Patient reported outcome measures for soft tissue facial reconstruction: a systematic review and evaluation of the quality of their measurement properties

Authors

Thomas D. Dobbs BM BCh, MA, MRCS[1,2]*, John A. G. Gibson MBBcH, MRCS[1], Sarah Hughes BSc, MHSc, MRCSLT[3,4], Arron Thind BA[5], Benjamin Patel BA[5], Hayley A Hutchings BSc, PhD[3], Iain Whitaker MBBChir PhD FRCS(Plast)[1,2]

Affiliations

(1) Reconstructive Surgery and Regenerative Medicine Research Group (ReconRegen), Institute of Life Science 2, Swansea University Medical School, Swansea, UK

(2) The Welsh Centre for Burns and Plastic Surgery, Morriston Hospital, Swansea, UK

(3) Patient and Population Health and Informatics, Institute of Life Science 2, Swansea University Medical School, Swansea, UK

(4) Abertawe Bro Morgannwg University Health Board, Princess of Wales Hospital, Bridgend, UK

(5) Oxford University Medical School, Oxford, UK

This is not the final published version

Full citation:

Dobbs TD, Gibson JAG, Hughes S, Thind A, Patel B, Hutchings, Whitaker I. Patient reported outcome measures for soft tissue facial reconstruction: a systematic review and evaluation of the quality of their measurement properties. *Plastic and Reconstructive Surgery* 2018

Corresponding Author

Mr Thomas Dobbs

The Welsh Centre for Burns and Plastic Surgery,

Morriston Hospital,

Swansea.

SA6 6NL

e: tomdobbs@doctors.org.uk

Meetings: This work has not yet been presented at any meetings.

Key Words: Facial reconstruction; patient-reported outcome measures; PROMs; systematic review; COSMIN; Quality of Life

Acknowledgements: We would like to acknowledge the help of Anne Powell, retired librarian at Abertawe Bro Morgannwg University Health Board.

Running Title: Systematic review of facial reconstruction PROMs

Words: 3128

Figures: 1

Tables: 8

Conflicts of Interest and Funding

None of the authors has a financial interest in any of the products, devices, or drugs mentioned in this manuscript. This work has received no specific funding. Mr Thomas Dobbs is funded by the Welsh Clinical Academic Training Fellowship.

Author role:

TD, HH and ISW developed the idea for this systematic review. TD, JG, AT, BP performed the literature review and data extraction. TD and SH assessed all included papers according to the methodology. TD produced the first draft of the manuscript and all authors were then involved in editing to reach the final, submitted version.

Abstract

Background

A patient's health-related quality of life (HRQoL) can be significantly impacted by facial scarring and disfigurement. Facial soft tissue reconstruction should aim to improve HRQoL, with outcomes measured from the patient's perspective using patient-reported outcome measures (PROMs). This systematic review identifies PROMs for soft tissue facial reconstruction and appraises their methodological and psychometric properties using up-to-date methods.

Methods

A systematic search of MEDLINE, EMBASE, PsychINFO and Cochrane was performed in line with the PRISMA guidelines. Identified PROMs were assessed using the updated COnsensus-based Standards for the Selection of Health Measurement INstruments (COSMIN) checklist. Psychometric properties were also assessed and a modified GRADE analysis was performed to aid in recommendations for future PROM use.

Results

Thirty-four studies covering 9 PROMs were included. Methodological quality and psychometric evidence was variable. FACE-Q, Skin Cancer Index (SCI), Patient Outcome of Surgery – Head/Neck (POS-Head/Neck) and the Derriford Appearance Scale 59/24 all demonstrated high enough evidence to be recommended as having potential for inclusion in future studies.

Conclusion

This is the first systematic review to identify and critically appraise PROMs for soft tissue facial reconstruction using internationally accepted criteria. Four PROMs were deemed to have adequate levels of methodological and psychometric evidence, although further studies should be conducted before their routine use in patients undergoing facial reconstruction. Through the use of psychometrically well-validated PROMs it is hoped that patients' concerns can be truly appreciated, level of care improved, and the quality of reconstructive options offered progressed.

Introduction

Health Related Quality of Life (HRQoL) is broadly defined as an individual's perception of the effects of an illness and/or treatment on the physical, psychological and social aspects of their life.[1,2] HRQoL can change over time, varying with changes in the condition itself, support network available, or other extrinsic factors.[3] The face plays an important role in social interactions[4,5] and therefore all three aspects of HRQoL can be affected by facial scarring and deformity. Unsurprisingly, facial scarring and disfigurement can lead to a number of psychosocial difficulties[3,6] and significantly reduce HRQoL.[7] In order to improve HRQoL in these patients, it is important that soft tissue reconstructive options address both form and function. Furthermore the reconstructive options offered should be appropriately appraised by the patients who will ultimately benefit from them. Traditionally, the outcomes of facial reconstruction have been assessed using non-objective or clinician reported measures. However, this is beginning to change.[8]

Patient reported outcome measures (PROMs) are standardized and validated questionnaires that are completed by patients to capture one or more aspects of their health and wellbeing.[9,10] They are broadly described as being generic (assessing general aspects of health) or disease-specific (covering aspects that are specific and pertinent to someone with that condition), with benefits and disadvantages to the use of either type.[11] The use of PROMs for the measurement of HRQoL has increased in recent years, with the UK Department of Health routinely collecting PROMs data on four surgical conditions[10] and the US Food and Drug Administration mandating their use in drug labeling.[2] Furthermore, the use of PROMs in clinical trials has become commonplace in many specialties, with recent consensus-based recommendations for

the inclusion of PROMs in the design of clinical trial protocols designed to further increase their use.[12] Despite their increasing use, there is a paucity of psychometrically robust PROMs as demonstrated by a number of systematic reviews.[13-16] This is particularly important if treatment decisions, study outcomes or adverse event reporting are to be based on their results. Psychometric validation of a PROM is complex, testing the questionnaire and its individual items for validity, reliability, responsiveness to change and clinical meaning. This validation process is described in greater detail elsewhere,[11,17] with some of the important terminology explained in *Table 1*.

Choosing the correct PROM to use based on its applicability to the condition of interest and its validity is therefore crucially important, especially if selecting instruments for inclusion in a Core Outcome Set (COS), where an agreed minimum set of outcomes is expected when reporting research in a specific disease area.[18,19]

The importance of soft tissue facial reconstruction in helping to restore form and function, whilst limiting the impact of facial scarring and deformity on HRQoL, mandates the need for reconstructive options to be assessed with appropriately designed and validated PROMs. This systematic review therefore aims to: (1) identify PROMs that have been designed for and/or validated in patients undergoing soft tissue facial reconstruction, (2) assess their psychometric properties and risk of bias using internationally agreed 'gold standards', (3) assess the adequacy of questions related to reconstruction and (4) make recommendations regarding appropriate PROMs for the inclusion in the future development of a COS in soft tissue facial reconstruction.

