Supplementary material

Development and validation of a patient reported outcome measure for systemic sclerosis: the EULAR Systemic sclerosis Impact of Disease (ScleroID) questionnaire

Mike O. Becker*[1], Rucsandra Dobrota*[1], Alexandru Garaiman[1], Rudolf Debelak[2,3] ,Kim Fligelstone[4], Ann Tyrell Kennedy[4], Annelise Roennow[4], Yannick Allanore[5], Patricia E. Carreira[6], László Czirják[7], Christopher P. Denton[8], Roger Hesselstrand[9], Gunnel Sandqvist[9], Otylia Kowal-Bielecka[10], Cosimo Bruni[11], Marco Matucci-Cerinic[11,12], Carina Mihai[1,13], Ana Maria Gheorghiu[13], Ulf Müller-Ladner[14], Joe Sexton[15], Tore K Kvien[15], Turid Heiberg≠[16], Oliver Distler≠[1]
*≠equal contributions

1. Department of Rheumatology, University Hospital Zurich, University of Zurich, Zurich, Switzerland
2. Department of Psychology, Psychological Methods, Evaluation and Statistics, University of Zurich, Zurich, Switzerland
3. Department of Psychology, Psychological Methodology, University of Leipzig, Leipzig, Germany
4. Federation of European Scleroderma Associations (FESCA)
5. Department of Rheumatology, University Paris Descartes and Cochin Hospital, Paris, France
6. Department of Rheumatology, Hospital Universitario 12 de Octubre, Madrid, Spain
7. Department of Immunology and Rheumatology, Medical School, University of Pécs, Pécs, Hungary
8. Centre for Rheumatology, University College London, Royal Free Campus, London, United Kingdom
9. Department of Rheumatology, Lund University, Lund, Sweden
10. Department of Rheumatology and Internal Medicine, Medical University of Bialystok, Bialystok, Poland
11. Department of Experimental and Clinical Medicine, Division of Rheumatology AOUC, University of Florence, Florence, Italy

Becker MO, *et al. Ann Rheum Dis* 2022; 81:507–515. doi: 10.1136/annrheumdis-2021-220702

12. Unit of Immunology, Rheumatology, Allergy and Rare diseases (UnIRAR), IRCCS San Raffaele Hospital, Milan, Italy
13. Department of Internal Medicine and Rheumatology, Cantacuzino Hospital, Carol Davila University of Medicine and Pharmacy, Bucharest, Romania
14. Department of Rheumatology and Clinical Immunology, Justus-Liebig University Giessen, Campus Kerckhoff, Bad Nauheim, Bad Nauheim, Germany
15. Division of Rheumatology and Research, Diakonhjemmet Hospital, Oslo, Norway
16. Regional Research Support, Oslo University Hospital, Oslo, Norway

Corresponding author:
Prof. Dr. med. Oliver Distler
Department of Rheumatology, University Hospital Zurich, Gloriastr. 25, 8091 Zurich
E-Mail: Oliver.Distler@usz.ch; Phone +41 44 255 29 70; Fax: +41 44 255 89 79

1. SUPPLEMENTARY METHODS

Main concept:

The ScleroID aims to specifically capture the global burden of disease of systemic sclerosis (SSc) as perceived by the patients themselves. In other words, it aims to provide an integrated and standardized overall assessment of the multiple health dimensions affected by SSc that are most important to patients. Hence, it aims to function similarly to the already successfully developed RAID and PsAID tools for rheumatoid arthritis and psoriatic arthritis, respectively [1-4].

ScleroID aims to meet an unmet need in the current assessment of the patients' disease experience in SSc. The current medical practice consists of using several existing PROM tools, which are either generic (e.g. SF-36) or somewhat adapted for SSc (e.g. SHAQ), or specifically focussing on one aspect of the disease (e.g. UCLA GIT for gastrointestinal involvement). This is in general important to detail certain aspects of the disease, but may burden the patients with lengthy and time-consuming questionnaires which however fail to capture the complexity of SSc. A specific, brief but also comprehensive questionnaire could considerably improve the inclusion of the patient perspective in clinical practice and clinical research in SSc.

We have validated the ScleroID questionnaire following the Outcome Measures in Rheumatology (OMERACT) filter, a widely acknowledged framework for development

Becker MO, *et al. Ann Rheum Dis* 2022; 81:507–515. doi: 10.1136/annrheumdis-2021-220702

of PROMs in Rheumatology [5]. By the OMERACT filter, a candidate outcome measure is evaluated according to three main pillars which are represented by *truth*, *discrimination* and *feasibility* [6]. *Truth* essentially means that the PROM measures what it is intended to, hereby including content validity, face validity and construct validity, which we investigated for ScleroID (as detailed in the main manuscript). Further, *discrimination* refers to whether the instrument can differentiate between situations of interest (either different states at one time or states at different times).[6] For this, we tested ScleroID for test/retest reliability and sensitivity to change in a clinical setting. Lastly, the feasibility of applying ScleroID in practice has been addressed in terms of translation, practicability, concision and easiness of use.

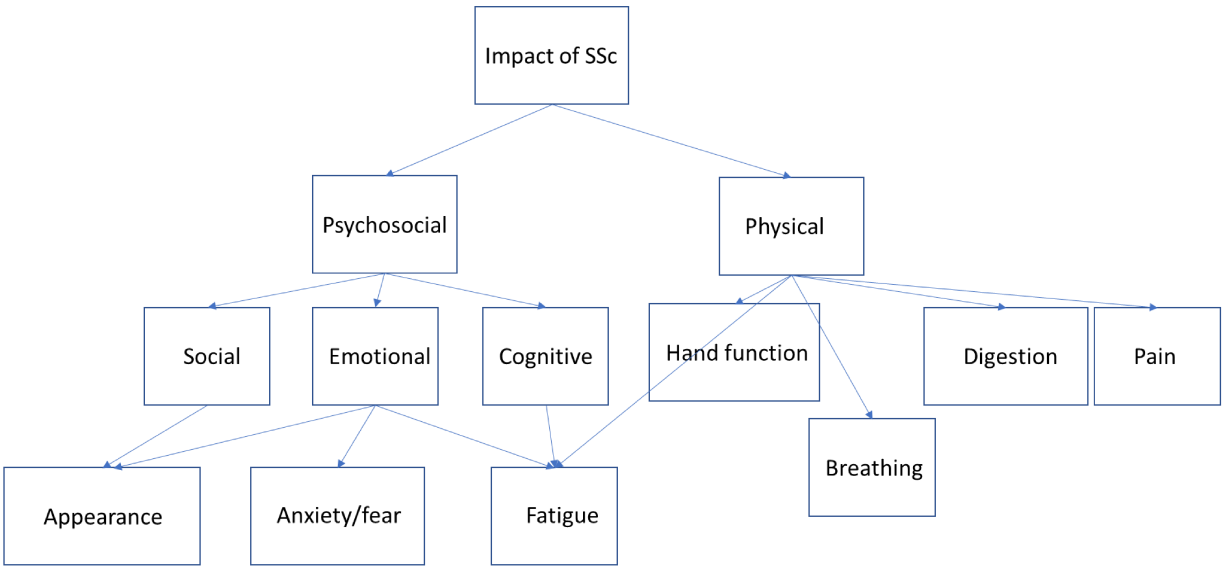
The clinical data were collected following generally accepted EUSTAR (European Scleroderma Trials and Research group) standards. Accordingly, detailed clinical, laboratory and imaging data from the patients' regular visits at the participating expert SSc centres are collected following a standardized protocol and datasheet. This includes yearly assessments with screening for organ involvement as well as potentiallly additional follow-up visits, according to the treating physician [7]. The data are systematically uploaded in a joint electronic database which undergoes periodic quality checks.

Development of the ScleroID questionnaire

Expert investigators from each centre, representing 11 European countries, invited one to three English-speaking patients, each with complementary disease features, as to cover the different aspects of the disease. Only one patient per centre was required on site, whereas the 1-2 additional patients joined via webinar/telephone conference. Given the heterogeneity of SSc, the availability of patient research partners for this first step was essential. Although there is no definitive need to calculate sample size in qualitative approaches, the principle of saturation, i.e. to reveal the full range of important perceptions, is regarded as an indicator.

