

Measurement properties of selected patient-reported outcome measures for use in randomised controlled trials in patients with systemic lupus erythematosus: a systematic review

Vibeke Strand ¹, Lee S Simon,² Alexa Simon Meara,³ Zahi Touma ⁴

To cite: Strand V, Simon LS, Meara AS, *et al.* Measurement properties of selected patient-reported outcome measures for use in randomised controlled trials in patients with systemic lupus erythematosus: a systematic review. *Lupus Science & Medicine* 2020;**7**:e000373. doi:10.1136/lupus-2019-000373

Received 27 November 2019
Revised 17 April 2020
Accepted 24 May 2020



© Author(s) (or their employer(s)) 2020. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

¹Division of Immunology/Rheumatology, Stanford University, Palo Alto, California, USA

²SDG, LLC, Cambridge, Massachusetts, USA

³Rheumatology, The Ohio State University, Columbus, Ohio, USA

⁴Department of Medicine, University of Toronto, Toronto, Ontario, Canada

Correspondence to

Dr Vibeke Strand;
vibekestrand@me.com

ABSTRACT

Objective The heterogeneous multisystem manifestations of SLE include fatigue, pain, depression, sleep disturbance and cognitive dysfunction, and underscore the importance of a multidimensional approach when assessing health-related quality of life. The US Food and Drug Administration has emphasised the importance of patient-reported outcomes (PROs) for approval of new medications and Outcome Measures in Rheumatology has mandated demonstration of appropriate measurement properties of selected PRO instruments.

Methods Published information regarding psychometric properties of the Medical Outcomes Survey Short Form 36 (SF-36), Lupus Quality of Life Questionnaire (LupusQoL) and Functional Assessment of Chronic Illness Therapy-Fatigue Scale (FACIT-F), and their suitability as end points in randomised controlled trials (RCTs) and longitudinal observational studies (LOS) were assessed. A search of English-language literature using MEDLINE and EMBASE identified studies related to development and validation of these instruments. Evidence addressed content validity, reliability (internal consistency and test-retest reliability), construct validity (convergent and divergent) and longitudinal responsiveness, including thresholds of meaning and discrimination.

Results All instruments demonstrated strong internal consistency, reliability and appropriate face/content validity, indicating items within each instrument that measure the intended concept. SF-36 and LupusQoL demonstrated test-retest reliability; although not published with FACIT-F in SLE supported by evidence from other rheumatic diseases. All instruments demonstrated convergent validity with other comparable PROs and responsiveness to treatment.

Conclusion The measurement properties of PRO instruments with published data from RCTs including: SF-36, LupusQoL and FACIT-F indicate their value as secondary end points to support labelling claims in RCTs and LOS evaluating the efficacy of SLE treatments.

INTRODUCTION

SLE is a chronic autoimmune disease that affects multiple organ systems and significantly impacts patient-reported health-related

quality of life (HRQoL). The clinical manifestations of SLE are heterogeneous, vary over time, and may include fatigue, pain, depression, sleep disturbance and cognitive dysfunction.¹ The multisystemic nature of SLE poses a challenge for evaluating treatment benefit and underscores the importance of using a multidimensional approach when assessing HRQoL. In 1998, the Outcome Measures in Rheumatology (OMERACT) international consensus effort recommended five domains for assessment in all randomised controlled trials (RCTs) and longitudinal observational studies (LOS) in SLE, including disease activity, damage, HRQoL, adverse events and economic costs.² OMERACT also recommended that both generic and disease-specific instruments be used to assess HRQoL.

Since the release of the OMERACT recommendations, some RCTs in SLE have included patient-reported outcomes (PROs), such as the Medical Outcomes Survey Short Form 36 (SF-36),³ the Lupus Quality of Life questionnaire (LupusQoL)⁴ and the Functional Assessment of Chronic Illness Therapy-Fatigue Scale (FACIT-F).⁵ The US Food and Drug Administration guidance for PRO measures outlines the methodology and evidence needed to support labelling claims for new treatments⁶ and emphasises the importance of demonstrating content validity, reliability, construct validity and responsiveness of the measure among the target population.

Existing PROs in SLE measure patient perceptions of their health conditions and assess a spectrum of HRQoL—pain, fatigue, anxiety, depression, physical function, cognitive function and others. Current SLE PROs can be grouped as disease-specific and generic. Among the generic, SF-36 is most commonly used in research settings and RCTs, as well

as EuroQol Five-Dimensional Questionnaire (EQ-5D).⁷ While several SLE-specific HRQoL questionnaires have been developed and validated, including LupusQoL, SLE-specific Quality of Life Questionnaire (SLE-QoL),⁸ SLE Quality of Life Questionnaire (L-QoL),⁹ LupusPRO and Lupus Impact Tracker (LIT),¹⁰ we focused on legacy measures and only those with publicly available data from RCTs: SF-36, LupusQoL and FACIT-F.

The objective of this analysis was to summarise available evidence supporting the psychometric properties of SF-36, LupusQoL and FACIT-F in SLE and to assess their suitability as secondary end points in RCTs to support labelling claims for SLE treatments.

METHODS

Search strategy

This review used the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines.¹¹ The search strategy was developed in consultation with a medical librarian with expertise in systematic reviews (online supplementary appendix 1). A search of English-language published literature using MEDLINE (1995 to June 2017) and EMBASE (1964 to June 2017) was conducted to identify:¹ studies related to development and validation of SF-36, LupusQoL and FACIT-F in SLE; and² RCTs and LOS in SLE that included these instruments. We excluded studies if they were non-English articles, publications written only in abstract form, conference letters, editorials, dissertations and case reports with <20 patients. Search terms were individualised for each database and for MEDLINE. Titles and abstracts of initially identified studies were screened and reviewed to further identify articles that could be included in the final literature review and synthesis.

Psychometric properties of included instruments

To assess the psychometric properties of each instrument, evidence of content validity, reliability (internal consistency and test-retest reliability), construct validity (convergent, divergent and known-group validity) and longitudinal responsiveness was extracted.^{12–13} Convergent validity was judged appropriate if positive correlations between instrument were present and >0.6 and discriminant validity if correlations were <0.3. For internal consistency, Cronbach's α >0.7 was considered acceptable.¹² An intraclass correlation coefficient (ICC) >0.7 for test-retest reliability was interpreted as acceptable.¹² Longitudinal responsiveness was evaluated using standardised response means (SRMs) and interpreted as poor if SRMs<0.5, moderate if SRMs \geq 0.5 and high if SRMs \geq 0.8 (14).¹⁴ Thresholds of meaning, particularly minimum clinically important difference (MCID) and minimum important differences (MIDs) are also presented. In discussions of clinically relevant thresholds for outcomes scores, MCIDs refer to approaches based on the patient perspective/perception^{15–16} while MIDs are not based on clinical judgement (eg, perceptions of patients or clinicians) and

generally use approaches that are anchored to a statistical change defined as $p<0.05$ or based on a change in a laboratory marker or a functional test ≥ 0.5 SD.^{17–18} It is important to note that anchor-based methods are recommended for derivation of MCID definitions.¹⁸ Discrimination of studied instruments in RCTs was also assessed. Instrument measurement properties of SF-36 (table 1), LupusQoL (table 2) and FACIT-F are included in the online appendix table A1. Selected RCTs and LOS in SLE that included these instruments are presented in online appendix tables A2 and A3, respectively.

RESULTS

Content validity

Content validity of SF-36 in patients with SLE was examined in two studies that included factor analysis of its domains. Results from the first study indicated that there were four significant factors with eigenvalues >3.3 that could be meaningfully interpreted, including 'physical functioning', 'physical and emotional role functioning', 'mental and social health' and 'general health'.¹⁹ The second assessed a Chinese version of SF-36 and demonstrated that the eight-domain structure of SF-36 was supported with the overall factor loadings as 'Physical Functioning', 'Role-Physical', 'Bodily Pain' and 'Role-Emotional'. These scales loaded cleanly onto one factor, while all other domains loaded on two factors.

Reliability (internal consistency and test-retest)

Evidence of internal consistency for SF-36 was demonstrated in three studies (online supplementary table S1), based on three different language versions, including Japanese,²⁰ Chinese²¹ and English.¹⁹ Cronbach's α for domains ranged from 0.72 to 0.96. Test-retest reliability of SF-36 (Spearman's rank correlation) measured in these three studies ranged from 0.65 to 0.90 (table 1).

