Measuring the patient perspective on latissimus dorsi donor site outcomes following breast reconstruction.

Author list: John P Browne, PhD[1]; Ranjeet Jeevan, MRCS[2]; Andrea L Pusic, MD, MHS[3]; Anne F Klassen, DPhil[4]; Carmel Gulliver-Clarke, PhD[5]; Jerome Pereira, FRCS[6]; Christopher M Caddy, FRCS(Plast)[7]; Stefan J Cano, PhD[8]

1. John P. Browne, PhD. Department of Epidemiology and Public Health, University College Cork; Cork, Ireland.
2. Ranjeet Jeevan, MRCS. Clinical Effectiveness Unit, Royal College of Surgeons of England; London, UK.
3. Andrea L. Pusic, MD, MHS. Division of Plastic and Reconstructive Surgery, Memorial Sloan-Kettering Cancer Center; New York, New York.
4. Anne F. Klassen, DPhil. McMaster University; Hamilton, Ontario, Canada.
5. Carmel Gulliver-Clarke, PhD. Integrated Breast Service, Western Sussex Hospitals NHS Foundation Trust; Worthing, UK.
6. Jerome Pereira, FRCS. Department of General Surgery, James Paget University Hospitals NHS Foundation Trust; Great Yarmouth, UK.
7. Christopher M Caddy, FRCS(Plast). Department of Plastic Surgery, Sheffield Teaching Hospitals NHS Foundation Trust; Sheffield, UK.
8. Stefan J. Cano, PhD. Modus Outcomes; Letchworth Garden City, UK.

The work should be attributed to the Clinical Effectiveness Unit of the Royal College of Surgeons of England.

Corresponding author

Professor John Browne, PhD

Department of Epidemiology and Public Health, University College Cork.

Western Gateway Building, Western Rd, Cork, Ireland.

Tel: +353 21 420 5510

Fax: +353 21 420 5469

E-mail: j.browne@ucc.ie

SUMMARY

Background: There is little evidence about the long-term donor site outcome of latissimus dorsi breast reconstruction and no patient-reported outcome measures designed specifically for the procedure.

Methods: A prospective cohort of breast cancer patients having latissimus dorsi reconstruction after a mastectomy were recruited from 270 hospitals in the United Kingdom. An 18-month follow up questionnaire containing two novel scales was sent to consenting patients. The prevalence of aesthetic and functional morbidity at the donor site was described. The two new scales were refined using the Rasch measurement model and subsequently validated.

Results: 1,096 women completed the new scales. 78% of patients reported that no back appearance issues had bothered them “most of the time” or “all of the time” in the past two weeks. The equivalent figure for functional morbidity was 60%. Four items were eliminated following initial psychometric testing. This produced an 8-item Back Appearance scale and an 11-item Back and Shoulder Function scale. Both scales showed adequate fit to the Rasch measurement model. Higher levels of aesthetic and functional bother were observed for completely autologous procedures versus those where latissimus dorsi reconstruction was used to cover an implant ($p < 0.05$). Higher levels of aesthetic bother were observed in women who had suffered a perioperative complication at the donor site ($p = 0.003$).

Conclusion: These results can inform patients of the morbidity associated with latissimus dorsi reconstruction. The new scales can be used to compare groups undergoing different variations of the procedure and to monitor individual patients.

Key words: Latissimus Dorsi Myocutaneous Flap; Breast reconstruction; Outcome Measures; Psychometrics.