The meeting took place on two days during the EULAR congress in Rome in June 2015. Eleven patients joined on site and 13 through a telephone conference. On the first day, the expert investigators (RD, MB, TH) presented a review of the literature on PROMs used in SSc to the patient research partners. The best-established and well-known questionnaires, the SHAQ and SF-36, were used as examples. The patient representatives thereafter suggested health dimensions on which the disease has an important impact, according to their personal perception, using the nominal group technique [8]. Only neutral moderating from the experts was permitted at this stage. On day one, 66 health dimensions were collected. On the second day, these were discussed, grouped and finally reduced to 17 candidate dimensions, which were approved unanimously by whole group. An example: patients initially freely reported areas in which they felt themselves affected by their SSc. The first brainstorming exercise brought up (among others) musculoskeletal aspects like “body stiffness” and “muscular weakness” as main issues for the patients. After a subsequent discussion with the group, the patients felt that these would best be captured under the term “body mobility”, understood as the general subjective perception of impaired body movement comprising both flexibility and strength aspects. A visualization of a conceptual model for a ScleroID PROM is given below.

Figure S1: A conceptual model for the ScleroID PROM



Development of study materials and translation protocol

All study materials intended for patients (prioritization sheet, ScleroID questionnaire, cohort study case report forms, CRFs) have been developed in English (RD, MB, TH, OD). For the development of the ScleroID questionnaire, the questions and NRS scales were constructed by the steering committee, including patient research partners (RD, MB, TH, OD, ATK) and agreed upon by the patient representatives who participated in the nominal group exercise in Rome (see main methods).

All study materials intended for patients (ScleroID, CRFs) were translated from English into the local language by each centre under the supervision of the local PI. The standardized translation protocol, which was recommended, required that two bilingual persons (one preferably a patient) separately translated from English into the target language, then met and reached consensus. A third person subsequently did the back translation from the local language to English. Finally, they all met to agree for a final version. The PI was advised to at least participate at this last meeting with the translation team.

The study CRFs are provided as Annex 1.

A standardized excel template for data collection was provided to the centres. All data were after completion sent to the lead centre. Where appropriate, queries were sent by the steering committee (MB, RD) to the PIs.

Selection of other PROMs as comparators for ScleroID

After literature review and discussion within the steering committee (MB, RD, OD, TH), the following questionnaires were initially selected as potential validation instruments for ScleroID and its constituting dimensions: SF-36, SSc-HAQ, EQ-5D, EUSTAR activity index, Cochin Hand Function Scale, ULCA GIT 2.0, FACIT, Raynaud's Condition Score.

Consistent with the experiences from the earlier successful EULAR projects on patient reported outcomes (RAID, PsAID) [1-3], PIs then agreed that single dimensions of the ScleroID questionnaire were not to be tested for concurrent validity. Instead, it was decided that the whole ScleroID questionnaire will be validated by comparison to other overall scores, i.e. questionnaires that evaluate the disease status of SSc patients more broadly. These were chosen to be the SF-36, the SSc-HAQ, the EQ-5D and the EUSTAR SSc activity index, based on the available data from the literature validating

Becker MO, *et al. Ann Rheum Dis* 2022; 81:507–515. doi: 10.1136/annrheumdis-2021-220702

their use in SSc. Translations for the comparator PROMs were retrieved from the literature, as available.

Table S1. Weighting exercise, as presented to patients

(extract from patient's baseline CRF, see Annex 1)

We want you to indicate how much your systemic sclerosis (scleroderma) impacts your health in the following selected health dimensions, shown below.

Please distribute 100 points between the dimensions according to their impact; the sum should be 100.

Please read all dimensions before starting to distribute your points.

You can spend your points in sets of 5. Give more points to dimensions which have important impact and less to dimensions that are not so important. You do not have to spend points in every area. You cannot spend more than 100 points.

Please take into account your whole disease history, not only how you feel today, when distributing the points.

In this table, you have to distribute your 100 points between 10 domains of health:

Domain/dimension	POINTS
Raynaud's Phenomenon	I__I__I
Hand function	I__I__I
Pain	I__I__I
Fatigue *(being tired physically, but also mental fatigue, lack of energy)*	I__I__I
Upper gastrointestinal tract symptoms *(e.g. swallowing difficulties, reflux, vomiting)*	I__I__I
Lower gastrointestinal tract symptoms *(e.g. bloating, diarrhea, constipation, anal incontinence)*	I__I__I
Limitations of life choices and activities *(e.g. social life, personal care, work)*	I__I__I
Body mobility	I__I__I
Breathlessness	I__I__I
Digital ulcers	I__I__I
TOTAL POINTS: Remember must add up to 100 points	100

2. Sample size considerations

For the initial group of patients who selected the main candidate health dimensions there was no formal sample size calculation, based upon the rationale that there is no definitive need to calculate sample size in qualitative approaches. Nonetheless, the principle of saturation, i.e. to reveal the full range of important dimensions is regarded as an indicator. A critical review from Yamazaki et al. identified a median sample size of 36 (range 9-383) in 80 qualitative studies published in the 5 most influential medical journals [9]. We also took into consideration that SSc has a wide range of clinical phenotypes, which requires diverse patient representation. As a result, the experts recruited SSc patients with a wide range of disease phenotypes and demographic characteristics, and a total of 24 took part to the nominal group exercise in Rome in 2015. For comparison, the number of participants in the initial phase of the RAID and PsAID studies for identification of candidate dimensions were 10 and 12, respectively [1, 2]. Focus groups were reported to usually contain 6 -12 participants [10]. Hence, we considered 24 SSc patients for the focus group to be sufficient.

For the prioritisation and weighting exercises, and for the validity study, formal power calculations were not performed. The literature suggested that a patient population of around 500 or more was estimated to be sufficient and we used the studies behind PsAID and RAID as models [10]. Numbers are very similar across the three studies [1-3].

Becker MO, *et al. Ann Rheum Dis* 2022; 81:507–515. doi: 10.1136/annrheumdis-2021-220702

2. SUPPLEMENTARY RESULTS

Table S2. Clinical and demographic characteristics of the patients who performed the prioritization step (N=108)

Variable	Frequency
Age (years, median (IQR))	53 (17)
Gender (n, %)	
Female	82 (76%)
Male	25 (24%)
Disease duration* (years, median (IQR))	10 (10)
Disease subset according to Le Roy (n, %)	
Limited skin involvement	53 (49.5%)
Diffuse skin involvement	54 (50.5%)
Distribution per country (alphabetically, n)	
France	9
Germany	10
Hungary	9
Italy	10
Netherlands	10
Poland	10
Romania	11
Spain	10
Sweden	7
Switzerland	12
UK	10
*time since onset of the first non-Raynaud symptom of the disease Abbreviations: IQR, interquartile range; UK, United Kingdom.	

Performance of ScleroID by the OMERACT filter – additional results

Table S3. Number and percentage of missing values for scores in the cross-sectional study.

Questionnaire	Patients with missing items, n(%)	Mean of missing items (SD)
ScleroID	10 (2.1)	3.3 (3.0)
Physician Global Assessment	23 (4.9)	1.0 (0.00)
Patient Global Assessment	3 (0.6)	1.0 (0.00)
SF-36 Physical component score	36 (7.6)	3.0 (3.7)
SF-36 Mental component score	37 (7.8)	3. 9 (3.7)
EQ-5D	8 (1.7)	1.8 (0.5)
HAQ-DI	12 (2.5)	8.0 (0.00)
SSc-HAQ	16 (3.4)	3.8 (0.8)

Abbreviations: n, number; SD, standard deviation; SF-36, 36-Item Short Form Survey; HAQ-DI, Health Assessment Questionnaire Disability Index; SSc-HAQ, Systemic Sclerosis Health Assessment Questionnaire Disability Index;

The table illustrates the number (and percentage) of patients who had at least one missing item per questionnaire and the mean number of missing items per questionnaire in those patients.

