The Bone & Joint Journal
Formerly known as JBJS (Br)

Five year follow-up results of the PROFHER randomised clinical trial comparing surgical versus non-surgical treatment of adults with displaced fractures of the proximal humerus

Journal:	*The Bone & Joint Journal*
Manuscript ID	Draft
Manuscript Type:	Original Article
Keywords:	Proximal humeral fractures, Randomised controlled trial, Surgical versus non-surgical treatment, Long-term follow-up

Abstract

Aims

To determine the long-term treatment effects beyond the two year follow-up reported in the PROFHER randomised clinical trial that compared surgery versus non-surgical treatment of adults with displaced proximal humeral fractures involving the surgical neck.

Patients and Methods

Trial participants consenting to the extended follow-up were sent postal questionnaires at three, four and five years after trial recruitment. The Oxford Shoulder Score (OSS; the primary outcome), EQ-5D-3L, and recent shoulder operations and fractures data were collected. Statistical and economic analyses, consistent with those of the main trial were applied.

Results

OSS data were available for 164, 155 and 149 participants at three, four and five years. There were no statistically or clinically significant differences between surgical and non-surgical treatment at each follow-up in the OSS. No participant had secondary shoulder surgery for a new complication. Analyses of EQ-5D data showed no significant between-group differences in quality of life over time.

Discussion

These results confirm that the main findings of the PROFHER trial over two years following fracture occurrence persist in the long term.

Introduction

This article reports on the extended follow-up, up to five years, of participants randomised into the PROximal Fracture of the Humerus Evaluation by Randomisation (PROFHER) trial. PROFHER was a pragmatic, multi-centre randomised controlled trial (RCT), funded by the UK National Institute for Health Research (NIHR), that compared surgical with non-surgical treatment of adults with displaced fractures of the proximal humerus involving the surgical neck [1]. It recruited 250 adults between September 2008 and April 2011, with data available for the Oxford Shoulder Score (OSS)[2,3], the primary outcome, from 215 participants at two year follow-up.[4] The results for the main pre-specified reporting period for PROFHER showed no significant difference between surgical treatment compared with non-surgical treatment in the OSS and other patient-reported clinical outcomes over two years following fracture [4,5] and that surgery cost significantly more over this period [6].

The initial choice of a two year follow-up for PROFHER was a pragmatic one that balanced feasibility and the expectation that any differences in the OSS between the two treatment groups by two year follow-up would represent a true and enduring effect. However, there is insufficient evidence from other RCTs to confirm this assumption [7]. Recovery from serious injuries such as proximal humeral fractures is a long and often incomplete process that can be hindered by complications. A substantial proportion of participants of a trial with less severe fractures than in PROFHER had continuing disability at two years follow-up, although reduced from the proportion at one year.[8] We reasoned that a five year follow-up would allow for delays in recovery, potential functional deterioration, and subsequent operations resulting from complications, such as avascular necrosis and complications of surgical fixation or humeral head replacement, which could arise or become symptomatic later on. The extension made sense practically as the infrastructure in place and the potential availability of a large group of patients – we anticipated there would be 200 followed up at two years – presented an unprecedented opportunity to gain important insights and reliable evidence on patient-reported longer-term outcome as well as insights on the feasibility of future research in this patient population. Hence, we set

up the extended follow-up study, securing ethics approval in September 2010 [5]; thus, prior to end of recruitment and without knowledge of the main study results.

Our primary aim was to obtain three, four and five year data on key outcomes (OSS, EQ-5D-3L (EuroQol Group, Rotterdam, Netherlands [9]) and subsequent surgery) in order to determine whether there was persistence or alteration of the treatment effect detected at two year follow-up, the definitive end point for the main trial. An associated aim, linked with the collection of EQ-5D-3L data and information on secondary surgery, was to examine the potential effects on our economic analysis [6] of any changes in quality of life and costs related to this key cost driver.

Our secondary aims were to:

- Obtain longer term condition specific data on shoulder function that would provide reference data for informing the interpretation of the findings of PROFHER and future studies of proximal humeral fractures.

- Inform future research in this area on the appropriate duration of follow-up.

Patients and Methods

Trial methods relating to the main study are reported elsewhere [1,5]. The trial inclusion and exclusion criteria are shown in Table 1. The trial protocol facilitated the implementation of good standards of treatment with the surgeons and physiotherapists using surgical interventions and procedures with which they were familiar. As reported, both interventions, including choice of surgical implants, and associated care programmes were representative of good current practice.[5] The final version of the extended study protocol (version 3.0; 09/08/12) is published on the NIHR website [10]; all related amendments were reviewed and approved by the Leeds West Research Ethics Committee (08/H1311/12).

Data collection

Outcome collection for the extended follow-up comprised postal questionnaires sent at three, four and five years after trial recruitment to all participants who had completed and returned a consent form sent within receipt of their 24 month questionnaire. A pre-notification letter was sent prior to the questionnaire and, where necessary, reminders were sent after two and four weeks, with the option to complete questionnaires via telephone after six weeks. To maximise collection of data at the three time points, patients were asked to complete a short questionnaire restricted to the OSS, EQ-5D-3L, recent operations on their shoulder, and recent fractures. Patients were also sent an unconditional £5 incentive payment with each questionnaire. We also collected data from the NHS Information Centre on patient mortality at regular intervals prior to dispatch of questionnaires; this was to prevent distressing bereaved families or friends.

Outcomes

The primary outcome was the OSS, a shoulder-specific outcome assessing pain, function and activities of daily living.[2,3] It contains 12 items, each with 5 categories of response, and a range of total scores of 0 (worst outcome) to 48 (best outcome).[3] Secondary outcomes were the EQ-5D-3L, used to estimate utilities ('health related quality of life weights')[9], further shoulder surgery and further fractures. While mortality was a secondary outcome in the main follow-up, it was reported solely as a reason for loss to follow-up in the extended follow-up because mortality and definitive treatment of the fracture after two years could not reasonably be expected to be linked. The OSS and EQ-5D-3L were collected at 6, 12, 24, 36, 48 and 60 months; EQ-5D-3L data were also collected at baseline and three months. Secondary shoulder surgery and further fractures were collected via hospital forms at one and two years; and via patient questionnaires at three, four and five years follow-up.