Introduction

Latissimus dorsi (LD) breast reconstruction involves rotating a flap of muscle, skin, fat and blood vessels from the upper back to the mastectomy site. There are two main types of LD reconstruction. The first involves the use of LD tissue to cover an implant. The second involves a pedicled flap of completely autologous tissue and is commonly known as an extended LD reconstruction. The largest study of LD reconstruction to date remains the UK National Mastectomy and Breast Reconstruction Audit, which recruited patients in 2008 and 2009. This found that both types of LD reconstruction were associated with higher patient-reported breast appearance scores than implant-only procedures, but slightly worse breast appearance scores than reconstruction with abdominal tissue.[1] Morbidity at the donor site must also be considered when comparing different types of breast reconstruction. The LD muscle can be functionally impaired when it is used in a breast reconstruction, pulling the arm back into the body, and turning it inward. There may also be aesthetic damage to the back which can be exacerbated by wound infection and skin necrosis. Two systematic reviews, both published in 2014, have synthesised the available literature on functional outcomes.[3,4] The reviews, which were limited by a reliance on small, single-centre studies, found that LD procedures lead to measurable reductions in shoulder and upper back strength and function in the short term. There was insufficient evidence to provide clear guidance on the extent of functional morbidity beyond six months. There is little published literature on aesthetic outcomes at the LD donor site. This may be due to an untested assumption that women are unconcerned by the appearance of their back because it is rarely visible to them. For both functional and aesthetic outcomes there are no patient-reported outcome measures that have been developed specifically for LD patients. It is possible, therefore, that the measures used in previous studies have lacked content validity.

In this study, we describe the long-term donor site morbidity arising from LD breast reconstruction after mastectomy in a large prospective cohort study using the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines. The psychometric properties of two new measurement scales developed specifically for LD reconstruction patients are also described.

Methods

The data presented in this paper are from the National Mastectomy and Breast Reconstruction Audit, which recruited patients between 1 January 2008 and 31 March 2009 from 270 public and private hospitals in the United Kingdom.[5] Data on surgical procedures and patient characteristics were prospectively recorded for women aged 16 years and over with a diagnosis of invasive carcinoma of the breast, or ductal carcinoma in-situ, undergoing mastectomy with immediate reconstruction or primary delayed reconstruction following a previous mastectomy. Written consent to participate in a follow-up survey was also obtained.

Questionnaires were sent to the home address of consenting patients 18 months after surgery and included two new scales designed to evaluate the aesthetic and functional outcomes of LD flap reconstruction. The scales are part of the BREAST-Q family of patient-reported outcome measures.[6,7] They were developed in qualitative work with patients who had undergone LD flap reconstruction in the United States, and pre-tested with English breast cancer patients to ensure acceptability. The resulting Back Appearance (9 items) and Back and Shoulder Function scales (14 items) asked patients to record how often in the past

two weeks they had been bothered by a set of problems, using five response options: none of the time, a little of the time, some of the time, most of the time and all of the time. Endorsement frequencies were used to quantify the morbidity of an LD procedure.

The new scales were tested using two distinct measurement paradigms. The dominant paradigm in quality of life measurement has traditionally been Classical Test Theory (CTT).[8,9] In CTT, observed patient responses are considered equal to a theoretical true score plus random error. The observed score on a scale is assumed to be a random variable which produces a bell-shaped curve around the true score. The error score is taken to have a value of zero as positive and negative errors cancel each other. A major difficulty with CTT is the need to measure repeatedly in order to reduce the size of random errors around individual patient scores. In practice, CTT is rarely used to measure individual patients, and error is dealt with by focusing on groups of patients only. CTT also does not evaluate the extent to which scales have interval level properties and this may lead to inappropriate usage when scores are analysed. Although limited, CTT provides a useful measure of the group-level reliability of a scale through a statistic known as 'Cronbach's alpha'.[10] This produces a coefficient which varies between 0 and 1 where higher values indicate higher levels of internal consistency among scale items and, by extension, higher levels of reliability. A second CTT analysis was performed to assess whether the items in each scale measure a common underlying construct. This was evaluated by examining the correlation between each item and scale score computed from the remaining items in that scale. Corrected item-total correlation ≥ 0.30 were taken as sufficient to satisfy this criterion.[11]

In a second stage of scale development, RUMM2030 software was used to test the extent to which patient responses fit the Rasch measurement model.[12] The Rasch model can be used to develop scales with invariant, interval level measurement properties. The answers to individual questions in a health outcome measure are usually summed to produce a total score, but to do this one must have measurement invariance. This requires that the relative location of any two persons on a scale is independent of the items used and conversely the relative location of any two items on the continuum is independent of the person on which they are measured.[13] If a scale is invariant one can treat the interval between, for example, scores of 50 and 60 on a 100 point scale, as equivalent to any other 10 point interval on the scale when performing a statistical test. This greatly increases the range of analyses that can be used with the data and allows for the precise measurement of individual patients. This is a major advantage over most existing measures which can only be used at the group level.