Imputation of missing ScleroID items

Two approaches to imputing a single missing component of ScleroID were investigated. The first is the approach that was used for PsAID, where the missing item of the ScleroID score is replaced by the average of the other components of the ScleroID score of the same patient ('PsAID Imputation'). The second method imputes the missing ScleroID item in one patient using the average value for this item across all patients, ('Mean Imputation'). Both methods were compared by setting one item as missing and using both methods to impute the missing item. Results were compared to the “true” ScleroID score. The difference between the imputed and true ScleroID is measured using the mean absolute error. The table below suggests that both approaches seem to work adequately, with the PsAID Imputation yielding slightly lower mean absolute errors.

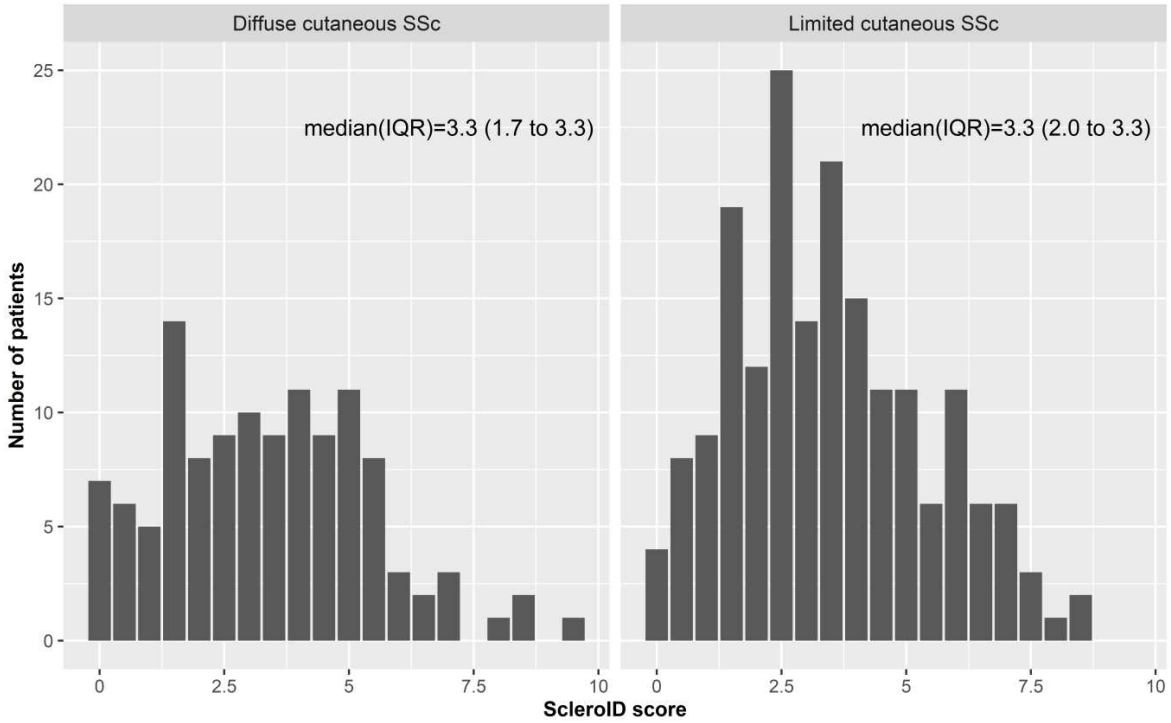
Supplementary Table S4: Imputation of single ScleroID component. Mean absolute error.

ScleroID component	Mean absolute error (PsAID Imputation)	Mean absolute error (Mean Imputation)
Raynaud`s phenomenon	0.27	0.29
Hand function	0.19	0.26
Pain	0.17	0.26
Fatigue	0.21	0.28
Upper gastrointestinal symptoms	0.17	0.21
Lower gastrointestinal symptoms	0.19	0.23
Life choices	0.16	0.24
Body mobility	0.14	0.21
Dyspnoea	0.17	0.20
Digital ulcers	0.23	0.18
Abbvreviations: PsAID, Psoriatic Arthritis Impact of Disease		

The table illustrates the mean absolute error when any ScleroID component is imputed by the PsAID or Mean Imputation method (see above). Given the ScleroID score range from 0 to 10 and the median and interquartile range (IQR) of 3.2 (1.9-4.7), the errors vary from 0.14 (4.3%) to 0.29 (9.1%).

Becker MO, *et al. Ann Rheum Dis* 2022; 81:507–515. doi: 10.1136/annrheumdis-2021-220702

Figure S2. Distribution of ScleroID scores across 472 patients at baseline.



The graphs show the distribution of final ScleroID scores amongst dcSSc (left) and lcSSc (right) patients with the respective median and IQR.

Table S5. Internal consistency of ScleroID analysed by Cronbach's alpha.

Health dimension	**Value***
Raynaud	0.87
Hand function	0.85
Pain	0.84
Fatigue	0.85
Upper GI symptoms	0.85
Lower GI symptoms	0.86
Life choices	0.84
Body mobility	0.84
Dyspnea	0.85
Digital ulcers	0.87
Cronbach's alpha	**0.87**

*Table gives Cronbach's alpha (last row) of components of ScleroID, and the value of Cronbach's alpha with individual dimension removed. For comparison, Cronbach's alpha for SSc-HAQ was 0.88, for HAQ 0.92, and for EQ5D 0.77.
Abbreviations: GI, gastrointestinal.

For Cronbach's alpha, a cut-off of 0.7-0.8 usually is regarded as satisfactory, and we interpreted values > 0.8 as strong[11, 12]. However, acceptable levels might be different and even lower depending on the actual study. Similarly, cut-off levels have been provided for correlation coefficients such as Pearson's r: "0-0.19 is regarded as very weak, 0.2-0.39 as weak, 0.40-0.59 as moderate, 0.6-0.79 as strong and 0.8-1 as very strong correlation"[13].

Further instruments to assess construct validity are methods that measure the relationship between a latent trait to be measured and the items of a questionnaire, such as principal component analysis, factor analysis or a Rasch model. We decided to implement a confirmatory factor analysis (CFA) as we a) had assumptions

concerning the possible internal structure of the questionnaire and b) thought it likely that preconditions for a Rasch model would be violated (e.g. the a priori assumption that all items measure the same latent trait and that correlations of items with the latent trait are equally distributed). With missings of no more than 3% we did a complete case analysis. The Kaiser-Maier-Olkin Measure of Sampling Adequacy was close to 1 with 0.89, Bartlett`s test suggested the variables were not completely uncorrelated ($p < 0.001$) and the determinant of the data regarded as a matrix was 0.019, all of which supported a confirmatory factor analysis. Because we hypothesised that a common latent trait might be important for all items, we tested a one factor structure as well as a bifactor/2 factors and a bifactor/3 factors structure. For comparison, structures with 2 and 3 factors were also evaluated. The model fit indices indicated slightly mixed results that in general favoured a bifactor model with either 2 or 3 factors (2 factors: hand – encompassing Raynaud`s, hand function, pain and ulcers, systemic: the remaining items; 3 factors: hand – as for the bifactor/2 factors model, GIT – lower and upper GI symptoms, life – the remaining items; see Supplementary Table S6 and Supplementary Figure S3).

Table S6. Model fit indices of the confirmatory factor analysis models.