Methods

Search strategy and selection criteria

A systematic review protocol was developed *a priori* in accordance with the Preferred Reporting for Items for Systematic Reviews and Meta-Analyses-Protocols (PRISMA-P) guidance.[20,21] The search strategy was constructed in line with PRISMA guidelines,[22] the Cochrane handbook[23] and guidance from Terwee et al.[24] A sensitive, rather than specific, approach was taken to the search strategy, with three separate constructs used (target condition, target body area and measurement instrument). Key words or MeSH terms were used where available. The search strategy was trialed and modified in collaboration with an experienced librarian, with an example of the final search strategy seen in *Supplementary Figure 1*.

All searches were performed by two independent researchers (TD and AP) on the same day in February 2017 using; MEDLINE (Ovid), Embase (Ovid), PsychINFO (Ovid) and Cochrane. Results were uploaded to Distiller SR (Evidence Partners, Ontario, Canada) and duplicates removed. Grey literature (non-traditional or non-peer reviewed publications such as annual reports, government documents and unpublished literature) searching using Google, Google Scholar and known PROM based websites was also conducted. All studies were screened according to the inclusion and exclusion criteria (*Table 2*) by four reviewers (TD, JG, AT, BP), ensuring that all papers were screened by at least two reviewers. Articles that matched the inclusion criteria were downloaded in full-text format and re-screened (TD and JG). References were also searched to identify any previously missed studies. Discrepancies were discussed between the two reviewers and a third (HH) consulted

if required. The search strategy was re-run prior to submission in February 2018 to identify any new articles.

Data extraction and analysis

Data required for the following analyses were extracted from each paper and collated in Word and Excel for Mac (V14.5.7). Inter-rater reliability statistics were calculated using the Statistical Package for Social Sciences (SPSS) V.22 (IBM Corp., New York, USA). Results are presented as tables and a narrative synthesis.

Assessment of the methodological quality and psychometric properties of included studies

The COnsensus-based Standards for the Selection of Health Measurement INstruments (COSMIN) steering committee recently published guidelines on conducting systematic reviews of PROMs.[25] These include an updated version of the COSMIN checklist for assessing the methodological quality and risk of bias in studies reporting on PROM development and validation.[26-28] The updated COSMIN risk of bias checklist assesses 10 specific areas: PROM development, content validity, structural validity, internal consistency, cross-cultural validity and measurement invariance, reliability, measurement error, criterion validity, hypothesis testing for construct validity and responsiveness.[28,29] Each section is scored on a 5-category scale (very good, adequate, doubtful, inadequate and not-applicable), with the lowest score in each category considered the final overall rating for the methodological quality in that category for the paper assessed (i.e. if internal consistency is rated as ‘very good’ on one question, but ‘doubtful’ on another, the overall score for internal consistency in the paper being assessed is ‘doubtful’). All papers included in this review were assessed against these criteria, with summary scores presented for each PROM.

The original COSMIN checklist demonstrated reasonable inter-rater reliability[30] with the new version being produced to try and improve this further. However, due to there still being a degree of subjectivity, it is considered good practice to compare the results of two independent reviewers. A randomly-selected 30% sample of studies were assessed by two reviewers (TD and SH) and the category scores compared using percentage agreement and intraclass coefficient.[31] It was decided *a priori* that if agreement were low, all studies would be doubly reviewed.

Each study was also assessed for its psychometric quality using criteria developed by Terwee et al[32] and recently updated[25] (*Supplementary Figure 2*). The measurement properties assessed closely mirror those in the COSMIN checklist and are rated as either positive (+), negative (-) or indeterminate (?).

Evidence synthesis and GRADE analysis

The results of the two assessments described above were pooled and used to produce a global score for each measurement property of each PROM as outlined in Prinsen et al.[25] Results can be positive (+), negative (-), inconsistent (+/-) or indeterminate (?), with a '75% in agreement' rule used (i.e. for a positive outcome on structural validity, 75% or more of the studies reporting structural validity must be positive).[29] The quality of the evidence contributing to this outcome was graded using a modified version of the Grading of Recommendation Assessment, Development and Evaluation (GRADE) approach for systematic reviews of clinical trials.[25,33] Those measurement categories that score an indeterminate (?) cannot be graded as no evidence has been presented in the studies assessed. Finally, the combined results of each measurement category and GRADE analysis were used to formulate

recommendation on the appropriateness of each PROM for use in a soft tissue facial reconstruction population.

Assessment of reconstructive relevance

Studies were selected based on their relevance to soft tissue facial reconstruction. Despite this, a secondary assessment of the face validity, specifically relating to soft tissue reconstruction was performed. No precedent exists; therefore, the authors made a subjective assessment of all items in each included PROM, allowing recommendations for future item and PROM generation to be made where required.

Results

Following the removal of duplicates, 16,165 individual title and abstracts were screened. Seventeen additional papers were added following reference screening, leading to 34 studies being included (*Figure 1*).[34-67] These 34 studies presented evidence for the design and/or validation of 9 PROMs for soft tissue facial reconstruction: FACE-Q, Patient Outcomes of Surgery-Head/Neck (POS-Head/Neck), Patient Scar Assessment Questionnaire (PSAQ), Nasal Appearance and Function Evaluation Questionnaire (NAFEQ), Lip Reanimation Outcome Questionnaire, Rhinoplasty/Facelift/Blepharoplasty/Skin Rejuvenation Outcomes Evaluation (ROE/FOE/BOE/SROE), Patient and Observer Scar Assessment Scale (POSAS), Skin Cancer Index (SCI) and Derriford Appearance Scale (DAS 59/24). A summary of these 9 PROMs is presented in *Table 3*.

Methodological quality and psychometric properties of included studies

Table 4 presents a summary of the cumulative COSMIN outcomes for each measurement property for those included PROMs. PROM development and content validity was deemed 'doubtful' or 'inadequate' for all but FACE-Q, SCI and DAS 59/24 and even then only SCI scored 'adequate' or 'very good' for both. Internal consistency was examined in all PROMs and was deemed 'very good' for all. Structural validity and reliability were also assessed in all PROMs; however, the other measurement properties were reported sporadically.

Average percentage agreement between the two independent COSMIN reviewers was 93.6%, with an ICC of 0.844 (95% CI, 0.808 – 0.874), demonstrating good agreement.

The psychometric properties of each study were also assessed as detailed in the methods. *Table 5* presents a summary of the cumulative score for each measurement category for each PROM, based on the 'best score' wins approach to summarizing each individual paper for each PROM into a summary score. A number of papers reported very little detail on psychometric validation and therefore a significant number have been given an indeterminate "?" result as there is neither enough to give a "+" or "-" result. FACE-Q and DAS 59/24 are the two PROMs with the highest number of positive ratings.

Evidence synthesis and GRADE analysis

In order to provide an overall assessment of each individual PROM and adjust for poor quality evidence, the results of *table 4* and *5* were pooled and a modified GRADE analysis performed as per the method described previously. Four PROMs, FACE-Q, SCI, POSAS and DAS 59/24 had high levels of evidence quality for those measurement properties that could be assessed. All the remaining PROMs were

downgraded in terms of evidence quality, mainly due to small participant numbers or only single studies of adequate quality on an individual PROM. The results of this are presented in *Table 6*. Finally, in order to provide recommendations for the use of PROMs in soft tissue facial reconstruction in the future, each PROM was categorized according to its potential (*Table 7*). FACE-Q, SCI, POS-Head/Neck and DAS 59/24 all demonstrated enough high-quality evidence of their methodological and psychometric properties to be considered an 'A' grade PROM.

