Coordinating outcomes measurement in ataxia research: do some widely used generic rating scales tick the boxes?

Riazi A[1,2], Cano SJ[1], Cooper JM[3], Bradley JL[3], Schapira AHV[3], Hobart JC[1,4]

[1]Neurologial Outcome Measures Unit, Institute of Neurology, London, UK

[2]Department of Psychology, Royal Holloway, University of London, Surrey, UK

[3]Department of Clinical Neurosciences, Royal Free and University College Medical School, London UK

[4]Department of Clinical Neurosciences, Peninsula Medical School, Plymouth, UK

Correspondence: Dr. Jeremy Hobart, Senior Lecturer and Honorary Consultant Neurologist, Department of Clinical Neuroscience, Peninsula Medical School, Room N16 ITTC Building, Tamar Science Park, Davy Road, Plymouth, Devon PL6 8BX, UK

T: +44 (0) 1752 315272;

F: +44 (0) 1752 315254;

E: Jeremy.Hobart@pms.ac.uk

Word count: Abstract 192; **Text** 2935; **Title** 88 characters

Abstract

Objectives: To examine the psychometric properties of four widely-used generic health status measures in Friedreich's ataxia (FA), to determine their suitability as outcome measures.

Methods: Fifty-six people with genetically confirmed FA completed the Barthel Index (BI), General Health Questionnaire (GHQ-12), EuroQol (EQ-5D) and Medical Outcomes Study 36-item Short Form Health Survey (SF-36) via postal survey. Six psychometric properties (data quality, scaling assumptions, acceptability, reliability, validity and responsiveness) were examined.

Results: The response rate was 97%. In general, the psychometric properties of the four measures satisfied recommended criteria. However, closer examination highlighted limitations restricting their use for treatment trials. For example, the BI had high missing data, EQ-5D had poor discriminant ability and five SF-36 scales had high floor and/or ceiling effects. Most scale scores did not span the entire scale range, had means that differed notably from the scale mid-point, and had wide confidence intervals. Effect sizes (ES) were small for all four measures raising questions about their ability to detect clinically significant change.

Conclusions: Results highlight the potential limitations of these four scales for evaluating health outcomes in FA, and suggest the need for new disease-specific patient-based measures of the impact.

Keywords: Friedreich's ataxia (FA), health status measures, psychometric, patient-based outcome measures

Introduction

Evaluations of therapeutic interventions should include measurement of patient-based outcomes.[1] These outcomes must be measured rigorously if they are to influence patient welfare and the expenditure of public funds.[2] Currently, no disease-specific patient-based rating scale exists for Friedreich's ataxia (FA). Commonly used ataxia rating scales include the International Cooperative Ataxia Rating Scale (ICARS)[3] and the International Ataxia Clinical Rating Scale (IACRS).[4] These are observer rated, and quantify the neurological examination of people with ataxia. Neither scale incorporates patients' perceptions or evaluates the impact of FA on daily life. A recently developed observer-rated scale for FA[5] also incorporates ADL (activities of daily living) assessment. However, none of these three scales were developed using recognised psychometric methods of scale construction (item generation, scale formation and testing). Perhaps not surprisingly, therefore, the ICARS has recently been shown to have limitations to its use in the assessment of FA.[6]

Apart from one previous study that found disability has an influence on work and social activities for people with FA,[7] a comprehensive literature search generated no quality of life studies in FA using standardised measures. Therefore an important first step in advancing patient-based outcomes measurement in FDRA is to evaluate the potential usefulness of some existing widely used rating scales. As such we conducted a postal survey of people with FA using four widely used generic health status measures: Barthel Index (BI[8]), General Health Questionnaire (GHQ[9]), EuroQol (EQ-5D[10]) and Medical Outcomes Study 36-item Short Form Health Survey (SF-36[11]). The aim of this study was to evaluate the potential suitability of these measures in treatment trials and epidemiological studies in FA.

Methods

Samples

The sample was 58 people with genetically confirmed FA at the Royal Free Hospital. Appropriate ethical committee approval was obtained. Rating scales were administered by postal survey in a booklet with some demographic questions. Non-responders received a single reminder at 8 weeks.[12] A second postal survey was conducted one year after the first survey to assess responsiveness.

