ELSEVIER

Journal of Clinical Epidemiology 59 (2006) 984–993

Journal of Clinical Epidemiology

Patients undergoing knee surgery provided accurate ratings of preoperative quality of life and function 2 weeks after surgery

Dianne Bryant[a,*], Geoff Norman[b], Paul Stratford[b], Robert G. Marx[c], S.D. Walter[b], Gordon Guyatt[b]

[a]*School of Physical Therapy, Faculty of Health Sciences, University of Western Ontario, Elborn College, Room 1438, London, Ontario N6G 1H1, Canada*
[b]*Department of Clinical Epidemiology & Biostatistics, McMaster University, Hamilton Health Sciences, Room 2C12, Hamilton, Ontario L8N, Canada*
[c]*Foster Center for Clinical Outcome Research, Sports Medicine and Shoulder Service, Hospital for Special Surgery, 535 East 70th Street, New York, NY 10021, USA*

Accepted 22 January 2006

Abstract

Objective: To evaluate patients' ability to recall their preoperative self-reported quality of life, function, and general health 2 weeks postoperatively.

Study Design and Setting: We randomized consecutive patients undergoing arthroscopic knee surgery to group I (assessments at 4 weeks preoperatively, on the day of surgery, and 2 weeks and 12 months postoperatively) or group II (assessments at 2 weeks and 12 months postoperatively). At each visit patients completed disease-specific, knee-specific, and generic health rating scales. At a median of 2 weeks postoperative (range, 0.6 to 14 weeks), patients completed questionnaires according to their recollection of their health 2 weeks before surgery.

Results: Agreement between actual and recalled data was excellent for disease-specific ($ICC_{WOMET} = 0.88$ (95% CI 0.82–0.91), $ICC_{ACL\text{-}QOL} = 0.86$ (95% CI 0.75–0.91)), knee-specific ($ICC_{IKDC} = 0.92$ (95% CI 0.90–0.94), $ICC_{KOOS} = 0.93$ (95% CI 0.91 to 0.95), and general physical health ($ICC_{SF\text{-}36(PCS)} = 0.81$ (95% CI 0.75–0.86)) instruments. Agreement for general mental health was moderate ($ICC_{SF\text{-}36(MCS)} = 0.67$ (95% CI 0.58–0.75). Greater error associated with recalled ratings contributed to small increases in sample size requirements or small decreases in power to detect differences between groups.

Conclusion: Patients recalled their preoperative quality of life, function, and general health at 2 weeks postoperative with sufficiently high accuracy to warrant substituting prospectively collected baseline data with recalled ratings.

Keywords: Retrospective; Baseline; Data collection; Memory; Mental recall; Orthopedic procedures

1. Introduction

When reporting the results of clinical trials comparing two interventions, investigators often describe patients' baseline health status to illustrate pretreatment similarity between groups. As well, these data often provide a covariate for statistical comparisons between groups to control for any differences that were present before treatment commenced and to thus increase the power to demonstrate the intervention effect.

Often times, the investigators for clinical trials involving surgical interventions are unable to fully assess a patient's eligibility before surgical evaluation. Thus, trialists and research staff collect preoperative (baseline) data for patients who according to history, imaging studies, and clinical examination appear to meet eligibility criteria for the study, but who nevertheless have the potential to be disqualified following surgical evaluation. Depending on the specificity of preoperative evaluations to diagnose the disease of interest and to identify concomitant pathology, the number of patients who prove to be ineligible following surgical examination can be high introducing huge inefficiencies to the process of data collection and the allocation of research resources.

For example, in a recent randomized trial to compare the effectiveness of inside-out suturing to bioabsorbable Arrows for reparable meniscal lesions, 700 patients undergoing anterior cruciate ligament (ACL) reconstruction or isolated arthroscopy who were suspected of having a meniscal tear gave their consent to participate and completed baseline assessments before surgery (each requiring approximately 40 minutes to complete). Following arthroscopic evaluation, however, only 100 of these patients had

* Corresponding author. Tel.: 519-661-2111x83947; fax: 519-661-3866.
E-mail address: dianne.bryant@uwo.ca (D. Bryant).

0895-4356/06/$ – see front matter
doi: 10.1016/j.jclinepi.2006.01.013

a meniscal tear amenable to repair using either intervention, making them eligible for inclusion into the study, which meant that 85.7% of consented patients were excluded. Therefore, as the use of quality-of-life instruments to measure treatment effects gains in popularity for orthopedic populations, the question arises as to whether investigators can collect baseline quality-of-life data retrospectively following surgical determination of patient eligibility, and whether these data will accurately represent data that would have otherwise been collected preoperatively.

Implicit theories of memory postulate that people possess beliefs about the stability of personal attributes as well as the conditions that might produce a change in these attributes such as a treatment or intervention in the instance of disease [1]. Implicit theories and present status may then be used as guides to judge previous states [1,2] when information cannot be recalled. If the situation is one in which patients believe that no change has occurred (e.g., chronic illness), they may assume a rating of their current status is an accurate reflection of their past status. In situations where no change has actually occurred, we might expect the appearance of accurate recall. However, if only gradual change has occurred, so that patients are unaware of the change, those who have improved will tend to overestimate their previous health, whereas patients who have declined will tend to underestimate their previous health [1].

In situations where patients believe their status has changed or should change, for example, following an intervention (surgery, rehabilitation, or drug therapy), they may again use their current status as a point from which to judge their prior state. If they perceive that an improvement has occurred, they may recall their previous state as being worse than their current state and vice versa if they perceive that their condition has been made worse by treatment. If reality is that there was little to no actual change, the recalled rating will again be an over- or underestimation of the previous rating [1].

It is possible that memory for acute situations differs from memory for more chronic conditions, especially if the situation occurs suddenly, so that it is distinct and carries with it a uniqueness that may make it easier to recall [3]. In the case of postsurgical patients who are asked to recall preoperative health, the surgery itself may be a sufficiently salient event to provide an anchor for accurate recollections of health before the event. It is also possible that soon after surgery patients discount their current pain and disability to the effects of surgery, assuming it to be variable and nonpermanent, and therefore do not use their current state as a reference from which to judge their previous state. Thus, it is possible that patients who have had recent surgery may be able to recall their prior health more accurately than if pain and discomfort associated with surgery has abated.

In the interest of efficient use of research resources and easing patient burden, we conducted a trial to investigate patients' ability to recall presurgical quality of life, function, and general health at 2 weeks postoperative.