Construct and known-group validity

Six studies^{22–27} assessed convergent validity of SF-36 (online supplementary table S2). Convergent validity with the LupusQoL was tested and established in five studies and demonstrated correlations of 0.48–0.83 between comparable domains in both questionnaires (SF-36/LupusQoL: Physical Functioning and Physical Health, Role-Emotional and Emotional Health, Bodily Pain and Pain, and Vitality and Fatigue).^{22–25–27} Comparison of component summary scores (Physical Component Subscale (PCS) and Mental Component Subscale (MCS)) and EQ-5D values indicated that PCS Score correlated more strongly than MCS Score with both EQ-5D ($r=0.72$ vs 0.49) and EQ-5D Visual Analogue Scale (VAS) ($r=0.61$ vs 0.37).²⁶ Two studies examined SF-36s divergent validity comparing individual domain scores with SLE Disease Activity Index 2000 (SLEDAI-2K) Scores and showed no significant correlations between the two measures.^{20–27} Using Systemic Lupus International Collaborating Clinics Damage Index (SDI), Baba *et al*²⁰ reported weak-to-moderate inverse correlations between SDI and

Table 1 SF-36: assessment of instrument properties in SLE

Author, year	Description of study	Reliability validity			Known groups	Responsiveness and ability to detect change	MCID/MID
		Internal consistency	Test-retest	Construct			
Baba <i>et al</i> , 2018 ²⁰	Prospective study in Japanese patients with SLE (n=233). SF-36 completed at baseline and 1 year follow-up	✓	✓	✓		✓	
Furie <i>et al</i> 2014 ³⁵	Secondary analysis of pooled data from the BLISS Trials in patients with SLE (n=1684). Changes in clinical and HRQoL measures from baseline to week 52 were compared between SRI responders and non-responders					✓	
Nantes <i>et al</i> 2018 ²²	Prospective study in patients with SLE (n=78). SF-36 and LupusQoL completed at baseline and follow-up			✓			
McElhone <i>et al</i> 2016 ²³	Prospective, longitudinal study in patients with SLE (n=101) experiencing a flare. SF-36 completed every 4 weeks for 9 months			✓		✓	✓
Devilliers <i>et al</i> 2015 ²⁹	Prospective study in 185 patients with SLE (n=185). SF-36 completed monthly for 3 months					✓	✓
Yilmaz- Oner <i>et al</i> , 2016 ²⁴	Cross-sectional study in Turkish patients (n=113) with SLE. SF-36 and LupusQoL completed once during a single visit			✓			
Garcia-Carrasco <i>et al</i> , 2012 ²⁵	Cross-sectional study in women with SLE (n=127). SF- 36 and LupusQoL completed once during study initiation			✓			
Hanly <i>et al</i> , 2011 ²⁸	Prospective international study (n=274) to evaluate change in HRQoL in association with neuropsychiatric events in newly diagnosed patients with SLE				✓	✓	
Touma <i>et al</i> , 2011 ²⁷	Longitudinal study in SLE (n=41). SF-36 and LupusQoL completed monthly for 12 months			✓		✓	
Wolfe <i>et al</i> , 2010 ²⁶	Longitudinal study in SLE (n=1316). Patients followed semi-annually for 10 years. A single random observation from each patient was included in analysis for SF-36 and EQ-5D			✓			
Colangelo <i>et al</i> , 2009 ³⁰	Prospective study in patients with SLE (n=202). SF-36 completed at two consecutive visits annually						✓
Strand <i>et al</i> , 2005 ³¹	Post hoc analysis of 2 clinical trials of abetimus sodium (n=298, phase III; n=189, phase II/III). SF-36 completed at 6 months and 12 months; responders and non- responders were identified from each trial						✓
Thumboo <i>et al</i> , 2000 ²¹	Cross-sectional study in Chinese patients with SLE (n=69), Chinese version of SF-36 completed twice within 7–14 days	✓	✓	✓			
Thumboo <i>et al</i> , 1999 ¹⁹	Cross-sectional study in Asian patients with SLE (n=118). SF-36 completed twice within a 14-day period	✓	✓	✓			

✓=Instrument property assessed in study.

EQ-5D, EuroQol Five-Dimensional Questionnaire; HRQoL, health-related quality of life; LupusQoL, Lupus Quality of Life questionnaire; MCID, minimum clinically important difference; MID, minimum important difference; SF-36, 36 Item Health Survey-Short Form; SRI, SLE Responder Index .

SF-36 domain scores ($r=-0.08$ to -0.47 , $p<0.05$). Two other studies that used SDI as a comparator indicated no significant correlations with SF-36 domains.^{19 26} Two studies correlated results from the British Isles Lupus Assessment Group (BILAG) general and organ subscales

with those from SF-36. Thumboo *et al*²¹ reported correlations ranging from -0.34 to 0.17 across SF-36 domains and BILAG General; and Thumboo *et al*¹⁹ reported correlations between -0.07 and -0.36 (all $p<0.05$ except for Physical Functioning and General Health) (online

Table 2 SF-36: responsiveness and ability to detect change in SLE: (within-group)

Reference	SF-36 domain	Anchor (change in disease status)	Findings		
			Improved	Unchanged	Worsened
				<3 change in SLEDAI-2K (n=208) ES; SRM; p	≥3 change in SLEDAI-2K (n=25) ES; SRM; p
Baba <i>et al</i> , 2018 ²⁰	PF	Clinical worsening defined by a change ≥3 in the SLEDAI-2K. The difference within each first-year and second-year pair assessed by paired t-test	Not reported	-0.04; -0.05 ;0.53	0.04; 0.07; 0.71
	RP			0.00; 0.00; 0.97	0.19; 0.27; 0.18
	BP			0.15; 0.16 ;0.02	-0.12; -0.11; 0.58
	GH			-0.04; -0.04; 0.58	-0.06; -0.06 ;0.76
	VT			0.10; 0.11; 0.11	0.12; 0.14; 0.49
	SF			0.11; 0.11 ;0.11	-0.31; -0.38; 0.07
	RE			0.08; 0.08; 0.26	-0.08; 0.07 ;0.74
	MH			0.06; 0.07; 0.29	0.00; 0.00; 0.99
	PCS			0.00; 0.00; 0.95	0.06; 0.10; 0.63
	MCS			0.12; 0.12; 0.08	-0.21; -0.20; 0.33
				0 change in SDI (n=204) ES; SRM; p	≥1 change in SDI (n=29) ES; SRM; p
	PF	Clinical worsening defined by a change ≥1 in the SDI. The difference within each first-year and second-year pair assessed by paired t-test	Not reported	-0.02; -0.03 ;0.72	-0.08; -0.10; 0.61
	RP			0.05; 0.05; 0.45	-0.09; -0.08; 0.65
	BP			0.16; 0.18; 0.01	-0.17; -0.13; 0.47
	GH			0.01; 0.01; 0.88	-0.40; -0.30; 0.12
	VT			0.15; 0.17; 0.02	-0.22; -0.21; 0.26
	SF			0.11; 0.12; 0.19	-0.28; -0.22; 0.24
	RE			0.09; 0.09; 0.20	-0.11; -0.11; 0.58
	MH			0.08; 0.09; 0.20	-0.09; -0.10; 0.60
	PCS			0.03; 0.03; 0.65	-0.11; -0.11; 0.54
MCS			0.15; 0.15; 0.03	-0.28; -0.24; 0.21	
		(2 to 7); mean (95% CI)	(-1 to 1); mean (95% CI)	(-7 to -2); mean (95% CI)	
McElhone <i>et al</i> , 2016 ²³	PF	Patient completed Global Rating of Change, ranging from 7 (a very great deal better) to -7 (a very great deal worse) with 0 indicating no change	5.6 (3.9 to 7.3)	1.2 (0.3 to 2.2)	-3.0 (-4.3 to -1.6)
	RP		14.7 (9.9 to 19.5)	1.4 (-1.3 to 4.0)	-9.9 (-15.3 to -4.5)
	BP		13.0 (10.6 to 15.4)	2.8 (1.1 to 4.5)	-7.0 (-9.3 to -4.7)
	GH		3.4 (2.2 to 4.6)	0.3 (-0.6 to 1.2)	-2.0 (-3.2 to -0.8)
	VT		11.2 (8.4 to 14.0)	0.9 (20.8 to 2.6)	-4.6 (-6.3 to -2.8)
	SF		10.1 (7.0 to 13.2)	1.6 (0.1 to 3.2)	-7.0 (-10.8 to -3.1)
	RE		11.3 (6.6 to 15.9)	2.6 (-0.3 to 5.4)	-10.1 (-15.9 to -4.3)
	MH		7.6 (5.9 to 9.4)	-0.1 (-1.6 to 1.4)	-5.5 (-7.5 to -3.6)
		SRM; mean variation; *p<0.05	(SRM; mean variation) *p<0.05		
Devilliers <i>et al</i> , 2015 ²⁹	PF	Patient completed 7-point VAS of change in lupus-related health status over 3 months. A difference +0.5 SD or more was considered worsening; a VAS with a difference of -0.5	0.40; +5.3*		-0.59; -7.8*
	RP		0.50; +16.9*		-0.34; -11.7*
	BP		0.57; +12.2*		-0.62; -13.2*
	GH		0.37; +4.9*		-0.72; -9.5*
	VT		0.32; +5.9*		-0.41; -7.6*
	SF		0.36; +8.1*		-0.27; -6.1

