

# 1        **RECOVERY, RESPONSIVENESS AND INTERPRETABILITY OF PATIENT-** 2        **REPORTED OUTCOME MEASURES AFTER SURGERY FOR DUPUYTREN'S** 3        **DISEASE**

4        JN Rodrigues<sup>1</sup>, W Zhang<sup>2</sup>, BE Scammell<sup>2</sup>, D Davidson<sup>3\*</sup>, S Fullilove<sup>4\*</sup>, I  
5        Chakrabarti<sup>5\*</sup>, PG Russell<sup>6\*</sup>, TRC Davis<sup>2</sup>

6        <sup>1</sup>Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences,  
7        University of Oxford, Botnar Research Centre, Nuffield Orthopaedic Centre, Windmill  
8        Road, Oxford, OX3 7HE

9        <sup>2</sup>Division of Rheumatology, Orthopaedics and Dermatology, University of Nottingham  
10       & Nottingham University Hospitals NHS Trust, Queen's Medical Centre, Derby Road,  
11       Nottingham, UK

12       <sup>3</sup>St John's Hospital at Howden, Howden Road South, Livingston, UK

13       <sup>4</sup>Derriford Hospital, Derriford Road, Plymouth, UK

14       <sup>5</sup>Rotherham General Hospital, Moorgate Road, Rotherham, UK

15       <sup>6</sup>Pulvertaft Hand Centre, Royal Derby Hospital, Uttoxeter New Road, Derby, UK

16       \*These authors contributed equally to this work

## 18       **Corresponding author**

19       JN Rodrigues, Nuffield Department of Orthopaedics, Rheumatology and  
20       Musculoskeletal Sciences, University of Oxford, Botnar Research Centre, Nuffield  
21       Orthopaedic Centre, Windmill Road, Oxford, OX3 7HE  
22       07929296069

23       [j.n.rodrigues@doctors.org.uk](mailto:j.n.rodrigues@doctors.org.uk)

25       **Keywords:** Dupuytren's contracture; Dupuytren's disease; Patient reported outcome  
26       measures; DASH; URAM; responsiveness; interpretability; MID; MIC

29

30

31

## SUMMARY

32 This prospective cohort study investigated the responsiveness and interpretability of  
33 the Disabilities of the Arm, Shoulder and Hand (DASH) and Unité Rhumatologique  
34 des Affections de la Main (URAM) outcome measures for assessing recovery after  
35 fasciectomy and dermofasciectomy for Dupuytren's disease.

36 DASH outcome scores at 1 year were significantly better than at 6 weeks, suggesting  
37 that recovery is not complete by 6 weeks. Of the 101 patients recruited to the DASH  
38 cohort, 71 completed preoperative, 6 week and 1 year postoperative DASH scores;  
39 68 of them completed preoperative and 1 year postoperative DASH scores and an  
40 external anchor question. In the URAM cohort, 30/44 completed the preoperative  
41 and the 1 year postoperative URAM scores and the anchor question. The DASH  
42 score exhibited moderate responsiveness but poor interpretability on receiver  
43 operating characteristic (ROC) curve analysis such that a minimal important change  
44 (MIC) could not be estimated. The URAM score showed acceptable responsiveness,  
45 and an MIC of 10.5 on ROC analysis.

46

47 Level of evidence: II

48

49

50

## INTRODUCTION

51

Evaluating the outcome Dupuytren's disease treatment requires use of an

52

appropriate outcome measure, an appreciation of what constitutes a clinically

53

important change following treatment and consideration of the timing of assessment

54

in relation to recovery.

55

Interpretability is concerned with the changes over time in a score that are

56

meaningful to patients or differences between patients that are relevant. The

57

interpretability of an outcome measure is distinct from its validity. Applying the

58

Consensus-based standards for the selection of health status measurement

59

instruments (COSMIN) to Dupuytren's disease might define validity as whether a

60

single time point measurement reflects hand function at that time and

61

responsiveness as whether changes in scores over time are 'appropriate' (Mokkink

62

et al., 2010). A relevant time period for studying responsiveness is often from before

63

treatment to after treatment, and a large effect size is considered desirable, as this

64

demonstrates that the items in the measure are appropriate for the score to be

65

sensitive to change, such as the change from the preoperative state to the

66

postoperative state. However, treatments do not always work. All patients do not

67

experience marked improvement in deformity or function from the treatment of their

68

Dupuytren's disease, so this is not appropriate. In contrast, interpretability is

69

concerned with defining what a patient considers a meaningful change. The smallest

70

change in state that individual patients consider important is called the minimal

71

important change (MIC). In contrast, the smallest difference in the net change in

72

scores between individual patients that they consider important is called the minimal

73

important difference (MID) (Mokkink et al., 2010). For example, a patient may need

74

experience an improvement in their function from a treatment of at least 15/100

75

points using a patient-reported outcome measure (PROM) to consider the treatment

76

'worthwhile'. This would be the MIC. In contrast, if one patient experiences an

improvement in hand function by 20/100 points on a PROM, but another patient who experiences an improvement of 15/100 believes that they have fared significantly worse, then the MID is 5/100 (i.e.  $20/100 - 15/100$ ). These can be calculated in different ways (Rodrigues et al., 2014), and are likely to be context-specific. Thus, for a particular outcome measure, both the MIC and MID may vary between different conditions and between treatments for one condition (Revicki et al., 2008). A range of MIC estimations exist for hand surgery. These include values for patient-reported outcome measures (PROMs) and for objective measures, such as angular deformity or grip strength (Rodrigues et al., 2014).

The only PROM with published interpretability data in Dupuytren's disease is the Unité Rhumatologique des Affections de la Main (URAM) Dupuytren's disease-specific scale. The MIC for this was estimated in a cohort of patients being treated with needle aponeurotomy (Beaudreuil et al., 2011); the MIC after open surgery (fasciectomy or dermofasciectomy) may be different. The Disabilities of the Arm, Shoulder and Hand (DASH) PROM has the most MIC estimations across all hand surgery conditions (Rodrigues et al., 2014). It is the most widely used PROM in Dupuytren's disease (Ball et al., 2013) but its interpretability in the treatment of Dupuytren's disease has not been studied.

