CHALLENGES IN OUTCOME-BASED CLEFT CARE

Inge Apon

Challenges in Outcome-based Cleft Care

Inge Apon

Inge Apon

Challenges in outcome-based cleft care

ISBN:	978-94-6483-374-4
Cover design:	Arina van Londen, Publiss | www.publiss.nl
Lay-out:	Publiss | www.publiss.nl
Print:	Ridderprint | www.ridderprint.nl

Challenges in Outcome-based Cleft Care

Uitdagingen in uitkomstgerichte schisiszorg

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de rector magnificus

Prof. dr. A.L. Bredenoord
en volgens het besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
dinsdag 10 oktober 2023 om 10:30 uur

door

Inge Apon
geboren te Sliedrecht

Erasmus University Rotterdam

Promotiecommissie

Promotor	Prof. dr. E.B. Wolvius
Overige leden	Prof. dr. H.F. Lingsma
	Prof. dr. J. Busschbach
	Prof. dr. A.B. Mink van der Molen
Copromotoren	Dr. S.L. Versnel
	Dr. N. van Leeuwen
	Dr. M.J. Koudstaal

List of Commonly Used Abbreviations

CAT	Computerized adaptive test
CL	Cleft lip only
CLA	Cleft lip and alveolus
CLP	Cleft lip and palate
COHIP-OSS	Child Oral Health Impact Profile - Oral Symptoms Scale
CP	Cleft palate only
CROMs	Clinical-reported outcome measures
EHR	Electronic health records
GRM	Graded response model
HRQOL	Health related quality of life
ICHOM	International Consortium for Health Outcomes Measurement
ICS	Intelligibility in Context Scale
IRT	Item response theory
MAE	Mean absolute error
NOSE	Nasal Obstruction Symptom Evaluation
PRO	Patient-reported outcome
PROMs	Patient-reported outcome measures
RE-AIM	Reach, effectiveness, adoption, implementation and maintenance
RMSE	Root mean squared error
RMT	Rasch measurement theory
VBHC	Value-based healthcare
WHO	World Health Organization

Aan mijn ouders en broer

“I did it my way”
- Frank Sinatra

Table of Contents

Chapter 1

General Introduction

Cleft lip and palate

Cleft lip and/or palate is the most common congenital craniofacial anomaly with an average prevalence of approximately 11 per 10,000 live births per year in The Netherlands.[1,2] Worldwide, the overall prevalence rates vary between 8 and 10 per 10,000 live births.[3-5] The anomaly results from a non-completed fusion of the lip, jaw (alveolus), palate, or a combination of these, between the fourth and twelfth week of gestation.[6] Due to the possibility of a plethora of cleft type combinations, various classifications have been proposed.[6,7] In this thesis, a classification distinguishing four main cleft categories will be utilized: 1) cleft lip only, 2) cleft palate only, 3) cleft lip and palate, and 4) cleft lip and alveolus (**Figure 1**).[8] According to literature, cleft palate has a female predominance, while cleft lip and palate, and cleft lip occur more frequently in males.[9] Further, a cleft can present unilaterally or bilaterally. The unilateral cleft presents itself twice as much on the left side as on the right side and is nine times more frequent than bilateral clefts.[6]

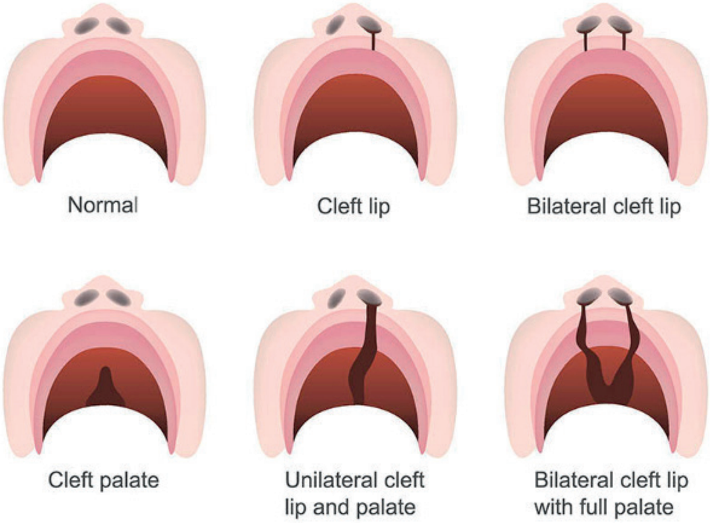

Normal
Cleft lip
Bilateral cleft lip
Cleft palate
Unilateral cleft lip and palate
Bilateral cleft lip with full palate

Figure 1 Various types of cleft lip and/or palate. Source: *https://www.healthdirect.gov.au/cleft-lip-and-cleft-palate.*

The occurrence of orofacial clefts is multifactorial and subject to various environmental and genetic influences.[6] For example, maternal exposure to alcohol or tobacco, poor prenatal nutrition, vitamin B6 and folic acid deficiency, anti-convulsant use, and maternal obesity are all associated with an increased risk of a new born with a cleft, although the degree of these associations remain unclear.[6,10-15]

With 71 percent, the isolated, non-syndromic cleft is the most common form, while the other 29 percent of the cleft malformations are part of a (genetic) syndrome and associated with other craniofacial differences, musculoskeletal, cardiovascular, or central nervous system defects.[16,17] To detect congenital disorders in a timely fashion, an ultrasound anomaly scan is offered to all pregnant women at the twentieth week of their pregnancy. The introduction of this scan has resulted in an improvement of the prenatal detection rate of orofacial clefts in The Netherlands.[18,19]

After the prenatal diagnosis of cleft lip and palate, parents are referred to a cleft centre with specialized healthcare professionals to receive detailed information on the care and prognosis of the condition.[18,20] Because of the heterogeneous character of this anomaly, patients can present with multiple functional and aesthetic problems with varying severity. In the first years of life, growth in the cleft patient can be affected by the inability to create suction for adequate feeding.[6,21] Later in life, feeding difficulties can be caused by mastication problems due to dental and orthodontic disturbances such as congenitally missing teeth, malocclusion, clefting of the alveolus and disturbed growth of the maxilla. In addition, articulation disorders and velopharyngeal insufficiency are frequently observed and hearing problems can occur due to chronic ear infections because of Eustachian tube dysfunction.[6,21] Furthermore, deformities of the nose, lip, teeth, and jaws can be visible and aesthetically disturbing.

The importance of normal speech and facial appearance for successful socialization cannot be underestimated.[6] In fact, psychosocial impairment has been commonly reported in patients with cleft lip and palate. The facial difference is associated with lower levels of self-esteem[22-24] and is, together with speech aberrations, a common source for teasing or bullying.[22,25-28] Also, being misunderstood because of speech difficulties can create frustration and feelings of embarrassment, and hearing difficulties can leave children feeling isolated.[22,25,29] Therefore, from birth until young adulthood, patients undergo multiple surgical procedures and non-surgical treatments by a multidisciplinary cleft team consisting of a maxillofacial surgeon, plastic surgeon, ear nose and throat specialist, speech therapist, dentist and orthodontist, psychologist and social worker, geneticist, and nurse specialist.

The choice, timing and order of treatments are of potential influence on the patient's functional and aesthetic outcomes. However, treatment protocols for cleft lip and palate show large variations between centres. Each cleft unit, even within a small country as The

Netherlands, has its own guidelines based on a variety of literature and their own clinical experiences. A lack of consensus is best discernible in the optimal timing of the surgical closure of the hard palate. When the best possible speech is pursued, it is advised to close the hard palate together with the soft palate in the first year of life. In contrast, when the maximum possible growth of the upper jaw (maxilla) is pursued, postponing the closure of the hard palate to a later age is advised.[30] Reviewing the treatment protocols of ten Dutch cleft teams, the timing of palatal closure ranges from 3 months up to 12 years of age.[30,31]

These large discrepancies between treatment protocols were not only detected during the development of the Dutch guideline for cleft care[30], but were also found by the Eurocleft project, an extensive inter-centre comparison study within Europe.[32] This study revealed that among 201 cleft centres, 194 different treatment protocols were being followed for unilateral clefts.[32] The follow-up Americleft study, consisting of a study design comparable with the Eurocleft project, concluded that reaching more favourable outcomes was associated with simpler, less burdensome treatment protocols.[33] This project was a promising first step towards a major inter-centre collaborative research endeavour to gain more insights in the cause-effect relationships of favourable versus unfavourable outcomes with the complex aspects of cleft care and varying patient characteristics.[33] These inter-centre research efforts are supported by a report from the World Health Organization (WHO) on ‘Global strategies to reduce the healthcare burden of craniofacial anomalies’ stating that “great confusion surrounds the optimal management for even the most common conditions”.[5] There are numerous worldwide registries with data on cleft deformities, however the validity and comparability of these data is challenging due to the heterogeneity of the cleft patient population, a lack of uniform definitions of cleft subtypes and case-mix variables, and diversity in treatment timing, method of data collection and follow-up time. Since it is not feasible to change treatment strategies worldwide to one and the same protocol, and as it is unknown which treatment protocol would lead to best results in the long-term, we need to consider another approach to decide on best treatment practices. Consequently, the WHO stated that there is an “urgent need to create collaborative groups in order to develop and standardize outcome measures, and there is an especially urgent need for work on psychosocial and quality of life measures, and economic outcomes”.[5]

Therefore, we have to define and establish an international consensus on the outcomes essential for determining the quality of cleft care. This should be done from both a healthcare provider’s perspective as from a patient’s point of view. In the end, uniform outcome collection would allow cleft teams to evaluate their own quality of care and stimulate local quality improvement endeavours.[32] Further, it provides cleft teams the opportunity to learn from each other by comparing outcomes ('benchmarking'), perform research on cause-effect relationships, and ultimately define best treatment strategies for obtaining the best possible outcomes for patients with a cleft.

Value-based healthcare

A possible solution for the worldwide improvement of quality of care was found in the theory of value-based healthcare (VBHC), as being introduced by the book "Redefining Health Care", written by professors Porter and Teisberg.[34] This publication aimed to initiate a paradigm shift in healthcare from focusing on volume of services delivered towards creating value for patients.[35] Achieving high value should become the new overarching goal of healthcare delivery, in which value is defined as health outcomes relative to costs.[35,36] The theory states that value should always be defined around the patient and should cover the full cycle of care for a patient's medical condition.[37] Therefore, it is proposed to organize care for one specific medical condition into integrated practice units (IPUs), including all necessary specialized professionals[37], and to move to a 'bundled payment' reimbursement system.[38] Another main aspect of VBHC includes the universal measurement of outcomes. In 2008, Porter wrote in the *Annals of Surgery* "Only by measuring patient outcomes over the cycle of care for each medical condition will it be possible to optimize overall value for the patient and to drive value improvement."[37]

Within the VBHC framework, outcome measurement should not be limited to only one outcome, but it should be a hierarchy of the various aspects of care. **Figure 2** shows the three levels of outcome metrics that should be included in a comprehensive set of outcomes measurement for any medical condition.[37] The first level includes outcomes directly related to the health status of a patient, such as survival, quality of life and physical functioning. The second level involves treatment-related outcomes, such as recovery speed, complications and re-occurrences. The third outcomes level captures the sustainability and long-term consequences of care.[37] It is theorized that improvements in level 2 and 3 will not only improve patient's health, but will also lead to a reduction in costs. A faster recovery, a minimization of complications and a decrease in re-occurrence rate should lead to better health and less care utilization in the long run.[37]

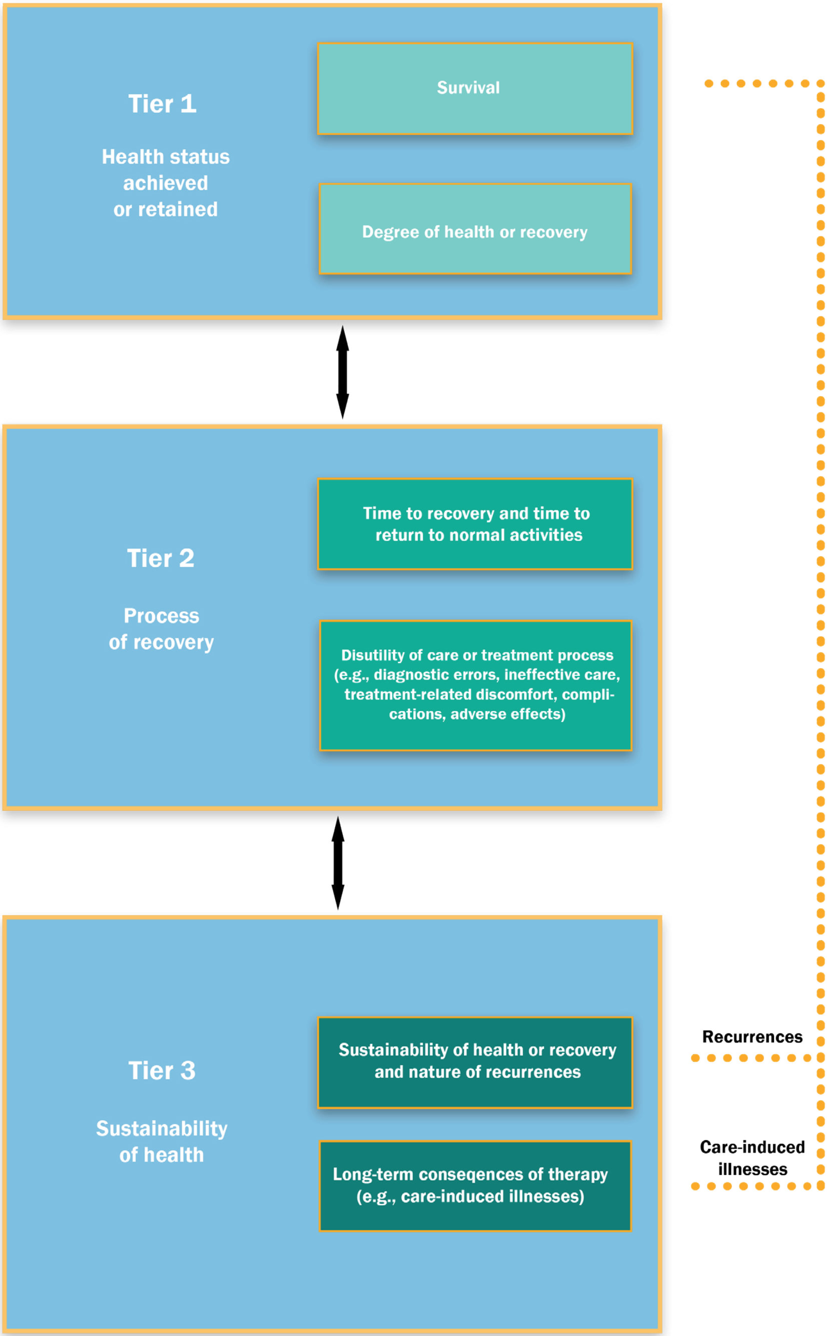

Tier 1
Health status achieved or retained
Survival
Degree of health or recovery
Tier 2
Process of recovery
Time to recovery and time to return to normal activities
Disutility of care or treatment process (e.g., diagnostic errors, ineffective care, treatment-related discomfort, compli-cations, adverse effects)
Tier 3
Sustainability of health
Sustainability of health or recovery and nature of recurrences
Long-term conseqences of therapy (e.g., care-induced illnesses)
Recurrences
Care-induced illnesses

Figure 2 Outcome measures hierarchy according to Porter. Source: Porter M.E. *What is Value in Health Care?* NEJM 2010.[35]

To accelerate the transition into VBHC delivery systems worldwide and to support the measurement of outcomes, the International Consortium for Health Outcomes Measurement (ICHOM) was founded by three institutions: the Harvard Institute for Strategy and Competitiveness, The Boston Consulting Group and the Karolinska Institute in Stockholm.[39] By convening various groups of experts and patient representatives, they act as a steward in developing standardized outcome measurement frameworks, named Standard Sets, for various diseases to use internationally and across cultures.[39] So far, 40 Standard Sets have been developed for a wide variety of medical conditions, such as breast cancer, heart failure, stroke, and craniofacial differences such as cleft lip and palate and craniofacial microsomia. A Standard Set is intended to provide a uniform foundation for a “learning healthcare system” to support the continuous quality improvement of care.[40]

Since 2018, the Dutch Ministry of Health, Welfare and Sport also encourages hospitals to move towards an outcome-based healthcare system to improve patient’s quality of life, job satisfaction for healthcare providers, and to support shared decision-making.[41] The Ministry’s goal is to provide insight and access to relevant outcome information for at least half of the disease burden in curative care, from which cleft is one of them.[41,42] The transition plan contains four work streams: 1) more insight into outcomes, 2) more shared decision-making, 3) more outcome-based organisation and payment, and 4) better access to relevant and up-to-date outcomes information.[41] The use of the ICHOM Standard Sets as the foundation for the national outcome registration is accepted and encouraged by the Dutch Ministry.[43]

The ICHOM Standard Set for Cleft Lip and Palate

In 2014, the ICHOM Cleft Lip and Palate working group gathered to develop a Standard Set for the comprehensive appraisal of cleft care.[40] The working group consisted of patients and parents and internationally recognized cleft experts from multiple disciplines related to cleft care to establish an international consensus on the outcomes that should be measured routinely as part of clinical practice. The Standard Set describes a method for classifying patients and defines the measurements, timings and registration of each outcome. Outcomes are collected during the first encounter with the cleft team, at the age of 3 months, and at 5, 8, 12 and 22 years of age.[8] Data collection includes multiple case-mix variables and clinical indicators to enable meaningful case-mix adjustment for comparison of outcomes globally[8], which are presented in **Table 1**.

Demographic factors	Baseline clinical status	Treatment	Other	Burden of care to patient	Clinical indicators
Age at first encounter	Phenotype	Operation	Adoption	Post-operative complications	Body weight
Sex	Syndrome/ genetic diagnosis	Loss to follow-up	Language	Death	Dental health (Decayed, Missing and Filled Teeth index)
Age	Comorbidities	Transferral of care	Insurance	Hospital stay	Occlusion (Overjet assessment)
Parent education			Distance to hospital	Oronasal fistula	Velopharyngeal competence
				Repeated speech surgery	Otologic Health questions
				Number of interventions for palate	Pure Tone Audiometry
				Number of interventions for lip & nose	
				Number of interventions for alveolus	

Table 1 Overview of case-mix variables and clinical indicators as defined by the ICHOM Standard Set for cleft lip and palate.[8]

Moreover, the patient's perspective on health is included by the frequent use of patient-reported outcome measures (PROMs).[8,40] A PROM provides a report on the status of a patient's health condition that comes directly from the patient (or from the parents, then it is referred to as parent- or proxy-reported outcome measures), without interpretation of the patient's response by the clinician or anyone else.[44] A systematic review by Eckstein et al.[45] has identified over forty patient-reported outcome instruments to measure quality of life or satisfaction in patients with cleft lip and palate.[45] However, only five instruments were validated in a patient population with cleft lip and palate, and none of these measures were initially developed with a focus on cleft lip and palate.[45-47] In response to this finding, the Q-Portfolio group developed the CLEFT-Q, a specific and unique PRO instrument to measure outcomes that matter most to children and young adults with cleft lip and palate.[46,47] The CLEFT-Q is now responsible for eighty percent of the PROMs in the ICHOM Standard Set for cleft lip and palate.[8] All outcome domains and patient-reported outcome instruments, including number of items, patient population, timing and example questions, are presented in **Table 2**.

Domain	Outcome instrument	Items	Cleft type	Timing (years)	Examples of questions	Responses
Facial appearance	CLEFT-Q face	9	CL, CP, CLA, CLP	8, 12, 22	How much do you like… - …how your face looks when you look your best? - …how your face looks when you smile?	Not at all, a little bit, quite a bit, very much
	CLEFT-Q teeth	8	CL, CP, CLA, CLP	8, 12, 22	How much do you like… - …the size of your teeth? - …how straight your teeth look?	Not at all, a little bit, quite a bit, very much
	CLEFT-Q jaws	7	CL, CP, CLA, CLP	12, 22	How much do you like… - …the size of your jaws? - …how your jaws look from the side?	Not at all, a little bit, quite a bit, very much
Psychosocial function	CLEFT-Q psychological	10	CL, CP, CLA, CLP	12	How do you feel? - I am happy with my life. - I feel confident.	Never, sometimes, often, always
	CLEFT-Q social	10	CL, CP, CLA, CLP	8, 22	How is your social life? - I have fun with friends. - I feel like I fit in.	Never, sometimes, often, always
	CLEFT-Q school	10	CL, CP, CLA, CLP	12	How is your school life? - I like seeing my friends at school - I feel safe at school (not bullied)	Never, sometimes, often, always
Speech	CLEFT-Q speech distress	10	CP, CLP	12, 22	How do you feel about speaking? - I get teased about my speech. - I get upset when I need to repeat myself	Always, sometimes, never
	CLEFT-Q speech function	12	CP, CLP	12, 22	How is your speech? - It's hard for my family to understand my speech. - I need to concentrate to speak well.	Always, sometimes, never
	Intelligibility in Context Scale (ICS)	7	CP, CLP	12	Think about your child's speech intelligibility over the past month and identify the degree of understanding. - Do you understand your child? - Do immediate members of your family understand your child?	Never, rarely, sometimes, usually, always

Domain	Outcome instrument	Items	Cleft type	Timing (years)	Examples of questions	Responses
Facial function	CLEFT-Q eating and drinking	9	CP, CLA, CLP	8, 12, 22	How is your eating and drinking? - Food falls out of my mouth when I eat. - Food or drinks go up my nose.	Always, often, sometimes, never
	Nasal Obstruction Symptom Evaluation (NOSE)	5	CL, CP, CLA, CLP	8, 12	How much of a problem were the following conditions for you? - …Nasal blockage or obstruction. - …Trouble breathing through my nose.	No problem, mild, moderate, fairly bad, severe problem
Oral health	Child Oral Health Impact Profile – Oral Symptoms Scale (COHIP-OSS)	5	CP, CLA, CLP	8, 12	In the past 3 months, have you… - …Had pain in your teeth? - …Had bleeding gums?	Never, almost never, sometimes, fairly often, almost all of the time

Table 2 Overview of the PROMs in the ICHOM Standard Set for Cleft Lip and Palate. CL = cleft lip, CP = cleft palate, CLA = cleft lip and alveolus, CLP = cleft lip and palate. Source: Adapted from Apon *et al.*[48] *(this thesis).*

Measurement properties

Historically, quality and health measurement mainly focused on process measures and clinical outcomes such as mortality, complication rate or laboratory test results, as these measures often emerged from clinical trials with a primarily clinical endpoint.[39] Since the 1970, the focus of health care evaluation shifted towards measuring broader, more complex and subjective aspects of health, due to a change in the WHO-definition of health as a "complete state of physical, mental and social well-being and not merely the absence of disease or infirmity".[44,49] In the 1980s, patient report rating scales, now known as patient-reported outcome measures (PROMs), were introduced and have been increasingly used in research, policy-making and quality improvements.[49] PROMs can be classified in the following four scale types:

- **Generic outcome measures**: these measures capture health aspects important to many disease populations, allowing direct comparison of various patient populations, and thus providing the possibility to make policy decisions across a variety of diseases. However, generic measures might be limited in addressing important outcome aspects of a particular disease, and they might not be sensitive enough to detect changes in outcome over time.[49-51]
- **Disease-/condition-specific measures**: these measures include items that are directly relevant to the condition and the scales are most likely shorter and more appropriate, which helps reducing patient burden and increase the acceptability of outcome measurement. Also, these measures are generally more sensitive in detecting changes in treatment outcomes.[49-51]
- **Site-specific measures**: these measures focus on health problems in a specific part of the body, for example the CLEFT-Q teeth, jaws or face, and thus are shorter and appear to be less burdensome to patients.[49-51]
- **Dimension-specific measures**: these measures are a general evaluation of one specific aspect ("domain") of health, for example the CLEFT-Q psychological or speech scale, and thus providing more detailed information on the area of concern. These measures can be applicable across different patient populations and treatments.[49-51] Dimension-specific measures can be unidimensional by measuring only one domain, or multidimensional by measuring multiple domains.[44]

Since there are many measurement instruments available, choosing the most appropriate instrument for a given situation might be challenging. The aspect of health, or the 'what' is being measured, is referred to as construct, and sometimes as domain or concept.[52]

PROMs can be administered with pen and paper, or with the help of an electronic device such as a tablet, mobile phone or personal computer. Further, PROMs can differ in language, the number of questions ('items'), the wording of questions (negative, positive), and the scoring system. As a result, the quality between measurement instruments may vary considerably. Extracting the construct and assessing the following measurement properties responsible for the psychometric performance of instruments might help to find an adequate outcome measure for research or clinical practice:[44,52-54]

- **Validity:** the degree to which an outcome measurement instrument measures the construct it purports to measure.[44,54] There are three main types of validity. *Content validity* is the degree to which the content of the instrument is an adequate reflection of the construct to be measured. *Construct validity* is the degree to which the scores of the instrument are in agreement with hypotheses, and *criterion validity* is the degree to which the scores of the instrument are an adequate reflection of the 'gold standard'.[44,54]
- **Reliability**: the degree to which the measurement is free from measurement error and the scores for patients who have not changed are the same for repeated measurement under varying conditions.[44,54]
- **Responsiveness:** the ability to detect change over time in the construct to be measured.[54]
- **Interpretability**: the degree to which one can assign a qualitative meaning to an instrument's quantitative scores or change in scores.[54]

The use of PROMs in cleft care

As every change in work routine comes with challenges, the introduction of the ICHOM Standard Set for cleft is no different. Transforming care into value-based healthcare and the use of outcome instruments in cleft practice is a relatively new phenomenon, and therefore in need of an evaluation to determine what it delivers and where and how to improve. In this thesis, we discern two levels in which we face challenges: firstly, the measurement of outcomes, and secondly, the implementation of the Standard Set in practice.

Measurement challenges

The development of an outcomes set is made up of two important components: decisions about *what* to measure (domains or constructs) and then decisions about *how* and *when* to measure each of the chosen domains.[55] In this process, it is essential to first define the scope and applicability of the set, i.e. define the patient population or condition, the setting in which measurement is going to take place, the geographical scope and the relevant stakeholders.[56,57] For decision-making on measurement instruments, the assessment of the measurement properties and the feasibility of the instruments is of importance.[56] For the assessment of feasibility, the practical use of an instrument should be evaluated, for example are there translations available, what is the length of the instrument, and what are costs of using the instruments in practice (i.e. licenses).[56]

In 2016, consensus within the ICHOM working group for cleft lip and palate was reached on the core outcome domains to measure, and on the instruments to measure the relevant outcomes.[40] Even though all these outcome instruments have undergone some degree of validity testing during their development phase, it is required to conduct research on the instruments' performances after implementation in a different setting and to a new patient population.[44] Furthermore, since the use of patient-reported outcome instruments as part of an outcomes framework in clinical cleft setting is a new phenomenon, knowledge on the various aspects of outcome measurement is still limited. Especially in the light of informing clinical-decision making, and facilitating future learning and quality improvement initiatives, it is necessary to verify that each of the included instruments is robust enough to accurately and reliably appraise the corresponding outcome domains. To ensure the feasibility and sustainment of the implementation of the Standard Set, the outcome measures need to be appraised for undue burden for patients and clinicians so that redundant measures could potentially be de-implemented or replaced and compliance can be improved.[58]

In addition, to limit patient burden, increase compliance and stimulate the uptake of the Standard Set in practice, the use of a so-called Computerized Adaptive Test (CAT) is proposed.[58] A CAT version could reduce the number of questions in a scale, while maintaining the same degree of accuracy as the full-length questionnaire. Recently, a CAT version for the CLEFT-Q scales is developed with the use of simulated patient data.[59,60] Since the CLEFT-Q scales with a total of 85 items form a large proportion of the PROMs in the Standard Set, the CLEFT-Q CAT could be a promising instrument to enhance the uptake of PROMs in clinical practice. Before actual implementation of the CAT, more information should be gathered about the external validity of the CLEFT-Q CAT when used in a real patient population and in a clinical setting. Also, assessing the user-experiences of patients and healthcare providers with the CAT is of great importance.

Implementation challenges

Over the past years, the cleft teams of the Erasmus University Medical Center (NL), Boston children's Hospital (USA), Duke Children's Hospital (USA), and Karolinska University Hospital (SE) have completely implemented the Standard Set in their routine clinical practice. At multiple other institutions, nationally and internationally, implementation efforts are ongoing, but often challenged by a lack of a defined strategy or clear understanding of conditions that promote or hinder routine outcome measurement.[61] The long, multidisciplinary and complex cleft trajectory serves as an additional challenge. So far, information on implementing outcome sets in clinical practice is limited or mainly focused on specific healthcare areas such as palliative care or orthopedic surgery.[61-65] Therefore, knowledge on the implementation process of the Standard Set for cleft and its hindering and promoting factors is needed to support other cleft teams during their implementation endeavors.

Further, there is a feeling among care providers that with the implementation of outcome measurements in practice, medical consultations might become more time-consuming and expenses might increase.[66,67] Also, problems related to health information technology integrations might complicate registration and consultation.[66] However, these assumptions have not been thoroughly investigated yet and are mainly based on gut feelings and personal expressions. Understanding the patterns of healthcare use and medical costs before and after implementation can help teams adapt their care pathways efficiently and adequately and could provide first insights in the possibilities for bundled payment strategies in cleft. For example, the use of PROMs in clinical practice might result in an increase of a clinician's burden on the short term due to early recognition of problems and needs, while decreasing the complication rates and associated costs in the long run.

The above-mentioned elements of outcome measurement and healthcare organization including medical costs as part of the value-based healthcare approach for cleft will be explored by multiple collaborative research projects described in this thesis. Only with broad implementation of a valid outcomes framework, outcomes can be adequately measured and best practices can be determined to provide high quality of cleft care that matches the needs and wishes of the individual patient.

Aims and outline of this thesis

This thesis covers multiple aspects concerning the measurement of patient-reported outcomes in patients with a cleft (**Part I)** and the implementation of these outcome measures in clinical cleft practice (**Part II**). The following research questions will be addressed:

1. How can we optimize the measurement of patient-reported outcomes in the ICHOM Standard Set for Cleft Lip and Palate?
    a. How is the psychometric performance and concept coverage of the patient-reported outcome measures of the ICHOM Standard Set for Cleft Lip and Palate?
    b. How can we maximize information while reducing burden when measuring psychosocial function within the ICHOM Standard Set for Cleft Lip and Palate?
    c. What is the external validity of the CLEFT-Q Computerized Adaptive Test in patients with cleft lip and palate?
2. How can we optimize the implementation of the ICHOM Standard Set for Cleft Lip and Palate in clinical cleft care?
    a. What are facilitators and barriers to the implementation of the ICHOM Standard Set for Cleft Lip and Palate in clinical practice?
    b. What are the healthcare use and medical costs patterns of clinical cleft care and how is this influenced by the use of the ICHOM Standard Set for Cleft lip and Palate?

Different methodologies and datasets were used to answer the research questions stated and are presented in **Table 3**. This thesis is divided in two parts, namely measurement challenges and implementation challenges.

Part I of this thesis studies varied aspects of the measurement instruments utilized in clinical practice as proposed by the ICHOM Standard Set for Cleft Lip and Palate (research question 1a, b, and c). **Chapter 2** presents a multicentre study in which the psychometric performance of the patient-reported outcome measures as included in the Standard Set is evaluated with Rasch measurement theory (RMT). RMT is a modern statistical approach to gain insights in the strengths and limitations of an outcome instrument.

Chapter 3 focuses on the usefulness of the PROMs on psychosocial function in terms of clinical-decision making and explores possible redundancy amongst the three psychosocial function scales. Also, associations between patient characteristics and their psychosocial outcome scores and referral to psychosocial care are explored.

Chapter 4 reports on the external validity of the Computerized Adaptive Test (CAT) version of the CLEFT-Q scales in the ICHOM Standard Set and describes the CAT user-experiences at various cleft teams internationally. **Chapter 5** presents a follow-up study on the CLEFT-Q CAT investigating whether item response theory (IRT) could improve the performance of the RMT CAT algorithms for the CLEFT-Q appearance scales.

Part II of this thesis focuses on the implementation of the Standard Set in clinical cleft practice and provides answers to research question 2a and 2b. **Chapter 6** describes a qualitative study on the facilitators and barriers to the implementation of the Standard Set in clinical cleft practice. The identified themes are presented in relation to the dimensions of reach, effectiveness, adoption, implementation and maintenance, as described by the RE-AIM framework. **Chapter 7** provides a closer view on the implementation experiences across four cleft centres internationally in a more conversational fashion.

Chapter 8 presents healthcare utilization patterns and medical costs of 40 patients with unilateral cleft lip and palate. The additional medical costs generated by the use of the ICHOM Standard Set in practice are explored based on the old and new treatment protocols.

Chapter 9 provides an overall discussion and reflects on the challenges of outcome measurement and implementation of the ICHOM Standard Set in clinical cleft practice can and how these aspects can be optimized.

Chapter 10 concludes this dissertation with English and Dutch summaries of the presented studies.