Continued

Table 2 Continued

Reference	SF-36 domain	Anchor (change in disease status)	Findings		
			Improved	Unchanged	Worsened
	RE	SD considered an improvement Patient-completed 4-point Likert Symptom Scale ranging from 0 (no problems) to 3 (severe problems)	0.58; +20.9*		-0.14; -4.9
	MH		0.32; +5.5*		-0.35; -5.9*
	PCS		0.44; +2.8*		-0.69; -4.4*
	MCS		0.43; +4.2*		-0.2; -1.9*
	PF		0.21; 3.0		-0.44; -6.2*
	RP		0.39; +13.2*		-0.51; -17.2*
	BP		0.58; +11.8*		-0.63; -13.0*
	GH		0.16; 2.2		-0.43; -6.1*
	VT		0.28; +5.0*		-0.39; -6.9*
	SF		0.47; +10.6*		-0.34; -7.7*
	RE		0.36; +13.0*		-0.26; -9.3*
	MH		0.41; +7.0*		-0.31; -5.3*
	PCS		0.29; +1.9*		-0.61; -3.9*
	MCS		0.41; +3.9*		-0.25; -2.4*
				Mean (SD)	Mean (SD)
Hanly <i>et al</i> , 2011 ²⁸	PCS	Physician neuropsychiatric event questionnaire*	1.73 (SD=0.71)	0	-0.62 (SD=1.58)
	MCS		3.66 (SD=0.89)	0	-4.00 (SD=1.97)
			Improved ES; SRM	Remission/unchanged ES; SRM/ES; SRM	Flare ES; SRM
Touma <i>et al</i> , 2011 ²⁷	PF	SLEDAI-2K 30 days, where improvement was defined as reduction in SLEDAI-2K \geq 4 from the previous visit, flare as an increase in SLEDAI-2K \geq 4 from the previous visit, remission as SLEDAI-2K=0, and unchanged for the rest of the patient visits	0.05; 0.23	0.03; 0.03/0.00; 0.00	0.07; 0.12
	RP		0.16; 0.30	0.05; 0.10/0.04; 0.05	0.50; 0.64
	BP		0.02; 0.06	0.02; 0.02/0.01; 0.02	0.03; 0.04
	GH		0.10; 0.40	0.00; 0.00/0.00; 0.01	0.02; 0.08
	VT		0.15; 0.30	0.02; 0.02/0.00; 0.01	0.09; 0.18
	SF		0.09; 0.24	0.04; 0.05/0.00; 0.01	0.20; 0.42
	RE		0.00; 0.00	0.02; 0.03/0.01; 0.02	0.16; 0.18
	MH		0.20; 0.43	0.05; 0.09/0.01; 0.03	0.03; 0.04
	PCS		0.04; 0.09	0.02; 0.03/0.00; 0.00	0.20; 0.30
	MCS		0.14; 0.60	0.02; 0.03/0.05; 0.10	0.02; 0.03

*A physician-generated 7-point Likert scale for NP events comparing the change in NP status between the onset of the event and time of study assessment was available for each NP event (1=patient demise, 2=much worse, 3=worse, 4=no change, 5=improved, 6=much improved, 7=resolved).

†MCID 1 by Strand *et al*; MCID 2 by Devilliers *et al*; MCID 3 by McElhone *et al*.

BP, bodily pain; ES, effect size; GH, general health; MCID, minimal clinically important difference; MCS, Mental Component Subscale; MH, mental health; PCS, Physical Component Subscale; PF, physical function; RE, role emotional; RP, role physical; SDI, Systemic Lupus International Collaborating Clinics Damage Index; SF, social functioning; SF-36, 36 Item Health Survey-Short Form; SLEDAI-2K, Systemic Lupus Erythematosus Disease Activity Index 2000; SRM, standardised response means; VAS, Visual Analogue Scale; VT, vitality.

supplementary table S3). One study reported known-group validity of SF-36 (online supplementary table S4). Results indicated that SF-36 scores could differentiate between neuropsychiatric events attributed to SLE and non-SLE causes. Changes in SF-36 component summary and domain scores, particularly those related to mental health, were strongly associated with the clinical outcome of neuropsychiatric events in SLE ($p < 0.05$ except for Role Physical) (table 1).²⁸

Longitudinal responsiveness

Five identified studies assessed the ability of SF-36 to detect changes over time (table 2).

Baba *et al*²⁰ assessed responsiveness of SF-36 over 1 year to clinical worsening defined by a change of ≥ 3 in the SLEDAI-2K or damage accrual defined by a change of ≥ 1 in SDI. Using SLEDAI-2K as an anchor of change in disease activity, effect sizes (ES) and SRM for SF-36 domain and component summary scores were generally < 0.20 in patients with clinical worsening and those who remained clinically unchanged. ES and SRM for social functioning and MCS Scores (≤ -0.20) suggested low responsiveness in patients whose SLEDAI-2K Scores changed by ≥ 3 . Using the SDI ≥ 1 (reflecting the accrual of damage compared with last assessment) as an anchor of

accrued damage over time, ES and SRM values for SF-36 domain and component summary scores were generally <0.19 in patients with evidence of damage worsening and patients who remained without damage change.²⁰

In another study, patients completed a Global Rating of Change (GRC) assessment and mean SF-36 domain scores were calculated for those with worsening (scores -7 to -2) versus improvement (scores 2 to 7). In all SF-36 domains, scores were significantly lower in the deterioration and unchanged groups versus those with improvements.²³ Devilliers *et al*²⁹ also assessed responsiveness of SF-36 domains using a 100 mm VAS of change in lupus-related health status over the past 3 months as the anchor. Patients reporting a difference $\geq+0.5$ SD were considered to report worsening health and those with a VAS difference ≥-0.5 SD improving health. For patients reporting improving health, significant improvements in Role Physical, Bodily Pain, Social Functioning, Role Emotional, and both PCS and MCS Scores were reported. In those reporting worsening health, significant decreases in Physical Functioning, Role Physical, Bodily Pain, Mental Health, Vitality, General Health domains PCS score were evident. Hanly *et al*²⁸ examined the ability to detect change in SF-36 using a physician-completed 7-point scale assessing neuropsychiatric events. Patients with neuropsychiatric improvement reported significant increases in PCS and MCS Scores. The responsiveness of SF-36 has also been assessed using SLEDAI-2K with improvements defined as reductions ≥ 4 from the previous visit, and worsening as increases ≥ 4 . Among patients with clinical worsening, SRMs of 0.64 were noted for Role Physical, 0.42 for Social Functioning and 0.30 for PCS Scores. Among those who improved, SRMs were 0.60 for MCS, 0.43 for Mental Health, 0.40 for General Health, 0.30 for Vitality, 0.30 for Role Physical, 0.24 for Social Functioning and 0.23 for Physical Functioning.²⁷

Thresholds of meaning of SF-36 Scores

Four studies assessed MCIDs (online supplementary table S5) or MID (online supplementary table S6) of SF-36. In one study, MCID was estimated using a patient-reported overall health status anchor: 'How would you describe your overall status since your last visit?' Response options included much better, somewhat better, about the same, somewhat worse and much worse. Those self-rated as somewhat better or somewhat worse were considered the 'minimally changed' subgroups. MCID for SF-36 was 2.1 (somewhat better) and -2.2 (somewhat worse) for PCS and 2.4 (somewhat better) and -1.2 (somewhat worse) for MCS Scores.³⁰ In a second study, MCIDs for domains and component summary scores of SF-36 were based on the 15-point global change scale (Guyatt feeling thermometer) corresponding to an improvement by a score of 6: 'a little better' and worsening by a score of 10: 'a little worse'. Clinically important improvement ranged from 6.7 to 11.4 points for domain scores and 3.4 to 4.9 for PCS. Clinically important worsening ranged from -14.7 to -1.7 points for domain scores and from -2.1 to

