Dermatology

Outcome Measures – Review Article

Dermatology 2021;237:416–432
DOI: 10.1159/000511438

Received: May 3, 2020
Accepted: September 3, 2020
Published online: December 3, 2020

Treatment Outcome Measurement Instruments for Port Wine Stains: A Systematic Review of Their Measurement Properties

M. Ingmar van Raath[a–d] Sandeep Chohan[e] Albert Wolkerstorfer[e]
Chantal M.A.M. van der Horst[f] Jacqueline Limpens[g] Xuan Huang[d]
Baoyue Ding[d] Gert Storm[c] René R.W.J. van der Hulst[a, b] Michal Heger[c, d]

[a]Department of Plastic, Reconstructive and Hand Surgery, Maastricht University Medical Center, Maastricht University, Maastricht, The Netherlands; [b]NUTRIM School of Nutrition and Translational Research in Metabolism, Maastricht University Medical Center, Maastricht University, Maastricht, The Netherlands; [c]Department of Pharmaceutics, Utrecht Institute for Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands; [d]Department of Pharmaceutics, Jiaxing Key Laboratory for Photonanomedicine and Experimental Therapeutics, College of Medicine, Jiaxing University, Jiaxing, PR China; [e]Department of Dermatology, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands; [f]Department of Plastic, Reconstructive and Hand Surgery, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands; [g]Medical Library, Amsterdam UMC, University of Amsterdam, Amsterdam, The Netherlands

Keywords

Capillary malformation · Clinical assessment · Clinimetric study · Psychometric evaluation · Quality of life

Abstract

Background: A plethora of outcome measurement instruments (OMIs) are being used in port wine stain (PWS) studies. It is currently unclear how valid, responsive, and reliable these are. ***Objectives:*** The aim of this systematic review was to appraise the content validity and other measurement properties of OMIs for PWS treatment to identify the most appropriate instruments and future research priorities. ***Methods:*** This study was performed using the updated Consensus-Based Standards for the Selection of Health Measurement Instruments (COSMIN) methodology and adhered to PRISMA guidelines. Comprehensive searches in Medline and Embase were performed. Studies in which an OMI for PWS patients was developed or its measurement properties were evaluated were included. Two investigators independently extracted data and assessed the quality of included studies and instruments to perform qualitative synthesis of the evidence. ***Results:*** In total, 1,034 articles were screened, and 77 full-text articles were reviewed. A total of 8 studies were included that reported on 6 physician-reported OMIs of clinical improvement and 6 parent- or patient-reported OMIs of life impact, of which 3 for health-related quality of life and 1 for perceived stigmatization. Overall, the quality of OMI development was inadequate (63%) or doubtful (37%). Each instrument has undergone a very limited evaluation in PWS patients. No content validity studies were performed. The quality of evidence for content validity was very low (78%), low (15%), or moderate (7%), with sufficient comprehensibility, mostly sufficient comprehensiveness, and mixed relevance. No studies on responsiveness, minimal important change, and cross-cultural va-

PROSPERO Registration No.: CRD42019119252.

karger@karger.com
www.karger.com/drm

Karger
OPEN ACCESS

Michal Heger
Department of Pharmaceutics
Utrecht Institute of Pharmaceutical Sciences
Utrecht University, NL–3584 CG Utrecht (The Netherlands)
m.heger@uu.nl

lidity were retrieved. There was moderate- to very low-quality evidence for sufficient inter-rater reliability for some clinical PWS OMIs. Internal consistency and measurement error were indeterminate in all studies. ***Conclusions:*** There was insufficient evidence to properly guide outcome selection. Additional assessment of the measurement properties of OMIs is needed, preferentially guided by a core domain set tailored to PWS.

Introduction

Port wine stains (PWS) are congenital capillary malformations resulting from differentiation-impaired endothelial cells with a progressive dilation of immature, venule-like vasculature [1, 2]. These lesions occur in 0.3–0.5% of the population, enlarge and darken proportionally to age [3–5], and can present a significant psychological burden [6, 7]. Pulsed dye laser (PDL) treatment is the current gold standard but fails to achieve optimal results in a significant proportion of patients [8].

Although much effort has been devoted to developing technical tools to objectively quantify PWS blanching [9–12], (physician- or patient-reported) clinical outcome measurement instruments (OMIs) appear to have undergone far less scrutiny, despite the fact that robust and uniform clinical PWS OMIs are crucial in clinical trials to steer the field towards progress. To date, progress has not materialized [8] notwithstanding the medical need and patient demand [13]. A recent systematic review of all prospective PWS studies performed since 2005 by our group unveiled the wide variety of clinical scoring systems [14], a phenomenon that has hampered study comparisons and meta-analyses. An appropriate OMI must have good measurement properties: it has to be valid (i.e., it measures what it purports to measure), reliable, and responsive to change. This information is derived from (high-quality) clinimetric studies.

The aim of this systematic review was to critically appraise the content validity and other measurement properties of all patient-, parent-, and physician-reported OMIs for the evaluation of clinical outcome or life impact of PWS treatments using the Consensus-Based Standards for the Selection of Health Measurement Instruments (COSMIN) framework [15]. Secondary aims were to establish the most appropriate OMI for measuring effectiveness in PWS and to identify areas for future research.

Methods

This systematic review was reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines, and the study protocol was registered in PROSPERO (CRD42019119252) [16]. A medical librarian (J.L.) performed a systematic search in OVID Medline and OVID Embase, using controlled terms and free-text words, combined with a clinimetric search filter developed for PubMed by Terwee et al. [17], adapted for Medline and Embase, and extended with feasibility, acceptability, and practicability (online suppl. Table 1; see www.karger.com/doi/10.1159/000511438 for all online suppl. material). The search was last performed on April 1, 2019. Nonessential methodological details are provided in the online supplementary Methods for purposes of brevity.

The aim of the included studies had to be the development of an OMI or the evaluation of one or more of its measurement properties. The studies had to have either an observer-/clinician-reported OMI or patient-/parent-reported outcome measure (PROM) for the evaluation of effectiveness in PWS. This included clinical OMIs as well as measures of life impact (in accordance with COMET classification; life impact includes the domains (psychosocial) functioning, perceived health status, and quality of life) [18], even if these OMIs had not yet been used as an outcome of effectiveness but could be so used (i.e., cross-sectional studies were also eligible). Further exclusion and inclusion criteria as well as data extraction procedures are addressed in the online supplementary Methods [19–22].

The updated COSMIN methodology was employed to assess the measurement properties (listed and defined in accordance with COSMIN taxonomy in online suppl. Table 2) [15, 23–26]. Although the COSMIN checklists have been developed primarily for the evaluation of PROMs, they are widely implemented for clinician-reported OMIs too [27–29]. Accordingly, some standards were ignored or adapted where necessary (online suppl. Methods). Studies on the comprehensibility of included OMIs in other, yet similar populations were also retrieved and considered inasmuch as these can provide meaningful evidence.

For each included OMI the original development study, manual, and/or additional resources were also retrieved online and by contacting authors (online suppl. Methods) and evaluated because proper item construction during development helps to ensure content validity. It was decided a priori that there is no gold standard in the assessment of PWS, therefore criterion validity and criterion approach for responsiveness could not be assessed (except to compare short versions to the original (long) questionnaires). For the assessment of construct validity, the generic COSMIN hypotheses were applied [15]. In order to help establish whether OMIs adhered to a formative and/or a reflective model, we used the checklist of Fleuren et al. [30].

Assessment of the methodological quality of the OMI's development study, additional content validity studies, and studies on measurement properties was performed independently by 2 reviewers (M.I.R. and S.C.) using the corresponding COSMIN risk of bias checklists [23] (and rated "very good," "adequate," "doubtful," or "inadequate"). A score per section per study was determined using the lowest rating of any item. The content validity and measurement properties themselves were assessed using the predefined COSMIN criteria for good content validity and the updated criteria for good measurement properties, respectively, and