Measures

Four standardised measures were administered: the self-report BI (a measure of personal activities of daily living[8,13,14]), the GHQ-12 (a measure of psychological well-being[9,15]), the EuroQol (EQ-5D Health State and Thermometer[10]) and the SF-36 (a measure of health status in 8 scales – Table 1 for description[11]). These scales were chosen as they are all widely used generic measures[16-21] recommended for a range of health care settings.

Analyses

The psychometric properties of the measures were evaluated using standard methods that are fully described in previous publications.[22-24] Six psychometric properties were examined: data quality, scaling assumptions, acceptability, reliability (internal consistency), convergent and discriminant construct validity and responsiveness.

Data quality (the extent to which an instrument can be used successfully in a clinical setting)[25] was determined to be high if items had low missing data (<10%), and if a high percentage of scale scores were computable for each patient.

Scaling assumptions (the legitimacy to sum item scores without weighting or standardisation to generate a total score) were examined by determining whether items in each scale had roughly similar response-option frequency distributions, equivalent mean and variances, and equivalent item-total correlations ($r > 0.30$).[26,27]

Acceptability (the targeting of a scale to a sample so that score distributions adequately represent the true distribution of health status in the sample[28]) was examined to determine that observed scores were well distributed,[29] mean scores were near the scale mid-point[30] floor and ceiling effects were low, and skewness statistics ranged from -1 to +1.[31]

Reliability (the extent to which an instrument is free from random error[32]) was examined using internal consistency, using Cronbach's alpha coefficient.[33] It is recommended that alpha >0.80.[34] We also computed the 95% CI limits around individual patient scores as: +/-1.96SEM, where SEM (standard error of measurement) = SD x $\sqrt{}$ (1 – alpha).

Validity (the extent to which an instrument measures the concept it purports to measure) was examined using convergent and discriminant construct validity.[35]

Correlations between scales were examined to determine the extent to which each instrument: 1) measures what it is supposed to measure (convergent construct validity) and 2) does not measure what it is not designed to measure (discriminant construct validity). Table 1 shows the predicted correlations between measures, based on the clinically expected associations between the constructs they purport to measure, using broad criteria of: <0.30 = low (L); 0.30 – 0.70 = moderate (M); >0.70 = high (H).

Responsiveness (the ability of an instrument to measure clinically important change over time) was examined by scales being administered at Time 1 (T1) and Time 2 (T2; 12 months later). Responsiveness was determined by calculating effect sizes (ES),[36] as defined as the mean change in score (T1 minus T2) divided by the standard deviation of T1 scores. These are interpreted as either small (ES < 0.20), medium (ES = 0.50) or large (ES > 0.80).[37]

Results

Samples

Fifty-six people returned completed questionnaires (response rate 97%). The group represented a broad range of adult patients with FA covering the disease spectrum from mild to severe. More than half the sample was single, used a wheelchair indoors, was not in employment, and was educated past the age of 16 (Table 2).

Measures

a) Self-report Barthel Index

One item (mobility) had missing data > 10%. Scale scores were computable for <90% of the sample, suggesting limited data quality. Item mean scores and standard deviations were variable. Scale scores were well-distributed but did not span the entire scale range. The mean BI score (16.3 point) was notably lower than the scale mid-point (50) although skewness was within the recommended range. The ceiling effect was acceptable.[38]

Internal consistency estimates exceeded recommended criteria (>= 0.80). However, the 95% CI for individual patient scores were quite wide (16 points) indicating limited usefulness for individual-level measurement (Table 3a). The direction and pattern of correlations were generally consistent with predictions. However, the magnitude of correlations with other physical scales was not as high as predicted (eg BI correlated highest with Euroqol health state and years since diagnosis; Table 4).

Change scores indicated minimal worsening of scores, implying worsening health between T1 and Time 2, although this was not statistically significant. ESs were small, implying low responsiveness (Table 5).

b) GHQ-12

Missing item data was low and data quality was good (91% computable scale scores). Frequency distributions for items were quite symmetrical. Item-total correlations (range 0.45 - 0.83) exceeded the criteria of 0.40. Scale scores were well distributed but did not span the entire scale range. The mean score (36.1) differed somewhat from the mid-point 50, although skewness was acceptable. The ceiling effect was minimal (2%).