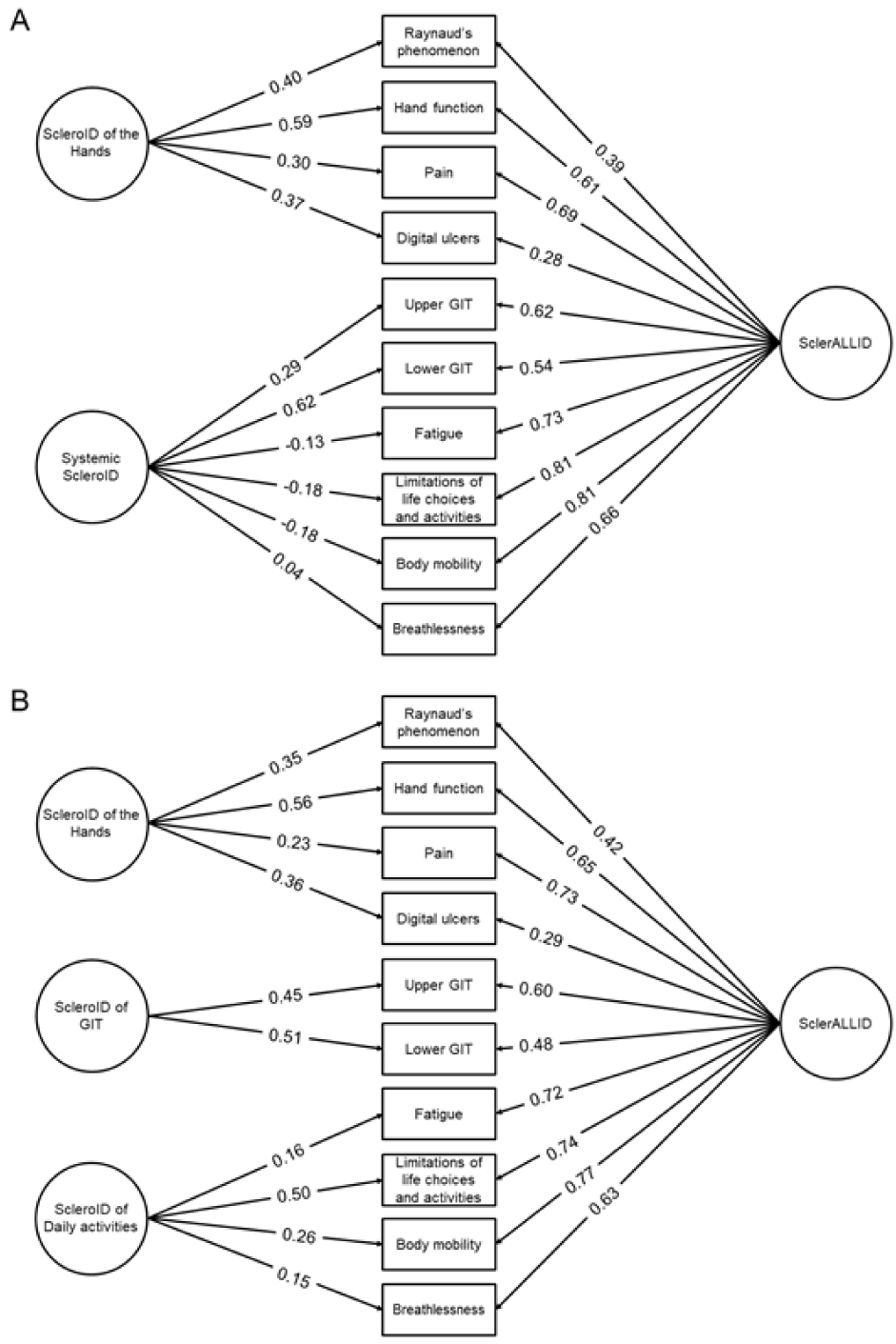
Model	Chisq	DF	Chisq/DF	CFI	RMSEA	SRMR	AIC	TLI
1 factor	236.38	35	6.8	0.89	0.11 (0.10-0.13)*	0.06	21112.05	0.86
2 factors	153.91	34	4.5	0.93	0.09 (0.07 – 0.10)*	0.05	21031.58	0.91
bifactor, 2 factors	50.78	25	2.0	0.99	0.05 (0.03 - 0.07)	0.02	20946.44	0.97
3 factors	92.74	32	2.9	0.97	0.06 (0.05 – 0.08)	0.04	20974.41	0.95
bifactor, 3 factors	62.95	25	2.5	0.98	0.06 (0.04 – 0.08)	0.03	20958.61	0.96

Chisq - chi-square statistic (all $p < 0.05$); DF – degrees of freedom; CFI - comparative fit index; RMSEA - root mean square error of approximation; SRMR - standardized root mean square residual; AIC - Akaike information criterion, TLI - Tucker Lewis. * indicates RMSEA p values < 0.05.

There are rules of thumb in the literature to assess model fit with indices: large sample sizes will almost always give significant chi-square statistics by default, therefore the ratio of the chi-square test statistic to the degrees of freedom is calculated, where a model fit is indicated by values smaller than 3 [14]. CFI should be > 0.9, better > 0.95 [15, 16]. RMSEA should be ≤ 0.6 [16], the SRMR ≤ 0.5 or at least ≤ 0.8 [16, 17]. AIC should be as low as possible with lower values indicating better fit (no absolute cut-offs). TLI should be ≥ 0.95 [16].The two bifactor models also showed the lowest local misfit in the variance-covariance matrix of standardised residuals (bifactor/2 factors: -0.474 to 0.542; bifactor/3 factors -0.474 to 0.640; compared to 1 factor: -0.652 to 1.805; 2 factors: -0.697 to 1.825; 3 factors: -0.677 to 0.800), see data in Annex 3.

Becker MO, *et al. Ann Rheum Dis* 2022; 81:507–515. doi: 10.1136/annrheumdis-2021-220702

Figure S3. Factor structure of ScleroID: A. Bifactor/2 factors model, B. Bifactor/3 factors model.



Standardized parameter estimates for the factor structure of the ScleroID. GIT: gastro-intestinal tract;

Factor loadings on the general factor for both models were meaningful for all items but digital ulcers (with loadings > 0.32 being meaningful according to Tabachnick and Fidell [18]; see Annex 4. However, as model fit measures alone are suggested to be insufficient to assess the validity of a model (see [19]) and bifactor models were suggested, we calculated additional indices, namely omegaH, omega and the reliable variance (i.e. not due to error) of the scores attributable to a general factor (i.e. possible SSc impact; calculated as omegaH divided by omega; see also [20]). Omega estimates are thought to be superior to Cronbach`s alpha, especially in the face of some multidimensionality as in bifactor models [21-25]. Although the superiority of the bifactor models speaks for (at least some) multidimensionality, we agree with Dunn et al. [26] that "an important question that the bifactor model can help the researcher to answer is: "Is this test unidimensional enough to be reported on a single scale, and relatedly, does it make sense to also report domain sub-scores?" In some respects, the bifactor model fleshes out the insight gained from the unidimensional model in cases where the researcher knows that there are likely to be dependencies between sub-groups of items within the test. Researchers in other disciplines suggest that this factor structure can, in fact, lead to greater conceptual clarity than alternative CFA model structures (e.g., Chen et al., 2012 [27]) and are particularly valuable for evaluating the plausibility of subscales (Reise et al., 2010, 2018: [28, 29]). The omega indices are given in supplemental Table S7. **With omegaH > 0.8, PUC <0.8 and ECV > 0.6, we conclude in analogy to Pretorius [21] despite some evidence of multidimensionality, there is largely reasonable evidence to claim unidimensionality and compute a single summary scale, because the large majority of variance in scores can be attributed to a general factor and 89% (bifactor/3 factors model) or 93% (bifactor/2 factors model) of the reliable variance can be accounted for by this general factor** (see also [20, 30]).

Table S7: Omega estimates of the explained variance from the two bifactor models.

