Responsiveness of quality-of-life measures in patients with peripheral vascular malformations: the OVAMA project*

M.M. Lokhorst [ID],[1] S.E.R. Horbach [ID],[1] M. Waner,[2] T.M. O,[2] C.J.M. van der Vleuten,[3] L.B. Mokkink,[4] C.M.A.M. van der Horst[1] and P.I. Spuls[5]

[1]Department of Plastic, Reconstructive and Hand Surgery, Amsterdam University Medical Centers, University of Amsterdam, Amsterdam, the Netherlands

[2]Department of Otolaryngology, Vascular Birthmark Institute of New York, Lenox Hill and Manhattan Eye, Ear, and Throat Hospitals, New York, NY, U.S.A.

[3]Department of Dermatology, Radboud University Medical Center, Nijmegen, the Netherlands

[4]Department of Epidemiology and Biostatistics and Amsterdam Public Health Research Institute, Amsterdam University Medical Centers, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands

[5]Department of Dermatology, Amsterdam University Medical Centers, University of Amsterdam, Amsterdam Public Health, Infection and Immunity, Amsterdam, the Netherlands

Correspondence
Max M. Lokhorst.
E-mail: m.m.lokhorst@amc.uva.nl

Summary

Background The OVAMA (Outcome Measures for Vascular Malformations) project determined quality of life (QoL) as a core outcome domain for patients with vascular malformations. In order to measure how current therapeutic strategies alter QoL in these patients, a patient-reported outcome measurement (PROM) responsive to changes in QoL is required.

Objectives To assess the responsiveness of two widely used generic QoL PROMs, the Medical Outcomes Study Short Form 36 (SF-36) and Skindex-29, in adult patients with vascular malformations.

Methods In an international multicentre prospective study, treated and untreated patients completed the SF-36 and Skindex-29 at baseline and after a follow-up period of 6–8 weeks. Global rating of change (GRC) scales assessing various QoL-related outcome domains were additionally completed. Per subscale, responsiveness was assessed using two methods: by testing hypotheses on expected correlation strength between change scores of the questionnaires and the GRC scales, and by calculating the area under the receiver operating characteristics curve (AUC). The questionnaires were considered responsive if $\geq$ 75% of the hypotheses were confirmed or if the AUC was $\geq$ 0·7.

Results Eighty-nine participants were recruited in three centres in the Netherlands and the U.S.A., of whom 67 completed all baseline and follow-up questionnaires. For all subscales of the SF-36 and Skindex-29, < 75% of the hypotheses were confirmed and the AUC was < 0·7.

Conclusions Our findings suggest that the SF-36 and Skindex-29 seemed unresponsive to change in QoL. This suggests that alternative PROMs are needed to measure – and ultimately improve – QoL in patients with vascular malformations.

What's already known about this topic?

- Quality of life is often impaired in patients with vascular malformations.
- Quality of life is considered a core outcome domain for evaluating treatment of vascular malformations.
- To measure the effect of treatment on quality of life, a patient-reported outcome measure is required that is responsive to changes in quality of life.

> **What does this study add?**
>
> - This is the first study assessing the responsiveness of quality-of-life measures in patients with vascular malformations.
> - The results seem to indicate that the Medical Outcomes Study Short Form 36 (SF-36) and Skindex-29 are not responsive to changes in quality of life in patients with vascular malformations.

> **What are the clinical implications of this work?**
>
> - Medical Outcomes Study Short Form 36 (SF-36) and Skindex-29 are not ideal to assess the effect on quality of life over time, of treatment strategies for peripheral vascular malformations.

Vascular malformations are rare congenital vascular anomalies, which grow proportionally with age, and can be of venous, lymphatic, arteriovenous, capillary or combined origin.[1] Patients with peripheral vascular malformations commonly experience disfigurement, pain, bleeding, impaired mobility, growth disturbances and bleeding and thrombotic complications.[1–5] Patients have overall poorer quality of life (QoL) than the general population.[6] Some lesions can be life-threatening, but most patients seek treatment to improve aspects of their QoL.

QoL and other patient-reported outcomes (PROs) were therefore recently included in the core domain set (CDS) for clinical research in vascular malformations (excluding capillary).[3] A CDS is a minimum set of outcome domains that should be measured when evaluating treatment outcomes in a certain health condition.[7] The CDS results came from an international consensus project, the Outcome Measures for Vascular Malformations (OVAMA) project,[3,8] which aims at uniform outcome reporting by determining *what* and *how* to measure.