Scale development and testing

Scale development and testing was carried out in three stages. In stage one, instances where item responses were inter-dependent were identified, as this violates the requirement of item invariance. Inter-item residual correlations greater than 0.4 were considered for elimination, and qualitative considerations such as wording or causal relationship were used to select which item in the correlated pair should be eliminated.

In stage two, the extent to which each scale covered the continuum of severity and discriminated between different levels of severity was assessed. Person-item threshold distributions and item 'locations' were used to examine the extent to which each scale was properly aligned with patient burden. Location is measured on the logit scale and lower

scores represent items that are more likely to be associated with a poorer outcome. A further test, known as the Person Separation Index, evaluated the extent to which the scales reliably discriminated between patients with different outcomes. A minimum value of 0.7 is recommended.

In stage three, items on each scale were examined for adherence to the Rasch model. The model posits that the probability of a person responding in a certain way to an item in a health outcome scale is a logistic function of the difference between that person's health status and the item's severity.[14] First, the standardised residuals for each item were calculated to estimate the extent to which the observed variance deviated from the Rasch measurement model. Fit residual values between +/−2.5 demonstrate adequate fit. A separate Chi-Square test assessed whether responses to each item were invariant across the continuum of difficulty. The *p*-value for the test was Bonferroni adjusted, and reduced sample sizes of 500 were used to avoid the risk of overdetecting misfit. Item characteristic curves were examined to determine the severity of misfit when both of these tests were failed. Second, the hierarchical order of response options was examined to ensure that it was in accordance with the underlying latent variable in question. This test compares the difficulty threshold in logits for each response option. Third, the extent to which patients undergoing immediate reconstruction answered items differently to patients undergoing delayed reconstruction was estimated. This test was performed because of a concern that delayed reconstruction patients might, because of their prolonged adjustment to the aesthetic and functional impact of a mastectomy, answer questions about LD morbidity differently to patients undergoing an immediate reconstruction. This problem, known as differential item functioning, was evaluated with an analysis of variance of the standardised

response residuals for each item between surgical groups. A Bonferroni adjusted *p*-value was again used to determine statistical significance.

Validation

We posited that a properly constructed scale of LD morbidity would find higher levels of impairment in patients undergoing a completely autologous procedure versus those who were receiving LD reconstruction to cover an implant. This is because the autologous procedure is more invasive with respect to harvesting material around the LD muscle. We also posited that women who had suffered a perioperative complication at the LD donor site would have worse outcomes than women who had not. Clinicians recorded all donor site complications requiring some form of treatment during the hospital admission. These comprised wound infection requiring intravenous antibiotics or surgical debridement, wound dehiscence requiring re-closure, skin flap necrosis requiring surgical debridement, and haematoma or seroma at the donor site requiring aspiration or drainage. To perform these analyses the overall score on each measure for each patient was transformed from a logit scale to a 0-100 scale, where higher scores represent a better outcome. The outcomes of different groups were then compared using linear multiple regression models, adjusting for baseline differences in prognostic variables (age, fitness for surgery[15] and ethnicity) that were significantly associated with scale scores at the 0.05 significance level. When performing these comparisons we defined the minimum clinically important difference as 0.5 of a standard deviation.[16]

At the time of the study national cancer audits were exempt from obtaining approval from the National Research Ethics Service. Approval to prospectively collect patient identifiable

data for analysis and reporting was obtained from the Patient Information Advisory Group under Section 60 of the Health and Social Care Act 2001.

Results

3,389 patients underwent a mastectomy with immediate reconstruction and 1,731 underwent a delayed reconstruction. 1,579 (47%) of the immediate reconstruction patients and 790 (46%) of the delayed reconstruction patients had a LD procedure. 1,551 of the women in this combined group were invited to take part in the follow-up survey. The remaining women were not invited, largely because of problems with the consent process in some hospitals. 1,383 (89%) of the invited women consented to follow-up and 1,109 (80%) of these women returned a questionnaire at 18 months after their surgery. 13 patients did not complete the new scales leaving a final sample of 1,096. Patient characteristics are shown in Table 1.