Internal consistency was high. The 95% CI for individual patient scores was less than 10 points (Table 3a). Construct validity was supported by the direction, magnitude and pattern of correlations with other scales (Table 4).

Change score were significantly lower at Time 2 implying a significant improvement over time. However ESs were low, raising the question of limited responsiveness (Table 5).

c) EuroQol (EQ-5D health state and Thermometer)

Only 80% of the sample completed all the items of the EQ-5D health state. The mobility item had 20% missing data suggesting poor data quality. EQ-5D health state did not span the entire scale range. Cronbach's alpha=0.58 for the unweighted items indicating limited reliability. Correlations with other scales were in the expected range (0.30 - 0.70), except with the SF-36 BP which was low (r=0.13). Many of the correlations were in a narrow range (0.29-0.54), indicating limited discriminant ability.

The EuroQoL thermometer was completed by 98% of the sample, was well-distributed but did not span the entire scale range. The mean score (64.3) was somewhat above the scale mid-point (50 points) indicating that the average response tended towards better health. No ceiling or floor effect was found, and scores were not notably skewed (Table 3a). Correlations with other scales were in the expected range (0.30 - 0.70), except for the SF-36 PF which was low (r=0.12). However, many were in the range 0.24 – 0.49 indicating limited discriminant ability (Table 4).

Change scores indicated very little change in both Thermometer and Health State between Time 1 and Time 2. ESs were low. This suggests limited responsiveness (Table 5).

d) SF-36

Missing data was low. Total scores could be computed for >94% of the sample, and frequency distributions for items were quite symmetrical. Item-total scale correlations were satisfactory (>0.30) for all scales except PF (0.24 for one item). Five out of the 8 scale scores (PF, BP, GHP, VT and MH) were well-distributed but did not span the entire scale range. The mean score of the PF, BP, SF, RL-E and MH scales differed substantially from the mid-point, by at least 20 points. PF and RL-E were outside the skewness range of -1 to +1. GHP, VT and MH scales had small floor and ceiling effects. The floor effects of PF, RL-P and RL-E scales were > 20%. The ceiling effects of RL-P, BP, SF, and RL-E scales were > 20%. These results suggest that some SF-36 scales have limited targeting to this sample (Table 3b).

Internal consistency estimates exceeded the recommended criteria for all scales except GHP. The 95% CIs for individual patients were smallest for PF, and widest for the RL-E. The correlation between the two psychological scales was substantial (MH and RL-E = 0.62). However, the correlations between the physical scales were lower than expected: PF and RL-P (0.11), PF and BP (0.08), PF and GHP (0.01), RL-P and BP (0.25), RL-P and GHP (0.27), and BP and GHP (0.37). These findings suggest less than adequate validity for the physical scales (Table 4).

Change scores for VT, SF and MH showed statistically non-significant improvement between Time 1 and Time 2. Change scores for RL-E, BP, and PF showed statistically non-significant worsening between Time 1 and Time 2. Change scores for GHP showed statistically significant worsening of scores. The scores for RL-P did not change between Time 1 and Time 2. For all scales, ESs were small, implying limited responsiveness (Table 6).

Discussion

This study examined the suitability of four widely used generic scales for use as outcome measures in FA research. In general, they satisfied basic criteria. As such we are able to make inferences about patients' perceptions of the impact of FA. It has a substantial impact on physical function, psychological well-being, general health perceptions, vitality, and overall quality of life. In addition, a comparison with SF-36 data in Multiple Sclerosis (MS) patients,[39] indicate that the physical impact in FA is greater. However, larger samples with age, sex and disability-matched comparisons with MS and other disease groups are required to make detailed and specific comparisons.

Although scales satisfied basic psychometric criteria, closer examinations highlight limitations that restrict their use in treatment trials. Most scales were not completed by the whole sample. BI Mobility and EQ5D had the highest missing data. This has implications for dropout rates, which in turn impact on the interpretation of studies. Unfortunately, our data do not provide an explanation for this finding. As the BI does not have a category for mobility without aid a proportion of these patients may have felt unable to answer this question. The ambiguity of response options is another

possible, but unlikely, reason. High missing data implies an item is of limited relevance and that it should be considered for removal. However, in this case, we would recommend qualitative work to uncover the true cause as mobility is an important aspect of disabling neurological disorders.