Assessment of reconstructive relevance

The items included in each PROM were assessed for their specific relevance to soft tissue reconstruction as judged by the authors. Summary findings are presented in *Table 8*.

Discussion

This systematic review has been designed to identify PROMs that have either been designed for, or validated in, a soft tissue facial reconstruction population. Internationally recognized best practice was used to appraise the quality of evidence and risk of bias in studies reporting on the design and validation of those included PROMs.[25,27,28] Other methods for assessing the psychometric properties of a PROM exist.[2,68] However, the COSMIN checklist is now routinely used in systematic reviews of PROMs across many specialities such as orthopaedics,[69] paediatrics,[70] dermatology[71] and neurology[72] and should be incorporated into all PROMs-based systematic reviews in plastic and reconstructive surgery.

Of the nine PROMs identified as having been designed for or validated in an appropriate population, there are a range of conditions or facial areas which they

focus on. All are condition-specific PROMs as it was felt that generic PROMs, while useful, would not have items that sufficiently covered aspects relevant to soft tissue facial reconstruction and were therefore excluded. However, of those condition-specific PROMs included, some are narrowly focused (e.g. NAFEQ on nasal reconstruction), while some are more broadly applicable (e.g. FACE-Q) and others are on the cusp of being non-specific but still relevant (e.g. DAS 59/24). PROMs specifically designed for rhinoplasty were excluded for two reasons: firstly because it was determined that a rhinoplasty involves more extensive tissue manipulation than just the soft tissues and secondly because there has been a recent systematic review that addresses this area.[73]

The methodological quality of the included studies as assessed using the COSMIN checklist varied widely, suggesting a significant risk of bias for many of the studies. When results were collated across studies for each PROM, it was revealed that while some aspects of design and validation were done well (e.g. internal consistency), many were done poorly (e.g. content reliability and responsiveness) and some were only sporadically reported (e.g. measurement error and criterion validity).

The measurement properties of 'PROM development' and 'content validity' scored poorly across all PROMs. This was likely the result of poor quality qualitative work in the generation of items (such as insufficiently sized qualitative interview groups and inappropriate coding methods for theme generation) leading to poor ratings on the COSMIN checklist, as well as a general lack of good quality reporting across studies.

As with any risk of bias assessment tool, one is reliant on the information being reported in the manuscript in order to give a positive or negative result. However, it appears that the majority of older studies reported poorly on many

aspects of PROM design and validation that are now considered to be important. Therefore, by definition, these studies will score poorly in many of these categories as scored using the COSMIN checklist. This makes it difficult to differentiate between those PROMs that have good content validity but lost points due to errors of omission in the reporting versus those that were poorly developed and lacked content validity. Evidence for psychometric validity was variable across all of the included PROMs, with many scoring 'indeterminate' for the quality of a psychometric property due to a lack of reporting as described above.

Research performed with poor quality PROMs constitutes a waste of resources.[74] Poorly validated studies with little clinical meaning and high responder burden are not suitable for routine clinical practice and limit the benefit of PROMs for the surgeon in terms of the critical appraisal of outcomes. For these reasons the combination of the COSMIN checklist[28] and the updated Terwee et al checklist[25] to form a summary of the evidence base for each PROM, as performed here, is crucial. In this systematic review four PROMs were identified as having sufficient methodological rigor and psychometric validity, combined with high quality evidence to be placed in grade 'A'. These PROMs (FACE-Q, SCI, POS-Head/Neck and DAS 59/24) all therefore have the potential to be recommended as the most suitable PROMs for inclusion in a COS for facial reconstruction. They do, however, all have deficiencies in their design and validation, which should be addressed through further large-scale psychometric evaluation. Furthermore, as can be seen from the assessment of their item focus on reconstruction, none are able to cover the full spectrum of likely concerns of a patient undergoing soft tissue facial reconstruction. FACE-Q and the scar related PROMS (PSAQ and POSAS) have the greatest number of relevant questions (despite being designed for a cosmetic facial population and scarring

respectively), but all are still lacking in a number of key areas. Further item generation and validation is therefore required, either as a new PROM or as additional items to one of the identified PROMs. Soft tissue facial reconstruction also encompasses a wide range of patients, from those with minor defects to those requiring large functional and aesthetic reconstructions. It is likely that a 'one-size-fits-all' PROM will not be able to address this spectrum of concerns and therefore multiple PRO instruments or a split design PROM is required.

The use of the COSMIN checklist and guidance by Prinsen et al[25] is a strength of this study. Despite the COSMIN checklist being considered the 'gold standard' for appraising the PROM literature, it has its limitations. The checklist is extensive and requires knowledge of the health-outcomes literature, potentially making it inaccessible to the non-specialist reader. Some sections are also subjective in parts, requiring the user to "read between the lines" of the assessed studies on occasions. To overcome this, two reviewers reviewed a 30% sample of papers in order to confirm that the percentage agreement and ICC between them was sufficient. We appreciate that other review teams could score sections differently, altering the final outcome.

A broad search strategy was used to identify all pertinent studies; however, only studies that demonstrated aspects of PROM design or validation were included. Because PROM validity was considered to be of utmost important, this could mean that PROMs which include useful items but that have not been validated were missed. Furthermore, the decision to exclude both generic and paediatric PROMs was based on the aim of identifying those PROMs that would have items most relevant to the adult soft tissue facial reconstruction patient. We appreciate that this decision may lead to potentially useful items being missed.

Conclusion

This is the first systematic review to identify PROMs for soft tissue facial reconstruction. This review has identified a number of different PROMs, which have all to some degree been designed for, or validated in, patients undergoing soft tissue facial reconstruction. Unfortunately, there is great variability in the quality of the validation process and, despite suggesting four PROMs that would potentially be suitable for inclusion in a COS for facial reconstruction, all of these instruments require further validation studies. In addition, for inclusion in a COS, decisions with regard to delivery medium, pre-operative and post-operative assessment timing would need to be made. Therefore, a PROM including an amalgamation of items from all those identified PROMs, plus newly designed items, would best address the concerns of patients undergoing reconstructive procedures for soft tissue facial deformities. The findings of this review suggest there is the need for a new PROM that includes items that measure functional, psycho-relational and cosmetic components of quality of life in these patients. All those involved in facial reconstruction are urged to take on the challenge of developing and validating such a PROM. In time this will allow a COS can be agreed upon, with treatments evaluated and improved according to the wishes of our patients.

References

1. Karimi M, Brazier J. Health, Health-Related Quality of Life, and Quality of Life: What is the Difference? *PharmacoEconomics*. 2016;34(7):645-649.