	Ch.	Study title	Study methodology	Data source	Cleft type	Outcome measures
Part I	2	Rasch Analysis of Patient- and Parent-Reported Outcome Measures in the International Consortium for Health Outcomes Measurement (ICHOM) Standard Set for Cleft Lip and Palate. *Value Health. 2021 Mar;24(3):404-412*	Rasch measurement theory (RMT)	Patients treated in the Erasmus University Medical Center, Boston Children's Hospital, Duke Children's Hospital between 2015 and 2019, and participants of the CLEFT-Q Phase 3 study (n=714)	CL, CP, CLA, CLP	- CLEFT-Q: face, teeth, jaws, psychological function, social function, school function, speech function, speech-related distress, eating and drinking - NOSE - COHIP-OSS - ICS
	3	Optimizing the Psychosocial Function Measures in the International Consortium for Health Outcomes Measurement Standard Set for Cleft. *Plast Reconstr Surg. 2023 Feb 1;151(2):274e-281e*	Spearman's correlation coefficients Univariable linear and logistic regression analyses	Patients treated in the Erasmus University Medical Center, Boston Children's Hospital, Duke Children's Hospital between 2015 and 2019, and participants of the CLEFT-Q development and validation project (n=3067)	CL, CP, CLA, CLP	- CLEFT-Q: social function, psychological function, school function, face, speech function, speech-related distress - Referrals to psychosocial care
	4	Personalized, Person-Centered Health Assessments for Cleft Lip and/or Palate: The Development, Deployment, and Evaluation of the CLEFT-Q Computerized Adaptive Test. *J Med Internet Res. 2023 Apr 27;25:e41870*	Pearson correlation coefficients for agreement, root mean squared error (RMSE), 95% limits of agreement Inductive thematic analyses	Calibration dataset: Patients treated in 30 centers in 12 countries between 2014 and 2016, data collected as part of the CLEFT-Q field test study (n=2434) Validation dataset: Patients treated in the Erasmus University Medical Center, Boston Children's Hospital, Duke Children's Hospital between 2015 and 2019 (n=536)	CL, CP, CLA, CLP	- CLEFT-Q: face, teeth, jaws, psychological function, social function, school function, speech function, speech-related distress
	5	Modern Test Theory Techniques for Adaptive Testing in Short Scales Comprising Polytomous Items: A Monte Carlo Simulation Study Comparing Rasch Measurement Theory to Unidimensional and Multidimensional Graded Response Models. *Submitted*	Unidimensional and multidimensional model calibration Pearson correlation coefficients for agreement, RMSE, mean absolute error (MAE), 95% limits of agreement	Calibration dataset: Sample of patients treated in 30 centers in 12 countries between 2014 and 2016, data collected as part of the CLEFT-Q field test study (n=895) Validation dataset: Patients treated in the Erasmus University Medical Center, Boston Children's Hospital, Duke Children's Hospital between 2015 and 2019 (n=551)	CL, CP, CLA, CLP	- CLEFT-Q: face, jaw, teeth (appearance domain)

Part II	6	Barriers and Facilitators to the International Implementation of Standardized Outcome Measures in Clinical Cleft Practice. *Cleft Palate Craniofac J. 2022 Jan;59(1):5-13*	Thematic content analysis of exploratory surveys and semi-structured interviews	Clinicians, health information technology professionals, project coordinators and project leaders of Erasmus University Medical Center, Boston Children's Hospital, Duke Children's Hospital and Karolinska University Hospital (n=13)	N/A	RE-AIM framework: reach, effectiveness, adoption, implementation, maintenance
	7	Whitepaper: Implementation of the ICHOM Standard Set for Cleft Lip and/or Palate Across Four Centers. *Published online: https://conference.ichom.org/wp-content/uploads/2021/09/24915-CLP-Cleft-Lip-whitepaper.pdf*	Narrative presentation of experiences	Project coordinators of Erasmus University Medical Center, Boston Children's Hospital, Duke Children's Hospital, Karolinska University Hospital and ICHOM	N/A	N/A
	8	Healthcare Use and Direct Medical Costs in a Cleft Lip and Palate Population: An Analysis of Observed and Protocolized Care and Costs. *Accepted at IJOMS*	Description of counts and (mean) costs per person year and for the total cycle of care	Sample of patients treated in the Erasmus University Medical Center between 2012 and 2019 (n=40)	Unilateral CLP	- (as summed up, see previous points in column) Observed healthcare use and medical costs - (as summed up, see previous points in column) Expected protocol costs without and with ICHOM

Table 3 Overview of the studies described in this thesis. CL = cleft lip, CP = cleft palate, CLA = cleft lip and alveolus, CLP = cleft lip and palate N/A = not applicable.

References

1. Rozendaal AM, Luijsterburg AJ, Ongkosuwito EM, de Vries E, Vermeij-Keers C. Decreasing prevalence of oral cleft live births in the Netherlands, 1997-2006. *Arch Dis Child Fetal Neonatal Ed.* 2011;96(3):F212-216.
2. Luijsterburg AJ, Vermeij-Keers C. Ten years recording common oral clefts with a new descriptive system. *Cleft Palate Craniofac J.* 2011;48(2):173-182.
3. International Perinatal Database of Typical Oral Clefts Working Group. Prevalence at birth of cleft lip with or without cleft palate: data from the International Perinatal Database of Typical Oral Clefts. *Cleft Palate Craniofac J.* 2011;48(1):66-81.
4. Tanaka SA, Mahabir RC, Jupiter DC, Menezes JM. Updating the epidemiology of cleft lip with or without cleft palate. *Plast Reconstr Surg.* 2012;129(3):511e-518e.
5. World Health Organization. *Global strategies to reduce the health-care burden of craniofacial anomalies : report of WHO meetings on International Collaborative Research on Craniofacial Anomalies, Geneva, Switzerland, 5-8 November 2000 ; Park City, Utah, U. S. A., 24-26 May 2001.* 2002.
6. Kirschner RE, LaRossa D. Cleft lip and palate. *Otolaryngol Clin North Am.* 2000;33(6):1191-1215, v-vi.
7. Allori AC, Mulliken JB, Meara JG, Shusterman S, Marcus JR. Classification of Cleft Lip/Palate: Then and Now. *Cleft Palate Craniofac J.* 2017;54(2):175-188.
8. International Consortium for Health Outcomes Measurement (ICHOM). Data collection reference guide. https://ichom.org/files/medical-conditions/cleft-lip-palate/cleft-lip-palate-reference-guide.pdf. Published 2018. Accessed December 1, 2020.
9. Martelli DR, Machado RA, Swerts MS, Rodrigues LA, Aquino SN, Martelli Junior H. Non syndromic cleft lip and palate: relationship between sex and clinical extension. *Braz J Otorhinolaryngol.* 2012;78(5):116-120.
10. Honein MA, Rasmussen SA, Reefhuis J, et al. Maternal smoking and environmental tobacco smoke exposure and the risk of orofacial clefts. *Epidemiology.* 2007;18(2):226-233.
11. Itikala PR, Watkins ML, Mulinare J, Moore CA, Liu Y. Maternal multivitamin use and orofacial clefts in offspring. *Teratology.* 2001;63(2):79-86.
12. Shaw GM, Lammer EJ, Wasserman CR, O'Malley CD, Tolarova MM. Risks of orofacial clefts in children born to women using multivitamins containing folic acid periconceptionally. *Lancet.* 1995;346(8972):393-396.
13. Shaw GM, Lammer EJ. Maternal periconceptional alcohol consumption and risk for orofacial clefts. *J Pediatr.* 1999;134(3):298-303.
14. Munger RG, Sauberlich HE, Corcoran C, Nepomuceno B, Daack-Hirsch S, Solon FS. Maternal vitamin B-6 and folate status and risk of oral cleft birth defects in the Philippines. *Birth Defects Res A Clin Mol Teratol.* 2004;70(7):464-471.
15. Izedonmwen OM, Cunningham C, Macfarlane TV. What is the Risk of Having Offspring with Cleft Lip/Palate in Pre-Maternal Obese/Overweight Women When Compared to Pre-Maternal Normal Weight Women? A Systematic Review and Meta-Analysis. *J Oral Maxillofac Res.* 2015;6(1):e1.
16. Calzolari E, Pierini A, Astolfi G, Bianchi F, Neville AJ, Rivieri F. Associated anomalies in multi-malformed infants with cleft lip and palate: An epidemiologic study of nearly 6 million births in 23 EUROCAT registries. *Am J Med Genet A.* 2007;143A(6):528-537.
17. Stoll C, Alembik Y, Dott B, Roth MP. Associated malformations in cases with oral clefts. *Cleft Palate Craniofac J.* 2000;37(1):41-47.

18. Fleurke-Rozema JH, van de Kamp K, Bakker MK, Pajkrt E, Bilardo CM, Snijders RJ. Prevalence, diagnosis and outcome of cleft lip with or without cleft palate in The Netherlands. *Ultrasound Obstet Gynecol.* 2016;48(4):458-463.

19. Ensing S, Kleinrouweler CE, Maas SM, Bilardo CM, Van der Horst CM, Pajkrt E. Influence of the 20-week anomaly scan on prenatal diagnosis and management of fetal facial clefts. *Ultrasound Obstet Gynecol.* 2014;44(2):154-159.

20. Nusbaum R, Grubs RE, Losee JE, Weidman C, Ford MD, Marazita ML. A qualitative description of receiving a diagnosis of clefting in the prenatal or postnatal period. *J Genet Couns.* 2008;17(4):336-350.

21. Baylis AL, Pearson GD, Hall C, et al. A Quality Improvement Initiative to Improve Feeding and Growth of Infants With Cleft Lip and/or Palate. *Cleft Palate Craniofac J.* 2018;55(9):1218-1224.

22. Stock NM, Feragen KB. Psychological adjustment to cleft lip and/or palate: A narrative review of the literature. *Psychol Health.* 2016;31(7):777-813.

23. Cheung LK, Loh JS, Ho SM. Psychological profile of Chinese with cleft lip and palate deformities. *Cleft Palate Craniofac J.* 2007;44(1):79-86.

24. Kramer FJ, Gruber R, Fialka F, Sinikovic B, Schliephake H. Quality of life and family functioning in children with nonsyndromic orofacial clefts at preschool ages. *J Craniofac Surg.* 2008;19(3):580-587.

25. Berger ZE, Dalton LJ. Coping with a cleft: psychosocial adjustment of adolescents with a cleft lip and palate and their parents. *Cleft Palate Craniofac J.* 2009;46(4):435-443.

26. Noor SN, Musa S. Assessment of patients' level of satisfaction with cleft treatment using the Cleft Evaluation Profile. *Cleft Palate Craniofac J.* 2007;44(3):292-303.

27. Lorot-Marchand A, Guerreschi P, Pellerin P, et al. Frequency and socio-psychological impact of taunting in school-age patients with cleft lip-palate surgical repair. *Int J Pediatr Otorhinolaryngol.* 2015;79(7):1041-1048.

28. Hunt O, Burden D, Hepper P, Stevenson M, Johnston C. Self-reports of psychosocial functioning among children and young adults with cleft lip and palate. *Cleft Palate Craniofac J.* 2006;43(5):598-605.

29. Tierney S, O'Brien K, Harman NL, Sharma RK, Madden C, Callery P. Otitis media with effusion: experiences of children with cleft palate and their parents. *Cleft Palate Craniofac J.* 2015;52(1):23-30.

30. Mink van der Molen AB, van Breugel JMM, Janssen NG, et al. Clinical Practice Guidelines on the Treatment of Patients with Cleft Lip, Alveolus, and Palate: An Executive Summary. *J Clin Med.* 2021;10(21).

31. Nederlandse Vereniging voor Schisis en Craniofaciale Aandoeningen (NVSCA). Teams - behandelprotocollen. https://www.schisis.nl/nvsca2/teams. Published 2022. Accessed March 1, 2022.

32. Shaw WC, Semb G, Nelson P, et al. The Eurocleft project 1996-2000: overview. *J Craniomaxillofac Surg.* 2001;29(3):131-140; discussion 141-132.

33. Russell K, Long RE, Jr., Hathaway R, et al. The Americleft study: an inter-center study of treatment outcomes for patients with unilateral cleft lip and palate part 5. General discussion and conclusions. *Cleft Palate Craniofac J.* 2011;48(3):265-270.

34. E. PM, O. TE. *Redefining Health Care: Creating Value-Based Competition on Results.* Harvard Business Review Press; 2006.

35. Porter ME. What is value in health care? *N Engl J Med.* 2010;363(26):2477-2481.

36. Porter ME, Teisberg, E. *Redefining healthcare; creating value-based competition on results.* Harvard Business Review Press; 2006.
37. Porter ME. Value-based health care delivery. *Ann Surg.* 2008;248(4):503-509.
38. Porter ME. A strategy for health care reform--toward a value-based system. *N Engl J Med.* 2009;361(2):109-112.
39. Porter ME, Larsson S, Lee TH. Standardizing Patient Outcomes Measurement. *N Engl J Med.* 2016;374(6):504-506.
40. Allori AC, Kelley T, Meara JG, et al. A Standard Set of Outcome Measures for the Comprehensive Appraisal of Cleft Care. *Cleft Palate Craniofac J.* 2017;54(5):540-554.
41. Ministry of Healthcare Welfare and Sport. *Outcome-based healthcare 2018-2022.* 2018.
42. Zorginstituut Nederland. Meer patiëntregie door meer uitkomstinformatie in 2022. 2018. https://www.zorginstituutnederland.nl/actueel/nieuws/2018/08/14/meer-patientregie-door-meer-uitkomstinformatie-in-2022. Accessed April 18, 2022.
43. Zorginstituut Nederland. *ICHOM de heilige graal of routekaart naar meer uitkomstinformatie?* 2018.
44. de Vet HC, Terwee CB, Mokkink LB, Knol DL. *Measurement in medicine: a practical guide.* Cambridge University Press; 2011.
45. Eckstein DA, Wu RL, Akinbiyi T, Silver L, Taub PJ. Measuring quality of life in cleft lip and palate patients: currently available patient-reported outcomes measures. *Plast Reconstr Surg.* 2011;128(5):518e-526e.
46. Tsangaris E, Wong Riff KWY, Goodacre T, et al. Establishing Content Validity of the CLEFT-Q: A New Patient-reported Outcome Instrument for Cleft Lip/Palate. *Plast Reconstr Surg Glob Open.* 2017;5(4):e1305.
47. Wong Riff KW, 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(1):e015467.
48. Apon I, van Leeuwen N, Allori AC, et al. Rasch Analysis of Patient- and Parent-Reported Outcome Measures in the International Consortium for Health Outcomes Measurement Standard Set for Cleft Lip and Palate. *Value Health.* 2021;24(3):404-412.
49. Cano SJ, Hobart JC. The problem with health measurement. *Patient Prefer Adherence.* 2011;5:279-290.
50. Patrick DL, Deyo RA. Generic and disease-specific measures in assessing health status and quality of life. *Med Care.* 1989;27(3 Suppl):S217-232.
51. Bergner M, Rothman ML. Health status measures: an overview and guide for selection. *Annu Rev Public Health.* 1987;8:191-210.
52. Mokkink LB, Boers M, van der Vleuten C, et al. *COSMIN Risk of Bias tool to assess the quality of studies on reliability and measurement error of outcome measurement instrument: user manual.* 2021.
53. 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.
54. 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.

55. Boers M, Kirwan JR, Tugwell P, et al. *The OMERACT Handbook.* 2018.

56. Schmitt J, Apfelbacher C, Spuls PI, et al. The Harmonizing Outcome Measures for Eczema (HOME) roadmap: a methodological framework to develop core sets of outcome measurements in dermatology. *J Invest Dermatol.* 2015;135(1):24-30.

57. Prinsen CA, Vohra S, Rose MR, et al. How to select outcome measurement instruments for outcomes included in a "Core Outcome Set" - a practical guideline. *Trials.* 2016;17(1):449.

58. Porter I, Goncalves-Bradley D, Ricci-Cabello I, et al. Framework and guidance for implementing patient-reported outcomes in clinical practice: evidence, challenges and opportunities. *J Comp Eff Res.* 2016;5(5):507-519.

59. Harrison CJ, Geerards D, Ottenhof MJ, et al. Computerised adaptive testing accurately predicts CLEFT-Q scores by selecting fewer, more patient-focused questions. *J Plast Reconstr Aesthet Surg.* 2019;72(11):1819-1824.

60. Harrison CJ, Rodrigues JN, Furniss D, et al. Optimising the computerised adaptive test to reliably reduce the burden of administering the CLEFT-Q: A Monte Carlo simulation study. *J Plast Reconstr Aesthet Surg.* 2021;74(6):1355-1401.

61. Basch E, Barbera L, Kerrigan CL, Velikova G. Implementation of Patient-Reported Outcomes in Routine Medical Care. *Am Soc Clin Oncol Educ Book.* 2018;38:122-134.

62. Antunes B, Harding R, Higginson IJ, Euroimpact. Implementing patient-reported outcome measures in palliative care clinical practice: a systematic review of facilitators and barriers. *Palliat Med.* 2014;28(2):158-175.

63. Boyce MB, Browne JP, Greenhalgh J. The experiences of professionals with using information from patient-reported outcome measures to improve the quality of healthcare: a systematic review of qualitative research. *BMJ Qual Saf.* 2014;23(6):508-518.

64. Bausewein C, Simon ST, Benalia H, et al. Implementing patient reported outcome measures (PROMs) in palliative care--users' cry for help. *Health Qual Life Outcomes.* 2011;9:27.

65. Foster A, Croot L, Brazier J, Harris J, O'Cathain A. The facilitators and barriers to implementing patient reported outcome measures in organisations delivering health related services: a systematic review of reviews. *J Patient Rep Outcomes.* 2018;2:46.

66. Amini M, Oemrawsingh A, Verweij LM, et al. Facilitators and barriers for implementing patient-reported outcome measures in clinical care: An academic center's initial experience. *Health Policy.* 2021;125(9):1247-1255.

67. van Egdom LSE, Oemrawsingh A, Verweij LM, et al. Implementing Patient-Reported Outcome Measures in Clinical Breast Cancer Care: A Systematic Review. *Value Health.* 2019;22(10):1197-1226.

PART I

Measurement Challenges

Chapter 2

Rasch Analysis of the Patient- and Parent-Reported Outcome Measures in the ICHOM Standard Set for Cleft Lip and Palate

Apon I, MD, MHS[1]; van Leeuwen N, PhD[2]; Allori AC, MD, PhD[3]; Koudstaal MJ, MD, DMD, PhD[1]; Rogers-Vizena CR, MD[4]; Wolvius EB, MD, DMD, PhD[1]; Cano SJ, PhD[5]; Klassen AF, DPhil[6]; Versnel SL, MD, PhD[7]

[1] Department of Oral and Maxillofacial Surgery, Erasmus University Medical Center, Rotterdam, The Netherlands
[2] Department of Public Health, Erasmus University Medical Center, Rotterdam, The Netherlands
[3] Department of Plastic, Maxillofacial and Oral Surgery, Duke Children's Hospital, Durham, North Carolina, USA
[4] Department of Plastic and Oral Surgery, Boston Children's Hospital, Boston, Massachusetts, USA
[5] Modus Outcomes, Letchworth Garden City, United Kingdom
[6] Department of Pediatrics, McMaster University, Hamilton, Ontario, Canada
[7] Department of Plastic and Reconstructive Surgery, Erasmus University Medical Center, Rotterdam, The Netherlands

Value Health. 2021 Mar;24(3):404-412

Abstract

Objectives: The aim of this study was to evaluate the psychometric performance of the patient- and parent-reported measures in the ICHOM Standard Set for Cleft Care, and to identify ways of improving concept coverage.

Methods: Data from 714 patients with cleft lip and/or palate, aged 8-9, 10-12.5 and 22 years were collected between November 2015 and April 2019 at Erasmus University Medical Center, Boston Children's Hospital, Duke Children's Hospital, and from participating sites in the CLEFT-Q Phase 3 study. The Standard Set includes nine CLEFT-Q scales, the NOSE questionnaire, the COHIP-OSS and the Intelligibility in Context Scale (ICS). Targeting, item-fit statistics, thresholds for item responses and measurement precision (PSI) were analysed using Rasch Measurement Theory.

Results: The proportion of the sample to score within each instruments range of measurement varied from 69% (ICS) to 92% (CLEFT-Q teeth and COHIP-OSS). Specific problems with individual items within the NOSE and COHIP-OSS questionnaires were noted, such as poor item fit to the Rasch model and disordered thresholds (6 out of 10). Reliability measured with PSI was above 0.82 for the ICS and all but one CLEFT-Q scale (speech distress). PSIs were lowest for the COHIP-OSS (0.43) and NOSE questionnaire (0.35).

Conclusion: The patient- and parent-reported components within the facial appearance, psychosocial function and speech domains are valid measures. However, the facial function and oral health domains are not sufficiently covered by the CLEFT-Q eating and drinking, NOSE and COHIP-OSS, and these questionnaires may not be accurate enough to stratify cleft-related outcomes.

Keywords: cleft lip and palate, ICHOM, patient-reported outcomes, psychometric performance, Rasch measurement theory.

Introduction

Cleft lip and/or palate (CL/P) is the most prevalent congenital craniofacial anomaly affecting approximately 7.94 per 10,000 live-births worldwide.[1,2] This complex disorder can negatively influence an individuals' appearance and psychosocial well-being, and cause functional disabilities such as problems with feeding, dentition, hearing and speech.[3,4] Patients may need to undergo many surgical and non-surgical procedures from infancy through young adulthood to improve physical and psychosocial function and well-being. To date, almost every cleft center has its own treatment protocol based on various literature and own experiences, resulting in differences in outcomes and quality of care.[5,6] Research into the psychosocial consequences of different treatment strategies for CL/P has been conducted without a uniform strategy.[7]

Traditionally, the success or failure of a cleft treatment has been evaluated and interpreted by clinicians.[6,8,9] However, clinician-reported outcomes fail to encompass the perspective of patients and their parents or caregivers, especially with regard to quality of life. In 2016, the Cleft Lip and Palate Working Group of the International Consortium for Health Outcomes Measurement (ICHOM) proposed a Standard Set of cleft-specific outcome measures for the comprehensive appraisal of cleft care. This Set has been implemented over the past few years in several centers worldwide.[10-13] It stresses the importance of the patient's perspective by incorporating parent- and patient-reported outcome measures (PROMs). Specifically, the Set includes nine CLEFT-Q scales[14-16], the Child Oral Health Impact Profile – Oral Symptoms Scale (COHIP-OSS)[17], the Nasal Obstruction Symptom Evaluation (NOSE) questionnaire[18], and the parent-reported Intelligibility in Context Scale (ICS)[19]. These instruments were chosen to cover core concepts of facial appearance, psychosocial function, speech, facial function (including eating/drinking and breathing) and oral health. Each of these conceptual domains should be assessed using clinically relevant, reliable and valid scales to properly inform clinical decision-making and to facilitate future comparative effectiveness research and quality-improvement projects.

In encouraging the adoption of any standardized outcomes-assessment framework, it is essential to verify that each of the included measures is robust enough to accurately and reliably appraise the corresponding conceptual construct or outcome domain. To that end, the aim of this study was to evaluate the psychometric performance of the patient- and parent-reported outcome measures in the ICHOM Standard Set for Cleft Care, such that we might gain insight into potential gaps of concept coverage.

Methods

Study setting

De-identified CL/P outcome data was collected prospectively in clinical practice between November 2015 and April 2019 at Erasmus University Medical Center, Duke Children's Hospital, Boston Children's Hospital, and at international centers participating in the CLEFT-Q Phase 3 study (Canada, United States, United Kingdom) led by researchers of McMaster University. The aim of the Phase 3 study was to measure change in outcomes following four specific cleft-related procedures (alveolar bone grafting, secondary cleft lip revision, jaw surgery and rhinoplasty). Research ethical approvals were obtained at the Institutional Review Board of each center.

Patient population

All patients with orofacial clefts were eligible for data collection. They were all treated by a multidisciplinary cleft team. Cleft phenotypic categories were specified as the following: cleft lip (CL); cleft palate (CP); cleft lip and alveolus (CLA); and cleft lip and palate (CLAP). Outcomes were measured at time points defined by patient's age: T8 (range 8-9), T12 (range 10-12.5) and T22 (22 years or end of treatment, whichever is soonest).[10] Outcome data was collected electronically via home-based computer, an iPad at clinics or paper and pencil and stored with REDCap[20,21] or Gemstracker[11], dependent on the site's preferences **(Supplemental Material – Table 1)**. All scales were administered in the native language of the country where each institution is located using approved translations of the instruments.

Patient-reported outcome measures

The outcome measures assessed in this study include nine patient-reported CLEFT-Q scales, the patient-reported COHIP-OSS and NOSE questionnaire, and the parent-reported ICS.

The CLEFT-Q is a rigorously developed, cleft-specific instrument focusing on three major domains: appearance, facial function and health-related quality of life.[14-16] Each major domain was further broken down conceptually into subdomains, based upon thematic content analysis of extensive focus groups and semi-structured interviews.[16] The CLEFT-Q face, jaws, teeth, psychological, school, social, speech function and speech distress scales and the CLEFT-Q eating and drinking checklist were adopted as part of the ICHOM Standard Set. For the assessment of oral health, the Child Oral Health Impact

Profile – Oral Symptoms Scale (COHIP-OSS) was included. The COHIP-OSS is a subscale of the larger COHIP, which was developed to measure various outcomes on oral health in school-aged children with different oral conditions, including CL/P.[17,22]

For assessing the quality of life related to nasal breathing, the Nasal Obstructive Symptom Evaluation (NOSE) questionnaire was adopted.[18,23] This questionnaire was developed to evaluate breathing outcomes of rhinoplasty and/or septoplasty treatment in adults.[24]

For speech, the Intelligibility in Context Scale (ICS) scale was developed to discriminate children with speech difficulties.[19] Since parents and family play an important role in representing the young patient with cleft, they were invited to complete the ICS by rating the degree of their children's intelligibility when speaking to various communication partners.

More information on the scales, including the core concepts measured, timing for completion, and example questions can be found in **Table 1**.

Concept	Scale	Number of items	Cleft phenotypes assessed	Time points (years)	Examples of questions	Response options
Facial appearance	CLEFT-Q face *(patient-reported)*	9	CL, CP, CLA, CLAP	8, 12, 22	How much do you like… - …how your face looks when you look your best? - …how your face looks when you smile?	Not at all, a little bit, quite a bit, very much
	CLEFT-Q teeth *(patient-reported)*	8	CL, CP, CLA, CLAP	8, 12, 22	How much do you like… - …the size of your teeth? - …how straight your teeth look?	Not at all, a little bit, quite a bit, very much
	CLEFT-Q jaws *(patient-reported)*	7	CL, CP, CLA, CLAP	12, 22	How much do you like… - …the size of your jaws? - …how your jaws look from the side?	Not at all, a little bit, quite a bit, very much
Psychosocial function	CLEFT-Q psychological *(patient-reported)*	10	CL, CP, CLA, CLAP	12	How do you feel? - I am happy with my life. - I feel confident.	Never, sometimes, often, always
	CLEFT-Q social *(patient-reported)*	10	CL, CP, CLA, CLAP	8, 22	How is your social life? - I have fun with friends. - I feel like I fit in.	Never, sometimes, often, always
	CLEFT-Q school *(patient-reported)*	10	CL, CP, CLA, CLAP	12	How is your school life? - I like seeing my friends at school - I feel safe at school (not bullied)	Never, sometimes, often, always
Speech	CLEFT-Q speech distress *(patient-reported)*	10	CP, CLAP	12, 22	How do you feel about speaking? - I get teased about my speech. - I get upset when I need to repeat myself	Always, sometimes, never
	CLEFT-Q speech function *(patient-reported)*	12	CP, CLAP	12, 22	How is your speech? - It's hard for my family to understand my speech. - I need to concentrate to speak well.	Always, sometimes, never
	Intelligibility in Context Scale *(parent-reported)*	7	CP, CLAP	12	Think about your child's speech intelligibility over the past month and identify the degree of understanding. - Do you understand your child? - Do immediate members of your family understand your child?	Never, rarely, sometimes, usually, always

Facial function	CLEFT-Q eating and drinking *(patient-reported)*	9	CP, CLA, CLAP	8, 12, 22	How is your eating and drinking? - Food falls out of my mouth when I eat. - Food or drinks go up my nose.	Always, often, sometimes, never
	NOSE *(patient-reported)*	5	CL, CP, CLA, CLAP	8, 12	How much of a problem were the following conditions for you? - …Nasal blockage or obstruction. - …Trouble breathing through my nose.	No problem, mild, moderate, fairly bad, severe problem
Oral health	COHIP-OSS *(patient-reported)*	5	CP, CLA, CLAP	8, 12	In the past 3 months, have you… - …Had pain in your teeth? - …Had bleeding gums?	Never, almost never, sometimes, fairly often, almost all of the time

Table 1 Overview of the patient-and parent-reported outcome measures in the ICHOM Standard Set for Cleft Care. The measured core concepts including measurement instruments, number of items per scale, phenotypic groups, age groups for scale completion and examples of questions with their response options are presented. CL indicates cleft lip; CLA, cleft lip and alveolus; CLAP, cleft lip and palate; CP, cleft palate.

Statistical analysis

Descriptive analyses were performed using SPSS software (IBM SPSS Statistics for Windows, Version 25.0, released 2017, IBM Corp.). To provide insights into the performances of the PROMs, we applied Rasch Measurement Theory using RUMM 2030 software (RUMM version 2030, 1997-2020, RUMM Laboratory Pty Ltd) to our dataset with polytomous response options. Rasch analysis is a method that examines the extent to which the patient's responses match the predictions of the responses from the mathematical, logistic Rasch model. The difference between the expected and observed responses indicate the degree of rigorous measurement.[25-29] Within RUMM, we used the Partial-Credit Model, as this places no constraints on the threshold parameters. For this study, the following four keystones of Rasch Measurement Theory were assessed:

Targeting: The extent to which the distribution of the responses of the sample matches the range that can be measured by a specific scale is called targeting. Targeting is evaluated both graphically as with the percentage of the sample to score within the scale's range. When the sample is normally distributed and matches the construct as defined by the sample, a high percentage will be reached. A lower percentage corresponds with more mismatch and suggests that some patients' real ability cannot be determined with the scale.

Item-fit statistics: To evaluate whether responses are consistent with the expectations of the Rasch model, three fit indicators were examined: the c^2 values (item-trait interaction), the log residuals (item-person interaction) and the item characteristic curves. The ideal fit residuals are between -2.5 and +2.5 with c^2 values non-significant after Bonferroni adjustment. Inconsistent use of response options or multidimensionality can contribute to individual item misfit.

Thresholds for item response options: The thresholds between the response options of the scales were examined to determine whether they were used in an orderly fashion. Disordered thresholds can occur as a consequence of unclear definitions, too many response options or underutilization of an option.[27]

Measurement precision: For each scale, the estimated measurement precision is given by the person separation index (PSI). Extreme values were withdrawn from the analyses. A higher PSI indicates higher reliability and a better discrimination amongst patients with different outcomes. A PSI of 0.7 is the lowest level of acceptability and is able to differentiate two groups.[30]

While Rasch Measurement Theory may also be applied towards the exploration of differences in item functioning between centers or countries, we did not address differential item functioning in this study, given that the Set is intended for international use.

Results

A total of 714 unique patients with CL/P completed at least one of the scales (as appropriate based on cleft phenotype and age), resulting in 748 assessments available for analysis. In total, 60% (n=425) of patients were diagnosed with CLAP, and 55% (n=391) of patients were male. Further demographics are presented in **Table 2**.

Characteristics	**Number of patients (%)** **Total N=714**
Cleft type	
Cleft lip only	51 (7)
Cleft palate only	165 (23)
Cleft lip and alveolus	73 (10)
Cleft lip and palate	425 (60)
Sex	
Male	391 (55)
Female	323 (45)
Sample	
Erasmus Medical Center	362 (51)
Duke Children’s Hospital	105 (15)
Boston Children’s Hospital	95 (13)
CLEFT-Q Phase 3 study	152 (21)
Time points	**Number of measurements (%)** **Total N=748**
8 years (range 7-9)	379 (51)
12 years (range 10-13)	244 (32)
22 years (range 20-24)	125 (17)

Table 2 Patient characteristics.

Results of the Rasch analyses are presented in **Table 3**. With regard to *targeting,* the highest percentage of participants to score within the scales’ measurement ranges were the CLEFT-Q teeth and the COHIP-OSS (both 92%). The CLEFT-Q jaws scale and the ICS were the least targeted (70% and 69% respectively). This is depicted in **Figure 1**, where an example is given of the person-item threshold distribution for the ICS showing that the instrument’s items did not cover the ability of persons at the higher end of the continuum.

Scale	Sample size	Targeting (% within range)	Items outside ± 2.5	Number of χ^2 significant p-values	Number of disordered thresholds	Person separation index
CLEFT-Q face	695	86	3	1	0	0.86
CLEFT-Q teeth	665	92	2	1	0	0.86
CLEFT-Q jaws	322	70	0	0	0	0.84
CLEFT-Q psychological	399	77	0	0	0	0.88
CLEFT-Q social	508	81	2	1	0	0.83
CLEFT-Q school	355	81	2	1	0	0.82
CLEFT-Q speech distress	257	76	0	0	0	0.61
CLEFT-Q speech function	274	81	1	0	0	0.83
Intelligibility in Context Scale	210	69	1	0	1	0.86
CLEFT-Q eating and drinking	501	74	1	1	7	0.49
NOSE	454	72	1	1	1	0.35
COHIP-OSS	426	92	0	0	5	0.43

Table 3 Scale performance statistics determined with Rasch Measurement Theory.