Five of the eight SF-36 scales had floor and/or ceiling effects. This suggests poor scale-to-sample targeting and has potential implications for detecting change in treatment trials. High floor or ceiling effects at baseline almost certainly underestimate change over time, and due to treatments. Thus small but clinically meaningful changes may go undetected. Floor and ceiling effects may also attenuate correlations between measures, as correlations are sensitive to scale ranges. This may explain why some of the expected correlations (Table 1) were not observed.

The mean score of the majority of scales (BI, GHQ-12, EuroQol, SF-36 PF, BP, SF, RL-E and MH) differed notably from the scale mid-point, suggesting limited targeting in this sample. The wide confidence intervals observed suggest that the measures are not applicable for individual patient monitoring, supporting previous work.[38] In addition, correlations between the physical scales and other SF-36 scale were lower than expected, suggesting limited validity. An alternative explanation is this may reflect the way FA patients view their illness as a unique set of physical features of FA, but to clarify this requires reliable and valid measures of each domain.

Effect sizes for all four measures were small. FA is a progressive disorder, so we might expect deterioration, at least in physical function. Low ESs imply limited responsiveness, which is an important consideration for treatment trials as they may

not be sensitive to detect small but clinically important changes. However, further responsiveness studies are needed that compare changes in scores to an external criterion of change, such as a transition question. Furthermore, assessment of change in FA is complicated by two factors: 1) the diverse nature of FA may mean that different functions may be affected at different stages of the disease and 2) the slow disease progression means that one year or less may be too short to detect any change in disease progression. This has implications for evaluating treatments that attempt to slow down the natural history of FA and studies may need to be over a long period of time if we are to detect true differences.

There are several limitations to the study. First, our sample size was small. However, FA is relatively rare, and sixty patients represents one of the largest studies of its type. Also, there is evidence to suggest that useful reliability and validity estimates can be obtained from small (even non representative) samples,[40,41] and that Cronbach's alpha is considered a conservative estimate of reliability.[42] Second, the sample representativeness is unknown as it may have been skewed towards the advanced stages of the disease, and patients under 18 years were not represented. Third, this study only examined a few of the many available scales. However, the inherent physical and mental fatigability of people with FA, and time taken to complete some of these scales, limits the number of instruments that can be administered at one time. Also, additional measures may have led to them being completed over several days, adding the complication of day-to-day variability, which is known to be substantial in FA. It would be valuable to form a UK register of people with FA, identify a list of potentially valuable measures, and systematically evaluate small groups using these measures over time. This would underpin evidence-based measurement in FA. A

fourth limitation is that we did not compare normative data for each scale. However, the primary aim of this paper was to psychometrically evaluate each scale, which is a prerequisite for meaningful comparisons. The problems uncovered may suggest that any comparison would be limited. In addition, there is no normative data available for the BI and GHQ-12. Finally, we did not determine whether patients were able to self-complete each questionnaire, or if they required assistance, and therefore do not know how this may have affected our findings.

Despite these limitations, this study supports the use of self-report questionnaires in FA to capture aspects of outcome not captured by objective measures, and providing a clearer picture of the wider impact of FA. Although the responsiveness of such measures is in question, the development of a new scale tailored specifically to FA patients may address such a shortfall. In real terms, the psychometric shortfalls of each measure question their appropriateness in cross-sectional descriptive studies (due to poor targeting), longitudinal studies (due to potential problems of assessing clinical change) and individual patient monitoring (due to wide confidence intervals around scale scores).

This study aimed to evaluate the potential for existing generic scales to measure the impact of FA in studies. This evidence has been thus far lacking. If these measures had passed this first hurdle, which they did not, the next step would have been to carry out careful and meaningful cross sectional and longitudinal examinations of the relationship between quality of life and disease severity. The limitations of these four scales suggest that new disease-specific patient-based measures of the impact of FA are needed if accurate evaluations of its natural history are required, and if unique

studies of the relationships between molecular genetics, clinical manifestations and health status are to be undertaken.

Acknowledgments

JLB was funded by a grant from the National Lottery UK. During the writing of this paper JH was on secondment to the School of Education, Murdoch University, Perth Western Australia. This research attachment was supported by the Royal Society of Medicine, through an Ellison-Cliffe Travelling Fellowship and the MS Society of Great Britain and Northern Ireland.