Six weeks following treatment has been used as the time point when the early outcome of surgery for Dupuytren's disease is measured (van Rijssen et al., 2006).

However, full recovery from open surgery, such as fasciectomy and dermofasciectomy, takes longer (Ullah et al., 2009). If a series of time points are studied, then differences between treatments may become apparent e.g. less

invasive treatments have quicker recovery. However, if only one time point is to be studied to summarise the effect of treatment, measuring outcome too early might underestimate the benefit of treatment or may bias comparisons towards treatments that have quicker recovery. It is not logical to determine whether a patient considers a treatment to be worthwhile when they are still rehabilitating from the treatment.

A range of outcome measures have been used to study Dupuytren's disease (Ball et al., 2013), and can be broadly grouped into generic, domain-specific and disease-specific measures (Szabo, 2001). Recently, Dupuytren's disease-specific measures have been developed (Beaudreuil et al., 2011, Mohan et al., 2014); the suitability of the DASH score has been questioned (Packham, 2011). Studies investigating the relationship between the DASH score and angular deformity show poor correlation between them (Degreef et al., 2009, Engstrand et al., 2009, Jerosch-Herold et al., 2011, Zyluk and Jagielski, 2007), but this is only detrimental to the DASH score if angular deformity is the 'gold standard' of patient-centred outcome in Dupuytren's disease. Such an assumption is inappropriate when assessing validity in general (Mokkink et al., 2010); angular deformity is unlikely to be the best standard on which to base patient-centred outcome in Dupuytren's disease (Rodrigues et al., 2014). A recent study has also identified issues with the construct validity of the DASH score (Forget et al., 2014). Despite these concerns, the DASH score has been used extensively. Understanding its interpretability would assist the interpretation of previous studies and clarify its suitability for continued use.

The aim of this prospective cohort study was to investigate the responsiveness and interpretability of the DASH and URAM PROMs following fasciectomy and dermofasciectomy.

## **METHODS**

### **Patient recruitment and data collection**

The data presented in this study were gathered as part of larger service evaluation project. Patient recruitment took place between September 2011 and April 2013. The exclusion criteria were: cognitive impairment preventing informed consent; and refusal to participate. The inclusion criteria were: primary or recurrent Dupuytren's disease in patients awaiting fasciectomy or dermofasciectomy at one UK hand surgery centre (for the cohort study).

Preoperatively patients were recruited at the routine preadmission clinic visit before surgery. Those who were eligible and consented to participate completed the DASH before surgery. Demographic details were also captured, including the patient's age, gender, which digit was treated and whether or not more than one digit on the same hand was being treated in the same procedure (described as "multiple digits treated"). These patients were also sent questionnaires for completion by post at 3 weeks, 6 weeks and 1 year post-operatively. Patients who were scheduled for surgery to the left and right hand at different times during the study recruitment period were eligible for recruitment twice. This happened on four occasions. The URAM scale was published during the study period. Patients recruited later in the cohort (August 2012 onwards) also completed the URAM pre-operatively and by postal questionnaires at 6 week and 1 year post-operatively. They were not sent URAM questionnaires at 3 weeks based on an interim analysis of the DASH scores at this point suggesting that recovery was far from complete, so as to minimise questionnaire burden.

After one year all patients were also posted the Global Rating of Change questionnaire (GRC) to complete (see Appendix). This is a single item questionnaire with 15 response options ranging from +7 ("a very great deal better") through 0 ("about the same") to -7 ("a very great deal worse") (Jaeschke et al., 1989). This was used as the anchor for determining the MIC.

#### **Handling of incomplete questionnaires**

The DASH score is reliable as long as at least 27/30 items are complete (Kennedy et al., 2011). Therefore, all returned questionnaires with  $\geq 27$  completed items were included. So some DASH questionnaires had one, two or three responses missing. Pairwise exclusion was the preferred method for handling unreturned or more incomplete questionnaires. This involves exclusion of the individual in analyses involving the missing piece of data, but including them in all other analyses where

possible. It therefore minimises data exclusion. If required (e.g. for repeated measures ANOVA), listwise exclusion was used, with exclusion of the individual with missing data from all analyses.

As clear guidance for handling incomplete questionnaires was not available for the URAM scores, all URAM scores with any missing entries were excluded.

## **Data handling**

The DASH summary score was calculated using the standard formula provided:

$$\text{DASH} = ((a/b) - 1) \times 25$$

Where “a” is the sum of the scores for the responses completed (each response could be scored between one and five), and “b” is the number of responses the patient completed.

The URAM summary score was calculated by adding the responses to all items.

As the PROM summary scores are virtually continuous scales (the DASH is scored 0-100; the URAM 0-45) and the sample comprised a large number of independent observations, parametric analyses were used to compare them, in keeping with the central limit theorem (Norman, 2010).

Recovery time was analysed by comparing DASH scores at different time points using repeated measures ANOVA with Tukey’s multiple comparison test.

Responsiveness was studied by calculating the effect size, defined as the mean change in score divided by the standard deviation of the baseline (preoperative) scores across the cohort (Kazis et al., 1989). When interpreting the effect sizes, 0.2 to 0.59 was considered small, 0.6 to 0.99 moderate and over 1.0 large (Testa, 1987).

Interpretability was studied using receiver operating characteristic (ROC) curves, as this is the most common method used in hand surgery (Rodrigues et al., 2014).