Sample size

The main study was designed to detect a standard effect size of 0.4 (approximating to 5 OSS points) with 80% power using 5% significance level, requiring approximately

200 participants (at 2 years)[1]. We considered that a 20% attrition rate at 5 years was a reasonable assumption to make for the population under study. We therefore based our proposal on a final sample size of 160 at the end of the five year follow-up period, which would provide 71% power to detect a standard effect size of 0.4 using 5% significance level. Given the reduced statistical power for the extended follow-up, significance testing was limited to the primary outcome alone.

Statistical and economics analyses

Analyses followed pre-specified formal analysis plans, the statistical analysis plan also being endorsed by the Data Monitoring and Ethics Committee, and were performed using Stata version 13.1 (StataCorp). All analyses were on an intention-to-treat basis, with participants being analysed in the groups to which they were randomised. Significance tests were 2-sided at the 5% significance level.

Primary analysis

OSS data from the extended follow-up time points were added to the PROFHER trial primary analysis model. The model compared OSS data from surgical and non-surgical treatment groups over all follow-up assessments. A multilevel model was fitted with time points nested in patients to allow for clustering of data within each patient. The model adjusted for the fixed effects of treatment group, time (6 months, 1, 2, 3, 4 and 5 years), interaction between treatment group and time, tuberosity involvement at baseline (yes or no), age (less than 65 years, 65 years or older), gender and health status at baseline (EQ-5D). The unstructured covariance pattern was retained from the primary analysis model. Patients with valid OSS data at one or more follow-up points for the standard or extended follow-up as well as complete covariate data were included in the analysis. Estimates of the difference in OSS between treatment groups, 95% confidence intervals and p-values were obtained for the extended follow-up time points at three, four and five years.

In a sensitivity analysis, the multilevel model was repeated substituting missing data with data derived by multiple imputation by chained equations. Missing outcome and

covariate data were imputed from age, gender, tuberosity involvement, EQ-5D-3L index at baseline and available OSS data at other follow-up points.

Subgroup Analyses

In line with the main trial, the possibility of differential long-term treatment responses for older patients (subgroups: less than 65 years versus 65 years or older) and more complex fractures (subgroups: involvement of no tuberosities versus one or both tuberosities) was explored. Prior expectations of the benefit of surgery over conservative treatment, established before knowledge of the main trial results, were that this was larger in patients under 65 years and larger in patients with fractures involving one or both tuberosities [1], and that these benefits may only emerge in the longer term [10]. Unadjusted OSS means by subgroups and treatment arm were therefore explored. In light of the substantially reduced statistical power for the subgroups, no statistical testing was performed.

Secondary outcomes

Annual and overall frequencies of shoulder surgery and fractures in each treatment group occurring within the previous year were calculated. Extended follow-up data were combined with those of the main trial to establish the number of participants by treatment group incurring secondary shoulder surgery or a further fracture over five years. Free text providing details of further surgery and non-pre-specified fractures was categorised, blind to treatment group, by two independent raters.

Economics analyses

The economic analysis aimed to explore whether the results obtained in the PROFHER trial were robust over a five year time horizon by determining the between-group differences in Health-Related Quality of Life (HRQoL, measured via the EQ-5D-3L) at set times (3, 4 and 5 years) and examining how this difference evolved over time (5 years). We also planned to estimate costs of any further shoulder surgery and report these descriptively.

The methods used to process the EQ-5D-3L data and calculate QALY (quality of life) scores were as in the main cost-effectiveness report [6]. In summary, the EQ-5D-3L data were transformed into 'health related quality of life weights' (utilities) using the

UK general population tariff which assigns societal values to each health state [11]. QALYs were calculated by combining the utility estimates by the duration of time in each health state using the area under the curve method (AUC) following the trapezium rule which assumed linear interpolation between follow-up points [12]. A discount rate was applied to QALYs after 12 months, at an annual rate of 3.5% [13].

In the main trial, the base-case analysis was conducted for the imputed dataset by means of multiple imputation (MI) with chained equations, using seemingly unrelated regression [6]. This method accounts for the correlation between costs and effects from the same individuals and imputes the missing data. However, other regression-based methods are available for handling missing data in longitudinal studies, principally mixed models, and results may be sensitive to the methods used [14]. A multilevel model similar to the primary OSS analysis was therefore conducted to investigate whether the results obtained in the main trial were robust to this alternative method of analysis. Thus the mean difference in utilities and QALYs (with 95% confidence intervals) between the two groups was estimated using a multilevel model that adjusted for the fixed effects of treatment group, time (3 and 6 months, 1, 2, 3, 4 and 5 years), interaction between treatment group and time, tuberosity involvement at baseline, age, gender and baseline utility.

Uncertainty around the results was explored by means of sensitivity analysis that used multiple imputation by chained equations to replace missing data on QALYs in the multilevel model where missing outcome and covariate data were imputed from age, gender, tuberosity involvement, and baseline utility.

Results

Of 176 patients (81% of the 218 who returned questionnaires at two years; 70% of 250 randomised trial participants) who consented to long term follow-up at two years following randomisation, valid Oxford Shoulder Scores (OSS) were received for 164 (93%) at three years, 155 (88%) at four years, and 149 (85%) at five years follow-up (Figure 1). Retention was therefore slightly lower than anticipated in the extended

follow-up; however, additional power was gained by the multilevel analysis. Ten patients died during the long term follow-up, five in each trial arm.

As found at baseline (except for smoking status, which did not affect the OSS results) and two year follow-up [4], patient characteristics were balanced between groups at five year follow-up in the 149 patients with complete OSS data (Tables IIa and b). Furthermore, the characteristics of the RCT population remained representative, as none of the baseline characteristics differed meaningfully between participants at the start of the trial and those remaining at the end.

Primary outcome (Oxford Shoulder Score)

Unadjusted OSS outcomes for patients with valid data were very similar in both groups for the extended follow-up period (Figure II). This featured a trend of small score increases between two and four years, with little difference in the fifth year. OSS scores were skewed towards maximum OSS shoulder functioning, with over half the population having stable and satisfactory shoulder function [3] at all three follow-up times: three years (median 42, IQR 35 to 47.5); four years (median 43, IQR 37 to 48); five years (median 44, IQR 36 to 48).