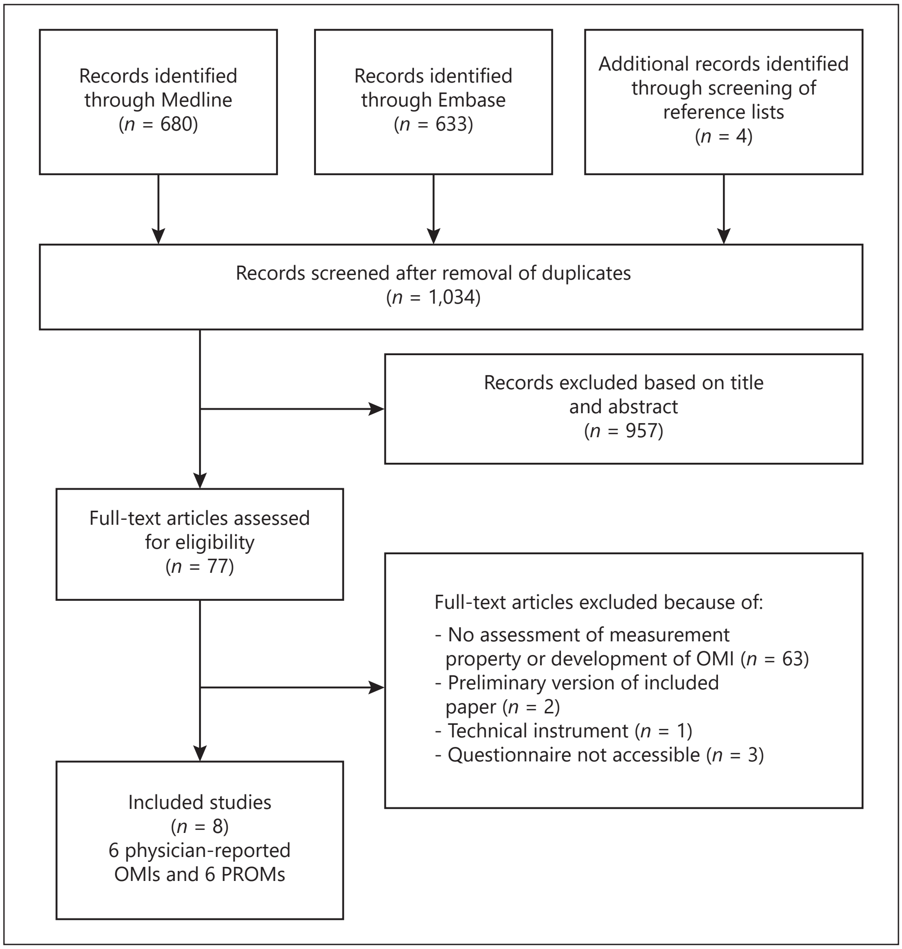
Records identified through Medline (n = 680)
Records identified through Embase (n = 633)
Additional records identified through screening of reference lists (n = 4)
Records screened after removal of duplicates (n = 1,034)
Records excluded based on title and abstract (n = 957)
Full-text articles assessed for eligibility (n = 77)
Full-text articles excluded because of:
- No assessment of measurement property or development of OMI (n = 63)
- Preliminary version of included paper (n = 2)
- Technical instrument (n = 1)
- Questionnaire not accessible (n = 3)
Included studies (n = 8) 6 physician-reported OMIs and 6 PROMs

Fig. 1. PRISMA flow chart showing the study selection and exclusion process. OMI, outcome measurement instrument; PROM, patient-reported outcome.

rated as sufficient (+), insufficient (–), or indeterminate (?) (online suppl. Table 2) [31, 32]. For all reviewer ratings of content validity, an expert panel comprising a plastic surgeon (C.M.A.M.H.) and dermatologist (A.W.) was consulted.

For each OMI the measurement properties and corresponding quality of evidence were summarized and pooled if possible. The overall ratings for measurement properties were rated as sufficient (+), insufficient (–), indeterminate (?), or inconsistent (±) [32]. Then, the overall rating of the quality of evidence was given using the modified GRADE approach ("high," "moderate," "low," or "very low" quality of evidence) [33].

Results

Study Characteristics

The search retrieved 1,030 unique hits (Fig. 1). After reviewing 77 full-text articles, 8 studies were included [34–41]. Reference checking yielded 4 additional eligible articles: 1 was not analyzed because it was a preliminary version of another included study [42] and 3 articles with some clinimetric assessment (albeit limited in reporting and/or quality) of self-constructed questionnaires regarding PWS-related stress and stigmatization were excluded because the questionnaires proved impossible to source [6, 43, 44]. Searches for studies on the comprehensibility of included OMIs performed in other populations yielded 1 additional study [45].

The characteristics of included studies are presented in Table 1. In these studies, 6 physician-reported clinical OMIs and 6 PROMs of life impact were evaluated (summarized in Table 2).

Included Clinical (Physician-Reported) OMIs

Koster et al. [39] have developed the most comprehensive PWS-specific instrument (Table 2). This physician-reported questionnaire contains 8 items (PWS color, patchiness, boundary, pigmentation, size, shape, surface, and hypertrophy) and requires pre- and post-treatment measurement.

Sajan et al. [37] developed a questionnaire for facial infantile hemangiomas and PWS in children that covers

Table 1. Characteristics of the included studies and study populations

OMI(s)	Reference	Population				Disease characteristics			Study design	Instrument administration			
		number	mean age (SD, range)	gender, % female	country	disease	lesion location	previous treatment		treatment	setting	language	assessor background
Koster's PWS questionnaire	Koster et al. [39] (1998)	70	11.4 (8.6, 0–30) years	65.7	The Netherlands	PWS	Head/neck	None	Prospective	PDL (5 sessions)	Rating of standardized pairs of pre- and posttreatment photographs	English	The treating physician, 2 plastic surgeons, a dermatologist, and a clinical photographer
Sajan's facial PWS and IH questionnaire	Sajan et al. [37] (2013)	22	Mean 10 months (exclusively pediatric patients)	NL	USA	PWS (n = 5, of which 1 with SWS), IH (n = 17)	Face	NL	Retrospective	PDL (average 3.6 sessions) (+ surgical reconstruction for 17 IH patients)	Assessment of pre- and posttreatment photographs	English	3 pediatric otolaryngologists and facial plastic surgeons
Poor/fair/good/excellent; 0–24; 25–49; 50–74 and 75–100% lightening; class 0–IV (Achauer)	Currie et al. [34] (2000)	20	10.5 (range 2–31) years	NL	UK	PWS	Head and neck (90%), upper limb (5%), and lower limb (5%)	None	Prospective	PDL	Assessment of standardized paired pre- and post-treatment photographs	English	1 plastic surgeon, 1 dermatologist, and 2 plastic surgery research fellows (from the Laser Treatment Center), 2 nurses; all experienced in laser treatment
Percentage PWS clearance	Pérez et al. [36] (1997)	80	NL	NL	Spain	PWS (25.4% pink, 53.1% red, 21.5% purple)	NL	NL	Prospective	PDL (1 or several sessions)	Assessment of standardized pre- and posttreatment photographs	NL	3 dermatologists usually dedicated to PWS management)
Percentage PWS treatment success	Szychta et al. [35] (2013)	48	Adults		UK	PWS	Face or neck	NL	Retrospective	PDL	Rating of pairs of pre- and post-treatment photographs	English	3 plastic surgeons highly experienced in laser treatment; 3 lay people
Lighter/darker/did not change	Naran et al. [38] (2008)	21	Median 5 (range 1–15) years	NL	USA	PWS	Head and neck	At least 4 prior PDL sessions (median: 5)	Prospective	PDL	Assessment of standardized paired pre- and post-treatment photographs	English	A surgeon, medical student, and data manager
DLQI	Wang et al. [40] (2017)	197	>16 years, 29.4% >60 years	64.4	China	PWS	In exposed skin	NL	Cross-sectional	NA	Questionnaires	Chinese	NA
DLQI[1]	Safikhani et al. [45] (2013)	21	48.8 (16.9, 18–85) years	52.4	USA	Psoriasis (moderate-to-severe)	NA	NA	Prospective	NA	Cognitive interview	English	NA
PSQ (child-[2] and parent-reported), CBCL/1.5–5[2], CBCL/4–18[2], TAPQOL[2], KIDSCREEN-27 (child- and parent-reported[2])	Masnari et al. [41] (2013)	88	6.31 (4.66, 0.75–15.75) years	45.5	Swiss	Burn scar (n = 25), PWS (n = 19), IH (n = 36), or congenital melanocytic nevus (n = 8; lesion ≥1 cm[2)])	Face	NL	Cross-sectional	NA	Questionnaires	German	NA

CBCL, Child Behavior Checklist; DLQI, Dermatology Life Quality Index; IH, infantile hemangioma; NA, not applicable; NL, not listed; OMI, outcome measurement instrument; PSQ, Perceived Stigmatization Questionnaire; PDL, pulsed dye laser; PWS, port wine stain; SD, standard deviation; SWS, Sturge-Weber syndrome; TAPQOL, TNO-AZL questionnaire for preschool children's health-related quality of life. [1] Study performed in non-PWS patients, included because it provides evidence on comprehensibility. [2] Questionnaires were completed in an age-specific subset of patients only.