2. Methods

2.1. Study design

This study was a randomized clinical trial with two centers (London Health Sciences Centre and Hamilton Health Sciences Center, Ontario, Canada) with eight orthopedic surgeons participating in patient recruitment. Research assistants at each center systematically reviewed incoming referrals and the charts of patients scheduled for arthroscopic surgery with or without ACL reconstruction. The research assistant contacted patients at least 4 weeks prior to surgery to determine their willingness to participate in the study. Eligible patients who gave consent were allocated into one of two groups. Group I underwent assessment at 4 weeks preoperatively, on the day of surgery, and at 2 weeks (both current and recalled health status) and 12 months postoperatively. Group II underwent assessment at 2 weeks (both current and recalled health status) and 12 months postoperatively (Fig. 1). The purpose of including group II in the study was to be able to compare recalled ratings between group I and group II, which were assumed to have similar characteristics by virtue of randomization, to determine whether there was evidence that group I patients' prior exposure to the instruments (4 weeks preoperative and day of surgery) influenced their recalled ratings.

2.2. Eligibility criteria

To decrease any learning effect, we excluded patients who had previous experience with the measurement tools being used in this study; thus, participants in past and present clinical trials with similar questionnaires were excluded. We also excluded patients undergoing minor procedures (manipulation or removal of hardware) or those with rare diseases or conditions that would usually preclude their participation in clinical trials (past or present history of metabolic bone, collagen, crystalline, or neoplastic disease). During surgical investigation, the surgeon conducted an arthroscopic examination of the knee joint and surrounding structures to rule out concomitant pathology that would render the patient ineligible. We also excluded patients who were unwilling or unable to be assessed according to study protocol including those with no fixed address, those with plans to move outside of the vicinity of a participating center, those with a major psychiatric illness, those who are intellectually challenged, and those unable to speak or understand English.

2.3. Patient randomization

To facilitate the balance of potential prognostic characteristics between groups, randomization was stratified by surgeon, the presence or absence of ACL reconstruction (to address possible differential reliance on current status as a reference for recalling preoperative status—and thus differentially accurate recall—related to the ACL reconstruction

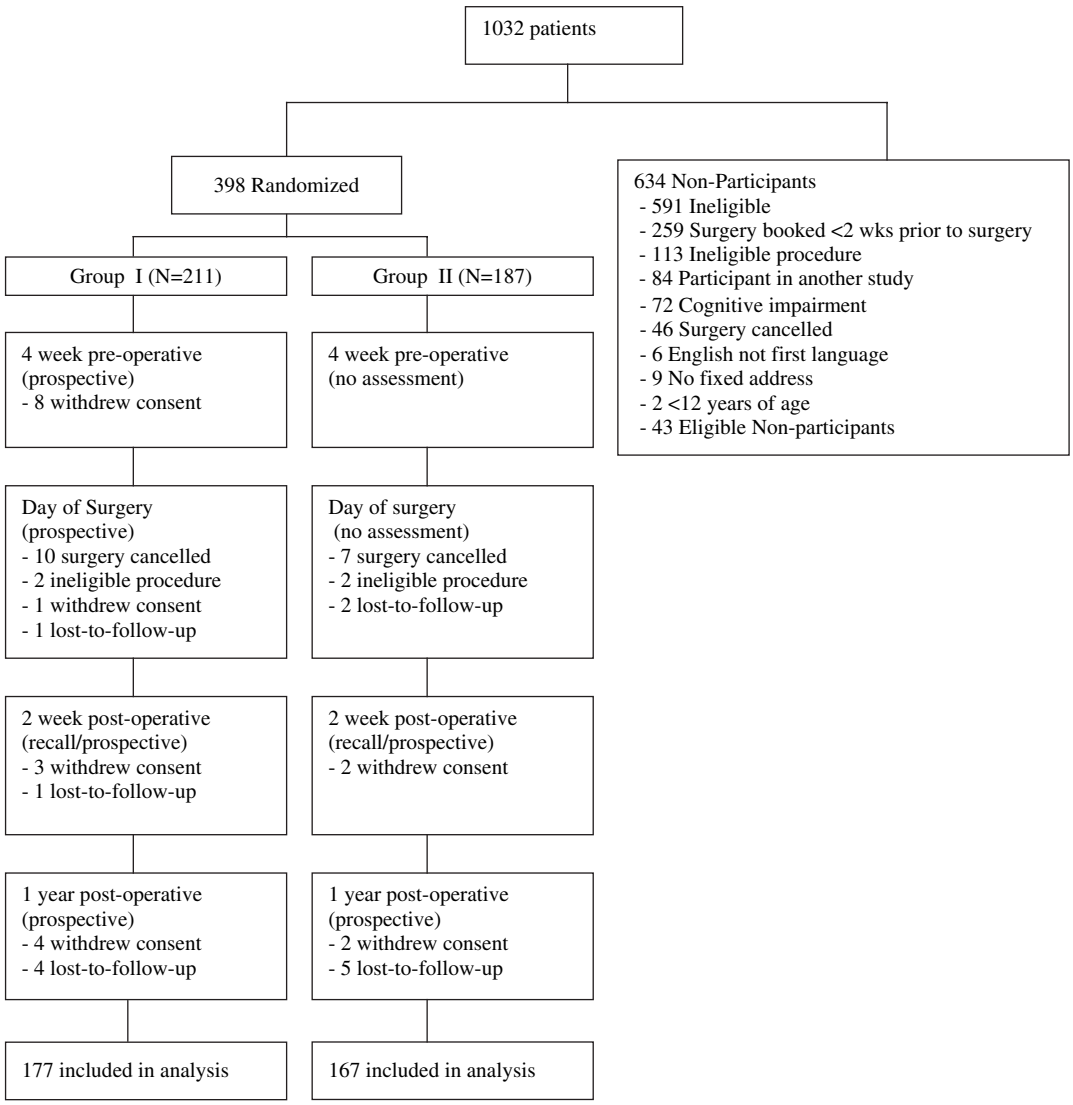

Fig. 1. Flow of patients through the trial.

resulting in a longer duration of disability in the early postoperative period than isolated arthroscopy), and presence or absence of previous knee surgery (to address a similar issue of possibly differential recall related to patients who have undergone previous knee surgery being more accepting of postoperative limitations). The randomization sequence was constructed using a computer algorithm with permuted block sizes of two and four. To ensure adequate concealment of allocation, the research assistant established patient eligibility and obtained verbal consent prior to contacting an independent researcher (uninvolved in patient recruitment) who consulted the randomization code and informed the research assistant of patient allocation to group I or group II.

2.4. Patient assessments

At 4 weeks preoperative (group I only), on the day of surgery (group I only) and at 12 months postoperative (both groups I and II), the research assistant asked each patient to complete the questionnaires to assess his or her current quality of life, general health, and functional status over the past 2 weeks. At 2 weeks postsurgery, the research assistant provided the patient with two sets of questionnaires (both groups I and II). For the first set of questionnaires, the research assistant asked the patients to recall their quality of life, general health, and function during the 2 weeks prior to surgery and to complete the questionnaires according to that recall. For the second set of questionnaires, the research assistant asked the patient to assess his or her current quality of life, general health, and functional status over the past 2 weeks. Figure 1 provides an illustration of the timing and type of assessments in each group (prospective/retrospective).