When adding the long term OSS follow-up data to the existing multilevel analysis, group differences were not statistically significant at any of the long term follow-up time points. This was true for the primary analysis model including all patients with available outcome data at any time point as well as the sensitivity analysis including all patients using data derived by multiple imputation (Table III and Table IV). None of the estimated mean differences were clinically meaningful; almost all were smaller than one OSS score point in magnitude with no consistent trend for the direction of the treatment effect.

The substantial overlap of the confidence intervals for the unadjusted OSS scores indicated there were no marked treatment group differences for patient subgroups based on age (Figure III) or tuberosity involvement (Figure IV). In both subgroups, the patterns of OSS score differences were not consistent with prior expectations.

Secondary Outcomes

Only one patient reported further shoulder surgery during extended follow-up. This was a reverse shoulder replacement in year three in a non-surgical group patient who already had surgery in the main follow-up. Thus the number of participants requiring secondary surgery remained at 11 in each treatment group [4].

A total of 81 further fractures were reported by 52 patients over the five year follow-up period. A small number of fractures are likely to be duplicated from one year to the next; however, as this could not be known definitively, patient data were accepted as submitted, with the exception of one participant who provided the date of their fracture. There were more fractures in the non-surgery group (50 fractures, 33 patients) than the surgery group (31 fractures, 19 patients); in particular, fractures of the spine and hip (Table V).

Economic analyses

Inevitably, relative to the 125 randomised into each treatment group, the extent of missing EQ-5D-3L data increased considerably in the extended follow-up period. For the 176 participants that consented to long term follow-up, complete EQ-5D-3L scores were available for 159 (90%) at three years, 153 (86%) at four years and 151 (86%) at 5 years.

Figure V shows the distribution of mean utilities (EQ-5D-3L scores) for all the available cases across the five years for the two groups. Patients in the surgery group started from a higher baseline utility on average (0.43, surgery *versus* 0.38, not surgery). However, at the end of the second year there was little difference in EQ-5D-3L scores between treatment groups. This finding was consistent at three, four and five years with the 95% confidence intervals overlapping at each assessment point. The same pattern applied for the analysis of utilities when adjusted for baseline utility or for all covariates (Table VI).

Between group mean difference in QALYs based on individual patient's utilities are shown in Table VII. At the end of the five years, patients allocated to the not surgery group obtained on average a marginally higher QALY gain than patients allocated to the surgery group. Hence the QALY gain for not surgery patients is maintained over time whether data are adjusted for baseline utility or for all covariates. The mixed

model, which included 200 cases, was repeated substituting missing data with data derived by multiple imputation by chained equations. For both analyses, there were negligible differences in the QALYs between the two groups at the different follow-up times (Table VII).

Discussion

The extension of follow-up has found no statistically or clinically significant differences between surgical and non-surgical treatment of displaced fractures of the proximal humerus involving the surgical neck at three, four or five years in the OSS, our primary outcome. Visual inspection of two subgroup analyses did not suggest any trend for group differences relating to age or fracture type. These findings mirror those of the main trial. No trial participant had secondary shoulder surgery for new complication during the extended follow-up period. The between group differences in utilities, based on EQ-5D-3L data, at three, four or five years were very small with the 95% confidence intervals overlapping at each assessment point. The same lack of statistically significant between-group differences applied to the quality of life analysis that showed the trend for a QALY gain for participants in the not surgery group is maintained over time. Sensitivity analyses indicated minimal differences between the two groups at each follow-up time.

By exceeding the original target of 200 participants at two year follow-up, PROFHER was sufficiently powered at the main follow-up. In contrast, we were 11 short of the 160 participants with OSS data at five years, and thus did not meet the revised statistical power criteria for the extended follow-up. However, we believe this is unlikely to affect the validity of the results. First, loss-to-follow-up including identical mortality (5 in each group) was balanced in both groups. Second, baseline characteristics at five years were comparable between groups as well as being representative of the original population. Third, much of the missing data was accounted for in the multilevel analysis, which included 231 cases. Fourth, the between group differences were small with the 95% confidence intervals at each follow-up time less than the minimal clinically important difference of 5 points. Fifth, the between group differences in the EQ-5D-3L were also very small, again reflecting

comparability of the groups. Last, there were no new complications warranting surgery.

Although there were no cost data to replicate the incremental cost-effectiveness analysis conducted for the PROFHER trial, the analyses of the health utility data for the five year period produced results that are consistent with the main trial analysis; on average, patients allocated to surgery reported lower HRQoL. The very small differences in HRQoL between the two groups found for the mixed model and multiple imputation analyses indicate negligible differences in quality of life between the treatment groups. The costs of the only shoulder operation reported for the extended follow-up would not have impacted on the findings of the main trial.

We consider that it is unsafe to draw any conclusions from the observed differences in participants incurring further fractures between the two groups in relation to treatment group allocation. We suggest this is primarily a chance effect. In terms of known risk factors for fractures (such as higher age, female gender, previous fracture and smoking), the two groups were at similar risk of further fractures at baseline except for smoking status, where there was a higher incidence of smokers in the not surgery group. This may partly explain a higher number of fractures in that group. Known inaccuracies, relating to both under- and over-reporting, of self-reported fractures are of some concern and indeed, based on additional participant commentary, we confirmed one instance of duplicate reporting over time. We also have no information on whether there was any difference in the advice offered and medication provided for preventing further fractures in the two groups.

Our findings of an absence of treatment differences in the OSS in the extended follow-up underpin the main findings for the two year follow-up. The only case of further surgery over the extended follow-up was repeat surgery for a complication that occurred within the two year follow-up. Given that most (15 of 22) secondary surgery occurred in the first year, this finding and the lack of difference in the OSS provide reassurance that late symptomatic complications are rare. The health-related quality of life results show that the PROFHER economic analysis persisted over a five year time horizon. The overall OSS results show that the majority of patients have attained

satisfactory shoulder function by two years, which is sustained subsequently. Thus the two year follow-up would have been sufficient for the PROFHER trial, and this could inform follow-up for future RCTs on these fractures.

Acknowledgements

We are grateful to the patients who generously completed questionnaires for the extended follow-up. We thank …. and …. for checking the Summary Care Records of patients for mortality, and staff at various participating sites for their help in locating missing patients.