2. Health USDO, Human Services FDA Center for Drug Evaluation, Research USDOH, et al. Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance. *Health Qual Life Outcomes*. 2006;4(1):79.

3. Rumsey N, Harcourt D. Body image and disfigurement: issues and interventions. *Body Image*. 2004;1(1):83-97.

4. Roberts RM, Gierasch A. The effect of visible facial difference on personal space during encounters with the general public. *Plast Surg Nurs*. 2013;33(2):71–80.

5. Kish V, Lansdown R. Meeting the Psychosocial Impact of Facial Disfigurement: Developing a Clinical Service for Children and Families. *Clinical Child Psychology and Psychiatry*. 2016;5(4):497-512.

6. Macgregor FC. Facial disfigurement: problems and management of social interaction and implications for mental health. *Aesthetic Plastic Surgery*. 1990;14(4):249-257.

7. Dey JK, Ishii LE, Joseph AW, et al. The Cost of Facial Deformity: A Health Utility and Valuation Study. *JAMA Facial Plast Surg*. 2016;18(4):241-249.

8. Most SP, Moubayed SP. Patient-Reported Outcome Measures for Facial Plastic Surgery: A Specialty Finally Gets to Go to the PROM. *JAMA Facial Plast Surg*. 2017;19(2):101-101.

9. McGrail K, Bryan S, Davis J. Let's all go to the PROM: the case for routine patient-reported outcome measurement in Canadian healthcare. *Healthc Pap*. 2011;11(4):8–18.

10. Devlin NJ, Appleby J. Getting the most out of PROMS. Putting health outcomes at the heart of NHS Decision-making. *The King's Fund. London.* 2010.

11. Wormald JCR, Rodrigues JN. Outcome measurement in plastic surgery. *J Plast Reconstr Aesthet Surg*. 2018;71(3):283-289..

12. Calvert M, Kyte D, Mercieca-Bebber R, et al. Guidelines for Inclusion of Patient-Reported Outcomes in Clinical Trial Protocols: The SPIRIT-PRO Extension. *JAMA*. 2018;319(5):483-494..

13. Pusic AL, Chen CM, Cano S, et al. Measuring Quality of Life in Cosmetic and Reconstructive Breast Surgery: A Systematic Review of Patient-Reported Outcomes Instruments. *Plast Reconstr Surg*. 2007;120(4):823-837.

14. Barone M, Cogliandro A, Coppola MM, et al. Patient-reported outcome measures following gynecomastia correction: a systematic review. *Eur J Plast Surg*. 2017;41(2):109-118.

15. Barone M, Cogliandro A, Di Stefano N, Tambone V, Persichetti P. A Systematic Review of Patient-Reported Outcome Measures Following

Transsexual Surgery. *Aesthetic Plastic Surgery*. 2017;41(3):700-713.

16. Kosowski TR, McCarthy C, Reavey PL, et al. A Systematic Review of Patient-Reported Outcome Measures after Facial Cosmetic Surgery and/or Nonsurgical Facial Rejuvenation. *Plast Reconstr Surg*. 2009;123(6):1819-1827.

17. Dobbs T, Hughes S, Mowbray N, Hutchings HA, Whitaker IS. How to decide which patient-reported outcome measure to use? A practical guide for plastic surgeons. *Journal of Plastic, Reconstructive & Aesthetic Surgery*. March 2018. doi:10.1016/j.bjps.2018.03.007.

18. Williamson P, Altman D, Blazeby J, Clarke M, Gargon E. Driving up the quality and relevance of research through the use of agreed core outcomes. *J Health Serv Res Policy*. 2012;17(1):1-2.

19. Porter ME, Larsson S, Lee TH. Standardizing Patient Outcomes Measurement. *N Engl J Med*. 2016;374(6):504-506.

20. Moher D, Shamseer L, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews 2015 4:1*. 2015;4(1):1.

21. Shamseer L, Moher D, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ*. 2015;349(jan02 1):g7647-g7647..

22. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *Ann Intern Med*. 2009;151(4):264-269.

23. Higgins J, Green S. *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1. 0 [Updated March 2011]*. The Cochrane Collaboration; 2011.

24. Terwee CB, Jansma EP, Riphagen II, de Vet HCW. Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Qual Life Res*. 2009;18(8):1115-1123.

25. Prinsen CAC, Mokkink LB, Bouter LM, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res*. 2018;41(2):212–11.

26. 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(4):539-549.

27. 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(7):737-745.

28. 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*. 2017;383(9912):166-169.

29. Mokkink LB, Prinsen CAC, Patrick DL, et al. COSMIN methodology for systematic reviews of Patient- Reported Outcome Measures (PROMs); user manual. February 2018:1-78. Last accessed on 10th March 2018 at

http://www.cosmin.nl.

30. Mokkink LB, Terwee CB, Gibbons E, et al. Inter-rater agreement and reliability of the COSMIN (COnsensus-based Standards for the selection of health status Measurement Instruments) Checklist. *BMC Med Res Methodol*. 2010;10(1):82.

31. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420-428.

32. Terwee CB, Bot SDM, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol*. 2007;60(1):34-42.

33. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *BMJ*. 2008;336(7650):924-926..

34. Klassen AF, Cano SJ, Scott A, Snell L, Pusic AL. Measuring patient-reported outcomes in facial aesthetic patients: development of the FACE-Q. *Facial plast Surg*. 2010;26(4):303-309.

35. Pusic AL, Klassen AF, Scott AM, Cano SJ. Development and Psychometric Evaluation of the FACE-Q Satisfaction with Appearance Scale. *Clin Plast Surg*. 2013;40(2):249-260.

36. Klassen AF, Cano SJ, Scott AM, Pusic AL. Measuring outcomes that matter to face-lift patients: development and validation of FACE-Q appearance appraisal scales and adverse effects checklist for the lower face and neck. *Plast Reconstr*

Surg. 2014;133(1):21-30.

37. Panchapakesan V, Klassen AF, Cano SJ, Scott AM, Pusic AL. Development and Psychometric Evaluation of the FACE-Q Aging Appraisal Scale and Patient-Perceived Age Visual Analog Scale. *Aesthet Surg J*. 2013;33(8):1099-1109.

38. Klassen AF, Cano SJ, Schwitzer JA, Scott AM, Pusic AL. FACE-Q scales for health-related quality of life, early life impact, satisfaction with outcomes, and decision to have treatment: development and validation. *Plast Reconstr Surg*. 2015;135(2):375-386.

39. Klassen AF, Cano SJ, Schwitzer JA, et al. Development and Psychometric Validation of the FACE-Q Skin, Lips, and Facial Rhytids Appearance Scales and Adverse Effects Checklists for Cosmetic Procedures. *JAMA Dermatol*. 2016;152(4):443-451.

40. Klassen AF, Cano SJ, Alderman A, et al. Self-Report Scales to Measure Expectations and Appearance-Related Psychosocial Distress in Patients Seeking Cosmetic Treatments. *Aesthet Surg J*. 2016;36(9):1068-1078.