-0.8 for PCS and MCS, respectively.³¹ MCIDs for improvement were then defined as 2.5 for PCS and MCS and 5.0 for domain scores; for deterioration -1.8 for PCS and MCS and -2.5 for domain scores, respectively. McElhone *et al*²³ examined MID using both anchor-based and distribution-based methods. For deterioration, mean MID ranged from -2.0 for General Health to -11.1 for Role Physical domain scores. For improvement, they ranged from 2.8 for General Health to 10.9 for both Bodily Pain and Vitality. MIDs were larger using distributional versus anchor-based approaches. The MID for SF-36 has also been estimated as the mean change observed in the minimally improved and the minimally worse categories defined by a 7-point Likert Scale (-3 , much improved; -2 , moderately improved; -1 , minimally improved; 0 , the same; $+1$, minimally worse; $+2$, moderately worse and $+3$, much worse). MID for global improvement ranged from 1.9 to 11.3 for SF-36 domain scores. In patients reporting worsening, MIDs ranged from -4.4 to -15.6 .²⁹

Discrimination of SF-36 in RCTs and LOS

SF-36 has been the most frequently used HRQoL instrument in SLE trials (online appendix table A2). Twenty-six RCTs examined the impact of treatment on SF-36 results. Several examples are summarised in this section. Results from two of three trials that evaluated abetimus sodium indicated that SF-36 reflected clinical improvements in SLE accompanied by improvements \geq MCID in SF-36.³¹⁻³³ Furie *et al*³⁴ studied the safety and efficacy of belimumab 1 mg/kg and 10 mg/kg in patients with SLE and found significantly more SLE Responder Index (SRI-4) responders in the 10 mg/kg group versus placebo ($p=0.017$), but this difference was not sustained over 76 weeks. Results from a post hoc analysis of SRI-4 responders versus non-responders in these trials indicated that PCS, MCS and all SF-36 domain scores were significantly greater in SRI responders, across treatment groups, versus non-responders ($p<0.001$).³⁵ Both belimumab groups also reported similar improvements in SF-36 domain scores at week 52 versus placebo. Secondary analyses of these two RCTs indicated that changes from baseline to week 52 in SF-36 PCS Scores were significantly greater ($p<0.05$) in the belimumab arms versus placebo.¹³⁶

Seven RCTs examined the impact of physical activity, psychotherapy and alternate medicine interventions in patients with SLE with the objective of exploring improvements in SF-36. In three, physical training showed significant improvements in SF-36 Vitality and Role Physical Scores; however, significant differences in clinical improvements between treatment and control groups were reported in only three trials.³⁷⁻³⁹ Psychotherapy and cognitive behavioural therapy were tested in three RCTs. Improvements in SF-36 MCS Scores were demonstrated in two trials and clinical improvement demonstrated in one.⁴⁰⁻⁴² Greco *et al*⁴³ studied the benefits of acupuncture in reducing pain and fatigue in patients with SLE and reported significant improvements in SF-36 Bodily Pain and Vitality domains.

Table 3 LupusQoL: assessment of instrument properties in SLE

Author, year	Description of study	Reliability validity			Known-group	Responsiveness and ability to detect change	MCID/MID
		Internal consistency	Test-retest	Construct			
Meseguer <i>et al</i> , 2017 ⁴⁴	Cross-sectional study in Spanish patients with SLE (n=223). LupusQoL SLAQ, EQ-5D and NRS disease severity administered by postal survey	✓		✓			
Nantes <i>et al</i> , 2018 ²²	Prospective study in SLE (n=78). LupusQoL and SF-36 completed at baseline and follow-up visit			✓			
Anindito <i>et al</i> , 2016 ⁴⁵	Cross-sectional study of Indonesian patients with SLE (n=65)	✓	✓	✓			
McElhone <i>et al</i> , 2016; 2014a; 2014b ²³	Prospective study in SLE (n=101) experiencing a flare. LupusQoL administered at each of the 10 monthly visits					✓	✓
Devilliers <i>et al</i> , 2017 ⁴⁷	Prospective study in SLE (n=324), with data obtained 3 months apart					✓	
Devilliers <i>et al</i> , 2015 ²⁹	Prospective study in SLE (n=185). LupusQoL completed monthly for 3 months					✓	✓
Touma <i>et al</i> , 2011 ²⁷	Longitudinal study in SLE (n=41). LupusQoL and SF-36 completed monthly for 12 months			✓	✓	✓	
Jolly <i>et al</i> , 2010 ⁴⁶	Cross-sectional study in SLE (n=185) to adapt and assess the validity and reliability of the UK LupusQoL for use in USA. SF-36, EQ-5D and disease status assessed	✓	✓	✓			
McElhone <i>et al</i> , 2007 ⁴	Development and validation of the LupusQoL		✓	✓	✓		

✓=Instrument property assessed in study.

EQ-5D, EuroQol Five-Dimensional Questionnaire; LupusQoL, Lupus Quality of Life; MCID, minimum clinically important difference; MID, minimum important difference; NRS, Numerical Rating Scale; SF-36, 36-Item Health Survey-Short Form; SLAQ, Systemic Lupus Activity Questionnaire.

Content validity

The LupusQoL was developed from qualitative interviews with patients with SLE, as well as inputs from clinical experts and refined through cognitive interviews supporting content, followed by two rounds of psychometric testing.⁴ Patients reported that most items were relevant, easy to understand and answer, and reflected their HRQoL. Factor analysis in both English-speaking and Spanish-speaking populations confirmed the eight-domain structure.^{4 44} In the original derivation of LupusQoL, although only women of two racial ethnicities were involved, white and South Indians, subsequent validations included a wider population as well as men.⁴

Reliability (internal consistency and test-retest)

Internal consistency of LupusQoL domains was assessed in three studies⁴⁴⁻⁴⁶ and Cronbach's α ranged from 0.85 to 0.94 across studies and domains (online supplementary table S7). Assessment of the test-retest reliability of LupusQoL indicated intraclass correlations (ICCs) ranging from 0.68 to 0.95 (online supplementary table S7),^{4 45 46} (table 3).

Construct and known-group validity

Six studies assessed the convergent validity of LupusQoL (online supplementary table S8). Five used SF-36 for

assessment of construct validity, finding moderate-to-strong correlations between LupusQoL and the corresponding SF-36 domains ($r=0.38$ to 0.83). The LupusQoL also correlated strongly with the Systemic Lupus Activity Questionnaire (SLAQ) Symptom Scale ($r=-0.70$ to -0.76), EQ-5D Analogic Scale ($r=0.76$ to 0.80) and comparable EQ-5D domains ($r=0.50$ to 0.68). Associations between LupusQoL domain scores and SLEDAI-2K Scores were small and non-significant ($r=-0.02$ to 0.25) confirming divergent validity with SLE disease activity.²⁷

Two studies examined known-group validity of LupusQoL in patients with SLE (online supplementary table S9). In one, mean LupusQoL domain scores, with the exception of Intimate Relationships, did not significantly differentiate between improved versus same/worsened groups on SLEDAI-2K ($p>0.05$ for all domains).²⁷ Results from a second study indicated that the LupusQoL discriminated among groups of patients in different disease activity categories based on either BILAG Index or SDI Scores (those with SDI=0 and SDI \geq 1).⁴

Longitudinal responsiveness

Four validation studies assessed ability of LupusQoL to detect change (table 4).

Table 4 LupusQoL: responsiveness and ability to detect change in SLE

Reference	LupusQoL domain	Anchor (clinical severity measure)	Findings	
			Improved	Worsening
Devilliers <i>et al</i> , 2017 ⁴⁷			SRI non-responders SRM; mean variation	SRI responders SRM; mean variation
	PH	SRI responders were defined as ≥ 4 -point reduction in SELENA-SLEDAI Scores from baseline	0.42; 5.7	0.03; 0.5*
	PA		0.65; 12.6	0.08; 1.6*
	PL		0.18; 3.5	0.07; 1.2
	IR		-0.06; -0.7	0.02; 0.4
	BU		0.24; 6.0	0.18; 3.9
	EH		0.38; 4.5	0.01; 0.3
	BI		0.29; 4.7	0.13; 2.7
FA	0.26; 3.5		0.13; 2.5	
McElhone <i>et al</i> , 2016 ²³			(GRC Score 2 to 7); mean (95% CI)	(GRC Score -7 to -2); mean (95% CI)
	PH	Patient completed GRC, ranging from 7 (a very great deal better) to -7 (a very great deal worse) with 0 indicating no change	5.6 (4.2 to 7.1)	-3.7 (-5.2 to -2.1)
	PA		9.3 (7.1 to 11.5)	-6.5 (-8.9 to -4.1)
	PL		6.3 (3.9 to 8.8)	-4.6 (-7.0 to -2.2)
	IR		8.3 (4.3 to 12.4)	-7.7 (-14.7 to -0.6)
	BU		10.4 (7.7 to 13.1)	-4.6 (-6.9 to -2.3)
	EH		6.2 (4.7 to 7.8)	-4.4 (-6.0 to -2.7)
	BI		6.4 (3.6 to 9.2)	-2.5 (-4.2 to -0.8)
FA	8.9 (6.8 to 11.0)		-4.6 (-6.5, -2.8)(81, 256)	
			SRM; Mean Variation	SRM; Mean Variation
		Patient completed a 100 mm VAS to rate their health during the last 3 months. A difference +0.5 SD or more was considered worsening; a VAS with a difference of -0.5 SD was considered an improvement		
PH	0.47; +5.9*		-0.33; -4.1*	
PA	0.43; +7.5*		-0.27; -4.8	
PL	0.25; +4.4		-0.52; -9.2*	
IR	0.16; +3.2		-0.32; -6.7	
BU	0.24; +4.9		-0.01; -0.1	
EH	0.37; +6.0*		-0.23; -3.7	
BI	0.02; +0.3		-0.07; -1.4	
FA	0.67; +11.1*	-0.33; -5.4*		
		Patient-completed 4-point Likert symptom scale ranging from 0 (no problems) to 3 (severe problems)		
PH	0.42; +5.1*		-0.46; -5.6*	
PA	0.47; +7.8*		-0.25; -4.2*	
PL	0.40; +6.7*		-0.55; -9.3*	
IR	0.28; +5.7*		-0.45; -9.3*	
BU	0.34; +6.9*		-0.15; -3.0	
EH	0.35; +5.4*		-0.36; -5.5*	
BI	0.17; +3.1		-0.10; -1.9	
FA	0.32; +5.5*	-0.08; -1.4		