The median patient age was 52 years (inter-quartile range = 14 years). 69% underwent an immediate reconstruction. Risk factors known to be associated with poor surgical outcomes were restricted to a minority of patients. 72% had the highest level of fitness for surgery, 89% were non-smokers and 83% had a body mass index less than 30. Slightly less than half the sample (46%) had their LD reconstruction to cover an implant. 76 of the 1,096 women who completed an 18-month questionnaire (6.9%) suffered a donor site complication. Reassuringly, this was similar to the proportion seen in the 2,369 patients who were eligible for participation in the study (8.7%). This is one indication that the sub-group who completed a follow-up questionnaire are generally representative of the larger group of eligible patients.

Aesthetic morbidity was rare (Table 2). 32% of the sample reported that they had not been bothered by any back appearance issues at any time in the past two weeks. 78% reported that none of the nine items in the scale bothered them most or all of the time. The most commonly reported problems related to clothing restrictions: either having to wear certain clothes to hide a back scar (12%) or not being able to wear certain clothes (14%).

Back and shoulder morbidity was slightly more frequent (Table 3) and only 8% of patients reported that they had had no functional bother on any item at any time in the past 2 weeks. However, severe morbidity was confined to a minority and 60% reported that none of the 14 items in the scale were bothersome most or all of the time. The items where patients most frequently experienced bother most, or all of the time, were carrying heavy objects (23%), lifting heavy objects (22%) and reaching for objects (21%).

Psychometric analysis of the Back Appearance scale

The 9-item version of the Back Appearance showed good internal consistency (Cronbach's alpha coefficient = 0.95) and all items had a high correlation with corrected total scores (range = 0.74 to 0.86). A residual correlation of 0.63 was observed between item 8 ('Wear certain clothes to hide back scar') and item 9 ('Not being able to wear certain clothes'). We eliminated item 9 because the wording used a double negative (being bothered by not being able to do something), which might confuse some patients. The 8-item version of the Back Appearance scale also had high internal consistency (Cronbach's alpha coefficient = 0.95) and all items continued to be highly associated with the underlying construct (corrected item-total correlation range = 0.75 to 0.86).

Figure 1 shows the person-item threshold distribution for the 8-item version of the Back Appearance Scale. The histogram above the x-axis represents the distribution of patients and bars to the right of the scale represent patients with lower levels of aesthetic morbidity. The histogram below the x-axis represents the severity of items, and bars to the left of the scale represent clinical problems that are more severe. The green curved line is the information plot and can be interpreted as the point where the measure has the most power to discriminate between patients with different levels of aesthetic morbidity. Figure 1 shows that the scale as a whole, and individual items, are well aligned with the burden reported by patients with at least some aesthetic bother from their surgery, but provide less coverage of patients with very mild, or no bother.

Table 4 shows the individual item locations for the 8-item Back Appearance scale. Item 5 ('Location of your back scar') is the most severe item on the scale and item 8 ('Wear certain clothes to hide back scar') is the least severe. This implies that patients with the worst aesthetic outcomes are likely to report experiencing the full range of issues covered by the scale, up to and including the location of their back scar. Conversely, patients with the best overall outcomes are only likely to be bothered by the need to choose certain clothes, or to report no problems at all. The Person Separation Index was acceptable (0.80).

Six items had variance that did not demonstrate ideal fit with the Rasch model but none performed inconsistently across ten class intervals of difficulty (Bonferroni adjusted significance threshold = 0.00125). There were no instances of threshold disordering or differential item functioning.

The mean total score on the new Back Appearance scale was 76.7 (SD = 22.0). This implies a minimum clinically important difference of 11 points. Patients undergoing a completely autologous LD procedure (mean = 75.3) had significantly (adjusted mean difference = -3.4; 95% CI, -6.0 to -0.7; *p* = 0.01) worse scores on the Back Appearance scale than those undergoing the procedure to cover an implant (mean = 78.4). There was a much larger (adjusted mean difference = -8.2; 95% CI, -13.5 to -2.9; *p* = 0.003) difference between women who had suffered a donor site complication (mean = 68.6) and those who had not (mean = 77.3), again in the hypothesised direction. In both instances these differences were less than our predefined minimum clinically important threshold.