References

1. Handoll H, Brealey S, Rangan A, Torgerson D, Dennis L, Armstrong A et al. Protocol for the ProFHER (PROximal Fracture of the Humerus: Evaluation by Randomisation) trial: a pragmatic multi-centre randomised controlled trial of surgical versus non-surgical treatment for proximal fracture of the humerus in adults. *BMC Musculoskelet Disord* 2009; 10(140).
2. Dawson J, Fitzpatrick R, Carr A. Questionnaire on the perceptions of patients about shoulder surgery. *J Bone Joint Surg Br*. 1996;78(4):593-600.
3. Dawson J, Rogers K, Fitzpatrick R, Carr A. The Oxford Shoulder Score revisited. *Arch Orthop Trauma Surg*. 2009;129(1):119-123.
4. Rangan A, Handoll H, Brealey S, Jefferson L, Keding A, Martin BC, et al. Surgical vs nonsurgical treatment of adults with displaced fractures of the proximal humerus: the PROFHER randomized clinical trial. *JAMA*. 2015;313(10):1037-47.
5. Handoll H, Brealey S, Rangan A, Keding A, Corbacho B, Jefferson L, et al. The ProFHER (PROximal Fracture of the Humerus: Evaluation by Randomisation) trial - a pragmatic multicentre randomised controlled trial evaluating the clinical effectiveness and cost-effectiveness of surgical compared with non-surgical treatment for proximal fracture of the humerus in adults. *Health Technol Assess* (Winchester, England). 2015;19(24):1-280.
6. Corbacho B, Duarte A, Keding A, Handoll H, Chuang LH, Torgerson D, Brealey S, Jefferson L, Hewitt C, Rangan A. Cost effectiveness of surgical versus nonsurgical treatment of adults with displaced fractures of the proximal humerus. Economic evaluation alongside the PROFHER TRIAL. *Bone Joint J* 2016;98-B:2;152-9
7. Handoll HHG, Brorson S. Interventions for treating proximal humeral fractures in adults. *Cochrane Database Sys Rev* 2015, Issue 11. Art. No.: CD000434. DOI: 10.1002/14651858.CD000434.pub4.
8. Hodgson SA, Mawson SJ, Saxton JM, Stanley D. Rehabilitation of two-part fractures of the neck of the humerus (two-year follow-up). *J Shoulder Elbow Surg* 2007;16(2):143-5.
9. No authors listed. EQ-%D-3L User Guide. http://www.euroqol.org/fileadmin/user_upload/Documenten/PDF/Folders_Flyers/EQ-5D-3L_UserGuide_2015.pdf (date last accessed 11 August 2016)
10. No authors listed. Protocol for an extended follow-up of patients for the ProFHER trial. http://www.nets.nihr.ac.uk/projects/hta/0640453 (date last accessed October 2016)
11. Dolan P, Gudex C, Kind P, Williams A. A social tariff for Euroqol: Results from a UK general population survey. Centre for Health Economics Discussion Paper Sep; 138, 1995. http://www.york.ac.uk/che/pdf/DP138.pdf (date last accessed 24 July 2016).
12. Billingham LJ, Abrams KR, Jones DR. Methods for the analysis of quality-of-life and survival data in health technology assessment. Health Technol Assess 1999;3:1–152
13. No authors listed. National Institute for Health and Care Excellence. Guide to the methods of technology appraisal 2013

http://www.nice.org.uk/article/pmg9/chapter/Foreword (date last accessed 29 September 2015).

14. Peters SA, Bots ML, den Ruijter HM, Palmer MK, Grobbee DE, Crouse JR 3rd, et al. Multiple imputation of missing repeated outcome measurements did not add to linear mixed-effects models. *J Clin Epidemiol* 2012;65(6):686-695.

Table I. Inclusion and exclusion criteria of the PROFHER trial

Inclusion and exclusion criteria
Inclusion criteria • Adults (aged 16 or above) presenting within three weeks of their injury with a radiologically confirmed displaced fracture of the humerus involving the surgical neck. This should include all two part surgical neck fractures; three part (including surgical neck) and four part fractures of proximal humerus (Neer Classification).9-11 It may also include displaced surgical neck fractures that do not meet the exact displacement criteria of the Neer Classification (1 cm or/and 45° angulation of displaced parts) where this reflects an individual surgeon's uncertainty (e.g., whether or not the surgical neck fracture should be treated surgically). **Exclusion criteria** • Associated dislocation of the injured joint of the shoulder • Open fracture • Mentally incompetent patient: unable to understand trial procedure or instructions for rehabilitation; significant mental impairment that would preclude compliance with rehabilitation and treatment advice • Comorbidities precluding surgery/anaesthesia • A clear indication for surgery such as severe soft-tissue compromise requiring surgery/emergency treatment (nerve injury/dysfunction) • Multiple injuries: same limb fractures; other upper limb fractures • Pathological fractures (other than osteoporotic) and terminal illness • Participant not resident in catchment area of trauma centre

Table IIa. Baseline characteristics (demographics) at randomisation and five years follow-up

	All randomised PROFHER patients		**Patients with OSS data at 5 years**	
Characteristic	**Surgery (N = 125)**	**Not Surgery (N = 125)**	**Surgery (N = 76)**	**Not Surgery (N = 73)**
Gender				
Male, n (%)	28 (22.4)	30 (24.0)	19 (25.0)	15 (20.6)
Female, n (%)	97 (77.6)	95 (76.0)	57 (75.0)	58 (79.5)
Age (years)				
Mean (SD)	66.60 (11.80)	65.43 (12.09)	65.80 (10.12)	65.33 (11.35)
Median (min, max)	67.42 (27.04,	66.12 (24.63,	65.69 (37.09,	65.37 (31.33,