PROs included in the CDS are 'appearance', 'overall symptom severity', 'pain', 'satisfaction with treatment outcome', 'satisfaction with treatment' and multiple aspects of QoL, namely 'activities of daily living', 'mobility', 'ability to work/study', 'confidence and self-esteem' and 'emotional wellbeing'. The next step of the OVAMA project is to decide how to measure these core outcome domains. This comprises selecting or developing outcome measurement instruments measuring the core domains, in other words, developing a core outcome measurement set (COMS).[7,9]

PROs are measured by patient-reported outcome measures (PROMs). There is currently no consensus on how to measure QoL and other core PROs in vascular malformations. Consequently, much is unclear about whether and how current treatment strategies improve QoL in this patient population. To improve QoL outcome, the ability to determine correctly the effect of treatment on different aspects of QoL is essential. This, in turn, requires a measurement instrument that can detect change in the desired outcome domains over time.[10] Identifying such an instrument involves evaluation of the measurement property 'responsiveness'. Responsiveness is defined as 'the ability of an instrument to detect change over time in the construct to be measured.'[11,12] Responsiveness can be assessed by comparing change scores of the measurement instrument and change scores of instruments measuring change in similar outcome domains.[10,12]

In adult patients with peripheral vascular malformations, we assessed the responsiveness of two widely used QoL questionnaires: the Medical Outcomes Study Short Form 36 (SF-36) and the Skindex-29. We thereby evaluated the suitability of these questionnaires to assess treatment effect on the core aspects of QoL in this population, and therefore whether they should be considered for inclusion in the COMS for vascular malformations.

Patients and methods

Patients and data collection

Data were collected prospectively from October 2016 to September 2017 from adult patients (age $\geq$ 18 years) with peripheral vascular malformations (capillary, venous, lymphatic, arteriovenous or combined vascular malformations) visiting the outpatient clinics of the Amsterdam University Medical Centre in Amsterdam, Radboud University Medical Centre in Nijmegen and the Vascular Birthmark Institute in New York. Informed consent was obtained from all participants.

Data on patient characteristics were extracted from the electronic patient files, including sex, age at start of treatment, type of vascular malformations, size (< 5 cm, 5–10 cm, $\geq$ 10 cm), location (head/neck, trunk, upper extremities, lower extremities, combined), lesion depth (skin/subcutaneous tissue, muscle, organs/bone) and previous treatments. We registered whether the patient received conservative treatment (watchful waiting or with compression stockings) or invasive treatment (including sclerotherapy, laser therapy and surgery). Follow-up occurred 6–8 weeks post-treatment, which is customary for evaluating the effect of the treatment in these clinics.

Outcome measures

Medical Outcomes Study Short Form 36

The SF-36 is a generic QoL questionnaire consisting of an item measuring health transition and 35 items forming eight subscales: (i) physical functioning, (ii) social functioning, (iii) role physical, (iv) role emotional, (v) mental health, (vi) vitality, (vii) bodily pain and (viii) general health.[13] Multiple versions are available; we used the freely available RAND-36. Scoring was according to the RAND-36 scoring instructions.[14] Higher scores indicate a better QoL.

Using all items, a physical health component score (PCS) and mental health component score (MCS) can be derived. We computed Z-scores using age- and sex-specific Dutch population normative data.[15] For the SF-36 missing items were handled following the rules of the SF-36 scoring manual. For a scale with < 50% of missing items, an imputation method is provided per subscale for the missing items. If a scale has ≥ 50% missing items, the scale scores were not calculated.

Skindex-29

The Skindex-29 is a 'dermatology-specific' QoL questionnaire, containing 29 items forming three subscales: (i) symptoms, (ii) functioning and (iii) emotions.[17] It also contains an additional item measuring adverse events. Questions refer to the previous 4 weeks. Higher scores indicate worse QoL. For easy interpretation, we calculated Skindex-29 change scores so that a positive change score means improvement, and a negative change score means worsening. Missing items for the Skindex-29 were computed following the Skindex-29 scoring instructions.[18] When ≥ 25% of the items were missing from a scale, the scale score was not calculated.

Global rating of change scales

At follow-up, patients filled in global rating of change (GRC) scales, asking about changes in various QoL domains since the baseline measurement. We formulated seven GRC scales corresponding with the constructs measured by subscales of the SF-36 and Skindex-29, based on the subjective significance questionnaire of Osoba (Appendix S1; see Supporting Information).[19] Patients were asked how much change they had experienced in: 'overall health', 'pain', 'physical condition', 'social relationships and activities', 'emotional wellbeing', 'ability to perform work/study' and 'symptom severity'. Questions were formulated by thoroughly examining individual items of the SF-36 and Skindex-29 subscales.

Patients also filled in two GRC scales about other core PROs for vascular malformations not explicitly covered by the SF-36 and Skindex-29, namely 'appearance' and 'mobility of the affected body part'. These were based on preliminary results of the CDS development study.[3] No GRC scale for the SF-36 subscale 'vitality' was formulated, because it was not considered a core outcome domain for patients with vascular malformations.[3,8]

The change was reported on a seven-point Likert scale, based on previously reported GRC scales and the subjective significance questionnaire of Osoba.[19,20] Response options included 'very much worse', 'worse', 'somewhat worse', 'no change', 'somewhat better', 'better' and 'very much better'. Differences between the conservative and invasive treatment groups were assessed using the Mann–Whitney U-test.

