BJR

■ **HIP AND KNEE**

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

T. F. M. Yeoman,
N. D. Clement,
D. Macdonald,
M. Moran

*Department of
Orthopaedics
and Trauma,
Royal Infirmary of
Edinburgh, Edinburgh,
United Kingdom*

Objectives
The primary aim of this study was to assess the reproducibility of the recalled preoperative Oxford Hip Score (OHS) and Oxford Knee Score (OKS) one year following arthroplasty for a cohort of patients. The secondary aim was to assess the reliability of a patient's recollection of their own preoperative OHS and OKS one year following surgery.

Methods
A total of 335 patients (mean age 72.5; 22 to 92; 53.7% female) undergoing total hip arthroplasty (n = 178) and total knee arthroplasty (n = 157) were prospectively assessed. Patients undergoing hip and knee arthroplasty completed an OHS or OKS, respectively, preoperatively and were asked to recall their preoperative condition while completing the same score one year after surgery.

Results
A mean difference of 0.04 points (95% confidence intervals (CI) -15.64 to 15.72, p = 0.97) between the actual and the recalled OHS was observed. The mean difference in the OKS was 1.59 points (95% CI -11.57 to 14.75, p = 0.10). There was excellent reliability for the 'average measures' intra-class correlation for both the OHS (r = 0.802) and the OKS (r = 0.772). However, this reliability was diminished for the individuals OHS (r = 0.670) and OKS (r = 0.629) using single measures intra-class correlation. Bland–Altman plots demonstrated wide variation in the individual patient's ability to recall their preoperative score (95% CI ± 16 for OHS, 95% CI ± 13 for OKS).

Conclusion
Prospective preoperative collection of OHS and OKS remains the benchmark. Using recalled scores one year following hip and knee arthroplasty is an alternative when used to assess a cohort of patients. However, the recall of an individual patient's preoperative score should not be relied upon due to the diminished reliability and wide CI.

■ T. F. M. Yeoman, MRCS, MSc, Specialist Trainee, Department of Orthopaedics and Trauma, Royal Infirmary of Edinburgh, Edinburgh, UK.

■ N. D. Clement, PhD, FRCSEd (Tr&Orth), Consultant Orthopaedic Surgeon, Department of Orthopaedics and Trauma, Royal Infirmary of Edinburgh, EdinburghA, UK.

■ D. Macdonald, BA(Hons), Database and Clinical Research Manager, Department of Orthopaedics and Trauma, University of Edinburgh, Edinburgh, UK.

■ M. Moran, FRCSEd (Tr&Orth), Consultant Orthopaedic Surgeon, Department of Orthopaedics and Trauma, Royal Infirmary of Edinburgh, Edinburgh EH16 4SA, UK.

Correspondence should be sent to T. F. M. Yeoman; email: tyeoman@nhs.net

Article focus
■ Can patients accurately recall preoperative status at 1 year after their total knee or total hip arthroplasty?

Key messages
■ Recalled Oxford joint scores one year following hip and knee arthroplasty are an alternative when used to assess a cohort of patients.

Strengths and limitations
■ A large number of patients from a single unit were included, using reliable and ubiquitous Oxford joint scores that are widely adopted in arthroplasty research and national arthroplasty databases.

■ We did not assess patient recall at different intervals, or the variation in recall according to their demographics (i.e., age and gender), which may influence their answers.

Introduction

Patient-reported outcome measures (PROMs) are commonly used to assess the outcome of total hip and knee arthroplasty (THA, TKA).[1,2] The Oxford Hip Score (OHS) and the Oxford Knee Score (OKS) are two such PROMs, which are validated, reliable, and well-established assessment tools.[3-8] These are routinely collected as part of national joint registry outcome assessments and are frequently used to assess the outcome of cohort studies. Assessment of any changes in these scores after surgery will indicate as to whether a minimally import change/difference has been achieved.[6,9-13] A limitation of these tools is that the preoperative score is often not available due to logistic reasons or when a retrospective assessment has been performed, and hence a change in score cannot be calculated from the postoperative score in isolation. Such patients are then often excluded from further follow-up and the absence of their preoperative data weakens the confidence of the data.[13] A retrospective collection of a preoperative OHS or OKS would enable the change in score to be calculated and allow inclusion into the study cohort, provided this recalled score was reliable.