ROC curves treat an outcome measure (such as the preoperative-postoperative change in DASH or URAM scores) as a diagnostic test, and assess its sensitivity and specificity for detecting “improvement” (Deyo and Centor, 1986). To do this, patients

are categorised as “improved”, “stable” or “worse” using an external criterion, or anchor (Revicki et al., 2008). In this case, the anchor used was the Global Rating of Change (GRC) scale at 1 year (Jaeschke et al., 1989). The ROC curve displays a range of different possible cut off values (e.g. improvement in DASH score of 20/100 compared to 30/100 or 40/100) with the sensitivity and specificity of each for diagnosing improvement plotted to form the curve. The MIC is the point on the ROC curve that combines sensitivity and specificity for identifying improvement the best, which is referred to as Youden’s J index (Youden, 1950). For this, patients with a GRC of +4 to +7 were considered “improved”, those with a GRC of -3 to +3 “stable” and those with a GRC of -7 to -4 “worse”, as previously described (Mintken et al., 2009). ROC curves were generated using Prism 6.0 for Mac OS X (GraphPad® Software, 2012).

## RESULTS

### Patients and procedures

One hundred and one patients were recruited to the study of the DASH. Forty-four were sent URAM questionnaires. The demographics of the cohort are shown in Table 1. The summary of the demographics of the URAM score subgroup was similar to the overall group that completed the DASH scores. Sixty-five of 101 patients completed preoperative, 3 week, 6 week and 1 year postoperative DASH scores. Some of those who failed to return complete 3 week DASHs did return preoperative, 6week and 1-year postoperative scores and the GRC; 71 of 101 were available to study responsiveness at 1 year. Sixty-eight had complete preoperative and 1 year postoperative DASHs and the GRC and so were included in the interpretability analysis. Thirty of 44 completed URAMs and the GRC for the interpretability analysis.

### Recovery



The mean DASH summary score was significantly different between time points ( $p < 0.0001$ , repeated measures ANOVA). The scores at different time points are shown in Table 2. Tukey's multiple comparisons test demonstrated statistically significant differences between all time points, apart from the preoperative and 6 week postoperative scores. Statistically, the DASH score rose between preoperative and 3 week postoperative assessments, indicating worsening in function. The difference between the DASH scores at 6 weeks and 1 year post-operatively was also significant ( $p = 0.001$ ). The developers of the DASH advise that a DASH summary score of 15 or more is consistent with a symptomatic upper limb; the mean DASH summary score only fell below 15 by 1 year (Chart 1). There was no statistically significant difference between scores for fasciectomy and scores for dermofasciectomy at any time point (unpaired t test p values: 0.054, 0.39, 0.69, 0.55 for preoperative, 3-week, 6-week and 1-year time points respectively). As a result the data relating to both procedures were combined for all further analyses.

### **Responsiveness and interpretability**

As functional state was significantly better at 1 year than 6 weeks post-operatively, responsiveness and interpretability analyses were performed using change between preoperative and 1 year post-operative PROMs.

Responses from all 71 patients with adequately completed preoperative and 1-year postoperative DASH questionnaires were included in responsiveness analysis, even if their 3 week or 6 week scores were incomplete. Thirty had appropriately completed URAMs for responsiveness analysis.

Both the DASH and URAM showed significant changes in scores from preoperative to the 1 year postoperative assessments. The DASH exhibited a moderate effect size of 0.58. The effect size for the URAM score was 0.87 (Table 3).

For the 68 patients included in the interpretability analysis, the mean GRC in the DASH cohort was +4.3 (95% confidence intervals (95% CIs): +3.4, +5.2). When DASH outcomes were sub grouped into "worse", "stable" or "improved" (using the

244 GRC scores of -4 to -7, +3 to -3, and +4 to +7 respectively), five patients were worse,  
 245 eleven were stable and 52 were improved. The mean change in DASH score in the  
 246 improved subgroup was 13.0/100; the mean change in DASH score in the stable  
 247 subgroup was 10.8/100. The difference between them (2.2/100 (95% CIs: -13.3 to  
 248 8.9)) was not statistically significant ( $p=0.69$ , unpaired t test), so an MID was not  
 249 identifiable. The ROC curve for the DASH is shown in Chart 2. The area under the  
 250 curve (AUC) was 0.51 (95% CIs: 0.33, 0.69), indicating that the DASH could not  
 251 identify meaningful change in function, as defined by the GRC. Consequently, an  
 252 MIC could not be estimated for the DASH at 1 year after fasciectomy or  
 253 dermofasciectomy.

254 For those who completed the URAM, the mean GRC was +2.9 (95% CIs: +1.2, +4.6).  
 255 When GRC-based sub grouping was performed, 18 were improved, eight were  
 256 stable and four were worse. The mean URAM change in the improved subgroup  
 257 was 11.9/45 (95% CIs: 6.7, 17.0); the mean URAM change in the stable subgroup  
 258 was 3.6 (9.3, -2.0). The difference between the URAM scores for these subgroups  
 259 (8.3, 95% CIs: 0.04, 16.5) was just significant ( $p=0.049$ , unpaired t test) and might  
 260 constitute an MID. The ROC curve for the URAM is shown in Chart 3. The area  
 261 under the curve (AUC) was 0.74 (95% CIs: 0.55, 0.93), and the MIC for the URAM  
 262 for fasciectomy and dermofasciectomy at 1 year (defined as Youden's j index)  
 263 corresponded to an improvement in the URAM of greater than 10.5, which has a  
 264 sensitivity of 56% and a specificity of 88%. The likelihood ratio for an improvement in  
 265 the URAM of 10.5 was 4.4, i.e. a patient with a URAM improvement over 10.5 was  
 266 4.4 times more likely to be 'improved' than a patient whose URAM had improved by  
 267 less than 10.5.

268 The GRC was hypothesised to correlate with 'change in DASH', as both were  
 269 expected to assess change from before surgery to after recovery. Instead, the GRC  
 270 correlated significantly with the 1-year DASH (Pearson's  $r$ : -0.48,  $p<0.0001$ ), and did  
 271 not correlate with the change in DASH (Pearson's  $r$ : -0.22,  $p=0.07$ ) (Chart 4).