Continued

Table 4 Continued

Reference	LupusQoL domain	Anchor (clinical severity measure)	Findings	
			Improved	Worsening
Touma <i>et al</i> , 2011 ²⁷			ES; SRM	ES; SRM
	PH	SLEDAI-2K 30 days, where improvement was defined as reduction in SLEDAI-2K \geq 4 from the previous visit, flare as an increase in SLEDAI-2K \geq 4 from the previous visit	0.35;0.51	0.02; 0.03
	PA		0.41; 0.73	0.02; 0.04
	PL		0.16; 0.36	0.06; 0.17
	IR		0.00; 0.00	0.04; 0.14
	BU		0.28; 0.37	0.24; 0.49
	EH		0.30; 0.45	0.05;0.08
	BI		0.27; 0.39	0.04; 0.07
	FA		0.30; 0.53	0.21; 0.67

MCID 2 definition by Devilliers *et al*; MCID 3 definition by McElhone *et al*

*p<0.05.

BI, body image; BU, burden; EH, emotional health; FA, fatigue; GRC, Global Rating of Change; IR, intimate relationships; LupusQoL, Lupus Quality of Life; MCID, minimal clinically important difference; PA, pain; PH, physical health; PL, planning; SELENA, Safety of Estrogens in Lupus Erythematosus National Assessment; SLEDAI-2K, Systemic Lupus Erythematosus Disease Activity Index 2000; SRI, SLE Responder Index ; SRM, standardised response means; VAS, Visual Analogue Scale.

Assessment of responsiveness of LupusQoL in SRI-4 responders versus non-responders indicated that only Physical Health and Pain domains of LupusQoL were responsive.⁴⁷ Results from patients who completed a GRC assessment and LupusQoL indicated that all LupusQoL domain scores were significantly worse in those with deterioration versus improvement.²³ Evaluation of responsiveness of LupusQoL domains in patients with improved or worsened health status measured by a 100 mm VAS indicated that patients with improving health reported significant improvements in LupusQoL Physical Health, Pain, Emotional Health and Fatigue. Physical Health, Planning and Fatigue domain scores declined significantly in those with worsening health.²⁹

Evaluation of LupusQoL using SLEDAI-2K over 30 days as an anchor indicated that LupusQoL displayed responsiveness in some domains determined by ES and SRM estimates. There were moderate effects in Pain, Fatigue and Physical Health; and small effects in Emotional Health, Body Image, Burden to Others and in Planning among patients whose SRM improved. Among patients who worsened, a moderate-effect SRM was found in Fatigue and small effect in Burden to Others.²⁷

Thresholds of meaning of LupusQoL Scores

Two studies calculated MCIDs for LupusQoL domains (online supplementary table S10). In one anchor-based analysis, patient GRC was used as the anchor (improvement MCID (McElhone)=GRC of 2 or 3; deterioration MCID (McElhone)=GRC of -3 or -2). For deterioration, mean LupusQoL domain scores ranged from -2.4 for Body Image to -8.7 for Intimate Relationships, and for improvement from 3.5 for Body Image to 7.3 for Burden to Others. Using distribution-based approaches based on 0.5 SD, LupusQoL domain MCIDs (McElhone) ranged from 12.9 (Emotional Health) to 16.7 (Intimate

Relationships).²³ Results from a second study by Devilliers *et al* that used a patient-reported anchor-based approach (7-point Likert Scale of change in health status over the past 3 months, a 100 mm VAS assessing impact of illness, and Likert Scale from 0 (no problem) to 3 (severe problem) exploring patient-reported symptoms) indicated minimally improved domain scores ranging from 1.1 to 9.2 while minimally worsened scores ranged from -0.5 to -6.4.²⁹

The different MCIDs defined by McElhone *et al*²³ were used in a recent prospective study of 78 clinically active patients with SLE²² to compare the performance of each MCID in determining worsening and improvement measured by LupusQoL. Results indicated that the percentage of patients reporting improvements or worsening across domains varied between different MCID definitions. For most domains, percentages of patients reporting changes (improvement or worsening) were greater for MCID defined by Devillier *et al*²⁹ versus those from McElhone *et al*.²³

Discrimination of LupusQoL in RCTs and LOS

Only two RCTs used LupusQoL (online appendix table A2). In the EMBODY 1 and EMBODY 2 phase III trials in which the primary end point was not achieved (ie, no significant differences between groups in BILAG-based Combined Lupus Assessment responses at week 48), there were also no significant between-group differences with LupusQoL.⁴⁸ Results from a small trial of Acthar Gel in 10 patients with SLE indicated improvements SLEDAI-2K and LupusQoL scores over 28 days.⁴⁹

Content validity

While FACIT-F⁵⁰ was not developed in patients with SLE, the content validity of the instrument has been confirmed in this patient population.⁵¹ Three 90 min focus groups,

Table 5 FACIT-F: assessment of instrument properties in SLE

Author, year	Description of study	Reliability		Validity		
		Internal consistency	Test-retest/Retest Discriminant	Known-groups	Responsiveness and ability to detect change	MCID
Furie <i>et al</i> , 2014 ³⁵	Secondary analysis of pooled data from the BLISS Trials in SLE (n=1684). Changes in clinical, laboratory and health-related quality of life measures from baseline to week 52 were compared between SRI responders and non-responders				✓	
Strand <i>et al</i> , 2014 ⁵⁶	Secondary analysis of two phase III RCTs in SLE, including BLISS-52 (52 weeks' duration, n=865) and BLISS-76 (76 weeks' duration, n=819)		✓			
Lai <i>et al</i> , 2011 ⁵²	Longitudinal validation study in SLE (n=254). The FACIT-F and other measures were administered at four time points from baseline to week 52	✓	✓	✓	✓	✓
Goligher <i>et al</i> , 2008 ⁵⁷	Cross-sectional study in SLE (n=80). Seven fatigue instruments administered to derive the MCID of the FACIT-F. Interviews were conducted to compare fatigue levels between participants		✓			✓

✓=Instrument property assessed in study.

FACIT-F, Functional Assessment of Chronic Illness Therapy-Fatigue Scale; MCID, minimal clinically important difference; SRI, SLE Responder Index.

each including six to eight patients with SLE, were conducted to determine if FACIT-F included all aspects of fatigue relevant to these patients. Overall, the content of FACIT-F was relevant for capturing fatigue in patients with SLE and no changes to the instrument were suggested.

Reliability

Internal consistency testing of FACIT-F indicated that Cronbach's α was >0.95 ⁵² (online supplementary table S11) (table 5).

Test-retest reliability of FACIT-F has been studied in other disease states. Yellin *et al*⁵ developed and validated a measurement system for oncology patients with anemia-related concerns. The FACT-Fatigue (FACT-F), consisting of the Cancer Therapy General (FACT-G) plus 13 fatigue items and the FACT-Anaemia (FACT-An), consisting of the FACT-F plus seven non-fatigue items were found to be stable (test-retest $r=0.87$ for both) in the 50 patients studied. Chandran *et al*,⁵³ studied the reliability and validity of the FACIT-F Scale in psoriatic arthritis. The ICC for first and repeat FACIT-F Scores was 0.95 in 73 patients. The FACIT-F was also studied in patients with inflammatory bowel disease by Tinsley *et al*.⁵⁴ The ICC for first and repeat FACIT-F Scores assessed within 180 days without change in disease state was 0.81 (CD 0.78; UC 0.87).