	All randomised PROFHER patients		Patients with OSS data at 5 years	
Characteristic	**Surgery (N = 125)**	**Not Surgery (N = 125)**	**Surgery (N = 76)**	**Not Surgery (N = 73)**
	92.04)	89.02)	87.76)	84.56)
Age (group)				
Less than 65 years, n (%)	51 (40.8)	57 (45.6)	34 (44.7)	36 (49.3)
Equal or greater than 65 years, n (%)	74 (59.2)	68 (54.4)	42 (55.3)	37 (50.7)
Ethnicity				
White, n (%)	124 (99.2)	125 (100.0)	75 (98.7)	73 (100.0)
Black, n (%)	1 (0.8)	0 (0.0)	1 (1.3)	0 (0.0)
Education				
No formal qualifications, n (%)	66 (52.8)	68 (54.4)	35 (46.1)	35 (48.0)
Some qualifications but no degree, n (%)	47 (37.6)	43 (34.4)	34 (44.7)	25 (34.3)
Degree or higher, n (%)	12 (9.6)	14 (11.2)	7 (9.2)	13 (17.8)
Employment				
Part-time, n (%)	12 (9.6)	7 (5.6)	10 (13.2)	5 (6.9)
Full-time, n (%)	17 (13.6)	22 (17.6)	12 (15.8)	15 (20.6)
Self-employed, n (%)	1 (0.8)	3 (22.4)	0 (0.0)	3 (4.1)
Retired, n (%)	78 (62.4)	82 (65.6)	43 (56.6)	45 (61.6)
Not employed but seeking work, n (%)	3 (2.4)	1 (0.8)	2 (2.6)	1 (1.4)
Other, n (%)	12 (9.6)	9 (7.2)	9 (11.8)	3 (4.1)
Missing, n (%)	2 (1.6)	1 (0.8)	0 (0.0)	1 (0.7)
Diabetes				
Yes, n (%)	18 (14.4)	13 (10.4)	8 (10.5)	8 (11.0)
No, n (%)	106 (84.8)	111 (88.8)	67 (88.2)	64 (87.7)
Missing, n (%)	1 (0.8)	1 (0.8)	1 (1.3)	1 (1.4)
Smoking status				
Yes, n (%)	24 (19.2)	40 (32.0)	13 (17.1)	16 (21.9)
No, n (%)	96 (76.8)	81 (64.8)	61 (80.3)	55 (75.3)
Missing, n (%)	5 (4.0)	4 (3.2)	2 (2.6)	2 (2.7)
Steroid use				
Yes, n (%)	6 (4.8)	7 (5.6)	4 (5.3)	6 (6.9)
No, n (%)	118 (94.4)	116 (92.8)	72 (94.7)	67 (91.8)
Missing, n (%)	1 (0.8)	2 (1.6)	0 (0.0)	1 (1.4)
Health Status (EQ-5D Index)				

	All randomised PROFHER patients		Patients with OSS data at 5 years	
Characteristic	**Surgery (N = 125)**	**Not Surgery (N = 125)**	**Surgery (N = 76)**	**Not Surgery (N = 73)**
N	123	121	75	70
Mean (SD)	0.43 (0.37)	0.38 (0.37)	0.43 (0.36)	0.34 (0.36)
Median (min, max)	0.59 (-0.36,1)	0.26 (-0.35,1)	0.59 (-0.35, 1)	0.24 (-0.35, 1)

Table IIb. Baseline characteristics (fracture data) at randomisation and five years follow-up

	All randomised PROFHER patients		Patients with OSS data at 5 years	
Characteristic	**Surgery (N = 125)**	**Not surgery (N = 125)**	**Surgery (N = 76)**	**Not surgery (N = 73)**
Time since injury (days)				
Mean (SD)	5.78 (4.90)	5.69 (4.89)	5.93 (5.17)	5.82 (4.59)
Median (min, max)	4 (0, 19)	4 (0, 21)	4.5 (0, 19)	4 (0, 18)
Affected shoulder				
Left, n (%)	57 (45.6)	68 (54.4)	32 (42.1)	40 (54.8)
Right, n (%)	68 (54.4)	57 (45.6)	44 (57.9)	33 (45.2)
Tuberosity involvement				
Yes, n (%)	99 (79.2)	94 (75.2)	58 (76.3)	58 (79.5)
No, n (%)	26 (20.8)	31 (24.8)	18 (23.7)	15 (20.6)
Tuberosity involvement (detail)				
Tuberosity not involved or missing, n (%)	26 (20.8)	31 (24.8)	18 (23.7)	15 (20.6)
Greater tuberosity, n (%)	58 (46.4)	61 (48.8)	34 (44.7)	36 (49.3)
Lesser tuberosity, n (%)	7 (5.6)	3 (2.4)	4 (5.3)	1 (1.4)
Greater and lesser tuberosity, n (%)	34 (20.8)	30 (24.0)	20 (26.3)	21 (28.8)
Fractures in the past 10 years				
Yes, n (%)	33 (26.4)	33 (26.4)	19 (25.0)	19 (26.0)
No, n (%)	92 (73.6)	90 (72.0)	57 (75.0)	53 (72.6)
Missing, n (%)	0 (0.0)	2 (1.6)	0 (0.0)	1 (1.4)
Previous surgery for fractures				

	All randomised PROFHER patients		Patients with OSS data at 5 years	
Characteristic	**Surgery (N = 125)**	**Not surgery (N = 125)**	**Surgery (N = 76)**	**Not surgery (N = 73)**
Yes, n (%)	8 (6.4)	12 (9.6)	3 (4.0)	9 (12.3)
No, n (%)	23 (18.4)	21 (16.8)	14 (18.4)	10 (13.7)
Missing, n (%)	2 (1.6)	0 (0.0)	2 (2.6)	0 (0.0)
No previous fractures, n (%)	92 (73.6)	92 (73.6)	57 (75.0)	54 (74.0)
Shoulder on dominant side				
Yes, n (%)	67 (53.6)	61 (48.8)	40 (52.6)	36 (49.3)
No, n (%)	56 (44.8)	62 (49.6)	34 (44.7)	35 (48.0)
Missing, n (%)	2 (1.6)	2 (1.6)	2 (2.6)	2 (2.7)
Injury mechanism				
Fall or trip from standing height or less, n (%)	90 (72.0)	96 (76.8)	55 (72.4)	58 (79.5)
Fall downstairs/steps or from a height, n (%)	18 (14.4)	17 (13.6)	11 (14.5)	9 (12.3)
Other, n (%)	15 (12.2)	9 (7.2)	8 (10.5)	5 (6.9)
Missing, n (%)	2 (1.6)	3 (2.4)	2 (2.6)	1 (1.4)

