscopic Bankart repair for anterior-inferior shoulder instability at 13-year follow-up. *Am J Sports Med*. 2016;45(4): 782-787.
2. Ahern S, Ruseckaite R, Ackerman IN. Collecting patient-reported outcome measures. *Intern Med J*. 2017;47(12):1454-1457.
3. Bland J, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986; 1(8476):307-310.
4. Bryant D, Norman G, Stratford P, Marx RG, Walter SD, Guyatt G. Patients undergoing knee surgery provided accurate ratings of preoperative quality of life and function 2 weeks after surgery. *J Clin Epidemiol*. 2006;59(9):984-993.
5. Danckert SL, Craik FIM. Does aging affect recall more than recognition memory? *Psychol Aging*. 2013;28(4):902-909.
6. de Vet HC, Terwee CB, Ostelo RW, Beckerman H, Knol DL, Bouter LM. Minimal changes in health status questionnaires: distinction between minimally detectable change and minimally important change. *Health Qual Life Outcomes*. 2006;4(1):54.
7. Howard GS, Dailey PR. Response-shift bias: a source of contamination of self-report measures. *J Appl Psychol*. 1979;64(2):144-150.
8. Howell J, Xu M, Duncan CP, Masri BA, Garbuz DS. A comparison between patient recall and concurrent measurement of preoperative quality of life outcome in total hip arthroplasty. *J Arthroplasty*. 2008; 23(6):843-849.
9. Imhoff AB, Ansah P, Tischer T, et al. Arthroscopic repair of anterior-inferior glenohumeral instability using a portal at the 5:30-o'clock position: analysis of the effects of age, fixation method, and concomitant shoulder injury on surgical outcomes. *Am J Sports Med*. 2010; 38(9):1795-1803.
10. Kirkley A, Griffin S, McLintock H, Ng L. The development and evaluation of a disease-specific quality of life measurement tool for shoulder instability. The Western Ontario Shoulder Instability Index (WOSI). *Am J Sports Med*. 1998;26:764-772.
11. Kwong E, Black N. Retrospectively patient-reported pre-event health status showed strong association and agreement with contemporaneous reports. *J Clin Epidemiol*. 2017;81:22-32.
12. Lafosse L, Lejeune E, Bouchard A, Kakuda C, Gobezie R, Kochhar T. The arthroscopic Latarjet procedure for the treatment of anterior shoulder instability. *Arthroscopy*. 2007;23(11):1242.
13. Lingard EA, Wright EA, Sledge CB; the Kinemax Outcomes Group. Pitfalls of using patient recall to derive preoperative status in outcome studies of total knee arthroplasty. *J Bone Joint Surg Am*. 2001;83(8): 1149-1156.
14. Linton SJ, Melin L. The accuracy of remembering chronic pain. *Pain*. 1982;13(3):281-285.
15. Lowe JT, Li X, Fasulo SM, Testa EJ, Jawa A. Patients recall worse preoperative pain after shoulder arthroplasty than originally reported: a study of recall accuracy using the American Shoulder and Elbow Surgeons score. *J Shoulder Elbow Surg*. 2017;26(3):506-511.
16. Mancuso CA, Charlson ME. Does recollection error threaten the validity of cross-sectional studies of effectiveness? *Med Care*. 1995;33(4): AS77-AS88.
17. Marsh J, Bryant D, MacDonald SJ. Older patients can accurately recall their preoperative health status six weeks following total hip arthroplasty. *J Bone Joint Surg Am*. 2009;91(12):2827-2837.
18. McPhail S, Haines T. Response shift, recall bias and their effect on measuring change in health-related quality of life amongst older hospital patients. *Health Qual Life Outcomes*. 2010;8:65.
19. Murphy MT, Vardi R, Journeaux SF, Whitehouse SL. A patient's recollection of pre-operative status is not accurate one year after arthroplasty of the hip or knee. *Bone Joint J*. 2015;97(8):1070-1075.
20. Schmier JK, Halpern MT. Patient recall and recall bias of health state and health status. *Expert Rev Pharmacoecon Outcomes Res*. 2004; 4(2):159-163.
21. Small SA. Age-related memory decline: current concepts and future directions. *Arch Neurol*. 2001;58(3):360-364.
22. Small SA, Stern Y, Tang M, Mayeux R. Selective decline in memory function among healthy elderly. *Neurology*. 1999;52(7):1392.
23. Stepan JG, London DA, Boyer MI, Calfee RP. Accuracy of patient recall of hand and elbow disability on the QuickDASH questionnaire over a two-year period. *J Bone Joint Surg Am*. 2013;95(22): e176.
24. Taylor R. Interpretation of the correlation coefficient: a basic review. *J Diagn Med Sonogr*. 1990;6(1):35-39.
25. Watson L, Story I, Dalziel R, Hoy G, Shimmin A, Woods D. A new clinical outcome measure of glenohumeral joint instability: the MISS questionnaire. *J Shoulder Elbow Surg*. 2005;14(1):22-30.
26. Wilson J, Baker P, Rangan A. Is retrospective application of the Oxford Shoulder Score valid? *J Shoulder Elbow Surg*. 2009;18(4): 577-580.
27. Yeoman TFM, Clement ND, Macdonald D, Moran M. Recall of preoperative Oxford Hip and Knee Scores one year after arthroplasty is an alternative and reliable technique when used for a cohort of patients. *Bone Joint Res*. 2018;7(5):351-356.