Finally, in patients with cancer of the head and neck, Eden, *et al*⁵⁵ established the test-retest reliability and concurrent validity of FACIT-F in 65 patients. The

FACIT-F ICC was 0.866 (0.75–0.93) and internal consistency was 0.874. Nevertheless, test-retest reliability has not been studied in patients with SLE.

Construct and known-group validity

Three RCTs assessed construct validity of FACIT-F (online supplementary table S12). Secondary analysis of pooled data across BLISS-52 and BLISS-76 RCTs indicated a strong correlation ($r=0.70$) between FACIT-F and SF-36 Vitality domain.⁵⁶ Analysis of results from a longitudinal study demonstrated construct validity of FACIT-F across time with moderate-to-strong correlations between FACIT-F and SF-36 Vitality domain, PCS and MCS scores, as well pain intensity, pain interference, patient global assessment and SLAQ Scores ($r=0.52$ to 0.87).⁵² Results from a third study indicated moderate correlations between FACIT-F and both SLAQ and Patient Global Assessment scores ($r=0.49$ to 0.59)⁵⁷ (table 5).

Known-group validity for FACIT-F was assessed in a longitudinal study by calculating mean FACIT-F Scores after stratifying by BILAG Musculoskeletal and General scores at baseline and week 12 (online supplementary table S12). FACIT-F significantly differentiated groups defined using BILAG anchors at both time points with ESs ranging from 0.22 to 0.65.⁵²

Longitudinal responsiveness

Two validation studies assessed ability of FACIT-F to detect change in patients with SLE.

In one, patients were classified as SRI-4 responders or non-responders at week 52 across all treatment groups in two RCTs. FACIT-F Scores were significantly higher in SRI responders versus non-responders. Improvements in the responder group exceeded the 4-point MCID the authors defined as their meaningful threshold score.³⁵ In the second study, patients were classified as improved, worsened or unchanged using BILAG Musculoskeletal, BILAG General and Patient Global Assessment of Change Scores and those classified as improved also reported significant mean improvements in FACIT-F Scores.⁵⁰

Thresholds of meaning of FACIT-F Scores

Two studies calculated the MCID of the FACIT-F in patients with SLE (online supplementary table S13). Lai *et al*⁵² used FACIT-F Scores derived from responsiveness analyses, as well as multiple distribution-based measures and MCID from these analyses were estimated to be 3–7 points. The authors concluded the likely MCID of FACIT-F in SLE to be in the range of 3–4 points.⁵² Goligher *et al*⁵⁷ estimated MCID of FACIT-F in SLE as the mean difference between the fatigue instrument scores between patients reporting ‘a little bit more’ fatigue (referred to as Greater Fatigue) and their interview partner. MCID was calculated by estimating the mean difference between patients reporting ‘a little bit less’ fatigue (referred to as Less Fatigue) and their interview partner. Using this method, Greater Fatigue MCID was 17.5 and Less Fatigue MCID was –5.3. Regression analyses estimated MCID to be –5.9 points using the original FACIT-F scaling.⁵⁷ Methods in this analysis have the potential to include a self-reference bias and an interview order effect. Further, differences between patients used to estimate the MCID may not provide valid references to interpret differences within patients (ie, within individual change), which are more appropriately derived using longitudinal data (table 6).

Discrimination of FACIT-F in RCT and LOS

Online supplementary table A2 presents examples of seven published RCTs and one LOS in SLE that used FACIT-F. For some RCTs, significant improvements in

clinical efficacy measures across each trial corresponded with significant improvements in FACIT-F. For example, in a 52-week RCT of blisibimod versus placebo significant improvements in SELENA-SLEDAI and FACIT-F Scores were observed with a 200 mg dose. FACIT-F Scores were also improved in patients who received 100 mg of blisibimod.⁵⁸ Secondary analysis of results from the BLISS RCTs indicated that FACIT-F Scores were not significantly different across treatment groups at the week 24 prespecified secondary end point. However, FACIT-F Scores from baseline to week 52 improved significantly ($p < 0.05$) with belimumab 1 mg and 10 mg versus placebo in BLISS-52, and with 1 mg at weeks 52 and week 76 in BLISS-76. These findings corresponded with significant improvements in FACIT-F reported by SRI responders versus non-responders in a combined analysis across treatment groups of both RCTs.³⁵ In some other RCTs, significant improvements in FACIT-F Scores were not achieved, even when the primary end point was met.⁵⁹

DISCUSSION

Measurement properties of SF-36, LupusQoL and FACIT-F in patients with SLE were examined to support their use as secondary end points supporting labelling claims in RCTs evaluating the efficacy of treatments for SLE. All three instruments demonstrated strong internally consistent reliability in an SLE population (ranging from 0.72 to 0.95 across measures, domains and studies), indicating that items within each instrument measured the intended concept. In addition, both SF-36 and LupusQoL demonstrated test-retest reliability; test-retest of FACIT-F has not been assessed in patients with SLE but acceptable (>0.7) ICCs in other rheumatic diseases have been confirmed.⁵³ All measures also demonstrated convergent validity with other comparable PROs (correlations for SF-36: 0.37–0.83; LupusQoL: 0.38–0.83; FACIT-F: 0.52–0.68). In general, correlations between these PROs and measures of disease activity and damage such as SLEDAI-2K and SDI were low, as might be expected with MD-assessed outcomes confirming divergent validity. This

Table 6 FACIT-F: responsiveness and ability to detect change in SLE (within-group)

Reference	Anchor (clinical severity measure)	Findings		
		SRI responder mean	SRI non-responder mean	P value
Furie <i>et al</i> ³⁵	SRI responder	5.2	3.0	<0.001
		Improved	Unchanged	Worsened
		Mean (SD), ES, p value	Mean (SD), ES, p value	Mean (SD), ES, p value
Lai <i>et al</i> ⁵²	BILAG musculoskeletal change	7.1 (10.7), 0.66, <0.001	3.3 (11.0), 0.30, 0.004	2.5 (11.6), 0.22, 0.617
	BILAG general change	8.2 (11.9), 0.69, <0.001	4.2 (10.2), 0.41, <0.001	0.0 (9.1), 0.00, 1.00
	Patient global assessment of change	10.5 (12.9), 0.82, <0.001	3.1 (7.8), 0.40, <0.001	–3.6 (6.7), –0.53, p<0.016

BILAG, British Isles Lupus Assessment Group; ES, effect size; FACIT-F, Functional Assessment of Chronic Illness Therapy-Fatigue Scale; SRI, SLE Responder Index.

finding suggests that SF-36, LupusQoL and FACIT-F assess important underlying concepts distinct from disease activity measures.

Given the multisystemic nature of SLE, it is important to use a multidimensional approach to capture a broad array of symptoms and impacts when assessing HRQoL. Both SF-36 and LupusQoL evaluate a number of domains, including physical and mental impacts. Results from several RCTs have shown that SF-36 and FACIT-F are responsive to treatment benefit.¹ LupusQoL is disease specific and has the advantage of being more sensitive to anticipated changes in the health status of a patient with SLE. It has been included in only a few trials of patients with SLE.^{48–49} While SLE-specific concepts covered by LupusQoL have been reported to be important to patients with SLE,⁶⁰ some domains (eg, Body Image, Burden to Others) have not performed as well as those similar to SF-36 domains such as fatigue and pain. As such, both SF-36 and LupusQoL should be used together and whether LupusQoL may be more sensitive and appropriate for use in SLE subpopulations (eg, those with cutaneous manifestations)—can be studied in future trials. On the other hand, FACIT-F has been used in seven RCTs demonstrating responsiveness and providing evidence that FACIT-F is able to capture treatment-related benefits of fatigue in SLE (online supplementary table A2).

Responsiveness of instruments can be studied by assessing the correlations between changes in instruments and external anchors of change, and by the magnitude of statistics (eg, SRMs). Also, responsiveness should be interpreted very carefully in the context of a study's hypothesis since statistics have little meaning on their own. In this review, we found that SRMs of responsiveness are not met by all domains or only by small magnitudes of change in some SF-36 and LupusQoL domains. Interpreting responsiveness by just focusing on the magnitude of statistics (SRMs) is not appropriate. First, SRMs should be interpreted in the context of the expected magnitude of change in every particular study – being small, moderate or large. Sometimes a change is not expected and it is acceptable to have undetectable SRMs. If a therapeutic intervention doesn't align with a large SRM and the a priori hypothesis expected only a small SRM, a small SRM should be accepted as a valid result. In this situation, the null hypothesis is valid and it should be concluded that the instrument is responsive in the studied population despite small SRMs.