272 Similarly, for the URAM, the GRC correlated more closely with the 1 year  
273 assessment (Pearson's  $r$ : -0.68,  $p < 0.0001$ ) than with the change in the URAM  
274 (Pearson's  $r$ : -0.56,  $p = 0.001$ ) (Chart 5).  
275

## DISCUSSION

This study demonstrated that statistically, recovery following fasciectomy or dermofasciectomy takes longer than 6 weeks, as shown by a significant difference between the DASH scores at weeks compared to 1 year post-operatively. However, with the pre-existing absence of interpretability data, the clinical relevance of these differences is not clear, hence the importance of interpretability data in general. The statistically significant difference between the 6 week and 1 year outcomes may be clinically relevant, as the 1 year time point was the only one with a mean DASH score below 15, the threshold above which the patient is considered symptomatic (Kennedy et al., 2011). However, it is unclear from our data if recovery is complete earlier than 1 year or if it continues further.

Fewer patients completed the URAM than the DASH, as it was introduced once the study had been designed. As a result, direct comparison of the two PROMs based on the data in this study is not ideal. However, the responsiveness and interpretability of each of the PROMs can still be considered independently. Using 1 year follow up data, the DASH exhibited moderate responsiveness. The URAM displayed good responsiveness. However, common analyses for responsiveness, such as effect size as used in this study, are only appropriate if all patients studied have undergone a clinically meaningful improvement. Interpretability analyses, which generate MICs and MID, aim to separate those who have experienced meaningful improvement from those who have not. Here, a notable proportion of patients did not experience benefit, or even experienced worsening, as defined by the GRC. Although the DASH demonstrated moderate responsiveness, it could not distinguish those who had experienced meaningful change so an MIC could not be calculated. It is possible that many of the task-based items of the DASH might reflect limitation in shoulder function hence it was not interpretable in this study. In addition, how patients answer the GRC in the context of the treatment of Dupuytren's disease is not understood. Patients might consider "success" to be straightening of the flexed

304 finger or general improvement in hand function. So the anchor might be answered in  
305 relation to finger straightening rather than general hand function, with the DASH  
306 items perhaps reflecting the latter.

307 The URAM showed acceptable interpretability with an MIC for open surgery of 10.5  
308 out of the 45 points on the scale and an MID of 8.

309 Interpretability had been studied in Dupuytren's disease (Beaudreuil et al., 2011,  
310 Witthaut et al., 2011). Neither study considered open surgery, though one generated  
311 an MIC estimate of 2.7 for the URAM for needle aponeurotomy (Beaudreuil et al.,  
312 2011). Our study generated a considerably larger MIC, possibly as a different  
313 technique for assessing interpretability was used. However, the difference in MICs  
314 could reflect the differences in recovery between open surgery and aponeurotomy,  
315 as aponeurotomy will typically have a shorter recovery. If recovery from open  
316 surgery is more arduous and prolonged than after aponeurotomy, then greater  
317 improvement may be needed for the patient to consider it clinically meaningful or  
318 worthwhile. MICs may vary between treatments in general (Revicki et al., 2008), and  
319 this is the case in hand surgery (Rodrigues et al., 2014).

320 The DASH has been the most popular PROM for studying Dupuytren's disease  
321 treatment (Ball et al., 2013). However, its validity has been questioned in a study  
322 with contemporary design (Forget et al., 2014), as well as in previous studies, which  
323 have more limitations (Degreef et al., 2009, Engstrand et al., 2009, Jerosch-Herold et  
324 al., 2011, Packham, 2011, Zyluk and Jagielski, 2007). Our study raises further  
325 questions regarding its suitability for use in Dupuytren's disease, in terms of poor  
326 interpretability, an aspect of its behaviour that has not been studied previously for  
327 Dupuytren's disease. However, given the potential issues associated with  
328 retrospective anchors, we believe that the poor interpretability of the DASH should be  
329 confirmed using a prospective anchor. Measurement of the interpretability of other  
330 measures for use in Dupuytren's disease is also required to interpret existing  
331 research and ensure appropriate outcome measures are used in future research.

332 The methodology used to study interpretability has limitations. Interpretability of the  
333 outcome of treatment was studied at one year following surgery, as the only  
334 postoperative time points captured were at 6 weeks and 1 year. However, 1 year  
335 follow up is a pragmatic time point to measure the outcome as this study  
336 demonstrates that recovery takes longer than 6 weeks (hence studies reporting  
337 outcome at 6 weeks may have too short follow up) and final follow up at 1 year may  
338 be used in clinical practice. It could be performed by telecommunication using  
339 PROMs, if interpretability data were available to inform how to handle responses  
340 obtained at this time point. Although the GRC was developed for the purpose of  
341 anchoring outcomes, it is administered retrospectively; the use of such retrospective  
342 anchors has been criticised (Norman et al., 1997). In particular, the GRC may reflect  
343 the status of the hand at the time of assessment, rather than reflecting the change  
344 that has occurred from the preoperative state (Garrison and Cook, 2012, Schmitt and  
345 Di Fabio, 2005). It is possible that the interpretability of the PROMs studied would be  
346 different if a prospective anchor were used. One prospective approach has been to  
347 use the satisfaction domain of the Michigan Hand Questionnaire (MHQ) as the  
348 anchor (London et al., 2014, Malay et al., 2013, Shauver and Chung, 2009, Waljee  
349 and Chung, 2012). However, it is not clear whether it is appropriate to use one  
350 domain of the MHQ to study the interpretability of other domains of the same tool.  
351 Anchors, and the GRC in particular, ask the patient to determine whether they are  
352 'better'. However, it is not clear which domain patients use to judge non-specific  
353 'improvement'. Furthermore, it is not clear whether there is heterogeneity in this, with  
354 different patients defining 'better' in different ways. Further consideration of  
355 appropriate anchors may be required and, if appropriate, our findings would have to  
356 be confirmed using such methodology.