Psychometric analysis of the Back and Shoulder Function scale

The 14-item version of the Back and Shoulder Function scale had good internal consistency (Cronbach's alpha coefficient = 0.95) and all items were highly correlated with the underlying construct (corrected item-total correlation range = 0.65 to 0.82).

High residual correlations between items 1 and 3 ('Back pain' and 'An aching feeling in your back area', r = 0.59), items 2 and 4 ('Shoulder pain' and 'An aching feeling in your shoulder area', r = 0.69) and items 9 and 10 ('Difficulty lifting heavy objects' and 'Difficulty carrying heavy objects', r = 0.85) on the Back and Shoulder Function scale indicated a violation of the assumption of invariance that could be removed by eliminating one item from each pair. Items 3 and 4, which referred to 'an aching feeling' were eliminated as it was felt that items 1 and 2, which referred to 'pain' alone, were clearer for patients. Item 9 (lifting) was also eliminated as it was considered to be prior in the causal pathway to carrying heavy objects

and therefore did not measure the ultimate functional goal. The 11-item version of the Back and Shoulder Function scale also had high internal consistency (Cronbach's alpha coefficient = 0.94) and all items continued to be highly associated with the underlying construct (corrected item-total correlation range = 0.61 to 0.83).

The 11-item Back and Shoulder Function scale was well targeted at patients who reported average or high levels of functional morbidity but poorly targeted at those with few or no functional problems (Figure 2). Item 5 ('Shoulder stiffness') is the most severe item on the scale and item 10 ('Difficulty carrying heavy objects') is the least severe (Table 5).

The Person Separation Index was acceptable (0.86). Eleven of the 14 items had variance that did not fit with the expectations of the Rasch model but only one of these (item 9) performed inconsistently across 10 class intervals of difficulty (Chi-square = 28.53; p = 0.00078; Bonferroni adjusted significance threshold = 0.00091). Inspection of the item characteristic curve for this item showed little evidence of a misfit between observed and expected scores across different levels of difficulty (Figure 3). There were no instances of threshold disordering or differential item functioning.

The mean total score on the new Back and Shoulder Function scale was 66.3 (SD = 18.3). This implies a minimum clinically important difference of 9.15 points. Patients undergoing a completely autologous procedure (mean = 63.3) had significantly (adjusted mean difference = -2.4; 95% CI, -4.6 to -0.2; p = 0.04) worse scores on the Back and Shoulder Function scale than those undergoing the procedure to cover an implant (mean = 67.4). This difference was not clinically significant according to our predefined threshold for a minimally important

difference. No difference was observed between patients who had suffered a complication and those who had not (p = 0.37).

Discussion

This large national prospective cohort study provides detailed outcome data for 1,096 women undergoing LD reconstruction after mastectomy in 2008 and 2009. The proportion of women who received immediate LD reconstruction in our study is high compared to current practice. A recent analysis of UK practice found that the popularity of immediate LD reconstruction peaked in 2008 and 2009 and has steadily declined since, possibly because of improvements in both implant-only and autologous abdominal techniques.[17]

Two new measures of outcome after LD flap reconstruction were tested for various psychometric properties and met most of the criteria assessed. In our judgement, both the 8-item Back Appearance scale and the 11-item Back and Shoulder Function scale provide enough reliable and valid information about different levels of morbidity to allow for the calculation of summary scores and use at the individual patient level.

Severe aesthetic bother at the LD donor site was rare at 18 months after surgery. Severe functional morbidity was slightly more common but still confined to a minority of patients. This indicates that the short-term functional impairments previously reported [3,4] may diminish over time, but do not completely resolve for some patients. Patients undergoing completely autologous LD procedures had slightly more morbidity on both scales than the less invasive classical procedure. These differences were statistically, but not clinically

significant. These results provide evidence of the validity of the new scales but also indicate that any differences between the two surgical approaches to LD reconstruction are small.

The Back Appearance scale demonstrated that donor site complications have long lasting aesthetic consequences for the minority of women affected. This is an important finding and is consistent with other research on the impact of surgical complications.[18] However, it should be noted that the difference was not clinically significant according to our predefined threshold. The Back and Shoulder Function scale did not detect a similar effect which may be because the complications recorded were largely concerned with damage to the skin surface.