2.5. Measurement instruments

At each visit, all patients completed the International Knee Documentation Committee (IKDC) [4] Subjective

Form, the Knee Injury and Osteoarthritis Outcome Score (KOOS) [5], and the Short-Form Health Survey (SF-36) [6] according to the assessment schedule for each group. In addition, patients undergoing ACL reconstruction completed the Quality of Life Outcome Measure (Questionnaire) for Chronic Anterior Cruciate Ligament Deficiency (ACL-QOL) [7] and patients undergoing an isolated arthroscopy completed the Western Ontario Meniscal Evaluation Tool (WOMET) [8]. All questionnaires were self-administered under the supervision of a research assistant.

The ACL-QOL [9] is a 32-item disease-specific quality-of-life questionnaire for patients with ACL deficiency. It has five domains that query physical symptoms (five items), occupational concerns (four items), recreational activities (12 items), lifestyle (six items), and social and emotional aspects (five items). Each item has one 100 mm visual analogue scale response option, with labeled anchors at 0 mm (e.g., extremely difficult) and 100 mm (e.g., not difficult at all). A patient's score is calculated by converting the average of each of the five domain scores to a total average score out of 100% where 100% represents the best possible score. This instrument has face validity and has demonstrated content and construct validity, excellent test–retest reliability (standard error of measurement (S.E.M.) = 6%) and is responsive to change [9].

The WOMET [8] is a 16-item disease-specific quality-of-life questionnaire for patients with meniscal pathology. There are three domains that query physical symptoms (nine items), recreation, occupation and lifestyle (four items), and emotional well-being (three items). Response options are framed as 100 mm visual analogue scales, with anchors at 0 mm (e.g., not bothered) and 100 mm (e.g., extremely bothered). A patient's score is determined by calculating the sum of each item to attain a raw score, which is subtracted from 1,600 (worst possible score), and divided by 16 to obtain a score out of 100%, where 100% represents the best possible score. This instrument has face validity and has been shown to have content and construct validity, excellent test–retest reliability (intraclass correlation coefficient [ICC] = 0.79, 95% confidence interval (CI) 0.59 to 0.87), and is responsive to change [8].

The IKDC [4] is an 18-item knee-specific questionnaire designed to detect change in patients with a variety of knee conditions including queries into physical symptoms (seven items), sports activities (10 items), and function prior to injury (one item). The number of response options per question varies between two options (one item), five options (14 items), and 11 options (three items). A patient's score is determined by calculating the difference between the raw score and lowest possible score (18) and then dividing this difference by the range of possible scores (87), multiplied by 100. The resulting total score is out of 100 possible points, which represents perfect knee function. This method of scoring weights each item according to the number of response options (where an item with only two response options can contribute a maximum of two points and an item with 11 response options can contribute up to 11 points). This instrument has face validity and has demonstrated construct validity, excellent test–retest reliability (ICC = 0.94, 95% CI 0.88 to 0.97), and is responsive to change [4,10].

The KOOS [5] is a 42-item knee-specific questionnaire with five separately reported domains of pain (eight items), physical symptoms (seven items), activities of daily living (17 items), function in recreation (five items), and quality of life (four items). Each item has five response options. Domain scores represent the average of all items in the domain standardized to a maximum score of 100%, which is the best possible score. In this study, each domain yielded similar results, and thus, for simplicity we present the overall aggregate score only. This instrument has face validity and demonstrated construct validity, excellent test–retest reliability for each domain (range, 0.75 to 0.93), and has been shown to be responsive to change [5,11].

The SF-36 [6] is a 36-item generic general health instrument that evaluates eight domains including restrictions or limitations on physical and social activities, normal activities and responsibilities of daily living, pain, mental health and well-being, and perceptions of health. The SF-36 can be reported as eight domain scores (physical functioning, role physical, bodily health, social functioning, role social, mental health, general health, vitality) or as two overall scores (Physical Component Score (PCS) or Mental Component Score (MCS)). Component scores are computed aggregates of all eight domain scores where the weight of each domain score is different for each component score. The SF-36 has been extensively used, and has been shown to be valid, reliable, and responsive in a wide variety of populations and contexts [12,13] including patients with orthopedic conditions [11,14,15]. For simplicity we present the PCS and the MCS only.

2.6. Statistical analysis

Because participants in group I had two prior exposures to the questionnaires before being asked to recall their baseline status, we explored the possibility that recalled ratings were influenced by their previous responses. Because patients were randomly assigned to groups, we expected that both groups were similar with respect to their expected true baseline status so that if accurate recall was possible, the scores for the recalled data would also be similar between groups. To explore this possibility and to avoid multiple testing, we conducted independent samples *t*-tests where the dependent variable was the recalled rating for each instrument and the independent variable was group allocation with Bonferroni correction for multiple comparisons so that $P < 0.025$ indicated a significant difference between the recalled ratings between groups.

We conducted an analysis of the validity and reliability of recalled ratings using two conceptual approaches. First,

we assumed that the preoperative measurement on the day of surgery provides a gold or criterion standard of patients' preoperative health and that if valid, recalled ratings collected 2 weeks postoperatively will accurately predict ratings provided on the day of surgery. Second, we assumed that both time points measure the same construct and should thus have high agreement or reliability.

To determine the validity of recalled ratings as compared to actual preoperative ratings (day of surgery), we used linear regression to determine the ability of patients' recalled data (the independent variable) to predict actual ratings (the dependent variable) for all questionnaires. We constructed scatterplots of these data with 95% group and individual prediction intervals around each regression line to illustrate the between- and within-subject variability and agreement between the two time points that one could expect for groups or an individual. We conducted several graphical and numerical tests to verify the assumptions of linear regression (linearity, normality, homoscedasticity). Specifically, from the scatterplot of actual vs. recalled data, we visually verified linearity and looked for potential outliers. Using a scatterplot of the standardized residuals against the prediction value we visually verified homoscedasticity and constructed Q-Q plots and histograms of the standardized residuals and used the Shapiro–Wilk test of the normality of residuals where $P < 0.05$ indicates significant deviations from the normal distribution [16].

To determine the reliability, we conducted a repeated measures analysis of variance (ANOVA) to determine whether there were significant systematic differences between actual and recalled ratings (main effect of time). Using the mean square values from the ANOVA (between-subjects (patients), within-subjects (time), and error), we estimated the magnitude of the association or accuracy of the recalled to actual data, by constructing an intraclass correlation coefficient (ICC) (two-way mixed model with measures of absolute agreement) for each instrument and its 95% CI [17]. Because the ICC is difficult to interpret in that it does not provide direct information about the error associated with measurement, we calculated the S.E.M. from the ANOVA (square root of the mean square error) and its 95% CIs to provide an easily defined estimate of the reproducibility of individual measurements [18].

