BMJ Open

International multiphase mixed methods study protocol to develop a cross-cultural patient-reported outcome instrument for children and young adults with cleft lip and/or palate (CLEFT-Q)

Karen W Y Wong Riff,[1] Elena Tsangaris,[2] Tim Goodacre,[3] Christopher R Forrest,[1] Andrea L Pusic,[4] Stefan J Cano,[5] Anne F Klassen[6]

To cite: Wong Riff KWY, Tsangaris E, Goodacre T, *et al*. International multiphase mixed methods study protocol to develop a cross-cultural patient-reported outcome instrument for children and young adults with cleft lip and/or palate (CLEFT-Q). *BMJ Open* 2017;**7**:e015467. doi:10.1136/bmjopen-2016-015467

► Prepublication history for this paper is available online. To view these files please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2016-015467).

Received 7 December 2016
Accepted 13 December 2016

For numbered affiliations see end of article.

Correspondence to
Professor Anne F Klassen;
aklass@mcmaster.ca

ABSTRACT

Introduction: Patient-reported outcome (PRO) instruments should be developed according to rigorous guidelines in order to provide clinically meaningful, scientifically sound measurement. Understanding the methodology behind instrument development informs the selection of the most appropriate tool. This mixed methods protocol describes the development of an internationally applicable PRO instrument, the CLEFT-Q, for evaluating outcomes of treatment for cleft lip and/or palate (CL/P).

Methods and analysis: The study includes three main phases that occur iteratively and interactively. In phase I, we determine what concepts are important to patients regarding their outcome. A conceptual framework for the CLEFT-Q is formed through a systematic review and an extensive international qualitative study. The systematic review ascertains what concepts have previously been measured in patients with CL/P. The qualitative study employs interpretive description and involves in-depth interviews with patients in high-income and lower-middle income countries. Preliminary items are generated from the qualitative data. Preliminary scales are then created for each theme in the framework. Cognitive debriefing interviews and expert clinician input are used to refine the scales in an iterative process. In phase II, the preliminary scales are administered to a large international group of patients with CL/P. The modern psychometric method of Rasch Measurement Theory analysis is employed to define the measurement characteristics. The preliminary scales are shortened based on these results. In phase III, further tests assess reliability, validity and responsiveness of the instrument.

Ethics and dissemination: The study is approved by Research Ethics Boards for each participating site. Findings from this study will be published in open access peer-reviewed journals and presented at national and international conferences. Integrated knowledge translation is employed to engage stakeholders from the outset of the study. Successful execution of the CLEFT-Q will result in an internationally applicable PRO instrument for children and young adults with CL/P.

Strengths and limitations of this study

- Multicentre, international study that includes patients in high-income and lower-middle income countries will ensure the CLEFT-Q is internationally applicable.
- Extensive qualitative component of the study will ensure content validity of the CLEFT-Q.
- Adherence to rigorous guidelines of instrument development and use of modern psychometric methods will make the CLEFT-Q as scientifically sound and clinically relevant as possible.
- The scope of the study, which includes participants from high-income and lower-middle income countries, necessitates a long time frame to completion.
- The CLEFT-Q field-test will not include children with CL/P aged under 8 years.

INTRODUCTION

Patient-reported outcomes (PROs) are increasingly important in the assessment of treatment effectiveness.[1 2] If PRO data are to be used to drive quality improvement and treatment decisions, PROs should be evaluated in a scientifically sound manner using a PRO instrument developed according to rigorous guidelines.[3] The methodology behind the development or 'validation' of an instrument can be complex. A clear description and understanding of the methods can

help to inform researchers selecting an appropriate PRO instrument for their target patient population.

Cleft lip and/or palate (CL/P) is the most common congenital craniofacial anomaly, with 7.94 cases per 10 000 live births annually.[3] The condition affects individuals worldwide and impacts an individual's appearance, dentition, hearing and speech. Treatment protocols vary widely, within and between countries.[4 5] Observer-reported or clinician-reported outcomes form the majority of clinical outcome assessments (COAs) to date.[6–8] However, the goal of treatment of CL/P is to improve the patient's physical, psychological and social health, all of which are difficult to evaluate accurately with observer-reported or clinician-reported outcomes. Measuring these outcomes requires the patient perspective, but there is currently no comprehensive, specific PRO instrument for patients with CL/P available.[9]

Beyond the scope of CL/P, few scales exist that measure appraisal of appearance from the patient perspective.[10] Congenital anomalies, trauma and other benign and malignant conditions can cause facial or other differences that are stigmatising and may lead to social isolation. The treatment of these conditions addresses form and function, yet the outcomes of treatment cannot be measured without appropriate PRO instruments that evaluate these concerns specifically and directly. The current study begins to fill this gap in measurement of appraisal of appearance from the patient perspective.

Many clinical conditions are prevalent around the world in high-income as well as low-income and middle-income countries. Multinational studies are increasingly common, and PROs are frequently used as primary or secondary end points. The Consolidated Standards of Reporting Trials (CONSORT) recommendations for reporting randomised controlled trials have included a PRO extension to guide PRO reporting.[2] However, PRO instruments have typically been developed in a single language and often in a single country.[11] Few PRO instruments have been designed for use in low-income and middle-income countries.[4] While COAs such as clinician-reported or observer-reported outcomes are more easily compared between countries, it is difficult to compare PROs globally in the absence of instruments designed for global use. While guidelines exist for translation and cross-cultural adaptation of PRO instruments,[11] the optimal design would be to develop the instrument in a cross-cultural manner from the outset.