This is the largest study of its kind to date and reflects the experiences of women treated in a wide range of hospital settings in both the immediate and delayed reconstruction context. Outcomes were measured at 18 months after surgery, allowing patients to completely recover from the procedure. The outcome scales used in the study were developed with and for patients undergoing LD surgery and have been tested using modern psychometric methods. Both scales have high levels of completion. The main weakness of the study is the failure to recruit a large proportion of eligible patients. This reflects the logistical problems associated with the administration of patient-reported outcome measures in more than 270 treatment settings simultaneously. However, there is no evidence that the recruited patients differ significantly from those who were not invited to participate.[5]

Clinicians can now communicate the frequency of a range of problems associated with LD breast reconstruction surgery to prospective patients and be confident that this information

is generalizable and covers the full period of recovery. The findings are reassuring: donor site morbidity following LD reconstruction is limited and similar to that seen with alternative reconstructive options such as a TRAM flap transfer.[19] The two scales presented in this paper are available on a licensed basis from Memorial Sloan Kettering Cancer Center (www.breast-q.org).

Conflict of interest statement

Andrea Pusic, Anne Klassen and Stefan Cano are co-developers of the BREAST-Q patient-reported outcome measure which is owned by Memorial Sloan Kettering Cancer Center and the University of British Columbia. They receive a share of license revenues as royalties when the instrument is used in for-profit industry-sponsored clinical trials. None of the other authors have a conflict to declare.

Acknowledgements

This publication is based on data collected by or on behalf of the Healthcare Quality Improvement Partnership, who have no responsibility or liability for the accuracy, currency, reliability and/or correctness of this publication.

We have only been able to report these findings due to the tremendous participation across NHS and independent hospital organisations, the support of the professional bodies and patient groups involved in breast cancer care, and the engagement and participation of women with breast cancer who completed questionnaires to report on their post-operative outcomes, and we would like to thank them for this.

References

1. Jeevan R, Browne JP, Gulliver-Clarke C, et al. Surgical determinants of patient-reported outcomes following post-mastectomy reconstruction in women with breast cancer. Plast Reconstr Surg 2017;139:1036-1045.

2. Cohen W, Mundy L, Ballard T, et al. The BREAST-Q in surgical research: a review of the literature 2009-2015. J Plast Recon Aesth Surg 2016;69:149-62.

3. Lee KT, Mun GH. Plast Reconstr Surg. A systematic review of functional donor-site morbidity after latissimus dorsi muscle transfer. Plast Reconstr Surg 2014;134:303-14.

4. Smith SL. Functional morbidity following latissimus dorsi flap breast reconstruction. J Adv Pract Oncol 2014;5:181-7.

5. Jeevan R, Cromwell DA, Browne JP, et al. Findings of a national comparative audit of mastectomy and breast reconstruction surgery in England. J Plast Recon Aesth Surg 2014;67:1333-44.

6. Pusic A, Klassen A, Scott A, Klok J, Cordeiro P, Cano S. Development of a new patient reported outcome measure for breast surgery: The BREAST-Q. Plast Reconstr Surg 2009;124:345–353.

7. Cano SJ, Klassen AF, Scott AM, et al. The BREAST-Q: further validation in independent clinical samples. Plast Reconstr Surg 2012;129:293-302.

8. Borsboom D. The attack of the psychometricians. Psychometrika 2006;71:425–440.

9. Cano SJ, Hobart JC. The problem with health measurement. Patient Preference and Adherence 2011;5:279–290.

10. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297-334.

11. Ware JEJ, Harris WJ, Gandek B, Rogers BW, Reese PR: MAP-R for windows: multitrait / multi-item analysis program – revised user's guide. Boston, MA, Health Assessment Lab.; 1997.

12. Andrich D, Sheridan B. Rasch RUMM2030: unidimensional models for measurement. Perth: RUMM Laboratory; 2010.

13. Hobart JC, Cano SJ: Improving the evaluation of therapeutic interventions in multiple sclerosis: the role of new psychometric methods. Health Technol Assess 2009;13:12.