Table III. Difference in mean OSS over time by treatment group; extended primary analysis model*

	Surgery Mean (95% CI)	Not surgery Mean (95% CI)	Difference (95% CI)	p-value
No. patients†‡	114	117	231	
6 months‡	37.84 (35.93, 39.65)	35.59 (33.62 to 37.45)	2.25 (-0.07 to 4.57)	0.058
1 year‡	39.23 (37.38, 40.99)	38.80 (36.99 to 40.53)	0.42 (-1.78 to 2.63)	0.706
2 years‡	40.11 (38.24, 41.90)	40.40 (38.59 to 42.13)	-0.29 (-2.53 to 1.95)	0.800
No. patients†	114	117	231	
3 years	40.53 (38.73 to 42.25)	40.36 (38.58 to 42.06)	0.17 (-2.02 to 2.35)	0.880
4 years	40.87 (39.04 to 42.62)	41.45 (39.67 to 43.16)	-0.58 (-2.81 to 1.64)	0.607
5 years	40.89 (39.99 to 42.70)	41.98 (40.14 to 43.74)	-1.09 (-3.41 to 1.23)	0.356

* Multilevel model of Oxford Shoulder Score (OSS, score range 0 to 48, higher scores indicate better outcomes) adjusted for treatment group, time (6, 12, 24, 36, 48 and 60 months), group × time interaction, baseline EQ-5D index, gender, age group (< 65 years/≥ 65 years) and tuberosity involvement at baseline (yes/no)

† Number of patients included in the analyses (complete baseline characteristics and valid OSS score for at least one follow-up, same for primary and long-term analyses)

‡ Rows obtained from original ProFHER trial analysis

Table IV. Difference in mean OSS over time by treatment group; data derived by multiple imputation*

	Surgery Mean (95% CI)	Not surgery Mean (95% CI)	Difference (95% CI)	p-value
No. patients†	125	125	250	
6 months†	37.96 (36.07 to 39.76)	35.67 (33.71 to 37.54)	2.28 (-0.04 to 4.61)	0.054
1 year†	39.29 (37.48 to 41.03)	38.84 (37.03 to 40.56)	0.46 (-1.72 to 2.64)	0.680
2 years†	40.18 (38.36 to 41.93)	40.54 (38.72 to 42.28)	-0.36 (-2.58 to 1.87)	0.752
No. patients	125	125	250	
3 years	40.59 (38.79 to 42.31)	40.22 (38.46 to 41.91)	0.36 (-1.86 to 2.58)	0.748
4 years	40.97 (39.14 to 42.71)	41.52 (39.84 to 43.13)	-0.55 (-5.64 to 1.53)	0.602
5 years	40.96 (39.10 to 42.75)	41.90 (40.13 to 43.59)	-0.93 (-3.19 to 1.32)	0.416

* Missing OSS and covariate data derived by multiple imputation. Multilevel model adjusted for treatment group, time (6, 12, 24, 36, 48 and 60 months), group × time interaction, baseline EQ-5D index, gender, age group (< 65 years/≥ 65 years) and tuberosity involvement at baseline (yes/no)

† Rows obtained from original ProFHER trial analysis

Table V. Further fractures by treatment arm

	Surgery			Not surgery			Total		
	M0-24*	M24-60†	Total	M0-24	M24-60	Total	M0-24	M24-60	Total
	N	N	N	N	N	N	N	N	N
Shoulder/ upper arm	1	5	6	2	4	6	3	9	12
Wrist	3	6	9	5	7	12	8	13	21
Hip	3	1	4	7	2	9	10	3	13
Spine	1	0	1	1	10	11	2	10	12
Elbow	0	0	0	1	2	3	1	2	3
Ankle	2	1	3	0	1	1	2	2	4
Other	0	8	8	2	6	8	2	14	16
Total fractures	10	21	31	18	32	50	28	53	81
Total patients	10	12	19	15	21	33	25	33	52

* M0-24: Follow-up up to 2 years; † M24-60: Extended follow-up from 2 to 5 years

Table VI. Difference in mean EQ-5D-3L over time by treatment group derived by multilevel model*

	Surgery Mean (se)	Not Surgery Mean (se)	Difference (95% CI) (Surgery - Not Surgery)
No. patients	123	121	244
3m	0.61 (0.03)	0.60 (0.03)	0.01 (-0.06 to 0.08)
6m	0.66 (0.03)	0.63 (0.03)	0.03 (-0.04 to 0.10)
12 m	0.63 (0.03)	0.66 (0.03)	-0.02 (-0.09 to 0.05)
2 years	0.66 (0.03)	0.66 (0.03)	-0.00 (-0.08 to 0.07)
3 years	0.65 (0.03)	0.63 (0.03)	0.02 (-0.06 to 0.10)
4 years	0.67 (0.03)	0.62 (0.04)	0.05 (-0.04 to 0.14)
5 years	0.65 (0.04)	0.62 (0.04)	0.03 (-0.07 to 0.13)

*Multilevel model for EQ-5D-3L (score range 0 to 1, higher scores indicate better HRQoL) adjusted for treatment allocation, time (3, 6,12, 24, 36, 48 and 60 months), group-time interaction, baseline EQ-5D-3L index, gender, age group and tuberosity involvement at baseline (yes/no). Number of patients included in the analyses (complete baselines characteristics and EQ-5D score for at least one follow-up).