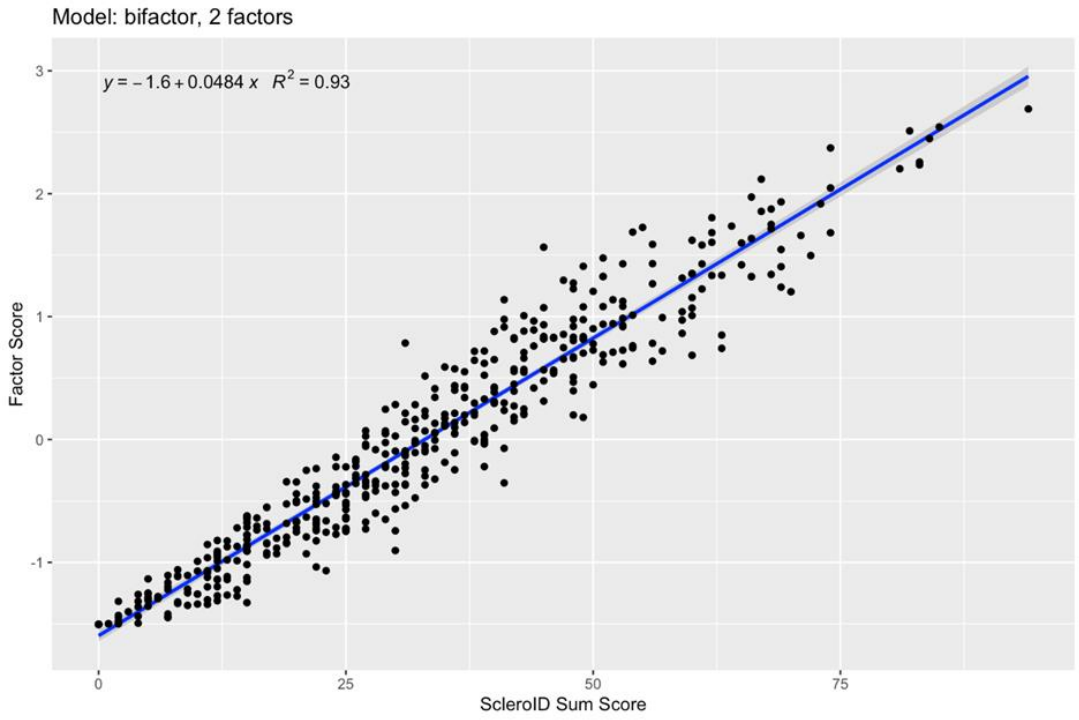
Model	Omega	OmegaH	OmegaH/Omega	ECV	PUC
Bifactor/2 factors	0.896	0.830	0.927	0.758	0.533
Bifactor/3 factors	0.895	0.800	0.894	0.727	0.711

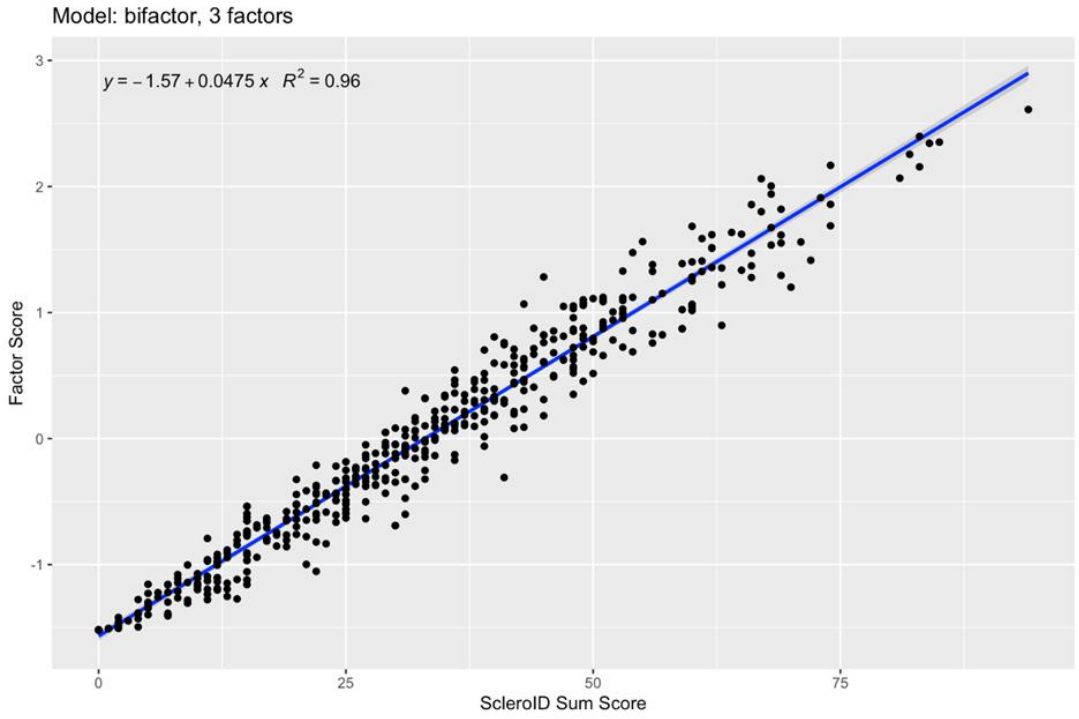
Omega - McDonald's omega: a model-based estimate of reliability; OmegaH – omega hierarchical; ECV - explained common variance; PUC - percentage of uncontaminated correlations.

If we assume that a summary score is justified, it remains to be clarified how to calculate the summary score that ideally represents the SSc impact on the life of patients as the latent trait measured by the questionnaire. Several methods exist to determine weights from a factor analysis and even using "unweighted" items (or unit-based weighting) for a sum score would have to be justified by the model [31]. One model-driven approach is for example, to use factor scores of the factor analysis model as weights for a sum score [32].

Our chosen patient-centred approach calculated weights by assigning item importance as reported by the patients and calculated a summary score. When we correlated the ScleroID sum scores with the calculated factor scores of the bifactor/2 factors model and the bifactor/3 factors model, the correlation was very high ($R^2 = 0.93$ and $R^2 = 0.96$, respectively; see Supplemental Figure S4) indicating only small differences between our weighted ScleroID sum scores and weights based on factor scores.

Figure S4: Correlation of Factor Scores with ScleroID sum scores for the two bifactor models (bifactor/2 factors above, bifactor/3 factors below).





Reliability – additional results

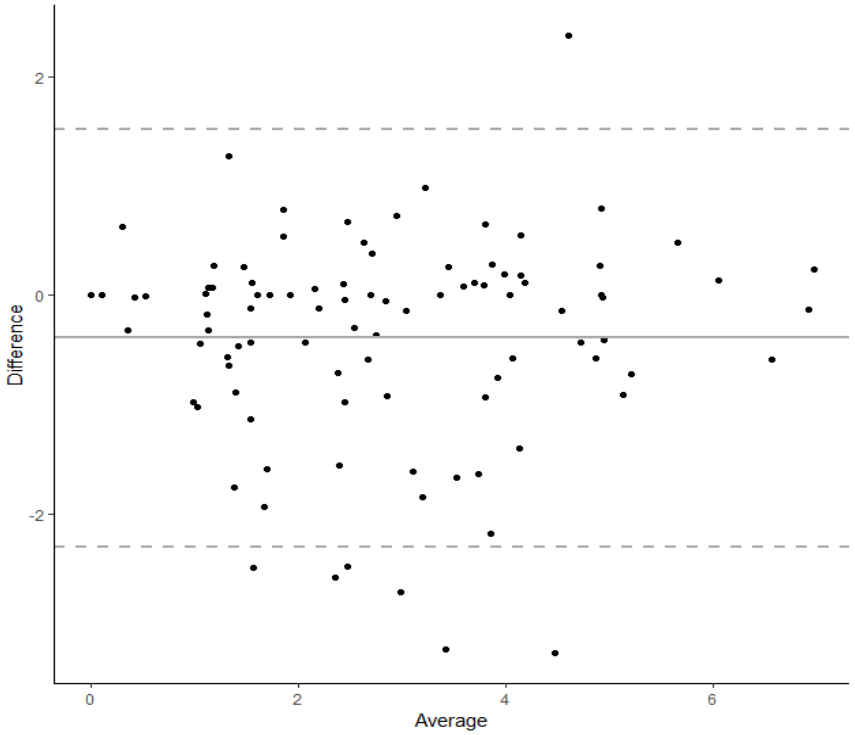
Patients' distribution per centre was: France (none), Italy (n=10), Hungary (n=20), Poland (n=3), Romania (n=10), Spain (none), Sweden (n=16), Switzerland (n=42), United Kingdom (n=8). All patients reporting a stable disease status were analysed (Table S8).