There are conflicting conclusions regarding the reliability of the recalled preoperative functional status, as assessed by PROMs, after total joint arthroplasty. Lingard et al[14] demonstrated only moderate agreement for the Western Ontario and McMaster University Osteoarthritis Index (WOMAC) score after TKA. In contrast Howell et al[15] found the recalled WOMAC score, in addition to the OHS, to be highly reliable after THA. More recently, Murphy et al[16] concluded that a "patient's recollection of pre-operative status is not accurate one year after arthroplasty of the hip or knee" when using the OHS and OKS. They illustrated that the difference between the recollected and actual preoperative scores for each patient was five points, and that this method was inconsistent and should not be used. However, these scores demonstrate variation in the reproducibility when an individual patient is assessed. The original designers of these scores showed that the reproducibility according to 95% confidence levels varied by ± 7 points for the OHS and ± 6 points for the OKS.[3-6] Hence, the variation demonstrated by Murphy et al[16] is expected and is an intrinsic property of the scores. The overall mean score for a cohort of patients should be used to assess the reproducibility of the recalled OHS and OKS, which was used by the original designers.[3,4]

The primary aim of this study was to assess the reproducibility of the recalled preoperative OHS and OKS one year following arthroplasty surgery for a cohort of patients. The secondary aim was to assess the reliability of a patient's recollection of their own preoperative OHS and OKS one year following arthroplasty surgery. The null hypothesis was that recall of the preoperative OHS and OKS at one year, for a cohort and individual patient, would have poor agreement with the actual preoperative score.

Patients and Methods

Ethical approval was obtained for this study from the regional ethics committee.

This prospective study included a consecutive group of patients that underwent a primary or revision THA or TKA over a six-month period between October 2012 and March 2013 at the study centre. Those included were under the care of one of 14 consultant arthroplasty surgeons (MM). Patients who had other surgery between the two data collection points, a history of hip or knee prosthetic infection, periprosthetic fracture, or cognitive impairment were excluded.

The study identified and included 335 of 410 patients in this time-period (81%): 178 primary or revision THA and 157 primary or revision TKA. All patients completed the relevant Oxford score preoperatively at the pre-assessment clinic. Patients were then asked to recall their preoperative condition while completing a further Oxford score one year after surgery. To increase participation, patients were followed up with a single telephone call to encourage them to complete and return the written postal questionnaire. The Oxford questionnaires have the advantage of being short and reliable while being practical, reproducible, valid, and sensitive to clinically important change when compared with other validated scores.[3,4] The OHS and OKS are 12-item questionnaires designed for patients undergoing joint arthroplasties to capture their outcomes.[3,4,17] Each of the 12 questions are assessed on a Likert scale with values from 0 to 4: a summative score is then calculated, where 48 is the best possible score (less symptomatic) and zero is the worst possible score (most symptomatic).[3,4] The minimal clinically important difference for both the Oxford hip and knee scores is expected to be between three and five points.[6,17,18]

Two implants were used for TKA: the cemented Triathlon (Stryker, Newbury, United Kingdom) and the cemented PFC Sigma (DePuy, Johnson & Johnson Professional Inc., Raynham, Massachusetts). The common implants for THA were a cemented Exeter femoral component (Stryker) with a cemented Contemporary polyethylene acetabular component (Stryker). Postoperatively all patients had a standardized rehabilitation protocol, with active mobilization on day one postoperatively. Patients were reviewed at six weeks, six months, and 12 months postoperatively.[19]

Statistical analysis. The statistical Package for Social Sciences Version 17.0 (SPSS Inc., Chicago, Illinois) was used. Descriptive statistics were used to define the patient's characteristics. A paired t-test determined whether there was an actual difference between Oxford