References

1. Ware JE Jr. Measuring patients' views: the optimum outcome measure. Br Med J 1993;306:1429-1430.

2. McDowell I, Jenkinson C. Development standards for health measures. J Health Serv Res Policy 1996;1:238-246.

3. Trouillas P, Takayanagi T, Hallet M, et al. International Cooperative Ataxia Rating Scale for pharmacological assessment of the cerebellar syndrome. J Neurol Sci 1997;145:205-211.

4. Filla A, De Michele G, Caruso G, Marconi R, Campanella G. Genetic data and natural history of Friedreich’s disease: a study of 80 Italian patients. J Neurol 1990;237:345–51.

5. Subramony SH, May W, Lynch D et al. Measuring Friedreich's ataxia: Interrater reliability of a neurologic rating scale. Neurology 2005;64:1261–1262.

6. Cano SJ, Hobart JC, Hart PE, Kolipara LVP, Schapira AHV, Cooper JM. The International Co-operative Ataxia Rating Scale (ICARS): An appropriate rating scale for Friedreich's Ataxia? Movement Disorders 2005;20(12):1585-1591.

7. D'Ambrosio R, Leone M, Rosso MG, Mittino D, Brignolio F. Disability and quality of life in hereditary ataxias: a self-administered postal questionnaire. Intl Disabil Stud 1987; 9:10-4.

8. Gompertz P, Pound P, Ebrahim S. A postal version of the Barthel Index. Clin Rehab 1994;8:233-239.

9. Goldberg DP. Manual of the General Health Questionnaire. Windsor: NFER-Nelson, 1978.

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

11. Ware JE Jr, Snow KK, Kosinski M, Gandek B. SF-36 Health Survey manual and interpretation guide. Boston, Massachusetts: Nimrod Press, 1993.

12. Dillman DA. Mail and telephone surveys: the total design method. New York: Wiley: 1978.

13. Collin C, Wade DT, Davies S, Horne V. The Barthel ADL Index: A reliability study. International Disability Studies 1988;10:61-63

14. Hobart JC, Lamping DL, Thompson AJ. Measuring disability in neurological disease: validity of the self-report Barthel index. J Neurol 1996; 243(suppl 2):S25.

15. Goldberg, DP Williams PA User's Guide to the GHQ. Windsor: NFER-Nelson;1988.

16. Shiely J-C, Bayliss MS, Keller SD, Tsai C, Ware JE. SF-36 Health Survey Annotated Bibliography:First Edition (1998-1995) Boston, MA: The Health Institute, New England Medical Center, 1996.

17. Rabin R, de Charro F. EQ-5D: a measure of health status from the EuroQol Group. Ann Med 2001;33(5):337-43.

18. Goldberg DP, Gater R, Sartorius N et al. The validity of two versions of the GHQ in the WHO study of mental illness in general health care. Psychol Med 1997; 27:191-7.

19. Wade DT. Measurement in neurological rehabilitation. Oxford: Oxford University Press, 1992.

20. Garrat AM, Ruta DA, Abdalla MI, Buckingham JK, Russell IT. The SF-36 health survey questionnaire: an outcome measure suitable for routine use within the NHS? Br Med J 1993; 306: 1440-4.

21. EuroQol group. EuroQol a new facility for the measurement of health related quality of life. Health Policy 1990; 16: 199-208.

22. Hobart JC, Lamping DL, Fitzpatrick R, Riazi A, Thompson AJ. The Multiple Sclerosis Impact Scale (MSIS-29); a new patient based outcome measure. Brain 2001;124:962-73.

23. Hobart JC, Freeman J, Lamping DL, Fitzpatrick R, Thompson AJ. The SF-36 in multiple sclerosis: why basic assumptions must be tested. J Neurol Neurosurg Psychiatry 2001;71:363-70.

24. Hobart JC, Freeman J, Thompson A. Kurtzke scales revisited: the application of psychometric methods to clinical intuition. Brain 2000;123:1027-1040.

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

26. Likert RA. A technique for the development of attitudes. Arch Psyhol 1932;140:5-55.

27. Ware JE, Harris WJ, Gandek B, Rogers BW: MAP-R for Windows: Multitrait- Multi-Item Analysis Program- Revised User’s Guide. Item Analysis Program - Revised User's Guide. MA: Boston, Health Assessment Lab 1997.