14. Mills RJ, Young CA, Pallant JF, Tennant A. Rasch analysis of the Modified Fatigue Impact Scale (MFIS) in multiple sclerosis. J Neurol Neurosurg Psychiatry 2010;81:1049–51.

15. Dripps RD. New classification of physical status. Anesthesiol 1963;24:111.

16. Cohen J. Statistical power analysis for the behavioral sciences (2nd ed.). 1988. Hillsdale, NJ: Lawrence Earlbaum Associates.

17. Leff DR, Bottle A, Mayer E, et al. Trends in immediate postmastectomy breast reconstruction in the United Kingdom. Plast Reconstr Surg Glob Open 2015;3:e507.

18. Pinto A, Faiz O, Davis R, Almoudaris A, Vincegt C. Surgical complications and their impact on patients' psychosocial well-being: a systematic review and meta-analysis. BMJ Open 2016;6:e007224.

19. Jeevan R, Cromwell D, Browne JP, et al. Fourth Annual Report of the National Mastectomy and Breast Reconstruction Audit. London: Information Centre for Health and Social Care. 2011.

Figure legends

Figure 1: Person-item threshold distribution for the 8-item version of the Back Appearance scale.

Figure 2: Person-item threshold distribution for the 11-item Back and Shoulder Function scale.

Figure 3: Item characteristic curve for item 9 of the Back and Shoulder Function scale.

Table 1: Patient characteristics (N = 1,096).

Patient characteristic		N	%
Age (years)	18-50	507	46.3
	>50	589	53.7
BMI 30+	No	874	82.6
	Yes	184	17.4
	Not known	38	
Smoker	No	954	89.0
	Yes	118	11.0
	Not known	24	
ASA grade	I	770	71.6
	II	293	27.3
	III/IV	12	1.1
	Not known	21	
Contralateral surgery	No	964	88.0
	Yes	132	12.0
Tumour grade	DCIS low grade	27	2.6
	DCIS intermediate grade	54	5.2
	DCIS high grade	186	17.9
	Well differentiated/invasive	82	7.9

	Moderately differentiated/invasive	380	36.6
	Poorly differentiated/invasive	310	29.8
	Not known	57	
Procedure	IR Pedicle with implant	325	29.7
	IR Autologous pedicle	426	38.9
	DR Pedicle with implant	180	16.4
	DR Autologous pedicle	165	15.1

Table 2: Endorsement frequencies and missing data levels: 9-item version of the Back Appearance Scale.

Item	Number choosing each option (% of those who gave a response)					Number (%) of patients who did not give a response
	None of the time	A little of the time	Some of the time	Most of the time	All of the time	
1. How your back looks	568 (52.7)	243 (22.6)	153 (14.2)	75 (7.0)	38 (3.5)	19 (1.7)
2. The shape (contour) of your back	612 (56.9)	213 (19.8)	141 (13.1)	66 (6.1)	44 (4.1)	20 (1.8)
3. The sides of your back not matching	619 (57.8)	207 (19.3)	139 (13.0)	68 (6.4)	37 (3.5)	26 (2.4)
4. How your back scar looks	535 (50.4)	279 (26.3)	129 (12.2)	71 (6.7)	48 (4.5)	34 (3.1)
5. The location of your back scar	729 (68.6)	169 (15.9)	91 (8.6)	44 (4.1)	30 (2.8)	33 (3.0)
6. The length of your back scar	699 (65.6)	198 (18.6)	80 (7.5)	57 (5.4)	31 (2.9)	31 (2.8)
7. How noticeable your back scar is to others	613 (57.7)	250 (23.5)	110 (10.4)	48 (4.5)	42 (3.9)	33 (3.0)
8. Wear certain clothes to hide back scar	606 56.5)	216 (20.2)	115 (10.7)	70 (6.5)	64 (6.0)	26 (2.4)
9. Not being able to wear certain clothes	574 (53.7)	211 (19.8)	129 (12.1)	63 (5.9)	91 (8.5)	28 (2.6)

Table 3: Endorsement frequencies and missing data levels: 14-item version of the Back and Shoulder Function Scale.