357 The methodology for sub-grouping outcomes to construct the ROC charts has been  
358 used in previous studies, with exclusion of patients who experienced deterioration.  
359 However, recent work suggests inclusion of such patients may improve the precision

of MIC analyses (Turner et al., 2009). This might be considered in the future work proposed above, in which interpretability would be studied with a prospective anchor. Our cohort of patients is heterogeneous, as it included both those undergoing either fasciectomy or dermofasciectomy. These data were collected as service evaluation of standard clinical practice and on several instances, the type of procedure to be performed was changed in the preadmission clinic, based on surgeon preference or during surgery itself for technical reasons. It is possible that the interpretability of the DASH and the URAM may differ between patients undergoing fasciectomies and dermofasciectomies. In particular, complications such as cold intolerance may be more common after dermofasciectomy. However, given that the recovery following the two procedures were was not different, and that MICs are considered estimates rather than exact values, we believe that this is unlikely to be of significance. Whilst early recovery from surgery for Dupuytren's disease has been assessed at 6 weeks (van Rijssen et al., 2006), and studies of other treatments have assessed outcome at 30 days (Hurst et al., 2009), our data support previous findings that suggest that the recovery from open surgery takes longer than 6 weeks (Ullah et al., 2009). Therefore, confirmation of when recovery after surgery plateaus is needed to determine the best time point for studying recovery and optimising length of follow up of clinical trials.

In conclusion, while the DASH score exhibited moderate responsiveness, its poor interpretability on receiver operating characteristic (ROC) curve analysis meant that an MIC could not be estimated. This suggests that it is not interpretable in clinical practice and research in surgery for Dupuytren's disease. The URAM score showed acceptable responsiveness, and was interpretable. It had an MIC of 10.5 on ROC analysis.

**Acknowledgements:**

388 JNR received educational support from a Scholarship from the National Institute for  
389 Health and Care Excellence (NICE) during this project.

390 **Conflict of Interest**

391 The authors declared no potential conflicts of interest with respect to the research,  
392 authorship, and/or publication of this article.

393 **Funding statement**

394 This work was supported by a BSSH Research Fellowship, Nottingham Hospitals  
395 Charity and Nottingham Orthopaedic Walk.

396 **Ethical approval**

397 This study was a minor element of a larger service evaluation project studying  
398 treatment outcome in Dupuytren's disease. In keeping with UK National Research  
399 Ethics Service guidance, it is exempt from ethical approval. Approval as service  
400 evaluation was prospectively obtained.

401



## References

- Ball C, Pratt AL, Nanchahal J. Optimal functional outcome measures for assessing treatment for Dupuytren's disease: a systematic review and recommendations for future practice. *BMC Musculoskelet Disord*. 2013, 14: 131.
- Beaudreuil J, Allard A, Zerkak D et al. Unite Rhumatologique des Affections de la Main (URAM) scale: development and validation of a tool to assess Dupuytren's disease-specific disability. *Arthrit Care Res*. 2011, 63: 1448-55.
- Degreef I, Vererfve PB, De Smet L. Effect of severity of Dupuytren contracture on disability. *Scand J Plast Recons*. 2009, 43: 41-2.
- Deyo RA, Centor RM. Assessing the responsiveness of functional scales to clinical change: an analogy to diagnostic test performance. *J Chron Dis*. 1986, 39: 897-906.
- Engstrand C, Boren L, Liedberg GM. Evaluation of activity limitation and digital extension in Dupuytren's contracture three months after fasciectomy and hand therapy interventions. *J Hand Ther*. 2009, 22: 21-6.
- Forget NJ, Jerosch-Herold C, Shepstone L, Higgins J. Psychometric evaluation of the Disabilities of the Arm, Shoulder and Hand (DASH) with Dupuytren's contracture: validity evidence using Rasch modeling. *BMC Musculoskelet Disord*. 2014, 15: 361.
- Garrison C, Cook C. Clinimetrics corner: the Global Rating of Change Score (GRoC) poorly correlates with functional measures and is not temporally stable. *J Man Manip Ther*. 2012, 20: 178-81.
- Hurst LC, Badalamente MA, Hentz VR et al. Injectable collagenase clostridium histolyticum for Dupuytren's contracture. *N Engl J Med*. 2009, 361: 968-79.
- Jaeschke R, Singer J, Guyatt GH. Measurement of health status. Ascertaining the minimal clinically important difference. *Control Clin Trials*. 1989, 10: 407-15.
- Jerosch-Herold C, Shepstone L, Chojnowski A, Larson D. Severity of contracture and self-reported disability in patients with Dupuytren's contracture referred for surgery. *J Hand Ther*. 2011, 24: 6-10.

429 Kazis LE, Anderson JJ, Meenan RF. Effect sizes for interpreting changes in health  
 430 status. *Med Care*. 1989, 27: S178-89.

431 Kennedy C, Beaton D, Solway S, McConnell S, Bombardier C. *The DASH and*  
 432 *QuickDASH Outcome Measure User's Manual*, Third ed. Toronto, Institute for Work  
 433 and Health, 2011.

434 London DA, Stepan JG, Calfee RP. Determining the Michigan Hand Outcomes  
 435 Questionnaire minimal clinically important difference by means of three methods.  
 436 *Plast Reconstr Surg*. 2014, 133: 616-25.

437 Malay S, Chung KC. The minimal clinically important difference after simple  
 438 decompression for ulnar neuropathy at the elbow. *J Hand Surg Am*. 2013, 38: 652-9.

439 Mintken PE, Glynn P, Cleland JA. Psychometric properties of the shortened  
 440 disabilities of the Arm, Shoulder, and Hand Questionnaire (QuickDASH) and Numeric  
 441 Pain Rating Scale in patients with shoulder pain. *J Shoulder Elb Surg*. 2009, 18: 920-  
 442 6.

443 Mohan A, Vadher J, Ismail H, Warwick D. The Southampton Dupuytren's Scoring  
 444 Scheme. *J Plast Surg Hand Surg*. 2014, 48: 28-33.

445 Mokkink LB, Terwee CB, Knol DL et al. The COSMIN checklist for evaluating the  
 446 methodological quality of studies on measurement properties: a clarification of its  
 447 content. *BMC Med Res Methodol*. 2010, 10: 22.

448 Norman G. Likert scales, levels of measurement and the "laws" of statistics. *Adv*  
 449 *Health Sci Educ Theory Pract*. 2010, 15: 625-32.