28. Ware JE Jr, Davies-Avery A, Donald CA. Conceptualization and measurement of health for adults in the health insurance study: Vol. V, general health perceptions. Santa Monica, California: The Rand Corporation; 1978.

29. Stewart AL, Ware JR Jr eds. Measuring functioning and well-being: the medical outcomes study approach. Duke University Press: Durham, North Carolina; 1992.

30. Eisen M, Ware JE Jr, Donald CA, Brook RH. Measuring components of children's health status. Med Care 1979;17:902-21.

31. Holmes WC, Bix B, Shea JA. SF-20 score and item distributions in a human immunodeficiency virus-seropositive sample. Med Care 1996;34:562-69.

32. Guilford JP. Psychometric methods, second edition. New York: McGraw-Hill;1954.

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

34. Nunnally JC, Bernstein IH. Psychometric theory., 3rd ed. New York: McGraw-Hill; 1994.

35. Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychol Bull 1955;52:281- 302.

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

37. Cohen J. Statistical power analysis for the behavioural sciences, First edition. Hillside, New Jersey: Lawrence Erlbaum; 1969.

38. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res 1995;4 293-307.

39. Riazi A, Hobart JC, Lamping DL, Fitzpatrick R, Freeman JA, Jenkinson C, Peto V, Thompson AJ. Using the SF-36 measure to compare the health impact of multiple sclerosis and Parkinson's disease with normal population health profiles. J Neurol Neurosurg Psychiatry 2003;74:710-714.

40. Hobart JC, Cano SJ, Thompson AJ. What sample sizes for reliability and validity studies? Qual Life Res 2002;11:636.

41. Cano SJ, Warner TT, Hobart JC. What sample sizes for reliability and validity studies II: a prospective study. Qual Life Res 2003;12:771.

42. Nunnally J, Bernstein I. Psychometric theory. Third ed. New York: McGraw-Hill; 1994.

Table 1. Correlations between measures predicted *a priori**

Instrument	Scale / dimension / variable	**Barthel**	**GHQ-12**	**EuroQol-Thermometer**	**EuroQol-Health State**	**SF-36 PF**	**SF-36 RL-P**	**SF-36 BP**	**SF-36 GHP**	**SF-36 VT**	**SF-36 SF**	**SF-36 RL-E**	**SF-36 MH**
GHQ-12		L	-	-	-	-	-	-	-	-	-	-	-
EuroQol	Thermometer	M	M	-	-	-	-	-	-	-	-	-	-
	Health state	H	L	M	-	-	-	-	-	-	-	-	-
SF-36 [1]	PF	H	L	M	M	-	-	-	-	-	-	-	-
	RL-P	H	L	M	M	H	-	-	-	-	-	-	-
	BP	M	M	M	M	M	M	-	-	-	-	-	-
	GHP	L	M	M	M	M	M	M	-	-	-	-	-
	VT	L	M	M	M	M	M	M	M	-	-	-	-
	SF	L	M	M	M	M	M	M	M	M	-	-	-
	RL-E	L	H	M	M	L	L	L	M	M	M	-	-
	MH	L	H	M	M	L	L	L	M	M	M	H	-
Demographic variables	Age	L	L	L	L	L	L	L	L	L	L	L	L
	Sex	L	L	L	L	L	L	L	L	L	L	L	L
	Years since diagnosis	M	L/M	M	M	M	M	L	L/M	L/M	L/M	L/M	L/M

*SF-36 scales are: physical functioning (PF), role limitations [physical problems] (RL-P), bodily pain (BP), general health perception (GHP), vitality (VT), social functioning (SF), role limitations [emotional problems] (RL-E), and mental health (MH).

[1] Medical Outcomes Study 36-item Short Form Health Survey: high scores = better health.