41. Klassen AF, Cano SJ, Grotting JC, et al. FACE-Q Eye Module for Measuring Patient-Reported Outcomes Following Cosmetic Eye Treatments. *JAMA Facial Plast Surg*. 2017;19(1):7-14.

42. Albornoz CR, Pusic AL, Reavey P, et al. Measuring health-related quality of life outcomes in head and neck reconstruction. *Clin Plast Surg*. 2013;40(2):341-349.

43. Cano SJ, Browne JP, Lamping DL, Roberts AHN, McGrouther DA, Black NA. The Patient Outcomes of Surgery-Head/Neck (POS-head/neck): a new patient-based outcome measure. *J Plast Reconstr Aesthet Surg*. 2006;59(1):65-73.

44. Durani P, McGrouther DA, Ferguson MW. The Patient Scar Assessment Questionnaire: A Reliable and Valid Patient-Reported Outcomes Measure for Linear Scars. *Plast Reconstr Surg*. 2009;123(5):1481-1489.

45. Economopoulos KP, Petralias A, Linos E, Linos D. Psychometric Evaluation of Patient Scar Assessment Questionnaire Following Thyroid and Parathyroid Surgery. *Thyroid*. 2012;22(2):145-150. doi:10.1089/thy.2011.0265.

46. Moolenburgh SE, Mureau MAM, Duivenvoorden HJ, Hofer SOP. Validation of a questionnaire assessing patient's aesthetic and functional outcome after nasal reconstruction: the patient NAFEQ-score. *J Plast Reconstr Aesthet Surg*. 2009;62(5):656-662.

47. de Almeida JR, Alexander AJ, Shrime MG, Gilbert RW, Goldstein DP. Development and preliminary validation of the lip reanimation outcomes questionnaire. *Otolaryngol Head Neck Surg*. 2010;143(3):361-366.

48. Alsarraf R. Outcomes research in facial plastic surgery: a review and new directions. *Aesthetic Plastic Surgery*. 2000;24(3):192-197.

49. Alsarraf R, Larrabee WF, Anderson S, Murakami CS, Johnson CM. Measuring cosmetic facial plastic surgery outcomes: a pilot study. *Arch Facial Plast Surg*. 2001;3(3):198-201.

50. Draaijers LJ, Tempelman FRH, Botman YAM, et al. The Patient and Observer

Scar Assessment Scale: A Reliable and Feasible Tool for Scar Evaluation. *Plast Reconstr Surg*. 2004;113(7):1960-1965.

51. van de Kar AL, Corion LUM, Smeulders MJC, Draaijers LJ, van der Horst CMAM, van Zuijlen PPM. Reliable and Feasible Evaluation of Linear Scars by the Patient and Observer Scar Assessment Scale. *Plast Reconstr Surg*. 2005;116(2):514-522.

52. van der Wal MBA, Tuinebreijer WE, Bloemen MCT, Verhaegen PDHM, Middelkoop E, van Zuijlen PPM. Rasch analysis of the Patient and Observer Scar Assessment Scale (POSAS) in burn scars. *Qual Life Res*. 2011;21(1):13-23.

53. Liu X, Nelemans PJ, Van Winden M, Kelleners Smeets NWJ, Mosterd K. Reliability of the Patient and Observer Scar Assessment Scale and a 4‐point scale in evaluating linear facial surgical scars. *J Eur Acad Dermatol Venereol*. 2017;31(2):341-346.

54. Rhee JS, Matthews BA, Neuburg M, Burzynski M, Nattinger AB. Creation of a Quality of Life Instrument for Nonmelanoma Skin Cancer Patients. *Laryngoscope*. 2005;115(7):1178-1185.

55. Matthews BA, Rhee JS, Neuburg M, Burzynski ML, Nattinger AB. Development of the facial skin care index: a health-related outcomes index for skin cancer patients. *Dermatol Surg*. 2006;32(7):924–34.

56. Rhee JS, Matthews BA, Neuburg M, Logan BR, Burzynski M, Nattinger AB. Validation of a Quality-of-Life Instrument for Patients With Nonmelanoma Skin Cancer. *Arch Facial Plast Surg.* 2006;8(5):314-318.

57. Rhee JS, Matthews BA, Neuburg M, Logan BR, Burzynski M, Nattinger AB. The Skin Cancer Index: Clinical Responsiveness and Predictors of Quality of Life. *Laryngoscope*. 2007;117(3):399-405.

58. de Troya-Martín M, Rivas-Ruiz F, Blázquez-Sánchez N, et al. A Spanish version of the Skin Cancer Index: a questionnaire for measuring quality of life in patients with cervicofacial nonmelanoma skin cancer. *Br J Dermatol*. 2015;172(1):160-168.

59. Klassen A, Jenkinson C, Fitzpatrick R, Goodacre T. Measuring quality of life in cosmetic surgery patients with a condition-specific instrument: the Derriford Scale. *Br J Plast Surg*. 1998;51(5):380-384.

60. Carr T, Harris D, James C. The Derriford Appearance Scale (DAS‐59): A new scale to measure individual responses to living with problems of appearance. *Br J Health Psychol*. 2000;5(2):201-215.

61. Harris DL, Carr AT. The Derriford Appearance Scale (DAS59): a new psychometric scale for the evaluation of patients with disfigurements and aesthetic problems of appearance. *Br J Plast Surg*. 2001;54(3):216-222.

62. Carr T, Moss T, Harris D. The DAS24: a short form of the Derriford Appearance Scale DAS59 to measure individual responses to living with problems of appearance. *B J Health Psychol*. 2005;10(Pt 2):285-298.

63. Moss TP, Lawson V, White P. Identification of the underlying factor structure of the Derriford Appearance Scale 24. *PeerJ*. 2015;3(6):e1070.

64. Singh RK, Moss T, Roy DK, et al. Translation and validation of the Nepalese

version of Derriford appearance scale (DAS-59). *MPS*. 2013;03(02):51-56.

65. Moss TP, Lawson V, Liu CY. The Taiwanese Derriford Appearance Scale: The translation and validation of a scale to measure individual responses to living with problems of appearance. *Psych J*. 2015;4(3):138-145.

66. Cogliandro A, Persichetti P, Ghilardi G, et al. How to assess appearance distress and motivation in plastic surgery candidates: Italian validation of Derriford Appearance Scale 59 (DAS 59). *Eur Rev Med Pharmacol Sci*. 2016;20(18):3732-3737.

67. Sadeghi-Bazargani H, Zare Z, Ranjbar F. Factor structure of the Persian version of general, social, and negative self-consciousness of appearance domains of Derriford Appearance Scale 59: an application in the field of burn injuries. *Neuropsychiatr Dis Treat*. 2017;13:147-154.

68. Aaronson N, Alonso J, Burnam A, et al. Assessing health status and quality-of-life instruments: attributes and review criteria. *Qual Life Res*. 2002;11(3):193-205.

69. Gagnier JJ, Mullin M, Huang H, et al. A Systematic Review of Measurement Properties of Patient-Reported Outcome Measures Used in Patients Undergoing Total Knee Arthroplasty. *J Arthroplasty*. 2017;32(5):1688–1697.e7.