Establishing scientifically sound, cross-cultural measurement tools involves a rigorous process. The following protocol describes the methodology for an international study to develop a cross-cultural PRO instrument for children and young adults with CL/P, called the CLEFT-Q. To the best of our knowledge, the CLEFT-Q will be the first international PRO measure that evaluates appraisal of appearance in addition to quality of life and function.

METHODS AND ANALYSIS

Development of the CLEFT-Q follows the guidelines set forth by the Scientific Advisory Committee of the Medical Outcomes Trust,[12] the USA Food and Drug Administration[13] and the International Society for Pharmacoeconomics and Outcomes Research.[14 15] The aim is to develop a self-report instrument for patients 8–29 years of age that is internationally applicable, multidimensional (eg, measures a number of different concepts of interest (COIs)) and useful in clinical practice as well as in clinical audits and research.

The study employs a multiphase mixed methods approach, with an iterative combination of qualitative and quantitative inquiries.[16] Measurement properties of instruments fall into the three categories of (1) reliability, (2) validity and (3) responsiveness. The Consensus-based Standards for the Selection of Health Status Measurement Instruments (COSMIN) checklist was designed to ensure and evaluate validity and reliability in measuring health-related PROs.[17 18] Similarly, a minimum standard for PRO instruments was outlined by members of the International Society for Quality of Life Research (ISOQOL).[19] There are three main phases to developing a PRO instrument, including item generation, item reduction and psychometric evaluation, and these phases are carried out in an iterative and interactive manner as opposed to a linear progression (figure 1). These three phases ensure that the resulting instrument fulfills the minimum standards outlined by ISOQOL as well as the COSMIN criteria for reliability and validity. The components of each phase are shown in figure 2.

Phase I: what should we measure?

The aims of phase I are to establish content validity of the CLEFT-Q and to generate preliminary scales. First, a systematic review of the literature was performed to ensure that there was indeed no existing instrument available and to define what PRO instruments have been validated and used in patients with CL/P in the past.[20] A comprehensive search following PRISMA guidelines yielded 4595 citations, of which 26 studies met inclusion criteria.[20] The studies were carried out in 9 high-income countries, confirming the lack of PRO measurement in low-income and middle-income countries. Twenty-nine different PRO instruments were used in the 26 studies, and 20 measures were used only once. On the basis of these findings, a

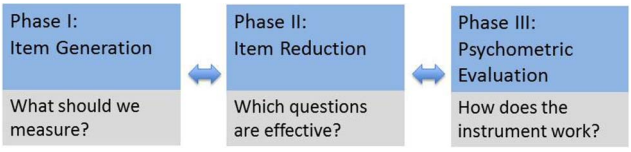

Figure 1 The phases of PRO instrument development. It is important to note that the phases can occur iteratively and interactively rather than in a linear progression. PRO, patient-reported outcome.