Item	Number choosing each option (% of those who gave a response)					Number (%) of patients who did not give a response
	None of the time	A little of the time	Some of the time	Most of the time	All of the time	
1. Back pain	528 (48.9)	227 (21.0)	207 (19.2)	82 (7.6)	36 (3.3)	16 (1.5)
2. Shoulder pain	626 (58.0)	224 (20.8)	150 (13.9)	56 (5.2)	23 (2.1)	17 (1.6)
3. An aching feeling in your back area	457 (42.4)	263 (24.4)	225 (20.9)	93 (8.6)	39 (3.6)	19 (1.7)
4. An aching feeling in your shoulder area	609 (56.4)	241 (22.3)	141 (13.1)	65 (6.0)	23 (2.1)	17 (1.6)
5. Shoulder stiffness	665 (61.3)	231 (21.3)	110 (10.1)	56 (5.2)	23 (2.1)	11 (1.0)
6. Tightness when you stretch your arm	412 (37.8)	302 (27.7)	174 (16.0)	124 (11.4)	77 (7.1)	7 (0.6)
7. A pulling feeling in your back	392 (36.1)	275 (25.3)	209 (19.3)	129 (11.9)	80 (7.4)	11 (1.0)
8. Weakness in your arm	482 (44.3)	277 (25.4)	160 (14.7)	100 (9.2)	70 (6.4)	7 (0.6)
9. Difficulty lifting heavy objects	404 (37.1)	261 (24.0)	188 (17.3)	127 (11.7)	108 (9.9)	8 (0.7)
10. Difficulty carrying heavy objects	393 (36.1)	257 (23.6)	191 (17.5)	133 (12.2)	115 (10.6)	7 (0.6)

11. Difficulty reaching for objects	382 (35.0)	285 (26.1)	197 (18.1)	124 (11.4)	102 (9.4)	6 (0.5)
12. Difficulty doing activities, arms outstretched	542 (49.7)	227 (20.8)	166 (15.2)	87 (8.0)	68 (6.2)	6 (0.5)
13. Difficulty doing activities, arms above head	520 (47.7)	266 (24.4)	157 (14.4)	87 (8.0)	60 (5.5)	6 (0.5)
14. Difficulty, repeat use of shoulder/back muscles	446 (41.3)	273 (25.2)	171 (15.8)	113 (10.5)	78 (7.2)	15 (1.4)

Table 4: Item fit statistics for the 8-item version of the Back Appearance scale, ordered by item location.

Item	Location	Standard error	Fit residual	Chi-square	Probability
5. The location of your back scar	-0.56	0.05	-3.10	21.83	0.005
6. The length of your back scar	-0.43	0.05	-3.22	15.12	0.057
7. How noticeable your back scar is to others	-0.04	0.05	1.11	13.14	0.107
3. The sides of your back not matching	0.05	0.05	2.58	12.32	0.137
1. How your back looks	0.12	0.05	-3.34	17.33	0.027
2. The shape (contour) of your back	0.15	0.05	1.27	8.55	0.382
4. How your back scar looks	0.30	0.05	-4.78	20.81	0.008

8. Wear certain clothes to hide back scar	0.39	0.05	6.07	25.14	<0.001

Table 5: Item fit statistics for the 11-item version of the Back and Shoulder Function scale, ordered by item location.

Item	Location	Standard error	Fit residual	Chi-square	Probability
3. Shoulder stiffness	-0.76	0.04	0.72	1.95	0.992
2. Shoulder pain	-0.64	0.04	2.00	9.76	0.371
1. Back pain	-0.20	0.04	5.41	26.98	<0.001
10. Difficulty doing activities, arms above head	-0.10	0.04	-4.90	18.68	0.028
9. Difficulty doing activities, arms outstretched	-0.07	0.04	-6.56	28.53	<0.001
6. Weakness in your arm	0.05	0.04	-2.91	10.23	0.332
11. Difficulty, repeat use of shoulder/back muscles	0.18	0.04	-5.06	17.74	0.038
4. Tightness when you stretch your arm	0.25	0.04	1.62	10.41	0.318
5. A pulling feeling in your back	0.34	0.04	5.66	22.80	0.007
8. Difficulty reaching for objects	0.45	0.04	-4.68	18.50	0.030
7. Difficulty carrying heavy objects	0.51	0.04	-2.78	11.00	0.276