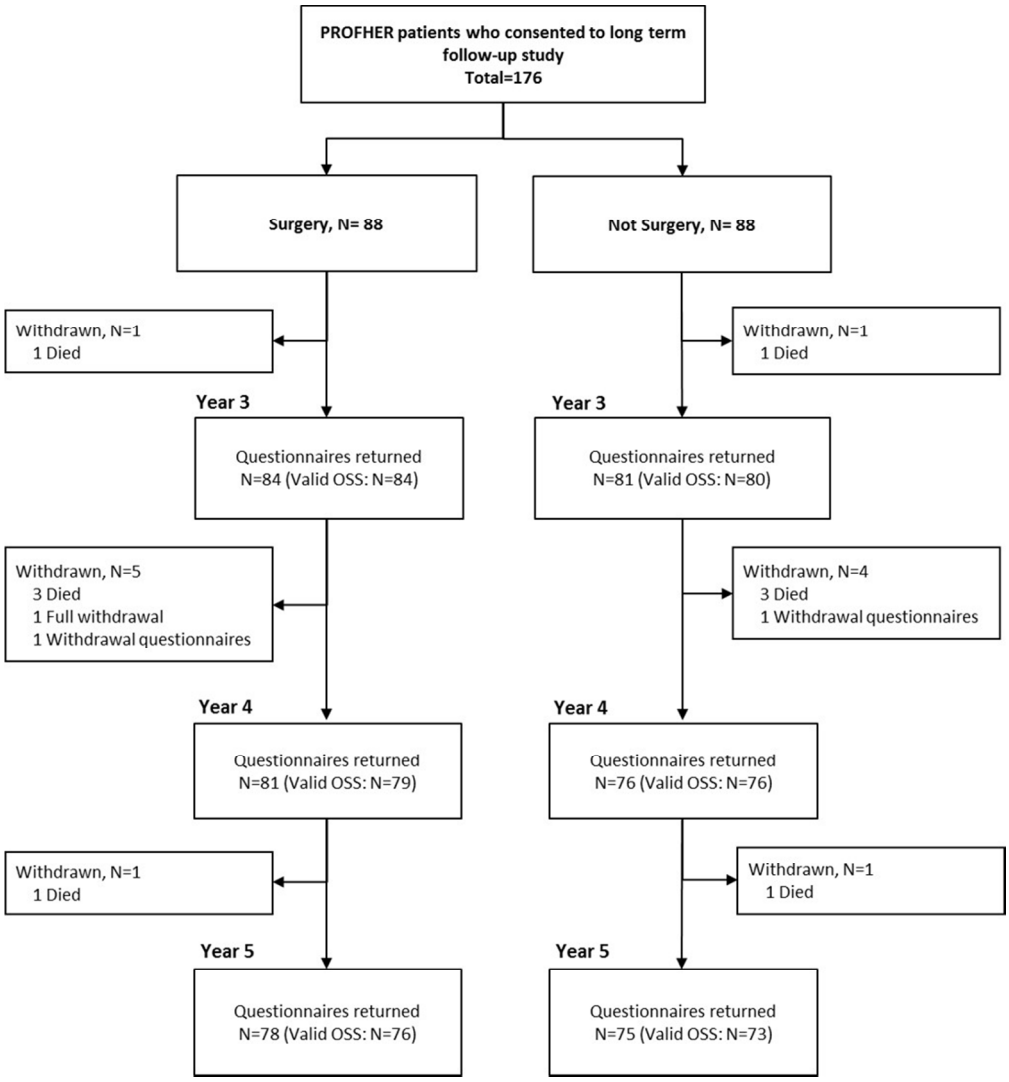
Table VII. HRQoL: Mixed model and multiple imputation sensitivity analyses at each follow-up time up to five years

Follow-up	Mixed model * Difference QALYs (adjusted for covariates) (Surgery - Not surgery) (95% CI) (N = 200) †	Multiple imputation ** Difference QALYs (adjusted for covariates) (Surgery - Not surgery) (95% CI) (N = 250)
3 months	-0.001 (-0.02 to 0.02)	-0.002 (-0.03 to 0.02)
6 months	0.028 (-0.03 to 0.04)	-0.000 (-0.03 to 0.03)
1 year	-0.004 (-0.06 to 0.05)	-0.004 (-0.06 to 0.05)
2 years	-0.031 (-0.15 to 0.09)	-0.024 (-0.15 to 0.10)
3 years	-0.061 (-0.25 to 0.12)	-0.034 (-0.23 to 0.16)
4 years	-0.063 (-0.32 to 0.19)	-0.027 (-0.29 to 0.24)
5 years	-0.042 (-0.36 to 0.28)	-0.013 (-0.35 to 0.32)

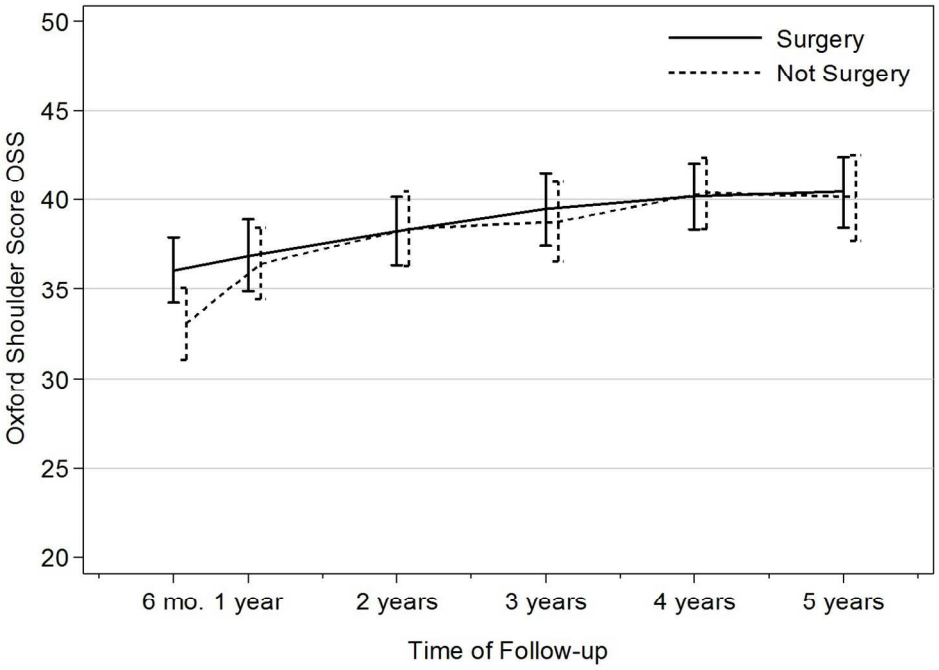
* Multilevel model for QALYs adjusted for treatment allocation, time (3,6,12, 24, 36, 48 and 60 months), group-time interaction, baseline utility, gender, age group and tuberosity involvement at baseline (yes/no).