Finally, to assess the impact of using recalled data in place of the prospectively collected baseline data, we estimated the sample size and power for each of the questionnaires corresponding to three common methods of making statistical comparisons between groups, (a) *t*-test of the posttest score only, (b) *t*-test of the change score (pretest–posttest), and (c) ANCOVA (analysis of covariance), where pretest (actual or recalled) is used as the covariate. For all calculations we considered an important difference as 20% of the mean preoperative score (actual or recalled) [19] and maintained the probabilities of type I and type II error at 0.05 and 0.20, respectively.

3. Results

3.1. Patient characteristics

We screened 1,032 consecutive patients who were scheduled for knee arthroscopy and/or ACL reconstruction. Of the 531 eligible patients, 133 did not participate (50 could not be contacted, 40 cancelled surgery, and 43 refused to give consent). Thus, 398 patients gave consent and were randomized. Two hundred and eleven patients were randomized to group I (73 ACL, 138 arthroscopy) and 187 patients were randomized to group II (60 ACL, 127 arthroscopy). Twenty-one patients were excluded post-randomization (17 cancelled surgery, four underwent ineligible procedures), 18 patients withdrew their consent, and 15 patients were lost-to-follow-up (group I = 6, group II = 9). The analyses include data from 344 patients. Figure 1 provides a more detailed description of the flow of patients through the trial.

Patient characteristics were similar between groups and for nonparticipants (missed, noneligible, and eligible nonparticipants) in age, gender, operative knee, procedure being performed (ACL reconstruction vs. arthroscopy), third party compensation status, prevalence of previous knee surgery, and time from injury to surgery (Table 1).

Table 1
Characteristics of participants and nonparticipants

Characteristic	Group I	Group II	Nonparticipants[a]
Males	62%	62%	59%
Age (years)	39 ± 13	36 ± 14	35 ± 14
Height (inches)	70 ± 14	70 ± 15	74 ± 23
Weight (pounds)	182 ± 38	174 ± 39	174 ± 42
Third party compensation[b]	16%	14%	14%
Smoking status[c]			
Never smoked	51%	57%	57%
Smoked, but quit	26% (11 ± 12)	24% (12 ± 15)	23% (11 ± 12)
Current smoker	23% (12 ± 11)	18% (12 ± 11)	21% (11 ± 12)
Right knee affected	56%	44%	50%
Scope only (no ACL reconstruction)	65%	68%	63%
Previous surgery	42%	41%	48%
Time from injury to surgery			
<3 months	3%	7%	7%
3 months to 1 year	30%	28%	28%
1 to 3 years	24%	26%	24%
3 to 5 years	5%	5%	9%
5 to 10 years	6%	5%	6%
>10 years	32%	28%	26%

Abbreviation: ACL = anterior cruciate ligament.

[a] Includes ineligible and eligible nonparticipants.

[b] Includes Workman's Safety and Insurance Board, disability, and litigation.

[c] Includes the proportion of patients who smoke and the average pack years (number of years of smoking multiplied by number of packs per day).

3.2. The influence of prior exposure to instruments on ability to recall

The difference (mean ± standard deviation) between recalled and actual ratings between group I and group II were not statistically different for any instrument, and the estimates of the difference between groups were not thought to represent a clinically meaningful difference (WOMET −5.60 ± 2.67, ACL-QOL −3.30 ± 3.24, IKDC −3.30 ± 2.05, KOOS −1.96 ± 2.03, SF-36 PCS −1.57 ± 1.19, SF-36 MCS 0.64 ± 2.75), suggesting that prior exposure to instruments did not affect recalled ratings (Table 2).

Table 2
Descriptive statistics of ratings at all time points for both groups

Time	Questionnaire	Group I (n = 177)	Group II (n = 167)
4-Week preoperative	WOMET[a]	34.3 ± 17.4	
	ACL-QOL[b]	32.2 ± 17.5	
	IKDC	46.2 ± 17.1	
	KOOS	57.2 ± 17.1	
	SF-36 PCS	38.6 ± 10.1	
	SF-36 MCS	51.7 ± 9.2	
Day of surgery	WOMET	35.8 ± 18.3	
	ACL-QOL	33.4 ± 17.5	
	IKDC	45.5 ± 17.7	
	KOOS	57.4 ± 17.3	
	SF-36 PCS	39.9 ± 9.9	
	SF-36 MCS	52.1 ± 9.2	
2-Week postoperative (recalled)[c]	WOMET	34.6 ± 19.5	40.2 ± 20.7
	ACL-QOL	36.6 ± 18.1	39.9 ± 16.8
	IKDC	45.6 ± 18.9	48.9 ± 19.1
	KOOS	57.9 ± 18.3	59.9 ± 19.4
	SF-36 PCS	39.0 ± 10.8	40.6 ± 11.2
	SF-36 MCS	52.6 ± 9.7	51.9 ± 10.0
2-Week postoperative (current)	WOMET	37.9 ± 23.5	38.2 ± 21.6
	ACL-QOL	20.9 ± 14.1	24.7 ± 15.9
	IKDC	33.3 ± 18.1	35.5 ± 19.5
	KOOS	47.2 ± 18.1	51.4 ± 18.1
	SF-36 PCS	32.2 ± 9.7	33.3 ± 10.3
	SF-36 MCS	50.2 ± 10.5	51.7 ± 10.6
1-Year postoperative	WOMET	56.1 ± 26.7	57.4 ± 29.6
	ACL-QOL	61.7 ± 20.9	64.5 ± 20.0
	IKDC	62.1 ± 22.4	65.9 ± 25.0
	KOOS	71.4 ± 19.2	73.0 ± 20.9
	SF-36 PCS	45.1 ± 10.8	45.7 ± 11.4
	SF-36 MCS	52.8 ± 9.1	52.2 ± 10.2

Abbreviations: IKDC = International Knee Documentation Committee; WOMET = Western Ontario Meniscal Evaluation Tool; KOOS = Subjective Form, the Knee Injury and Osteoarthritis Outcome Score; SF-36 = Short-Form Health Survey; PCS = Physical Component Score; and MCS = Mental Component Score; ACL-QOL = Quality of Life Outcome Measure (Questionnaire) for Chronic Anterior Cruciate Ligament Deficiency.

[a] There were 113 patients with WOMET scores in each group.

[b] There were 62 patients in group I and 55 patients in group II with ACL-QOL scores.

[c] The independent groups t-test comparison of recalled ratings between group I and group II were not significant.