70. Speyer R, Cordier R, Parsons L, Denman D, Kim J-H. Psychometric Characteristics of Non-instrumental Swallowing and Feeding Assessments in Pediatrics: A Systematic Review Using COSMIN. *Dysphagia*. 2017;31(1):1-14.

71. Gerbens LAA, Prinsen CAC, Chalmers JR, et al. Evaluation of the measurement properties of symptom measurement instruments for atopic eczema: a systematic review. *Allergy*. 2017;72(1):146-163.

72. Haywood KL, Mars TS, Potter R, Patel S, Matharu M, Underwood M. Assessing the impact of headaches and the outcomes of treatment: A systematic review of patient-reported outcome measures (PROMs). *Cephalalgia*. January 2017. Doi:10.1177/0333102417731348.

73. Barone M, Cogliandro A, Di Stefano N, Tambone V, Persichetti P. A systematic review of patient-reported outcome measures after rhinoplasty. *Eur Arch Otorhinolaryngol*. 2017;274(4):1807-1811.

74. Ioannidis JPA, Greenland S, Hlatky MA, et al. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet*. 2014;383(9912):166-175.

Figure 1: PRISMA flow diagram demonstrating the identification and inclusion of studies

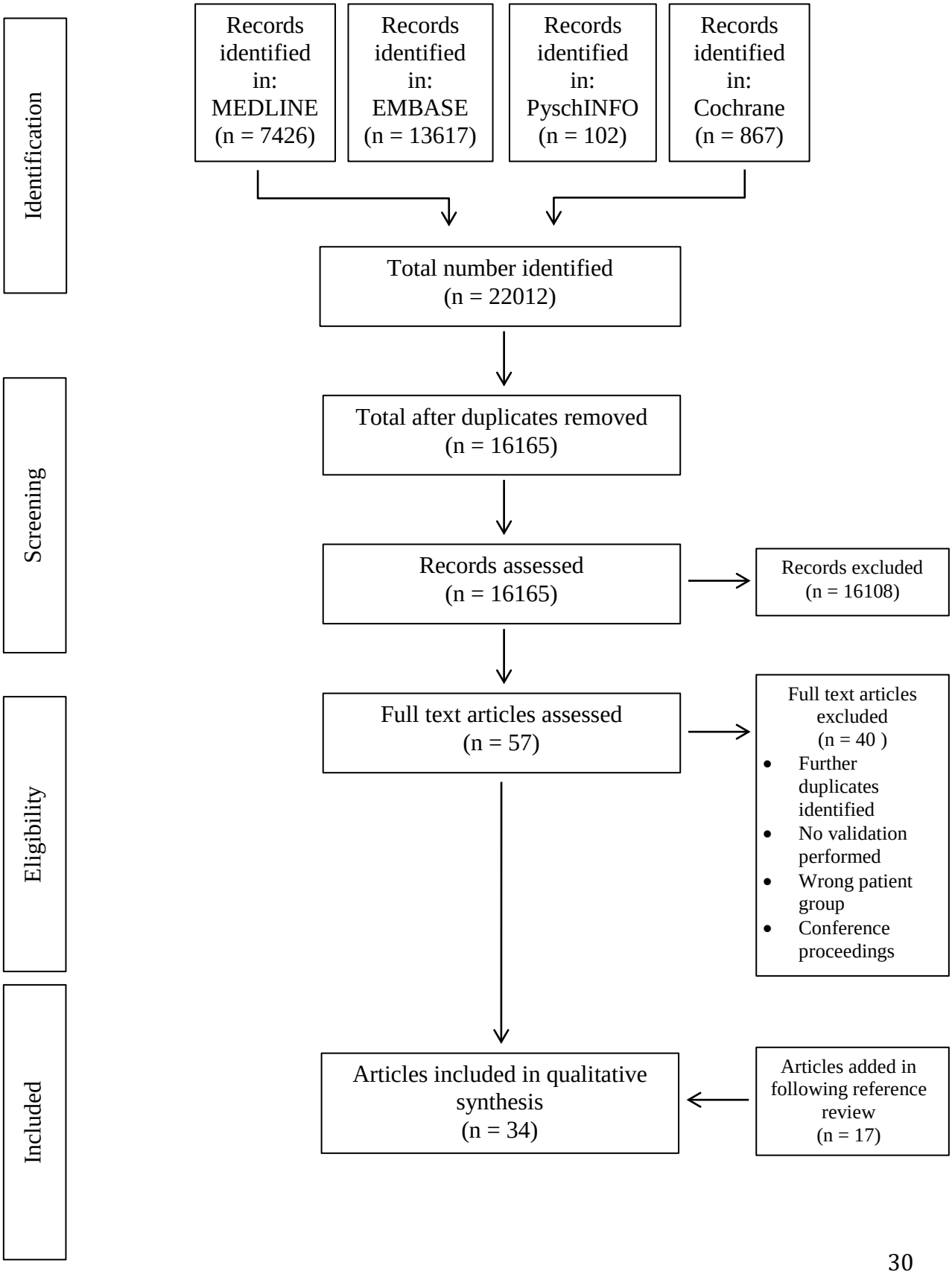

Table 1: Glossary of terms used in the psychometric validation of patient reported outcome measures. Reproduced from Dobbs et al, 2018.[17]

Term	Definition
Classical Test Theory	The traditional method of assessing the scientific robustness of a PROM.
Content validity	Refers to whether the whole instrument is measuring all that is relevant and important to the patient and their condition.
Criterion validity	Assessment of how well the instrument being studied correlates with another instrument (ideally considered to be the gold-standard).
Face validity	A subjective measure of whether the questions are actually measuring what they are meant to be.
Instrument	A method of capturing data. In the case of patient-reported outcome measures an instrument usually refers to a questionnaire.
Items	An item is an individual question. Multiple items make up an instrument.
Interpretability	The degree to which one can assign clinical meaning to the quantitative score given by an instrument.
Modern Test Theory	Rasch measurement theory and item response theory and two methods encompassed by the term 'modern test theory'. These are newer methods of statistical analysis, designed to address some of the flaws of classical test theory.

Patient-reported outcome measures	Standardised and validated questionnaires that are designed to capture one or more aspect of a person's health and wellbeing.
Reliability	Refers to how consistent the results are when the instrument is applied in different situations.
Responsiveness	Refers to the ability of an instrument to measure a clinically important change.
Sensitivity	Refers to the ability of an instrument to measure any change.

Table 2: Inclusion and exclusion criteria used when screening studies identified in the literature search.

Inclusion Criteria	1) Soft tissue facial reconstruction or aesthetic improvement 2) Papers discussing some aspect of PROM development or validation 3) English only articles
Exclusion Criteria	1) Questionnaires not developed or validated in patients undergoing soft tissue facial reconstruction or aesthetic surgery 2) Oropharyngeal head and neck cancer population 3) Bony reconstruction of the face (e.g. mandibular or maxillary reconstruction) 4) Questionnaires developed for the paediatric population 5) General oncology questionnaires unless specifically validated in a facial reconstruction population 6) General HRQoL questionnaires unless specifically validated in a facial reconstruction population 7) Meeting abstracts or letters

Table 3: Summary of included patient reported outcome measures (PROMs), including the population of interest and the domains covered by each PROM.