Table 2. Characteristics of the included outcome measurement instruments

OMI (presenting reference)	Construct(s)	Target population	Mode of administration	Recall period	(Sub)scale(s) (number of items)	Response options	Range of scores/ scoring options	Original language	Available translations	Completion time	Licensing/ costs
Koster's PWS questionnaire (Koster et al. [39])	PWS color, patchiness, boundary, hypo-/ hyperpigmentation, size, shape, surface, hypertrophy	PWS patients treated with PDL	Physician-reported	NA (photograph-based)	8	Varies per question	Total score	English	None	2–5 min	Free
Sajan's PWS and IH questionnaire (Sajan et al. [37])	Improvement in color, thickness, and size, side effects (atrophy, scarring, and hypopigmentation)	IH and PWS patients treated with PDL	Physician-reported	NA (photograph-based)	6	0, 1–24, 25–49, 50–75, 76–99, 100% (for improvement in color and thickness)	Individual item scores (no total score)	English	None	1–2 min	Free
Lighter/darker/ did not change (Naran et al. [38])	PWS color improvement	PWS patients treated with PDL	Physician-reported	NA (photograph-based)	1	Lighter, darker, did not change		English	NA	1 min	Free
Poor/fair/good/ excellent (Currie and Monk [34])	Overall PWS response	PWS patients/ undefined	Physician-reported	NA (photograph-based)	1	Poor (defined as almost no change); fair (partly cleared); good (much improved); or excellent (essentially gone)		English	NA	1–2 min	Free
Percentage lightening [ordinal] (Currie and Monk [34])	Degree of PWS lightening/ success	PWS patients/ undefined	Physician-reported	NA (photograph-based)	1	0–24%, 25–49%, 50–74%, 75–100% lightening		English	NA	1 min	Free
Percentage clearance [contin.] (Pérez et al. [36])	PWS clearance	PWS patients/ undefined	Physician-reported	NA (photograph-based)	1	0–100% clearance		English	NA	1 min	Free
Percentage clearance [ordinal] (Pérez et al. [36])	PWS clearance	PWS patients/ undefined	Physician-reported	NA (photograph-based)	1	0–24% (absence of clearance), 25–49% (poor clearance), 50–74% good clearance, and 75–100% (excellent clearance)		English	NA	1 min	Free
Percentage success [contin.] (Szychta et al. [35])	PWS clearance	PWS patients/ undefined	Physician-reported	NA (photograph-based)	1	0–100% success		English	NA	1 min	Free
Class 0–IV (Achauer et al., modified by Currie and Monk [34])	Overall PWS response	PWS patients/ undefined	Physician-reported	NA (photograph-based)	1	Class 0 (essentially gone); class I (faint, barely discernible borders); class II (well-defined borders with areas of normal skin interspersed within the lesion); class III (well-defined borders, solid lesion with no areas of normal skin); or class IV (all of class III plus raised and thickened). This was categorized as: class 0, excellent; class I, excellent; class II, good; class III, fair; class IV, poor		English	NA	1–2 min	Free

Table 2 (continued)

OMI (presenting reference)	Construct(s)	Target population	Mode of administration	Recall period	(Sub)scale(s) (number of items)	Response options	Range of scores/ scoring options	Original language	Available translations	Completion time	Licensing/ costs
DLQI (Finlay et al. [47])	Dermatology-specific quality of life	Dermatological patients	Patient-reported	1 week	10	Very much, a lot, a little, not at all	0–1, no effect; 2–5, small effect; 6–10, moderate effect; 11–20, very large effect; 21–30, extremely large effect	English	>110 translations	1–2 min	Free for most purposes
PSQ (Lawrence et al. [48])	Perceived stigmatization	People with physical distinction	Parent-/patient-reported	1 year	3 (21)	Never, almost never, sometimes, often, always	Total score and subscores (1–5) for absence of friendly behavior, confused/ staring behavior, and hostile behavior	English	German, Portuguese	<5 min	Free
TAPQOL (Fekkes et al. [49])	Parent-perceived HRQoL (domains: physical (incl. motor functioning[1]), social[1], cognitive[1], and emotional functioning)	Largely healthy children (9 months to 6 years old)	Parent-reported	3 months	12 (43)	Varies per question	T scores per (sub) scale	Dutch	English, French, German	10 min	License required for commercial studies
KIDSCREEN-27 (Ravens-Sieberer et al. [50])	HRQoL	Healthy and chronically ill children/ adolescents (8–18 years old)	Parent-/patient-reported	1 week	5 (27)	Varies per question	T scores and percentiles per (sub)scale	English	>40 translations	10–15 min	License required for commercial studies
CBCL/1.5–5 (Achenbach et al. [52])	Competencies, adaptive functioning, and problems	Children aged 18 months to 5 years	Parent-/proxy-reported	2 months	12 (103)	0 = not true (as far as you know), 1 = somewhat or sometimes true, 2 = very true or often true	T scores per (sub) scale	English	>50 languages (full list at aseba.org)	10–15 min	Fees per form apply
CBCL/4–18 (Achenbach et al. [52])	Competencies, adaptive functioning, and problems	Children aged 4–18 years	Parent-/proxy-reported	6 months	8 (122)	0 = not true (as far as you know), 1 = somewhat or sometimes true, 2 = Very true or often true	T scores per (sub) scale	English	>50 languages (full list at aseba.org)	10–20 min	Fees per form apply

CBCL, Child Behavior Checklist; contin., continuous scale; DLQI, Dermatology Life Quality Index; HRQoL, health-related quality of life; IH, infantile hemangioma; NA, not applicable; PDL, pulsed dye laser; OMI, outcome measurement instrument; PSQ, Perceived Stigmatization Questionnaire; PWS, port wine stain; TAPQOL, TNO-AZL questionnaire for preschool children's health-related quality of life. [1] Only for children older than 18 months.

(percentage) improvement in color, thickness, and size, and the appearance of (new) scarring, atrophy, and hypopigmentation.

The study of Currie and Monk [34] evaluated 3 previously employed clinical OMIs. In the first OMI, clinical results were categorized into poor (almost no change), fair (partly cleared), good (much improved), and excellent (essentially gone, barely discernible). Note that many trials have used the same categories with different definitions [19].

The second OMI in Currie's study assessed percentage lightening. A variety of percentage improvement scales are commonly used in PWS trials and given distinct designations (percentage "lightening," "improvement," "success," etc.) and reported as continuous variables or ranges (usually 0–24, 25–49, 50–74, and 75–100%). As these constitute highly intuitive measures, no published development paper exists. Their measurement properties have also been investigated in the studies of Pérez (percentage clearance; both continuous and ordinal scale) and Szychta (percentage success; continuous scale) [35, 36]. Although it could be argued that the construct "color improvement," "lightening," or "blanching" is much narrower than "improvement" or "success," in this review these were considered equivalent and the data were pooled depending on the scales.

The third OMI assessed by Currie and Monk [34] was originally described by Achauer et al. [46]. Treated PWS were categorized into class 0–IV and then also converted to poor, fair, good, or excellent (Table 2). Note that originally, Achauer et al. also included class V ("all of class IV plus nodularity") [46].

Naran et al. [38] used a very simple 3-point rating scale ("lighter," "darker," or "did not change") to assess additional lightening after additional PDL treatments in children.

Summary of Included Life Impact PROMs

Three instruments for health-related quality of life (HRQoL), 2 for detecting emotional and behavioral problems and 1 for perceived stigmatization, were included (Table 2). None of these PROMs has yet been used as an OMI in published PWS intervention studies. The average Flesch-Kincaid grade level for readability of the English translations was 3.5 (online suppl. Table 3). These and other characteristics of OMI feasibility and interpretability are listed in online supplementary Tables 3 and 4.

The Dermatology Life Quality Index (DLQI) was the first dermatology-specific HRQoL questionnaire [47]. It consists of 10 items (symptoms and feelings, daily activities, leisure, work, and school, personal relationships, and treatment) and has been employed for a range of dermatoses. In the included study using the DLQI's adult version, PWS patients had a small to moderate HRQoL impairment (online suppl. Table 4). A study that assessed the DLQI's content validity in adults with psoriasis was also included because its findings regarding comprehensibility are likely to apply to PWS patients as well [45].

The 21-item Perceived Stigmatization Questionnaire (PSQ) ascertains how often people are confronted by certain stigmatizing behavior [48]. Although it was designed for people with visible distinctions, its further development was limited mostly to adult burn survivors. The patient- or parent-reported instrument yields a total score and subscores (range 1–5) for "absence of friendly behavior," "confused/staring behavior," and "hostile behavior." In this study, preschool and school-age children with a facial difference (including PWS) had a PSQ score of 1.66 and 2.10, respectively (compared to, e.g., 2.2 in adult burn survivors [48]).

The TNO-AZL Questionnaire for Preschool Children's Health-Related Quality of Life (TAPQOL) measures parent's perception of HRQoL in children aged 9 months to 6 years and incorporates the (perceived) emotional reaction of the child to their health status problems [49]. This multidimensional questionnaire is presented primarily as an instrument for research and group-level data. In this study of children with a facial difference, TAPQOL scores were not impaired (online suppl. Table 4).