Table S8. Test-retest reliability of ScleroID compared to other PROM

Variable	**Intra-class correlation** [no. of valid cases]	**95% Confidence interval**
ScleroID	0.84 [98]	(0.77,0.89)
Raynaud	0.78 [100]	(0.68,0.84)
Hand function	0.79 [100]	(0.70,0.85)
Pain	0.67 [100]	(0.55,0.77)
Fatigue	0.66 [100]	(0.53,0.76)
Upper GI symptoms	0.67 [100]	(0.55,0.77)
Lower GI symptoms	0.61 [100]	(0.47,0.72)
Life Choices	0.72 [99]	(0.61,0.81)
Body Mobility	0.67 [101]	(0.54,0.76)
Dyspnoea	0.63 [100]	(0.50,0.74)
Digital ulcers	0.65 [101]	(0.52,0.75)
Patient's Global Assessment	0.78 [101]	(0.69,0.85)
SF-36 Physical component score	0.76 [100]	(0.66,0.83)
SF-36 Mental component score	0.69 [100]	(0.57,0.78)
HAQ-DI	0.72 [95]	(0.61,0.8)
SSc HAQ	0.72 [93]	(0.60,0.8)
EQ-5D	0.43 [97]	(0.25,0.58)

Abbreviations: SF-36: the short form (36) health survey; HAQ-DI: health assessment questionnaire disability index; SSc HAQ: systemic sclerosis health assessment questionnaire; EQ-5D: EuroQol five dimensional questionnaire. UK: United Kingdom; VAS: visual analogue scale.

Figure S5. Bland-Altman plot for agreement regarding test-retest reliability of ScleroID.



The Bland-Altman plot shows on the y-axis the mean difference between every pair of two tests (test and re-test, solid line) and the upper and lower levels of agreement (+/- 1.96 standard deviation of the difference). The x-axis depicts the average ScleroID score of the two tests (test and re-test).

Sensitivity to change – responsiveness statistics

The formula for SRM includes in the nominator the difference of the mean score at the follow up and mean scores at the baseline (so the change mean), while the denominator is represented by the standard deviation of this difference between follow up scores and baseline scores.[33] It can also be defined as a function of the paired t-test (or vice versa). Since there is no standard error of the mean in the denomination, the SRM remove the dependence on the sample size, which represents a big asset.[34] Moreover, the denominator is represented by the standard deviation of this difference and, and therefore it reflects the standard deviation of the change which

makes SRM to be more attractive than other effect size measures which are capable to reflect only the standard deviation of the baseline scores only and not the variability of the change scores. [35] Often, cut off values of 0.2, 0.5, 0.8 or greater have been proposed to distinguish small, moderate and large responsiveness, respectively.

REFERENCES:

1. Gossec L, de Wit M, Kiltz U, Braun J, Kalyoncu U, Scrivo R, et al. A patient-derived and patient-reported outcome measure for assessing psoriatic arthritis: elaboration and preliminary validation of the Psoriatic Arthritis Impact of Disease (PsAID) questionnaire, a 13-country EULAR initiative. Ann Rheum Dis. 2014 Jun; 73(6):1012-1019.
2. Gossec L, Dougados M, Rincheval N, Balanescu A, Boumpas DT, Canadelo S, et al. Elaboration of the preliminary Rheumatoid Arthritis Impact of Disease (RAID) score: a EULAR initiative. Ann Rheum Dis. 2009 Nov; 68(11):1680-1685.
3. Gossec L, Paternotte S, Aanerud GJ, Balanescu A, Boumpas DT, Carmona L, et al. Finalisation and validation of the rheumatoid arthritis impact of disease score, a patient-derived composite measure of impact of rheumatoid arthritis: a EULAR initiative. Ann Rheum Dis. 2011 Jun; 70(6):935-942.
4. Dougados M, Brault Y, Logeart I, van der Heijde D, Gossec L, Kvien T. Defining cut-off values for disease activity states and improvement scores for patient-reported outcomes: the example of the Rheumatoid Arthritis Impact of Disease (RAID). Arthritis Res Ther. 2012 May 30; 14(3):R129.
5. Tugwell P, Boers M, D'Agostino MA, Beaton D, Boonen A, Bingham CO, 3rd, et al. Updating the OMERACT filter: implications of filter 2.0 to select outcome instruments through assessment of "truth": content, face, and construct validity. J Rheumatol. 2014 May; 41(5):1000-1004.
6. Beaton DE, Maxwell LJ, Shea BJ, Wells GA, Boers M, Grosskleg S, et al. Instrument Selection Using the OMERACT Filter 2.1: The OMERACT Methodology. J Rheumatol. 2019 Aug; 46(8):1028-1035.
7. Walker UA, Tyndall A, Czirjak L, Denton C, Farge-Bancel D, Kowal-Bielecka O, et al. Clinical risk assessment of organ manifestations in systemic sclerosis: a report from the EULAR Scleroderma Trials And Research group database. Ann Rheum Dis. 2007 Jun; 66(6):754-763.
8. McMillan SS, King M, Tully MP. How to use the nominal group and Delphi techniques. Int J Clin Pharm. 2016 Jun; 38(3):655-662.
9. Yamazaki H, Slingsby BT, Takahashi M, Hayashi Y, Sugimori H, Nakayama T. Characteristics of qualitative studies in influential journals of general medicine: a critical review. Biosci Trends. 2009 Dec; 3(6):202-209.
10. Frost MH, Reeve BB, Liepa AM, Stauffer JW, Hays RD, Mayo FDAP-ROCMG. What is sufficient evidence for the reliability and validity of patient-reported outcome measures? Value Health. 2007 Nov-Dec; 10 Suppl 2:S94-S105.
11. Bland JM, Altman DG. Cronbach's alpha. BMJ. 1997 Feb 22; 314(7080):572.
12. Taber KS. The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. Research in Science Education. 2017; 48(6):1273-1296.
13. Swinscow T, Campbell, MJ. Statistics at Square One: BMJ Publishing Group, 1997.
14. Iacobucci D. Structural equations modeling: Fit Indices, sample size, and advanced topics. J Consum Psychol. 2010 Jan; 20(1):90-98.
15. Bentler PM, Yuan KH. Structural Equation Modeling with Small Samples: Test Statistics. Multivariate Behav Res. 1999 Apr 1; 34(2):181-197.

16. Hu LT, Bentler PM. Cutoff Criteria for Fit Indexes in Covariance Structure Analysis: Conventional Criteria Versus New Alternatives. Struct Equ Modeling. 1999; 6(1):1-55.
17. Byrne BM. Structural Equation Modeling With AMOS, EQS, and LISREL: Comparative Approaches to Testing for the Factorial Validity of a Measuring Instrument. International Journal of Testing. 2001 2001/03/01; 1(1):55-86.
18. Tabachnick BGF, L.S. Using Multivariate Statistics. 5th ed. ed. New York: Allyn and Bacon, 2007.
19. Morgan GB, Hodge KJ, Wells KE, Watkins MW. Are Fit Indices Biased in Favor of Bi-Factor Models in Cognitive Ability Research?: A Comparison of Fit in Correlated Factors, Higher-Order, and Bi-Factor Models via Monte Carlo Simulations. Journal of Intelligence. 2015; 3(1).
20. Neff KD, Whittaker TA, Karl A. Examining the Factor Structure of the Self-Compassion Scale in Four Distinct Populations: Is the Use of a Total Scale Score Justified? J Pers Assess. 2017 Nov-Dec; 99(6):596-607.
21. Pretorius TB. Over reliance on model fit indices in confirmatory factor analyses may lead to incorrect inferences about bifactor models: A cautionary note. 2021; 2021.
22. Zinbarg RE, Yovel I, Revelle W, McDonald RP. Estimating Generalizability to a Latent Variable Common to All of a Scale's Indicators: A Comparison of Estimators for ωh. Applied Psychological Measurement. 2006; 30(2):121-144.
23. Trizano-Hermosilla I, Galvez-Nieto JL, Alvarado JM, Saiz JL, Salvo-Garrido S. Reliability Estimation in Multidimensional Scales: Comparing the Bias of Six Estimators in Measures With a Bifactor Structure. Front Psychol. 2021; 12:508287.
24. Sijtsma K. On the Use, the Misuse, and the Very Limited Usefulness of Cronbach's Alpha. Psychometrika. 2009 Mar; 74(1):107-120.
25. Dunn TJ, Baguley T, Brunsden V. From alpha to omega: a practical solution to the pervasive problem of internal consistency estimation. Br J Psychol. 2014 Aug; 105(3):399-412.
26. Dunn KJ, McCray G. The Place of the Bifactor Model in Confirmatory Factor Analysis Investigations Into Construct Dimensionality in Language Testing. Front Psychol. 2020; 11:1357.
27. Chen FF, Hayes A, Carver CS, Laurenceau JP, Zhang Z. Modeling general and specific variance in multifaceted constructs: a comparison of the bifactor model to other approaches. J Pers. 2012 Feb; 80(1):219-251.
28. Reise SP, Bonifay, W., Haviland, M.G. Bifactor Modelling and the Evaluation of Scale Scores. In: Irwing P, Booth, T., Hughes, D.J., ed. *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development*: Wiley; 2018.
29. Reise SP, Moore TM, Haviland MG. Bifactor models and rotations: exploring the extent to which multidimensional data yield univocal scale scores. J Pers Assess. 2010 Nov; 92(6):544-559.
30. Reise SP, Bonifay WE, Haviland MG. Scoring and modeling psychological measures in the presence of multidimensionality. J Pers Assess. 2013; 95(2):129-140.
31. McNeish D, Wolf MG. Thinking twice about sum scores. Behav Res Methods. 2020 Dec; 52(6):2287-2305.
32. Grice J. Computing and evaluating factor scores. Psychological methods. 2002 01/01; 6:430-450.