Table I. Prospective preoperative and recall preoperative Oxford Hip Score

	Minimum	Maximum	Mean (sd)
Prospective preoperative	3	45	20.9 (8.89)
Recall preoperative	0	48	20.9 (10.00)
Difference	-39	30	0.04 (8.78)

The mean difference is 0.04 with a standard deviation of 8.78, so the 95% confidence interval is -15. 64 to 15.72

Table II. Prospective preoperative and recall preoperative Oxford Knee Score

	Minimum	Maximum	Mean (sd)
Prospective preoperative	3	43	21.85 (8.35)
Recall preoperative	2	42	20.26 (8.44)
Difference	- 13	21	1.59 (6.71)

The mean difference is 1.59 with a sd of 6.712, so the 95% confidence interval is -11. 57 to 14. 75

joint scores collected pre- and postoperatively. A p-value < 0.05 was considered statistically significant.

The recalled preoperative Oxford scores were plotted against the prospectively collected preoperative scores for each patient and the relationships were summarized using the intra-class correlation coefficient.[20] Relative reliability examines the relation between two or more sets of repeated measures.[20] Intra-class correlation coefficients (two-way random effects model with absolute agreement) were applied to determine the relative reliability of the recall scores compared with the prospectively collected scores.[20,21] Scores for intra-class correlation coefficient range from 0 to 1, where the former shows no reliability and the latter exhibits perfect reliability. Cicchetti and Sparrow[22] and Fleiss[23] have suggested that a score of < 0.40 is poor, 0.40 to 0.59 is fair, 0.60 to 0.74 is good, and > 0.74 is excellent, but it is generally recognized that a coefficient of < 0.70 is considered unacceptable.[24]

Bland and Altman's limits of agreement were calculated and plotted.[21] Bland and Altman recommend that the differences between each of the two scores be compared, plotting the differences against the means of the scores.[21,25] No linear relationship on the Bland and Altman plot indicates that the statistical variation was similar for individuals with low clinical measurement scores and high clinical measurement scores. Additionally, the variation in the values was not proportional to, or dependent on, the mean clinical measurement score.[25]

Results

The mean age of the cohort was 72.5 years old (22 to 92) and the majority were female (n = 180, 53.7%). The most common pathology affecting these patients was osteoarthritis, which affected 113 (88%) of the primary THA patients and 109 (89%) of primary TKA patients. A minority had revision procedures: 6 patients (3.8%) had revision TKA and 21 patients (11.8%) had revision THA.

The mean for the actual preoperative OHS and the recalled score were both 20.9 (Table I), with a mean difference of 0.04 (95% confidence interval (CI) -15.64 to 15.72, p=0.97). The mean recalled OKS was 20.26 (Table II), which was lower than the actual OKS preoperative score of 21.85 with a mean difference of 1.59 (95% CI -11.57 to 14.75, p=0.10). The mean difference in the actual preoperative score and the recalled score for both OHS and OKS were not statistically significant, nor were they greater than the defined minimally important difference.[10] The coefficient of reliability was calculated as 15.70 for the OHS using the Bland and Altman method (Fig. 1): 95% of the score differences were between ± 15.70 points. The coefficient of reliability was calculated as 13.16 for the OKS using the Bland and Altman method (Fig. 2).

The spread of data for the prospective and recalled OHS data is represented in Figure 3. The OHS results demonstrate that the prospective and recalled data follow a linear relationship with the points lying close to the best-fit line.[18] An intra-class correlation coefficient demonstrates good reliability for the mean preoperative score ('average measures' intra-class correlation) with a score of 0.802 (Table III); however, this is not the case for the individual's score of 0.670 ('single measures' intra-class correlation). This finding is also not affected by the absolute score (Bland–Altman plot) demonstrated in Figure 1.

The spread of data for the prospective and recall preoperative OKS is represented in Figure 4. The scatter plots for the OKS demonstrate that the prospective and recalled data follow a linear relationship, with the points lying close to the best-fit line.[25] An intra-class correlation coefficient demonstrates good reliability for the mean recall preoperative OKS ('average measures' intra-class correlation) with a score of 0.772 (Table IV) but not for an individual's score 0.629 ('single measures' intra-class correlation). This is also not affected by the absolute score (Bland–Altman plot) demonstrated in Figure 2.