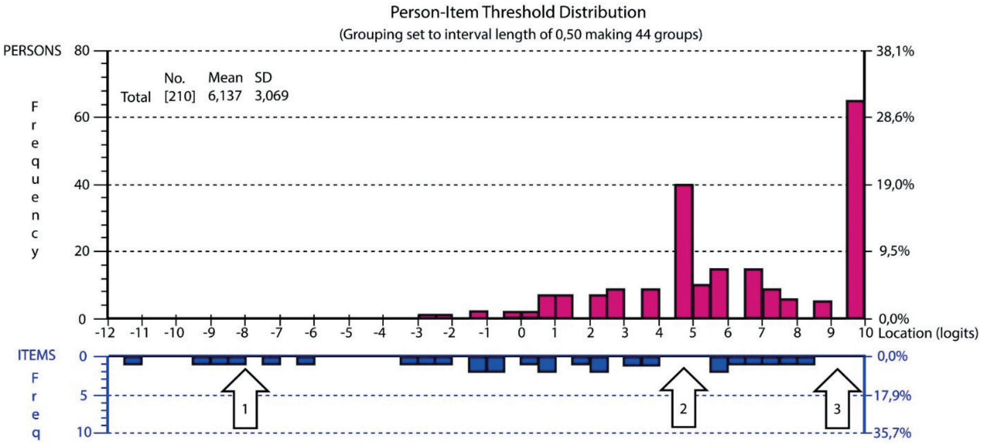


Figure 1 ICS person-item threshold distribution. This figure shows the targeting between the items, shown by the histogram in the lower half, and the patient sample, represented by the histogram in the upper half. At the lower end of the continuum the items are not covered by persons (arrow 1), whereas at +5 logit (arrow 2) and at the higher end of the continuum the persons are not covered by the items (arrow 3). This scale would benefit from including items that are more difficult.

Examination of *item-fit statistics* showed log residuals outside the ±2.5 for 13 of the 102 items for the entire Set, from which 6 of these items had a significant c^2-value. These items were all a marginal source of misfit with minor influence on the validity of the scale. None of the items in the Set failed all three criteria for fit. In **Table 4**, an example of model fit evaluation with item-characteristic curves is given for the CLEFT-Q face scale for two items.

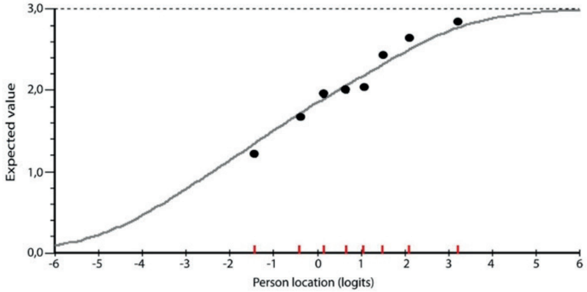
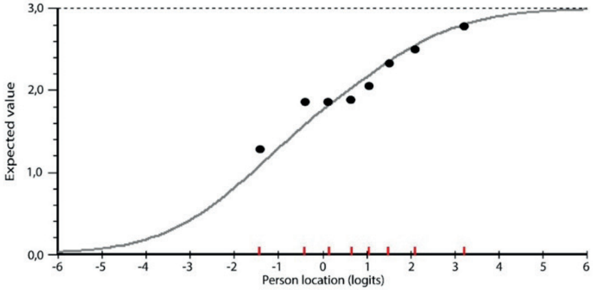
Item	Fit residual	χ^2	Item-characteristic curve	Interpretation
1	-2,640	22.34		Marginal overdiscrimination; the observed scores form a steeper curve than the expected scores. This item ('How your face looks when you look your best') is very similar to another item ('When you are ready to go out') and might therefore become redundant. However, this finding is not significant.
3	2,861*	25.25		Marginal underdiscrimination; the observed scores form a flatter curve than the expected scores. In clinical practice, a lot of patients consider this item ('how much do you like the shape of your face') as a more objective item in contrast with the other more subjective questions in the scale about 'how you look'. However, the deviation is very mild and is not considered clinically relevant.

Table 4 Examination of item fit of two CLEFT-Q face scale items. The observed values are represented by black dots and the expected values by the curve. High negative fit residuals are associated with redundancy or dependency of items, high positive fit residuals with multidimensionality.
* significant p-value.

For thresholds for item response options, 14 of the 102 items had disordered thresholds, including all 5 items of the COHIP-OSS. **Figure 2** illustrates this phenomenon of disordered thresholds with a characteristics probability curve of one item of the COHIP-OSS. The figure shows that the middle response options are never the most likely to be selected by this population in this specific clinical setting. The NOSE questionnaire and ICS both showed similar results for one disordered item ("trouble sleeping" and "understood by parents", respectively). Rescoring the NOSE questionnaire and the COHIP-OSS by combining the middle scores resulted in better threshold ordering. The CLEFT-Q eating and drinking checklist showed 7 disordered items.

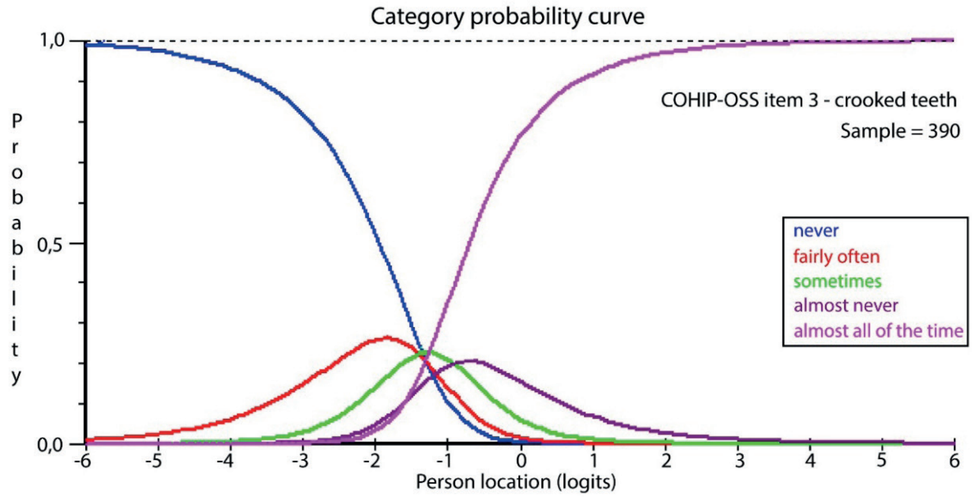


Figure 2 Category probability curve for item 3 'crooked teeth' of the COHIP-OSS showing disordered thresholds. The x-axis represents the construct with increasing severity to the right. The y-axis shows the probability of choosing the response categories. The middle categories were never the most likely to be selected.

For measurement precision, PSI values ranged from 0.82 to 0.88 for the CLEFT-Q scales, except for the speech distress scale (0.61) and eating and drinking (0.49). The analysis of the ICS revealed high reliability with a PSI value of 0.86. In contrast, the reliability scores for the NOSE and COHIP-OSS questionnaires were 0.35 and 0.43, respectively. This finding suggests that these scales were therefore not able to discriminate between patients with different qualities of nasal breathing and oral health.

Discussion

The ICHOM Cleft Lip and Palate Working Group acknowledged the importance of the patient perspective of health and included 12 patient- and parent-reported outcome scales in the ICHOM Standard Set for Cleft Care. These patient- and parent-reported instruments cover the core concepts of facial appearance, psychosocial function, speech, facial function (including eating/drinking and breathing) and oral health. The instruments were selected based on multiple criteria, including prior published evidence of instrument validation, clinical significance, practicality in implementation, availability, and translation into multiple languages. While the instruments were previously subject to some degree of validity testing, they have not yet undergone robust psychometric evaluation after implementation in "real world" clinical practice. Our study provides the first independent evaluation of the psychometric performance of these instruments as utilized within the context of the ICHOM Standard Set for Cleft Care. The Rasch analysis showed that the scales relating to the concepts of facial appearance, speech function, and psychosocial function worked properly with high reliability parameters.

Scales that lacked adequate resolution at the higher end of the continuum

The CLEFT-Q speech distress scale, which was incorporated in the Set for the evaluation of 12-year-old children and young-adult patients with CP or CLAP phenotypes, showed a slightly lower PSI value than the other CLEFT-Q scales. This is most likely due to some mis-targeting, since a lot of these patients have already completed intensive speech therapy and do not experience speech problems anymore. As a result, reliability of the scale is somewhat compromised without influencing the other psychometric characteristics.

The seven-item ICS is included as a parent-reported outcome measure. It has previously been tested and validated in pre-school aged children without cognitive or developmental disorders and has shown to be effective in discriminating children with speech difficulties.[19] In our study, the majority of patients scored high. As a result, a large group of patients is located at the upper extreme of the continuum, and these patients were not targeted by the scale items. This mis-calibration of the scale range has the effect of impairing the possibility of accurately determining the patient's intelligibility in context, or of being sensitive to change after speech-related interventions such as revision palatoplasty, pharyngoplasty or speech therapy. To improve the ICS, more items are needed at the higher end of the continuum.

Imbalanced scales that performed more like checklists

Facial function is covered by the CLEFT-Q eating and drinking checklist and the NOSE questionnaire. The developers of the CLEFT-Q previously reported that the reliability for the eating and drinking checklist was low (PSI < 0.60).[14] Our present study confirms these findings: most items in this questionnaire had disordered thresholds, which is why the creators of the CLEFT-Q emphasize the use of the term "checklist" rather than "scale".

Additionally, the NOSE questionnaire asks the patient how much of a problem some specific symptoms were for the patient over the past month, for example "nasal blockage" or "trouble breathing through my nose".[18,24] This is the first evaluation of the psychometric properties of the NOSE questionnaire in children with CL/P and revealed disordered thresholds for the item "trouble sleeping". Prior assessments in adults corroborate that this item contributed least in terms of measuring the construct of the scale.[24] Anecdotally, cleft clinicians at Erasmus University Medical Center experienced that the phrasing of the NOSE questions was too difficult to understand for children of this young age; parents were often asked to explain what "obstruction of the nose" means or whether they have "trouble sleeping". According to the category probability curves and item-threshold distribution, most children with CL/P experienced no problems breathing through their nose and thus respond at the end of the scale. Experiencing no problems might be incorrect in these patients since they don't know otherwise in view of their congenital nature. A small number of children with severe problems will score on the other end, whereas the middle options are not sensitive enough to measure small differences between patients. This finding was underlined by a very low PSI indicating no more than two groups can be discriminated with this questionnaire. A similar situation can be seen in a recent application of a modified NOSE questionnaire to investigate the prevalence of nasal obstruction symptoms in children with CL/P.[31] Modifications included a longer recall period of 12 months and questions and answers being rephrased from "problems" to "concerns". For the analysis of frequencies of NOSE scores and differences between cleft phenotypes, response categories were merged from five to three. With these response options, differences in nasal obstruction severity between unilateral and bilateral CLAP patients were found. This shows that with a small modification of the NOSE questionnaire, discriminative value can be slightly increased to enhance clinical utility. While this instrument might be useful as a screening tool or symptom checklist in clinical practice, we feel that the NOSE questionnaire in its current form is not sufficient as a pediatric PROM scale and suboptimal for the assessment of the young patient with cleft. In the same manner that the CLEFT-Q eating and drinking checklist is called a checklist rather than a scale, we would encourage that people refer to NOSE as a checklist rather than as a validated scale, as used in the pediatric cleft population.

This phenomenon of performing as a symptom screening tool, rather than a robust scale, also applies to the use of the COHIP-OSS for the assessment of oral health. This instrument measures the patient's view on oral health symptoms and was originally validated in a very heterogeneous sample of patients with diverse conditions affecting oral health, including patients with CL/P.[17,22] In our analysis of 8- and 12-year-old children with CL/P, the COHIP-OSS scale demonstrated low reliability, and all category thresholds were disordered. Most of the children responded at one end of the scale, reporting they "never" had any of the symptoms, except for the item "crooked teeth", which is most often scored as "almost all of the time". The latter can be explained by the fact that 8-year-old children are in mixed dentition and orthodontic treatment is awaiting. The middle response options of the COHIP-OSS were hardly used. Our findings suggest that either there are too many irrelevant response options, or the five options are not distinctive enough. Although this scale has been tested and validated in school-aged children with different types of clefts, our study confirmed the necessity to test and validate measurement instruments when used in different populations and under altered circumstances, since measurement characteristics can differ.[32]

To keep or to discard? That is the question

With regards to the use of PROM data for future comparative effectiveness research, it is important to minimize measurement error on outcomes. Therefore, it should be taken into consideration whether the use of poorly validated, not well understood instruments for children with CL/P, is sufficient enough for measuring the respective outcome domains. In a truly valid scale, all items should measure the same construct resulting in a sum score that informs patients and healthcare professionals on the overall well-being of the patient regarding the specific construct measured by the scale. The final sum score of a scale can then be used for comparative effectiveness research. When a scale measures subtly different constructs, resulting in a checklist, every single item may be appraised as an independent entity with a separate score, but no overall sum score should be calculated. A checklist can still be relevant for clinical decision-making, since individual elements can be intervened upon. However, due to its multidimensionality it is less suitable for outcome comparisons, such as comparing treatment techniques, protocols or centers, as sum scores are not interpretable.[33,34]

An attempt to improve the performances of the CLEFT-Q eating and drinking checklist, COHIP-OSS questionnaire and NOSE questionnaire by adding items or changing response options could be an option. On the other hand, it may be better to search for (or develop) a different scale that truly fits the concept. If the intended usage of these questionnaires is more akin to a screening tool than a diagnostic tool, then adding a quantitative measurement (for example nasometry measurement for the assessment of the nasal airway) for the corroboration of poorly scoring children could be considered. If the intended usage of these questionnaires is for outcome comparisons, a conservative option is to remove these three checklists from the Set. This will reduce burden on patients and will allow the clinicians to focus on the most useful PROMs.

Strengths and limitations

Since the ICHOM Standard Set is meant to be measured worldwide, a strength of this study is the international cohort of patients with CL/P resulting in a reflection of the cleft population that is eligible for completing the ICHOM Standard Set. However, a limitation of our study is that low income-countries were not represented in this cohort. Additionally, due to the clinical transition phase of implementing the Set, some 7-year-old children were asked to complete one or more of the outcome questionnaires, resulting in a slightly broader age range than advised by the ICHOM Reference Guide (age range 8-9).[10] The ages of eligibility were set at 8, since it is known that children as young as 8 years are able to report on well-being and psychosocial health.[35,36] However, given the small number of 7-year-old children included in this study and the large total sample size, we do not expect to find different results. Furthermore, we feel that including these patients in our sample gives a good reflection of daily clinical practice.

Conclusion

To improve patient-centered care and to facilitate future comparative effectiveness research and quality-improvement endeavours, it is important to include clinically meaningful and scientifically sound measurement instruments in an outcome set. This study found that most of the patient- and parent-reported components recommended by the ICHOM Standard Set for Cleft Care are valid tools for assessing cleft-specific outcomes. Importantly, the CLEFT-Q eating and drinking checklist, the COHIP-OSS and NOSE questionnaires were not found to be robust enough for outcomes comparisons, and instead work like a checklist rather than a measurement scale. As a result, the concepts of facial function (including eating/drinking and breathing) and oral health are not sufficiently covered by the PROMs included in the ICHOM Standard Set for Cleft Care.

Acknowledgements

We are grateful for all the help provided by Jet de Gier, Mona Haj, Anna Miroshnychenko, Charlene Rae, Karl Sanchez, Linda van der Sluijs, and Mariska van Veen-van der Hoek. We are thankful for the support provided by the Canadian Institutes of Health Research.

Conflicts of Interest

Dr. Cano is co-owner of Modus Outcomes and reported having a patent BODY-Q copyright will receive a share of any license revenues as royalties based on the inventor sharing policy. Dr. Klassen reported receiving grants from the Canadian Institutes of Health Research during the conduct of the study, and personal fees from Allergan outside the submitted work. In addition, Dr. Klassen reported having a patent CLEFT-Q and copyright is pending. No other disclosures were reported.

References

1. Tanaka SA, Mahabir RC, Jupiter DC, Menezes JM. Updating the epidemiology of cleft lip with or without cleft palate. *Plast Reconstr Surg.* 2012;129(3):511e-518e.
2. International Perinatal Database of Typical Oral Clefts Working Group. Prevalence at birth of cleft lip with or without cleft palate: data from the International Perinatal Database of Typical Oral Clefts. *Cleft Palate Craniofac J.* 2011;48(1):66-81.
3. Fadeyibi IO, Coker OA, Zacchariah MP, Fasawe A, Ademiluyi SA. Psychosocial effects of cleft lip and palate on Nigerians: the Ikeja-Lagos experience. *J Plast Surg Hand Surg.* 2012;46(1):13-18.
4. Kirschner RE, LaRossa D. Cleft lip and palate. *Otolaryngol Clin North Am.* 2000;33(6):1191-1215, v-vi.
5. Bearn D, Mildinhall S, Murphy T, et al. Cleft lip and palate care in the United Kingdom--the Clinical Standards Advisory Group (CSAG) Study. Part 4: outcome comparisons, training, and conclusions. *Cleft Palate Craniofac J.* 2001;38(1):38-43.
6. Shaw WC, Semb G, Nelson P, et al. The Eurocleft project 1996-2000: overview. *J Craniomaxillofac Surg.* 2001;29(3):131-140; discussion 141-132.
7. Stock NM, Feragen KB. Psychological adjustment to cleft lip and/or palate: A narrative review of the literature. *Psychol Health.* 2016;31(7):777-813.
8. Russell K, Long RE, Jr., Hathaway R, et al. The Americleft study: an inter-center study of treatment outcomes for patients with unilateral cleft lip and palate part 5. General discussion and conclusions. *Cleft Palate Craniofac J.* 2011;48(3):265-270.
9. Rautio J, Andersen M, Bolund S, et al. Scandcleft randomised trials of primary surgery for unilateral cleft lip and palate: 2. Surgical results. *J Plast Surg Hand Surg.* 2017;51(1):14-20.
10. International Consortium of Health Outcomes Measurement (ICHOM). Data collection reference guide. https://ichom.org/files/medical-conditions/cleft-lip-palate/cleft-lip-palate-reference-guide.pdf.
11. Arora J, Haj M. Implementing ICHOM's Standard Sets of Outcomes: Cleft Lip and Palate at Erasmus University Medical Centre in the Netherlands. *London, UK: International Consortium for Health Outcomes Measurement (ICHOM), December 2016 (available at www.ichom.org).*
12. Porter ME. A strategy for health care reform--toward a value-based system. *N Engl J Med.* 2009;361(2):109-112.
13. Allori AC, Kelley T, Meara JG, et al. A standard set of outcome measures for the comprehensive appraisal of cleft care. *Cleft Palate Craniofac J.* 2017;54(5):540-554.
14. Klassen AF, Wong Riff KW, Longmire NM, et al. Psychometric findings and normative values for the CLEFT-Q based on 2434 children and young adult patients with cleft lip and/or palate from 12 countries. *CMAJ.* 2018;190(15):E455-E462.
15. Tsangaris E, Wong Riff KWY, Goodacre T, et al. Establishing Content Validity of the CLEFT-Q: A New Patient-reported Outcome Instrument for Cleft Lip/Palate. *Plast Reconstr Surg Glob Open.* 2017;5(4):e1305.
16. Wong Riff KW, 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(1):e015467.
17. Broder HL, McGrath C, Cisneros GJ. Questionnaire development: face validity and item impact testing of the Child Oral Health Impact Profile. *Community Dent Oral Epidemiol.* 2007;35 Suppl 1:8-19.
18. Stewart MG, Witsell DL, Smith TL, Weaver EM, Yueh B, Hannley MT. Development and validation of the Nasal Obstruction Symptom Evaluation (NOSE) scale. *Otolaryngol Head Neck Surg.* 2004;130(2):157-163.

19. McLeod S, Harrison LJ, McCormack J. The intelligibility in Context Scale: validity and reliability of a subjective rating measure. *J Speech Lang Hear Res.* 2012;55(2):648-656.

20. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: Building an international community of software platform partners. *J Biomed Inform.* 2019;95:103208.

21. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform.* 2009;42(2):377-381.

22. 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(4):302-312.

23. Zhang RS, Lin LO, Hoppe IC, et al. Nasal Obstruction in Children With Cleft Lip and Palate: Results of a Cross-Sectional Study Utilizing the NOSE Scale. *Cleft Palate Craniofac J.* 2018:1055665618772400.

24. van Zijl F, Timman R, Datema FR. Adaptation and validation of the Dutch version of the nasal obstruction symptom evaluation (NOSE) scale. *Eur Arch Otorhinolaryngol.* 2017;274(6):2469-2476.

25. Hobart J. Improving the evaluation of therapeutic intervention in MS: the role of new psychometric methods. *Health Technol Assess* 2009;13: iii, ix-x:1-177.

26. Rasch G. Probabilistic models for some intelligence and attainment tests. Vol. 1 of studies in mathematical psychology. *Copenhagen: Danmarks Paedagogiske Institut.* 1960.

27. Pallant JF, Tennant A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *Br J Clin Psychol.* 2007;46(Pt 1):1-18.

28. Hagquist C, Bruce M, Gustavsson JP. Using the Rasch model in nursing research: an introduction and illustrative example. *Int J Nurs Stud.* 2009;46(3):380-393.

29. Cano SJ, Hobart JC. The problem with health measurement. *Patient Prefer Adherence.* 2011;5:279-290.

30. Fisher Jr W. Reliability, separation, strata statistics. *Rasch Measurement Transactions.* 1992;6(3):238.

31. Sobol DL, Allori AC, Carlson AR, et al. Nasal Airway Dysfunction in Children with Cleft Lip and Cleft Palate: Results of a Cross-Sectional Population-Based Study, with Anatomical and Surgical Considerations. *Plast Reconstr Surg.* 2016;138(6):1275-1285.

32. Terwee CB, Prinsen CAC, Chiarotto A, et al. COSMIN methodology for evaluating the content validity of patient-reported outcome measures: a Delphi study. *Qual Life Res.* 2018;27(5):1159-1170.

33. Prinsen CAC, Mokkink LB, Bouter LM, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res.* 2018;27(5):1147-1157.

34. Carle AC, Weech-Maldonado R. Validly interpreting patients’ reports: using bifactor and multidimensional models to determine whether surveys and scales measure one or more constructs. *Med Care.* 2012;50(9 Suppl 2):S42-48.

35. Varni JW, Limbers CA, Burwinkle TM. How young can children reliably and validly self-report their health-related quality of life?: an analysis of 8,591 children across age subgroups with the PedsQL 4.0 Generic Core Scales. *Health Qual Life Outcomes.* 2007;5:1.

36. Bevans KB, Riley AW, Moon J, Forrest CB. Conceptual and methodological advances in child-reported outcomes measurement. *Expert Rev Pharmacoecon Outcomes Res.* 2010;10(4):385-396.

Supplemental Material

	Erasmus University Medical Center, Rotterdam (The Netherlands)	Duke University Hospital, Durham (United States of America)	Boston Children's Hospital, Boston (United States of America)	CLEFT-Q Phase 3 Study, McMaster University, Hamilton (Canada)
Collection period	November 2015 - April 2019	January 2017 - April 2019	July 2016 - July 2018	January 2018 - April 2019
Collection method	Electronically at home, 2 weeks before clinic visit	iPad at day of clinic visit	iPad at the time of an in-person clinic visit	iPad or paper-based
Collection tool	Gemstracker and Limesurvey	REDCap	REDCap	REDCap

Table 1 Data collection methods per participating cleft center.

Chapter 3

Optimizing the Psychosocial Function Measures in the International Consortium for Health Outcomes Measurement Standard Set for Cleft

Apon I, MD, MHS[1]; van Leeuwen N, PhD[2]; Koudstaal MJ, MD, DMD, PhD[1]; Allori AC, MD, PhD[3]; Rogers-Vizena CR, MD[4]; Wolvius EB, MD, DMD, PhD[1], Klassen AF, DPhil[5], Versnel SL, MD, PhD[6]

[1] Department of Oral and Maxillofacial Surgery, Erasmus University Medical Center, Rotterdam, The Netherlands
[2] Department of Public Health, Erasmus University Medical Center, Rotterdam, The Netherlands
[3] Department of Plastic, Maxillofacial and Oral Surgery, Duke Children's Hospital, Durham, North Carolina, USA
[4] Department of Plastic and Oral Surgery, Boston Children's Hospital, Boston, Massachusetts, USA
[5] Department of Pediatrics, McMaster University, Hamilton, Ontario, Canada
[6] Department of Plastic and Reconstructive Surgery, Erasmus University Medical Center, Rotterdam, The Netherlands

Plast Reconstr Surg. 2023 Feb 1;151(2):274e-281

Abstract

Background: To ensure the feasibility of implementing patient-reported outcome measures in clinical practice, they must be continually appraised for undue burden placed on patients and clinicians and their usefulness for decision-making. This study assesses correlations between the CLEFT-Q psychosocial scales in the ICHOM Standard Set for cleft and explores their associations with patient characteristics and psychosocial care referral.

Methods: Spearman's correlation coefficients were calculated for CLEFT-Q psychological function, social function, school function, face, speech function and speech-related distress scales. Logistic regressions were used to assess the association of cleft phenotype, syndrome, sex, and adoption status on scale scores and clinical referral to psychosocial care for further evaluation and management.

Results: Data were obtained from 3,067 patients with cleft lip and/or palate at three centers. Strong correlations were observed between social and psychological ($r > 0.69$) and school function ($r > 0.78$) scales. Correlation between school and psychological function was lower ($r = 0.59 - 0.68$). Genetic syndrome (OR = 2.37, 95% CI 1.04-5.41), psychological function (OR = 0.92, 95% CI 0.88-0.97), school function (OR = 0.94, 95% CI 0.90-0.98), and face scale (OR = 0.96, 95% CI 0.94-0.98) were significant predictors for referral to psychosocial care.

Conclusion: Since CLEFT-Q social function showed strong correlations with both school and psychological function, its additional value for measuring psychosocial function within the Standard Set is limited, and it is reasonable to consider removing this scale from the Standard Set.

Keywords: cleft lip and palate, patient-reported outcome measures, psychosocial function.

Introduction

The treatment of cleft lip and palate remains complex and differs among treatment centers resulting in varying quality of care worldwide.[1-4] Recognizing the need for uniform outcome measurement in cleft care, the International Consortium for Health Outcomes Measurement (ICHOM) published its Standard Set for the comprehensive appraisal of cleft lip and palate, based on consensus recommendations of a large, international, multidisciplinary working group.[5-7] The objective of the Standard Set was to provide a starting point for all cleft teams to measure the same outcome domains, using the same methods and instruments, at the same time points, and to record those data in the same structured format. The ultimate goal is to apply these outcomes toward improving patient-centered care.

The Standard Set includes not only traditional, clinician-reported outcomes and clinical variables, but also condition-specific parent- and patient-reported outcome measures (PROMs) in the domains of speech, facial appearance, and psychosocial function.[8] The rationale for including these scales is intuitive: facial appearance and speech are major outcomes domains in the treatment of cleft lip and/or palate (CL/P), and poor outcomes in these domains may contribute to psychosocial distress. In fact, psychosocial impairment has been commonly reported in patients with CL/P, with main contributing factors including patients being teased or bullied, dissatisfaction with appearance, and dissatisfaction with speech.[9-14] Therefore, it is important for cleft teams to identify psychosocial problems early in order to provide timely and appropriate care within the team or, if indicated, by referral to a psychologist or psychiatrist for further evaluation and management. To this end, the ICHOM Standard Set includes three psychosocial CLEFT-Q scales: psychological function, social function and school function.[15-17]

To ensure the feasibility of implementing these scales in clinical practice and sustainability of that implementation over time, the Standard Set itself must be continually appraised for undue burden placed on patients and clinicians. At the same time, it is important to confirm that scales included in the Standard Set manifest useful information that can be used to inform clinical decision-making. The corollary to this statement is that scales deemed suboptimal, statistically invalid, redundant, or otherwise uninformative should be de-implemented. To see whether the psychosocial scales in the ICHOM Standard Set prove their worth or whether its inclusion in the Standard Set should be reconsidered, this study assesses correlations between the psychosocial scales and explores their associations with patient characteristics and referrals to psychosocial care.

Methods

Participants and recruitment

Three cleft teams participated in this study (Boston Children's Hospital, Duke Children's Hospital, and Erasmus University Medical Center). Patients with unilateral or bilateral cleft lip, cleft palate, cleft lip and palate, or cleft lip and alveolus, aged 8 to 22 years, were eligible for the measurement of CLEFT-Q psychosocial function scales. Data were prospectively collected according to the guidelines of the ICHOM Standard Set for Cleft Lip and Palate between November 2015 and April 2019. CLEFT-Q psychological and school function outcome data was collected at time point 't_{12}' (10-13 years), and data for the CLEFT-Q social function at time points 't_8' (8-9 years) and 't_{final}' (20-22 years), as recommended by the Standard Set.[18] Since the authors felt that there is a large measurement gap between the ages of 12 and 22 years, the time points 't_{15}' (14-16 years) and 't_{17}' (17-19 years) were added for research purposes. This additional data was collected as part of the CLEFT-Q development and validation project organized by McMaster University. Information on this project, its recruitment and data-collection procedures were described previously.[15,16,19] Institutional Review Board approval was obtained at every participating site (MEC-2016-156; IRB-P00030776; IRB-Pro00067808; REB Project #10-651).

Patient-reported outcome measures

The primary outcomes were the scores of the CLEFT-Q psychological function, social function and school function scales. These scales focus on the themes, "how do you feel?", "how is your social life?", and "how is your school life?", respectively. Each scale consists of ten items, with four possible responses of "never", "sometimes", "often", and "always".

In addition, it was hypothesized that psychosocial function is influenced by facial difference or speech dysfunction. Therefore, the CLEFT-Q face, speech function and speech-related distress scales, which are prescribed by the ICHOM Standard Set as well, were also evaluated. The CLEFT-Q face scale focuses on the theme, "how much do you like how your face looks?" and includes nine items with possible responses "not at all", "a little bit", "quite a bit", and "very much". The CLEFT-Q speech function (12 items) and speech-related distress (10 items) scales, assess "how is your speech?" and "how do you feel about speaking?", respectively, with possible responses "always", "sometimes", and "never", For each scale, a raw score was transformed to a scale ranging from 0 to 100, where higher scores represent better functioning.

Additional variables

The Standard Set also includes the collection of various patient characteristics (also known as case-mix variables or predictors), including sex, age (grouped according to Standard Set "time points"), cleft phenotypic group, the presence of a genetic syndrome, and whether or not a child has been adopted. For a subgroup analysis of the patients from the Netherlands, Dutch socioeconomic status scores of 2017 were added as an additional variable to explore its association with outcomes. These scores are based on postal codes and higher scores represent higher socioeconomic status.[20] Also, information on referral status of the patient to any type of psychosocial care (psychiatrist, psychologist, social care) was gathered retrospectively from the patient's medical files.

Statistical analysis

All analyses were performed using SPSS software (IBM SPSS Statistics for Windows, Version 25.0, released 2017, IBM Corp.) For the correlational analyses, Spearman's correlation coefficients (r) were calculated for every relationship between the different PROM scores per time point, since the PROM scores were not normally distributed based on the histograms. A priori, it was defined that a coefficient above 0.7 was to illustrate a strong correlation, whereas between 0.4 and 0.7 was considered moderate, and coefficients below 0.4 were considered weak. Stronger correlations between scales would indicate similar constructs are measured. Univariable linear regression was performed to assess the influence of the time points on the psychosocial outcome scores.

Subgroup analysis was performed on the Dutch sample and included univariable linear regression to investigate the influence of patient characteristics on the psychosocial function scores, and logistic regression to explore the associations of patient characteristics and psychosocial scores with referral to psychosocial care. All analyses were performed based on complete cases. The two-tailed significance level was set at $p < 0.05$.

Results

The complete dataset included 3,067 patients who provided a total of 3,103 assessments. The majority of patients were diagnosed with cleft lip and palate (n=1,773 (58%)) and 1,714 (56%) patients were male. In 1,080 (35%) cases, the PROMs were completed around the age of 12 (**Table 1**).

	Complete sample	Subset
Characteristics	**Patients, No. (%) Total N=3067**	**Patients, No. (%) Total N=353**
Sex		
Male	1714 (56)	200 (57)
Female	1353 (44)	153 (43)
Cleft type		
Cleft lip	301 (10)	34 (10)
Cleft palate	718 (23)	117 (33)
Cleft lip and palate	1773 (58)	172 (49)
Cleft lip and alveolus	275 (9)	30 (8)
Adoption		
No	-	292 (83)
Yes	-	61 (17)
Unknown	-	-
Genetic syndrome		
No	-	303 (86)
Yes	-	50 (14)
Socio-economic status (Mean (range))	**-**	0.06 (-3.63 – 2.31)
	Measurements, No (%) Total N=3103	**Measurements, No (%) Total N=365**
Timing of PROMs		
8 years (t_8)	735 (24)	134 (37)
12 years (t_{12})	1080 (35)	154 (42)
15 years (t_{15})	593 (19)	-
17 years (t_{17})	386 (12)	-
22 years (t_{final})	309 (10)	77 (21)
Psychosocial care referral		
No	-	330 (90)
Yes, after PROM scores	-	18 (5)
Yes, other reason than PROM scores	-	17 (5)

Table 1 Patient characteristics and variables.