33. Middel B, van Sonderen E. Statistical significant change versus relevant or important change in (quasi) experimental design: some conceptual and methodological problems in estimating magnitude of intervention-related change in health services research. Int J Integr Care. 2002; 2:e15.
34. Beaton DE, Hogg-Johnson S, Bombardier C. Evaluating changes in health status: reliability and responsiveness of five generic health status measures in workers with musculoskeletal disorders. J Clin Epidemiol. 1997 Jan; 50(1):79-93.
35. Husted JA, Cook RJ, Farewell VT, Gladman DD. Methods for assessing responsiveness: a critical review and recommendations. J Clin Epidemiol. 2000 May; 53(5):459-468.

ANNEX 1 (see .pdf file):

1) **Baseline patient CRF**
2) **Baseline physician CRF**
3) **Reliability patient CRF**
4) **Reliability physician CRF**
5) **Sensitivity to change patient CRF**
6) **Sensitivity to change physician CRF**

Annex 2: Item mapping of the health dimensions selected for ScleroID

Initial candidate health dimensions as freely reported by the patient research partners in the first step of the nominal group exercise:

1. Digestion - bloating
2. Oesophagus – difficulty swallowing and pain
3. Limitation of hand function – pain, loss of mobility, shortened fingers
4. Disability - change of face, hands and all physical aspects
5. Quality of life
6. Social and governmental support
7. Cold and aching fingers - due to Raynaud
8. Breathlessness
9. Fear of losing my job
10. Fatigue - Shortness of breath
11. Depression
12. Body stiffness
13. Hand limitation*
14. Fatigue
15. Anxiety
16. Fatigue
17. Fear of uncertainty
18. Hand disability
19. Vomiting
20. Cold fingers with loss of sensibility – due to Raynaud
21. Muscular weakness
22. Pain due to calcinosis
23. Painful feet – due to loss of tissue in the soles
24. Anal incontinence
25. Digestion problems - acidity, constipation
26. Hand function
27. Appearance
28. Exhaustion
29. Focusing attention
30. Managing changing symptoms
31. Uncertainty
32. Shortness of breath
33. Need to explain to others
34. Appearance - hands, face
35. Limitations of choice in everyday life
36. Anxiety (uncertainty)
37. Digestion problems – reflux, vomiting, anal incontinence, incl. social aspects
38. Fatigue – exhaustion after small efforts
39. Dryness of eyes and mouth
40. Forgetfulness
41. Cold and stiff fingers
42. Loss of time – due to the disease
43. Appearance
44. Limitation of hand and feet function due to ulcers
45. Digestion – reflux, cough
46. Loss of hand mobility and strength
47. Loss of weight
48. Eating problems – because of small mouth
49. Suffocation (shortness of breath), cough
50. Pain in bowels and anal incontinence
51. Frequent infections
52. Frequent infections
53. Fatigue, lack of energy - work impairment
54. Constipation
55. Coughing constantly
56. Short breath
57. Burden of taking medicines – esp. attention to risk of infection as a side effect
58. Oesophageal (GI) reflux
59. Painful and cold hands –due to Raynaud
60. Fear – of transplant rejection
61. Fear – of comorbidity e.g. cancer
62. Breathlessness – due to heart problems
63. Limitations of everyday life - due to reduced body mobility, incontinence
64. GI difficulty - with reflux, swallowing and digestion (as a whole)
65. Painful digital ulcers and calcinosis
66. Fatigue – due to musculoskeletal pain

Becker MO, *et al. Ann Rheum Dis* 2022; 81:507–515. doi: 10.1136/annrheumdis-2021-220702

Exercise to group the initial health dimensions according to their common concept:

GI:
1. Digestion - bloating
2. Esophagus – difficulty swallowing and pain
3. Vomiting
4. Anal incontinence
5. Digestion problems - acidity, constipation
6. Digestion problems – reflux, vomiting, anal incontinence, incl. social aspects
7. Digestion – reflux, cough
8. Loss of weight
9. Eating problems – because of small mouth
10. Pain in bowels and anal incontinence
11. Constipation
12. Esophageal (GI) reflux
13. GI difficulty - with reflux, swallowing and digestion (as a whole)

Hands and feet function:
14. Limitation of hand function – pain, loss of mobility, shortened fingers
15. Hand limitation
16. Hand disability
17. Hand function
18. Limitation of hand and feet function due to ulcers
19. Loss of hand mobility and strength

Mixed:
20. Disability - change of face, hands and all physical aspects
21. Quality of life
22. Fatigue - Shortness of breath
23. Limitations of choice in everyday life

Social:
24. Social and governmental support
25. Fear of losing my job
26. Appearance
27. Need to explain to others
28. Appearance - hands, face
29. Loss of time – due to the disease
30. Appearance
31. Limitations of everyday life - due to reduced body mobility, incontinence

Peripheral vascular:
32. Cold and aching fingers - due to Raynaud
33. Cold fingers with loss of sensibility – due to Raynaud
34. Cold and stiff fingers
35. Painful and cold hands –due to Raynaud

Breathlessness:
36. Breathlessness
37. Shortness of breath
38. Suffocation (shortness of breath), cough
39. Short breath
40. Breathlessness – due to heart problems

Fatigue:
41. Fatigue
42. Fatigue
43. Exhaustion
44. Fatigue – exhaustion after small efforts
45. Fatigue, lack of energy - work impairment
46. Fatigue – due to musculoskeletal pain

Mental:
47. Depression
48. Anxiety
49. Fear of uncertainty
50. Focusing attention
51. Managing changing symptoms
52. Uncertainty
53. Anxiety (uncertainty)
54. Forgetfulness
55. Fear – of transplant rejection
56. Fear – of comorbidity e.g. cancer

Musculoskeletal:
57. Body stiffness
58. Muscular weakness

Pain:
59. Pain due to calcinosis
60. Painful feet – due to loss of tissue in the soles
61. Painful digital ulcers and calcinosis

62. Dryness of eyes and mouth
63. Coughing constantly

Side effects of therapy:
64. Frequent infections
65. Frequent infections
66. Burden of taking medicines – esp. attention to risk of infection as a side effect

Selected 17 candidate health dimensions for the following prioritisation exercise:

1. Upper gastrointestinal tract symptoms (e.g. swallowing difficulties, reflux, vomiting)
2. Lower gastrointestinal tract symptoms (e.g. bloating, diarrhea, constipation, anal incontinence)
3. Pain
4. Raynaud
5. Hand function
6. Body mobility
7. Ulcers
8. Calcinosis
9. Appearance
10. Limitations of life choices and activities (e.g. social life, personal care, work)
11. Breathlessness
12. Cough
13. Fatigue
14. Depression
15. Anxiety (unpredictable course of disease, or infection as a side effect of therapy)
16. Concentration ability
17. Dryness (eyes, mouth, skin)

Final top 10 health dimensions to be included in ScleroID as a result of the prioritisation exercise:

1. Raynaud
2. Hand function
3. Upper GI symptoms
4. Pain
5. Fatigue
6. Lower GI symptoms
7. Limitation of life choices and activities

Becker MO, *et al. Ann Rheum Dis* 2022; 81:507–515. doi: 10.1136/annrheumdis-2021-220702