L = < 0.30; M = 0.30 -0.70; H = > 0.70

Table 2 Sample characteristics

Characteristics	
Age (n = 56)	
mean (SD)	31.0 (8.6)
range	18 - 57
Sex (n = 56)	
% female	57.1
Yrs since FA diagnosed (n = 55)	
mean (SD)	13.3 (8.5)
range	1 - 32
Yrs since FA started (n = 49)	
mean (SD)	18.1 (8.4)
range	4 - 38
Ethnicity (n = 56)	94.6
% white	
Employment status (n=54) %	
Employed/self employed	40.8
Retired due to FA	25.9
Retired for other reasons	1.9
Unemployed	18.5
Student	13.0
Education (n = 56) %	
Educated after minimum school leaving age	64.3
Degree / equivalent qualification	25.0
Marital status (n=56) %	
Single	58.9
Married	21.4
Divorced	5.4
With a partner	14.3
Mobility indoors (n = 56) %	
Walks unaided	14.3
Walks with an aid	30.4
Wheelchair user	55.4

Table 3a *Data quality, scaling assumptions, acceptability and reliability of health status measures*

	Health status measures			
	Barthel	**GHQ-12**	**EuroQol Thermometer**	**EuroQol Health State**
N	47	51	55	45
Data quality				
Item missing data %	0-14.3	1.8-3.6	1.8	0-19.6
Computable scale scores%	84%	91%	98%	80%
Scaling assumptions				
Item mean scores	0.43-2.34	1.78-2.31	N/A	N/A
Item SD	0.31-1.04	0.37-0.92	N/A	N/A
Item-total correlation	0.26-0.84	0.45-0.83	N/A	N/A
Acceptability				
Scale range	0-100	0-100	0-100	-0.59-1.00
Score range	10-100	0-78	20-95	-0.09-1.00
Mean score (SD)	67.3 (23.6)	36.1 (16.9)	64.3 (19.1)	0.53 (0.30)
Floor/ceiling %	0/8.5	0/2.0	0/0	0/2.2
Skewness[2]	-0.580	0.678	-0.748	-0.902
Reliability				
Alpha	0.87	0.92	N/A	0.58
SEM[3]	8.5	4.8	N/A	N/A
95% CI[4]	±16.7	±9.4	N/A	N/A

[2] It is recommended that skewness statistic ranges from -1 to +1
[3] Standard Error of Measurement = SD x √ 1 - alpha
[4] 95% Confidence Interval = ± 1.96 x SEM

Table 3b *Data quality, scaling assumptions, acceptability and reliability of health status measures (SF-36)*

	SF-36 Dimensions							
	PF	**RL-P**	**BP**	**GHP**	**VT**	**SF**	**RL-E**	**MH**
N	54	55	56	55	55	56	53	55
Data quality								
Item missing data %	1.8-5.4	0-1.8	0-3.6	0-1.8	1.8-3.6	0	3.6-5.4	1.8
Range computable scale scores%	96.4	98.2	100	98.2	98.2	100	94.6	98.2
Scaling assumptions								
Item mean scores	1.14-2.10	1.42-1.65	4.84-4.99	1.93-3.67	3.11-3.78	3.75-3.96	1.68-1.81	3.70-5.02
Item SD	0.45-0.78	0.48-0.51	1.19-1.19	1.04-1.46	1.26-1.33	1.15-1.21	0.40-0.47	1.11-1.41
Item-total correlation	0.24-0.87	0.60-0.85	0.81	0.44-0.69	0.70-0.80	0.69	0.56-0.70	0.61-0.88
Acceptability								
Scale range	0-100	0-100	0-100	0-100	0-100	0-100	0-100	0-100
Score range	0-90	0-100	22-100	5-100	0-85	0-100	0-100	24-100
Mean score (SD)	21.7 (23.7)	50.9 (42.5)	78.3 (22.6)	46.8 (23.4)	49.8 (22.3)	71.4 (27.3)	73.0 (37.6)	67.2 (21.3)
Floor/ceiling %	20.4/0	27.3/36.4	0/39.3	0/1.8	1.8/0	1.8/28.6	15.1/58.5	0/1.8
Skewness[5]	1.401	0.059	-0.814	0.217	-0.474	-0.885	-1.056	-0.404
Reliability								
Alpha	0.92	0.88	0.90	0.79	0.88	0.82	0.80	0.89
SEM[6]	6.7	14.7	7.1	10.7	7.7	11.6	16.8	7.1
95% CI[7]	± 13.1	± 28.9	± 14.0	± 21.0	± 15.1	± 22.7	± 33.0	± 13.8

[5] It is recommended that skewness statistic ranges from -1 to +1
[6] Standard Error of Measurement = SD x √ 1 - alpha
[7] 95% Confidence Interval = ± 1.96 x SEM