Strong correlations were found between psychological and social scales (r = 0.74 – 0.76) at all measured time points, except at t_8 (r = 0.69). The correlations between social and school function scales were between 0.78 and 0.85. This correlation could not be computed for t_{final} because the school scale was not completed at this time point. The correlations between psychological and school function scales varied between 0.59 and 0.68 (see **Supplemental Material – Figure 1**). The face scale was moderately correlated (r = 0.37 - 0.65) and the speech-related scales had low to moderate correlations (r = 0.14 – 0.53) with all three psychosocial scales (**Table 2**). Similar findings were found for each of the four cleft phenotypic groups (see **Supplemental Material – Table 1**).

	t_8	t_{12}	t_{15}	t_{17}	t_{final}
Psych - Social	0.69, p=0.00*	0.74, p=0.00*	0.76, p=0.00*	0.75, p=0.00*	0.75, p=0.00*
Psych - School	0.59, p=0.00*	0.68, p=0.00*	0.66, p=0.00*	0.67, p=0.00*	N/C
Social - School	0.80, p=0.00*	0.85, p=0.00*	0.82, p=0.00*	0.78, p=0.00*	N/C
Psych - Face	0.60, p=0.00*	0.61, p=0.00*	0.59, p=0.00*	0.61, p=0.00*	0.65, p=0.00*
Social - Face	0.46, p=0.00*	0.55, p=0.00*	0.48, p=0.00*	0.51, p=0.00*	0.58, p=0.00*
School - Face	0.37, p=0.00*	0.45, p=0.00*	0.40, p=0.00*	0.49, p=0.00*	N/C
Psych – Speech distress	0.24. p=0.00*	0.35, p=0.00*	0.34, p=0.00*	0.39, p=0.00*	0.28, p=0.00*
Psych – Speech function	0.22, p=0.00*	0.27, p=0.00*	0.26, p=0.00*	0.14, p=0.03*	0.14, p=0.09
Social – Speech distress	0.42, p=0.00*	0.49, p=0.00*	0.46, p=0.00*	0.53, p=0.00*	0.46, p=0.00*
Social – Speech function	0.37, p=0.00*	0.41, p=0.00*	0.40, p=0.00*	0.30, p=0.00*	0.26, p=0.00*
School – Speech distress	0.36, p=0.00*	0.41, p=0.00*	0.39, p=0.00*	0.46, p=0.00*	N/C
School – Speech function	0.33, p=0.00*	0.32, p=0.00*	0.30, p=0.00*	0.31, p=0.00*	N/C

Table 2 Spearman's rho correlation coefficient. N/C = could not be computed, because of too few observations in school function scale at t_{final}. * statistically significant.

Linear regression revealed a negative significant association between time points and outcome scores of the psychological function scale; a higher age group was associated with lower scores (t_{12}: β = -3.30, 95% confidence interval (-5.33, -1.56); t_{15}: -6.70 (-8.99, -4.41); t_{17}: -8.87 (-11.44, -6.29); t_{final}: -8.71 (-11.70, -5.72)). No significant associations between time points and the social and school scales were found (**Table 3**).

	CLEFT-Q psychological			**CLEFT-Q social**			**CLEFT-Q school**		
Time points	**B**	**CI (95%)**	**p-value**	**B**	**CI (95%)**	**p-value**	**B**	**CI (95%)**	**p-value**
t_8 (*ref*)	*78.48*			*73.04*			*75.18*		
t_{12}	-3.30	-5.33;-1.56	0.00*	1.27	-0.58;3.13	0.18	-0.06	-2.04;1.93	0.96
t_{15}	-6.70	-8.99;-4.41	0.00*	-0.04	-2.07;1.98	0.97	0.83	-1.43;3.08	0.47
t_{17}	-8.87	-11.44;-6.29	0.00*	-0.45	-2.75;1.86	0.71	-0.37	-3.80;3.06	0.83
t_{final}	-8.71	-11.70;-5.72	0.00*	-1.71	-4.25;0.83	0.19	-7.18	-32.08;17.72	0.57

Table 3 Linear univariable regression analysis for different time points per psychosocial function scale. B indicates regression coefficient, CI confidence interval, *ref* reference group. * statistically significant.

The Dutch subset included 353 patients who provided 365 measurements. The phenotypic group of cleft lip and palate included 172 (49%) patients, 200 (57%) were male and PROMs were mostly completed around the age of 12 (n=154, 42%) (**Table 1**). No statistically significant associations were found between patient characteristics and PROM scores (**Table 4**).

	CLEFT-Q psychological			CLEFT-Q social			CLEFT-Q school		
Variables	**B**	**CI (95%)**	**p-value**	**B**	**CI (95%)**	**p-value**	**B**	**CI (95%)**	**p-value**
Sex									
Male *(ref)*									
Female	-0.65	-6.86;5.57	0.84	0.40	-6.56;7.36	0.91	-0.47	-6.38;5.45	0.88
Cleft type									
Cleft lip and palate *(ref)*									
Cleft lip	-4.99	-18.12;8.14	0.45	0.71	-10.29;11.71	0.90	-2.97	-15.56;9.62	0.64
Cleft palate	2.78	-3.98;9.54	0.42	-1.91	-10.30;6.47	0.65	1.53	-4.95;8.01	0.64
Cleft lip and alveolus	-4.57	-15.24;6.09	0.40	3.63	-14.36;21.62	0.69	-0.78	-11.00;9.45	0.88
Adoption									
No *(ref)*									
Yes	-2.17	-10.08;5.74	0.59	0.71	-8.49;9.90	0.88	1.88	-5.65;9.40	0.62
Genetic syndrome									
No *(ref)*									
Yes	1.98	-6.91;10.87	0.66	-3.20	-13.80;7.43	0.55	0.99	-7.47;9.45	0.82
Socio-economic status	0.01	-2.64;2.66	0.99	2.95	-0.61;6.50	0.10	0.91	-1.61;3.43	0.48

Table 4 Univariable linear analysis for psychological function, social function and school function outcomes based on data collected at Erasmus University Medical Center. B indicates regression coefficient, CI confidence interval, *ref* reference group.

In total, 35 (10%) patients were referred to psychosocial care from which 18 (5%) patients were referred based on the result of a PROM score, and 17 (5%) for another reason (**Table 1**). The majority of the referred patients were diagnosed with cleft lip and palate (n=21, 60%), male (n=23, 66%) and approximately 8 years of age (n=17, 49%). Concerns about appearance or speech were detected by the PROMs, while non-PROM related reasons for referral consisted of anxiety, behavioral and coping problems. Patients referred due to PROM scores were most likely to score low on the psychological and face scales with mean scores of 26 (range 19 – 32) and 40 (range 0 – 59), respectively (see **Supplemental Material – Table 2;** see **Supplemental Material – Table 3**).

The score for the psychological function scale was significantly associated with referral to psychosocial care (odds ratio (OR) = 0.92, 95% confidence interval (95% CI) 0.88 - 0.97). Similar effects were found for the school (OR 0.94, 95% CI 0.90 - 0.98) and face (OR 0.96, 95% CI 0.94 - 0.98) scale scores. The presence of a genetic syndrome was significantly associated with referrals (OR 2.37, 95% CI 1.04 - 5.41), whereas other patient characteristics were not (**Table 5**).

	Referral to psychosocial work		
Variables	**OR**	**CI (95%)**	**p-value**
Sex			
Male *(ref)*			
Female	0.65	0.31;1.35	0.25
Cleft type			
Cleft lip and palate *(ref)*			
Cleft lip	0.96	0.31;2.99	0.94
Cleft palate	0.45	0.19;1.10	0.08
Cleft lip and alveolus	0.83	0.23;2.96	0.77
Adoption			
No *(ref)*			
Yes	1.22	0.51;2.94	0.65
Genetic syndrome			
No *(ref)*			
Yes	2.37	1.04;5.41	*0.04
Timing of PROMs			
t_8 *(ref)*			
t_{12}	0.69	0.33;1.46	0.33
t_{final}	0.38	0.12;1.17	0.09
CLEFT-Q Psychological function	0.92	0.88;0.97	*0.00
CLEFT-Q Social function	0.96	0.92;1.00	0.06
CLEFT-Q School function	0.94	0.90;0.98	*0.01
CLEFT-Q Face	0.96	0.94;0.98	*0.00
CLEFT-Q Speech distress	0.98	0.95;1.01	0.12
CLEFT-Q Speech function	0.99	0.97;1.02	0.56
Socio-economic status	0.83	0.61;1.13	0.24

Table 5 Univariable logistic analysis with odds ratios (OR) for referral to psychosocial care after completion of PROMs. OR indicates odds ratio, CI confidence interval, *ref* reference group. * Statistically significant.

Discussion

This study evaluated the correlations between the CLEFT-Q psychosocial scales that are recommended by the ICHOM Standard Set to determine whether each scale measures a unique construct or overlaps other scales. The CLEFT-Q social function scale measures a construct very similar to the CLEFT-Q school function scale and also has significant common ground with the psychological function scale, whereas the correlation between school function and psychological function was more modest. This suggests that the school function scale addresses a particular aspect of psychosocial function that the other instruments don’t capture, namely aspects related to the social environment at school.

In contrast, the social function scale does not contribute much unique information, as it overlaps much with the school and psychological function scales. In other words, the Standard Set might be limited to administering only the CLEFT-Q psychological and school function scales, without losing any relevant information. Dropping the social function scale will reduce the number of questions by ten, helping to reduce the burden for both patient and clinical team and making the outcomes measurement project more sustainable in the long run. In situations where a child does not attend school, the CLEFT-Q social function scale may serve as a reasonable alternative.

The finding of moderate correlations between the CLEFT-Q face and psychosocial scales suggests that a patient's psychosocial functioning is influenced by a patient's subjective appraisal of facial appearance. This finding was confirmed in the subgroup analysis performed on the Dutch dataset where patients with a visible cleft lip achieve lower scores on the psychological and school function scales than patients with cleft palate. The weak correlations between the two speech-related scales and the three psychosocial scales suggests that a patient is able to achieve high scores for psychosocial function while experiencing speech problems. This finding is in concordance with a large study on CLEFT-Q normative scores where only small differences in mean scores of psychological, school and social scales between patients with a clinically moderate to severe speech problem and patients with mild or no speech problems were found.[19]

Addressing the measurement gap during teenage years

The ICHOM Standard Set presently has a measurement gap, as there are no assessments done on patients between the ages of 12 and 22 years. The teenage years are very important since young people undergo puberty and experience many changes in their physical and psychosocial development. To improve the possibilities for longitudinal follow-up and future benchmarking projects, administering the psychological and school function scales at the age of 15 and 17 years of age would provide important additional information about psychosocial adjustment. Importantly, the CLEFT-Q psychological function scale could provide useful information regarding a patient's functioning around young adolescence, since regression analysis showed a decreasing trend in outcome score over time, suggesting that this scale is most sufficient to intervene upon.

Influence of patient and demographic characteristics on psychosocial functioning

The second part of this study performed on the Dutch subset of data, exploring associations between patient's clinical and demographic characteristics and the PROM scales, did not find any statistically significant associations. This finding may be due to limitations of the demographic and clinical information that is collected according to the Standard Set. A recently published study on language proficiency of parents from children with craniofacial anomalies, including patients with cleft lip and palate and cleft palate, demonstrated that parental limited English proficiency was a risk factor for the development of psychosocial distress in terms of higher anger, anxiety, depression, and poor peer relationships.[21] Other variables that could be thought to be of influence are family composition (such as siblings or divorce), parental income and level of education or the child's educational performance. The limitations of the present exploratory project precluded us from including these variables, but we recommend that future prospective studies dealing with psychosocial well-being or functioning take them into account.[13,22]

For the subgroup analysis of the patients from the Netherlands, the socioeconomic status scores were added to explore their value as a patient characteristic, since little is known about its influence on patient-reported outcome scores in patients with a cleft. A non-significant trend showed that children with lower socioeconomic status scores reported lower scores on the psychosocial function scales and were more likely to be referred to psychosocial care. Unfortunately, for international use in benchmarking projects the generalizability of this finding is limited, since these status scores are specific for Dutch regions and therefore not directly transferrable to other countries. Education, income and profession are three other indicators for socioeconomic status.[23] Collecting this data in future research could provide more generalizable insights.

The use of PROMs in referring patients to psychosocial care

Patients who reported poorer outcomes on either the psychological function, school function, and/or face scales were more likely to have been referred for psychosocial evaluation and management. The cleft team actively used the scales to review symptoms. A poor score on one of these psychosocial scales could prompt the clinician to investigate further and make the appropriate referral when concerned about the patient's psychosocial health. Patients who were referred because of other reasons than PROM scores mainly experienced anxiety or behavioral problems. Previous literature showed that Social Anxiety Disorder is more prevalent in children with cleft lip and palate compared to a healthy control group.[24] Also, higher levels of social anxiety were found in adults compared to adolescents with cleft lip and palate, while dental treatment anxiety was highest in children aged 4 to 6 years old.[25,26] Therefore, it could be taken into consideration to include a valid screening tool, parent-reported at the young age and patient-reported from the age of 8, for measuring anxiety problems.

Strengths and limitations

A strength of this study is the large international sample for the correlational analyses, though we recognize the second part of this study was limited to a smaller population recruited at one university hospital. Even though patient demographics were comparable between these two datasets, the smaller sample size could have resulted in less power to detect clinically relevant differences when evaluating associative relationships. When closing data collection for this study, the ICHOM Standard Set was implemented for four years. This has resulted in a cross-sectional study design, in which the maximum possible follow-up period between two measurements was 4 years (t_8 and t_{12}) and very few longitudinal data were gathered. Therefore, results should be interpreted on a group level rather than on an individual patient level and do not reflect a patient's psychosocial well-being in the long-term.

Conclusion

This is the first study to explore the psychosocial domain within the ICHOM Standard Set and specifically six CLEFT-Q scales of psychological function, social function, school function, face, speech function and speech-related distress. Since the CLEFT-Q social function scale showed strong overlap with both psychological and school scales, its value is limited and inclusion in the Standard Set should be reconsidered. Only including the CLEFT-Q psychological and school function scales is recommended. Further recommendations are expansion of required time points to include the teenage years (e.g., 15 and 17 years of age) and addition of expanded demographic and socioeconomic variables.

Acknowledgements

We would like to thank all members of the participating cleft teams for data collection.

Conflicts of Interest

The CLEFT-Q is owned by McMaster University and The Hospital for Sick Children. Anne F. Klassen is a co-developer of the CLEFT-Q and, as such, could potentially receive a share of any license revenues as royalties based on her institution's inventor sharing policy.

Data collected as part of the CLEFT-Q development and validation project was supported by a grant from the Canadian Institutes of Health Research. The other authors have nothing to disclose for.

References

1. Al-Omari I, Millett DT, Ayoub AF. Methods of assessment of cleft-related facial deformity: a review. *Cleft Palate Craniofac J.* 2005;42(2):145-156.
2. de Ladeira PR, Alonso N. Protocols in cleft lip and palate treatment: systematic review. *Plast Surg Int.* 2012;2012:562892.
3. Russell K, Long RE, Jr., Hathaway R, et al. The Americleft study: an inter-center study of treatment outcomes for patients with unilateral cleft lip and palate part 5. General discussion and conclusions. *Cleft Palate Craniofac J.* 2011;48(3):265-270.
4. Shaw WC, Semb G, Nelson P, et al. The Eurocleft project 1996-2000: overview. *J Craniomaxillofac Surg.* 2001;29(3):131-140; discussion 141-132.
5. Porter ME. A strategy for health care reform--toward a value-based system. *N Engl J Med.* 2009;361(2):109-112.
6. Porter ME. Value-based health care delivery. *Ann Surg.* 2008;248(4):503-509.
7. Porter ME. What is value in health care? *N Engl J Med.* 2010;363(26):2477-2481.
8. Allori AC, Kelley T, Meara JG, et al. A standard set of outcome measures for the comprehensive appraisal of cleft care. *Cleft Palate Craniofac J.* 2017;54(5):540-554.
9. Feragen KB, Borge AI, Rumsey N. Social experience in 10-year-old children born with a cleft: exploring psychosocial resilience. *Cleft Palate Craniofac J.* 2009;46(1):65-74.
10. Feragen KB, Saervold TK, Aukner R, Stock NM. Speech, Language, and Reading in 10-Year-Olds With Cleft: Associations With Teasing, Satisfaction With Speech, and Psychological Adjustment. *Cleft Palate Craniofac J.* 2017;54(2):153-165.
11. Feragen KB, Stock NM. Risk and Protective Factors at Age 10: Psychological Adjustment in Children With a Cleft Lip and/or Palate. *Cleft Palate Craniofac J.* 2016;53(2):161-179.
12. Hoek IH, Kraaimaat FW, Admiraal RJ, Kuijpers-Jagtman AM, Verhaak CM. [Psychosocial adjustment in children with a cleft lip and/or palate] Sociaal-emotionele gezondheid bij kinderen met schisis. *Ned Tijdschr Geneeskd.* 2009;153:B352.
13. Stock NM, Feragen KB. Psychological adjustment to cleft lip and/or palate: A narrative review of the literature. *Psychol Health.* 2016;31(7):777-813.
14. Hunt O, Burden D, Hepper P, Stevenson M, Johnston C. Parent reports of the psychosocial functioning of children with cleft lip and/or palate. *Cleft Palate Craniofac J.* 2007;44(3):304-311.
15. Tsangaris E, Wong Riff KWY, Goodacre T, et al. Establishing Content Validity of the CLEFT-Q: A New Patient-reported Outcome Instrument for Cleft Lip/Palate. *Plast Reconstr Surg Glob Open.* 2017;5(4):e1305.
16. Wong Riff KW, 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(1):e015467.
17. Harrison CJ, Rae C, Tsangaris E, et al. Further construct validation of the CLEFT-Q: Ability to detect differences in outcome for four cleft-specific surgeries. *J Plast Reconstr Aesthet Surg.* 2019;72(12):2049-2055.
18. International Consortium of Health Outcomes Measurement (ICHOM). Data collection reference guide. 2020; https://ichom.org/files/medical-conditions/cleft-lip-palate/cleft-lip-palate-reference-guide.pdf. Accessed July 1, 2020.

19. Klassen AF, Wong Riff KW, Longmire NM, et al. Psychometric findings and normative values for the CLEFT-Q based on 2434 children and young adult patients with cleft lip and/or palate from 12 countries. *CMAJ.* 2018;190(15):E455-E462.

20. Sociaal en Cultureel Planbureau (SCP). SCP Statusscores (2018). http://www.scp.nl/Formulieren/Statusscores_opvragen. Accessed April 4, 2019.

21. De Leon FS, Pfaff MJ, Volpicelli EJ, et al. Effect of Parental English Proficiency on Psychosocial Functioning in Children with Craniofacial Anomalies. *Plast Reconstr Surg.* 2020;145(3):764-773.

22. Stock NM, Hammond V, Hearst D, et al. Achieving Consensus in the Measurement of Psychological Adjustment to Cleft Lip and/or Palate at Age 8+ Years. *Cleft Palate Craniofac J.* 2020;57(6):746-752.

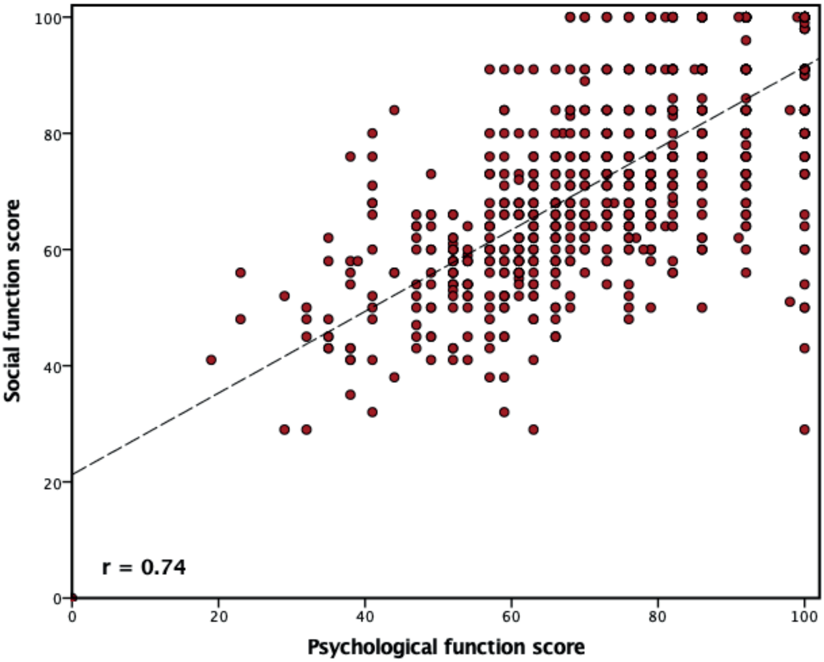
23. Shavers VL. Measurement of socioeconomic status in health disparities research. *J Natl Med Assoc.* 2007;99(9):1013-1023.

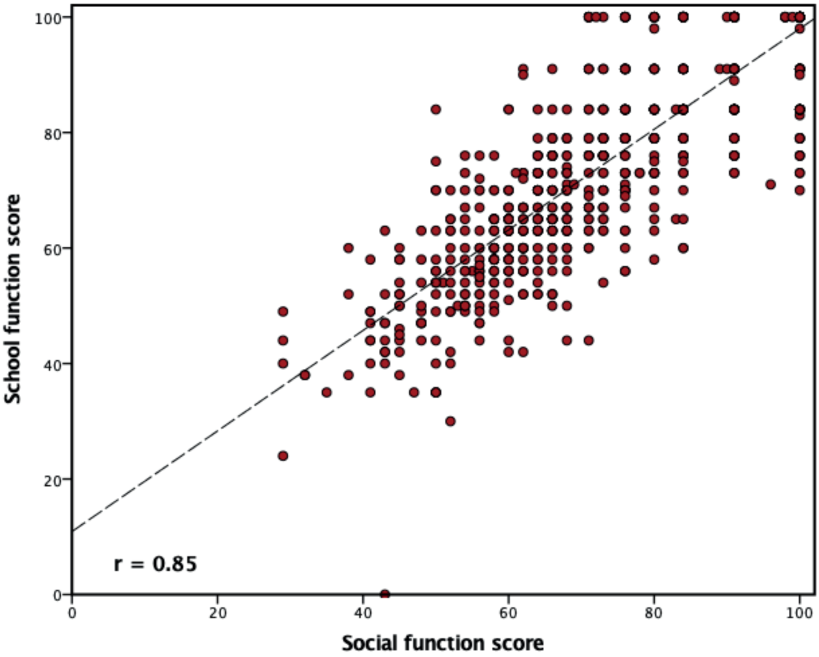
24. Demir T, Karacetin G, Baghaki S, Aydin Y. Psychiatric assessment of children with nonsyndromic cleft lip and palate. *Gen Hosp Psychiatry.* 2011;33(6):594-603.

25. Cheung LK, Loh JS, Ho SM. Psychological profile of Chinese with cleft lip and palate deformities. *Cleft Palate Craniofac J.* 2007;44(1):79-86.

26. Vogels WE, Aartman IH, Veerkamp JS. Dental fear in children with a cleft lip and/or cleft palate. *Cleft Palate Craniofac J.* 2011;48(6):736-740.

Supplemental Material





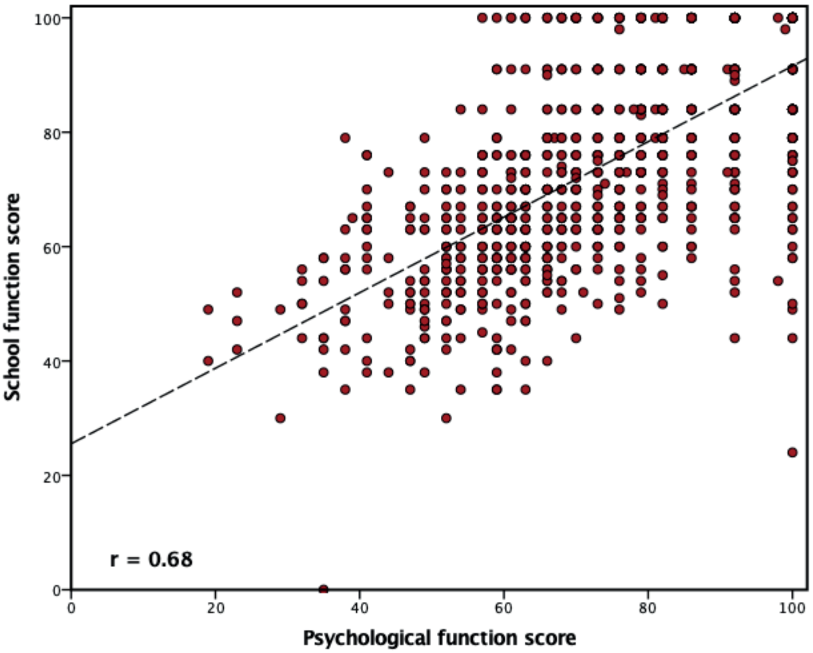


Figure 1 Spearman correlation plots for the psychosocial scales at t_{12}.

t_8	**CL**	**CP**	**CLAP**	**CLA**
Psych - Social	0.55, p=0.00*	0.68, p=0.00*	0.71, p=0.00*	0.63, p=0.00*
Psych - School	0.51, p=0.00*	0.53, p=0.00*	0.62, p=0.00*	0.52, p=0.00*
Social - School	0.65, p=0.00*	0.83, p=0.00*	0.80, p=0.00*	0.78, p=0.00*
Psych - Face	0.52, p=0.00*	0.61, p=0.00*	0.61, p=0.00*	0.60, p=0.00*
Social - Face	0.57, p=0.00*	0.42, p=0.00*	0.47, p=0.00*	0.50, p=0.00*
School - Face	0.28, p=0.06	0.31, p=0.00*	0.40, p=0.00*	0.37, p=0.01*
Psych – Speech distress	N/C	0.12, p=0.28	0.26, p=0.00*	0.26, p=0.16
Psych – Speech function	N/C	0.22, p=0.04*	0.21, p=0.00*	0.41, p=0.02*
Social – Speech distress	N/C	0.37, p=0.00*	0.41, p=0.00*	0.49, p=0.01*
Social – Speech function	N/C	0.38, p=0.00*	0.33, p=0.00*	0.64, p=0.00*
School – Speech distress	N/C	0.29, p=0.01*	0.39, p=0.00*	0.27, p=0.16
School – Speech function	N/C	0.42, p=0.00*	0.28, p=0.00*	0.49, p=0.01*
t_{12}	**CL**	**CP**	**CLAP**	**CLA**
Psych - Social	0.75, p=0.00*	0.72, p=0.00*	0.74, p=0.00*	0.75, p=0.00*
Psych - School	0.64, p=0.00*	0.67, p=0.00*	0.69, p=0.00*	0.70, p=0.00*
Social - School	0.84, p=0.00*	0.85, p=0.00*	0.85, p=0.00*	0.82, p=0.00*
Psych - Face	0.68, p=0.00*	0.52, p=0.00*	0.65, p=0.00*	0.46, p=0.00*
Social - Face	0.56, p=0.00*	0.47, p=0.00*	0.58, p=0.00*	0.61, p=0.00*
School - Face	0.41, p=0.00*	0.36, p=0.00*	0.51, p=0.00*	0.37, p=0.00*
Psych – Speech distress	N/C	0.34, p=0.00*	0.37, p=0.00*	0.23, p=0.18
Psych – Speech function	N/C	0.20, p=0.00*	0.32, p=0.00*	0.23, p=0.17
Social – Speech distress	N/C	0.51, p=0.00*	0.47, p=0.00*	0.47, p=0.01*
Social – Speech function	N/C	0.39, p=0.00*	0.42, p=0.00*	0.49, p=0.01*
School – Speech distress	N/C	0.41, p=0.00*	0.41, p=0.00*	0.38, p=0.02*
School – Speech function	N/C	0.30, p=0.00*	0.33, p=0.00*	0.47, p=0.00*
t_{15}	**CL**	**CP**	**CLAP**	**CLA**
Psych - Social	0.75, p=0.00*	0.76, p=0.00*	0.77, p=0.00*	0.74, p=0.00*
Psych - School	0.57, p=0.00*	0.65, p=0.00*	0.67, p=0.00*	0.64, p=0.00*
Social - School	0.75, p=0.00*	0.80, p=0.00*	0.83, p=0.00*	0.80, p=0.00*
Psych - Face	0.63, p=0.00*	0.66, p=0.00*	0.57, p=0.00*	0.57, p=0.00*
Social - Face	0.55, p=0.00*	0.46, p=0.00*	0.47, p=0.00*	0.57, p=0.00*
School - Face	0.34, p=0.01*	0.46, p=0.00*	0.38, p=0.00*	0.40, p=0.00*
Psych – Speech distress	N/C	0.26, p=0.01*	0.39, p=0.00*	0.19, p=0.26
Psych – Speech function	N/C	0.24, p=0.01*	0.26, p=0.00*	0.38, p=0.02*
Social – Speech distress	N/C	0.40, p=0.00*	0.51, p=0.00*	0.27, p=0.10
Social – Speech function	N/C	0.43, p=0.00*	0.40, p=0.00*	0.38, p=0.02*
School – Speech distress	N/C	0.32, p=0.00*	0.45, p=0.00*	0.19, p=0.28
School – Speech function	N/C	0.29, p=0.00*	0.34, p=0.00*	0.16, p=0.35
t_{17}	**CL**	**CP**	**CLAP**	**CLA**
Psych - Social	0.68, p=0.00*	0.68, p=0.00*	0.78, p=0.00*	0.75, p=0.00*
Psych - School	0.17, p=0.66	0.73, p=0.00*	0.67, p=0.00*	0.57, p=0.07
Social - School	0.73, p=0.03*	0.84, p=0.00*	0.78, p=0.00*	0.63, p=0.04*
Psych - Face	0.41, p=0.01*	0.72, p=0.00*	0.60, p=0.00*	0.67, p=0.00*
Social - Face	0.30, p=0.08	0.59, p=0.00*	0.51, p=0.00*	0.40, p=0.04*

School - Face	0.24, p=0.54	0.60, p=0.01*	0.48, p=0.00*	0.48, p=0.14
Psych – Speech distress	N/C	0.32, p=0.01*	0.41, p=0.00*	0.66, p=0.01*
Psych – Speech function	N/C	0.08, p=0.56	0.15, p=0.04*	0.15, p=0.58
Social – Speech distress	N/C	0.41, p=0.00*	0.54, p=0.00*	0.68, p=0.00*
Social – Speech function	N/C	0.30, p=0.03*	0.28, p=0.00*	0.30, p=0.25
School – Speech distress	N/C	0.54, p=0.02*	0.46, p=0.00*	0.77, p=0.04*
School – Speech function	N/C	0.60, p=0.01*	0.25, p=0.02*	0.20, p=0.66
t_{final}	**CL**	**CP**	**CLAP**	**CLA**
Psych - Social	0.80, p=0.00*	0.88, p=0.00*	0.72, p=0.00*	0.87, p=0.00*
Psych - School	N/C	N/C	N/C	N/C
Social - School	N/C	N/C	N/C	N/C
Psych - Face	0.74, p=0.00*	0.71, p=0.00*	0.61, p=0.00*	0.63, p=0.00*
Social - Face	0.68, p=0.00*	0.74, p=0.00*	0.56, p=0.00*	0.74, p=0.00*
School - Face	N/C	N/C	N/C	N/C
Psych – Speech distress	N/C	0.33, p=0.07	0.25, p=0.01*	0.38, p=0.31
Psych – Speech function	N/C	0.08, p=0.66	0.15, p=0.10	0.16, p=0.71
Social – Speech distress	N/C	0.58, p=0.00*	0.40, p=0.00*	0.58, p=0.10
Social – Speech function	N/C	0.26, p=0.10	0.24, p=0.00*	0.26, p=0.53
School – Speech distress	N/C	N/C	N/C	N/C
School – Speech function	N/C	N/C	N/C	N/C

Table 1 Spearman's correlation coefficients. N/C = could not be computed, because of too few observations. CL = cleft lip, CP = cleft palate, CLAP = cleft lip and palate, CLA = cleft lip and alveolus. * statistically significant.

	All referred patients, N (%)	Patients referred after PROMs, N	Patients referred for other reasons, N
Cleft type			
Cleft lip	4 (11)	3	1
Cleft palate	7 (20)	3	4
Cleft lip and palate	21 (60)	11	10
Cleft lip and alveolus	3 (9)	0	3
Sex			
Male	23 (66)	11	12
Female	12 (34)	6	6
Timing			
t_8	17 (49)	6	11
t_{12}	14 (40)	7	7
t_{final}	4 (11)	4	0
Genetic syndrome			
No	26 (74)	13	13
Yes	9 (26)	4	5
Adoption			
No	28 (80)	13	15
Yes	7 (20)	4	3

Table 2 Characteristics of patients referred to psychosocial care.