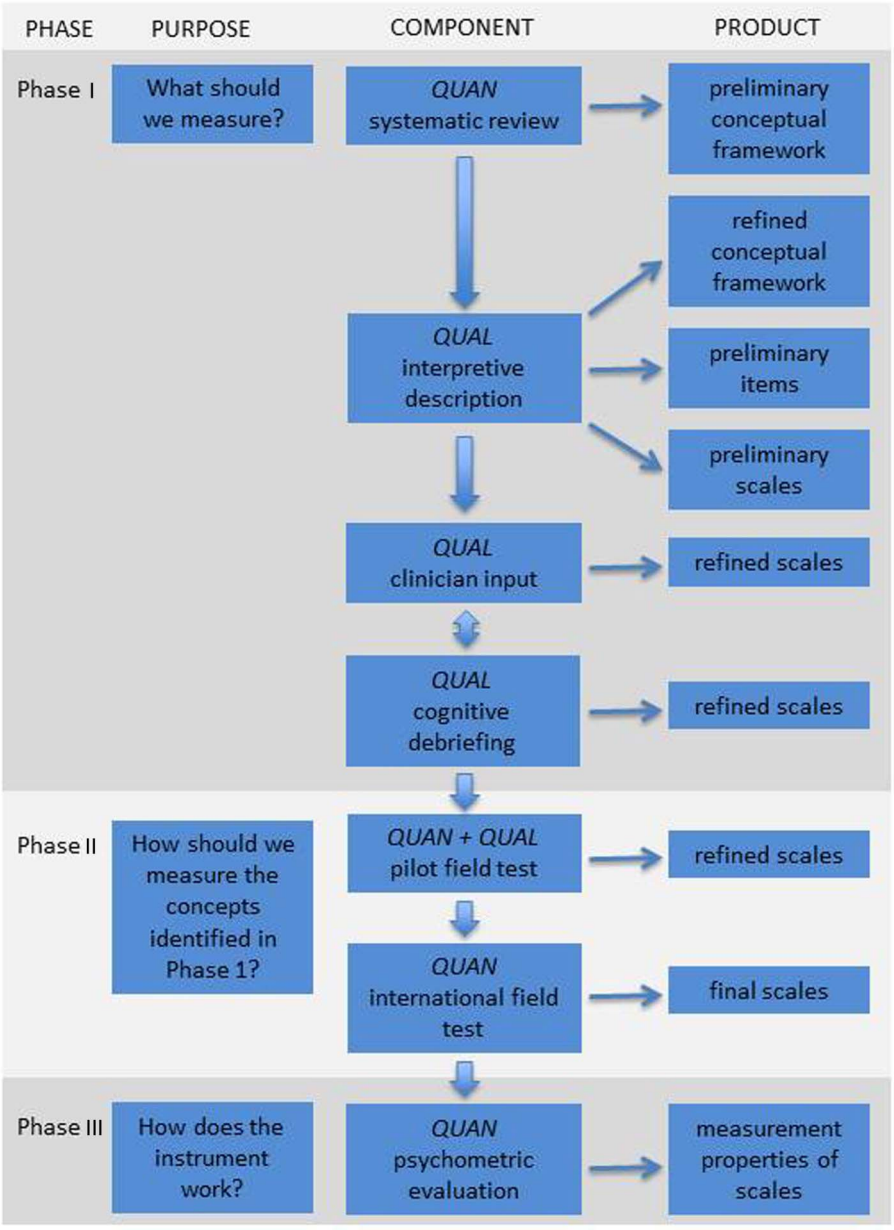

Figure 2 Flow diagram showing the multiphase mixed methods protocol for developing the CLEFT-Q. It is important to note that the process can be iterative and interactive as opposed to strictly linear. QUAN, quantitative study component; QUAL, qualitative study component.

need for a comprehensive PRO instrument for CL/P exists, and we proceeded with the current study.

Conceptual framework

The first step in phase I is to develop a conceptual framework, or 'a rationale for and description of the concepts and the populations that a measure is intended to assess and the relationship between those concepts'.[12] From the systematic review performed at the outset of the study,[20] COIs that were previously measured are mapped to create a preliminary conceptual framework.

Qualitative study

Next, a comprehensive qualitative study is carried out with participants with CL/P in high-income and lower-middle income countries. The qualitative methodology employed is Interpretive Description, which seeks to generate relevant knowledge for a clinical context presuming that there is theoretical and clinical knowledge informing the study.[21 22] For this study, the theoretical knowledge is derived from the systematic review, and clinical knowledge is derived from the team members carrying out the study. The philosophical underpinning of the qualitative study is pragmatism, meaning that the individual's understanding of a concept is of greatest importance, regardless of clinical explanations.[23]

Participants, setting and recruitment

Eliciting knowledge in high-income and lower-middle income countries allows for cultural differences to be identified from the outset, facilitating accurate targeting of the scales in subsequent phases. The participating centres in this phase of the study are in six countries (Canada, Kenya, India, Philippines, UK and USA). Recruitment takes place at cleft care centres. In the

high-income countries (Canada, UK, USA), participants are recruited either through posters in clinics (and contacted by telephone to arrange an interview), or face-to-face in the clinical setting. In the lower-middle income countries (Kenya, India, Philippines), a study team member recruits participants face-to-face in the clinical setting. Participants are eligible for inclusion if they have a diagnosis of CL/P. In the high-income countries, participants between 8–29 years of age are included. In the lower-middle income countries, participants of any age are included if they are presenting for clinical care to maximise the information gathered at these sites. In addition, parents of children with CL/P in the lower-middle income countries are invited to participate if the child prefers. This difference is important since a study team member, foreign to these countries, is present and working with a translator, which may make the child feel less comfortable if they are alone. Exclusion criteria include the inability to speak the language of the interviewer or translator in each country or a cognitive delay such that the individual cannot participate in a semistructured interview.

Sampling

Participants are purposively sampled to gain a heterogeneous sample based on age, gender and cleft type. Sampling continues until the point of saturation, when no further new concepts arise in subsequent interviews.[24]

Data collection

After obtaining written assent and/or consent as appropriate, a study team member trained in qualitative interviewing technique carries out individual, semistructured interviews that are audio-recorded, using a translator in the lower-middle income countries as needed.[24] Participant age, gender and cleft type is documented. An interview guide is developed based on the preliminary conceptual framework, providing a list of open-ended questions for the interview. The interviewer probes new concepts as they arise. As standard qualitative methods dictate, data from interviews are analysed on an ongoing basis, allowing for changes to be made to the interview guide for subsequent interviews to include new concepts that warrant further probing.

Data analysis

Interviews are transcribed verbatim. Interviews performed through a translator, which would have language in English and the target language, are again translated to English by a bilingual individual to confirm the translation. The interview data are then analysed within NVivo V.8 software (QSR International Pty, 2012) using the line-by-line approach to coding data, with constant comparison used to identify and classify the COIs identified. These concepts are then categorised into overarching domains with themes within the domains to refine the preliminary conceptual framework. Concurrent and iterative data collection and analysis are performed, allowing for changes to be made to the interview guide as new concepts arise. When no further new concepts are elicited from interviews, data collection ends and the conceptual framework is finalised. This conceptual framework represents all the COIs to patients in six different countries with CL/P.