Table 4 Convergent and discriminant construct validity of the health status measures

Instrument	Scale / dimension / variable	**Barthel**	**GHQ-12**	**EuroQol-Thermometer**	**EuroQol-Health State**	**SF-36 PF**	**SF-36 RL-P**	**SF-36 BP**	**SF-36 GHP**	**SF-36 VT**	**SF-36 SF**	**SF-36 RL-E**	**SF-36 MH**
GHQ-12		-0.03	-	-	-	-	-	-	-	-	-	-	-
EuroQol	Thermometer	0.08	-0.38	-	-	-	-	-	-	-	-	-	-
	Health state	0.70	-0.28	0.39	-	-	-	-	-	-	-	-	-
SF-36 [8]	PF	0.55	-0.07	0.12	0.54	-	-	-	-	-	-	-	-
	RL-P	0.14	-0.41	0.26	0.29	0.11	-	-	-	-	-	-	-
	BP	-0.06	-0.31	0.24	0.13	-0.08	0.25	-	-	-	-	-	-
	GHP	-0.00	-0.39	0.43	0.30	0.01	0.27	0.37	-	-	-	-	-
	VT	0.06	-0.65	0.49	0.48	0.23	0.39	0.42	0.38	-	-	-	-
	SF	0.00	-0.60	0.49	0.43	0.11	0.45	0.32	0.42	0.61	-	-	-
	RL-E	0.21	-0.57	0.38	0.29	-0.01	0.36	0.12	0.39	0.41	0.53	-	-
	MH	0.09	-0.75	0.32	0.35	0.14	0.22	0.24	0.40	0.59	0.57	0.62	-
Demographic variables	Age	-0.22	-0.17	-0.01	0.02	-0.07	-0.13	0.00	0.16	0.30	0.11	0.06	0.31
	Sex	0.11	-0.08	0.22	0.26	-0.00	0.24	0.05	0.35	0.10	0.32	0.31	0.04
	Years since diagnosis	-0.67	-0.11	-0.09	-0.47	-0.48	-0.15	0.13	0.14	0.18	0.02	-0.05	0.07

[1] Medical Outcomes Study 36-item Short Form Health Survey: high scores = better health.

Table 5

Barthel Index, GHQ-12, EurQol-5D (Transformed 0-100)

Responsiveness (T-Test and Effect Size)

	Measures			
	Barthel Index	**GHQ-12**	**EuroQol Thermometer**	**EuroQol Health State**
N	43	43	41	36
Mean score time 1 (SD)	69.8 (20.1)	36.9 (18.4)	63.4 (18.8)	0.58 (0.25)
Mean score time 2 (SD)	68.0 (18.5)	32.8 (15.7)	65.3 (19.1)	0.55 (0.25)
Time 1 – Time 2 difference (SD)	1.74 (8.16)	4.13 (13.20)	-1.88 (16.6)	0.03 (0.19)
t-test	1.40	2.05	-0.72	1.05
p	0.17	0.05	0.47	0.30
Effect size	0.09	0.23	0.10	0.13

Table 6

SF-36 (0-100)

Responsiveness (T-Test and Effect Size)

	SF-36 Dimensions							
	RL-E	**RL-P**	**BP**	**VT**	**GHP**	**SF**	**PF**	**MH**
N	42	43	43	42	43	43	42	42
Mean score time 1	76.2	51.2	79.0	47.9	48.5	69.5	22.1	66.5
(SD)	(34.8)	(41.9)	(21.7)	(22.8)	(23.8)	(29.4)	(22.9)	(21.6)
Mean score time 2	75.4	51.2	74.0	50.0	42.8	75.6	20.1	66.7
(SD)	(36.9)	(35.8)	(25.2)	(21.3)	(24.6)	(25.0)	(22.8)	(23.6)
Time 1 – Time 2	0.79	0.00	5.02	-2.02	5.67	-6.11	2.02	-0.21
difference (SD)	(34.9)	(42.6)	(21.5)	(18.5)	(16.8)	(29.9)	(19.6)	(15.5)
t-test	0.15	0.00	1.54	-0.71	2.22	-1.34	0.67	-0.09
p	0.88	1.00	0.13	0.48	0.03	0.19	0.52	0.93
Effect size	0.01	0.00	0.06	0.04	0.12	0.09	0.09	0.01