3.3. Patients' ability to recall preoperative quality of life and general health status

Scatterplots with prediction lines and the 95% CIs from group I patients' recalled vs. actual ratings of quality of life, general health, and functional status are suggestive of high levels of agreement for all questionnaires. Across all instruments, recalled ratings were a significant predictor of actual ratings ($P < 0.001$) and Pearson's correlation coefficient indicates excellent agreement between ratings (range, 0.67 to 0.88) (Table 3). Figure 2 shows scatterplots for the IKDC and SF-36 Mental Component Scale. Residual analysis verified that the data were consistent with the assumptions of linear regression.

The difference between the mean of recalled and actual ratings were consistently small across all instruments; only one of the within-group comparisons (ACL-QOL) reached statistical significance (−3.24 95% CI −5.43 to −1.04, $P = 0.01$) though this difference is not important. Further, there was no obvious pattern in the direction of the difference between recalled and prospective ratings among questionnaires.

Reliability of recalled and actual ratings for the WOMET (ICC = 0.88, 95% CI 0.82 to 0.91) and the ACL-QOL (ICC = 0.86, 95% CI 0.75 to 0.91) disease-specific questionnaires, the IKDC (ICC = 0.92, 95% CI 0.90 to 0.94) and KOOS (ICC = 0.93, 95% CI 0.91 to 0.95) knee-specific questionnaires, and the PCS of the SF-36 ($ICC_{PCS} = 0.81$, 95% CI 0.75 to 0.86) was excellent, whereas reliability of the SF-36 MCS was moderate (and $ICC_{MCS} = 0.67$, 95% CI 0.58 to 0.75) (Table 4).

The S.E.M. for each questionnaire (Table 4) was small and similar to that observed in test–retest situations. By examining the scatterplot of recalled to actual data, the lower level of agreement for the generic health measure cannot be attributed to greater measurement error but rather to smaller between-subject variability (greater homogeneity) for this population with respect to general health as measured by generic instruments compared to more specific instruments.

Table 3
Predictive validity of using retrospective (recalled) ratings in place of prospective (actual) ratings

Questionnaire	Pearson's r	Coefficient (β)
WOMET	0.88	0.82 (95% CI 0.74–0.91), $P < .001$
ACL-QOL	0.88	0.75 (95% CI 0.65–0.86), $P < .001$
KOOS	0.93	0.88 (95% CI 0.83–0.93), $P < .001$
IKDC	0.92	0.86 (95% CI 0.81–0.92), $P < .001$
SF-36 PCS	0.81	0.75 (95% CI 0.67–0.83), $P < .001$
SF-36 MCS	0.68	0.64 (95% CI 0.53–0.75), $P < .001$

Abbreviations: IKDC = International Knee Documentation Committee; WOMET = Western Ontario Meniscal Evaluation Tool; KOOS = Subjective Form, the Knee Injury and Osteoarthritis Outcome Score; SF-36 = Short-Form Health Survey; PCS = Physical Component Score; and MCS = Mental Component Score; CI = confidence interval; ACL-QOL = Quality of Life Outcome Measure (Questionnaire) for Chronic Anterior Cruciate Ligament Deficiency.

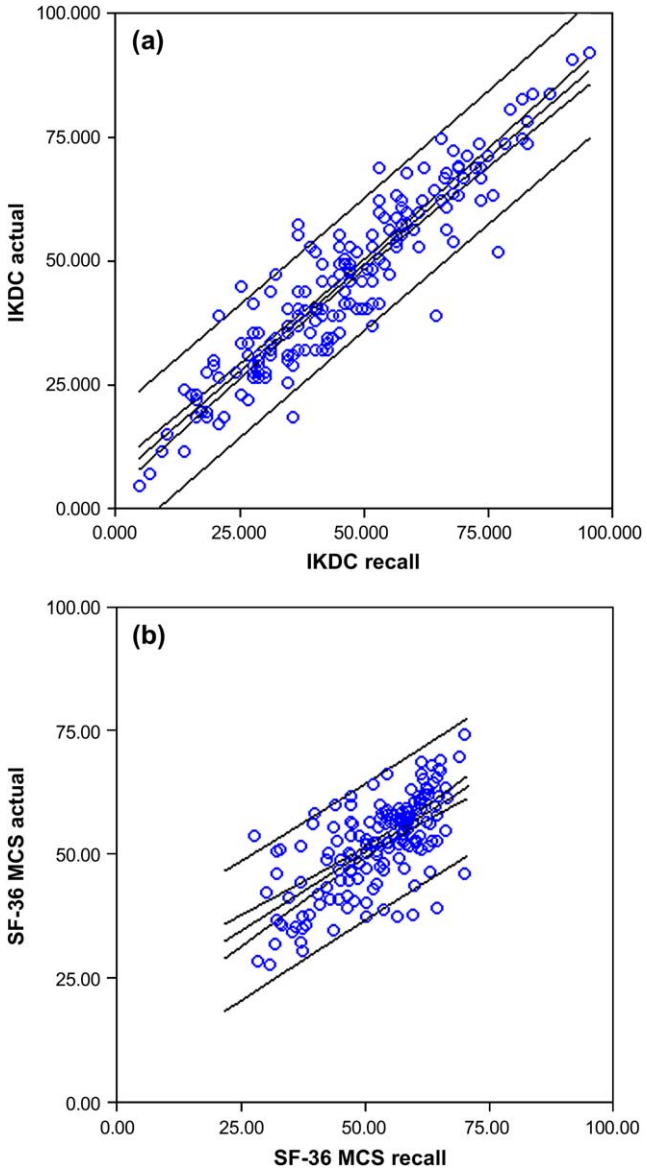

Fig. 2. Patients' recalled rating of quality of life compared to actual rating for the (a) International Knee Documentation Committee (IKDC) and (b) Short-Form Health Survey (SF-36) Mental Component Score (MCS).

3.4. The effect on sample size and power when using retrospective data

The correlation between actual preoperative ratings and 1-year postoperative ratings ranged from 0.40 to 0.59, whereas the correlation between recalled preoperative and 1-year postoperative ratings ranged from 0.39 to 0.56 across questionnaires. For each instrument, correlations between actual or recalled data and the 1-year postoperative scores did not differ significantly from each other (range of the difference between correlations, 0.00 to 0.05).

The required increase in sample size to detect an important difference when using actual vs. recalled data varied between 8% and 31% if the planned statistical comparisons involve a change score and between 8% and 27% if the planned comparisons involve an ANCOVA. None of the sample size estimates using recalled or actual data for a planned ANCOVA were greater than those that would be required for comparisons using a postscore only, whereas half of calculations for sample size using recalled or actual data for comparisons using change scores were greater than those using a postscore only (Table 5).

Similarly, estimates of reduction in power when using recalled data instead of prospective data with change scores ranges from 3% to 11% or from 2% to 7% with ANCOVA (Table 5). All estimates of power for comparisons using an ANCOVA were greater than the 80% power of a planned posttest only comparison (6% to 13% gain if using recalled ratings and 3% to 9% gain if using actual ratings).