	All referrals	Referred after PROMs	Referred for other reasons
Anxiety	9	1	8
Behavioral problems	4	0	4
Insecurities	5	5	0
Developmental problems	1	0	1
Coping problems	5	1	4
Dissatisfaction (with appearance or speech)	11	11	0
Total	35	18	17

Table 3 Reasons for referral to psychosocial care, collected by themes.

Chapter 4

The Development, Deployment, and Evaluation of the CLEFT-Q Computerized Adaptive Test: A Multi-methods Approach, Contributing to Personalized, Person-centered Health Assessments in Plastic Surgery

Harrison CJ, MD, PhD[1]; **Apon I, MD, MHS[2]**; Ardouin K, PhD[3,4]; Sidey-Gibbons CJ, PhD[5]; Klassen AF, DPhil[6]; Cano SJ, PhD[7]; Wong Riff KWY, MD, PhD[8]; Pusic AL, MD, PhD[9]; Versnel SL, MD, PhD[10]; Koudstaal MJ, MD, DMD, PhD[2]; Allori AC, MD, PhD[11]; Rogers-Vizena CR, MD[12]; Swan MC, PhD[13]; Furniss D, PhD[1]; Rodrigues JN, PhD[14,15]

[1] Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK
[2] Department of Oral and Maxillofacial Surgery, the Dutch Craniofacial Center, Erasmus University Medical Center, Rotterdam, the Netherlands
[3] Cleft Lip and Palate Association, London, UK
[4] Department of Psychology, Speech and Hearing, University of Canterbury, Christchurch, New Zealand
[5] MD Anderson Center for INSPiRED Cancer Care, the University of Texas, Houston, Texas, USA
[6] Department of Pediatrics, McMaster University, Hamilton, Ontario, Canada
[7] Modus Outcomes, Letchworth Garden City, UK
[8] Department of Plastic and Reconstructive Surgery, Hospital for Sick Children, Toronto, Ontario, Canada
[9] Patient-Reported Outcomes, Values & Experience Center, Department of Surgery, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA
[10] Department of Plastic and Reconstructive Surgery, the Dutch Craniofacial Center, Erasmus University Medical Center, Rotterdam, the Netherlands
[11] Division of Plastic, Maxillofacial & Oral Surgery, Duke University Hospital & Children's Health Center, Durham, North Carolina, USA
[12] Department of Plastic and Oral Surgery, Boston Children's Hospital, Boston, Massachusetts, USA
[13] The Spires Cleft Centre, John Radcliffe Hospital, Oxford University Hospitals, Oxford, UK
[14] Department of Plastic Surgery, Stoke Mandeville Hospital, Buckinghamshire Healthcare NHS trust, UK
[15] Warwick Clinical Trials Unit, Warwick Medical School, University of Warwick, Coventry, UK

J Med Internet Res. 2023 Apr 27;25:e41870

Abstract

Background: Routine use of patient-reported outcome measures (PROMs) and computerized adaptive tests (CATs) may improve care in a range of surgical conditions. However, most available CATs are not condition-specific, nor co-produced with patients, and lack clinically-relevant score interpretation. Recently, a PROM called the CLEFT-Q has been developed for use in the treatment of cleft lip and/or palate (CL/P), but assessment burden may be limiting its uptake into clinical practice.

Objectives: We aimed to develop a CAT for the CLEFT-Q, that could facilitate the uptake of the CLEFT-Q PROM internationally. We aimed to conduct this work with a novel patient-centered approach, and make source code available as an open-source framework for CAT development in other surgical conditions.

Methods: CATs were developed with Rasch measurement theory, using full-length CLEFT-Q responses collected during the CLEFT-Q field test (this included 2434 patients across 12 countries). These algorithms were validated in Monte Carlo simulations involving full-length CLEFT-Q responses collected from 536 patients. In these simulations, the CAT algorithms approximated full-length CLEFT-Q scores iteratively, using progressively fewer items from the full-length PROM. Agreement between full-length CLEFT-Q score and CAT score at different assessment lengths was measured by Pearson correlation coefficient, root mean squared error (RMSE) and 95% limits of agreement. CAT settings, including the number of items to be included in the final assessments, were determined in a multistakeholder workshop which included patients and healthcare professionals. A user interface was developed for the platform, and it was prospectively piloted in the UK and the Netherlands. Interviews were conducted with six patients and four clinicians to explore end-user experience.

Results: The length of all eight CLEFT-Q scales in the International Consortium for Health Outcomes Measurement Standard Set combined was reduced from 76 to 59 items, and at this length CAT assessments reproduced full-length CLEFT-Q scores accurately (with correlations between full-length CLEFT-Q score and CAT score exceeding 0.97, and RMSE ranging from 2-5 out of 100). Workshop stakeholders considered this the optimal balance between accuracy and assessment burden. The platform was perceived to improve clinical communication and facilitate shared decision making.

Conclusion: Our platform is likely to facilitate routine CLEFT-Q uptake, and this may have a positive impact on clinical care. Our free source code enables other researchers to rapidly and economically reproduce this work for other PROMs.

Keywords: cleft lip, cleft palate, patient-reported outcome measures, outcome assessment, CLEFT-Q, computerized adaptive test, CAT.

Introduction

Patient-reported outcome measures (PROMs) have gained widespread acceptance as tools for measuring the impact of treatments on elements of health that matter most to patients. There is also a rapidly growing body of evidence to suggest that adopting PROM feedback into surgical care improves outcomes by enhancing clinical communication and facilitating detection of previously unidentified issues. For many conditions, the use of PROMs is associated with improved health related quality of life (HRQOL), faster detection of clinical deterioration and even improved survival.[1-4] PROMs may be especially helpful in pediatric surgical care, where they may also deliver improved communication, more sensitive detection of HRQOL issues, higher referral rates, better patient experience, and faster consultations.[5-10]

A key group that would benefit from routine use of PROMs are those with cleft lip and/or palate (CL/P) and other craniofacial conditions. CL/P is one of the most common birth differences, affecting one in 700 internationally, with significant implications for a person's facial appearance, dentition, speech, psychosocial development and education.[11] The International Consortium for Health Outcomes Measurement (ICHOM) have recently proposed a Standard Set of outcome measures for the "comprehensive appraisal of cleft care", which largely comprises scales (questionnaires) from the CLEFT-Q, a condition-specific PROM for people aged 8 to 29, born with a CL/P or other craniofacial condition.[12,13]

The eight CLEFT-Q scales included in the ICHOM Standard Set for CL/P measure: the appearance of the face, teeth and jaws; speech function and speech distress; and school, social and psychological function. These scales contain between seven and 12 items (questions), equating to 76 items when all eight scales are administered simultaneously.[12,14]

Barriers to using PROMs such as CLEFT-Q in routine surgical practice include delays in obtaining scores, scores that are difficult to interpret, reference ranges that are difficult to interpret, and difficulties in data governance.[15] In addition, response burden may be an important barrier to implementing PROMs in pediatric settings as it is not always appropriate to administer lengthy assessments to young children in clinical practice. This has limited the uptake of the CLEFT-Q and ICHOM Standard Set for CL/P internationally.[16-19]

Computerized adaptive tests (CATs) are a potential way to overcome these barriers. CATs use algorithms that can make PROMs like CLEFT-Q shorter and more personalized by selecting the most relevant questions for an individual, based on the answers that person has already provided during the assessment. There are three components to a basic CAT algorithm: a score estimator, an item selection criterion, and a stopping rule. The score estimator predicts a person's score from the responses obtained so far during the assessment. The item selection criterion then selects the most useful question to ask, based on the score estimate. This approach avoids asking questions that are unlikely to improve measurement precision. To illustrate, consider an assessment of mobility. If we know that a patient has difficulty walking 100 meters, it would not be helpful to ask whether they have difficulty walking a mile. Instead, a CAT algorithm may select a question more appropriately targeted to that patient, for example whether they have difficulty walking from room to room in their house unaided. The stopping rule terminates the assessment when a prespecified criterion is met, for example after a certain number of questions or given level of measurement precision. This individually tailored approach balances a PROM's reliability with its length, to reduce response burden and is hoped to improve PROM uptake, both in routine clinical practice and research.

There are notable limitations to available CAT platforms in clinical surgery. Firstly, most surgical CATs are generic (as opposed to condition-specific) measures, which may fail to adequately capture the elements of health most important to patients with specific health needs.[20] Secondly, CAT scores are often interpreted through comparison with general population scores. A more useful approach may be to compare a person's score with the scores of people who have similar demographic and clinical characteristics.[21] Thirdly, the number of questions in most CATs is chosen based on psychometric heuristics relating to the assessment's standard error of measurement, an indicator of theoretical measurement reliability.[22] Finally, most CATs send a person's response from the electronic health records (EHR) platform to an external assessment center to select the next question. This is less efficient and secure than a locally implemented system.[23]

The aim of this project was to address these barriers and limitations with a novel system that can deploy person-centered CATs for the CLEFT-Q scales, and feed scores back to clinicians and patients in a rapid, engaging, and clinically useful way. We designed the platform to be open-source and transferrable so that it could be easily, cheaply, and rapidly adapted for any surgical PROM meeting contemporary psychometric standards.

Methods

CAT calibration

We developed CAT algorithms for each CLEFT-Q scale in the ICHOM Standard Set using responses to full-length CLEFT-Q scales that were obtained from the CLEFT-Q field test. This study recruited from October 2014 to November 2016 and collected CLEFT-Q responses from 2434 participants aged eight to 29 years from 30 cleft treatment units in 12 countries. Participants in the CLEFT-Q field test were at a variety of treatment stages for either isolated cleft lip (CL), isolated cleft palate (CP), cleft lip and alveolus (CLA), or cleft lip, alveolus and palate (CLP). Patients with a CL were not asked to complete Speech Function or Speech Distress scales, only children currently in school were asked to complete the School Function scale, and only participants aged 12 years and older were asked to complete the Jaw scale. Each respondent in this cohort completed the CLEFT-Q at one time point. Local Institutional Review Board approval was obtained from each centre. An in-depth report describing the methodology and results of the CLEFT-Q field test has been published previously.[14]

We performed Rasch analysis in R to calibrate CAT parameters from these data (see **Rasch Parameterization, Supplementary Appendix 1**). Rasch analysis is a framework for the development and evaluation of statistical models that describe the relationship between a person's level of measured construct and the probability that they will endorse a certain item response. For example, in the CLEFT-Q social function scale, Rasch models explain how likely a person is to respond to an item in a given way, based on their overall social function level. These models are used by CAT algorithms to estimate a person's overall score, and also to select the most useful item to pose, given the current score estimate. Specific CAT settings for score calculation and item selection were chosen based on previous optimization studies.[24]

CAT validation

We evaluated the performance of these CAT algorithms in an independent validation dataset that included the CLEFT-Q responses of 536 participants, during 561 clinic appointments. These were collected between November 2015 and April 2019 at Erasmus University Medical Center, the Netherlands, as well as Boston Children's Hospital and Duke Children's Hospital, both in the United States of America. Respondents were aged seven to 24 years and receiving care for either CL, CP, CLA or CLP. The timing of scale administration approximately followed the recommendations proposed in the ICHOM Standard Set: clinical teams aimed to assess patients at 8 years of age with the CLEFT-Q face, teeth, and social function scales; then again at approximately 12 and 22 years of age with scales that were pertinent to the patient's specific cleft type. For example, a 22-year-old with an isolated CP would complete the face, jaws, teeth, speech distress, speech function, and social function scales. Incomplete response sets were removed via listwise exclusion and outliers were determined by Mahalanobis distance (see **Missing Data and Outliers in the Validation Dataset, Supplementary Appendix 1**).

We ran a series of Monte Carlo simulations in which CAT algorithms aimed to estimate the full-length CLEFT-Q scale scores of each participant in the validation dataset, based on a predetermined number of their item responses, using an R package which we developed specifically for this study.[25] For example, the CAT for the CLEFT-Q face scale (nine items) first aimed to estimate each respondent's CLEFT-Q face score from all nine items, then from eight items only, then again from seven items. The algorithms used Bayesian statistics to choose which items to administer, and in which order (see **Computerized Adaptive Test Simulation Settings, Supplementary Appendix 1**). For each CAT, at each possible assessment length, concordance between CAT and full-length score was measured with Pearson's correlation coefficient, root mean square error (RMSE) and 95% limits of agreement. RMSE is a measure of the difference between full-length CLEFT-Q scale scores and CLEFT-Q CAT scores, averaged across the population, and 95% limits of agreement demonstrate the difference between full-length CLEFT-Q scale scores and CLEFT-Q CAT scores at the individual level. For example, if the 95% limits of agreement between full-length and CAT scores were -7.00 to +7.00, we would expect that 95% of the time, for any individual, the CAT score would fall within ± 7.00 points of the full-length scale score.

In secondary sensitivity analyses, these computations were repeated with and without outliers, and with both listwise inclusion and imputation of missing item responses (see **Missing Data and Outliers in the Validation Dataset, Supplementary Appendix 1**).

Multi-stakeholder consensus workshop

We discussed the findings of the validation study during a multi-stakeholder consensus workshop attended by three adults who were born with a CL/P, five current patients aged 11-16 years (accompanied by one parent each), two psychologists, two cleft surgeons, two speech and language therapists, one dentist, one orthodontist, and two cleft specialist nurses. Prior to the workshop, participants were asked to read through the full-length CLEFT-Q.

For each scale, the balance between accuracy and burden was discussed in virtual breakout rooms with experienced facilitators ensuring all voices were heard. Particular consideration was given to the scale length, the item wordings, participants' experiences of administering or completing the questionnaire, and the results of the validation study. Every participant voted on the assessment length they felt was most appropriate for each scale. CAT assessment lengths were chosen based on majority voting at this workshop. Ethical approval for this work was obtained from the University of Oxford Medical Sciences Interdivisional Research Ethics Committee (R74005/RE001).

User interface development

We built a user interface to administer each CLEFT-Q scale according to its respective CAT algorithm, using the Concerto platform.[26] Concerto can run CAT algorithms internally, and be installed locally, such that CATs can be administered via Concerto without data leaving a hospital's local server. We integrated the results into a Shiny app which we have called the Score Checker, to help patients and clinicians visualize and interpret CLEFT-Q CAT scores within the clinical context.

Pilot testing

The CLEFT-Q CAT platform was tested in outpatient cleft clinics in Oxford (UK) and Rotterdam (the Netherlands). Patients were asked to complete relevant CLEFT-Q CAT scales in the waiting room, prior to their clinical appointment. Scores were then reviewed by the clinical team before the patient entered the consultation room. Clinicians were then free to discuss and/or action these results as appropriate in the clinical situation.

A purposively diverse sample of UK patients and clinicians that had used the platform within the last seven days were recruited for semi-structured interviews that explored the platform's user experience. It was important to interview both patients and clinicians, as the platform is intended to be acceptable, usable, and of benefit to both of these stakeholder groups. The selection of patients for interviewing was made to be deliberately diverse by age, gender, cleft type, and ethnicity. The selection of clinicians was deliberately diversified by gender and occupation.

Interviews were recorded and transcribed verbatim, then coded with the NVivo platform (version 1.0 for Mac) under the following prespecified categories: experience of the CAT's content; experience of the software; barriers to implementing the CLEFT-Q CAT; and facilitators to implementing the CLEFT-Q CAT. Emergent themes within and outside of these categories were synthesized through an inductive approach. Clinicians involved in piloting the platform at both sites reviewed these themes to check that they accurately and comprehensively captured their experience. A completed Consolidated Criteria for Reporting Qualitative Research (COREQ) checklist[27] is provided in **Supplementary Table 1.** This provides a detailed and standardized report of the qualitative element to this work, including information about the research team, study design, and analysis. Interview schedules are provided in **Supplementary Table 2** and **Supplementary Table 3 (Supplementary Appendix 1**).

Results

Demographics

Clinical and demographic variables for the CAT calibration and validation datasets are presented in **Table 1**. Within both datasets, there was a preponderance towards the male sex and a diagnosis of CLP.

		Psychological	Social	School	Speech distress	Speech Function	Face	Teeth	Jaws
Calibration dataset									
Included participants		2187	2154	1527	1819	1764	2301	2227	1443
Age	Median (IQR)	14 (7)	14 (7)	12 (5)	14 (7)	14 (7)	14 (7)	14 (7)	16 (5)
	Missing data	1	0	0	0	0	1	0	0
Gender	Male	1217	1199	866	1007	973	1277	1231	775
	Female	968	954	661	812	791	1022	995	667
	Missing data	0	0	0	0	0	0	0	0
Country	Australia	23	24	15	20	20	23	25	12
	Canada	476	468	260	380	369	592	526	345
	England	312	304	233	263	253	309	309	205
	Ireland	95	93	57	87	90	96	96	79
	USA	354	351	312	317	316	350	348	178
	The Netherlands	197	195	129	160	153	198	194	138
	India	231	232	176	174	172	232	231	106
	Sweden	93	91	80	77	71	93	92	62
	Turkey	54	52	36	47	50	54	54	49
	Columbia	180	174	105	148	119	183	184	137
	Chile	84	81	57	74	76	84	85	71
	Spain	88	89	67	72	75	87	83	61
	Missing	0	0	0	0	0	0	0	0
Cleft type	CL	244	233	175	0	0	252	248	146
	CP	494	493	374	482	464	526	514	301
	CLA	179	178	139	128	127	195	191	122
	CLAP	1270	1250	839	1209	1173	1328	1274	874
	Missing data	0	0	0	0	0	0	0	0
Validation dataset									
Included participants		247	345	247	258	274	530	529	314
Age	Median (IQR)	12 (1)	9 (5)	12 (1)	12 (5)	12 (5)	11 (3)	11 (3)	12 (10)
	Missing data	0	0	0	0	0	0	0	0
Gender	Male	134	189	134	134	144	292	290	164
	Female	113	156	113	124	130	238	239	150
	Missing data	0	0	0	0	0	0	0	0
Country	The Netherlands	130	226	130	157	174	354	358	214
	USA	117	119	117	101	100	176	171	100
	Missing	0	0	0	0	0	0	0	0
Cleft type	CL	13	27	13	4	4	39	40	22
	CP	71	99	70	86	93	151	151	94
	CLA	24	29	24	7	7	51	50	29
	CLAP	139	190	140	161	170	289	288	169
	Missing data	0	0	0	0	0	0	0	0

Table 1 Clinical and demographic variables of the calibration and validation datasets for each computerized adaptive test. IQR: interquartile range; CL: cleft lip; CP: cleft palate; CLA: cleft lip and alveolus; CLAP: cleft lip, alveolus and palate.

CAT performance

Table 2 summarizes the correlation and agreement between CAT scores and full-length assessments for each scale in the validation dataset. As the number of items in a CAT decreased, so did the correlation and agreement of CAT and full-length scale scores (**Table 2**). A decrease in scale length of two items did not significantly affect accuracy, with correlations all 0.97 or above and RMSE ranging from 2-5 at this level of item reduction.

Scale	CAT length (items)	Correlation with full length	RMSE	Lower LoA	Upper LoA
Face **(9 items total)**	8	0.997	1.67	-3.52	2.92
	7*	0.989*	3.19*	-6.80*	5.46*
	6	0.983	4.01	-8.48	7.00
	5	0.972	5.07	-10.41	9.35
Jaw **(7 items total)**	6*	0.997*	2.24*	-4.09*	4.64*
	5	0.992	3.68	-6.82	7.57
	4	0.985	5.23	-9.70	10.72
	3	0.980	6.28	-11.68	12.86
Teeth **(8 items total)**	7	0.995	2.14	-3.87	4.46
	6*	0.989*	3.17*	-6.33*	6.12*
	5	0.982	4.13	-8.20	7.98
	4	0.968	5.47	-10.74	10.74
School **(10 items total)**	9	0.996	2.00	-4.20	3.54
	8	0.991	2.92	-6.07	5.28
	7*	0.987*	3.53*	-7.28*	6.52*
	6	0.975	4.97	-10.26	9.12
Psychological Function **(10 items total)**	9	0.997	1.98	-3.70	4.03
	8*	0.994*	2.72*	-5.27*	5.41*
	7	0.989	3.75	-7.20	7.53
	6	0.985	4.45	-8.23	9.15
Speech Distress **(10 items total)**	9*	0.995*	2.15*	-4.48*	3.85*
	8	0.973	5.44	-11.82	8.58
	7	0.947	7.61	-16.58	11.79
	6	0.904	10.44	-22.82	15.75
Speech Function **(12 items total)**	11	0.998	1.79	-3.86	2.90
	10	0.992	3.10	-6.59	5.31
	9	0.987	4.14	-8.91	6.83
	8*	0.981*	4.98*	-10.75*	8.12*
Social Function **(10 items total)**	9	0.998	1.40	-2.88	2.57
	8*	0.995*	2.16*	-4.14*	4.32*
	7	0.988	3.45	-7.19	6.22
	6	0.984	4.08	-8.61	7.19

Table 2 CAT performance in validation dataset. Correlation: between linear assessment and CAT scores; RMSE: root mean squared error between linear assessment and CAT scores (out of 100 points); LoA: 95% limit of agreement between linear assessment and CAT scores (out of 100 points), according to the Bland-Altman method. * Indicates CAT settings that were selected by stakeholders to represent the optimal balance between accuracy and assessment burden.

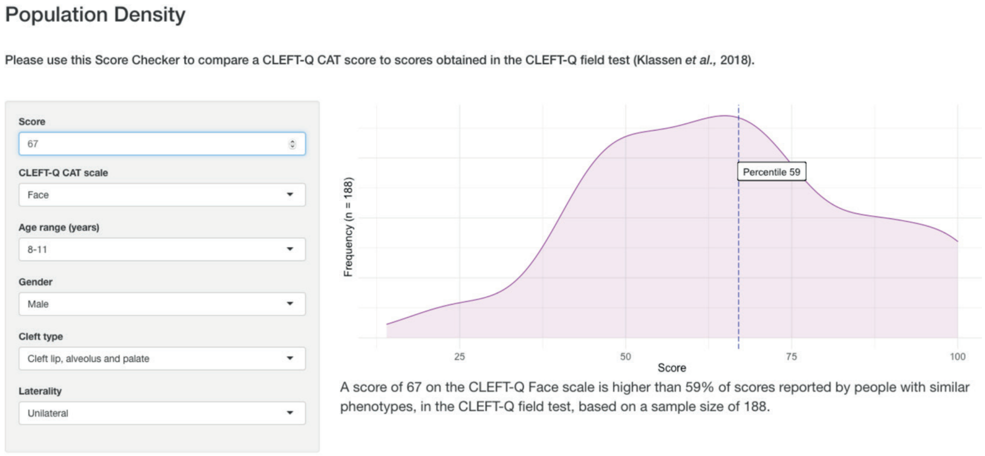
Exclusion of outliers and imputation of missing data did not significantly affect these results. Complete results tables, including those of the sensitivity analyses are available in **sheets 4** and **5 of the Supplementary Appendix 2**.

Multi-stakeholder workshop

The CAT lengths that were chosen to represent the optimal balance between accuracy and burden during the multistakeholder workshop are indicated in **Table 2**, and **Sheet 6 of Supplementary Appendix 2**. The RMSE of these CATs ranged from 2-5 points out of 100 from the full-length assessment scores.

User interface

Figure 1 demonstrates the population density tab of the Score Checker app. Scores are expressed as a percentile of CLEFT-Q field test scores from respondents with similar demographics (age, gender, cleft type and laterality). In the left panel, users can filter the CLEFT-Q field test data based on clinical and demographic variables. The magenta density plot demonstrates the distribution of scores achieved by individuals after filtering, with sample sizes displayed on the y-axis and in text below the plot. The vertical, blue dashed line superimposing the plot demonstrates where a given score would fall in this distribution.



Population Density

Please use this Score Checker to compare a CLEFT-Q CAT score to scores obtained in the CLEFT-Q field test (Klassen *et al.*, 2018).

Score: 67

CLEFT-Q CAT scale: Face

Age range (years): 8-11

Gender: Male

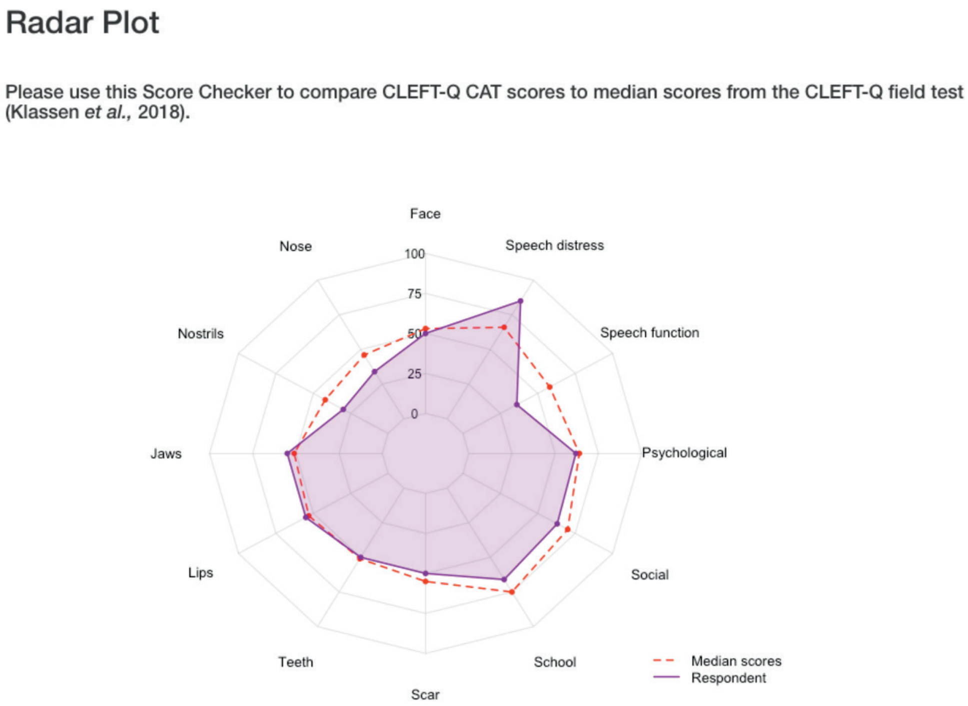
Cleft type: Cleft lip, alveolus and palate

Laterality: Unilateral

A score of 67 on the CLEFT-Q Face scale is higher than 59% of scores reported by people with similar phenotypes, in the CLEFT-Q field test, based on a sample size of 188.

Figure 1 Population density tab of the Score Checker web application.

Figure 2 demonstrates the output of the Radar plot tab of the Score Checker app. Magenta points represent an individual patient's scores, and red points are median field test scores from respondents with similar demographics, based on the filters applied (see left panel of **Figure 1**). Outermore points indicate higher (clinically better) CLEFT-Q scores. Illustrations of the patient-facing interface are provided in the **Supplementary Figure 1** and **Supplementary Figure 2, Supplementary Appendix 1.**



Radar Plot

Please use this Score Checker to compare CLEFT-Q CAT scores to median scores from the CLEFT-Q field test (Klassen *et al.*, 2018).

Figure 2 Radar plot tab of the Score Checker web application.

Semi-structured interviews

We recruited six patients and four clinicians for semi-structured interviews. This included three male and three female patients, aged eight to 28 years with a variety of diagnoses and ethnicities (see **Supplementary Table 4, Supplementary Appendix 1** for participant characteristics), and a cleft surgeon, a cleft specialist nurse, a speech therapist, and a dentist.

Positive themes relating to the content of the CAT included: its ability to cause patients to think about previously unconsidered health aspects, its person-centeredness, and its ability to normalize health concerns. Negative themes relating to the CAT content were repetitiveness and its potential to cause upset to patients who would rather not answer sensitive questions. Themes relating to the user interface were its ease of use and a preference for electronic tablets over pen-and-paper. No participant felt the CAT caused excessive response burden, even when asked directly. Quotes to illustrate these themes are provided in **Thematic Analysis, Supplementary Appendix 1.**

Potential barriers to implementing the system included integration across different EHR platforms, maintaining equality of care between hub and spoke services, a physical means of collecting data (e.g. electronic tablets, staffing and space), opportunity costs for patients and clinicians, reluctance to use technology, and change resistance. Facilitators included the opportunistic use of waiting room time, training, and education of the benefits. The option to complete the assessment at home was seen as a facilitator by some, but not by others.

The use of CAT as a clinical communication aid was an emergent theme within both patient and clinician interviews. Sub-themes, illustrated in **Table 3**, included: improving consultation focus; improving patient-to-clinician information flow; facilitating a multi-disciplinary approach to care; improving patient readiness; and facilitating shared decision making.

Sub-theme	Participants reporting sub-theme	Example quotes
Improving consultation focus	2 clinicians	"We had a patient where they have cleft as part of a complex craniofacial condition, and there had been planning and suggestion of doing really quite major surgery, reconstructing around the nose area, but when the patient did the CAT before they came into the clinic, actually that was one of their lowest priorities, which meant that we could then refocus the consultation to focus on their priorities rather than what had been perceived as what should be discussed" – clinician.
Improving patient-to-clinician information flow	2 patients and 2 clinicians	"I was forearmed and so forewarned so I could broach things differently with her [my patient], have a slightly different dialogue and then facilitate some supportive therapy with clinical psychologists" – clinician. "I found the questionnaire really helped me be more honest about what was bothering me" – patient.
Facilitating a multi-disciplinary approach to care	2 clinicians	"I think what it allows patients to do, is be seen as a whole patient rather than just an element of treatment" – clinician. "All of the patient's focus was on his teeth and jaws and I perhaps wouldn't have been thinking very much about that in my own sort of uni-disciplinary way, and it meant really that, you know, it's [the CLEFT-Q CAT is] actually guiding the treatment pathway for him" – clinician.
Improving patient readiness	2 patients and 3 clinicians	"It made me more alert as to why I was there" – patient.
Facilitating shared decision making	1 patient and 2 clinicians	"What it [the CLEFT-Q CAT] does do is open doors for patients and clinicians to rethink the direction sometimes they were taking [in their care plan]" – clinician. "It helps set the plan of what's more important and what we can do first [which heath interventions should be prioritized]" – patient.

Table 3 Sub-themes relating to the use of the CLEFT-Q CAT as a communication aid.

Discussion

We have developed, validated, deployed, and evaluated a system that can facilitate the uptake of high-quality, standardized outcome measurement for CL/P and other craniofacial conditions, and act as an open-source framework for the development of other surgical CATs. Our approach to CAT development has focused on person-centeredness, and elements of our methodology may be preferable to those used in popular alternatives. Firstly, we have co-produced our software with people who are undergoing, or have undergone, treatment for the condition of interest. We included patients in the setting of

CAT stopping rules, rather than deciding the acceptable level of response burden on their behalf. Secondly, the platform uses condition-specific measures, administered at fixed lengths chosen by stakeholders, and presents scores in comparison to clinically relevant populations (see **Figure 1**). In practice, we anticipate this translating into patients being more likely to complete our CAT than others developed with conventional methods that do not include patients. The platform can also run locally without internet access, meaning that data never have to be shared outside of the clinical environment. This may make our system more efficient and more secure than alternative platforms.

It is possible that these design elements will directly facilitate PROM uptake, as individualization of PROMs, assessment burden, and interpretability of results have all been identified as important "pinch points" for the PROM implementation pipeline.[28] We have made source code for our validation software and Score Checker app freely available for open appraisal and reproduction.[25,29] These can be quickly and cost-effectively translated into other outcome measurement systems.

Our thematic analysis suggests that the platform encourages patient reflection, improves the patient-to-clinician information flow, and facilitates clinical prioritization and shared decision making. These findings are consistent with frameworks derived from other qualitative research into PROM implementation.[10,28,30] While patients found the content of the CAT person-centered, it was also described as repetitive. This may be a generalizable finding of CATs for PROMs, as they aim to select the best-targeted (most salient) items from a scale, which may be similar in content to each other. While response burden is a well described barrier to CLEFT-Q implementation[16-19], none of our interview participants felt that the CAT was excessively burdensome, even on direct questioning: *"for me it was pretty quick, so anyone could fill in this form"* – patient.

There are some limitations to this work. The CLEFT-Q is a novel instrument, and there are no longitudinal, anchor-based estimates for CLEFT-Q scales' minimal important change or minimal important difference. This means our system is limited to interpreting a patient's score through comparison with cross-sectional data from matched populations, using for example, median scores. When a change has occurred in an individual (e.g. following treatment) it is difficult to relate this to real world change that is meaningful to the patient. Similarly, it is difficult to confidently say whether one treatment or one hospital achieves meaningfully better results than another. Ongoing work into CLEFT-Q interpretability, driven partly by ICHOM's promotion of the PROM, will support our platform's use in long-term monitoring and inter-departmental benchmarking.

Future work will look to address these limitations and the other implementation barriers described in our thematic synthesis. The extent to which clinical PROM integration improves patient outcomes in CL/P and other complex, long-term surgical conditions should also be explored in future research. Existing frameworks suggest that they may be most impactful as screening tools, clinical monitoring tools, and decision support systems for shared care planning[28], and this is consistent with our findings.