Rigor

Rigor in the qualitative study is ensured using several strategies. One team member performs data coding, and a second team member then confirms the analysis. By performing interviews in an iterative fashion, member-checking is employed to confirm that concepts identified are indeed valuable and important to participants with CL/P. Finally, peer debriefing is used to verify data analysis between members of the study team.

Item generation

Coding of the qualitative data creates an exhaustive list of potential items to include in scales. A list of scales to be created is derived from the conceptual framework arising from the qualitative study. Each theme within the domains of the conceptual framework is turned into an individual preliminary scale. In this way, the entire suite of scales should cover all the COIs to patients with CL/P. Individual scales are populated with items generated from the patients' own language whenever possible with the lowest feasible grade reading level (Fleisch-Kincaid level). Positive or neutral wording is adopted for the items in the scales as much as possible to limit any negative effects of filling out the CLEFT-Q in the future.

Refining the preliminary scales

The final stage of phase I aims to refine the preliminary scales through an iterative process of returning to the patients to perform cognitive debriefing interviews[14] and obtaining expert multidisciplinary clinician input.

Cognitive debriefing interviews

Once the preliminary scales are formed, further semi-structured individual interviews are carried out to ensure that patients with CL/P understand the items on the scales and to confirm that no concepts are missing. Recruitment is carried out in a similar fashion to the qualitative study with participants from multiple countries to ensure cross-cultural input. Participants go through all the items on the preliminary scales with the interviewer using the 'think aloud' technique. The interviewer records items that are problematic and the reasons why these items are problematic. Cognitive debriefing interviews are carried out iteratively alongside obtaining expert clinician input as described below. Data from both sources are analysed concurrently, again allowing for progressive improvements to the scales. Cognitive debriefing interviews follow a similar strategy as the qualitative study in that interviews continue until no further issues with the items on the scales arise.

Expert clinician input

Expert clinician input is sought to ensure that no further concepts should be included in the scales. Clinicians involved in cleft care from different disciplines (nursing, orthodontics, otolaryngology, paediatrics, psychology, social work, speech-language pathology, surgery) are purposively sampled from multiple countries through the networks of the study team. Focus groups with groups of clinicians are performed in a similar fashion to the cognitive debriefing interviews. In cases where focus groups cannot be performed, individual input is sought. The interviewer goes through all the items on the scales, looking for input on any missing items or on the wording of items. Again, data are analysed concurrently with the cognitive debriefing interviews to refine the scales.

Translation

In the next phase of the study, the scales are field-tested in a large population of patients from multiple countries. The preliminary scales are translated into the necessary target languages according to guidelines set forth by the International Society for Pharmacoeconomics and Outcomes Research[25] and Mapi Research Trust.[26] Briefly, each translation is performed using two translators whose mother tongue is in the target language and are fluent English. The two translators perform independent translations of the CLEFT-Q from English to the target language. Resulting translations are then reconciled to create a single translated version. A third individual whose mother tongue is English and is fluent in the target language then translates this version back into English, and this English version is compared to the original. The group then resolves the discrepancies together. The translated versions are taken back to the patient population in further cognitive debriefing interviews to ensure that the meaning of the items, response options and instructions are the same, and that the wording is appropriate. At the end of this phase, a complete set of CLEFT-Q scales is ready to be tested in a population of patients.

Phase II: what questions are effective in measuring the concepts identified in phase I?

The next phase of developing the CLEFT-Q involves field-testing the scales in a large population of patients with CL/P to determine which items on the scales are the most effective in measuring the COIs. We employ the modern psychometric method of Rasch Measurement Theory (RMT) analysis to identify which items perform well on scales and to determine the measurement properties of the scales.[27] In order to provide rigorous measurement, the data must fit the requirements of a mathematical model, that is, the Rasch model. Briefly, RMT creates a scale where an individual is placed along the scale based on the probability that he/she answered the questions or items in a certain way. This method contrasts with classical test theory, where scores are designed for group level analyses. This difference in mathematical modelling allows RMT analysis to provide an accurate individual person estimate. A RMT scale can be conceptualised as a ruler, with an ordered arrangement or hierarchy of items from a low to high 'amount' of the construct. RMT analysis creates interval-level measurement, or a scale where the notches on the scale are evenly spaced, as opposed to ordinal-level measurement, or a scale where the notches are not necessarily evenly spaced. Interval-level measurement allows for accurate tracking of change over time.[28] In addition, RMT analysis results in a scale that provides person estimates that are independent of the sampling distribution of the items. In other words, the scale functions the same way regardless of the people that it is measuring, meaning that the same scale can be used accurately in different subsets of the target population (eg, participants in different countries, or of different ages).

Through the RMT analysis, the psychometric properties of the scale are defined. Items that are effective in measurement within the preliminary scales are then kept, and items that do not function as well in measurement or items that are identified as being redundant can be dropped. The final scales are created through this process of item reduction as described below.

Pilot field-test

A large-scale field-test of the CLEFT-Q is planned to take place in multiple countries. Since a multicentred field-test is a resource-intensive endeavour, a pilot field-test is carried out at two sites in Ontario, Canada, to identify any logistic obstacles and to perform an early preliminary RMT analysis to troubleshoot any early issues with scale performance.