Table 4
Agreement between actual and recalled baseline data

Questionnaire	Statistic	Agreement
WOMET	Difference (95% CI), *P*-value	1.12 (95% CI −0.63 to 2.88), $P = 0.21$
	ICC	0.88 (95% CI 0.82 to 0.91)
	S.E.M.	6.65 (95% CI 5.88 to 7.65)
ACL-QOL	Difference (95% CI), *P*-value	−3.24 (95% CI −5.43 to −1.04), $P = 0.01$
	ICC	0.86 (95% CI 0.75 to 0.91)
	S.E.M.	6.12 (95% CI 5.20 to 7.44)
KOOS	Difference (95% CI), *P*-value	−0.50 (95% CI −0.50 to 1.50), $P = 0.33$
	ICC	0.93 (95% CI 0.91 to 0.95)
	S.E.M.	4.77 (95% CI 4.32 to 5.33)
IKDC	Difference (95% CI), *P*-value	−0.03(95% CI −1.04 to 1.10), $P = 0.96$
	ICC	0.92 (95% CI 0.90 to 0.94)
	S.E.M.	5.11 (95% CI 4.63 to 5.70)
SF-36 PCS	Difference (95% CI), *P*-value	0.90 (95% CI −0.05 to 1.85), $P = 0.06$
	ICC	0.81 (95% CI 0.75 to 0.86)
	S.E.M.	4.50 (95% CI 4.07 to 5.03)
SF-36 MCS	Difference (95% CI), *P*-value	−0.43 (95% CI −1.64 to 0.79), $P = 0.49$
	ICC	0.67 (95% CI 0.58 to 0.75)
	S.E.M.	5.52 (95% CI 4.98 to 6.21)

Abbreviations: IKDC = International Knee Documentation Committee; WOMET = Western Ontario Meniscal Evaluation Tool; KOOS = Subjective Form, the Knee Injury and Osteoarthritis Outcome Score; SF-36 = Short-Form Health Survey; PCS = Physical Component Score; and MCS = Mental Component Score; ICC = intraclass correlation coefficient and S.E.M. = standard error of measurement; CI = confidence interval; ACL-QOL = Quality of Life Outcome Measure (Questionnaire) for Chronic Anterior Cruciate Ligament Deficiency.

Note: The Pearson's correlation coefficient given in Table 3 and the ICC given in Table 4 are identical because the actual and recalled ratings are similar and the predominant source of error is random error.

4. Discussion

The primary reasons surgeons undertake orthopedic procedures for chronic conditions are to improve patients'

Table 5
Assessment of the affect of using recalled ratings on sample size and power for three common methods of making statistical comparisons

					Sample size estimations actual:recalled (change in sample size)			Power estimations actual:recalled (change in power)		
Questionnaire	SD_a	SD_r	r_{ao1}	r_{ro1}	Post	Change	ANCOVA	Post	Change	ANCOVA
WOMET	18.3	19.5	0.53	0.52	104	98:114 (16%)	75:86 (15%)	0.80	0.82:0.76 (−6%)	0.91:0.87 (−4%)
ACL-QOL	17.5	18.1	0.40	0.39	98	118:128 (9%)	82:89 (8%)	0.80	0.72:0.69 (−4%)	0.86:0.83 (−3%)
KOOS	17.3	18.3	0.59	0.56	36	29:35 (21%)	23:27 (19%)	0.80	0.87:0.80 (−7%)	0.93:0.89 (−4%)
IKDC	17.7	18.9	0.50	0.47	59	59:72 (22%)	44:53 (19%)	0.80	0.80:0.80 (−8%)	0.90:0.84 (−6%)
SF-36 (PCS)	9.9	10.8	0.54	0.49	25	23:30 (31%)	18:22 (27%)	0.80	0.83:0.72 (−11%)	0.91:0.84 (−7%)
SF-36 (MCS)	9.2	9.7	0.50	0.50	12	12:14 (8%)	9:10 (8%)	0.80	0.80:0.77 (−3%)	0.90:0.87 (−2%)

Abbreviations: IKDC = International Knee Documentation Committee; WOMET = Western Ontario Meniscal Evaluation Tool; KOOS = Subjective Form, the Knee Injury and Osteoarthritis Outcome Score; SF-36 = Short-Form Health Survey; PCS = Physical Component Score; and MCS = Mental Component Score; ACL-QOL = Quality of Life Outcome Measure (Questionnaire) for Chronic Anterior Cruciate Ligament Deficiency.

Note: SD_a = standard deviation of actual baseline data, SD_r = standard deviation of recalled baseline data, r_{ao1} = Pearson's correlation coefficient of actual baseline ratings (day of surgery) to posttest (1-year postoperative), r_{ro1} = Pearson's correlation coefficient of recalled baseline ratings (recall) to posttest (1-year postoperative). For all calculations, probability of type I error = 0.05, probability of type II error = 0.20. An important difference was calculated as 20% of the mean preoperative rating for each instrument [19].

quality of life and functional status. Thus, patient-rated quality of life and functional assessments are often used as the primary measure of a treatment's impact. Currently, the perceived necessity to measure quality of life and functional status prior to surgery, when potentially only a small proportion of patients will prove eligible for participation in a particular trial often introduces large inefficiencies in data collection (unnecessary patient burden and use of research resources to support research staff to collect data). The results of this study suggest that one can reasonably substitute retrospective recalled data for preoperative baseline data thus avoiding these inefficiencies.

This study supports the use of retrospective baseline self-ratings of health status in this population for explanatory purposes; exploring relationships between variables in a sample with the goal of generalizing to a population. In our regression analysis (assessment of predictive validity), we present both population (group) and individual prediction lines (Fig. 2). When trying to predict the preoperative health status on an individual patient basis using retrospective recalled ratings from that individual, it is not possible to predict with any certainty the magnitude or even the direction of the discrepancy in recall is (has the individual over- or underestimated his or her actual status and by how much?). This phenomenon is not unique to retrospective data collection but is a concern whenever predictions are made for individuals using population data.

This study shows that patients undergoing ACL reconstruction and/or arthroscopy can accurately recall their preoperative disease-specific quality of life, general health, and functional status at 2 weeks after surgery. Other studies that have assessed patients' ability to recall pain, functional status, and general health have had mixed results [20–38] although differences between studies in their design may help to explain these inconsistencies. For example, some studies asked patients to recall specific experiences (e.g., worst pain, least pain) [20,22,23,26,28,38], whereas others asked patients to recall their average experience (e.g., average pain over the past week) [29,31]. Patients were unable to accurately recall specific experiences but showed good recall of average experiences. In our study, all questions on all instruments were worded so that patients indicated their average experience over the past 2 weeks for prospective ratings and their average experience during the 2 weeks prior to surgery for the recall task.