Conclusion

We have provided an open-source framework for the development of condition-specific, person-centered CAT platforms, and used this to develop and implement a CAT for the CLEFT-Q. This novel approach may be more person-centered and clinically useful than alternatives. The platform was perceived to improve clinical communication and patient experience, and will facilitate the implementation of routine, standardized PROMs in CL/P care. Our methods are generalizable to other long-term, multi-system conditions. We have provided all necessary material for researchers to reproduce these tools for other PROMs.

Acknowledgements

Conrad J. Harrison is funded by a National Institute for Health Research (NIHR) Doctoral Research Fellowship (NIHR300684) for this research project. Jeremy N. Rodrigues is funded by a NIHR postdoctoral fellowship (PDF-2017-10-075). Dominic Furniss is funded by the NIHR Oxford Biomedical Research Centre. The CLEFT-Q field-test study was funded by a grant from the Canadian Institute of Health Research (CIHR, FRN-130515).

This document presents independent research funded by the NIHR and the CIHR. The views expressed are those of the authors and not necessarily those of the CIHR, the NHS, the NIHR or the Department of Health and Social Care.

Conflict of interests

The CLEFT-Q is owned by McMaster University and the Hospital for Sick Children. Anne F Klassen and Karen W.Y. Wong Riff are co-developers of the CLEFT-Q and, as such, could potentially receive a share of any license revenues based on their institutions inventor sharing policy. The other authors have no conflicts of interest to declare in relation to the content of this article.

References

1. Denis F, Lethrosne C, Pourel N, et al. Randomized Trial Comparing a Web-Mediated Follow-up With Routine Surveillance in Lung Cancer Patients. *JNCI: Journal of the National Cancer Institute*. 2017;109(9). doi:10.1093/jnci/djx029.
2. Basch E, Deal AM, Kris MG, et al. Symptom Monitoring With Patient-Reported Outcomes During Routine Cancer Treatment: A Randomized Controlled Trial. *JCO*. 2016;34(6):557-565. doi:10.1200/JCO.2015.63.0830.
3. Strasser F, Blum D, von Moos R, et al. The effect of real-time electronic monitoring of patient-reported symptoms and clinical syndromes in outpatient workflow of medical oncologists: E-MO AIC, a multicenter cluster-randomized phase III study (SAKK 95/06). *Annals of Oncology*. 2016;27(2):324-332. doi:10.1093/annonc/mdv576.
4. Stuck AE, Moser A, Morf U, et al. Effect of Health Risk Assessment and Counselling on Health Behaviour and Survival in Older People: A Pragmatic Randomised Trial. Basu S, ed. *PLoS Med*. 2015;12(10):e1001889. doi:10.1371/journal.pmed.1001889.
5. Engelen V, Detmar S, Koopman H, et al. Reporting health-related quality of life scores to physicians during routine follow-up visits of pediatric oncology patients: Is it effective?: Patient Reported Outcomes in Pediatric Clinic. *Pediatr Blood Cancer*. 2012;58(5):766-774. doi:10.1002/pbc.23158.
6. Engelen V, van Zwieten M, Koopman H, et al. The influence of patient reported outcomes on the discussion of psychosocial issues in children with cancer: Effect of PROs on Communication. *Pediatr Blood Cancer*. 2012;59(1):161-166. doi:10.1002/pbc.24089.
7. Wolfe J, Orellana L, Cook EF, et al. Improving the Care of Children With Advanced Cancer by Using an Electronic Patient-Reported Feedback Intervention: Results From the PediQUEST Randomized Controlled Trial. *JCO*. 2014;32(11):1119-1126. doi:10.1200/JCO.2013.51.5981.
8. Murillo M, Bel J, Pérez J, et al. Impact of monitoring health-related quality of life in clinical practice in children with type 1 diabetes mellitus. *Qual Life Res*. 2017;26(12):3267-3277. doi:10.1007/s11136-017-1682-6.
9. Haverman L, Engelen V, van Rossum MA, Heymans HS, Grootenhuis MA. Monitoring health-related quality of life in paediatric practice: development of an innovative web-based application. *BMC Pediatr*. 2011;11(1):3. doi:10.1186/1471-2431-11-3.
10. Bele S, Chugh A, Mohamed B, Teela L, Haverman L, Santana MJ. Patient-Reported Outcome Measures in Routine Pediatric Clinical Care: A Systematic Review. *Front Pediatr*. 2020;8:364. doi:10.3389/fped.2020.00364.
11. Mossey PA, Little J, Munger RG, Dixon MJ, Shaw WC. Cleft lip and palate. *The Lancet*. 2009;374(9703):1773-1785. doi:10.1016/S0140-6736(09)60695-4.
12. Allori AC, Kelley T, Meara JG, et al. A Standard Set of Outcome Measures for the Comprehensive Appraisal of Cleft Care. *The Cleft Palate-Craniofacial Journal*. 2017;54(5):540-554. doi:10.1597/15-292.
13. Klassen AF, Rae C, Wong Riff KW, et al. FACE-Q Craniofacial Module: Part 1 validation of CLEFT-Q scales for use in children and young adults with facial conditions. *Journal of Plastic, Reconstructive & Aesthetic Surgery*. Published online June 2021:S1748681521002928. doi:10.1016/j.bjps.2021.05.040.
14. Klassen AF, Riff KWW, Longmire NM, et al. Psychometric findings and normative values for the CLEFT-Q based on 2434 children and young adult patients with cleft lip and/or palate from 12 countries. *CMAJ*. 2018;190(15):E455-E462. doi:10.1503/cmaj.170289.
15. Hancock SL, Ryan OF, Marion V, et al. Feedback of patient-reported outcomes to healthcare professionals for comparing health service performance: a scoping review. *BMJ Open*. 2020;10(11):e038190. doi:10.1136/bmjopen-2020-038190.

16. Apon I, Rogers-Vizena CR, Koudstaal MJ, et al. Barriers and Facilitators to the International Implementation of Standardized Outcome Measures in Clinical Cleft Practice. *The Cleft Palate-Craniofacial Journal*. Published online March 5, 2021:105566562199766. doi:10.1177/1055665621997668.
17. Stock NM, Hammond V, Hearst D, et al. Achieving Consensus in the Measurement of Psychological Adjustment to Cleft Lip and/or Palate at Age 8+ Years. *The Cleft Palate-Craniofacial Journal*. 2020;57(6):746-752. doi:10.1177/1055665619898596.
18. Weidler EM, Britto MT, Sitzman TJ. Facilitators and Barriers to Implementing Standardized Outcome Measurement for Children With Cleft Lip and Palate. *The Cleft Palate-Craniofacial Journal*. 2021;58(1):7-18. doi:10.1177/1055665620940187.
19. Harrison CJ, Rodrigues JN, Furniss D, Swan MC. Response to Barriers and Facilitators to the International Implementation of Standardized Outcome Measures in Clinical Cleft Practice. *The Cleft Palate-Craniofacial Journal*. Published online May 11, 2021:105566562110150. doi:10.1177/10556656211015013.
20. Weldring T, Smith SMS. Article Commentary: Patient-Reported Outcomes (PROs) and Patient-Reported Outcome Measures (PROMs). *Health-Serv-Insights*. 2013;6:HSI.S11093. doi:10.4137/HSI.S11093.
21. Northwestern University. PROMIS Reference Populations. Published 2021. Accessed July 13, 2021. https://www.healthmeasures.net/score-and-interpret/interpret-scores/promis/reference-populations.
22. Varni JW, Magnus B, Stucky BD, et al. Psychometric properties of the PROMIS® pediatric scales: precision, stability, and comparison of different scoring and administration options. *Qual Life Res*. 2014;23(4):1233-1243. doi:10.1007/s11136-013-0544-0.
23. HL7 International - FHIR Infrastructure Work Group. PRO Overview. HL7 FHIR Implementation Guide. Published 2019. Accessed July 13, 2021. http://hl7.org/fhir/us/patient-reported-outcomes/2019May/pro-overview.html.
24. Harrison CJ, Rodrigues JN, Furniss D, et al. Optimising the computerised adaptive test to reliably reduce the burden of administering the CLEFT-Q: A Monte Carlo simulation study. *Journal of Plastic, Reconstructive & Aesthetic Surgery*. 2021;74(6):1355-1401. doi:10.1016/j.bjps.2020.12.029.
25. Harrison CJ. cleftqCATsim. Published 2021. Accessed July 13, 2021. https://github.com/MrConradHarrison/cleftqCATsim.
26. Harrison C, Loe BS, Lis P, Sidey-Gibbons C. Maximizing the Potential of Patient-Reported Assessments by Using the Open-Source Concerto Platform With Computerized Adaptive Testing and Machine Learning. *J Med Internet Res*. 2020;22(10):e20950. doi:10.2196/20950.
27. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care*. 2007;19(6):349-357. doi:10.1093/intqhc/mzm042.
28. Greenhalgh J, Dalkin S, Gooding K, et al. Functionality and feedback: a realist synthesis of the collation, interpretation and utilisation of patient-reported outcome measures data to improve patient care. *Health Serv Deliv Res*. 2017;5(2):1-280. doi:10.3310/hsdr05020.
29. Harrison C. CLEFT-Q-CAT-Score-Checker. GitHub. Published 2021. Accessed August 26, 2021. https://github.com/MrConradHarrison/CLEFT-Q-CAT-Score-Checker.
30. Dowrick C, Leydon GM, McBride A, et al. Patients' and doctors' views on depression severity questionnaires incentivised in UK quality and outcomes framework: qualitative study. *BMJ*. 2009;338(mar19 1):b663-b663. doi:10.1136/bmj.b663.

Supplementary Appendix 1

1.0 Supplementary Methods

1.1 Supplementary Methods: Rasch Parameterization

Rasch parameterization was conducted using *R* version 4.0.0 running under *macOS Mojave 10.14.6* with the mirt package (version 1.32.1).[1]

Rasch models were developed from calibration dataset responses following listwise exclusion of participants with incomplete response sets. To generate Rasch models, we used a fixed-quadrature expectation maximization (EM) algorithm.[1] Before generating Rasch models, the two middle response options in the Speech Distress and Speech Function scales were collapsed, and scoring reversed to represent the current version of the CLEFT- Q.[2]

1.2 Supplementary Methods: Missing Data and Outliers in the Validation Dataset

Subjects who declined to answer any CLEFT-Q items were not included in the validation dataset. For the included participants, there were missing responses to items in the Speech Distress scale and Social Function scale. In the Speech Distress scale, missing responses existed for one participant who did not answer nine of the ten items. In the Social scale, there were 141 missing responses to item nine, which was added to the scale during data collection. There were also nine missing responses to item seven in the Social Function scale, and all of these occurred at the same study center.

Missing responses were handled through listwise exclusion of respondents with incomplete response sets. For the Social Function scale, we repeated the analysis including all participants, and imputed the 150 missing responses using one iteration of multiple imputation by chained equations and a proportional odds model.[3] Each analysis was performed with and without outliers, who were identified by Mahalanobis distance.[4] Hereafter we describe these repeat analyses as sensitivity-type analyses.

1.3 Supplementary Methods: Computerized Adaptive Test Simulation Settings

To perform these computerized adaptive test (CAT) simulations, we developed an R package called cleftqCATsim, which contains 15 functions that allow readers to recreate these experiments with their own data. The key CAT simulation functions serve as convenience wrappers for Phil Chalmers' mirtCAT package.[5] We have made cleftqCATsim available through GitHub with an illustrative vignette.[6]

In the CAT simulations, factor scores were calculated for each validation dataset respondent with an expected *a posteriori* approach. Items were selected based on minimum expected posterior variance.

No	Item	Description
Domain 1: Research team and reflexivity		
Personal Characteristics		
1.	Interviewer/facilitator	Interviews were conducted by the manuscript's first author.
2.	Credentials	The interviewer holds a medical degree and is a current doctoral candidate.
3.	Occupation	Honorary plastic surgery registrar and doctoral candidate.
4.	Gender	Male
5.	Experience and training	The interviewer has undertaken a training course on qualitative research with NVivo and has had external supervision from experienced (> 15 years) qualitative researchers.
Relationship with participants		
6.	Relationship established	The interviewer met patient-participants during their routine clinic appointments, and has prior relationships with clinician-participants through clinical work.
7.	Participant knowledge of the interviewer	Participants were provided with a Participant Information Sheet which contained detailed information about the study. Participants knew that this work was conducted as part of the interviewer's doctoral thesis
8.	Interviewer characteristics	The interviewer has led the development of the CLEFT-Q computerized adaptive test as part of his doctoral research.
Domain 2: Study design		
Theoretical framework		
9.	Methodological orientation and theory	Grounded theory.
Participant selection		
10.	Sampling	Patient-participants were purposively selected for diversity in age, gender, ethnicity and diagnosis. Clinician-participants were purposively selected for diversity in occupation.
11.	Method of approach	Participants were approached following routine clinic appointments. All participants had recent (< 7 day) experience of using the CLEFT-Q computerized adaptive test.
12.	Sample size	6 patient-participants and 4 clinician-participants.
13.	Non-participation	Two patient-participants were unable to attend interviews within 7 days due to logistic and time constraints.
Setting		
14.	Setting of data collection	Data were collected either over videoconferencing software in the participant's home, or in the clinical environment following an appointment.
15.	Presence of non-participants	For participants aged < 18 years an adult with parental responsibility was present during the interview.
16.	Description of sample	Sample demographics are presented in Supplementary Table 4.

No	Item	Description
Data collection		
17.	Interview guide	Interview schedules were piloted with one of the interviewer's doctoral supervisors. These are presented in Supplementary Table 2 and Supplementary Table 3
18.	Repeat interviews	No repeat interviews were conducted.
19.	Audio/visual recording	The interviewer made audio recordings which were transcribed verbatim
20.	Field notes	Field notes were made where they were required to understand interview responses (e.g. non-verbal responses).
21.	Duration	Interviews ranged in duration from 6 minutes 49 seconds to 20 minutes 25 seconds.
22.	Data saturation	Participants were not deliberately recruited to reach thematic saturation, although no new themes emerged by the final interview
23.	Transcripts returned	Transcripts were not returned to participants.
Domain 3: Analysis and findings		
Data analysis		
24.	Number of data coders	Data were single coded by the interviewer.
25.	Description of the coding tree	An illustration of the coding tree is provided in Supplementary Figure 3.
26.	Derivation of themes	The following topics were specified *a priori*: experience of the computerized adaptive test's content, experience of the software, barriers to implementing the CLEFT-Q computerized adaptive test, and facilitators to implementing the CLEFT-Q computerized adaptive test. Themes within and in addition to these topics were emergent.
27.	Software	Data were managed in NVivo 1.4.
28.	Participant checking	Participants did not provide feedback on findings.
Reporting		
29.	Quotations presented	Quotations are presented in 2.8 Supplementary Results: Thematic Analysis.
30.	Data and findings consistent	Illustrative data are presented 2.8 Supplementary Results: Thematic Analysis.
31.	Clarity of major themes	Major themes are discussed in the main manuscript and are presented in 2.8 Supplementary Results: Thematic Analysis.
32.	Clarity of minor themes	Minor themes are discussed in the main manuscript and are presented in 2.8 Supplementary Results: Thematic Analysis.

Supplementary Table 1 Consolidated Criteria for Reporting Qualitative Research checklist[7] for the qualitative component of this study.

Focus area	**Opening question and examples of additional probes**
Introduction	Tell me about your last visit to see the cleft team.
Is the CLEFT-Q CAT a worthwhile adjunct to clinical practice?	What did you think of the CLEFT-Q CAT questionnaire?
	Do you think it changed anything about your conversation with the cleft team?
	What did it change?
	Did it change anything else?
	Did you like completing it? Do you think it’s a good idea to ask other people to complete the CLEFT-Q CAT at their appointments, just like you did?
	Why do you think that?
How burdensome is the CLEFT-Q CAT?	How difficult was it to complete the CLEFT-Q CAT?
	What was difficult about it?
	Did it take a long time?
	Was it boring?
	What was boring about it?
	Did it make you tired?
	If you had the choice, would you rather do the CLEFT-Q CAT on an iPad (just like you did) or would you rather have a pen-and-paper version of the questionnaire, with slightly more questions?
	Why?
Facilitators and barriers to CLEFT-Q CAT implementation	Can you think of anything that might make you less likely to use the CLEFT-Q CAT?
	Can you think of anything that might make you more likely to use the CLEFT-Q CAT?
Areas for CLEFT-Q CAT improvement	If you could change anything about the CLEFT-Q CAT, what would you change?
	Why?
	Is there anything you really liked about the CLEFT-Q CAT?

Supplementary Table 2 Interview schedule for patient-participants.

Focus area	Opening question and examples of additional probes
Introduction	Tell me about your role in the cleft team.
	Have you used the CLEFT-Q CAT a lot?
Is the CLEFT-Q CAT a worthwhile adjunct to clinical practice?	What do you think of the CLEFT-Q CAT questionnaire?
	Do you think it has changed any aspect of your clinical care, or that of your colleagues?
	What has it changed?
	Has it changed anything else?
	Do you think patients like completing it?
	How useful is it as an adjunct to clinical care?
	Do you think other cleft teams should be using it?
	Why do you think that?
How burdensome is the CLEFT-Q CAT?	How burdensome is the CLEFT-Q CAT, from your perspective?
	Has it changed your workload, or that of your colleagues?
	In what way?
	Was it boring?
	Does it make clinics faster or slower?
	Have patients given you feedback about the burden of completing it?
Facilitators and barriers to CLEFT-Q CAT implementation	Can you think of any barriers to cleft teams implementing the CLEFT-Q CAT?
	Can you think of anything that made it easier or more difficult to implement?
	What advice would you give other cleft teams that are thinking about using the CLEFT-Q CAT?
Areas for CLEFT-Q CAT improvement	If you could change anything about the CLEFT-Q CAT, what would you change?
	Why?
	Is there anything you really like about the CLEFT-Q CAT?

Supplementary Table 3 Interview schedule for clinician-participants.

2.0 Supplementary Results

2.1 Supplementary Results: Missing Data in the Calibration Dataset

An analysis of missing items was performed for the calibration dataset and is presented in **Sheet 1 of Supplementary Appendix 2**. Missing item responses were largely *missing at random* (explainable by other variables). For example, 84% (837/991) of participants missing one or more Jaw scale item(s) were under the age of 12 years. This is because only CLEFT-Q field test participants aged 12-29 years were asked to complete Jaw scale items.[2] Similarly, 50% (307/614) of participants missing one or more School scale items were not attending school (and therefore not administered these items in the CLEFT-Q field test). In the calibration sample, 43% (263/615) of participants missing Speech Distress items, and 39% (263/670) of those missing Speech Function items were born with a cleft lip only, and therefore unlikely to use these subscales in a real- world setting.

2.2 Supplementary Results: Rasch Parameterization

Rasch model parameters and fit statistics are presented in **Sheet 2 of Supplementary Appendix 2**.

2.3 Supplementary Results: Missing Data and Outliers in the Validation Dataset

The proportions of outliers for each scale are presented in **Sheet 3 of Supplementary Appendix 2**.

2.4 Supplementary Results: Computerized Adaptive Test Simulation Settings

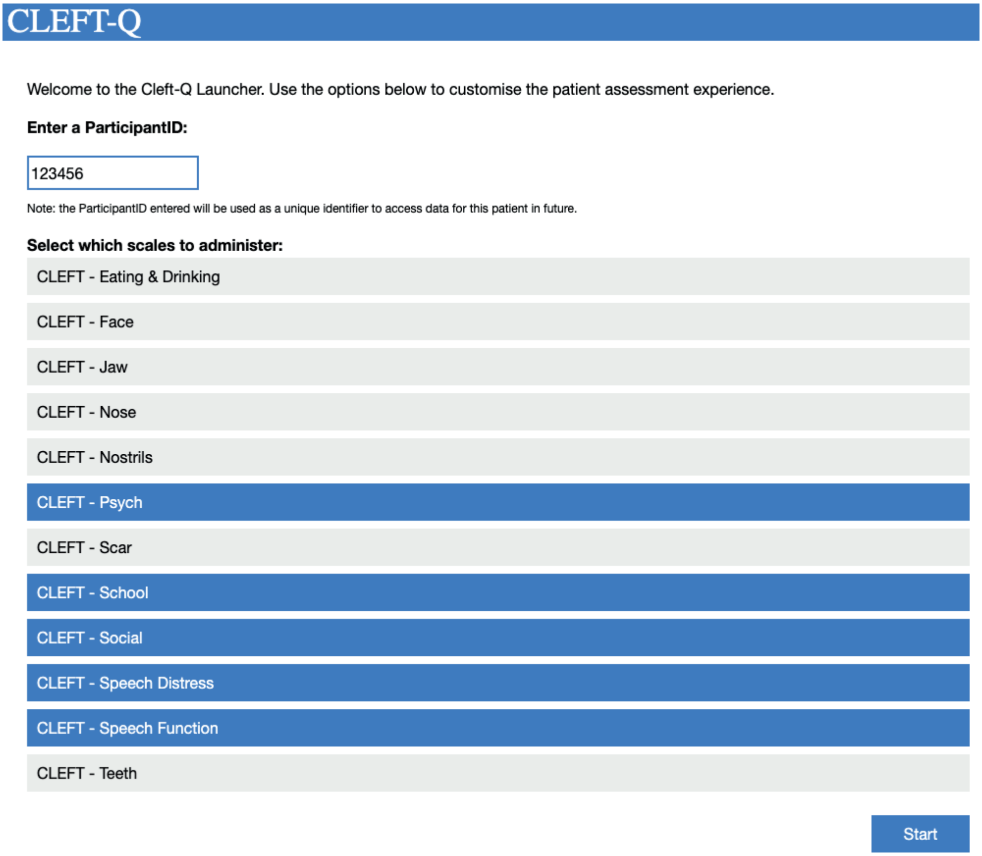
Full results from the computerized adaptive test (CAT) simulations are presented in **Sheet 4 of Supplementary Appendix 2**. In this sheet, root mean squared error (RMSE) and 95% limits of agreement are presented as person-location logits, and median values for standard error of measurement are presented for each assessment, with their inter-quartile ranges. This includes all sensitivity-type analyses. In **sheet 5 of Supplementary Appendix 2**, these results are presented as transformed (0-100) CLEFT-Q scores.

2.5 Supplementary Results: Multistakeholder Workshop

Voting results for stopping rules (CAT assessment lengths) at the multistakeholder workshop are presented in **Sheet 6 of Supplementary Appendix 2**.

2.6 Supplementary Results: Concerto Front-End

The patient-facing front-end of the Concerto-based CLEFT-Q CAT app is illustrated in **Supplementary Figure 1** and **Supplementary Figure 2**. **Supplementary Figure 1** shows the CLEFT-Q CAT launcher, where relevant scales for the patient can be selected, and **Supplementary Figure 2** shows an example item.



CLEFT-Q

Welcome to the Cleft-Q Launcher. Use the options below to customise the patient assessment experience.

Enter a ParticipantID:

123456

Note: the ParticipantID entered will be used as a unique identifier to access data for this patient in future.

Select which scales to administer:

CLEFT - Eating & Drinking

CLEFT - Face

CLEFT - Jaw

CLEFT - Nose

CLEFT - Nostrils

[✓ CLEFT - Psych]

CLEFT - Scar

[✓ CLEFT - School]

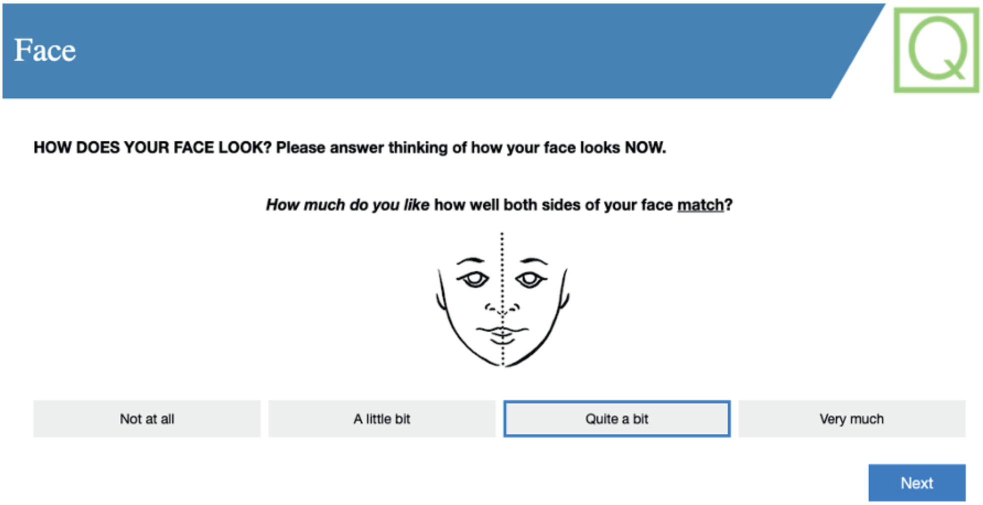
[✓ CLEFT - Social]

[✓ CLEFT - Speech Distress]

[✓ CLEFT - Speech Function]

CLEFT - Teeth

Supplementary Figure 1 The CLEFT-Q computerized adaptive test launcher.



Face

HOW DOES YOUR FACE LOOK? Please answer thinking of how your face looks NOW.

How much do you like how well both sides of your face match?

Supplementary Figure 2 An example item from the CLEFT-Q computerized adaptive test.

2.7 Supplementary Results: Interview Participants

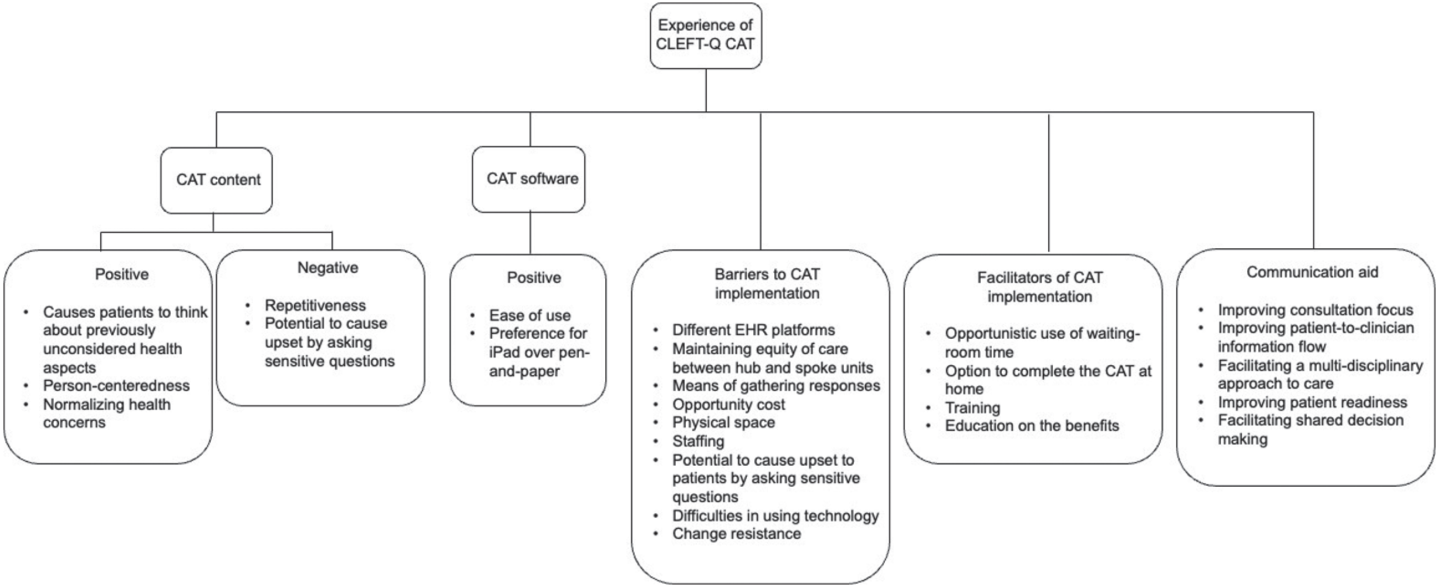
Characteristics of interview participants are displayed in **Supplementary Table 4.**

Patients				
Interview number	**Gender**	**Cleft type**	**Age**	**Ethnicity**
1	M	UCLP	28	White British
2	F	UCL	13	Kurdish
3	M	UCLP	16	Asian (other)
4	M	BCLP	8	Indian
5	F	BCLP	24	White British
6	F	BCLP	18	British Pakistani
Clinicians				
Interview number		**Occupation**		
7	M	Surgeon		
8	F	Specialist nurse		
9	F	Speech and language therapist		
10	M	Dentist		

Supplementary Table 4 Interview participant characteristics. UCLP: unilateral cleft lip and palate; UCL: unilateral, isolated, cleft lip; BCLP: bilateral cleft lip and palate.

2.8 Supplementary Results: Thematic Analysis

The interview transcript coding tree is displayed in **Supplementary Figure 3**. Below, we present quotes to illustrate themes not included in the main manuscript.


Experience of CLEFT-Q CAT
CAT content
CAT software
Positive
• Causes patients to think about previously unconsidered health aspects
• Person-centeredness
• Normalizing health concerns
Negative
• Repetitiveness
• Potential to cause upset by asking sensitive questions
Positive
• Ease of use
• Preference for iPad over pen-and-paper
Barriers to CAT implementation
• Different EHR platforms
• Maintaining equity of care between hub and spoke units
• Means of gathering responses
• Opportunity cost
• Physical space
• Staffing
• Potential to cause upset to patients by asking sensitive questions
• Difficulties in using technology
• Change resistance
Facilitators of CAT implementation
• Opportunistic use of waiting-room time
• Option to complete the CAT at home
• Training
• Education on the benefits
Communication aid
• Improving consultation focus
• Improving patient-to-clinician information flow
• Facilitating a multi-disciplinary approach to care
• Improving patient readiness
• Facilitating shared decision making

Supplementary Figure 3 The coding tree from interview transcripts.

Causing patients to think about previously unconsidered health aspects:
"I think it's really detailed, it makes you consider aspects that you don't really consider thinking about too often, so that can help the doctors as well as you" – patient.

Person-centeredness:
"The questions are I think on topic and are what are relevant to me personally" – patient.

"I found that that questionnaire, compared to the paper one I've done in the past, they were similar, but it was more about how I felt, rather than other people" – patient.

Normalizing health concerns:
"It made it feel way better, like I'm more like other people" – patient.

Repetitiveness:
"Some of them could have been a bit repetitive. It was almost like the same question but not" – patient.

Potential to cause upset to patients by asking sensitive questions:
"It's good, because I guess you get to express yourself and say what you feel, I think. But also, I feel like it makes you more conscious about how you look... I didn't find it upsetting personally, but I feel that for other people it probably might be" – patient.

Ease of use:
"It was really straightforward" – patient.

"For me it was pretty quick, so anyone could fill in this form" – patient.

"I think the visual appearance [of scores presented as a radar chart] is helpful. Rather than just having scores... you can absorb more information quite quickly" – clinician

A preference for using electronic tablets over pen-and-paper:
"The iPad is much more convenient [than pen and paper]... you get the results much quicker" – patient.

Integration across different electronic health record platforms:
"They [other clinicians] will all tend to have different electronic record needs, so it's trying to find something generic, or a platform that can be used across systems, which may be a challenge." – clinician.

Maintaining equality between hub and spoke services:
"You need to have that equality of care, because the reason for a lot of outlying clinics in cleft is for more deprived patient groups. So, in a way it's even more important that you can roll it out to those groups than it is for the hub patients" – clinician.

Means of gathering responses:
"Aside from the availability of iPads and then you know the logistics of them missing and signing them in and signing them out, I don't really see any barriers going forwards" – clinician.

Opportunity cost:
"If you're an adult, you're probably thinking about how you've taken a couple of hours out of your day and you've got your phone on you so you can do some work" – patient.

Interviewer: *"Anything that would make you want to do it less?"*
Patient: *"Yes"* Interviewer: *"Like what?"* Patient: *"Playing games"*
Interviewer: *"So if you could play games instead, you would rather do that? Is that what you're saying?"*
Patient: *Nods and grins cheekily.*

Physical space:
"You need a space for them [the patients] to use [to implement the CLEFT-Q CAT]" – clinician.

Staffing:
"You need a person who can explain to the patient or the family, what to do [to implement the CLEFT-Q CAT]"
– clinician

Difficulties in using technology:
"I don't know [what might make the CLEFT-Q CAT difficult to implement], it is straightforward to be fair. But it might put them [other patients] off when they see the iPad and they think "oh no", or unless you actually do it, because it was really straightforward, but it could put them off if they see the iPad. Maybe that's older people because a lot of young people know what an iPad is. I wouldn't be put off by doing it" – patient.

Change resistance:
"When we had similar projects, the barriers tend to be partly personality driven, so some centers will not adopt anything new, or certainly won't adopt anything they didn't develop, almost out of principle. So, there will be some late adopters, where hopefully they'll come on board later" – clinician.

Opportunistic use of waiting room time:
"If you're a child you're bored and you've got the option of a tablet in front of you, you'd probably do it, personally. It's either that or watching some terrible TV you get in a waiting room… if it's like the 3-4 years I had of brace treatment, I was stuck in the waiting room for 40 minutes, so 10 minutes doing that is probably a welcome distraction" – patient.