The KIDSCREEN-27 is a generic, parent- or self-reported HRQoL measure for children and adolescents (8–18 years old) that covers 5 domains (physical well-being, psychological well-being, autonomy and parents, peers and social support, and school environment) [50]. It is the short version of the KIDSCREEN-52 [51]. Both KIDSCREEN questionnaires were developed in 12 European countries to measure HRQoL in largely healthy children. The included study in children with a facial difference found statistically significant impairment of child- and parent-reported psychological well-being and parent-reported overall HRQoL and physical well-being.

The Child Behavior Checklists (CBCL) gauge parent-reported emotional and behavioral problems, competencies, and adaptive functioning in children and adolescents [52]. Throughout its existence, many revisions have been made. The CBCL/1.5–5 (the successor to the CBCL/2–3; for 18 months to 5 years old) and the CBCL/4–18 (the predecessor of the current CBCL/6–18; for 4–18 years old) were included here. Both offer 3 composite scales (internalizing, externalizing, and total problems) and differential syndrome- and DSM-oriented scales. Neither the included study (online suppl. Table 4) nor others have observed increased prevalence of psychopathology in children with PWS using the CBCL [53, 54].

OMI Development and Content Validity

The quality of OMI development is shown in Table 3. OMIs that have been presented without any information on their development were omitted. Overall scores were poor. Neither of the 2 clinical OMIs had performed adequate concept elicitation to identify relevant items (most importantly, it was unclear if and how professionals or patients were involved). The questionnaire of Koster had been pilot-tested but it was unclear whether problems regarding the comprehensibility of the instructions, items, response options, and recall period were properly addressed. Also, physicians were not asked about its comprehensiveness. The questionnaire of Sajan did not report any pilot testing. The PSQ's target population was people with physical distinction, yet its development was limited to adult burn survivors. None of the PROMs performed proper pilot testing for comprehensibility or any form of testing for comprehensiveness.

No (additional) content validity studies performed in PWS patients were found. As a result, the overall ratings for content validity (and its 3 subcomponents: relevance, comprehensiveness, and comprehensibility) depended highly on reviewer ratings (Table 4) with low levels of evidence. The lack of supportive data on what constitute relevant items for clinical assessment of PWS and measures of functioning for PWS patients negatively impacted relevance scores. For Koster's questionnaire, this resulted in an inconsistent score for relevance. Also, the answering options for the items "size" and "hypertrophy" were considered imprecise. Comprehensiveness and comprehensibility were rated as sufficient. The relevance of Sajan's questionnaire was deemed inadequate because question 4 (new atrophy) was considered irrelevant for PWS. As a result, the criterion for relevance ("reviewers consider ≥85% of items relevant for the population of interest") was not met. Comprehensiveness was considered insufficient because, for example, hyperpigmentation was not included. No paper has published on the development or content validity of the other clinical OMIs [34, 38, 46].

The DLQI and PSQ had sufficient ratings for all 3 subcomponents. The relevance was inconsistent for the TAPQOL and KIDSCREEN-27 (too many items were considered insufficiently relevant for PWS patients) and

Table 3. Quality of outcome measurement instrument development

OMI	OMI design					Concept elicitation[1]	Total OMI design	CI study[2] or other pilot testing				Total OMI development
	general design requirements							general design requirements	compreh-ensibility	comprehen-siveness	total CI study	
	clear construct	clear origin of construct	clear target population for which the OMI was developed	clear context of use	OMI developed in sample representing the target population			CI study performed in sample representing the target population				
Koster's PWS questionnaire	V	V	V	V	V	I	I	V	I	D	I	I
Sajan's PWS and IH questionnaire	V	D	V	V	V	I	I				I	I
DLQI	V	D	V	V	V	D	D	A	I	D	I	I
PSQ	V	V	V	V	I		I	I	I	D	I	I
TAPQOL	V	V	I	V	V	D	I	I	I	I	I	I
KIDSCREEN-27	V	V	V	V	V	A	A	A	D	D	D	D
CBCL/1.5-5	V	V	V	V	V	D	D	A	D	D	D	D
CBCL/4-18	V	V	V	V	A	D	D	A	D	D	D	D

Development was rated as very good (V), adequate (A), doubtful (D), or inadequate (I). CBCL, Child Behavior Checklist; DLQI, Dermatology Life Quality Index; IH, infantile hemangioma; OMI, outcome measurement instrument; PSQ, Perceived Stigmatization Questionnaire; PWS, port wine stain; TAPQOL, TNO-AZL questionnaire for preschool children's health-related quality of life. [1] When the patient-/parent-reported outcome measure was not developed in a sample representing the target population, concept elicitation was not further rated. [2] Empty cells indicate that a cognitive interview (CI) study (or part of it) was not performed.

Table 4. Ratings and quality of evidence for content validity

OMI	Relevance		Comprehensiveness		Comprehensibility		Overall content validity	
	rating	quality of evidence	rating	quality of evidence	rating	quality of evidence	rating	quality of evidence
Koster's PWS questionnaire	±	very low	+	very low	+	very low	+	very low
Sajan's PWS and IH questionnaire	–	very low	–	very low	+	very low	±	very low
DLQI	+	low	+	very low	+	moderate	+	very low
PSQ	+	very low	+	very low	+	very low	+	very low
TAPQOL	±	very low	+	very low	+	very low		
KIDSCREEN-27	±	moderate	+	low	+	low		
CBCL/1.5–5	?	very low	+	very low	+	very low		
CBCL/4–18	?	low	+	very low	+	very low		

The content validity was rated as sufficient (+), insufficient (–), inconsistent (±), or indeterminate (?). Outcome measures with very different results for relevance, comprehensiveness, and comprehensibility do not have an overall content validity rating. CBCL, Child Behavior Checklist; DLQI, Dermatology Life Quality Index; IH, infantile hemangioma; OMI, outcome measurement instrument; PSQ, Perceived Stigmatization Questionnaire; PWS, port wine stain; TAPQOL, TNO-AZL questionnaire for preschool children's health-related quality of life.

indeterminate for the CBCL questionnaires. All PROMs had sufficient comprehensiveness and comprehensibility.

Measurement Properties of PWS OMIs

The quality and results of the studies on measurement properties are summarized in Table 5. Most studies were of doubtful or inadequate quality.

Structural Validity and Internal Consistency

Wang et al. [40] have performed principal component analysis (a form of exploratory factor analysis) of the Chinese translation of the DLQI in patients with exposed PWS and identified 2 common factors (Table 5). This study was found to be of doubtful quality because no factor rotation method was applied. No criteria for good measurement properties have been defined by COSMIN for exploratory factor analysis, resulting in an indeterminate score (and low-grade quality of evidence score; Tables 5, 6). Because of the lack of evidence for sufficient structural validity and calculation of only an overall internal consistency despite suggesting a two-dimensional model (internal consistency can only be interpreted within a unidimensional scale), internal consistency of the DLQI was also scored as indeterminate (very low quality of evidence). Of note, studies in other diseases have yielded inconsistent results regarding unidimensionality of the DLQI [55–57].

Masnari et al. [41, 42] evaluated internal consistency of several PROMs in children with various facial blemishes. Note that internal consistency has no meaningful interpretation for questionnaires with a formative model. Unfortunately, not all OMIs have studied or reported their putative model (reflective vs. formative). Classifying questionnaires correctly can be difficult, and mixed models also exist. We assigned a formative model to the PSQ and CBCL, and a reflective or mixed model to the other PROMs. Even though most (sub-)scales had sufficient internal consistency, there was a lack of information on the structural validity in PWS patients resulting in an indeterminate score. The quality of evidence for these PROMs was downgraded to low or very low due to small sample sizes (imprecision) and indirectness (only 22% of patients had PWS).