PROM	Papers	Country of study	Population	Total population size (n = individuals)	Number of items	Domains
FACE-Q	Klassen et al, 2010[34] Pusic et al, 2013[35] Klassen et al, 2014[36] Panchapakesan et al, 2013[37] Klassen et al, 2015[38] Klassen et al, 2016[39] Klassen et al, 2016[40] Klassen et al, 2017[41] Albornoz et al, 2013[42]	Canada / USA / Europe	Facial aesthetic patients undergoing a range of surgical and non-surgical treatments	> 783	353 (across a wide range of subscales)	• Satisfaction with facial appearance • Quality of life • Adverse effects • Patient experience
POS-Head/Neck	Cano et al, 2006[43]	United Kingdom	Patients undergoing surgery for head and neck skin lesions	458	15 (6 pre-operative and 9 post-operative)	• Psychological functioning • Cosmetic appearance • Satisfaction

PSAQ	Durani et al, 2009[44] Economopoulos et al, 2012[45]	United Kingdom / Greece	Thyroid surgery	1252	39	• Scar appearance • Consciousness • Satisfaction with scar appearance • Satisfaction with scar symptoms
NAFEQ	Moolenburgh et al, 2009[46]	Netherlands / Canada	Nasal reconstruction	208	14	• Nasal function • Satisfaction with nasal appearance
Lip Reanimation Outcome Questionnaire	de Almeida et al, 2010[47]	Canada	Lip reconstruction and reanimation patients	20	15	• Appearance • Oral competence • Speech • Symmetry
ROE/FOE/BOE/SROE	Alsarraf et al, 2000[48] Alsarraf et al, 2001[49]	USA	Facial aesthetic patients	78	6 (in each instrument)	• Physical • Mental/emotional • Social
POSAS	Draaijers et al, 2004[50] van der Kar et al, 2005[51] van der Wal et al, 2012[52] Liu et al, 2017[53]	Netherlands	Patients with scars, both linear and burns	877	12 (+ 2 overall questions not scored)	• Scarring (patient rated) • Scarring (observer rated)
SCI	Rhee et al, 2005[54] Matthews et al, 2006[55] Rhee et al, 2006[56] Rhee et al, 2007[57] de Troya-	USA / Spain	Non-melanoma facial skin cancer	776	15	• Emotional well-being • Social well-being • Appearance issues

	Martin et al, 2015[58]					
DAS 59/24	Klassen et al, 1998[59] Carr et al, 2000[60] Harris et al, 2001[61] Carr et al, 2005[62] Moss et al, 2015[63] Singh et al, 2013[64] Moss et al all, 2015[65] Cogliandro et al, 2016[66] Sadeghi-Bazargani et al, 2017[67]	United Kingdom / Taiwan / Italy / Iran / Nepal	Patients with problems with appearance Normal controls	2741 (for DAS 59) 2907 (for DAS 24 1621 (for cross-cultural adaption)	59 in long version 24 in short version	• Self consciousness of appearance • Social self consciousness of appearance • Sexual and bodily self consciousness of appearance • Negative self concept • Facial self consciousness of appearance

Table 4: Summary of the cumulative scores for each PROM as assessed by the COSMIN checklist. The best

property across all papers contributing to the validation of the individual PROM was used to determine the cu

measurement property for the PROM in question.

PROM	PROM development	Content validity	Structural validity	Internal consistency	Cross-cultural validity/Measurement invariance	Reliability	Measurement error	Crite vali
FACE-Q	Doubtful	Adequate	Very good	Very good	Adequate	Adequate	--	--
POS-Head/Neck	Doubtful	Doubtful	Adequate	Very good	--	Adequate	--	--
PSAQ	Doubtful	Doubtful	Inadequate	Very good	--	Very good	--	--
NAFEQ	Inadequate	Inadequate	Very good	Very good	--	Inadequate	--	--
Lip Reanimation Outcome Questionnaire	Inadequate	Inadequate	Inadequate	Very good	--	Adequate	--	--
ROE/FOE/BOE/SROE	Inadequate	Inadequate	Inadequate	Very good	--	Adequate	--	--
POSAS	Inadequate	Inadequate	Very good	Very good	--	Adequate	--	--
SCI	Adequate	Very good	Very good	Very good	Very good	Adequate	--	--
DAS 59/24	Doubtful	Adequate	Very good	Very good	Very good	Adequate	--	Very

Very good/adequate/doubtful/inadequate/not-applicable are the 5-categories of the COSMIN checklist

"--" when no information was presented in the included studies to assess

Table 5: Summary of cumulative score for each category assessed per PROM using the modified Terwee et a

based on the best score for each measurement property in all studies contributing to a PROM in the same ma

PROM	Structural validity	Internal consistency	Reliability	Measurement error	Hypotheses testing for construct validity	Cross-cultural validity/Measurement invariance	Criterion validity	Responsiv
FACE-Q	+	+	+	?	+	?	?	+
POS-Head/Neck	?	+	+	?	+	?	?	+
PSAQ	-	?	+	?	+	?	?	?
NAFEQ	?	+	?	?	+	?	?	?
Lip Reanimation Outcome Questionnaire	-	?	+	?	+	?	?	?
ROE/FOE/BOE/SROE	?	?	+	?	?	?	?	+
POSAS	?	?	+	?	?	?	?	?
SCI	?	+	+	?	+	?	?	?
DAS 59/24	+	+	+	?	+	?	+	+

‘+’ = sufficient, “-” = insufficient, “?” = indeterminate

Table 6: Overall combined score for each measurement property per PROM taking into account their COSM analysis for the quality of evidence presented also demonstrated.

PROM		Structural validity	Internal consistency	Reliability	Measurement error	Hypotheses testing for construct validity	Cross-cultural validity/Measurement invariance	Criterion validity
FACE-Q	Overall quality	+	+	+	?	+	?	?
	GRADE result	High	High	High	NA	High	NA	NA
POS-Head/Neck	Overall quality	±	+	+	?	+	?	?
	GRADE result	NA	Moderate	Moderate	NA	Moderate	NA	NA
PSAQ	Overall quality	-	±	+	?	+	?	?
	GRADE result	Moderate	NA	Moderate	NA	Low	NA	NA
NAFEQ	Overall	±	+	±	?	±	?	?

	quality							
	GRADE result	NA	Low	NA	NA	NA	NA	NA
Lip Reanimation Outcome Questionnaire	Overall quality	-	±	+	?	+	?	?
	GRADE result	Very low	NA	Very low	NA	Very low	NA	NA
ROE/FOE/BOE/SROE	Overall quality	?	?	+	?	?	?	?
	GRADE results	NA	NA	Low	NA	NA	NA	NA
POSAS	Overall quality	?	?	+	?	?	?	?
	GRADE result	NA	NA	High	NA	NA	NA	NA
SCI	Overall quality	?	+	+	?	+	?	?
	GRADE result	NA	High	High	NA	High	NA	NA
DAS 59/24	Overall quality	+	+	+	?	+	?	+
	GRADE	High	High	High	NA	High	High	High

	result							

‘+’ = sufficient, “-” = insufficient, “±” = inconsistent, “?” = indeterminate

For the GRADE analysis the starting point is the assumption that the evidence is of high quality. It is then do

to low, to very low based on the deduction of points for the risk of bias, inconsistency, imprecision and indire

please consult the COSMIN manual[29].