450 Norman GR, Stratford P, Regehr G. Methodological problems in the retrospective  
 451 computation of responsiveness to change: the lesson of Cronbach. *J Clin Epidemiol*.  
 452 1997, 50: 869-79.

453 Packham T. Clinical commentary in response to: Severity of contracture and self-  
 454 reported disability in patients with Dupuytren's contracture referred for surgery. *J*  
 455 *Hand Ther*. 2011, 24: 12-4.

456 Revicki D, Hays RD, Cella D, Sloan J. Recommended methods for determining  
 457 responsiveness and minimally important differences for patient-reported outcomes. J  
 458 Clin Epidemiol. 2008, 61: 102-9.

459 Rodrigues JN, Zhang W, Scammell BE, Davis TR. What patients want from the  
 460 treatment of Dupuytren's disease -- is the Unité Rhumatologique des Affections de la  
 461 Main (URAM) scale relevant? J Hand Surg Eur. 2014, 40: 150-4.

462 Rodrigues JN, Mabvuure NT, Nikkhah D, Shariff Z, Davis TR. Minimal important  
 463 changes and differences in elective hand surgery. J Hand Surg Eur. 2014.

464 Schmitt J, Di Fabio RP. The Validity of Prospective and Retrospective Global  
 465 Change Criterion Measures. Arch Phys Med Rehabil. 2005, 86: 2270-6.

466 Shauver MJ, Chung KC. The minimal clinically important difference of the Michigan  
 467 hand outcomes questionnaire. J Hand Surg Am. 2009, 34: 509-14.

468 Szabo RM. Outcomes assessment in hand surgery: when are they meaningful? J  
 469 Hand Surg Am. 2001, 26: 993-1002.

470 Testa MA. Interpreting quality-of-life clinical trial data for use in the clinical practice of  
 471 antihypertensive therapy. J Hypertens. 1987, 5: S9-13.

472 Turner D, Schunemann HJ, Griffith LE et al. Using the entire cohort in the receiver  
 473 operating characteristic analysis maximizes precision of the minimal important  
 474 difference. J Clin Epidemiol. 2009, 62: 374-9.

475 Ullah AS, Dias JJ, Bhowal B. Does a 'firebreak' full-thickness skin graft prevent  
 476 recurrence after surgery for Dupuytren's contracture?: a prospective, randomised  
 477 trial. J Bone Joint Surg Br. 2009, 91: 374-8.

478 van Rijssen AL, Gerbrandy FS, Ter Linden H, Klip H, Werker PM. A comparison of  
 479 the direct outcomes of percutaneous needle fasciotomy and limited fasciectomy for  
 480 Dupuytren's disease: a 6-week follow-up study. J Hand Surg Am. 2006, 31: 717-25.

481 Waljee JF, Chung KC. Objective functional outcomes and patient satisfaction after  
 482 silicone metacarpophalangeal arthroplasty for rheumatoid arthritis. J Hand Surg Am.  
 483 2012, 37: 47-54.

Witthaut J, Bushmakina AG, Gerber RA, Cappelleri JC, Le Graverand-Gastineau MP.  
Determining clinically important changes in range of motion in patients with  
Dupuytren's Contracture: secondary analysis of the randomized, double-blind,  
placebo-controlled CORD I study. Clin Drug Invest. 2011, 31: 791-8.  
Youden WJ. Index for rating diagnostic tests. Cancer. 1950, 3: 32-5.  
Zyluk A, Jagielski W. The effect of the severity of the Dupuytren's contracture on the  
function of the hand before and after surgery. J Hand Surg Eur. 2007, 32: 326-9.

## Legends

## Tables

Table 1: Patient demographics in cohort study

Table 2: DASH scores at different timepoints for individuals who completed all  
timepoints (65 patients)

Table 3: Responsiveness of DASH scores and URAM scores at 1 year  
postoperatively

## Figures

Chart 1: Line chart of DASH summary scores from patients who completed all time  
points (n=65)

Chart 2: ROC curve of DASH score's ability to separate 'improved' from 'stable'  
outcomes, based on the GRC