8. Body mobility
9. Breathlessness
10. Digital ulcers

Annex 3 - Local misfit diagnostics with the variance-covariance matrix of standardised residuals

```
$`1 factor`
##          qraynd qhandf qpain  qulcrs qfatig qlifec qbodym qdyspn qlowrg
qupprg
## qraynaud  0.000
## qhandf    1.345  0.000
## qpain     0.708  0.716  0.000
## qulcers   0.321  1.296  0.489  0.000
## qfatigue  0.134 -0.327  0.029 -0.690  0.000
## qlifec   -0.444 -0.360 -0.508 -0.215  0.262  0.000
## qbodym   -0.536 -0.004 -0.173 -0.104  0.071  0.489  0.000
## qdyspnea -0.652 -0.641 -0.279 -0.488  0.302  0.278 -0.005  0.000
## qlowergi -0.203 -0.423  0.586 -0.472 -0.362 -0.447 -0.430  0.622  0.000
## quppergi -0.104 -0.145 -0.120 -0.029 -0.278 -0.025 -0.268  0.433  1.805
0.000
## $`2 factors`
##          qraynd qhandf qpain  qulcrs qfatig qlifec qbodym qdyspn qlowrg
qupprg
## qraynaud  0.000
## qhandf    0.462  0.000
## qpain    -0.059 -0.237  0.000
## qulcers  -0.216  0.591 -0.136  0.000
## qfatigue  0.210 -0.074  0.531 -0.697  0.000
## qlifec   -0.429 -0.179 -0.058 -0.272  0.063  0.000
## qbodym   -0.479  0.224  0.307 -0.125 -0.040  0.251  0.000
## qdyspnea -0.635 -0.500  0.059 -0.525  0.164  0.034 -0.170  0.000
## qlowergi -0.118 -0.205  0.970 -0.451 -0.354 -0.513 -0.442  0.580  0.000
## quppergi -0.034  0.065  0.279 -0.025 -0.316 -0.153 -0.331  0.346  1.825
0.000
## $`bifactor, 2 factors`
##          qraynd qhandf qpain  qulcrs qfatig qlifec qbodym qdyspn qlowrg
qupprg
## qraynaud  0.000
## qhandf    0.021  0.000
## qpain     0.266 -0.087  0.000
## qulcers  -0.462  0.096  0.080  0.000
```

```
## qfatigue  0.542  0.097  0.305 -0.327  0.000
## qlifec   -0.066 -0.003 -0.327  0.136 -0.041  0.000
## qbodym   -0.156  0.374  0.049  0.242 -0.155  0.065  0.000
## qdyspnea -0.393 -0.408 -0.180 -0.242  0.237  0.110 -0.110  0.000
## qlowergi -0.146 -0.474  0.388 -0.376 -0.039 -0.025  0.026 -0.005  0.000
## quppergi  0.122  0.046 -0.060  0.190 -0.118  0.117 -0.082  0.125 -0.016
0.000
## $`3 factors`
##          qraynd qhandf qpain  qulcrs qfatig qlifec qbodym qdyspn qlowrg
qupprg
## qraynaud  0.000
## qhandf    0.461  0.000
## qpain    -0.059 -0.235  0.000
## qulcers  -0.219  0.587 -0.140  0.000
## qfatigue  0.241 -0.028  0.580 -0.677  0.000
## qlifec   -0.416 -0.159 -0.035 -0.266 -0.032  0.000
## qbodym   -0.465  0.246  0.329 -0.117 -0.123  0.120  0.000
## qdyspnea -0.578 -0.417  0.146 -0.485  0.178  0.019 -0.181  0.000
## qlowergi -0.201 -0.322  0.849 -0.514 -0.126 -0.279 -0.229  0.800  0.000
## quppergi -0.169 -0.126  0.082 -0.126 -0.108  0.054 -0.141  0.555  0.000
0.000
$`bifactor, 3 factors`
##          qraynd qhandf qpain  qulcrs qfatig qlifec qbodym qdyspn qlowrg
qupprg
## qraynaud  0.000
## qhandf    0.030  0.000
## qpain     0.260 -0.089  0.000
## qulcers  -0.429  0.068  0.131  0.000
## qfatigue  0.365 -0.071  0.118 -0.403  0.000
## qlifec   -0.043  0.148 -0.150  0.199 -0.022  0.000
## qbodym   -0.232  0.364  0.041  0.234  0.001  0.010  0.000
## qdyspnea -0.474 -0.446 -0.221 -0.261  0.204  0.001 -0.090  0.000
## qlowergi -0.078 -0.295  0.597 -0.300 -0.330 -0.237 -0.313  0.640  0.000
## quppergi  0.042  0.006 -0.102  0.170 -0.235  0.221 -0.130  0.457  0.000
0.000
```

Annex 4 – Item loadings for the bifactor models

```
bifactor/2 factors
Latent Variables:
##                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   hand =~
##     qraynaud         1.142    0.190    6.006    0.000    1.142    0.399
##     qhandf           1.669    0.214    7.791    0.000    1.669    0.592
##     qpain            0.858    0.171    5.011    0.000    0.858    0.298
##     qulcers          1.003    0.173    5.807    0.000    1.003    0.367
##   systemic =~
##     qfatigue        -0.389    0.176   -2.215    0.027   -0.389   -0.134
##     qlifec          -0.517    0.239   -2.165    0.030   -0.517   -0.177
##     qbodym          -0.489    0.184   -2.661    0.008   -0.489   -0.183
##     qdyspnea         0.118    0.226    0.523    0.601    0.118    0.044
##     qlowergi         1.829    0.473    3.867    0.000    1.829    0.620
##     quppergi         0.770    0.324    2.377    0.017    0.770    0.286
##   all =~
##     qraynaud         1.107    0.138    8.028    0.000    1.107    0.386
##     qhandf           1.730    0.122   14.160    0.000    1.730    0.613
##     qpain            1.982    0.115   17.233    0.000    1.982    0.687
##     qulcers          0.753    0.150    5.009    0.000    0.753    0.275
##     qfatigue         2.108    0.110   19.078    0.000    2.108    0.727
##     qlifec           2.369    0.100   23.616    0.000    2.369    0.813
##     qbodym           2.158    0.108   19.899    0.000    2.158    0.809
##     qdyspnea         1.757    0.117   15.059    0.000    1.757    0.656
##     qlowergi         1.588    0.210    7.572    0.000    1.588    0.538
##     quppergi         1.673    0.135   12.412    0.000    1.673    0.621


Bifactor/3 factors
Latent Variables:
##                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
##   hand =~
##     qraynaud         1.013    0.236    4.302    0.000    1.013    0.354
##     qhandf           1.589    0.249    6.392    0.000    1.589    0.563
```

```
##     qpain            0.650    0.241    2.699    0.007    0.650    0.225
##     qulcers          0.974    0.203    4.788    0.000    0.974    0.356
##   life =~
##     qfatigue         0.461    0.467    0.988    0.323    0.461    0.159
##     qlifec           1.466    0.543    2.701    0.007    1.466    0.503
##     qbodym           0.685    0.417    1.642    0.101    0.685    0.257
##     qdyspnea         0.395    0.369    1.071    0.284    0.395    0.148
##   git =~
##     qlowergi         1.488    0.107   13.929    0.000    1.488    0.505
##     quppergi         1.207    0.180    6.710    0.000    1.207    0.448
##   all =~
##     qraynaud         1.204    0.153    7.867    0.000    1.204    0.420
##     qhandf           1.829    0.128   14.310    0.000    1.829    0.649
##     qpain            2.094    0.141   14.889    0.000    2.094    0.726
##     qulcers          0.798    0.158    5.047    0.000    0.798    0.292
##     qfatigue         2.085    0.157   13.311    0.000    2.085    0.719
##     qlifec           2.158    0.140   15.438    0.000    2.158    0.741
##     qbodym           2.046    0.136   15.068    0.000    2.046    0.767
##     qdyspnea         1.683    0.157   10.707    0.000    1.683    0.628
##     qlowergi         1.403    0.170    8.240    0.000    1.403    0.476
##     quppergi         1.604    0.131   12.222    0.000    1.604    0.595
```

Annex 5 : Model CRF for the collection of EUSTAR clinical data (see pdf)