Evaluating responsiveness

Responsiveness was assessed using two methods for seven of the SF-36 subscales (excluding 'vitality') and its two composite scores, and for the three Skindex-29 subscales. The first method involved the construct approach, by testing predefined hypotheses on comparison with other outcome measurement instruments (the GRC scales). Seven hypotheses were formulated per subscale (Table 1). These hypotheses were based on previous studies assessing responsiveness and methodology guidelines by COSMIN.[10,12,21–28]

The first four hypotheses referred to the strength of correlations between change scores of the PROMs, and the GRC scale scores. For each PROM subscale, correlation strength was calculated with a GRC scale measuring change in a similar construct (hypothesis 1), two GRC scales measuring related but dissimilar constructs (hypotheses 2 and 3), and one GRC scale measuring an unrelated construct (hypothesis 4). Two independent researchers (M.M.L. and S.E.R.H.) defined these expected relations beforehand (Table 2) by thoroughly reviewing the subscales. Disagreement was resolved by consensus (M.M.L. and S.E.R.H.). Correlations between PROM

Table 1 Hypotheses for testing responsiveness

1. High positive correlation between SF-36 or Skindex-29 change scores and the GRC scale measuring a similar construct
2. Moderate positive correlation between SF-36 or Skindex-29 change score and a GRC scale measuring a related, but dissimilar construct
3. Moderate positive correlation between SF-36 or Skindex-29 change score and a second GRC scale measuring a related, but dissimilar construct
4. Low positive or negative correlation between SF-36 or Skindex-29 change score and a GRC scale measuring an unrelated construct
5. Patients indicating improvement on the associated GRC scale should have a positive mean change score
6. Patients indicating worsening on the associated GRC scale should have a negative mean change score
7. The mean change score of patients indicating improvement should be higher than the mean change score of unchanged patients, which in turn should be higher than the mean change score of worsened patients

Per Medical Outcomes Study Short Form 36 (SF-36) and Skindex-29 subscale, these seven hypotheses were tested. If ≥ 75% of these hypotheses were confirmed for a subscale, the subscale was considered responsive to change. GRC, global rating of change.

change scores and GRC scale scores were assessed by calculating Spearman's rank correlation coefficients. Correlation was interpreted as high (> 0·5), moderate (0·3–0·5) or low (< 0·3), based on previous studies and guidelines for assessing responsiveness.[10,12,21–29] The next hypotheses (5, 6, 7) concerned the mean PROM change scores of improved, unchanged and worsened patients according to the associated GRC scale, namely the GRC scale measuring a similar construct. These hypotheses therefore use the same data as the first hypothesis, but add more detailed insight on the data. A subscale was considered responsive to change if ≥ 75% of the hypotheses were confirmed, as shown in Table 1.[10,12]

The second method for assessing responsiveness involved calculating the area under the receiver operating characteristics curve (AUC). The AUC indicates how good the instrument is at discriminating patients who improved from those unchanged. The unchanged and improved groups were based on the associated GRC scale. As it dichotomizes the data, it is a rougher but more comprehensible method of interpreting the data. An AUC ≥ 0·7 is considered appropriate.[12,30]

All data were analysed with SPSS statistics 25·0 (IBM, Armonk, NY, U.S.A.).

Results

In total, 109 patients were approached, of whom 89 were included. Sixty-seven (75%) of the included patients completed the SF-36 and Skindex-29 at baseline and follow-up, and also filled in the GRC scales.

Baseline characteristics

The baseline characteristics of all patients who completed the questionnaires at baseline and follow-up and the GRC scales are shown in Table 3. Compared with the 89 included patients, the 20 excluded patients had significantly fewer patients in the > 10cm size category (P = 0·003). No statistically significant differences in baseline characteristics were found between patients who completed both baseline and follow-up questionnaires and those who did not.

Table 2 Predefined expected correlations between Medical Outcomes Study Short Form 36 (SF-36) and Skindex-29 change scores and global rating of change (GRC) scales

	GRC scale measuring similar construct (expected high correlation)	GRC scale measuring related, but dissimilar construct #1 (expected moderate correlation)	GRC scale measuring related, but dissimilar construct #2 (expected moderate correlation)	GRC scale measuring unrelated construct (expected low correlation)
SF-36 GH	Overall health	Physical condition	Emotional wellbeing	Appearance
SF-36 BP	Pain	Overall health	Symptom severity	Appearance
SF-36 PF	Physical condition	Overall health	Pain	Appearance
SF-36 SF	Social relationships and social activities	Overall health	Ability to perform work, study or school-related activities	Mobility of affected body part
SF-36 MH	Emotional wellbeing	Overall health	Social relationships and social activities	Mobility of affected body part
SF-36 RP	Ability to perform work, study or school-related activities	Overall health	Physical condition	Appearance
SF-36 RE	Ability to perform work, study or school-related activities	Overall health	Emotional wellbeing	Mobility of affected body part
SF-36 PCS	Physical condition	Overall health	Pain	Appearance
SF-36 MCS	Emotional wellbeing	Overall health	Social relationships and social activities	Mobility of affected body part
SF-36 sum	Overall health	Physical condition	Symptom severity	Appearance
Skindex S	Symptom severity	Overall health	Pain	Appearance
Skindex F	Social relationships and social activities	Overall health	Ability to perform work, study or school-related activities	Appearance
Skindex E	Emotional wellbeing	Overall health	Social relationships and social activities	Mobility of affected body part
Skindex sum	Overall health	Emotional wellbeing	Symptom severity	Appearance