512 Chart 3: ROC curve of the URAM score's ability to separate 'improved' from 'stable'  
513 outcomes, based on the GRC

514

515 Chart 4: Scatterplot of GRC versus change in DASH score at 1 year (n=68)

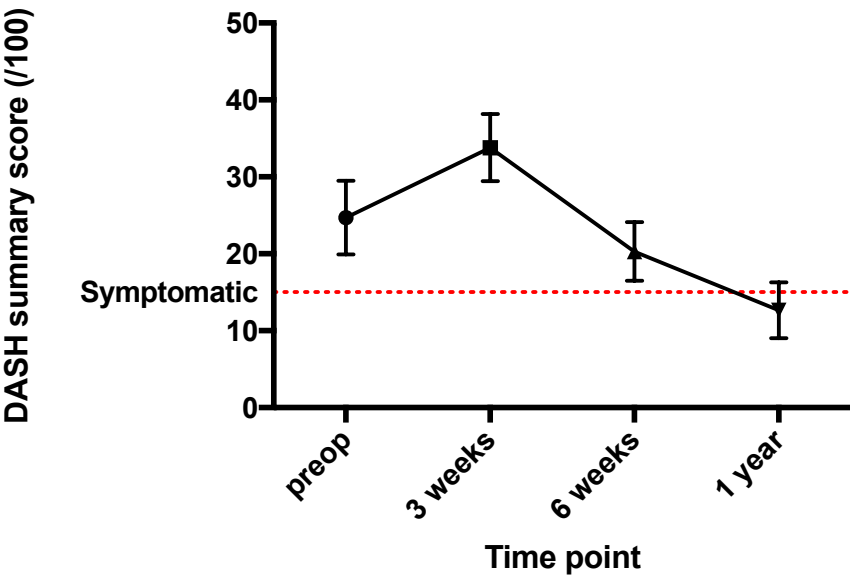
516

517 Chart 5: Scatterplot of GRC versus change in URAM score at 1 year (n=30)

518

519

1 Chart 1: Line chart of DASH summary scores from patients who completed all time  
2 points (n=65)  
3



4

5 N.B. The points are the mean DASH score, with 95% confidence intervals. The

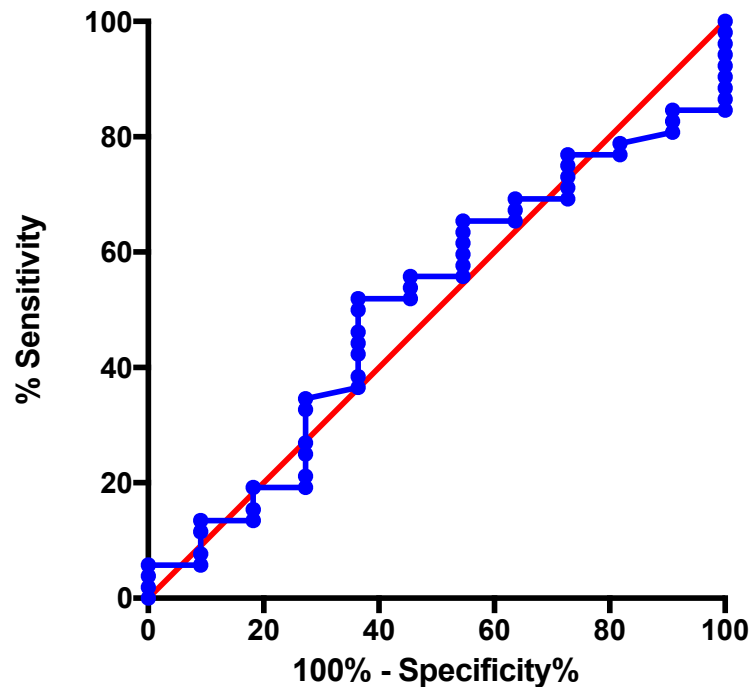
6 developers of the DASH anticipate that a patient will be “symptomatic” in his or her

7 upper limb if their DASH score is >15/100 (Kennedy, et al., 2011).

8

9

- 1 Chart 2: ROC curve of DASH score's ability to separate 'improved' from 'stable'
- 2 outcomes, based on the GRC
- 3



4

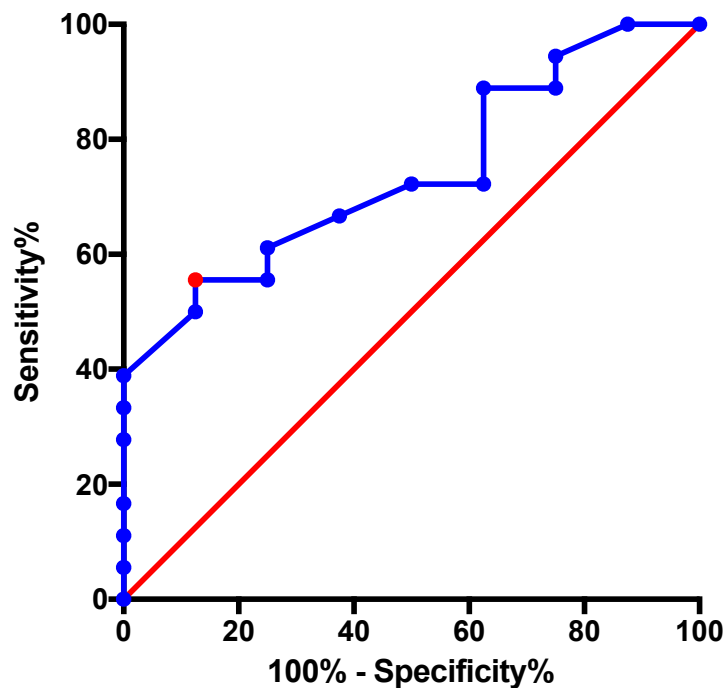
5

- 6 The red line indicates the line of identity, where sensitivity and specificity are both
- 7 50%, and corresponds to an area under the curve of 0.5. The blue points are
- 8 theoretical cut offs for change in DASH score that could be used to attempt to
- 9 separate 'improved' from 'stable' outcomes. Conventionally "100% - specificity %" is
- 10 plotted on the x axis as this represents the "False positive rate". Specificity on the y
- 11 axis is the true positive rate.

12

13

- 1 Chart 3: ROC curve of the URAM score's ability to separate 'improved' from 'stable'
- 2 outcomes, based on the GRC
- 3



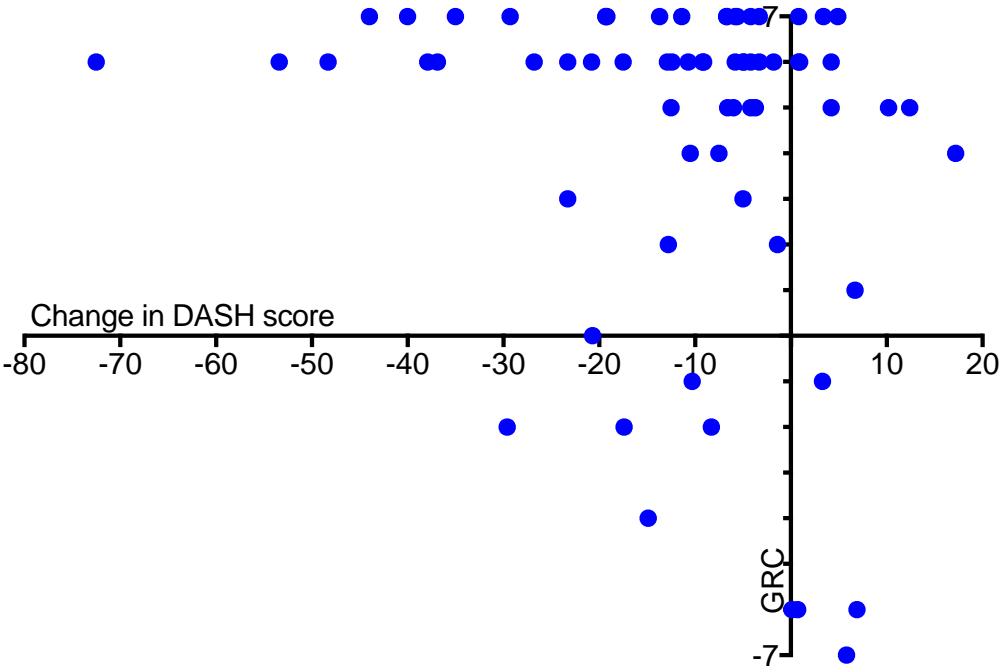
- 4
- 5 The red point on the blue line corresponds to Youden's index (highest value for
- 6 sensitivity + specificity -1). This corresponds to a cut point of an improvement of
- 7 >10.5 in the URAM, i.e. using an improvement of 10.5 is the most effective MIC, with
- 8 those who experience less than 10.5 improvement in their URAM score considered
- 9 stable, and those who experience more than 10.5 improvement in their URAM score
- 10 considered improved.