Reliability and Measurement Error

The inter-rater reliability of Koster's questionnaire was evaluated for a panel of 5 in children and (young) adults (Table 1). Insufficient reliability was found for all 8 items, particularly pigmentation and surface-structure. This can be explained by the low frequency of hypopigmentation and hyperpigmentation and uneven surface in the study population, as supported by the relatively high percentage of absolute agreement (average of 5 raters): 88 and 79%, respectively (Table 5). The Cronbach α-values on averaged ratings of the 5 panel members suggest that application of this method improves reliability, although

Table 5. Results and quality of studies on measurement properties

OMI	Language in which the OMI was evaluated	Structural validity			Internal consistency			Reliability (inter-rater)			Measurement error (inter-rater)			Measurement error (intrarater)			Construct validity		
		n	meth. qual.	result (rating)	*n*	meth. qual.	result (rating)	*n*	meth. qual.	result (rating)	*n*	meth. qual.	result (rating)	*n*	meth. qual.	result (rating)	*n*	meth. qual.	result (rating)
0–24; 25–49; 50–74 and 75–100% lightening (Currie and Monk [34])	English													20	A	Mean agreement[1] 52.8% (range 50–62) (?)			
0–24; 25–49; 50–74 and 75–100% clearance (Pérez et al. [36])	NL							80	A	κ_w = 0.77 (+)									
Percentage success (core physicians) (Szychta et al. [35])	English							48	I	*R* = 0.81 (?)	48	I	TEM = 9.91% (?)						
Percentage success (lay people) (Szychta et al. [35])	English							48	I	*R* = 0.63 (?)	48	I	TEM = 13.84% (?)						
0–100% clearance (Pérez et al. [36])	NL							80	A	ICC = 0.83 (+)									
0; 1–24; 25–49; 50–75; 76–99, and 100% color improvement (Sajan et al. [37])	English							22	I	κ_w = 0.92 (+)									
Summary result for all physician-reported percentage-improvement outcomes (overall rating)								**182**	**A/I**	**0.77–0.92 (+)2**	**48**	**I**	**TEM = 9.91% (?)**	**20**	**A**	**Mean agreement 52.8% (range 50–62) (?)**			
Koster's PWS questionnaire (Koster et al. [39])	English							70	A	Mean κ_w = 0.37 (range 0.11–0.64) (–)	70	A	Mean % exact agreement: 47 (color), 50 (patchiness), 64 (boundary), 88 (pigmentation), 66 (size), 52 (shape), 79 (surface structure), 71 (hypertrophy). Mean 64.6% (range 47–88) (?)						
Sajan's PWS and IH questionnaire (Sajan et al. [37])	English							24^3	I	ICC = 0.92 for color, thickness, size, and scarring (+), ICC = 0.70 for atrophy (+), ICC = 0.10 for hypopigmentation (–)									
Poor/fair/good/excellent (Currie and Monk [34])	English													20	A	Mean agreement[1] 77.1% (range 48.8–96.5) (?)			
Class 0–IV (Currie and Monk [34])	English													20	A	Mean agreement[1] 64.2% (range 34.7–82.6) (?)			

Table 5 (continued)

OMI	Language in which the OMI was evaluated	Structural validity			Internal consistency			Reliability (inter-rater)			Measurement error (inter-rater)			Measurement error (intrarater)			Construct validity		
		n	meth. qual.	result (rating)	*n*	meth. qual.	result (rating)	*n*	meth. qual.	result (rating)	*n*	meth. qual.	result (rating)	*n*	meth. qual.	result (rating)	*n*	meth. qual.	result (rating)
Lighter/darker/ did not change (Naran et al. [38])	English							21	I	Spearman correlations between 3 observers 0.54 ($p < 0.001$), 0.41 ($p < 0.01$), and 0.34 ($p = 0.02$) (?)							21	I	Spearman correlation between 3 observers and digital image analysis −0.33 to −0.09 ($p > 0.01$) (–)
DLQI (Wang et al. [40])	Chinese	197	D	CFA: 2 common factors with 47.5% cumulative explained variance (?)	197	D	Cronbach α = 0.740 (overall) (?)												
PSQ; patient-reported (Masnari et al. [42])	German				34	D	Cronbach α = 0.81 (?)												
PSQ; parent-reported (Masnari et al. [42])	German				86	V	Cronbach α = 0.88 (?)												
TAPQOL (Masnari et al. [42])	German				54	V	Cronbach α-values = 0.51–1.00 (mean 0.72) (?)												
KIDSCREEN-27; parent-reported (Masnari et al. [42])	German				32	V	Cronbach α-values = 0.49–0.92 (mean 78.7) (?)												
KIDSCREEN-27; child-reported (Masnari et al. [42])	German				31	V	Cronbach α-values = 0.44–0.86 (mean 0.68) (?)												
CBCL/1.5–5 (Masnari et al. [42])	German				21	V	Cronbach α-values = 0.73 (internalizing), 0.90 (externalizing), 0.92 (total) (?)												
CBCL/4–18 (Masnari et al. [42])	German				51	V	Cronbach α-values = 0.87 (internalizing), 0.93 (externalizing), 0.95 (total) (?)												

Study quality was very good (V), adequate (A), doubtful (D), or inadequate (I). Measurement properties were rated as sufficient (+), insufficient (–), inconsistent (±), or indeterminate (?). Results were pooled if possible. CBCL, Child Behavior Checklist; CFA, common factor analysis; DLQI, Dermatology Life Quality Index; ICC, intraclass correlation coefficient; IH, infantile hemangioma; $κ_w$, weighted kappa; meth. qual., methodological quality; NL, not listed; OMI, outcome measurement instrument; PSQ, Perceived Stigmatization Questionnaire; PWS, port wine stain; SWS, Sturge-Weber syndrome; TAPQOL, TNO-AZL questionnaire for preschool children's health-related quality of life; TEM, technical error of measurement. [1] Mean agreement for all outcome categories for all 6 assessors. [2] The results of Szychta et al. for reliability could not be pooled because it was reported using an incompatible parameter. [3] PWS ($n = 5$, of which 1 with SWS) and IH ($n = 19$) patients.

Table 6. Overall quality of evidence per measurement property

Outcome measurement instrument:	Koster's PWS questionnaire		Sajan's PWS and IH questionnaire		Lighter/darker/ did not change		Poor/fair/good/ excellent		0–24, 25–49, 50–74, and 75–100% lightening/clearance		0–100% clearance/ success		All percent color improvement or clearance-based outcomes	
Ref(s):	Koster et al. [39]		Sajan et al. [37]		Naran et al. [38]		Currie and Monk [34]		Currie and Monk [34], Pérez et al. [36]		Pérez et al. [36], Szychta et al. [35]		Pérez et al. [36], Sajan et al. [37], Currie and Monk [34], Szychta et al. [35]	
	overall rating	quality of evidence	overall rating	quality of evidence	overall rating	quality of evidence	overall rating	quality of evidence	overall rating	quality of evidence	overall rating	quality of evidence	overall rating	quality of evidence
Content validity	±	very low	±	very low	±	very low								
Relevance	+	very low	–	very low	+	very low								
Comprehensiveness	+	very low	–	very low	–	very low								
Comprehensibility	+	very low	+	very low	±	very low								
Structural validity														
Internal consistency														
Reliability (inter-rater)	–	low	+[1]	very low	?	very low			+	low	+	low	+	moderate
Measurement error (intrarater)							?	very low	?	very low			?	very low
Measurement error (inter-rater)	?	low									?	very low	?	very low
Construct validity					–	very low								

Outcome measurement instrument:	Class 0–IV		DLQI (Chinese version)		TAPQOL (German version)		KIDSCREEN-27 (parent- and self-reported; German version)		PSQ (parent- and self-reported; German version)		CBCL/1.5-5 (German version)		CBCL/4-18 (German version)	
Ref(s):	Currie and Monk [34]		Wang et al. [40], Safikhani et al. [45]		Masnari et al. [42]		Masnari et al. [42]		Masnari et al. [42]		Masnari et al. [42]		Masnari et al. [42]	
	overall rating	quality of evidence	overall rating	quality of evidence	overall rating	quality of evidence	overall rating	quality of evidence	overall rating	quality of evidence	overall rating	quality of evidence	overall rating	quality of evidence
Content validity			+	very low					+	very low				
Relevance			+	low	±	very low	±	moderate	+	very low	?	very low	?	low
Comprehensiveness			+	very low	+	very low	+	low	+	very low	+	very low	+	very low
Comprehensibility			+	moderate	+	very low	+	low	+	very low	+	very low	+	very low
Structural validity			?	low										
Internal consistency			?	very low	?	low	?	very low	?	very low	?	very low	?	low
Reliability (inter-rater)														
Measurement error (intrarater)	?	very low												
Measurement error (inter-rater)														
Construct validity														

Measurement properties were rated sufficient (+), insufficient (–), inconsistent (±), or indeterminate (?). The level of evidence was graded according to the modified GRADE approach. CBCL, Child Behavior Checklist; DLQI, Dermatology Life Quality Index; IH, infantile hemangioma; PSQ, Perceived Stigmatization Questionnaire; PWS, port wine stain; Ref(s)., reference(s); TAPQOL, TNO-AZL questionnaire for preschool children's health-related quality of life. [1] Sufficient for all items except "hypopigmentation."

this would have to be properly evaluated in a separate study. Agreement for the other items varied between 47 and 71% (Table 5). Unfortunately, no criteria for good measurement error have been defined by COSMIN for categorical parameters (i.e., percentage agreement) [15], and no alternative criteria were found in the literature resulting in an indeterminate score.

The questionnaire of Sajan was used in PDL-treated children with PWS and infantile hemangiomas. The inter-rater reliability was sufficient for all items except hypopigmentation. Because of the small number of (PWS) patients, this study provided a very low quality of evidence.