Table 7: Identified PROMs categorized according to recommendations for their future use.

Category	Explanation	PROM
A	PROMs that have the potential to be recommended as the most suitable PROM for the construct and population of interest (i.e., PROMs with evidence for sufficient content validity (any level) and at least low evidence for sufficient internal consistency)	• FACE-Q • Skin Cancer Index (SCI) • Patient Outcome of Surgery-Head/Neck (POS-Head/Neck) • Derriford Appearance Scale (DAS)
B	PROMs that may have the potential to be recommended, but further validation studies are needed (i.e., PROMS categorized not in A or C)	• Patient Scar Assessment Questionnaire (PSAQ) • Patient and Observer Scar Assessment Scale (POSAS)
C	PROMs that should not be recommended (i.e., PROMs with high quality evidence for insufficient measurement properties)	• Rhinoplasty/Facelift/Blepharoplasty/Skin Rejuvenation Outcomes Evaluation (ROE/FOE/BOE/SROE) • Nasal Appearance and Function Evaluation Questionnaire (NAFEQ) • Lip Reanimation Outcome Questionnaire

Table 8: Assessment of the relevance of items in each PROM to soft tissue facial reconstruction and post-treatment aesthetics.

PROM	Items focusing on aspects specific to soft tissue facial reconstruction	Global summary of face validity for soft tissue facial reconstruction
FACE-Q	Multiple relevant items	Good
POS-Head/Neck	Some attempt to address aspects of operation and outcomes	Average
PSAQ	Many scar questions which would be useful for assessing facial reconstruction	Good
NAFEQ	Very nasal specific with 7/14 questions relating to nasal appearance. Some could be of use	Good
Lip Reanimation Outcome Questionnaire	Aesthetic based questions but lacking on aspects of reconstruction	Average
ROE/FOE/BOE/SROE	Some questions of relevance	Average
POSAS	As with PSAQ scar questions which could be of use in a facial reconstruction PROM	Good
SCI	Two items relevant to scarring	Average
DAS 59/24	Focus in on appearance and therefore some items would be useful. Lack of specific reconstruction questions	Average

Supplementary Figure 1: Search strategy used from Medline (OVID), searched from inception until the date of search in February 2017.

1 reconstructive surgical procedures.mp. or exp Reconstructive Surgical Procedures/

2 exp Microsurgery/ or microsurgery.mp.

3 skin transplantation.mp. or exp Skin Transplantation/

4 surgical flaps.mp. or exp Surgical Flaps/

5 plastic surgery.mp. or exp Surgery, Plastic/

6 (reconstruct* or graft* or plastic or flap* or microsurg* or reanimation).mp.

7 1 or 2 or 3 or 4 or 5 or 6

8 exp Head/ or head.mp.

9 exp Neck/ or neck.mp.

10 (head or neck or face or facial or nose* or nasal or mouth or lip* or eye* or cheek* or ear or ears).mp.

11 (cervicofacial or maxillofacial).mp.

12 8 or 9 or 10 or 11

13 exp "Surveys and Questionnaires"/

14 (surveys or questionnaire*).mp.

15 patient satisfaction.mp. or exp Patient Satisfaction/

16 "quality of life".mp. or exp "Quality of Life"/

17 health status indicators.mp. or exp Health Status Indicators/

18 (patient reported outcome* or PRO or PROM).mp.

19 13 or 14 or 15 or 16 or 17 or 18

20 7 and 12 and 19

Supplementary Figure 2: Criteria for good measurement properties (psychometric quality of the study) as proposed by Terwee et al[32] and updated by Prinsen et al.[25] Figure copied from Prinsen et al.[25]

Measurement Property	Rating	Criteria
Structural validity	+	**CTT** CFA: CFI or TLI or comparable measure > 0.95 OR RMSEA < 0.06 OR SRMR < 0.08[a] **IRT/Rasch** No violation of unidimensionality[b]: CFI or TLI or comparable measure > 0.95 OR RMSEA < 0.06 OR SRMR < 0.08 AND no violation of local independence: residual correlations among the items after controlling for the dominant factor < 0.20 OR Q3's < 0.37 AND no violation of monotonicity: adequate looking graphs OR item scalability > 0.30 AND adequate model fit IRT: χ^2 > 0.001 Rasch: infit and outfit mean squares ≥ 0.5 and ≤ 1.5 OR Z-standardized values > -2 and < 2
	?	CTT: not all information for '+' reported IRT/Rasch: model fit not reported
	-	Criteria for '+' not met

Internal consistency	+	At least low evidence[c] for sufficient structural validity[d] AND Cronbach's alpha(s) ≥ 0.70 for each unidimensional scale or subscale[e]
	?	Criteria for "At least low evidence[c] for sufficient structural validity[d]" not met
	-	At least low evidence[c] for sufficient structural validity[d] AND Cronbach's alpha(s) < 0.70 for each unidimensional scale or subscale[e]
Reliability	+	ICC or weighted Kappa ≥ 0.70
	?	ICC or weighted Kappa not reported
	-	ICC or weighted Kappa < 0.70
Measurement error	+	SDC or LoA < MIC[d]
	?	MIC not defined
	-	SDC or LoA > MIC[d]
Hypotheses testing for construct validity	+	The result is in accordance with the hypothesis[f]
	?	No hypothesis is defined (by the review team)
	-	The result is not in accordance with the hypothesis[f]
Cross-cultural validity\measurement invariance	+	No important difference found between group factors (such as age, gender, language) in multiple group factor

		analysis OR no important DIF for group factors (McFadden's $R^2 < 0.02$)
	?	No multiple group factor analysis OR DIF analysis performed
	-	Important difference between group factors OR DIF was found
Criterion validity	+	Correlation with gold standard ≥ 0.70 OR AUC ≥ 0.70
	?	Not all information for '+' reported
	-	Correlation with gold standard <0.70 OR AUC < 0.70
Responsiveness	+	The result is in accordance with the hypothesis[f] OR AUC ≥ 0.70
	?	No hypothesis defined (by the review team)
	-	The result is not in accordance with the hypothesis[f] OR AUC < 0.70

AUC area under the curve, *CFA* confirmatory factor analysis, *CFI* comparative fit index, *CTT* classical test theory, *DIF* differential item functioning, *ICC* intraclass correlation coefficient, *IRT* item response theory, *LoA* limits of agreement, *MIC* minimal important change, *RMSEA* root mean square error of approximation, *SEM* standard error of measurement, *SDC* smallest detectable change, *SRMR* standardized root mean residuals, *TLI* Tucker-Lewis index