Per SF-36 and Skindex-29 subscale, a GRC scale measuring change in a similar construct was determined, with which a high correlation was expected. Moderate correlation was expected with GRC scales measuring change in a related, but dissimilar construct (two defined per subscale). Low correlation was expected with a GRC scale measuring change in an unrelated construct. Correlation was interpreted as high (> 0·5), moderate (0·3–0·5) or low (< 0·3). GH, general health; BP, bodily pain; PF, physical function; SF, social function; MH, mental health; RP, role physical; RE, role emotional; PCS, physical health component score; MCS, mental health component score; S, symptoms; F, functioning; E, emotions.

Descriptive data of Medical Outcomes Study Short Form 36 and Skindex-29

Table S1 (see Supporting Information) shows descriptive data of SF-36 and Skindex-29 scores at baseline and follow-up. This includes patients who filled in the questionnaires at both baseline and follow-up. For all SF-36 and Skindex-29 subscales, no significant change was observed between baseline and follow-up. No significant differences were found between the conservative and invasive management groups in baseline scores or change scores. At the baseline measurement, one patient had missing data on the SF-36 'role physical' subscale, and one different patient had missing data on the SF-36 'role emotional' subscale.

Global rating of change scale responses

An overview of the results of the GRC scales is shown in Table S2 (see Supporting Information). Patients who received an invasive treatment indicated more improvement than the conservative group in all GRC scale domains, of which the following were significant by Mann–Whitney U-test: overall health ($P = 0.01$), pain ($P = 0.04$), emotional wellbeing ($P = 0.02$), ability to perform work/study ($P = 0.02$), symptom severity ($P = 0.003$), appearance ($P = 0.008$) and mobility of the affected body part ($P = 0.005$).

Correlation between Medical Outcomes Study Short Form 36 and Skindex-29 score changes and global rating of change scale scores

An overview of the Spearman's rank correlation coefficients between the SF-36 and Skindex-29 scales and all GRC scales is presented in Table S3 (see Supporting Information). No SF-36 or Skindex-29 sum or subscale had high correlation with the GRC scale for which high correlation was expected (GRC scale measuring a similar construct). The Skindex-29 'symptoms' subscale came closest, with a moderate correlation (0·34) with its predefined GRC scale measuring a similar construct (symptom severity). No SF-36 or Skindex-29 scale reached moderate correlation with GRC scales for which moderate correlation was expected (GRC scales measuring related but dissimilar constructs). All SF-36 or Skindex-29 scales had low correlation with the GRC scales for which a low correlation was expected (GRC measuring an unrelated construct).

Responsiveness: hypothesis testing and area under the receiver operating characteristics curve

Table 4 shows the results of hypothesis testing and the AUC values for each subscale. For no sum or subscale were ≥ 75% of hypotheses correct. In addition, for no sum or subscale was the AUC ≥ 0·7. Consequently, all scales of both the SF-36 and Skindex-29 were considered not responsive to change. Table S4 (see Supporting Information) shows the exact values on which the hypotheses were confirmed or rejected.

Correlation between Medical Outcomes Study Short Form 36 change scores and Skindex-29 change scores

An overview of Spearman's rank correlation coefficients between SF-36 change scores and Skindex-29 change scores is presented in Table S5 (see Supporting Information). We chose

Table 3 Baseline characteristics of the 67 patients who completed the questionnaires at baseline and follow-up, and the global rating of change scales