Currie and Monk [34] intended to assess intrarater reliability of 3 common PWS OMIs after 1 month but calculated mean agreement ("concordance"), which is a measure of measurement error (see above). Because the exact distribution of scores is unknown, a post hoc calculation of reliability was impossible. Although no criteria for this parameter of measurement error have been defined, the 3 OMIs can be compared to each other. OMI 1 (poor, fair, good, or excellent) had the best overall results (67–89% mean agreement per outcome category; Table 5). OMI 2 (percentage lightening) performed less but was the most consistent across all 4 categories (50–62%). OMI 3 (class 0–IV) performed relatively well (52–76%) for "excellent," "good," and "poor" outcomes but exceptionally poorly (17%) in "fair" outcomes. For all 3 OMIs concordance was highest for the best and worse outcome and lowest in the intermediate categories.

Szychta et al. [35] assessed inter-rater reliability and measurement error of "percentage success" after PDL treatment for both core physicians as well as lay people. For this, they used parameters typically used in anthropometry for which also no COSMIN criteria for good measurement properties have been defined (TEM, technical error of measurement, and *R*, coefficient of reliability). Because TEM was calculated instead of the standard error of measurement (SEM; the gold standard), this study received an inadequate score. Anthropometric literature suggests an $R \geq 0.95$ as sufficient [58], which would yield an insufficient score for the current results for both core physicians and lay people (Table 5). The inter-rater TEM values (9.9% for expert physicians) appear acceptable but, because TEM is not the same as SEM, it is difficult to interpret the results [59]. In addition, no information is available on the minimal important change in PWS patients.

Pérez et al. [36] investigated inter-rater reliability of a 0–100% clearance scale among specialized dermatologists (Table 1). The continuous score was also converted to 0–24, 25–49, 50–74, and 75–100% clearance. The study was of only adequate quality because it could only be assumed that test conditions were similar for all 3 assessors. Both the categorized and continuous scores had sufficient reliability (Table 5). Overall, there was low quality of evidence due to the limited sample size and quality of the study.

If the percentage improvement scales are regarded as equivalent and the results of Sajan's (percentage color improvement) and Pérez' (percentage clearance) studies are pooled (even though Sajan uses 2 extra categories for 0 and 100% color improvement), there is moderate evidence for sufficient reliability (Tables 5, 6). Because of the differences in scales, no meta-analysis was performed.

Naran et al. [38] used correlation analysis to assess the inter-rater reliability of its OMI. However, correlations are not appropriate for this purpose, particularly because the OMI is an ordinal score.

Construct Validity

The OMI of Naran et al. was compared to ΔE (i.e., the change in color difference in comparison to normal skin) derived from digital image analysis. Because this study lacked a predefined hypothesis, it was considered to be of inadequate quality. Spearman correlations between the 3 observers and digital image analysis varied from −0.33 to −0.09 ($p > 0.01$), which is much smaller than what would be expected (Tables 5, 6).

COSMIN Recommendations

In line with COSMIN guidelines, instrument recommendations are provided below based on 2 conclusions of this review: (1) no OMI with evidence for sufficient content validity (any level) and at least low-quality evidence for sufficient internal consistency was included (this OMI would be recommended for use and produce results that can be trusted), and (2) no OMI with high-quality evidence for an insufficient measurement property was found (this OMI would not be recommended for use). Accordingly, all included instruments have the potential to be included as an OMI for PWS treatment but would require further assessment of quality. Based on content validity, the PSQ and DLQI are provisionally recommended. No clinical OMI can be recommended.

Discussion

This study was the first to systematically review the measurement properties of PWS OMIs. We have identified 6 physician-reported clinical OMIs and 6 PROMs of life impact. Each has undergone very limited evaluation in PWS patients. Although not all elements of the COSMIN are applicable for all OMIs, each OMI has important data missing. No study addressed responsiveness, minimal important change, or cross-cultural validity. The quality of the development (if any) of PWS-specific OMIs was inadequate because substandard methods were used to generate relevant items and assess the comprehensibility and comprehensiveness of the OMI. Most importantly, no study has asked PWS patients (or their parents) what matters to them, nor have professionals been adequately interviewed about relevant items. Together with the small number of PWS patients this generally resulted in low levels of evidence (quality was very low and low for 74 and 21% of all evidence, respectively). No study has assessed the content validity of OMIs originally developed for other populations and now used in PWS patients.

Considering the lack of high-quality evidence, it was impossible to recommend a clinical OMI. Koster's questionnaire is clearly the most elaborate clinical OMI. This approach is justified at least in part by another study of Koster et al. [60], in which Koster's questionnaire and a disfigurement score were used to study the disfiguring effect of PWS on the head and neck. According to physicians, PWS size, color, and boundary contributed the most to disfigurement, underscoring their relevance in clinical OMIs. The inter-rater reliability of Koster's questionnaire was insufficient. However, in evaluative studies measurement error is much more important than reliability [61]. This is based on the fact that reliability refers to how well patients can be discriminated, that is, the ability of an instrument to distinguish patients from each other despite measurement error [62]. As a result, reliability could be low because variability in the study population is low. The measurement error (percentage agreement) was <70% for a majority of items, indicating substantial measurement error. The measurement properties of Koster's questionnaire have not been investigated further since, nor has the instrument been employed in a published trial. The relevance of Sajan's questionnaire [37] suffered from the fact that it was also designed for infantile hemangiomas and included a question on new atrophy, which was considered irrelevant for PWS. Consequently, it did not meet the COSMIN criterion (≥85% relevant items). Perhaps this criterion is too strict for such brief questionnaires. Nevertheless, this questionnaire could be easily modified and retested. This instrument has not yet been used in a published PWS trial either. The OMI of Naran et al. [38] was considered to be far too limited.

The most common OMI in recent PWS trials is a percentage improvement or lightening, also referred to as blanching, clearance, or other similar terms [14]. It is unclear whether these terms are actually interchangeable. For example, a very lightened but patchy lesion may score better for "lightening" than for "success." Unfortunately, the content validity of these methods has not been investigated. When all studies are taken together there is moderate evidence for sufficient inter-rater reliability (when performed by experienced physicians). Note that these data are derived almost exclusively from patients with a PWS in the face or neck and may not apply to lesions elsewhere. When the percentages are categorized it is important to be aware of floor and ceiling effects (online suppl. Table 4), as these may obscure treatment superiority. The classification of Pérez [36] suffers from such a ceiling effect, which could be prevented by adding another top category (as in Sajan's questionnaire [37]). Unfortunately, the data on measurement error could not be compared to COSMIN criteria but the results do indicate that experienced physicians perform better than lay people.

In order to achieve treatment results that most closely align with what means the most for patients, it is important to measure corresponding outcomes, that is, PROMs. The emphasis on and inclusion of such patient-reported data has increased greatly, also in dermatology [63, 64], and helps to provide patient-centered care and guide clinical decision making. In terms of PROMs for PWS patients, many studies have demonstrated a significant psychosocial impact of PWS, stigmatization, and reduced HRQoL [6, 7, 43, 44, 53, 65]. However, the use of PROMs in PWS trials is limited [14]. In this study no clinimetric studies of PWS-specific PROMs or a (generic) PROM used as an outcome for a PWS intervention trial were found. Because PROMs not yet used in effectiveness studies were also eligible, we included 6 PROMs. The assessment of these PROMs was very restricted (limited to structural validity and/or internal consistency). Additionally, the information on structural validity was of poor quality or absent, which hampered the interpretability of the internal consistency. The sufficient overall content validity of the DLQI and PSQ do justify a provisional recommendation. The relevance of the TAPQOL, KIDSCREEN-27, and CBCL questionnaires was less certain, primarily because too many questionnaire items were deemed irrelevant for PWS patients as be supported by the lack of aberrant TAPQOL and CBCL scores (Table 1 and online suppl. Table 4).

This review was limited in several ways. First, 3 studies of 3 PROMs could not be included because the questionnaires themselves were not accessible. However, these studies were of low quality and would have provided little evidence (data not shown). Second, some OMIs were developed prior to the introduction of current methodological standards. Because of the COSMIN's worst score counts principle, this negatively impacted OMI development and content validity scores for older OMIs. Third, even though the COSMIN methodology specifies many criteria and possibilities, many questions (particularly for content validity) still require some form of subjective assessment. Other assessors may thus arrive at slightly different results. It should also be emphasized that this review is not a comprehensive index of all OMIs used in PWS studies, as OMIs without any assessment of measurement properties are not covered in this methodology.