** Missing EQ-5D and covariate data derived by multiple imputation. Multilevel model adjusted for treatment group, time (6, 12, 24, 36, 48 and 60 months), group × time interaction, baseline utility, gender, age group (< 65 years/≥ 65 years) and tuberosity involvement at baseline (yes/no)

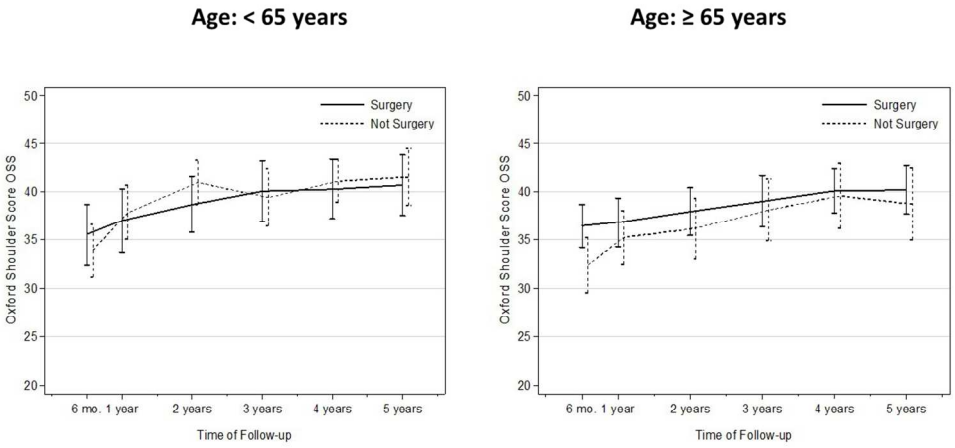
† Number of patients included in the analyses (complete baselines characteristics and QALYs score for at least one follow-up): 106 surgery; 94 not surgery



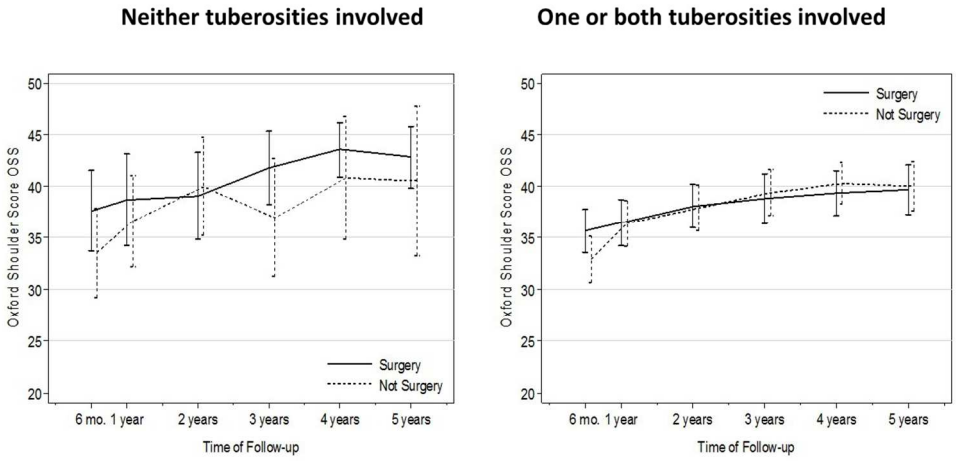
Participant flow diagram
Figure I
173x185mm (150 x 150 DPI)



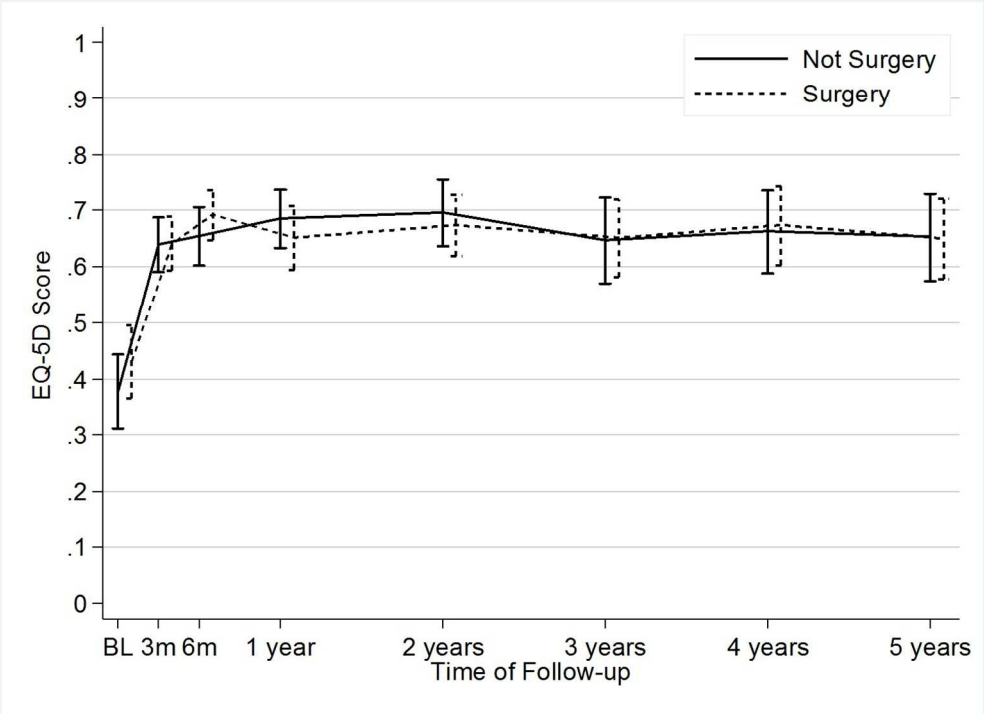
Unadjusted OSS Scores by allocated treatment (patients with available OSS only)
Figure II
259x190mm (150 x 150 DPI)



Unadjusted OSS Scores by allocation and age group (patients with available OSS only)
Figure III
338x190mm (96 x 96 DPI)



Unadjusted OSS Scores by allocation and tuberosity involvement group (patients with available OSS only)
Figure IV
338x190mm (96 x 96 DPI)



Mean EQ-5D-3L scores at baseline (BL) and follow-up points to five years
Figure V
259x190mm (150 x 150 DPI)