Study participants

Patients with CL/P who are 8–29 years of age and who do not have a cognitive delay resulting in an inability to fill out the scales are recruited from two clinical settings in Canada. A minimum of 200 patients is required to perform the preliminary RMT analysis. Since this pilot study is meant to optimise the scales prior to the large-scale field-test, the preliminary RMT analysis may trigger further data collection and recruitment prior to finalising the field-test versions of the scales.

Data collection

Participants are asked to fill out the CLEFT-Q scales on paper and to give qualitative feedback in written format on completion. Demographic characteristics including age, gender, cleft type and stage of treatment are collected. Participants are also asked if they feel that the length of the entire CLEFT-Q is 'about right', 'too long' or 'too short'. The time to complete the scales is recorded.

Data analysis

Qualitative feedback is analysed in a similar fashion to the cognitive debriefing interviews. Details of the RMT analysis are described in further detail below. The results from the qualitative and RMT analyses are used to further refine the scales. This iterative nature to scale development optimises the likelihood that the scales will function well with minimal logistical obstacles in the ensuing large-scale field-test.

International field-test and RMT analysis

The goal of the international field-test is to gather CLEFT-Q data from a large population of patients with CL/P internationally to define which items should be included in the final scales and to examine the measurement properties of the scales.

Study participants

The international field test includes participants from 12 countries (Australia, Canada, Chile, Colombia, England, Ireland, India, Netherlands, Spain, Sweden, Turkey and USA). Centres are included based on interest and feasibility of recruiting the sample size required in a reasonable time frame. Participants with CL/P between the ages of 8 and 29 years are recruited to fill out the CLEFT-Q scales. Exclusion criteria include a cognitive delay resulting in the inability to complete the scales. Recruitment takes place either face-to-face or by mail depending on each centre's preferences. The goal is to recruit a minimum of 108 from each country; a sample size from 108 to 200 results in item calibrations that are stable within 0.5 logits (person location estimates) with a 99% CI.[29]

Data collection

The demographic characteristics collected are listed in table 1. Participants will fill out the CLEFT-Q scales either on paper or on tablets in Research Electronic Data Capture (REDCap), a secure, web-based application for electronic data capture.[30]

Data analysis

Field-test data are entered into REDCap if participants filled out the scales on paper. Completed data files are then downloaded into IBM SPSS 22.0 (IBM Corp. IBM SPSS Statistics for Windows. 22.0 ed. Armonk, New York, USA: IBM Corp., Released 2013). The SPSS file is then imported into RUMM2030, the Rasch analysis software.[31] Each scale is analysed independently. The psychometric function of each scale is examined using a number of tests and various criteria. First, the thresholds for the item response options must be ordered, meaning that a '1' on a 4-point scale must sit lower in the continuum than a '2', and so on. The RMT analysis then defines the hierarchy of items on the scale, from the 'easiest' question for a patient to endorse to the 'hardest' question. Second, 3 item fit statistics are used to evaluate whether the items in a scale work together as a set: (1) log residuals, which represent item-person interaction; (2) χ^2 values, which represent item-trait interaction and (3) item characteristic curves. Items that are not functioning well with respect to these 3 statistics will be dropped from the scales unless they represent clinically important concepts. Third, the scale must be targeted to the population. The range of the construct measured by the scale is compared to the range of the construct experienced by the population, and maximal overlap is preferable to ensure that the scale can measure the construct in the population of interest.

The next component of the analysis ensures internal consistency, which refers to the interrelatedness among items on a scale. First, the scale is tested for unidimensionality, or whether the items on the scale all measure a single construct.[32] Second, the scale is evaluated using the Person Separation Index, a measure of the precision of a person estimate, which is a corollary of reliability (Cronbach's α) in classical test theory.[33] At any stage of the analysis, scales that are not functioning appropriately can be analysed with poorly functioning items dropped. This process continues until all the above statistics are within the acceptable range.

Differential item functioning

Since the Rasch model creates a fixed ruler that is independent of the individual person estimates, differences between subgroups can be identified. Differential item functioning (DIF) occurs when one subset of the target population answers a question differently than another subset.[33] In creating an international PRO instrument for children and young adults, differences based on country and age are an important consideration in creating scientifically sound instruments. In the field-test, DIF can be identified in RUMM2030 and items that show DIF can be dropped in the item reduction phase or kept in with adjustments made to the scoring to account for the differences.

Item reduction

From the item and threshold locations, the location of each question within the field-test scale on the overall ruler can be determined. Poorly functioning items can be dropped as described above, and extra items that measure in a similar fashion (showing residual correlations in the RMT analysis) can be dropped to develop a scale with the optimal number of items. It should be noted that at some point, further dropping of items will

Table 1 Demographic characteristics collected for participants in the international field-test

Demographic characteristics	
Age	Syndromes
Gender	Other craniofacial anomalies
Cleft type	Developmental disabilities
Country	Past treatments
Student status	Current treatments
Language spoken at home	Future treatments
Adopted	

result in less precise measurement. The final decision regarding the optimal number of items depends on the distribution of the item locations as well as some clinical indication of a requirement for a certain degree of precision. Once item reduction is complete, the scales are finalised. The RMT analysis then provides a scoring table for each scale, since calculating the score on each scale is more complex than simply summing the responses to each of the individual items.