Duration of time between actual and recalled ratings may provide another explanation for the inconsistency in results between studies. Studies that asked patients to recall over a longer timeframe (2 months to 10 years) [21,23,25,26,28,37,38] found that patients could not recall their pretreatment status accurately, whereas studies that asked patients to recall over a shorter timeframe (1 day to 2 weeks) [24,29,31,33] found that patients could provide accurate ratings of pain and function. In our study, patients were asked to recall preoperative quality of life, general health, and functional status at an average of 2 weeks following surgery. Our finding that patients can accurately recall their preoperative quality of life is consistent with other studies using a similar timeframe for recall.

Our purpose in investigating patients' ability to recall preoperative status differs from the majority of studies reported in the literature that assess ability to recall. Our objective was to determine whether, in a prospective randomized trial, we could plan a priori to collect recalled pretreatment quality of life, general health, and functional status with reasonable accuracy to improve the efficiency of data collection. The objective of the majority of studies we reviewed was to determine the accuracy of patient's ability to recall for conducting retrospective unplanned studies.

There are three important differences between these objectives. The first is that by planning in advance to use recalled baseline data, we can plan to collect these data within a relatively short time interval following treatment. Secondly, when planning a randomized trial, we can ensure that the task of recollection occurs prior to potential post-treatment differences between groups in current health or

functional status (i.e., the benefits or harms associated with treatment are realized) that have been shown to cause deviations in recall (i.e., over- or underestimates actual status) [1]. Thirdly, randomization should ensure that differences in ability to recall between patients will be equally distributed between groups, essentially removing the influence of these differences on the estimate of treatment effect.

In addition to our results supporting the ability of patients' to accurately recall preoperative health status, we also presented data indicating the effect on power and sample size estimates if one were to use recalled data in place of prospectively collected baseline data. Intuitively, we expect a greater degree of error when asking patients to recall a prior state, thus increasing the within-subject error, which will contribute to a greater overall error (variance). Because variance is directly related to the estimate of sample size, and inversely related to power, greater variance will lead to a larger estimated of sample size and a reduction in power (or increase in type II error rate). In fact, our results do demonstrate that recalled data do have greater associated variances, a finding that was consistent across questionnaires.

What implications for sample size follow from this increased variance? Several authors have discussed the inefficiencies of using postonly scores compared to change scores when making comparisons between independent groups when the magnitude of the correlation between pretest and final ratings is greater than 0.5 [39–41]. This criterion (correlation of >0.5 between preintervention and postintervention scores) was not met by three of the questionnaires and was barely met by the remaining questionnaires leading us to believe that inefficiencies in statistical comparisons (increased probability of type II error) in orthopedic surgical trials may exist if investigators use change scores. Further, a large body of literature exists to support the use of baseline measurements as a covariate in an ANCOVA, because this type of analysis takes into consideration the between-subject variations in baseline measurements, thus making its tolerance of small associations between pretest and posttest data much greater [39–45]. From the results of this study (Table 5), in comparing the expected power across statistical methods for between-group comparisons, use of ANCOVA provides considerable advantage. Further, if researchers plan to use recalled data as their pretest measurement for a planned comparison between groups using an ANCOVA but do not have sufficient pilot data to estimate the increase in variance or the magnitude of the association between pretest and posttest, they can generate a conservative estimate of sample size by using calculations for sample size meant for postonly comparisons. Thus, if the optimal statistical method for making comparisons between groups in a randomized trial (ANCOVA) is used, minimal losses in power ($<10\%$) can be expected if using recalled preoperative data.

One further issue relevant to interpreting our results is the opposing opinions of implicit theorists [1] and response shift theorists. Implicit theorists believe that recalling a previous state is difficult without other contextual features with which to associate the memory [1,3]. Without such a reference point people begin their ''recollection'' by asking themselves how they are currently, followed by asking themselves how they think things have changed, and then infer what their initial state must have been like [1]. The difference between actual and a recalled preoperative ratings is thought to be a reflection of the error in this process [1]. For implicit theorists, our results suggest that surgery is, at least for the subsequent 2 weeks, a sufficiently vivid event that it allows people to accurately recall their presurgical status [3].

Response shift theorists [46–48] argue that as a result of their changing status patients may change their internal standard of measurement, whereby 'most of the time' when thinking of pain, for example, means something different than it did previously based on more recent experiences (scale recalibration). Further, patients' values may change over time so that aspects of quality of life, such as physical health, social health, and psychological health, for example, change in the order of their importance. Finally, patients may redefine the target construct (e.g., quality of life) so that how they define quality of life changes (scale reconceptualization). Response shift theorists claim that differences between the actual and recalled preoperative ratings (or the thentest) represent a response shift and that the thentest or recalled rating is the only valid measure of preoperative status because it is rated using the same metric as the postoperative or final outcome rating. For response shift theorists, our data provide an example of a context in which response shift does not occur, at least to an appreciable degree.

Strengths of this study include its size (approximately 400 patients), its design (randomized, multisurgeon, multicenter clinical trial), the robustness of the results (patients' ability to recall was consistent across questionnaire types, including a generic health measure and knee- and disease-specific questionnaires, with differing forms of response options including Likert-type scales, yes/no, or visual analogue scale), and the diversity of the patient population. Specifically, our sample of patients included a broad spectrum of surgical procedures of the knee including ACL reconstruction, meniscal repair, and meniscectomy to debridement of articular cartilage for mild to severe osteoarthritis. In addition, our sample demonstrated a wide range in patient ages (from 13 to 78 years), levels of experience with prior knee surgery (0 to 8 previous surgeries), and acute vs. chronic injuries (time from injury to surgery ranged from 3 days to 3 years and 42 patients with gradual onset (and no specific injury)).

5. Conclusion

Patients undergoing knee surgery can accurately recall their preoperative quality of life, general health, and functional status at 2 weeks postoperative. The results suggest that investigators can improve the efficiency of data collection for randomized clinical trials in this patient population,

with minimal loss of statistical power, by obtaining retrospective ratings of preoperative function.

References

[1] Ross M. Relation of implicit theories to the construction of personal histories. Psychol Rev 1989;96:341–57.

[2] Bellezza FS, Brower GH. Person stereotypes and memory for people. J Pers Soc Psychol 1981;41:23–31.

[3] Baddeley A. Recollection and autobiographical memory human memory. Toronto: Allyn and Bacon; 1998.

[4] Irrgang JJ, Anderson AF, Boland AL, Harner CD, Kurosaka M, Neyret P, Richmond JC, Shelborne KD. Development and validation of the international knee documentation committee subjective knee form. Am J Sports Med 2001;29:600–13.

[5] Roos EM, Roos HP, Ekdahl C, Lohmander LS. Knee injury and Osteoarthritis Outcome Score (KOOS)—validation of a Swedish version. Scand J Med Sci Sports 1998;8:439–48.