Option to complete the CAT at home:
This was seen as a facilitator by some, but not by others:

"So, personally, I think we had that [the CLEFT-Q CAT] sent out with their appointment, I don't know, a few weeks ahead, that would be really useful because, then, to have that information going in, and we could have had it, if there was anything really significant that was coming up when we reviewed them, then we would be able to discuss that prior to their appointment, so I think, yeah, kind of getting ahead with time would be really good" – clinician

"I'm of the opinion that it should be done in the clinical setting with the clinician present at the time, or just before the appointment. Ideally patients would aim to arrive at their appointment a few minutes early anyway, and I think it's a good time to reflect" – clinician.

Training:
"It would probably be good to have somebody who understands it [the CLEFT-Q CAT] well to come in and have a short training session [with new users]" – clinician.

Education surrounding benefits:
"*A lot of the time you're asked to fill out a questionnaire, and they're [the questionnaire administrators] like 'win a £10 Amazon voucher' – you're not bothered about that. But actually, if they say this will actually really benefit you from a health perspective, and our consultants [attendings] will understand you as a person more, I think they'll go 'actually, yeah, let's not make this a dreaded experience, let's make this an experience we can benefit from'"* – patient.

References

1. Chalmers RP. mirt: A Multidimensional Item Response Theory Package for the R Environment. J Stat Softw. 2012;48(6). doi:10.18637/jss.v048.i06.
2. Klassen AF, Riff KWW, Longmire NM, et al. Psychometric findings and normative values for the CLEFT-Q based on 2434 children and young adult patients with cleft lip and/or palate from 12 countries. Can Med Assoc J. 2018;190(15):E455-E462. doi:10.1503/cmaj.170289.
3. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work?: Multiple imputation by chained equations. Int J Methods Psychiatr Res. 2011;20(1):40-49. doi:10.1002/mpr.329.
4. Filzmoser P, Ruiz-Gazen A, Thomas-Agnan C. Identification of local multivariate outliers. Stat Pap. 2014;55(1):29-47. doi:10.1007/s00362-013-0524-z.
5. Chalmers RP. Generating Adaptive and Non-Adaptive Test Interfaces for Multidimensional Item Response Theory Applications. J Stat Softw. 2016;71(5). doi:10.18637/jss.v071.i05.
6. Harrison CJ. cleftqCATsim. Published 2021. Accessed July 13, 2021. https://github.com/MrConradHarrison/cleftqCATsim.
7. Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. Int J Qual Health Care. 2007;19(6):349-357. doi:10.1093/intqhc/mzm042.

Supplementary Appendix 2

Sheet 1 Missing item responses for each scale in the calibration dataset.

Face scale	
Missing responses	445
Complete responses	21461
Number of missing items	Number of respondents
0	2301
1	78
2	12
3	3
4	2
5	2
6	2
7	0
8	2
9	32
Missing item	Number of missing responses
1	43
2	43
3	59
4	46
5	53
6	45
7	48
8	59
9	49

Jaw scale	
Missing responses	6749
Complete responses	10289
Number of missing items	Number of respondents
0	1443
1	29
2	1
3	2
4	0
5	0
6	1
7	958
Missing item	Number of missing responses
1	964
2	969
3	959
4	967
5	963
6	960
7	967

Teeth scale	
Missing responses	1118
Complete responses	18354
Number of missing items	Number of respondents
0	2227
1	67
2	7
3	2
4	1
5	1
6	3
7	4
8	122
Missing item	Number of missing responses
1	133
2	142
3	137
4	149
5	140
6	137
7	139
8	141

School scale	
Missing responses	7879
Complete responses	16461
Number of missing items	Number of respondents
0	1527
1	124
2	6
3	2
4	1
5	1
6	0
7	0
8	0
9	2
10	771
Missing item	Number of missing responses
1	862
2	775
3	778
4	782
5	779
6	780
7	776
8	780
9	787
10	780

Psychological function scale	
Missing responses	1912
Complete responses	22428
Number of missing items	Number of respondents
0	2187
1	52
2	5
3	3
4	3
5	1
6	0
7	0
8	2
9	2
10	179
Missing item	Number of missing responses
1	182
2	191
3	188
4	191
5	193
6	188
7	196
8	197
9	192
10	194

Speech Distress scale	
Missing responses	5553
Complete responses	18787
Number of missing items	Number of respondents
0	1819
1	56
2	7
3	4
4	0
5	0
6	1
7	1
8	0
9	2
10	544
Missing item	Number of missing responses
1	554
2	555
3	556
4	554
5	554
6	555
7	554
8	554
9	559
10	558

Speech Function scale	
Missing responses	6969
Complete responses	22239
Number of missing items	Number of respondents
0	1764
1	77
2	16
3	3
4	1
5	0
6	1
7	2
8	1
9	2
10	1
11	1
12	565
Missing item	Number of missing responses
1	579
2	580
3	576
4	582
5	579
6	589
7	584
8	581
9	583
10	581
11	580
12	575

Social Function scale	
Missing responses	2013
Complete responses	22327
Number of missing items	Number of respondents
0	2154
1	62
2	17
3	6
4	0
5	1
6	1
7	0
8	20
9	2
10	171
Missing item	Number of missing responses
1	201
2	200
3	206
4	203
5	200
6	208
7	198
8	205
9	193
10	199

Sheet 2 Rasch model fit statistics and item parameters.

Fit Statistics

Scale	**X2 p-value**	**RMSEA**	**SRMSR**	**TLI**	**CFI**
Face	< 0.001	0,067	0,062	0,945	0,948
Jaw	< 0.001	0,095	0,114	0,879	0,896
Teeth	< 0.001	0,105	0,054	0,867	0,878
School	< 0.001	0,069	0,089	0,923	0,926
Psychological Function	< 0.001	0,080	0,051	0,911	0,914
Speech Distress	< 0.001	0,110	0,072	0,930	0,932
Speech Function	< 0.001	0,083	0,075	0,970	0,970
Social Function	< 0.001	0,080	0,079	0,909	0,912

X2: Chi squared; RMSEA: root mean squared error; SRMSR: standardized root mean squared residual; TLI: Tucker-Lewis Index; CFI: comparative fit index.

Model Parameters

Face				
	a	b1	b2	b3
Item 1	1	-4,89	-2,197	0,073
Item 2	1	-4,444	-2,205	0,082
Item 3	1	-3,843	-1,721	0,478
Item 4	1	-3,164	-1,044	0,916
Item 5	1	-3,008	-1,013	0,707
Item 6	1	-2,941	-1,043	0,726
Item 7	1	-2,797	-0,978	0,763
Item 8	1	-2,35	-0,792	0,965
Item 9	1	-2,69	-0,735	1,384

Face scale item parameters.

Jaw				
	a	b1	b2	b3
Item 1	1	-4,831	-2,85	1,007
Item 2	1	-4,659	-2,741	0,898
Item 3	1	-4,614	-2,771	0,986
Item 4	1	-4,599	-2,678	0,832
Item 5	1	-4,502	-2,621	0,773
Item 6	1	-4,439	-2,591	0,965
Item 7	1	-3,9	-2,232	1,119

Jaw scale item parameters.

Teeth				
	a	b1	b2	b3
Item 1	1	-3,463	-1,462	0,787
Item 2	1	-2,572	-0,944	1,076
Item 3	1	-2,056	-0,633	0,963
Item 4	1	-2,109	-0,553	1,266
Item 5	1	-1,864	-0,302	1,254
Item 6	1	-1,343	-0,288	1,131
Item 7	1	-1,871	-0,08	1,663
Item 8	1	-1,461	-0,159	1,498

Teeth scale item parameters.

School				
	a	b1	b2	b3
Item 1	1	-4,908	-3,361	-2,216
Item 2	1	-4,718	-3,441	-1,775
Item 3	1	-4,172	-2,539	-1,241
Item 4	1	-5,041	-2,542	-0,502
Item 5	1	-4,487	-2,335	-0,922
Item 6	1	-4,552	-2,362	-0,606
Item 7	1	-4,652	-2,177	-0,259
Item 8	1	-3,786	-2,129	-1,241
Item 9	1	-3,561	-1,307	-0,35
Item 10	1	-3,193	-1,653	-0,414

Scool function scale item parameters.

Psychological function				
	a	b1	b2	b3
Item 1	1	-6,154	-3,053	-0,402
Item 2	1	-5,759	-3,184	-0,684
Item 3	1	-5,921	-2,882	0,081
Item 4	1	-5,853	-2,37	-0,181
Item 5	1	-5,408	-2,511	-0,528
Item 6	1	-5,472	-2,432	-0,293
Item 7	1	-4,749	-2,413	-0,383
Item 8	1	-5,277	-1,948	0,052
Item 9	1	-5,303	-2,12	0,442
Item 10	1	-4,241	-1,498	0,399

Psychological function scale item parameters.

Speech distress			
	a	b1	b2
Item 1	1	-4,292	-2,335
Item 2	1	-4,402	-1,65
Item 3	1	-3,983	-1,617
Item 4	1	-3,877	-0,751
Item 5	1	-3,766	-0,77
Item 6	1	-3,224	0,079
Item 7	1	-3,213	0,271
Item 8	1	-3,089	0,256
Item 9	1	-2,56	0,876
Item 10	1	-2,585	1,13

Speech distress scale item parameters.

Speech function			
	a	b1	b2
Item 1	1	-5,516	-1,572
Item 2	1	-5,535	-0,993
Item 3	1	-3,87	-0,627
Item 4	1	-4,526	0,132
Item 5	1	-3,94	-0,397
Item 6	1	-4,22	-0,314
Item 7	1	-4,309	-0,135
Item 8	1	-3,448	-0,189
Item 9	1	-4,324	0,77
Item 10	1	-3,659	0,098
Item 11	1	-3,166	0,078
Item 12	1	-3,987	1,089

Speech function scale item parameters.

Social function				
	a	b1	b2	b3
Item 1	1	-5,665	-2,863	-1,852
Item 2	1	-5,062	-2,936	-1,516
Item 3	1	-5,007	-2,209	-0,314
Item 4	1	-4,371	-2,008	-0,76
Item 5	1	-4,331	-1,855	-0,777
Item 6	1	-3,878	-1,488	-0,116
Item 7	1	-3,518	-1,512	-0,074
Item 8	1	-3,439	-1,232	-0,162
Item 9	1	-2,868	-1,375	-0,045
Item 10	1	-3,024	-1,119	-0,14

Social function scale item parameters.

Sheet 3 Outliers in validation dataset.

Number of outliers per scale, by Mahalanobis distance. X2: Chi squared value; DF: degrees of freedom.

Scale	Sample size	Outliers	X2	DF
Face	551	22	27,88	9
Jaw	317	15	24,32	7
Teeth	551	13	26,12	8
School	254	8	29,59	10
Psychological Function	255	9	29,59	10
Speech Distress	261	10	29,59	10
Speech Function	278	1	32,91	12
Social Function	219	6	29,59	10

Sheet 4 CAT simulation results (logits).

In this sheet, root mean squared error and limits of agreement are presented as person-location logits. Simulations results for the social function scale are presented following either listwise exclusion or imputation. SEM: Standard error of measurement; RMSE: root mean squared error; LoA: 95% limit of agreement.

		Outliers included					
Scale	**Items**	**Median SEM**	**SEM IQR**	**Correlation**	**RMSE**	**Lower LoA**	**Upper LoA**
Face (9 items total)	9	0,501	0,148	1,000	0,000	0,000	0,000
	8	0,516	0,133	0,997	0,135	-0,283	0,239
	7	0,546	0,176	0,989	0,268	-0,571	0,459
	6	0,586	0,170	0,983	0,335	-0,701	0,599
	5	0,634	0,141	0,972	0,422	-0,869	0,775
	4	0,695	0,114	0,963	0,484	-0,982	0,914
	3	0,778	0,178	0,950	0,565	-1,131	1,082
	2	0,922	0,088	0,912	0,736	-1,457	1,432
	1	1,164	0,161	0,812	1,058	-1,995	2,144
Jaw (7 items total)	7	0,697	0,430	1,000	0,000	0,000	0,000
	6	0,743	0,420	0,997	0,212	-0,398	0,430
	5	0,801	0,404	0,992	0,347	-0,652	0,707
	4	0,881	0,382	0,985	0,486	-0,900	0,997
	3	0,998	0,347	0,980	0,593	-1,126	1,198
	2	1,180	0,290	0,966	0,786	-1,441	1,623
	1	1,555	0,156	0,915	1,214	-2,210	2,518
Teeth (8 items total)	8	0,483	0,104	1,000	0,000	0,000	0,000
	7	0,512	0,093	0,995	0,168	-0,314	0,343
	6	0,554	0,119	0,990	0,254	-0,514	0,481
	5	0,591	0,100	0,982	0,333	-0,668	0,639
	4	0,655	0,133	0,968	0,442	-0,875	0,858
	3	0,747	0,105	0,942	0,591	-1,207	1,105
	2	0,879	0,134	0,894	0,801	-1,694	1,392
	1	1,149	0,158	0,821	1,019	-2,145	1,792

	Outliers included						
Scale	**Items**	**Median SEM**	**SEM IQR**	**Correlation**	**RMSE**	**Lower LoA**	**Upper LoA**
School (10 items total)	10	0,495	0,416	1,000	0,000	0,000	0,000
	9	0,518	0,400	0,996	0,150	-0,310	0,277
	8	0,537	0,390	0,992	0,226	-0,476	0,398
	7	0,552	0,368	0,988	0,273	-0,562	0,504
	6	0,605	0,351	0,975	0,386	-0,795	0,711
	5	0,633	0,321	0,959	0,490	-1,011	0,900
	4	0,705	0,459	0,938	0,608	-1,280	1,073
	3	0,816	0,461	0,911	0,718	-1,489	1,308
	2	0,999	0,371	0,858	0,904	-1,908	1,588
	1	1,147	0,207	0,768	1,130	-2,412	1,923
Psychological function (10 items total)	10	0,541	0,408	1,000	0,000	0,000	0,000
	9	0,564	0,393	0,997	0,172	-0,329	0,345
	8	0,587	0,377	0,994	0,239	-0,465	0,476
	7	0,635	0,362	0,989	0,333	-0,646	0,663
	6	0,667	0,460	0,985	0,393	-0,746	0,794
	5	0,737	0,558	0,977	0,480	-0,897	0,982
	4	0,782	0,516	0,969	0,566	-1,049	1,163
	3	0,901	0,461	0,948	0,726	-1,399	1,450
	2	1,121	0,382	0,909	0,948	-1,898	1,822
	1	1,311	0,217	0,807	1,334	-2,634	2,607
Speech distress (10 items total)	10	0,710	0,316	1,000	0,000	0,000	0,000
	9	0,716	0,305	0,995	0,179	-0,372	0,326
	8	0,727	0,477	0,973	0,452	-0,980	0,720
	7	0,809	0,448	0,947	0,643	-1,407	0,968
	6	0,830	0,414	0,904	0,876	-1,917	1,309
	5	0,858	0,362	0,883	0,957	-2,092	1,448
	4	0,914	0,325	0,864	0,994	-2,142	1,627
	3	1,078	0,262	0,811	1,146	-2,442	1,956
	2	1,195	0,190	0,747	1,294	-2,720	2,294
	1	1,531	0,095	0,622	1,499	-2,818	3,053
Speech function (12 items total)	12	0,582	0,194	1,000	0,000	0,000	0,000
	11	0,607	0,181	0,998	0,155	-0,326	0,276
	10	0,624	0,287	0,992	0,280	-0,595	0,482
	9	0,669	0,314	0,987	0,374	-0,804	0,618
	8	0,697	0,298	0,981	0,449	-0,962	0,748
	7	0,760	0,277	0,974	0,525	-1,137	0,844
	6	0,800	0,252	0,963	0,620	-1,342	0,999
	5	0,900	0,223	0,951	0,722	-1,572	1,121
	4	0,969	0,393	0,932	0,845	-1,840	1,308
	3	1,140	0,340	0,910	0,965	-2,091	1,540
	2	1,291	0,291	0,835	1,278	-2,770	2,037
	1	1,689	0,188	0,746	1,487	-3,073	2,726

				Outliers included			
Scale	**Items**	**Median SEM**	**SEM IQR**	**Correlation**	**RMSE**	**Lower LoA**	**Upper LoA**
Social function (10 items total)	10	0,482	0,247	1,000	0,000	0,000	0,000
	9	0,500	0,232	0,998	0,105	-0,213	0,199
	8	0,521	0,218	0,995	0,165	-0,324	0,326
	7	0,534	0,199	0,988	0,270	-0,566	0,484
	6	0,580	0,294	0,984	0,317	-0,670	0,555
	5	0,601	0,458	0,961	0,486	-1,045	0,805
	4	0,672	0,425	0,952	0,530	-1,119	0,932
	3	0,756	0,385	0,935	0,610	-1,236	1,156
	2	0,866	0,300	0,887	0,795	-1,629	1,483
	1	1,100	0,209	0,772	1,088	-2,115	2,158
Social function with imputed data (10 items total)	10	0,509	0,246	1,000	0,000	0,000	0,000
	9	0,514	0,232	0,998	0,103	-0,210	0,194
	8	0,521	0,216	0,994	0,181	-0,377	0,325
	7	0,567	0,319	0,988	0,263	-0,560	0,450
	6	0,581	0,294	0,983	0,316	-0,670	0,548
	5	0,645	0,262	0,962	0,468	-1,007	0,771
	4	0,686	0,422	0,952	0,520	-1,108	0,885
	3	0,776	0,379	0,936	0,589	-1,186	1,123
	2	0,932	0,300	0,887	0,777	-1,589	1,448
	1	1,100	0,168	0,766	1,069	-2,064	2,130

				Outliers excluded			
Scale	**Items**	**Median SEM**	**SEM IQR**	**Correlation**	**RMSE**	**Lower LoA**	**Upper LoA**
Face (9 items total)	9	0,501	0,143	1,000	0,000	0,000	0,000
	8	0,525	0,133	0,997	0,134	-0,281	0,237
	7	0,546	0,186	0,989	0,267	-0,572	0,450
	6	0,586	0,167	0,984	0,326	-0,684	0,577
	5	0,641	0,139	0,974	0,407	-0,834	0,757
	4	0,695	0,223	0,965	0,470	-0,950	0,892
	3	0,778	0,175	0,954	0,551	-1,105	1,053
	2	0,922	0,256	0,919	0,719	-1,425	1,395
	1	1,164	0,161	0,828	1,033	-1,922	2,115
Jaw (7 items total)	7	0,697	0,430	1,000	0,000	0,000	0,000
	6	0,743	0,420	0,998	0,195	-0,376	0,388
	5	0,801	0,404	0,995	0,291	-0,543	0,596
	4	0,881	0,382	0,989	0,439	-0,814	0,902
	3	0,998	0,347	0,983	0,571	-1,064	1,168
	2	1,180	0,290	0,971	0,761	-1,374	1,585
	1	1,555	0,156	0,930	1,169	-2,099	2,442

		Outliers excluded					
Scale	**Items**	**Median SEM**	**SEM IQR**	**Correlation**	**RMSE**	**Lower LoA**	**Upper LoA**
Teeth (8 items total)	8	0,495	0,104	1,000	0,000	0,000	0,000
	7	0,512	0,093	0,996	0,165	-0,311	0,335
	6	0,554	0,119	0,990	0,249	-0,503	0,470
	5	0,596	0,100	0,983	0,330	-0,662	0,632
	4	0,655	0,133	0,971	0,428	-0,851	0,830
	3	0,747	0,105	0,950	0,562	-1,152	1,041
	2	0,879	0,134	0,908	0,757	-1,605	1,306
	1	1,149	0,158	0,829	1,005	-2,119	1,761
School (10 items total)	10	0,524	0,416	1,000	0,000	0,000	0,000
	9	0,530	0,399	0,996	0,148	-0,304	0,275
	8	0,537	0,385	0,992	0,224	-0,470	0,399
	7	0,552	0,368	0,988	0,260	-0,528	0,489
	6	0,605	0,347	0,978	0,362	-0,739	0,678
	5	0,633	0,321	0,964	0,459	-0,938	0,858
	4	0,705	0,516	0,945	0,566	-1,178	1,023
	3	0,816	0,461	0,919	0,680	-1,397	1,259
	2	0,999	0,368	0,864	0,876	-1,837	1,559
	1	1,147	0,207	0,780	1,088	-2,299	1,905
Psychological function (10 items total)	10	0,559	0,408	1,000	0,000	0,000	0,000
	9	0,572	0,393	0,997	0,170	-0,327	0,338
	8	0,587	0,377	0,994	0,235	-0,451	0,473
	7	0,640	0,545	0,989	0,323	-0,621	0,647
	6	0,707	0,584	0,986	0,378	-0,713	0,769
	5	0,737	0,558	0,979	0,461	-0,854	0,946
	4	0,821	0,516	0,973	0,537	-0,975	1,116
	3	0,907	0,461	0,952	0,694	-1,304	1,415
	2	1,121	0,378	0,913	0,920	-1,825	1,789
	1	1,311	0,217	0,811	1,316	-2,574	2,595
Speech distress (10 items total)	10	0,571	0,063	1,000	0,000	0,000	0,000
	9	0,602	0,082	0,991	0,202	-0,411	0,378
	8	0,625	0,081	0,979	0,325	-0,596	0,672
	7	0,668	0,083	0,960	0,476	-0,778	1,026
	6	0,701	0,096	0,943	0,626	-0,881	1,377
	5	0,753	0,105	0,922	0,821	-0,926	1,802
	4	0,818	0,117	0,887	1,109	-0,975	2,387
	3	0,920	0,111	0,818	1,526	-1,066	3,205
	2	1,107	0,074	0,721	2,193	-0,845	4,345
	1	1,425	0,012	0,610	2,871	0,071	5,076

		Outliers excluded					
Scale	**Items**	**Median SEM**	**SEM IQR**	**Correlation**	**RMSE**	**Lower LoA**	**Upper LoA**
Speech function (12 items total)	12	0,571	0,081	1,000	0,000	0,000	0,000
	11	0,586	0,080	0,997	0,152	-0,255	0,327
	10	0,611	0,148	0,991	0,272	-0,467	0,578
	9	0,636	0,110	0,986	0,344	-0,527	0,751
	8	0,666	0,117	0,978	0,458	-0,620	1,008
	7	0,701	0,103	0,967	0,599	-0,670	1,314
	6	0,746	0,055	0,953	0,751	-0,684	1,620
	5	0,803	0,148	0,941	0,882	-0,698	1,878
	4	0,879	0,121	0,922	1,019	-0,739	2,150
	3	0,986	0,101	0,898	1,244	-0,733	2,568
	2	1,159	0,089	0,842	1,651	-0,729	3,311
	1	1,411	0,090	0,770	2,217	-0,396	4,171
Social function (10 items total)	10	0,486	0,247	1,000	0,000	0,000	0,000
	9	0,514	0,232	0,998	0,104	-0,209	0,199
	8	0,521	0,216	0,995	0,163	-0,318	0,324
	7	0,537	0,319	0,988	0,266	-0,555	0,482
	6	0,580	0,294	0,984	0,310	-0,653	0,549
	5	0,601	0,455	0,961	0,487	-1,046	0,808
	4	0,672	0,423	0,952	0,533	-1,121	0,943
	3	0,760	0,388	0,936	0,607	-1,225	1,154
	2	0,866	0,300	0,892	0,782	-1,597	1,465
	1	1,100	0,209	0,795	1,045	-1,999	2,101
Social function with imputed data (10 items total)	10	0,509	0,242	1,000	0,000	0,000	0,000
	9	0,514	0,229	0,998	0,103	-0,208	0,194
	8	0,547	0,336	0,994	0,182	-0,380	0,327
	7	0,567	0,319	0,988	0,260	-0,552	0,451
	6	0,583	0,294	0,984	0,308	-0,646	0,545
	5	0,645	0,454	0,961	0,476	-1,020	0,792
	4	0,741	0,422	0,948	0,544	-1,154	0,937
	3	0,776	0,376	0,930	0,615	-1,246	1,163
	2	0,953	0,300	0,881	0,794	-1,626	1,475
	1	1,100	0,168	0,789	1,031	-1,945	2,093

Sheet 5 CAT simulation results (transformed).

In this sheet, root mean squared error and limits of agreement are presented as transformed (0-100) CLEFT-Q scores. Simulations results for the social function scale are presented following either listwise exclusion or imputation. RMSE: Root mean squared error; LoA: 95% limit of agreement.

	Outliers included				
Scale	**Items**	**Correlation**	**RMSE**	**Lower LoA**	**Upper LoA**
Face (9 items total)	9	1,000	0,00	0,00	0,00
	8	0,997	1,67	-3,52	2,92
	7	0,989	3,19	-6,80	5,46
	6	0,983	4,01	-8,48	7,00
	5	0,972	5,07	-10,41	9,35
	4	0,964	5,71	-11,56	10,79
	3	0,949	6,76	-13,49	13,04
	2	0,912	8,80	-17,42	17,09
	1	0,812	12,64	-23,70	25,74
Jaw (7 items total)	7	1,000	0,00	0,00	0,00
	6	0,997	2,24	-4,09	4,64
	5	0,992	3,68	-6,82	7,57
	4	0,985	5,23	-9,70	10,72
	3	0,980	6,28	-11,68	12,86
	2	0,967	8,24	-15,04	17,07
	1	0,915	12,73	-23,20	26,39
Teeth (8 items total)	8	1,000	0,00	0,00	0,00
	7	0,995	2,14	-3,87	4,46
	6	0,989	3,17	-6,33	6,12
	5	0,982	4,13	-8,20	7,98
	4	0,968	5,47	-10,74	10,74
	3	0,942	7,37	-14,98	13,84
	2	0,894	9,94	-20,92	17,52
	1	0,823	12,57	-26,46	22,12
School (10 items total)	10	1,000	0,00	0,00	0,00
	9	0,996	2,00	-4,20	3,54
	8	0,991	2,92	-6,07	5,28
	7	0,987	3,53	-7,28	6,52
	6	0,975	4,97	-10,26	9,12
	5	0,959	6,32	-13,09	11,56
	4	0,937	7,84	-16,53	13,77
	3	0,910	9,23	-19,04	16,98
	2	0,857	11,59	-24,37	20,54
	1	0,766	14,52	-30,93	24,84
Psychological function (10 items total)	10	1,000	0,00	0,00	0,00
	9	0,997	1,98	-3,70	4,03
	8	0,994	2,72	-5,27	5,41
	7	0,989	3,75	-7,20	7,53
	6	0,985	4,45	-8,23	9,15
	5	0,977	5,43	-9,92	11,26
	4	0,969	6,32	-11,64	13,04
	3	0,948	8,18	-15,80	16,32
	2	0,909	10,65	-21,24	20,58
	1	0,809	14,89	-29,54	28,94

	Outliers included				
Scale	**Items**	**Correlation**	**RMSE**	**Lower LoA**	**Upper LoA**
Speech distress (10 items total)	10	1,000	0,00	0,00	0,00
	9	0,995	2,15	-4,48	3,85
	8	0,973	5,44	-11,82	8,58
	7	0,947	7,61	-16,58	11,79
	6	0,904	10,44	-22,82	15,75
	5	0,882	11,49	-25,09	17,46
	4	0,864	11,91	-25,63	19,61
	3	0,811	13,78	-29,38	23,52
	2	0,748	15,45	-32,40	27,55
	1	0,621	18,05	-33,75	36,86
Speech function (12 items total)	12	1,000	0,00	0,00	0,00
	11	0,998	1,79	-3,86	2,90
	10	0,992	3,10	-6,59	5,31
	9	0,987	4,14	-8,91	6,83
	8	0,981	4,98	-10,75	8,12
	7	0,974	5,86	-12,74	9,19
	6	0,964	6,84	-14,79	11,00
	5	0,951	8,02	-17,46	12,42
	4	0,933	9,33	-20,34	14,37
	3	0,910	10,66	-23,10	17,06
	2	0,835	14,12	-30,59	22,59
	1	0,746	16,46	-33,98	30,24
Social function (10 items total)	10	1,000	0,00	0,00	0,00
	9	0,998	1,40	-2,88	2,57
	8	0,995	2,16	-4,14	4,32
	7	0,988	3,45	-7,19	6,22
	6	0,984	4,08	-8,61	7,19
	5	0,961	6,28	-13,55	10,29
	4	0,952	6,81	-14,38	11,95
	3	0,935	7,89	-15,92	15,03
	2	0,887	10,23	-20,89	19,16
	1	0,770	14,01	-27,28	27,75
Social function with imputed data (10 items total)	10	1,000	0,00	0,00	0,00
	9	0,998	1,37	-2,82	2,54
	8	0,994	2,35	-4,84	4,31
	7	0,988	3,34	-7,08	5,77
	6	0,983	4,06	-8,56	7,11
	5	0,962	6,04	-13,05	9,82
	4	0,952	6,66	-14,19	11,34
	3	0,936	7,60	-15,24	14,57
	2	0,887	9,99	-20,35	18,73
	1	0,765	13,76	-26,62	27,36

Sheet 6 Voting results from stopping rule workshop. Green numbers indicate selected stopping rules.

Scale	Assessment length (items)	Number of votes
Face	3	1
	5	1
	6	7
	7	10
	9	1
Jaw	4	2
	5	6
	6	9
	7	3
Teeth	5	1
	6	15
	7	4
School	3	1
	5	1
	7	12
	8	5
	9	1
Psychological function	4	1
	5	3
	6	4
	8	11
	9	1
Speech distress	6	1
	7	1
	8	6
	9	11
	10	1
Speech function	5	1
	6	1
	8	12
	9	1
	10	4
	11	1
Social function	6	2
	7	4
	8	13
	9	1

Chapter 5

Modern Test Theory Techniques for Adaptive Testing in Short Scales Comprising Polytomous Items: A Monte Carlo Simulation Study Comparing Rasch Measurement Theory to Unidimensional and Multidimensional Graded Response Models

Harrison CJ, MD, PhD1; Sheng Loe B, PhD[2]; **Apon I, MD, MHS[3]**, MHS[3]; Sidey-Gibbons CJ, PhD[4];
Swan MC, PhD[5]; Furniss D, PhD[1]; Klassen AF, DPhil[6]; Wong Riff KWY, MD, PhD[7];
Versnel SL, MD, PhD[8]; Koudstaal MJ, MD, DMD, PhD[3]; Allori AC, MD, PhD[9];
Rogers-Vizena CR, MD[10]; Rodrigues JN, PhD[11,12]

[1] Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK
[2] The Psychometrics Centre, University of Cambridge, Cambridge, UK
[3] Department of Oral and Maxillofacial Surgery, the Dutch Craniofacial Center, Erasmus University Medical Center, Rotterdam, the Netherlands
[4] MD Anderson Center for INSPiRED Cancer Care, the University of Texas, Houston, Texas, USA
[5] The Spires Cleft Centre, John Radcliffe Hospital, Oxford University Hospitals, Oxford, UK
[6] Department of Pediatrics, McMaster University, Hamilton, Ontario, Canada
[7] Department of Plastic and Reconstructive Surgery, Hospital for Sick Children, Toronto, Ontario, Canada
[8] Department of Plastic and Reconstructive Surgery, the Dutch Craniofacial Center, Erasmus University Medical Center, Rotterdam, the Netherlands
[9] Division of Plastic, Maxillofacial & Oral Surgery, Duke University Hospital & Children's Health Center, Durham, North Carolina, USA
[10] Department of Plastic and Oral Surgery, Boston Children's Hospital, Boston, Massachusetts, USA
[11] Department of Plastic Surgery, Stoke Mandeville Hospital, Buckinghamshire Healthcare NHS Trust, Aylesbury, UK
[12] Warwick Clinical Trials Unit, University of Warwick, Coventry, UK

Manuscript submitted

Abstract

Background: There are two philosophical approaches to contemporary psychometrics: Rasch measurement theory (RMT) and item response theory (IRT). Either measurement strategy can be applied to computerized adaptive testing (CAT). There are potential benefits of IRT over RMT with regards to measurement precision, but also potential risks to measurement generalizability. RMT CAT assessments have demonstrated good performance with the CLEFT-Q, a patient-reported outcome measure for use in orofacial clefting.

Objectives: To test whether the post-hoc application of IRT (graded response models, GRMs, and multidimensional GRMs) to RMT-validated CLEFT-Q appearance scales could improve CAT accuracy at given assessment lengths.

Methods: Partial credit Rasch models, unidimensional GRMs and a multidimensional GRM were calibrated for each of the 7 CLEFT-Q appearance scales (which measure the appearance of the: face, jaw, teeth, nose, nostrils, cleft lip scar and lips) using data from the CLEFT-Q field test. A second, simulated dataset was generated with 1000 plausible response sets to each scale. Rasch and GRM scores were calculated for each simulated response set, scaled to 0-100 scores, and compared by Pearson's correlation coefficient, root mean square error (RMSE), mean absolute error (MAE) and 95% limits of agreement. For the face, teeth and jaw scales, we repeated this in a third, independent, real patient dataset. We then used the simulated data to compare the performance of a range of fixed-length CAT assessments that were generated with partial credit Rasch models, unidimensional GRMs and the multidimensional GRM. Median standard error of measurement (SEM) was recorded for each assessment. CAT scores were scaled to 0-100 and compared to linear assessment Rasch scores with RMSE, MAE and 95% limits of agreement. This was repeated in the independent, real patient dataset with the RMT and unidimensional GRM CAT assessments for the face, teeth and jaw scales to test the generalizability of our simulated data analysis.