Normative data and construct validity

Once the scale scoring has been determined, scores are calculated for the field-test participants. Normative data and basic associations between scores and demographic characteristics can then be calculated using analysis of variance (ANOVA) in SPSS.

Construct validity includes the aspects of structural validity, which assesses internal relationships, hypotheses testing and cross-cultural validity. Structural validity and cross-cultural validity are addressed in the RMT analysis with unidimensionality and DIF, respectively. Hypotheses testing is used to establish whether the responses either correlate or differ in different patient groups in a way that would be expected.[34] In the CLEFT-Q, we test the following hypotheses: (1) that patients with a visible difference, that is, CL and CL/P, will have lower scores on appraisal of appearance compared to those with an invisible difference (ie, CP only); (2) that patients undergoing speech therapy or speech surgery will have lower scores on the speech scales than those not requiring any further intervention; (3) that patients requiring further treatment to the nose, lip or jaw will have lower scores on the appearance scales as well as the quality of life scales compared to those not needing any further treatment; (4) that patients who rank their appraisal of their overall appearance or speech to be higher ('like' their appearance more) on a four-point scale will have higher scores on the appearance or speech and quality of life scales, respectively and (5) that patients who are receiving psychological counselling or therapy will have lower scores on the quality of life scales. ANOVA in SPSS will be used to test these hypotheses.

Phase III: how does the instrument work?

Several components of the COSMIN checklist are addressed in phases I and II of development. Additional tests to ensure reliability, validity and responsiveness comprise phase III. All tests of the CLEFT-Q employ the finalised scales in this phase.

Reliability

Reliability includes two measurement concepts: (1) internal consistency, which is evaluated in phase II; and (2) test–retest reliability, which is evaluated in phase III. To establish test–retest reliability, a smaller group of patients complete the CLEFT-Q scales and then complete the scales again 1 week after the first administration. Scales that are reliable will have a minimum test–retest reliability of 0.70 in studies including at least 50 patients.[35]

Validity

In the COSMIN checklist, the domain of validity includes three measurement properties, that is, content validity, construct validity and criterion validity.[17 18] Content validity is addressed in phase I of the study, and construct validity is addressed in phase II.

The final component of validity is criterion validity, or the degree to which the instrument reflects the findings on a 'gold standard' instrument.[17] When an instrument is comparable to similar instruments, concurrent validity is established. While we did not identify any single instrument as comprehensive as the CLEFT-Q, the aim of this substudy is to compare the results on the CLEFT-Q to two other instruments used in the past in patients with CL/P: (1) the Child Oral Health Impact Profile (COHIP),[36 37] and (2) the CHASQ.[38] We hypothesise that CLEFT-Q scores for similar constructs will moderately correlate with the scores on these other instruments.

Responsiveness

Responsiveness evaluates the instrument's ability to detect clinically meaningful change over time. The two main methods of evaluating responsiveness include an anchor-based and a distribution-based approach. In the anchor-based approach, patient-rated, clinician-rated or condition-specific variables are used to estimate a minimally important difference (MID) for a scale.[39] The distribution-based approach estimates the MID based on the distribution of scores from a target population.[39] Techniques to evaluate responsiveness are debated in the literature.[17] RMT analysis has been shown to allow for increased detection of responsiveness.[40] We employ a variety of methods to best define responsiveness.

Study participants

Participants for the test–retest reliability and criterion validity testing are recruited simultaneously. Again, participants from 8-29 years of age are recruited from the clinical setting with the same exclusion criteria as the field-test. Since this phase requires fewer numbers of participants (50), the number of participating centers is lower than the field-test (Canada, UK, USA). To study responsiveness, participants who are undergoing either (1) orthognathic surgery, (2) rhinoplasty or (3) lip revision are recruited.

Data collection

Participants fill out the CLEFT-Q scales in addition to the COHIP and the CHASQ on tablets through REDCap. Contact information is collected, and participants are sent a link to complete the CLEFT-Q scales online 1 week later. Similar demographic data to the field-test is collected. For the responsiveness substudy, participants fill out the CLEFT-Q scales preoperatively. Contact information is collected and participants are

sent a link to the complete the CLEFT-Q scales again at least 6 months later.

Data analysis

Test–retest reliability

CLEFT-Q scores are calculated from the two separate administrations of the scales for each participant. Test–retest reliability is then calculated in SPSS.

Criterion validity

CLEFT-Q, COHIP and CHASQ scores are calculated for each participant. Scores on each of the scales are then compared using a Pearson's r correlation in SPSS.

Responsiveness

Anchor-based techniques are used to calculate the MID from the transformed Rasch scores. To support the anchor-based methods, a distribution-based approach is used. The transformed Rasch scores are compared using paired t-tests, and then an effect size and standardised response means, two indicators of change, can be calculated.[39 40] One of the strengths of RMT analysis is the ability to perform individual person level analyses for responsiveness. The tests listed above provide group level comparisons. In individual person level comparison, the significance of a person's own change can be calculated using the individual person estimates, which are associated with bespoke standard errors.[40] Using group level and individual level comparisons, the responsiveness of the CLEFT-Q can be defined as clearly as possible.