Finally, it is important that we address the lack of evidence for the measurement properties of PWS OMIs because this hinders outcome selection, especially since PWS OMIs have thus far been highly heterogeneous [14]. Ideally, a PWS core outcome set is developed. This would drastically improve study quality and comparability across trials and enable interstudy comparison, which is direly needed considering the paucity of high-quality PWS trials [14]. First, consensus on the core domains (constructs) measured in therapeutic trials would need to be reached in a structured Delphi process that includes all stakeholders. Subsequently, new OMIs need to be developed, or existing OMIs could be repurposed. Finally, further assessment of all relevant measurement properties needs to be performed.

Conclusions

There was insufficient evidence to recommend a single clinical OMI or PROM for treatment of PWS. More research into the measurement properties of clinical OMIs and PROMs is needed, preferentially guided by the establishment of a set of core domains.

Key Message

The current literature provides insufficient evidence to guide outcome measurement instrument selection for port wine stains.

Acknowledgments

We are grateful to the COSMIN team for their assistance in performing this review.

Conflict of Interest Statement

C.M.A.M.H. co-authored one of the included papers (Koster's questionnaire) and was therefore excluded from review decisions regarding this paper. The authors have no other conflicts of interest to declare.

Funding Sources

This project was supported by a Stichting Technologische Wetenschap (STW) valorization grant (12064) and a preseed grant from the Academic Medical Center SKE Fund (Technostarter 20090812) to M.H. M.H. is currently supported by grants from the Dutch Cancer Foundation (KWF project 10666), National Natural Science Foundation of China (81872220), a Zhejiang Provincial Foreign Expert Program Grant, Zhejiang Provincial Key Natural Science Foundation of China (Z20H160031), and a grant for the establishment of the Jiaxing Key Laboratory for Photonanomedicine and Experimental Therapeutics. B.D. is supported by a basic public welfare research project of Zhejiang Province (LGF18H160034). Site-specific pharmaco-laser therapy, which is an experimental treatment modality for PWS and a core research activity of M.H. that thematically inspired this COSMIN study, was patented (M. Heger. Drug delivery system for use in the treatment of vascular and vessel-related pathologies; EP09151332.5, PCT/EP2010/050833). M.H. is cofounder of Nurish.Me and has equity in that company (whose business activities are unrelated to the present work).

Author Contributions

M.I.R., S.C., A.W., C.M.A.M.H., M.H., and J.L. contributed to the study design. J.L. performed all searches. M.I.R. and S.C. conducted qualitative analysis. Reviewer decisions were discussed by M.I.R., S.C., C.M.A.M.H., and A.W. All authors participated in the drafting and writing of the manuscript, contributed to the interpretation and critical revision of the manuscript, and gave final approval of this paper.

References

1 Tan W, Wang J, Zhou F, Gao L, Yin R, Liu H, et al. Coexistence of Eph receptor B1 and ephrin B2 in port-wine stain endothelial progenitor cells contributes to clinicopathological vasculature dilatation. Br J Dermatol. 2017 Dec;177(6):1601–11.
2 Nguyen V, Hochman M, Mihm MC Jr, Nelson JS, Tan W. The Pathogenesis of Port Wine Stain and Sturge Weber Syndrome: Complex Interactions between Genetic Alterations and Aberrant MAPK and PI3K Activation. Int J Mol Sci. 2019 May;20(9):2243.
3 Jacobs AH, Walton RG. The incidence of birthmarks in the neonate. Pediatrics. 1976 Aug;58(2):218–22.
4 Pratt AG. Birthmarks in infants. AMA Arch Derm Syphilol. 1953 Mar;67(3):302–5.

5 van Drooge AM, Beek JF, van der Veen JP, van der Horst CM, Wolkerstorfer A. Hypertrophy in port-wine stains: prevalence and patient characteristics in a large patient cohort. J Am Acad Dermatol. 2012 Dec;67(6): 1214–9.

6 Augustin M, Zschocke I, Wiek K, Peschen M, Vanscheidt W. Psychosocial stress of patients with port wine stains and expectations of dye laser treatment. Dermatology. 1998;197(4): 353–60.

7 Hagen SL, Grey KR, Korta DZ, Kelly KM. Quality of life in adults with facial port-wine stains. J Am Acad Dermatol. 2017 Apr;76(4): 695–702.

8 van Raath MI, Chohan S, Wolkerstorfer A, van der Horst CM, Storm G, Heger M. Port wine stain treatment outcomes have not improved over the past three decades. J Eur Acad Dermatol Venereol. 2019 Jul;33(7):1369–77.

9 Al-Janabi MH, Ismaeel Ali NT, Mohamed Al-Sabti KD, Al-Dhalimi MA, Abdul Wahid SN. A new imaging technique for assessment of the effectiveness of long pulse Nd:YAG 532 nm laser in treatment of facial port wine stain. J Cosmet Laser Ther. 2017 Nov;19(7):418–21.

10 Jung B, Kim CS, Choi B, Kelly KM, Nelson JS. Use of erythema index imaging for systematic analysis of port wine stain skin response to laser therapy. Lasers Surg Med. 2005 Sep; 37(3):186–91.

11 Lister T, Wright P, Chappell P. Spectrophotometers for the clinical assessment of port-wine stain skin lesions: a review. Lasers Med Sci. 2010 May;25(3):449–57.

12 Choi B, Tan W, Jia W, White SM, Moy WJ, Yang BY, et al. The Role of Laser Speckle Imaging in Port-Wine Stain Research: Recent Advances and Opportunities. IEEE J Sel Top Quantum Electron. 2016 May-Jun;2016(3): 1–12.

13 van Raath MI, Bambach CA, Dijksman LM, Wolkerstorfer A, Heger M. Prospective analysis of the port-wine stain patient population in the Netherlands in light of novel treatment modalities. J Cosmet Laser Ther. 2018 Apr; 20(2):77–84.

14 van Raath MI, Chohan S, Wolkerstorfer A, van der Horst CM, Limpens J, Huang X, et al. Clinical outcome measures and scoring systems used in prospective studies of port wine stains: A systematic review. PLoS One. 2020 Jul;15(7):e0235657.

15 Prinsen CA, Mokkink LB, Bouter LM, Alonso J, Patrick DL, de Vet HC, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. Qual Life Res. 2018 May;27(5):1147–57.

16 van Raath MI, Chohan S, Wolkerstorfer A, van der Horst C. Outcome measures for port wine stains: a systematic review of their measurement properties. PROSPERO 2019 CRD42019119252. Available from: https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42019119252.

17 Terwee CB, Jansma EP, Riphagen II, de Vet HC. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. Qual Life Res. 2009 Oct;18(8): 1115–23.

18 Dodd S, Clarke M, Becker L, Mavergames C, Fish R, Williamson PR. A taxonomy has been developed for outcomes in medical research to help improve knowledge discovery. J Clin Epidemiol. 2018 Apr;96:84–92.

19 Quaba AA. Results of argon laser treatment of port-wine stains: a method of assessment. Br J Plast Surg. 1989 Mar;42(2):125–32.

20 Ginsbach G. A Tool for the Evaluation of Colour in Port Wine Stains. Lasers Med Sci. 1991;6(1):49–52.

21 Rah DK, Kim SC, Lee KH, Park BY, Kim DW. Objective evaluation of treatment effects on port-wine stains using L*a*b* color coordinates. Plast Reconstr Surg. 2001 Sep;108(4): 842–7.

22 Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan-a web and mobile app for systematic reviews. Syst Rev. 2016 Dec; 5(1):210.

23 Mokkink LB, de Vet HC, Prinsen CA, Patrick DL, Alonso J, Bouter LM, et al. COSMIN Risk of Bias checklist for systematic reviews of Patient-Reported Outcome Measures. Qual Life Res. 2018 May;27(5):1171–9.

24 Terwee CB, Prinsen CA, Chiarotto A, Westerman MJ, Patrick DL, Alonso J, et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. Qual Life Res. 2018 May;27(5): 1159–70.

25 Terwee CB, Prinsen CA, Chiarotto A, De Vet HC, Bouter LM, Alonso J, et al. COSMIN methodology for assessing the content validity of PROMs: User manual version 1.0. 2018. pp. 1–72.

26 Mokkink LB, Terwee CB, Patrick DL, Alonso J, Stratford PW, Knol 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 Jul;63(7):737–45.

27 Dobson F, Choi YM, Hall M, Hinman RS. Clinimetric properties of observer-assessed impairment tests used to evaluate hip and groin impairments: a systematic review. Arthritis Care Res (Hoboken). 2012 Oct;64(10): 1565–75.

28 Balemans AC, Fragala-Pinkham MA, Lennon N, Thorpe D, Boyd RN, O'Neil ME, et al. Systematic review of the clinimetric properties of laboratory- and field-based aerobic and anaerobic fitness measures in children with cerebral palsy. Arch Phys Med Rehabil. 2013 Feb;94(2):287–301.