Age at baseline (years), median (IQR)	39 (26–50)
Sex female	43 (64)
Type	
Venous	37 (55)
Arteriovenous	8 (12)
Venous, capillary	6 (9)
Lymphatic	5 (7)
Venous, lymphatic	5 (7)
Venous, lymphatic, capillary	3 (4)
Arteriovenous, capillary	1 (1)
Capillary	1 (1)
Not specified	1 (1)
Localization	
Lower extremity	21 (31)
Head/neck	19 (28)
Trunk, lower extremity	9 (13)
Trunk	8 (12)
Upper extremity	6 (9)
Head/neck, trunk, upper extremity	2 (3)
Head/neck, trunk, upper extremity, lower extremity	1 (1)
Trunk, upper extremity, lower extremity	1 (1)
Size	
< 5 cm	20 (30)
5–10 cm	12 (18)
> 10 cm	35 (52)
Depth/extension	
Skin/subcutaneous tissue	23 (34)
Muscle	29 (43)
Bone/organs	15 (22)
Treatment history	
No prior treatment	22 (33)
Surgery	11 (16)
Elastic stockings	7 (10)
Embolization	3 (4)
Laser therapy	2 (3)
Sclerotherapy	2 (3)
Anticoagulants	1 (1)
A combination of the above	19 (28)
Treatment during study period	
Conservative	42 (63)
Expectant management	34 (51)
Elastic stockings	7 (10)
Invasive	25 (37)
Sclerotherapy	15 (22)
Laser therapy	4 (6)
Surgery	4 (6)
Embolization	1 (1)
Rapamycin	1 (1)
Sclerotherapy, laser therapy	1 (1)

Data are presented as n (%) except for age.

Table 4 Testing results of the hypotheses for evaluation of responsiveness. Exact definitions of the hypotheses are shown in Table 2. Exact values are presented in Table S4 (see Supporting Information)

	Spearman's rank correlation coefficients with GRC scales[a]				5. Mean change improved[b]	6. Mean change worsened[c]	7. Mean change improved > unchanged > worsened	Hypotheses confirmed	AUC	Responsive to change
	1. Similar (expected: high)	2. Related #1 (expected: moderate)	3. Related #2 (expected: moderate)	4. Unrelated (expected: low)						
SF-36 GH	Low	Low	Low	**Low**	Negative	Positive	No	14%	< 0·7	No
SF-36 BP	Low	Low	Low	**Low**	**Positive**	Positive	No	29%	< 0·7	No
SF-36 PF	Low	Low	Low	**Low**	**Positive**	**Negative**	**Yes**	57%	< 0·7	No
SF-36 SF	Low	Low	Low	**Low**	Negative	**Negative**	No	29%	< 0·7	No
SF-36 MH	Low	Low	Low	**Low**	**Positive**	**Negative**	**Yes**	57%	< 0·7	No
SF-36 RP	Low	Low	Low	**Low**	**Positive**	Positive	No	29%	< 0·7	No
SF-36 RE	Low	Low	Low	**Low**	**Positive**	Positive	No	29%	< 0·7	No
SF-36 PCS	Low	Low	Low	**Low**	**Positive**	Positive	No	29%	< 0·7	No
SF-36 MCS	Low	Low	Low	**Low**	**Positive**	**Negative**	**Yes**	57%	< 0·7	No
SF-36 sum	Low	Low	Low	**Low**	**Positive**	**Negative**	**Yes**	57%	< 0·7	No
Skindex S	Moderate	Low	Low	**Low**	**Positive**	**Negative**	**Yes**	57%	< 0·7	No
Skindex F	Low	Low	Low	**Low**	Negative	Positive	No	14%	< 0·7	No
Skindex E	Low	Low	Low	**Low**	**Positive**	Positive	No	29%	< 0·7	No
Skindex sum	Low	Low	Low	**Low**	**Positive**	Positive	No	29%	< 0·7	No

Confirmed hypotheses are shown in bold. A subscale was considered responsive to change if ≥ 75% of the hypotheses were confirmed. GRC, global rating of change; AUC, area under the receiver operating characteristics curve; GH, general health; BP, bodily pain; PF, physical function; SF, social function; MH, mental health; RP, role physical; RE, role emotional; PCS, physical health component score; MCS, mental health component score; S, symptoms; F, functioning; E, emotions. [a]Similar: correlation strength with the GRC scale measuring a similar construct as defined in Table 1. Related #1: correlation strength with the first GRC scale measuring a related, but dissimilar construct as defined in Table 1. Related #2: correlation strength with the second GRC scale measuring a related, but dissimilar construct as defined in Table 1. Unrelated: correlation strength with the GRC scale measuring an unrelated construct as defined in Table 1. Correlation was interpreted as high (> 0·5), moderate (0·3–0·5) or low (< 0·3). [b]The mean change score of the patients indicating improvement on the associated GRC scale. [c]The mean change score of the patients indicating worsening on the associated GRC scale.

to calculate these correlations with the Spearman's rank test, as the change scores were not normally distributed. The only moderate correlation was between the SF-36 'social functioning' subscale and the Skindex-29 'functioning' subscale (0·33). All other correlations were low.

Discussion

These data suggest that all subscales of both the SF-36 and Skindex-29 were unresponsive to change in QoL in this study population of patients with peripheral vascular malformations. Our study therefore implies that these PROMs are not ideal to assess the effect on QoL over time of treatment strategies for peripheral vascular malformations. As of yet, they seem unreliable for inclusion in the COMS for peripheral vascular malformations.