[6] Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. Med Care 1992;30:473–83.

[7] Altman DG, Bland JM. Statistics notes. Units of analysis. BMJ 1997;314:1874.

[8] Griffin S, Huffman H, Bryant D, Kirkley A. The development and validation of a quality of life measurement tool for patients with meniscal pathology: The Western Ontario Meniscal Evaluation Tool (WOMET). Manuscript in progress.

[9] Mohtadi N. Development and validation of the quality of life outcome measure (questionnaire) for chronic anterior cruciate ligament deficiency. Am J Sports Med 1998;26:350–9.

[10] Irrgang JJ, Ho H, Harner CD, Fu FH. Use of the International Knee Documentation Committee guidelines to assess outcome following anterior cruciate ligament reconstruction. Knee Surg Sports Traumatol Arthrosc 1998;6:107–14.

[11] Di Fabio RP, Boissonnault W. Physical therapy and health-related outcomes for patients with common orthopaedic diagnoses. J Orthop Sports Phys Ther 1998;27:219–30.

[12] McHorney CA, Ware JE Jr, Lu JF, Sherbourne CD. The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. Med Care 1994;32:40–66.

[13] McHorney CA, Ware JE Jr, Raczek AE. The MOS 36-Item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. Med Care 1993;31:247–63.

[14] Jette DU, Jette AM. Physical therapy and health outcomes in patients with knee impairments. Phys Ther 1996;76:1178–87.

[15] Shapiro ET, Richmond JC, Rockett SE, McGrath MM, Donaldson WR. The use of a generic, patient-based health assessment (SF-36) for evaluation of patients with anterior cruciate ligament injuries. Am J Sports Med 1996;24:196–200.

[16] Shapiro SS, Wilk MB. An analysis of variance test for normality. Biometrika 1965;52:591–9.

[17] McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. Psychol Methods 1996;1:30–46.

[18] Stratford PW, Goldsmith CH. Use of the standard error as a reliability index of interest: an applied example using elbow flexor strength data. Phys Ther 1997;77:745–50.

[19] Goldsmith CH, Boers M, Bombardier C, Tugwell P, for the OMERACT Committee. Criteria for clinically important changes in outcomes: development, scoring and evaluation of rheumatoid arthritis patient and trial profiles. J Rheumatol 1993;20:561–5.

[20] Haas M, Nyiendo J, Aickin M. One-year trend in pain and disability relief recall in acute and chronic ambulatory low back pain patients. Pain 2002;95:83–91.

[21] Pellise F, Vidal X, Hernandez A, Cedraschi C, Bago J, Villanueva C. Reliability of retrospective clinical data to evaluate the effectiveness of lumbar fusion in chronic low back pain. Spine 2005;30:365–8.

[22] Gedney JJ, Logan H, Baron RS. Predictors of short-term and long-term memory of sensory and affective dimensions of pain. J Pain 2003;4:47–55.

[23] Dawson EG, Kanim LE, Sra P, Dorey FJ, Goldstein TB, Delamarter RB, Sandhu HS. Low back pain recollection versus concurrent accounts: outcomes analysis. Spine 2002;27:984–93. discussion 994.

[24] Singer AJ, Kowalska A, Thode HC. Ability of patients to accurately recall the severity of acute painful events. Acad Emerg Med 2001;8: 292–5.

[25] Lingard EA, Wright EA, Sledge CB. Pitfalls of using patient recall to derive preoperative status in outcome studies of total knee arthroplasty. J Bone Joint Surg Am 2001;83-A:1149–56.

[26] Everts B, Karlson B, Wahrborg P, Abdon N, Herlitz J, Hedner T. Pain recollection after chest pain of cardiac origin. Cardiology 1999;92: 115–20.

[27] Bolton JE. Accuracy of recall of usual pain intensity in back pain patients. Pain 1999;83:533–9.

[28] Feine JS, Lavigne GJ, Dao TT, Morin C, Lund JP. Memories of chronic pain and perceptions of relief. Pain 1998;77:137–41.

[29] Zonneveld LN, McGrath PJ, Reid GJ, Sorbi MJ. Accuracy of children's pain memories. Pain 1997;71:297–302.

[30] Niven CA, Brodie EE. Memory for labor pain: context and quality. Pain 1995;64:387–92.

[31] Babul N, Darke AC, Johnson DH, Charron-Vincent K. Using memory for pain in analgesic research. Ann Pharmacother 1993;27:9–12.

[32] Rachman S, Eyrl K. Predicting and remembering recurrent pain. Behav Res Ther 1989;27:621–35.

[33] Hunter M, Philips C, Rachman S. Memory for pain. Pain 1979;6:35–46.

[34] Bryant RA. Memory for pain and affect in chronic pain patients. Pain 1993;54:347–51.

[35] Beese A, Morley S. Memory for acute pain experience is specifically inaccurate but generally reliable. Pain 1993;53:183–9.

[36] Eich E, Reeves JL, Jaeger B, Graff-Radford SB. Memory for pain: relation between past and present pain intensity. Pain 1985;23:375–80.

[37] Linton SJ, Melin L. The accuracy of remembering chronic pain. Pain 1982;13:281–5.

[38] Mancuso CA, Charlson ME. Does recollection error threaten the validity of cross-sectional studies of effectiveness? Med Care 1995; 33:AS77–88.

[39] Frison L, Pocock SJ. Repeated measures in clinical trials: analysis using mean summary statistics and its implications for design. Stat Med 1992;11:1685–704.

[40] Norman GR. Issues in the use of change scores in randomized trials. J Clin Epidemiol 1989;42:1097–105.

[41] Cronbach LJ, Furby L. How should we measure 'change'—or should we? Psychol Bull 1970;74:68–80.

[42] Knapp TR. The (un)reliability of change scores in counseling research. Meas Eval Guid 1980;13:149–57.

[43] Egger MJ, Coleman ML, Ward JR, Reading JC, Williams HJ. Uses and abuses of analysis of covariance in clinical trials. Control Clin Trials 1985;6:12–24.

[44] Oldham PD. A note on the analysis of repeated measurements of the same subjects. J Chronic Dis 1962;15:969–77.

[45] Lee J. A note on the comparison of group means based on repeated measurements of the same subject. J Chronic Dis 1980;33:673–5.

[46] Schwartz CE, Sprangers MA. Methodological approaches for assessing response shift in longitudinal health-related quality-of-life research. Soc Sci Med 1999;48:1531–48.

[47] Golembiewski RT, Billingsley K, Yeager S. Measuring change and persistence in human affairs: types of change generated by OLD designs. J Appl Behav Sci 1976;12:133–57.

[48] Howard GS, Dailey PR. Response-shift bias: a source of contamination of self-report measures. J Appl Psychol 1979;64:144–50.