Second, the magnitudes of statistics of change (SRMs) depends also on the baseline characteristics of the studied patients – for example, if improvements in SF-36 domains are hypothesised to be large after a specific intervention but patients have reported good levels of HRQoL at baseline across a majority of the domains, it is unlikely that moderate-large SRMs will be identified. Rather than indicating that the instrument is not responsive, these results confirm the null hypothesis is not valid in this group of patients. Therefore, it is difficult to assess the responsiveness of an instrument from a single study and it is unlikely

that all domains will demonstrate changes in all studies. In conclusion, the magnitude of statistics (eg, SRMs) of responsiveness should always be interpreted in the context of the research question – what magnitude of change was hypothesised and what magnitude was identified?

The use of PROs in SLE is essential and complements the assessment and management of patients with SLE. HRQoL in SLE is measured by generic questionnaires (eg, SF-36, FACIT, EQ-5D) and SLE-specific questionnaires (eg, LupusQoL, L-QoL, SLE-QoL, LupusPRO and LIT). While in this review we focused only on three measures, SF-36, LupusQoL and FACIT-F, we will again review the psychometric properties of all other HRQoL measures in future work under the auspices of OMERACT to reconsider the domains for the core outcome set in SLE.

In conclusion, available evidence of the measurement properties of SF-36, LupusQoL and FACIT-F in patients with SLE supports the use of these instruments as secondary end points to support labelling claims in RCTs evaluating the efficacy of treatments for SLE.

Contributors The work reviewed in this manuscript was conducted under the auspices of the SLE working group of OMERACT (VS, LSS, SM and ZT), with the SLE international experts (John Esdaile (Canada), Martin Aringer (Germany), Matthias Schneider (Germany), Anca Askenaze (USA), Rosalind Ramsey-Goldman (USA), George Karpouzas (USA), Alfred Kim (USA), Julian Thumboo (Singapore), Eric Morand (Australia), David Tunnicliffe (Australia), Roger Levy (Brazil/USA), Edward Vital (UK), Ian Bruce (UK)), the Patient Research Partners (PRPs) (Kirsten Lerstrom (Denmark), Francesca Marchiori (Italy), Davide Mazzoni (Italy), Nelma Nimaout (Brazil), Eduardo Ferreira Borba Neto (Brazil), Izabel Oliveiras (Brazil), Amy Reynolds (Australia), Corry Ang (Australia), Adwoa Parker (UK and MRC), Karina Svalya (Arthritis Research Canada), Debra Hurst (Canada), Rowena Rodriguez (Canada), Kelsey Schmitt (USA), Lucretia Taper (USA), Janeen Mays (USA)) and SLE sponsor reviewers from Amgen (Brian Ortemeier and Brad Stolshek), AstraZeneca (David Ginkel and Micki Hultquist and Ogun Sasova), EMD Serono (Amy Kao and Stephen Wax), Pfizer (Connie Chen, Sam Zwillich and Noriko Likuni), Janssen (Chetan Karyekar, Mark Chevrier and Pamela Barry), Lilly (Julie Birt) and Xencor (Debra Zack). All members from the SLE working group, the PRPs and SLE sponsor reviewers were involved in all steps of this review.

Competing interests None declared.

Patient and public involvement Patients and/or the public were involved in the design, or conduct, or reporting, or dissemination plans of this research. Refer to the Methods section for further details.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement All data relevant to the study are included in the article or uploaded as supplementary information. All data reviewed as part of this systematic review has been provided within the tables and supplementary materials.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

ORCID iDs

Vibeke Strand <http://orcid.org/0000-0003-4978-4072>

Zahi Touma <http://orcid.org/0000-0001-5177-2076>

REFERENCES

- Strand V, Galateanu C, Pushparajah DS, *et al*. Limitations of current treatments for systemic lupus erythematosus: a patient and physician survey. *Lupus* 2013;22:819–26.