There were no linear relationships evident in either Oxford score Bland–Altman plots, indicating that a statistical variation was similar for individuals with low scores and high scores, and the variation in the values were not
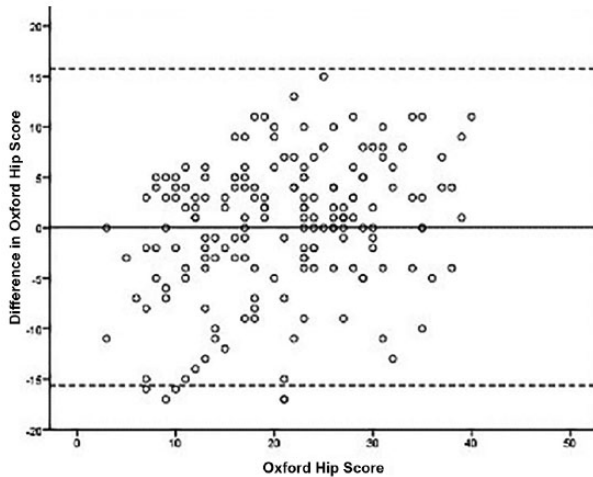
Fig. 1

Bland–Altman plot showing the mean Oxford Hip Score against the difference between prospective preoperative and recall preoperative Oxford Hip Scores.
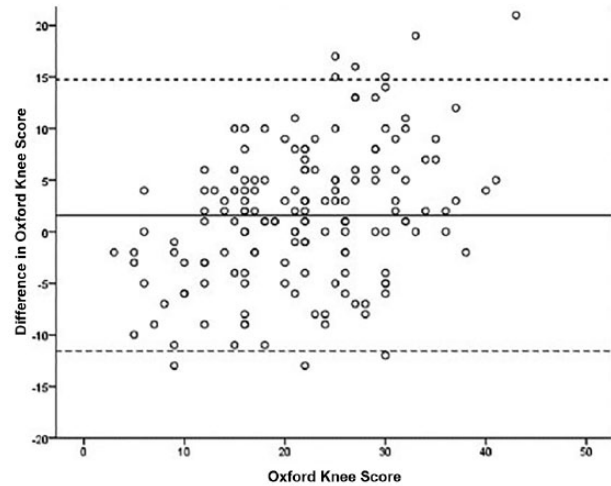

Fig. 2

Bland–Altman plot showing the mean Oxford Knee Score against the difference between prospective preoperative and recall preoperative Oxford Knee Scores.
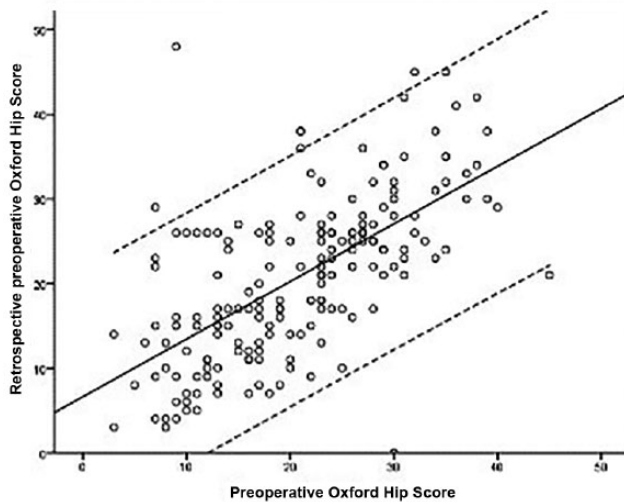