ETHICS

Throughout the study, participants may be asked to discuss or answer questions about issues that are sensitive and may experience distress as a result. To address this concern, study team members explain during the consent process that should this occur, an option to follow-up with a clinical team member will be provided. Participants are also assured that all information is kept confidential; in the qualitative phase, interviews are transcribed with no identifying data, and in the qualitative phase, identifying data are kept in a separate file at each institution.

DISSEMINATION

The intention of the study is not to directly compare different centres with respect to their outcomes. Any publications or presentations arising from this study will not identify specific centres.

An integrated knowledge translation approach is taken in this study. Collaborations with multiple sites internationally will hopefully result in increased uptake and the use of the CLEFT-Q in the future. All phases II and III results for participants from each site will be sent back to the individual sites for their own use.

Finally, results of the study will be published in open access journals as required by the granting agency. Study team members will present the results at international and national conferences.

Author affiliations
[1]Division of Plastic and Reconstructive Surgery, Department of Surgery, Hospital for Sick Children, University of Toronto, Toronto, Canada
[2]Department of Clinical Epidemiology and Biostatistics, McMaster University, Hamilton, Ontario, Canada
[3]Spires Cleft Center, Oxford Radcliffe Children's Hospital, Oxford, UK
[4]Department of Plastic and Reconstructive Surgery, Memorial Sloan-Kettering Cancer Center, New York, New York, USA
[5]Modus Outcomes, Letchworth Garden City, UK
[6]Department of Pediatrics, McMaster University, Hamilton, Ontario, Canada

Twitter Follow Anne Klassen @anneklassen and Karen Wong Riff @KWongRiffMD

Acknowledgements The authors acknowledge our partners in cleft care centres around the world for participating in data collection in various phases of the study.

Contributors KWYWR and AFK conceived of the study. KWYWR, AFK, ET, TG, CRF, ALP and SJC participated in the design of the study. All authors read and approved the final manuscript.

Funding This work is supported by the Canadian Institutes of Health Research Open Operating Grant number 299133 (to AFK, KWYWR, ALP and CRF), the Physicians' Services Incorporated Foundation Resident Research Grant (to KWYWR, AFK and CRF) and the Canadian Society of Plastic Surgeons Educational Foundation Grant (to KWYWR, AFK and CRF).

Competing interests KWYWR, ET, TG, CRF, ALP and AFK declare that they have no competing interests. SJC started Modus Outcomes, a company that designs PRO instruments, after the inception of this study.

Patient consent Obtained.

Provenance and peer review Not commissioned; peer reviewed for ethical and funding approval prior to submission.

Data sharing statement No additional data are available.

REFERENCES

1. Basch E. The missing voice of patients in drug-safety reporting. *N Engl J Med* 2010;362:865–9.
2. Calvert M, Blazeby J, Altman DG, *et al*. Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. *JAMA* 2013;309:814–22.
3. Tanaka SA, Mahabir RC, Jupiter DC, *et al*. Updating the epidemiology of cleft lip with or without cleft palate. *Plast Reconstr Surg* 2012;129:511e–18e.
4. Paltzer J, Barker E, Witt WP. Measuring the health-related quality of life (HRQoL) of young children in resource-limited settings: a review of existing measures. *Qual Life Res* 2013;22:1177–87.
5. Shaw WC, Semb G, Nelson P, *et al*. The Eurocleft project 1996–2000: overview. *J Craniomaxillofac Surg* 2001;29:131–40; discussion 41-2.
6. Semb G, Brattström V, Mølsted K, *et al*. The Eurocleft study: intercenter study of treatment outcome in patients with complete cleft lip and palate. Part 1: introduction and treatment experience. *Cleft Palate Craniofac J* 2005;42:64–8.
7. Long RE Jr, Hathaway R, Daskalogiannakis J, *et al*. The Americleft study: an inter-center study of treatment outcomes for patients with unilateral cleft lip and palate part 1. Principles and study design. *Cleft Palate Craniofac J* 2011;48:239–43.
8. Heliövaara A, Küseler A, Skaare P, *et al*. SCANDCLEFT randomised trials of primary surgery for unilateral cleft lip and palate: 6. Dental arch relationships at 5 years. *J Plast Surg Hand Surg* 2016;25:1–6.