- 2 Strand V, Gladman D, Isenberg D, *et al.* Endpoints: consensus recommendations from OMERACT IV. outcome measures in rheumatology. *Lupus* 2000;9:322–7.
- 3 Ware JE, Kosinski M, Dewey JE. *How to score version 2 of the SF-36 health survey*. QualityMetric Incorporated: Lincoln, RI, 2000.
- 4 McElhone K, Abbott J, Shelmerdine J, *et al.* Development and validation of a disease-specific health-related quality of life measure, the LupusQoL, for adults with systemic lupus erythematosus. *Arthritis Rheum* 2007;57:972–9.
- 5 Yellen SB, Cella DF, Webster K, *et al.* Measuring fatigue and other anemia-related symptoms with the functional assessment of cancer therapy (fact) measurement system. *J Pain Symptom Manage* 1997;13:63–74.
- 6 Food and Drug Administration (FDA). Guidance for industry patient-reported outcome measures: use in medical product development to support labeling claims. *Fed Regist* 2009;74:65132–3.
- 7 Wang S-li, Wu B, Zhu L-an, *et al.* Construct and criterion validity of the Euro QoL-5D in patients with systemic lupus erythematosus. *PLoS One* 2014;9:e098883.
- 8 Leong KP, Kong KO, Thong BYH, *et al.* Development and preliminary validation of a systemic lupus erythematosus-specific quality-of-life instrument (SLEQOL). *Rheumatology* 2005;44:1267–76.
- 9 Doward LC, McKenna SP, Whalley D, *et al.* The development of the L-QoL: a quality-of-life instrument specific to systemic lupus erythematosus. *Ann Rheum Dis* 2009;68:196–200.
- 10 Jolly M, Pickard AS, Block JA, *et al.* Disease-Specific patient reported outcome tools for systemic lupus erythematosus. *Semin Arthritis Rheum* 2012;42:56–65.
- 11 Moher D, Liberati A, Tetzlaff J, *et al.* Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *J Clin Epidemiol* 2009;62:1006–12.
- 12 Terwee CB, Bot SDM, de Boer MR, *et al.* Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol* 2007;60:34–42.
- 13 Aaronson N, Alonso J, Burnam A, *et al.* Assessing health status and quality-of-life instruments: attributes and review criteria. *Qual Life Res* 2002;11:193–205.
- 14 Guyatt GH, Deyo RA, Charlson M, *et al.* Responsiveness and validity in health status measurement: a clarification. *J Clin Epidemiol* 1989;42:403–8.
- 15 Wells GA, Tugwell P, Kraag GR, *et al.* Minimum important difference between patients with rheumatoid arthritis: the patient's perspective. *J Rheumatol* 1993;20:557–60.
- 16 Strand V, Boers M, Idzerda L, *et al.* It's good to feel better but it's better to feel good and even better to feel good as soon as possible for as long as possible. response criteria and the importance of change at OMERACT 10. *J Rheumatol* 2011;38:1720–7.
- 17 Norman GR, Sloan JA, Wyrych KW. Interpretation of changes in health-related quality of life: the remarkable universality of half a standard deviation. *Med Care* 2003;41:582–92.
- 18 Engel L, Beaton DE, Touma Z. Minimal clinically important difference: a review of outcome measure score interpretation. *Rheum Dis Clin North Am* 2018;44:177–88.
- 19 Thumboo J, Fong KY, Ng TP, *et al.* Validation of the mos SF-36 for quality of life assessment of patients with systemic lupus erythematosus in Singapore. *J Rheumatol* 1999;26:97–102.
- 20 Baba S, Katsumata Y, Okamoto Y, *et al.* Reliability of the SF-36 in Japanese patients with systemic lupus erythematosus and its associations with disease activity and damage: a two-consecutive year prospective study. *Lupus* 2018;27:407–16.
- 21 Thumboo J, Feng PH, Boey ML, *et al.* Validation of the Chinese SF-36 for quality of life assessment in patients with systemic lupus erythematosus. *Lupus* 2000;9:708–12.
- 22 Nantes SG, Strand V, Su J, *et al.* Comparison of the sensitivity to change of the 36-item short form health survey and the lupus quality of life measure using various definitions of minimum clinically important differences in patients with active systemic lupus erythematosus. *Arthritis Care Res* 2018;70:125–33.
- 23 McElhone K, Abbott J, Sutton C, *et al.* Sensitivity to change and minimal important differences of the LupusQoL in patients with systemic lupus erythematosus. *Arthritis Care Res* 2016;68:1505–13.
- 24 Yilmaz-Oner S, Oner C, Dogukan FM, *et al.* Health-Related quality of life assessed by LupusQoL questionnaire and SF-36 in Turkish patients with systemic lupus erythematosus. *Clin Rheumatol* 2016;35:617–22.
- 25 Garcia-Carrasco M, Mendoza-Pinto C, Cardiel MH, *et al.* Health related quality of life in Mexican women with systemic lupus erythematosus: a descriptive study using SF-36 and LupusQoL(C). *Lupus* 2012;21:1219–24.
- 26 Wolfe F, Michaud K, Li T, *et al.* EQ-5D and SF-36 quality of life measures in systemic lupus erythematosus: comparisons with rheumatoid arthritis, noninflammatory rheumatic disorders, and fibromyalgia. *J Rheumatol* 2010;37:296–304.
- 27 Touma Z, Gladman DD, Ibañez D, *et al.* Is there an advantage over SF-36 with a quality of life measure that is specific to systemic lupus erythematosus? *J Rheumatol* 2011;38:1898–905.
- 28 Hanly JG, Urowitz MB, Jackson D, *et al.* SF-36 summary and subscale scores are reliable outcomes of neuropsychiatric events in systemic lupus erythematosus. *Ann Rheum Dis* 2011;70:961–7.
- 29 Devilliers H, Amoura Z, Besancenot J-F, *et al.* Responsiveness of the 36-item short form health survey and the lupus quality of life questionnaire in SLE. *Rheumatology* 2015;54:940–9.
- 30 Colangelo KJ, Pope JE, Peschken C. The minimally important difference for patient reported outcomes in systemic lupus erythematosus including the HAQ-DI, pain, fatigue, and SF-36. *J Rheumatol* 2009;36:2231–7.
- 31 Strand V, Crawford B. Improvement in health-related quality of life in patients with SLE following sustained reductions in anti-dsDNA antibodies. *Expert Rev Pharmacoecon Outcomes Res* 2005;5:317–26.
- 32 Cardiel MH, Tumlin JA, Furie RA, *et al.* Abetimus sodium for renal flare in systemic lupus erythematosus: results of a randomized, controlled phase III trial. *Arthritis Rheum* 2008;58:2470–80.
- 33 Strand V, Aranow C, Cardiel MH, *et al.* Improvement in health-related quality of life in systemic lupus erythematosus patients enrolled in a randomized clinical trial comparing LJP 394 treatment with placebo. *Lupus* 2003;12:677–86.
- 34 Furie R, Petri M, Zamani O, *et al.* A phase III, randomized, placebo-controlled study of belimumab, a monoclonal antibody that inhibits B lymphocyte stimulator, in patients with systemic lupus erythematosus. *Arthritis Rheum* 2011;63:3918–30.
- 35 Furie R, Petri MA, Strand V, *et al.* Clinical, laboratory and health-related quality of life correlates of systemic lupus erythematosus Responder index response: a post hoc analysis of the phase 3 belimumab trials. *Lupus Sci Med* 2014;1:e000031.
- 36 Navarra SV, Guzmán RM, Gallacher AE, *et al.* Efficacy and safety of belimumab in patients with active systemic lupus erythematosus: a randomised, placebo-controlled, phase 3 trial. *Lancet* 2011;377:721–31.
- 37 Boström C, Elfving B, Dupré B, *et al.* Effects of a one-year physical activity programme for women with systemic lupus erythematosus - a randomized controlled study. *Lupus* 2016;25:602–16.
- 38 Abrahão MI, Gomiero AB, Peccin MS, *et al.* Cardiovascular training vs. resistance training for improving quality of life and physical function in patients with systemic lupus erythematosus: a randomized controlled trial. *Scand J Rheumatol* 2016;45:197–201.
- 39 Tench CM, McCarthy J, McCurdie I, *et al.* Fatigue in systemic lupus erythematosus: a randomized controlled trial of exercise. *Rheumatology* 2003;42:1050–4.
- 40 Navarrete-Navarrete N, Peralta-Ramírez MI, Sabio JM, *et al.* Quality-Of-Life predictor factors in patients with SLE and their modification after cognitive behavioural therapy. *Lupus* 2010;19:1632–9.
- 41 Karlson EW, Liang MH, Eaton H, *et al.* A randomized clinical trial of a psychoeducational intervention to improve outcomes in systemic lupus erythematosus. *Arthritis Rheum* 2004;50:1832–41.
- 42 Dobkin PL, Da Costa D, Joseph L, *et al.* Counterbalancing patient demands with evidence: results from a pan-Canadian randomized clinical trial of brief supportive-expressive group psychotherapy for women with systemic lupus erythematosus. *Ann Behav Med* 2002;24:88–99.
- 43 Greco CM, Kao AH, Maksimowicz-McKinnon K, *et al.* Acupuncture for systemic lupus erythematosus: a pilot RCT feasibility and safety study. *Lupus* 2008;17:1108–16.
- 44 Meseguer-Henarejos A-B, Gascón-Cánovas J-J, López-Pina J-A. Components of quality of life in a sample of patients with lupus: a confirmatory factor analysis and Rasch modeling of the LupusQoL. *Clin Rheumatol* 2017;36:1789–95.
- 45 Anindito B, Hidayat R, Koesnoe S, *et al.* Validity and reliability of lupus quality of life questionnaire in patients with systemic lupus erythematosus in Indonesia. *Indones J Rheumatol* 2016;8.
- 46 Jolly M, Pickard AS, Wilke C, *et al.* Lupus-specific health outcome measure for us patients: the LupusQoL-US version. *Ann Rheum Dis* 2010;69:29–33.
- 47 Devilliers H, Bonithon-Kopp C, Jolly M. The lupus impact tracker is responsive to changes in clinical activity measured by the systemic lupus erythematosus Responder index. *Lupus* 2017;26:396–402.
- 48 Clowse MEB, Wallace DJ, Furie RA, *et al.* Efficacy and safety of Epratuzumab in moderately to severely active systemic lupus erythematosus: results from two phase III randomized, double-blind, placebo-controlled trials. *Arthritis Rheumatol* 2017;69:362–75.

- 49 Fiechtner JJ, Montroy T. Treatment of moderately to severely active systemic lupus erythematosus with adrenocorticotropic hormone: a single-site, open-label trial. *Lupus* 2014;23:905–12.
- 50 Cella D, Yount S, Sorensen M, *et al.* Validation of the functional assessment of chronic illness therapy fatigue scale relative to other instrumentation in patients with rheumatoid arthritis. *J Rheumatol* 2005;32:811–9.
- 51 Kosinski M, Gajria K, Fernandes AW, *et al.* Qualitative validation of the FACIT-fatigue scale in systemic lupus erythematosus. *Lupus* 2013;22:422–30.
- 52 Lai J-S, Beaumont JL, Ogale S, *et al.* Validation of the functional assessment of chronic illness therapy-fatigue scale in patients with moderately to severely active systemic lupus erythematosus, participating in a clinical trial. *J Rheumatol* 2011;38:672–9.
- 53 Chandran V, Bhella S, Schentag C, *et al.* Functional assessment of chronic illness therapy-fatigue scale is valid in patients with psoriatic arthritis. *Ann Rheum Dis* 2007;66:936–9.
- 54 Tinsley A, Macklin EA, Korzenik JR, *et al.* Validation of the functional assessment of chronic illness therapy-fatigue (FACIT-F) in patients with inflammatory bowel disease. *Aliment Pharmacol Ther* 2011;34:1328–36.
- 55 Eden MM, Kunkel K. Psychometric properties of the modified brief fatigue inventory and FACIT-Fatigue in individuals with cancer of the head and neck. *Rehabil Oncol* 2016;34:97–103.
- 56 Strand V, Levy RA, Cervera R, *et al.* Improvements in health-related quality of life with belimumab, a B-lymphocyte stimulator-specific inhibitor, in patients with autoantibody-positive systemic lupus erythematosus from the randomised controlled trials. *Ann Rheum Dis* 2014a;73:838–44.
- 57 Goligher EC, Pouchot J, Brant R, *et al.* Minimal clinically important difference for 7 measures of fatigue in patients with systemic lupus erythematosus. *J Rheumatol* 2008;35:635–42.
- 58 Petri MA, Martin RS, Scheinberg MA, *et al.* Assessments of fatigue and disease activity in patients with systemic lupus erythematosus enrolled in the phase 2 clinical trial with blisibimod. *Lupus* 2017;26:27–37.
- 59 Khamashta M, Merrill JT, Werth VP, *et al.* Sifalimumab, an anti-interferon- α monoclonal antibody, in moderate to severe systemic lupus erythematosus: a randomised, double-blind, placebo-controlled study. *Ann Rheum Dis* 2016;75:1909–16.
- 60 Mathias SD, Berry P, De Vries J, *et al.* Patient experience in systemic lupus erythematosus: development of novel patient-reported symptom and patient-reported impact measures. *J Patient Rep Outcomes* 2017;2:11.