Fig. 3

Scatter plots showing the correlation of preoperative prospective Oxford Hip Score compared with the recalled preoperative score at one year.
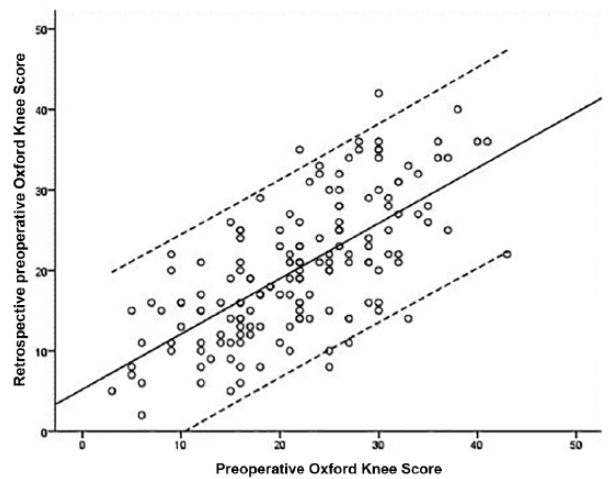

Fig. 4

Scatter plots showing correlation of preoperative prospective Oxford Knee Score compared with the recalled preoperative score at one year.

proportional to or dependent on the mean OHS or OKS (Figs 2 and 4).

Discussion

This study has demonstrated that the reproducibility of the recalled preoperative OHS and OKS one year following arthroplasty surgery for a cohort of patients is reliable, with no statistical difference in the mean recalled and actual preoperative scores. In contrast, the reliability of an individual patient to recollect their own preoperative OHS or OKS one year following arthroplasty surgery is poor. There was an excellent correlation between the recalled and actual scores when assessed as a mean, but when assessed individually the correlation was reduced to good. Interestingly, the variation of the patient's

recalled preoperative score was not influenced by the actual score, suggesting this is an intrinsic variability of the scores.

Several studies investigating recollection of symptoms in arthroplasty patients have demonstrated accurate recollection of preoperative symptoms up to six weeks postoperatively.[14,22,26,27] However, beyond six weeks the accuracy is mixed. Mancuso and Charlson[26] reported that patients after THA using the Hip Rating Questionnaire and found poor to fair agreement between prospective preoperative scores and recalled preoperative scores at a mean of 2.5 years. They concluded that relying on a patient's recollection does not provide an accurate measure of the preoperative state. Marsh et al[17] assessed the WOMAC, OHS, and 12-Item Short-Form (SF-) 12 Health

Table III. Prospective preoperative Oxford Hip Score compared with recall preoperative scores. Intra-class correlation (ICC): two-way mixed effects model (people effects are random and measures effects are fixed) with absolute agreement

	ICC[*]	95% confidence interval	F-test with true value 0			
			Value	df1	df2	Sig
Single measures	0.670[†]	0.569 to 0.750	5.262	156	156	< 0.0001
Average measures	0.802[‡]	0.725 to 0.857	5.262	156	156	< 0.0001

*Type A intra-class correlation coefficients using an absolute agreement definition
†The estimator is the same, whether the interaction effect is present or not
‡This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise

Table IV. Prospective preoperative Oxford Knee Score compared with recall preoperative scores. Intra-class correlation (ICC): two-way mixed effects model (people effects are random and measures effects are fixed) with absolute agreement

	ICC[*]	95% confidence interval	F-test with true value 0			
			df1	df2	Sig	Sig
Single measures	0.629[†]	0.531 to 0.710	4.368	178	178	< 0.0001
Average measures	0.772[‡]	0.694 to 0.830	4.368	178	178	< 0.0001

*Type A intra-class correlation coefficients using an absolute agreement definition
†The estimator is the same, whether the interaction effect is present or not
‡This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise

Survey and found that THA patients can accurately recall their preoperative health status at six weeks postoperatively but not at three months. Howell et al[15] studied 104 THA patients using several scores (WOMAC, OHS, and SF-12) and found that the reliability of patient recollection of preoperative function remained accurate for up to three months postoperatively.[15,17] Lingard et al[14] report the largest series (n = 770) of patients after TKA and found a moderate agreement between recalled scores and prospectively collected preoperative data at three months after surgery using the WOMAC and SF-36 health survey. They supported this retrospective collection of preoperative data but highlighted the fact that this method of data collection is not a direct substitute for prospectively collected data. It was suggested that patient experience during the interval between the two data collection points could recalibrate their internal standards of pain and functional ability leading to an over- or underestimation of the benefit of surgery.[14,28,29]