To our knowledge, this is the first study assessing the responsiveness of QoL questionnaires in patients with vascular malformations. In a systematic review on measurement instruments used in patients with vascular malformations,[31] we found no studies assessing responsiveness in this population or similar populations. Very little was found on the assessment of responsiveness of the SF-36 and Skindex-29 in other patient populations, and in the studies that we found, mostly outdated methods were used.[32–36]

Our hypotheses for testing responsiveness were formulated according to the COSMIN methodology and previous studies assessing responsiveness. The hypotheses for determining responsiveness are somewhat subjective; however, as none of the numerous hypotheses were confirmed, the results seem indisputable.

Although GRC scales are viewed as the best single measure of the importance of change from the patient's perspective,[37] it is difficult to establish evidence that this is the correct assessment of change. It is impossible to determine the responsiveness to change of GRC scales, as they directly assess change with a single measurement. What supports the accuracy of the GRC scales in this study is that patients treated invasively indicated significantly more improvement in almost all GRC scales than patients treated conservatively. Also, the low correlation coefficients between SF-36 and Skindex-29 scales measuring similar or related constructs support our conclusion that these questionnaires may not correctly detect change in the constructs they aim to measure.

Symptoms of vascular malformations are known to vary according to type, size and location. This, combined with the rarity of the disease, makes a heterogeneous study population unavoidable. The responsiveness to change of these questionnaires might differ for certain subgroups. However, our objective is to find a PROM suitable for all peripheral vascular

malformations. The COMS for vascular malformations should include outcome measurements responsive to change for all patients with peripheral vascular malformations.

A sample size of > 50 is considered adequate for studies on responsiveness,[38] although in another guideline, sample size standards were explicitly removed.[21] Either way, our sample is more than sufficient, especially considering that a study population of 67 is exceptionally high for studies on this rare disease. If a much larger study population is needed in order to observe differences, this would mean that the questionnaire might not be a feasible option to measure treatment effect in this patient population.

A possible explanation for the difference in size between the included and excluded patients might be that patients with larger vascular malformations are more willing to fill in the questionnaires, as their life might be impacted more by the disease. Future research with an adequate QoL measurement instrument must uncover the relationship between malformation size and QoL.

Even though a period of 6–8 weeks post-treatment is standard for evaluating treatment effect, a longer follow-up period might be necessary to measure some QoL aspects. This may especially be the case for the outcome domains 'social relationships and social activities' and 'ability to perform work/school/study'. On the contrary, the same constructs measured by the GRC scales were often considered improved by the participating patients. In seven of the outcome domains there were more than 10 patients indicating improvement. Moreover, a large group of patients underwent invasive treatment; change could be expected in this group. Although significant differences were found between the invasive group and the conservatively managed group in seven of nine GRC scales, there were no significant differences in the corresponding SF-36 and Skindex-29 change scores.

Interpretability, mostly determined by the minimal important change (MIC) and smallest detectable change (SDC) of the questionnaire, is an important characteristic of a PROM,[10] defined as 'the degree to which one can assign qualitative meaning to an instrument's quantitative scores or change in scores.'[12] Only one subscale satisfied the criteria for calculating the MIC and SDC (≥ 0.3 correlation with its anchor, the associated GRC scale): the Skindex-29 'symptoms' scale. Using different methods, we found MICs of 5–8 points, next to a very large SDC (37 points). However, interpretability is only meaningful if the measurement instrument shows agreeable responsiveness. In other words, in order to calculate which 'changes' in questionnaire scores are clinically important, there have to be actual measurable 'changes' in the questionnaire scores, which we did not find in this study. Thus, the MIC values are not reliable and would be unfitting to show in the results section. However, in light of the large measurement error it suggests that, if this scale were responsive to change, it would probably not be able to detect meaningful change on an individual level.

We found similar large measurement errors for the other subscales, which possibly is the reason why these PROMs showed insufficient responsiveness. It is possible that in larger study populations, responsiveness might improve because of a smaller measurement error. However, clinical research on vascular malformations mostly involves smaller patient groups, so a PROM with smaller measurement error is needed. Causes of the large measurement errors might be that the PROMs were developed for use in larger populations, or that the items and response options are not appropriate for this population. Another reason for low responsiveness might be that the subscales of the SF-36 and Skindex-29 measure multiple constructs of interest, averaging out possible change in one of these constructs. In addition, the scales might contain irrelevant questions for patients with vascular malformations, masking change in the items of interest.

If we want to improve the most important outcomes from the perspective of patients with vascular malformations, we need a PROM that can detect changes in the core PROs (including QoL) over time to evaluate effects of treatment strategies. The next step in improving assessment of QoL in patients with vascular malformations is therefore identifying PROMs that are capable of detecting change in the core PROs over time. Other existing PROMs may be explored, or a new disease-specific PROM will have to be developed. The solution might be using PROMs that have smaller measurement error, have good content validity and measure the core PROs for vascular malformations more 'unidimensionally'.