29 Vrijman C, Linthorst Homan MW, Limpens J, van der Veen W, Wolkerstorfer A, Terwee CB, et al. Measurement properties of outcome measures for vitiligo. A systematic review. Arch Dermatol. 2012 Nov;148(11):1302–9.

30 Fleuren BP, van Amelsvoort LG, Zijlstra FR, de Grip A, Kant I. Handling the reflective-formative measurement conundrum: a practical illustration based on sustainable employability. J Clin Epidemiol. 2018 Nov;103:71–81.

31 Prinsen CA, Vohra S, Rose MR, Boers M, Tugwell P, Clarke M, et al. How to select outcome measurement instruments for outcomes included in a "Core Outcome Set"– a practical guideline. Trials. 2016 Sep;17(1):449.

32 Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol. 2007 Jan;60(1):34–42.

33 Mokkink LB, Prinsen CA, Patrick DL, Alonso J, Bouter LM, de Vet HC, et al. COSMIN methodology for systematic reviews of Patient-Reported Outcome Measures (PROMs) - user manual version 1.0. 2018. p. 78.

34 Currie CL, Monk BE. Can the response of port-wine stains to laser treatment be reliably assessed using subjective methods? Br J Dermatol. 2000 Aug;143(2):360–4.

35 Szychta P, Al-Nakib K, Anderson W, Stewart K, Quaba A. Quantitative method for evaluation of aesthetic results after laser treatment for birthmarks. Lasers Med Sci. 2013 Nov; 28(6):1567–72.

36 Pérez B, Abraira V, Núñez M, Boixeda P, Perez Corral F, Ledo A. Evaluation of agreement among dermatologists in the assessment of the color of port wine stains and their clearance after treatment with the flashlamp-pumped dye laser. Dermatology. 1997;194(2): 127–30.

37 Sajan JA, Tibesar R, Jabbour N, Lander T, Hilger P, Sidman J. Assessment of pulsed-dye laser therapy for pediatric cutaneous vascular anomalies. JAMA Facial Plast Surg. 2013 Nov-Dec;15(6):434–8.

38 Naran S, Gilmore J, Deleyiannis FW. The assessment of port wine stains in children following multiple pulsed-dye laser treatments. Ann Plast Surg. 2008 Apr;60(4):426–30.

39 Koster PH, Bossuyt PM, van der Horst CM, Gijsbers GH, van Gemert MJ. Assessment of clinical outcome after flashlamp pumped pulsed dye laser treatment of portwine stains: a comprehensive questionnaire. Plast Reconstr Surg. 1998 Jul;102(1):42–8.

40 Wang J, Zhu YY, Wang ZY, Yao XH, Zhang LF, Lv H, et al. Analysis of quality of life and influencing factors in 197 Chinese patients with port-wine stains. Medicine (Baltimore). 2017 Dec;96(51):e9446.

41 Masnari O, Schiestl C, Rössler J, Gütlein SK, Neuhaus K, Weibel L, et al. Stigmatization predicts psychological adjustment and quality of life in children and adolescents with a facial difference. J Pediatr Psychol. 2013 Mar;38(2):162–72.

42 Masnari O, Landolt MA, Roessler J, Weingaertner SK, Neuhaus K, Meuli M, et al. Self- and parent-perceived stigmatisation in children and adolescents with congenital or acquired facial differences. J Plast Reconstr Aesthet Surg. 2012 Dec;65(12):1664–70.

43 Lanigan SW, Cotterill JA. Psychological disabilities amongst patients with port wine stains. Br J Dermatol. 1989 Aug;121(2):209–15.

44 Troilius A, Wrangsjö B, Ljunggren B. Patients with port-wine stains and their psychosocial reactions after photothermolytic treatment. Dermatol Surg. 2000 Mar;26(3):190–6.

45 Safikhani S, Sundaram M, Bao Y, Mulani P, Revicki DA. Qualitative assessment of the content validity of the Dermatology Life Quality Index in patients with moderate to severe psoriasis. J Dermatolog Treat. 2013 Feb;24(1):50–9.

46 Achauer BM, Vander Kam VM, Padilla JF 3rd. Clinical experience with the tunable pulsed-dye laser (585 nm) in the treatment of capillary vascular malformations. Plast Reconstr Surg. 1993 Dec;92(7):1233–41.

47 Finlay AY, Khan GK. Dermatology Life Quality Index (DLQI)–a simple practical measure for routine clinical use. Clin Exp Dermatol. 1994 May;19(3):210–6.

48 Lawrence JW, Fauerbach JA, Heinberg LJ, Doctor M, Thombs BD. The reliability and validity of the Perceived Stigmatization Questionnaire (PSQ) and the Social Comfort Questionnaire (SCQ) among an adult burn survivor sample. Psychol Assess. 2006 Mar;18(1):106–11.

49 Fekkes M, Theunissen NC, Brugman E, Veen S, Verrips EG, Koopman HM, et al. Development and psychometric evaluation of the TAPQOL: a health-related quality of life instrument for 1-5-year-old children. Qual Life Res. 2000;9(8):961–72.

50 Ravens-Sieberer U, Gosch A, Rajmil L, Erhart M, Bruil J, Power M, et al.; KIDSCREEN Group. The KIDSCREEN-52 quality of life measure for children and adolescents: psychometric results from a cross-cultural survey in 13 European countries. Value Health. 2008 Jul-Aug;11(4):645–58.

51 Ravens-Sieberer U, Gosch A, Rajmil L, Erhart M, Bruil J, Duer W, et al. KIDSCREEN-52 quality-of-life measure for children and adolescents. Expert Rev Pharmacoecon Outcomes Res. 2005 Jun;5(3):353–64.

52 Achenbach TM, Edelbrock C, Howell CT. Empirically based assessment of the behavioral/emotional problems of 2- and 3- year-old children. J Abnorm Child Psychol. 1987 Dec;15(4):629–50.

53 van der Horst CM, de Borgie CA, Knopper JL, Bossuyt PM. Psychosocial adjustment of children and adults with port wine stains. Br J Plast Surg. 1997 Sep;50(6):463–7.

54 Miller AC, Pit-Ten Cate IM, Watson HS, Geronemus RG. Stress and family satisfaction in parents of children with facial port-wine stains. Pediatr Dermatol. 1999 May-Jun;16(3):190–7.

55 He Z, Lu C, Basra MK, Ou A, Yan Y, Li L. Psychometric properties of the Chinese version of Dermatology Life Quality Index (DLQI) in 851 Chinese patients with psoriasis. J Eur Acad Dermatol Venereol. 2013 Jan;27(1):109–15.

56 Mazzotti E, Barbaranelli C, Picardi A, Abeni D, Pasquini P. Psychometric properties of the Dermatology Life Quality Index (DLQI) in 900 Italian patients with psoriasis. Acta Derm Venereol. 2005;85(5):409–13.

57 Nijsten T, Meads DM, McKenna SP. Dimensionality of the dermatology life quality index (DLQI): a commentary. Acta Derm Venereol. 2006;86(3):284–5.

58 Ulijaszek SJ, Lourie JA. Intra- and inter-observer error in anthropometric measurement. In: Ulijaszek SJ, Mascie-Taylor CG, editors. Anthr. Individ. Popul. Cambridge: Cambridge University Press; 1994. pp. 30–55.

59 Gore C, Norton K, Olds T, Whittingham N, Birchall K, Clough M, et al. Accreditation in Anthropometry: an Australian Model. In: Norton K, Olds T, editors. Anthr. a Textb. body Meas. Sport. Heal. courses. Sydney: University of New South Wales Press; 1996. pp. 395–421.

60 Koster PH, Bossuyt PM, van der Horst CM, Gijsbers GH, van Gemert MJ. Characterization of portwine stain disfigurement. Plast Reconstr Surg. 1998 Sep;102(4):1210–6.

61 de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. J Clin Epidemiol. 2006 Oct;59(10):1033–9.

62 de Vet HC, Terwee CB, Mokkink LB, Knol DL. Measurement in medicine. Cambridge: Cambridge University Press; 2011.

63 Gottlieb AB, Levin AA, Armstrong AW, Abernethy A, Duffin KC, Bhushan R, et al. The International Dermatology Outcome Measures Group: formation of patient-centered outcome measures in dermatology. J Am Acad Dermatol. 2015 Feb;72(2):345–8.

64 Calvert M, Kyte D, Mercieca-Bebber R, Slade A, Chan AW, King MT, et al.; the SPIRIT-PRO Group. Guidelines for Inclusion of Patient-Reported Outcomes in Clinical Trial Protocols: The SPIRIT-PRO Extension. JAMA. 2018 Feb;319(5):483–94.

65 Kurwa H, Mills C, Lanigan S. Improvement in the psychological impact of a port wine stain after successful pulsed dye laser Improvement in the psychological impact of a port wine stain after successful pulsed dye laser therapy. J Dermatolog Treat. 1999;10(4):277–82.