- 11
- 12



1    Chart 4: Scatterplot of GRC versus change in DASH score at 1 year (n=68)

2



3

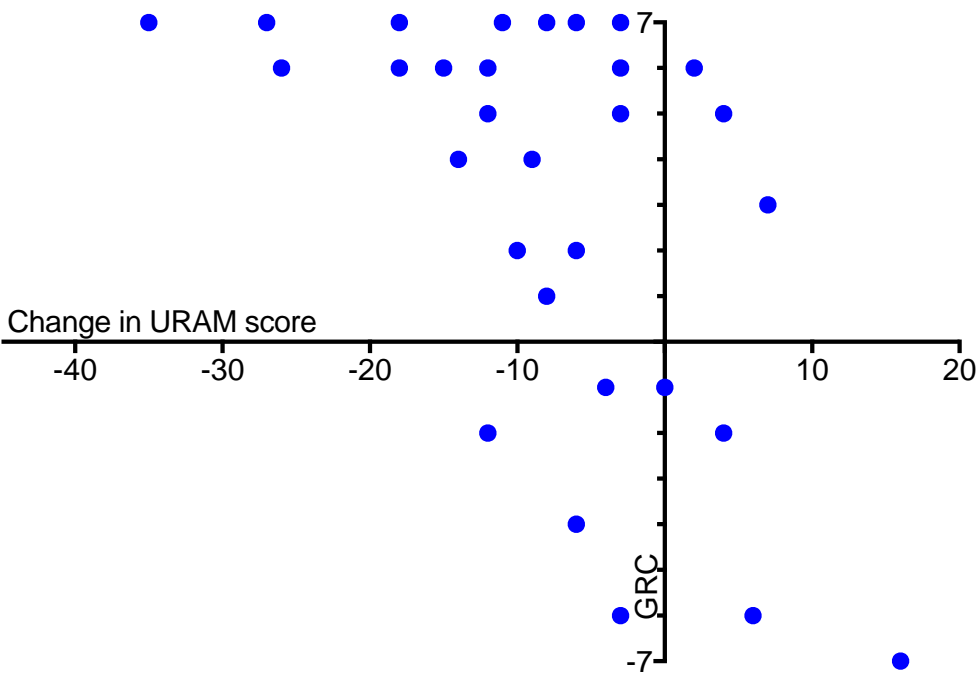
4

5

6

1 Chart 5: Scatterplot of GRC versus change in URAM score at 1 year (n=30)

2



3

4

5

6

Table 1: Patient demographics in cohort study

Demographic	DASH scores	URAM scores
Age at recruitment (years)	Mean 67, Range 34-90	Mean 66, Range 38-90
Gender	83/101 men (82%)	38/44 men (86%)
Procedure types	73 fasciectomies	29 fasciectomies
	28 dermofasciectomies	15 dermofasciectomies
Hand treated	61/101 right (60%)	25/44 right (57%)
Multiple digits treated	27/101 (27%)	12/44 (28%)
Digits treated	135 digits in 101 patients	59 digits in 44 patients
Little	80 (59%)	34 (58%)
Ring	39(29%)	15 (25%)
Middle	10 (7%)	7 (12%)
Index	4 (3%)	2 (3%)
Thumb	2 (2%)	1 (1%)

1    Table 2: DASH scores at different timepoints for individuals who completed all timepoints (65 patients)

		Tukey's multiple comparisons test results		
	Mean (95%CI)	Versus 3 weeks postoperative	Versus 6 weeks postoperative	Versus 1 year postoperative
Preoperative	24.7/100 (19.9, 29.5)	p=0.002	p=0.205	p<0.0001
3 weeks postoperative	33.8/100 (29.4, 38.2)		p<0.0001	p<0.0001
6 weeks postoperative	20.3/100 (16.5, 24.1)			p=0.001
1 year postoperative	12.7/100 (9.0, 16.3)			

2

3    Repeated measures one way ANOVA: p<0.0001

4    Incomplete results excluded listwise (i.e. if an individual missed one or more timepoints, all data for that individual was not analysed)

5

6

7

1 Table 3: Responsiveness of DASH scores and URAM scores at 1 year  
2 postoperatively

	DASH score (71 patients)	URAM score (30 patients)
Preoperative	24.5/100 (19.9, 29.0)	17.8/45 (14.5, 21.1)
[mean (95% CIs)]		
1-year postoperative	12.4/100 (8.9, 16.0)	10.1/45 (6.1, 14.1)
[mean (95% CIs)]		
Difference post-preop	12.0/100 (8.2, 15.9)	7.7/45 (3.7, 11.7)
[mean (95% CIs)],	p<0.0001	p=0.0005
paired t test		
Effect size	0.58	0.87
(mean/SD preop)		

3  
4 Incomplete data were excluded pairwise, i.e. all individuals who completed  
5 preoperative and 1-year postoperative outcomes were included, even if they did not  
6 adequately complete 3 week or 6 week postoperative outcomes (i.e. 71 patients  
7 completed preoperative and 1-year postoperative DASH scores, whereas 65  
8 completed all timepoints).

9

10