The current study seems to contradict the conclusions, but mirrors the findings of the recent study by Murphy et al.[16] Our study supports the findings of Murphy et al,[16] in that an individual patient recollection of their preoperative OHS or OKS is not accurate when assessed one year following surgery. We demonstrated a wide 95% CI for the recalled OHS (± 16 points) and OKS (± 14 points) and, similarly to Murphy et al,[16] a correlation of 0.6 to 0.7 between the recalled and actual scores. However, the OHS and OKS has a recognized variation when assessed for an individual patient. The reproducibility, according to the 95% confidence levels, vary by ± 7 points for the OHS and ± 6 points for the OKS, when re-assessed 24 hours apart. We showed that at one year this individual patient variation is double that observed after 24 hours.

The primary aim of the current study was to assess the reproducibility of recalled OHS and OKS for a cohort, as would be used in clinical practice, and have shown this to be a reproducible mean score. This finding is also supported by the results of Murphy et al,[16] who also demonstrated a similar mean score for the recalled and actual preoperative score for both the OHS and OKS.

This study's strength was the recruitment of a large cohort of arthroplasty patients from a single centre with standardized rehabilitation and follow-up for all patients. Good patient numbers were recruited in both the hip and knee groups to investigate their corresponding OHS and OKS. The study focused on the reliable and ubiquitous Oxford joint scores that are widely adopted in arthroplasty research and national arthroplasty databases. The study's weakness is that recalled preoperative data was collected at a single interval postoperatively. Collection of recalled data at frequent time intervals over a longer period could give information as to whether recalled data in a cohort of patients becomes more or less reliable over time. An additional weakness was the fact we did not assess the patient recall variation of their preoperative score according to their demographics, such as age and gender, which may influence their answers. Factors including other comorbidities, mental health, or the ceiling effect of the OHS and OKS could have also affected their recall ability.[8]

The results of this study demonstrate that individually collected recalled preoperative OHS and OKS do not agree sufficiently with the prospectively collected scores to be reliable. However, the average measure (mean of a cohort) of the retrospectively collected recalled preoperative OHS and OKS did not differ significantly from the prospectively collected preoperative

data.[26] Based on these results, the null hypothesis can be rejected in part. It can be concluded that, when applied to a large group, the recollection of pre-operative hip and knee symptoms by hip and knee arthroplasty patients using the OHS and the OKS was not subject to recall bias.[26] It is, therefore, possible to use the mean values of retrospectively collected pre-operative OHS and OKS scores at one year after surgery, within a population, to collect accurate data to assess the impact of an arthroplasty procedure.[26]

Prospective preoperative collection of OHS and OKS remains the benchmark. Using recalled scores one year following arthroplasty surgery is an alternative when used to assess a cohort of patients. However, the recall of individual patient preoperative score should not be relied upon due to the diminished reliability and wide confidence intervals.

References

1. **Ramkumar PN, Harris JD, Noble PC.** Patient-reported outcome measures after total knee arthroplasty: a systematic review. *Bone Joint Res* 2015;4:120-127.

2. **Paulsen A, Odgaard A, Overgaard S.** Translation, cross-cultural adaptation and validation of the Danish version of the Oxford hip score: assessed against generic and disease-specific questionnaires. *Bone Joint Res* 2012;1:225-233.

3. **Dawson J, Fitzpatrick R, Carr A, Murray D.** Questionnaire on the perceptions of patients about total hip replacement. *J Bone Joint Surg [Br]* 1996;78-B:185-190.

4. **Dawson J, Fitzpatrick R, Murray D, Carr A.** Questionnaire on the perceptions of patients about total knee replacement. *J Bone Joint Surg [Br]* 1998;80-B:63-69.

5. **Dawson J, Fitzpatrick R, Frost S, et al.** Evidence for the validity of a patient-based instrument for assessment of outcome after revision hip replacement. *J Bone Joint Surg [Br]* 2001;83-B:1125-1129.