9. Eckstein DA, Wu RL, Akinbiyi T, *et al*. Measuring quality of life in cleft lip and palate patients: currently available patient-reported outcomes measures. *Plast Reconstr Surg* 2011;128:518e–26e.
10. Wickert NM, Wong Riff KW, Mansour M, *et al*. Content Validity of Patient Reported Outcome Instruments Used with Pediatric Patients with Facial Differences: A Systematic Review. *Cleft Palate Craniofac J* 2016 In Press.
11. Wild D, Eremenco S, Mear I, *et al*. Multinational trials-recommendations on the translations required, approaches to using the same language in different countries, and the approaches to support pooling the data: the ISPOR Patient-Reported Outcomes Translation and Linguistic Validation Good Research Practices Task Force report. *Value Health* 2009;12:430–40.
12. Aaronson N, Alonso J, Burnam A, *et al*. Assessing health status and quality-of-life instruments: attributes and review criteria. *Qual Life Res* 2002;11:193–205.
13. U.S. Department of Health and Human Services Food and Drug Administration. Guidance for Industry Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. http://www.fda.gov/downloads/Drugs/Guidances/UCM193282.pdf[2].
14. Patrick DL, Burke LB, Gwaltney CJ, *et al*. Content validity—establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO Good Research Practices Task Force report: part 2—assessing respondent understanding. *Value Health* 2011;14:978–88.
15. Patrick DL, Burke LB, Gwaltney CJ, *et al*. Content validity—establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: part 1—eliciting concepts for a new PRO instrument. *Value Health* 2011;14:967–77.
16. Creswell JW, Plano Clark VL. *Designing and conducting mixed methods research*. 2nd edn. Los Angeles (CA): Sage, 2011.
17. Mokkink LB, Terwee CB, Knol DL, *et al*. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. *BMC Med Res Methodol* 2010;10:22.
18. Mokkink LB, Terwee CB, Patrick DL, *et al*. The COSMIN checklist for assessing the methodological quality of studies on measurement properties of health status measurement instruments: an international Delphi study. *Qual Life Res* 2010;19:539–49.
19. Reeve BB, Wyrwich KW, Wu AW, *et al*. ISOQOL recommends minimum standards for patient-reported outcome measures used in patient-centered outcomes and comparative effectiveness research. *Qual Life Res* 2013;22:1889–905.
20. Klassen AF, Tsangaris E, Forrest CR, *et al*. Quality of life of children treated for cleft lip and/or palate: a systematic review. *J Plast Reconstr Aesthet Surg* 2012;65:547–57.
21. Thorne S, Kirkham SR, MacDonald-Emes J. Interpretive description: a noncategorical qualitative alternative for developing nursing knowledge. *Res Nurs Health* 1997;20:169–77.
22. Thorne SE. *Interpretive description*. Developing qualitative inquiry, vol 2. Walnut Creek (CA): Left Coast Press, 2008.
23. Welford C, Murphy K, Casey D. Demystifying nursing research terminology. Part 1. *Nurse Res* 2011;18:38–43.
24. Sandelowski M. Theoretical saturation. In: Given LM, ed. *The sage encyclopedia of qualitative methods*. Thousand Oaks (CA): Sage, 2008:875–6.
25. Wild D, Grove A, Martin M, *et al*., ISPOR Task Force for Translation and Cultural Adaptation. Principles of Good Practice for the Translation and Cultural Adaptation Process for Patient-Reported Outcomes (PRO) Measures: report of the ISPOR Task Force for Translation and Cultural Adaptation. *Value Health* 2005;8:94–104.
26. Acquadro C, Conway K, Giroudet C, *et al*. *Linguistic validation manual for health outcome assessments*. Lyon: Mapi Institute, 2012.
27. Rasch G. *Probabilistic models for some intelligence and attainment tests*. vol 1. Copenhagen: Studies in Mathematical Psychology, 1960.
28. Cano SJ, Hobart JC. The problem with health measurement. *Patient Prefer Adherence* 2011;5:279–90.
29. Linacre JM. Sample size and item calibration (or person measure) stability. *Rasch Meas Trans* 1994;7:328.
30. Harris PA, Taylor R, Thielke R, *et al*. Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform* 2009;42:377–81.
31. RUMM Lab. http://www.rummlab.com.au/ (accessed 10 Oct 2014).
32. Fayers PM, Hand DJ, Bjordal K, *et al*. Causal indicators in quality of life research. *Qual Life Res* 1997;6:393–406.
33. Andrich D. *Rasch models for measurement*. Sage university papers series quantitative applications in the social sciences, vol 07-068. Newbury Park (CA): Sage, 1988.
34. Strauss ME, Smith GT. Construct validity: advances in theory and methodology. *Annu Rev Clin Psychol* 2009;5:1–25.
35. Giraudeau B, Mary JY. Planning a reproducibility study: how many subjects and how many replicates per subject for an expected width of the 95 per cent confidence interval of the intraclass correlation coefficient. *Stat Med* 2001;20:3205–14.
36. Broder HL, Wilson-Genderson M. Reliability and convergent and discriminant validity of the Child Oral Health Impact Profile (COHIP Child's version). *Community Dent Oral Epidemiol* 2007;35(Suppl 1):20–31.
37. Broder HL, Wilson-Genderson M, Sischo L. Reliability and validity testing for the Child Oral Health Impact Profile-Reduced (COHIP-SF 19). *J Public Health Dent* 2012;72:302–12.
38. Emerson M, Spencer-Bowdage S, Bates A. Relationships between self-esteem, social experiences and satisfaction with appearance: standardisation and construct validation of two cleft audit measures. *Presented at the Annual Scientific Conference of the Craniofacial Society of Great Britain and Ireland*, Bath, UK, 2004.
39. Revicki D, Hays RD, Cella D, *et al*. Recommended methods for determining responsiveness and minimally important differences for patient-reported outcomes. *J Clin Epidemiol* 2008;61:102–9.
40. Hobart JC, Cano SJ, Thompson AJ. Effect sizes can be misleading: is it time to change the way we measure change? *J Neurol Neurosurg Psychiatry* 2010;81:1044–8.