New methodologies such as item response theory and computer adaptive tests enable tailoring of questionnaires to the individual, which improves precision and accuracy.[39] Especially in a heterogeneous patient population such as this, these might have better measurement properties. We plan on exploring the measurement properties of the unidimensional Patient-Reported Outcomes Measurement Information System (PROMIS) item banks, which were developed with item response theory and can be administered as computer adaptive tests.[40] Because not all outcome domains from the CDS can potentially be covered by PROMIS scales, we also plan on developing a disease-specific questionnaire, with a focus on appearance and disease-specific symptoms for vascular malformations.

References

1 Wassef M, Blei F, Adams D *et al.* Vascular anomalies classification: recommendations from the International Society for the Study of Vascular Anomalies. *Pediatrics* 2015; **136**:e203–14.

2 Dasgupta R, Fishman SJ. ISSVA classification. *Semin Pediatr Surg* 2014; **23**:158–61.

3 Horbach SER, van der Horst C, Blei F *et al.* Development of an international core outcome set for peripheral vascular malformations: the OVAMA project. *Br J Dermatol* 2018; **178**:473–81.

4 Horbach SE, Lokhorst MM, Oduber CE *et al.* Complications of pregnancy and labour in women with Klippel-Trenaunay syndrome: a nationwide cross-sectional study. *BJOG* 2017; **124**:1780–8.

5 Horbach SE, Lokhorst MM, Saeed P *et al.* Sclerotherapy for low-flow vascular malformations of the head and neck: a systematic review of sclerosing agents. *J Plast Reconstr Aesthet Surg* 2016; **69**:295–304.

6 Nguyen HL, Bonadurer GF 3rd, Tollefson MM. Vascular malformations and health-related quality of life: a systematic review and meta-analysis. *JAMA Dermatol* 2018; **154**:661–9.

7 Boers M, Kirwan JR, Wells G *et al.* Developing core outcome measurement sets for clinical trials: OMERACT filter 2.0. *J Clin Epidemiol* 2014; **67**:745–53.

8 Lokhorst MM, Horbach SER, van der Horst C *et al.* Finalizing the international core domain set for peripheral vascular malformations: the OVAMA project. *Br J Dermatol* 2019; **181**:1076–8.

9 Schmitt J, Apfelbacher C, Spuls PI *et al.* The Harmonizing Outcome Measures for Eczema (HOME) roadmap: a methodological framework to develop core sets of outcome measurements in dermatology. *J Invest Dermatol* 2015; **135**:24–30.

10 Mokkink LB, Terwee CB, Patrick DL *et al.* The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res* 2010; **19**:539–49.

11 Mokkink LB, Terwee CB, Patrick DL *et al.* The COSMIN study reached international consensus on taxonomy, terminology, and definitions of measurement properties for health-related patient-reported outcomes. *J Clin Epidemiol* 2010; **63**:737–45.

12 De Vet HC, Terwee CB, Mokkink LB *et al. Measurement in Medicine.* Cambridge: Cambridge University Press, 2011.

13 Breugem CC, Merkus MP, Smitt JH *et al.* Quality of life in patients with vascular malformations of the lower extremity. *Br J Plast Surg* 2004; **57**:754–63.

14 RAND Health Care. 36-Item Short Form Survey (SF-36). Available at: https://www.rand.org/health-care/surveys_tools/mos/36-item-short-form.html (last accessed 31 October 2019).

15 Aaronson NK, Muller M, Cohen PD *et al.* Translation, validation, and norming of the Dutch language version of the SF-36 Health Survey in community and chronic disease populations. *J Clin Epidemiol* 1998; **51**:1055–68.

16 Ware JE, Snow KK, Kosinski M *et al. SF-36 Health Survey. Manual and Interpretation Guide.* Boston: The Health Institute, New England Medical Center, 1993.

17 Oduber CE, Khemlani K, Sillevis Smitt JH *et al.* Baseline quality of life in patients with Klippel-Trenaunay syndrome. *J Plast Reconstr Aesthet Surg* 2010; **63**:603–9.

18 Handleiding Nederlandstalige. Skindex-29. Available at: https://docplayer.nl/19876454-Handleiding-nederlandstalige-skindex-29-een-dermatologiespecifieke-kwaliteit-van-leven-vragenlijst.html (last accessed 31 October 2019).

19 Osoba D. Interpreting the meaningfulness of changes in health-related quality of life scores: lessons from studies in adults. *Int J Cancer Suppl* 1999; **12**:132–7.

20 Kamper SJ, Maher CG, Mackay G. Global rating of change scales: a review of strengths and weaknesses and considerations for design. *J Man Manip Ther* 2009; **17**:163–70.

21 Mokkink LB, de Vet HCW, Prinsen CAC *et al.* COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. *Qual Life Res* 2018; **27**:1171–9.