6. **Murray DW, Fitzpatrick R, Rogers K, et al.** The use of the Oxford hip and knee scores. *J Bone Joint Surg [Br]* 2007;89-B:1010-1014.

7. **Murray DW, Pandit H, Weston-Simons JS, et al.** Does body mass index affect the outcome of unicompartmental knee replacement? *Knee* 2013;20:461-465.

8. **Hamilton DF, Giesinger JM, MacDonald DJ, et al.** Responsiveness and ceiling effects of the Forgotten Joint Score-12 following total hip arthroplasty. *Bone Joint Res* 2016;5:87-91.

9. **EuroQol Group.** EuroQol—a new facility for the measurement of health-related quality of life. *Health Policy* 1990;16:199-208.

10. **Beard DJ, Harris K, Dawson J, et al.** Meaningful changes for the Oxford hip and knee scores after joint replacement surgery. *J Clin Epidemiol* 2015;68:73-79.

11. **Sims AL, Parsons N, Achten J, et al.** The World Hip Trauma Evaluation Study 3: hemiarthroplasty evaluation by multicentre investigation - WHITE 3: HEMI - An Abridged Protocol. *Bone Joint Res* 2016;5:18-25.

12. **Griffin XL, Parsons N, McArthur J, Achten J, Costa ML.** The Warwick Hip Trauma Evaluation One: a randomised pilot trial comparing the X-Bolt Dynamic Hip Plating System with sliding hip screw fixation in complex extracapsular hip fractures: WHiTE (One). *Bone Joint J* 2016;98-B:686-689.

13. **Clement ND, MacDonald DJ, Hamilton DF, Burnett R.** Posterior condylar offset is an independent predictor of functional outcome after revision total knee arthroplasty. *Bone Joint Res* 2017;6:172-178.

14. **Lingard EA, Wright EA, Sledge CB 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-A:1149-1156.

15. **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;23:843-849.

16. **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-B:1070-1075.

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-A:2827-2837.

18. **Clement ND, Jenkins PJ, MacDonald D, et al.** Socioeconomic status affects the Oxford knee score and short-form 12 score following total knee replacement. *Bone Joint J* 2013;95-B:52-58.

19. **Clement ND, MacDonald D, Howie CR, Biant LC.** The outcome of primary total hip and knee arthroplasty in patients aged 80 years or more. *J Bone Joint Surg [Br]* 2011;93-B:1265-1270.

20. **Alkjaer T, Henriksen M, Dyhre-Poulsen P, Simonsen EB.** Forward lunge as a functional performance test in ACL deficient subjects: test-retest reliability. *Knee* 2009;16:176-182.

21. **Bland JM, Altman DG.** Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1:307-310.

22. **Cicchetti DV, Sparrow SS.** Developing criteria for establishing the interrater reliability of specific items in a given inventory. *Am J Ment Defic* 1981;86:127-137.

23. **Fleiss JL.** *Statistical Methods for Rates and Proportions.* Second ed. New York: Wiley, 1981.

24. **Howell DC.** *Statistical Methods for Psychology.* Fourth ed. Belmont: Duxbury Press, 1997:490-493.

25. **Wilson J, Baker P, Rangan A.** Is retrospective application of the Oxford Shoulder Score valid? *J Shoulder Elbow Surg* 2009;18:577-580.

26. **Mancuso CA, Charlson ME.** Does recollection error threaten the validity of cross-sectional studies of effectiveness? *Med Care* 1995;33(suppl):AS77-AS88.

27. **Bryant D, Norman G, Stratford P, et al.** Patients undergoing knee surgery provided accurate ratings of preoperative quality of life and function 2 weeks after surgery. *J Clin Epidemiol* 2006;59:984-993.

28. **Howard GS, Dailey PR.** Response-shift bias: a source of contamination of self-report measures. *J Appl Psychol* 1979;64:144-150.

29. **Sprangers M, Hoogstraten J.** Pretesting effects in retrospective pretest-posttest designs. *J Appl Psychol* 1989;74:265-272.