22 Lee AC, Driban JB, Price LL *et al.* Responsiveness and minimally important differences for 4 patient-reported outcomes measurement information system short forms: physical function, pain interference, depression, and anxiety in knee osteoarthritis. *J Pain* 2017; **18**:1096–110.

23 Assas M, Wiriyakijja P, Fedele S *et al.* Measurement properties of patient-reported outcome measures in radiotherapy-induced trismus. *J Oral Pathol Med* 2019; **48**:351–7.

24 Stuge B, Jenssen HK, Grotle M. The Pelvic Girdle Questionnaire: responsiveness and minimal important change in women with pregnancy-related pelvic girdle pain, low back pain, or both. *Phys Ther* 2017; **97**:1103–13.

25 de Boer MR, Terwee CB, de Vet HC *et al.* Evaluation of cross-sectional and longitudinal construct validity of two vision-related quality of life questionnaires: the LVQOL and VCM1. *Qual Life Res* 2006; **15**:233–48.

26 Dommasch ED, Shin DB, Troxel AB *et al.* Reliability, validity and responsiveness to change of the Patient Report of Extent of Psoriasis Involvement (PREPI) for measuring body surface area affected by psoriasis. *Br J Dermatol* 2010; **162**:835–42.

27 Thoomes-de Graaf M, Scholten-Peeters W, Duijn E *et al.* The responsiveness and interpretability of the Shoulder Pain and Disability Index. *J Orthop Sports Phys Ther* 2017; **47**:278–86.

28 van der Linde JA, van Kampen DA, van Beers L *et al.* The responsiveness and minimal important change of the Western Ontario Shoulder Instability Index and Oxford Shoulder Instability Score. *J Orthop Sports Phys Ther* 2017; **47**:402–10.

29 Mokkink LB, Prinsen CAC, Patrick DL *et al.* COSMIN methodology for systematic reviews of Patient-Reported Outcome Measures (PROMs). Available at: https://www.cosmin.nl/wp-content/uploads/COSMIN-syst-review-for-PROMs-manual_version-1_feb-2018.pdf (last accessed 31 October 2019).

30 Terwee CB, Bot SD, de Boer MR *et al.* Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol* 2007; **60**:34–42.

31 Horbach SER, Rongen APM, Elbers RG, et al. Outcome measurement instruments for peripheral vascular malformations and an assessment of the measurement properties: a systematic review. *Qual Life Res* 2019; https://doi.org/10.1007/s11136-019-02301-x.

32 Chren MM, Lasek RJ, Quinn LM *et al.* Skindex, a quality-of-life measure for patients with skin disease: reliability, validity, and responsiveness. *J Invest Dermatol* 1996; **107**:707–13.

33 Vasquez D, Aguirre DC, Sanclemente G. Construct validity and responsiveness of the Colombian version of Skindex-29. *Br J Dermatol* 2019; **181**:770–7.

34 Erez G, Selman L, Murtagh FE. Measuring health-related quality of life in patients with conservatively managed stage 5 chronic kidney disease: limitations of the Medical Outcomes Study Short Form 36: SF-36. *Qual Life Res* 2016; **25**:2799–809.

35 Garratt AM, Ruta DA, Abdalla MI *et al.* Responsiveness of the SF-36 and a condition-specific measure of health for patients with varicose veins. *Qual Life Res* 1996; **5**:223–34.

36 Devilliers H, Amoura Z, Besancenot JF *et al.* Responsiveness of the 36-item Short Form Health Survey and the Lupus Quality of Life questionnaire in SLE. *Rheumatology (Oxford)* 2015; **54**:940–9.

37 Crosby RD, Kolotkin RL, Williams GR. Defining clinically meaningful change in health-related quality of life. *J Clin Epidemiol* 2003; **56**:395–407.

38 Mokkink LB, Prinsen CAC, Patrick DL *et al.* COSMIN study design checklist for patient-reported outcome measurement instruments. Available at: https://www.cosmin.nl/wp-content/uploads/COSMIN-study-designing-checklist_final.pdf (last accessed 31 October 2019).

39 Cella D, Gershon R, Lai JS *et al.* The future of outcomes measurement: item banking, tailored short-forms, and computerized adaptive assessment. *Qual Life Res* 2007; **16** (Suppl. 1):133–41.

40 Cella D, Riley W, Stone A *et al.* The Patient-Reported Outcomes Measurement Information System (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *J Clin Epidemiol* 2010; **63**:1179–94.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix S1 Global rating of change scales.

Table S1 Descriptive data of the Short Form 36 and Skindex-29.

Table S2 Global rating of change scale responses.

Table S3 Spearman's rank correlation coefficients between Short Form 36 and Skindex-29 change scores and global rating of change scales.

Table S4 The exact values on which the hypotheses for evaluating responsiveness were confirmed or rejected.

Table S5 Spearman's rank correlation coefficients between Short Form 36 and Skindex-29 change scores.

Powerpoint S1 Journal Club Slide Set.

Video S1 Author video.