Results: Linear assessment scores generated by Rasch models and unidimensional GRMs showed close agreement, with RMSE ranging from 2.2 to 6.1, and MAE ranging from 1.5 to 4.9 in the simulated dataset. These findings were closely reproduced in the real patient dataset. Unidimensional GRM CAT algorithms achieved lower median SEM than Rasch counterparts, but reproduced linear assessment scores with very similar accuracy (RMSE, MAE and 95% limits of agreement). The multidimensional GRM had poorer accuracy than the unidimensional models at comparable assessment lengths.

Conclusion: Partial credit Rasch models and GRMs produce very similar CAT scores. GRM CAT assessments achieve a lower SEM, but this does not translate into better accuracy. Commonly used SEM heuristics for target measurement reliability should not be generalized across CAT assessments built with different psychometric models. In this study, a relatively parsimonious multidimensional GRM CAT algorithm performed more poorly than unidimensional GRM comparators.

Keywords: Rasch measurement theory, item response theory, graded response model, computerized adaptive testing, patient-reported outcome measures.

Introduction

Modern test theory

There are two competing philosophies within modern test theory: Rasch measurement theory (RMT)[1] and item response theory (IRT).[2] RMT is prescriptive. In this school of thought, items are selected, modified or discarded based on their fit to a strict model in which items only differ by *difficulty* and respondents only differ by *ability*. IRT is descriptive. A range of models, including the Rasch model[1] and graded response model (GRM),[3] can be calibrated to describe patterns within item responses. In addition to *difficulty*, IRT models can evaluate *discrimination* (the discriminative potential of an item at different levels of the measured construct), among other item properties.

Proponents of the Rasch model argue for its unique properties of sufficiency, separability and specific objectivity.[4] Some argue that this allows RMT validated tools to uniquely transcend study samples and generalize to different populations where other IRT models would not.[5] On the other hand, Rasch sceptics argue that the simplicity of the Rasch model is not a realistic representation of real-world item performance.[6] The added complexity of IRT models may predispose them to overfitting, but if these models did hold true across samples, their complexity could provide more reliably and precise measurement at the item level. More complex still are the family of models belonging to multidimensional item response theory (MIRT).[7] In MIRT, a single item may measure more than one latent construct, and statistical relationships between different factors can be accounted for at the item level. As with the debate between RMT and IRT, potential gains in accuracy and efficiency with MIRT are weighed against risks to generalizability.

Over the last decade this discourse has slowly diffused into clinical research, where the use of patient-reported outcome measures (PROMs) to quantify latent traits such as depression, pain, and functional ability has become popular.

The CLEFT-Q

The CLEFT-Q is a PROM for use in orofacial clefting. Within the CLEFT-Q appearance domain, there are 7 scales that measure appearance of the: face, nose, nostrils, lips, cleft lip scar, jaws and teeth. These scales range from 6 to 12 items in length, with 4 response options per item. The CLEFT-Q was developed in line with RMT and Rasch model fit has been described for these scales previously.[8]

Additionally, we have shown that these scales fit a second order structural model in which the second order factor (appearance of the face) is measured both by face scale items and the 6 first order factors, with residual covariance between the nose and nostrils factors, and between the lips and cleft lip scar factors (article in press).

Computerized adaptive testing

Computerized adaptive testing (CAT) describes the process of using modern test theory algorithms to selectively administer items from a full-length scale based on an individual's previous responses.[9] After each item, a respondent's latent trait is measured with increasing precision until a stopping rule is met - often a prespecified number of items or standard error of measurement (SEM, the expected resampling error over repeated measurement in an individual with an unchanging true score).[10] This can make PROM administration less burdensome and more personalized.[11]

Unidimensional RMT-based CAT algorithms are capable of item reduction in each of the CLEFT-Q appearance scales with good accuracy (agreement between CAT and linear assessment scores).[12] But these algorithms have a relatively high SEM compared to CAT algorithms built with other IRT models, for example those used in the Patient-Reported Outcomes Measurement Information System (PROMIS).[13] This is because closer-fitting IRT models allow for greater measurement reliability (lower SEM) within the constraints of potentially less generalizable models. IRT- or MIRT-based CAT algorithms might achieve lower SEM values at a given assessment length, but this does not guarantee that they could reproduce linear assessment scores more accurately.

Hypotheses

In this series of experiments, we aimed to explore whether IRT or MIRT could improve on the performance of RMT CAT algorithms for the CLEFT-Q appearance scales, and we examine the usefulness of SEM as a statistic for comparing the criterion validity of CAT algorithms built with different psychometric models.

We hypothesized that:

1. CLEFT-Q appearance scales would demonstrate statistical fit to unidimensional and multidimensional GRMs.
2. Linear assessment scores generated through GRMs would be similar to scores generated through partial credit Rasch models.

3. At similar assessment lengths, unidimensional GRM CAT assessments would reproduce linear assessment scores more accurately than RMT CAT assessments, and multidimensional GRM CAT assessments would provide more accuracy still.

Methods

Software

We conducted all analyses in *R 4.0.3* with the following packages: *foreign 0.8-81*, *plyr 1.8.6*, *dplyr 1.0.4*, *mirt 1.33.2*,[14] *mirtCAT 1.10,*[15] and *ggplot2 3.3.*

Study participants

We used two independent datasets in this study. Models were calibrated from CLEFT-Q responses collected during the CLEFT-Q field test (calibration dataset). This was a prospective international study that sampled 2434 participants in 12 countries, who were aged 8 to 29 years and born with an orofacial cleft. The study recruited between October 2014 and November 2016.[8]

For external validation, we used a sample of 561 CLEFT-Q response sets collected as part of routine care at three sites across the USA and the Netherlands between November 2015 and April 2019 (validation dataset). Within this dataset, response sets were available for face, teeth and jaw scales. These were collected from 7- to 28-year-olds at time frames specified in the International Consortium for Health Outcomes Measurement (ICHOM) Standard Set for cleft lip and/or palate.[16]

Sample sizes and missing data

After listwise exclusion we included 895 out of 2434 response sets from the calibration dataset that had no missing responses to the 58 items in the CLEFT-Q appearance domain. This sample size is considered large enough for definitive model calibration[17] but is likely to cause type I errors in the c^2 fit statistic.[18]

Missing calibration data were largely explained by data collection processes in the CLEFT-Q field test. Participants who were not born with a cleft lip did not complete the cleft lip scar scale, and the jaw scale was only administered to those aged 12 years or older.[8] This precluded 1179 participants from these analyses.

The calibration dataset contained 551 responses to the face and teeth scales, and 317 responses to the jaw scale, which was not administered to patients aged under 12 years.

Unidimensional model calibration

Partial credit Rasch models and GRMs were fitted for each scale in the calibration dataset using an expectation maximization (EM) approach.[14] GRM fit was assessed using the following statistics and thresholds to suggest good fit: c^2 (p value > 0.05), root mean square error of approximation (RMSEA, < 0.06), standardized root mean square residual (SRMR, ≤ 0.08), Tucker-Lewis index (TLI, ≥ 0.95), comparative fit index (CFI, ≥ 0.95).[19]

Multidimensional model calibration

To calibrate the multidimensional GRM, we first conducted a full-information exploratory factor analysis (EFA) using a quasi-Monte Carlo EM algorithm with 7 factors and oblimin rotation.[14] A confirmatory multidimensional GRM was then calibrated based on the EFA results, with item cross-loadings and factor correlations included where they made sense clinically. The *mirt* package does not readily support hierarchical models (such as the second order model previously proposed for the CLEFT-Q appearance domain). We therefore adopted a 7 factor first order confirmatory model. The bifactor model, which assumes group factor orthogonality, was ruled out on clinical grounds.[20]

Simulated data

We tested our second and third hypotheses with simulated data for all scales, and then with real patient data from the validation dataset for the face, teeth and jaw scales. To generate simulated response sets with realistic score distributions, expected *a posteriori* (EAP) factor scores (logits) were first calculated for each calibration dataset participant based on Rasch models and scaled into 0-100 scores for each scale. We randomly resampled these scores (with replacement) 1000 times for each scale and generated plausible response sets for each resampled score using the *mirtCAT* package.[15] For the subsequent CAT experiments that used the simulated dataset, these resampled scores were taken as ground truth. CAT algorithms aimed to reproduce these scores from the plausible response sets (**Figure 1**).

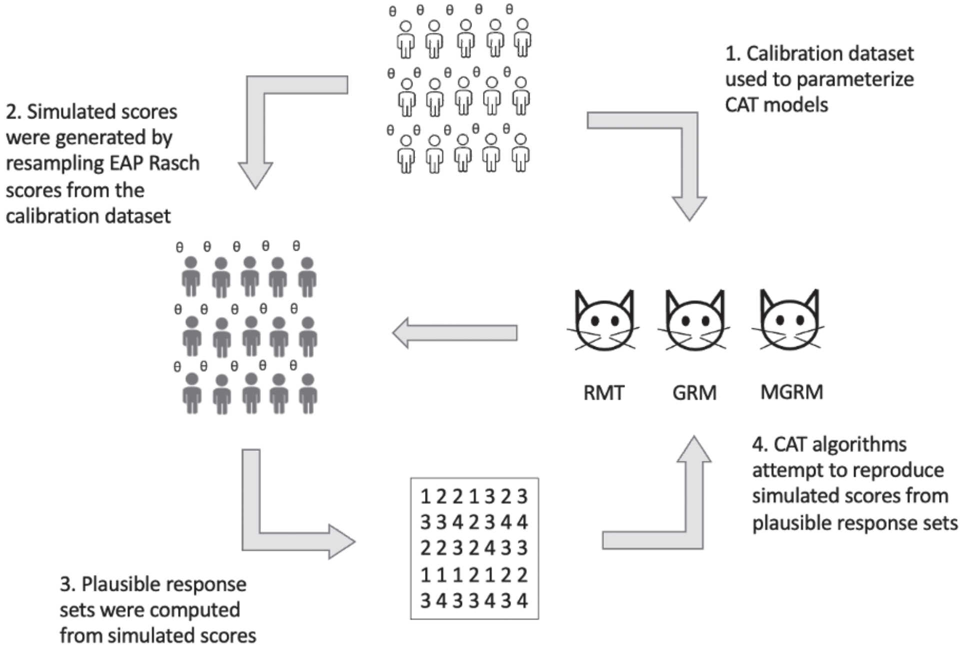

1. Calibration dataset used to parameterize CAT models
2. Simulated scores were generated by resampling EAP Rasch scores from the calibration dataset
RMT
GRM
MGRM
4. CAT algorithms attempt to reproduce simulated scores from plausible response sets
1221323
3342344
2232433
1112122
3433434
3. Plausible response sets were computed from simulated scores

Figure 1 Schematic illustrating simulated data flow in CAT comparison. CAT: computerized adaptive test, EAP: expected *a posteriori*; RMT: Rasch measurement theory; GRM: graded response model; MGRM: multidimensional graded response model.

Comparison of linear assessment scores between models

For each respondent in the simulated dataset, we administered two linear assessments. The first was scored according to the Rasch models and the second was scored according to the unidimensional GRMs. Scores were scaled into 0-100 form for both assessments and compared with the following statistics: Pearson's correlation coefficient, root mean square error (RMSE), mean absolute error (MAE) and 95% limits of agreement.[21] For the face, teeth and jaw scales, this was repeated with the validation dataset.

Comparison of CAT algorithms with simulated data

We then simulated CAT assessments for each of the 1000 respondents in the simulated dataset. The first set of assessments used unidimensional Rasch models for each scale and the second set of assessments used unidimensional GRMs. These CAT algorithms were programmed to terminate at all possible fixed-length stopping rules (i.e. for a scale of *n* items, *n* CAT assessments were used with stopping rules of 1, 2, 3, … *n* items). Participants were scored with an EAP approach and item selection was based on the minimum expected posterior variance.

In a third set of assessments, we used the multidimensional GRM with maximum *a posteriori* scoring (which is recommended for higher dimensional models) and item selection based on the determinant rule, including the score's posterior weight.[22] The following fixed length stopping rules were set for the whole CAT assessment (combining all scales): 10, 20, 30, 40, 50 and 58 items.

All CAT assessments then aimed to reproduce ground truth scores from the plausible response sets (**Figure 1**). We recorded median SEM for each assessment, and the mean number of items used from each scale in the case of the multidimensional GRM CAT assessment. CAT scores were scaled to 0-100 and compared to the resampled scores from which simulated response sets were initially derived. Comparisons were made using Pearson's correlation coefficient, RMSE, MAE and 95% limits of agreement.

Comparison of CAT algorithms with real patient data

We calculated linear assessment scores to the face, teeth and jaw scales, for each respondent in the validation dataset, based on the Rasch models that had been parameterized with the calibration dataset. Rasch CAT and unidimensional GRM CAT algorithms were then evaluated with the validation dataset responses. In each case, CAT scores were compared to the (Rasch) linear assessment scores for each individual. These CAT assessments used the same settings as those used with the simulated dataset. In the absence of real responses to the other scales, multidimensional CAT was not assessed with the validation dataset.

Results

Model calibration

Fit statistics suggested reasonable fit of each of the unidimensional GRMs, although only the face and teeth scales met all 5 fit statistic thresholds (**Sheet 1, Supplemental Material**).

The results of the EFA (**Sheet 2, Supplemental Material**) suggested the following, clinically plausible, multidimensional model specifications:

- Face item 6 (which relates to smiling) cross loads onto the teeth factor and face factor;
- Face item 7 (which relates to laughing) cross loads onto the teeth factor and face factor;
- Face item 8 (which relates to facial profile) cross loads onto the nose factor and face factor;
- Nose item 12 (which relates to nasal symmetry) cross loads onto the nostrils factor and nose factor;
- Nose and nostrils factors are correlated;
- Lips and cleft lip scar factors are correlated.

In addition to the primary item loadings, these parameters were freely estimated in the confirmatory multidimensional GRM (the parameters of the confirmatory model are available in **Sheet 3, Supplemental Material**). The confirmatory model demonstrated moderate fit (c^2 $p < 0.001$, RMSEA 0.076, SRMR 0.39, TLI 0.96, CFI 0.96).

Linear assessments

In the simulated dataset, scaled Rasch and unidimensional GRM scores from the linear assessments were similar, with RMSE ranging from 2.2 to 6.1, and MAE ranging from 1.5 to 4.9 (**Table 1**). The nose scale (the longest scale, with 12 items) had the closest concordance between Rasch and GRM scores, followed by the face and teeth scales (which had demonstrated best fit to the unidimensional GRM).

Scale	Pearson's correlation coefficient	RMSE	MAE	Upper 95% limit of agreement	Lower 95% limit of agreement
Face	1.00	2.2	1.7	2.4	-4.8
Nose	1.00	2.0	1.5	3.1	-4.3
Nostrils	1.00	3.4	2.6	7.0	-6.1
Lips	1.00	3.7	3.0	4.6	-8.1
Scar	0.99	4.6	3.5	6.8	-10.0
Teeth	1.00	2.7	2.1	3.8	-5.9
Jaw	1.00	6.1	4.9	5.5	-13.1

Table 1 Correlation and agreement between Rasch and unidimensional graded response model scores following linear assessments in the simulated dataset. Scores have been scaled to 0-100 form. RMSE: root mean square error; MAE: mean absolute error.

These findings were reproduced for the three scales in the validation dataset (**Table 2**) and are illustrated for the face scale in **Figure 2**.

Scale	Pearson's correlation coefficient	RMSE	MAE	Lower 95% limit of agreement	Upper 95% limit of agreement
Face	1.00	1.9	1.3	2.8	-4.1
Teeth	0.99	2.8	2.1	2.8	-6.1
Jaw	1.00	5.4	4.2	3.6	-11.3

Table 2 Correlation and agreement between Rasch and unidimensional graded response model scores following linear assessments in the validation dataset. Scores have been scaled to 0-100 form. RMSE: root mean square error; MAE: mean absolute error.

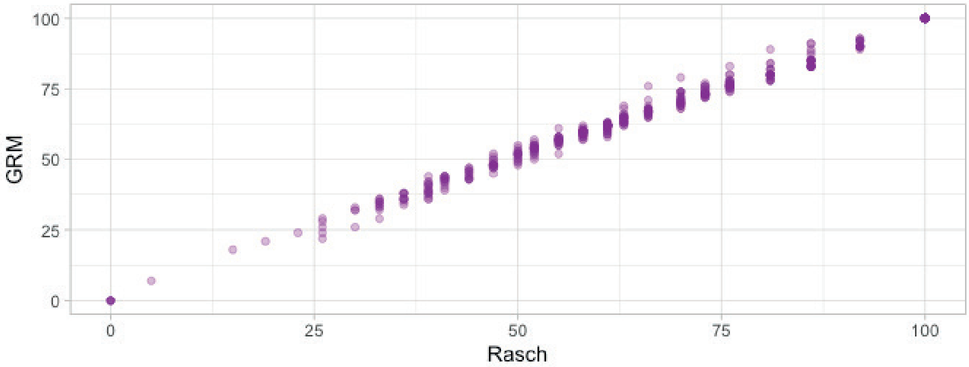


Figure 2 Scatterplot demonstrating the relationship between Rasch and unidimensional graded response model scores following face scale linear assessments in the validation dataset. Scores have been scaled to 0-100 form. GRM: graded response model.

CAT assessments with simulated data

The results of the CAT assessments in the simulated dataset are presented in **Sheet 4, Supplemental Material**. Results for the unidimensional face scale CAT assessments with the simulated dataset are presented in **Table 3** for illustration.

At a given assessment length, GRM CAT algorithms consistently achieved lower SEM. However, accuracy (MAE, RMSE and 95% limits of agreement) was remarkably similar between both unidimensional approaches.

Number of items	Model	Median SEM	MAE	RMSE	Correlation	Lower 95% limit of agreement	Upper 95% limit of agreement	95% limit of agreement range
9	Rasch	0.51	0.0	0.0	1.00	0.0	0.0	0.0
	GRM	0.27	1.3	1.9	1.00	-4.1	2.8	6.9
8	Rasch	0.53	1.3	1.8	1.00	-3.9	2.7	6.6
	GRM	0.27	1.8	2.4	0.99	-5.3	3.9	9.2
7	Rasch	0.56	2.4	3.2	0.99	-6.8	5.4	12.2
	GRM	0.28	2.2	3.2	0.99	-6.8	5.8	12.6
6	Rasch	0.59	3.3	4.1	0.98	-8.6	7.1	15.7
	GRM	0.30	2.9	4.0	0.98	-8.0	7.6	15.6
5	Rasch	0.64	3.8	4.9	0.97	-10.3	8.5	18.8
	GRM	0.32	3.6	4.7	0.97	-9.6	8.8	18.4
4	Rasch	0.71	5.0	6.2	0.96	-12.3	11.8	24.1
	GRM	0.35	4.4	5.7	0.96	-11.3	11.0	22.3
3	Rasch	0.78	5.6	6.9	0.95	-13.8	13.3	27.2
	GRM	0.39	5.7	7.2	0.94	-14.4	13.9	28.3
2	Rasch	0.91	7.2	9.0	0.91	-17.4	17.8	35.2
	GRM	0.47	6.8	8.4	0.92	-17.0	16.0	33.0
1	Rasch	1.14	10.1	12.6	0.81	-22.6	26.4	49.0
	GRM	0.57	10.1	13.0	0.79	-23.8	26.7	50.5

Table 3 Results of the face scale computerized adaptive testing assessments in the simulated dataset. The full-length face scale contains 9 items. Results are compared to prespecified Rasch scores that were resampled from the calibration dataset and used to generate plausible response options. SEM: standard error of measurement; RMSE: root mean square error; MAE: mean absolute error; GRM: graded response model; MGRM: multidimensional graded response model.

The multidimensional GRM showed poorer performance than both unidimensional models (**Figure 3 and Sheet 4, Supplemental Material),** with lower correlation and higher MAE, RMSE and 95% limits of agreement than unidimensional assessments with similar length.

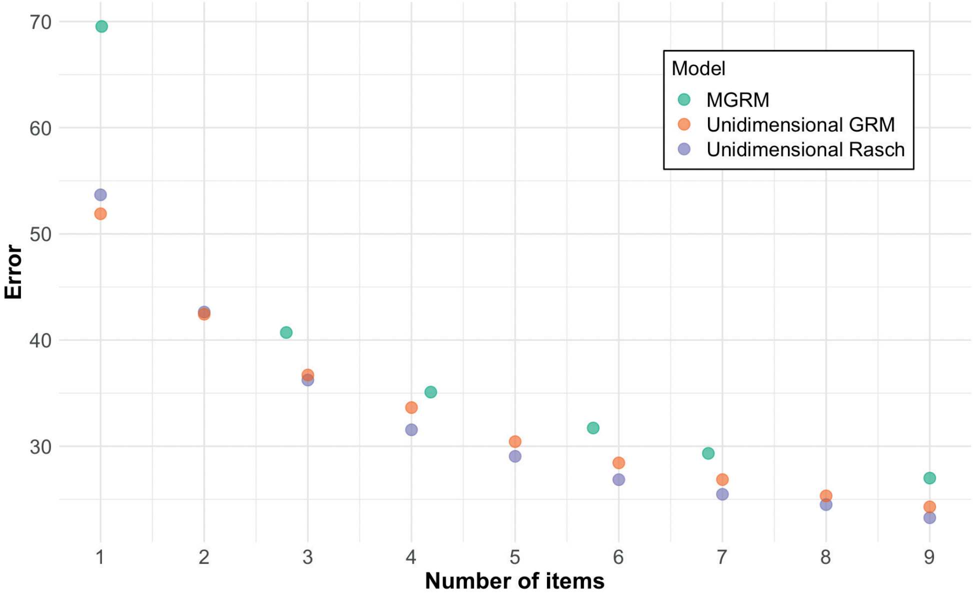


Figure 3 Scatterplot comparing assessment length (mean number of items) and error (the range between 95% limits of agreement) in face scale CAT algorithms built with MGRM, unidimensional GRMs and unidimensional Rasch models. MGRM: multidimensional graded response model; GRM: graded response model.

CAT assessments with real patient data

In the validation dataset, unidimensional GRM CAT assessments for the face, jaw and teeth scales achieved lower SEM than their RMT counterparts, but as with the simulated data, both models achieved very similar accuracy statistics (**Sheet 4, Supplemental Material**).

Discussion

Key findings

We found a close concordance between linear assessments scored with Rasch models and unidimensional GRMs, using simulated datasets that mimicked real world response distributions. Score agreement was highest in the 12-item nose scale (the longest scale tested) and the face and teeth scales (which demonstrated closest GRM fit). Agreement was good even in scales as short as 6 or 7 items with moderate GRM fit. Linear assessment score agreement was confirmed for the face, teeth and jaw scales in an independent sample of real patient responses. This suggests that unidimensional GRMs can be used to reliably score RMT-validated PROM scales with similar structural properties to those tested here.

At a given assessment length, unidimensional GRM CAT assessments achieved considerably lower SEM than RMT comparators in both simulated and real datasets. The unconstrained *discrimination* parameter of the GRM allows closer model fit, and if one assumes GRM parameters hold true across samples, this facilitates more reliable and precise measurement than RMT. This effect may be pronounced further in 3 and 4 parameter IRT models which capture additional response properties such as *inattention* and *guessing*. However, SEM did not reflect the accuracy of CAT algorithms (their ability to reproduce full-length linear assessment scores) in our experiments. SEM, a resampling error prediction based on cross-sectional data, is difficult to conceptualize in real-world terms and we propose that the ability of a CAT assessment to accurately reproduce full-length assessment scores at an individual- and population-level is more clinically important than SEM, which is influenced by model flexibility. This study challenges the generalizability of commonly used SEM heuristics for target measurement reliability across CAT assessments built with different psychometric models.[11,23]

For each scale, at comparable assessment lengths, the multidimensional GRM CAT algorithm achieved poorer accuracy than the corresponding unidimensional GRM CAT algorithms, contrary to our third hypothesis. There are a number of possible reasons for this. Different item parameter estimation methods, item selection criteria and score calculation techniques are required for high-dimensional adaptive testing, and these may have biased our comparison.[14,22]

Furthermore, the EFA suggested a relatively parsimonious multidimensional model, with clinically plausible cross loading of only 4 items out of 58, onto a maximum of 2 factors each. This is probably because items in the CLEFT-Q were selected for independence (in keeping with RMT), but there may also be a contributing framing effect: items in the CLEFT-Q field test were presented in a booklet of independent scales, with one scale to a page. A multidimensional CAT assessment would not necessarily group items by scale. It may start with an item from the face scale, then pose an item from the jaw scale, then the teeth scale and so on. It is possible that this mode of administration could result in more cross loading than was demonstrated in our datasets.

Strengths and limitations

This study examines the post-hoc application of IRT to scales that have been developed with RMT. We make no conclusions about the choice of RMT or IRT for scale development. In this study, we have chosen linear assessment scores calculated with Rasch models as ground truth to compare CAT scores with. All statistical models are imperfect representations of real-world phenomena, and while we can measure the agreement between scores obtained through different approaches, we cannot infer whether GRM or RMT scores are *truer*.

There are strengths and limitations of the simulated data used in this study. Where real data are unavailable, Monte Carlo CAT simulations studies often use randomly generated datasets with prespecified distributions, usually $N(0,1)$.[15] To create a more realistically distributed dataset, we resampled Rasch factor scores from the calibration data and computed plausible response sets to match these scores. Theoretically, there is a chance of data leakage with this approach that could have favored the performance of Rasch CAT assessments in the simulated data. However, we were able to reproduce simulated data results in all 3 scales included in the independent, real patient, validation data.

Comparison to other literature

Previous studies have suggested similarities between measurements obtained through RMT, IRT and CTT.[24,25] The International Society for Quality of Life Research (ISOQOL) Psychometrics Special Interest Group recently published three papers that aimed to create and evaluate scales from the PROMIS depression item bank using RMT,[26] IRT[27] and CTT.[28] These techniques achieved similar results, and an undogmatic approach to PROM development has been suggested in response.[29]

Our study focuses on the choice of model to adopt following scale development, for the purposes of CAT. This is timely, as the application of CAT to PROMs has been made recently popular through high-profile CAT initiatives such as PROMIS,[30] and RMT-validated PROMs are becoming increasingly available. It is likely that the interchangeability of GRM and RMT measurement described in this study is generalizable to RMT-validated PROMs in other clinical fields.

Conclusion

This study suggests CAT assessments can be built for RMT-validated item banks using RMT or IRT. Scoring is very similar between both approaches, and lower SEM values found in IRT CATs do not represent better CAT accuracy. Our relatively parsimonious MIRT CAT algorithm performed more poorly than unidimensional GRM CAT algorithms.

Acknowledgements

Conrad J. Harrison is funded by a National Institute for Health Research (NIHR) Doctoral Research Fellowship (NIHR300684) for this research project. Jeremy N. Rodrigues is funded by a NIHR postdoctoral fellowship (PDF-2017-10-075). Dominic Furniss is funded by the NIHR Oxford Biomedical Research Centre. The CLEFT-Q field-test study was funded by a grant from the Canadian Institute of Health Research (CIHR, FRN-130515).

This document presents independent research funded by the NIHR and the CIHR. The views expressed are those of the authors and not necessarily those of the CIHR, the NHS, the NIHR or the Department of Health and Social Care.

Conflict of interests

The CLEFT-Q is owned by McMaster University and the Hospital for Sick Children. Anne F Klassen and Karen W.Y. Wong Riff are co-developers of the CLEFT-Q and, as such, could potentially receive a share of any license revenues based on their institutions inventor sharing policy. The other authors have no conflicts of interest to declare in relation to the content of this article.

References

1. Andrich D. Rating scales and Rasch measurement. *Expert Rev Pharmacoecon Outcomes Res*. 2011;11(5):571-585. doi:10.1586/erp.11.59.
2. Cai L, Choi K, Hansen M, Harrell L. Item Response Theory. *Annu Rev Stat Its Appl*. 2016;3(1):297-321. doi:10.1146/annurev-statistics-041715-033702.
3. Samejima F. Graded Response Model. In: van der Linden WJ, Hambleton RK, eds. *Handbook of Modern Item Response Theory*. Springer New York; 1997:85-100. doi:10.1007/978-1-4757-2691-6_5.
4. Van Breukelen GJP. Separability of item and person parameters in response time models. *Psychometrika*. 1997;62(4):525-544. doi:10.1007/BF02294641.
5. Shaw F. Descriptive IRT vs. Prescriptive Rasch. *Rasch Meas Trans*. 1991;5(1):131.
6. Harvey RJ. Improving Measurement via Item Response Theory: Great Idea, But Hold the Rasch. *Couns Psychol*. 2016;44(2):195-204. doi:10.1177/0011000015615427.
7. Reckase MD. The Past and Future of Multidimensional Item Response Theory. *Appl Psychol Meas*. 1997;21(1):25-36. doi:10.1177/0146621697211002.
8. Klassen AF, Riff KWW, Longmire NM, et al. Psychometric findings and normative values for the CLEFT-Q based on 2434 children and young adult patients with cleft lip and/or palate from 12 countries. *Can Med Assoc J*. 2018;190(15):E455-E462. doi:10.1503/cmaj.170289.
9. Weiss DJ, Vale CD. Adaptive Testing. *Appl Psychol*. 1987;36(3-4):249-262. doi:10.1111/j.1464-0597.1987.tb01190.x.
10. Dudek FJ. The continuing misinterpretation of the standard error of measurement. *Psychol Bull*. 1979;86(2):335-337. doi:10.1037/0033-2909.86.2.335.
11. Gibbons C, Bower P, Lovell K, Valderas J, Skevington S. Electronic Quality of Life Assessment Using Computer-Adaptive Testing. *J Med Internet Res*. 2016;18(9):e240. doi:10.2196/jmir.6053.
12. Harrison CJ, Rodrigues JN, Furniss D, et al. Optimising the computerised adaptive test to reliably reduce the burden of administering the CLEFT-Q: A Monte Carlo simulation study. *J Plast Reconstr Aesthet Surg*. 2021;74(6):1355-1401. doi:10.1016/j.bjps.2020.12.029.
13. Hung M, Baumhauer JF, Latt DL, Saltzman CL, SooHoo NF, Hunt KJ. Validation of PROMIS® Physical Function Computerized Adaptive Tests for Orthopaedic Foot and Ankle Outcome Research. *Clin Orthop*. 2013;471(11):3466-3474. doi:10.1007/s11999-013-3097-1.
14. Chalmers RP. mirt : A Multidimensional Item Response Theory Package for the R Environment. *J Stat Softw*. 2012;48(6). doi:10.18637/jss.v048.i06.
15. Chalmers RP. Generating Adaptive and Non-Adaptive Test Interfaces for Multidimensional Item Response Theory Applications. *J Stat Softw*. 2016;71(5). doi:10.18637/jss.v071.i05.
16. Allori AC, Kelley T, Meara JG, et al. A Standard Set of Outcome Measures for the Comprehensive Appraisal of Cleft Care. *Cleft Palate Craniofac J*. 2017;54(5):540-554. doi:10.1597/15-292.
17. Linacre J. Sample Size and Item Calibration [or Person Measure] Stability. *Rasch Meas Trans*. 1994;7(4):328.
18. David Kenny. Measuring model fit. Published 2015. Accessed August 26, 2021. http://www.davidakenny.net/cm/fit.htm.
19. Schreiber JB, Nora A, Stage FK, Barlow EA, King J. Reporting Structural Equation Modeling and Confirmatory Factor Analysis Results: A Review. *J Educ Res*. 2006;99(6):323-338. doi:10.3200/JOER.99.6.323-338.

20. Mansolf M, Reise SP. When and why the second-order and bifactor models are distinguishable. *Intelligence*. 2017;61:120-129. doi:10.1016/j.intell.2017.01.012.

21. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Int J Nurs Stud*. 2010;47(8):931-936. doi:10.1016/j.ijnurstu.2009.10.001.

22. Chalmers P. Package “mirt.” Published 2021. Accessed August 26, 2021. https://cran.r-project.org/web/packages/mirt/mirt.pdf.

23. Loe BS, Stillwell D, Gibbons C. Computerized Adaptive Testing Provides Reliable and Efficient Depression Measurement Using the CES-D Scale. *J Med Internet Res*. 2017;19(9):e302. doi:10.2196/jmir.7453.

24. Fan X. Item Response Theory and Classical Test Theory: An Empirical Comparison of their Item/Person Statistics. *Educ Psychol Meas*. 1998;58(3):357-381. doi:10.1177/0013164498058003001.

25. Stewart J. Does IRT provide more sensitive measures of latent traits in statistical tests? An empirical examination. Shiken Research Bulletin. Accessed August 26, 2021. Does IRT provide more sensitive measures of latent traits in statistical tests? An empirical examination.

26. Cleanthous S, Barbic SP, Smith S, Regnault A. Psychometric performance of the PROMIS® depression item bank: a comparison of the 28- and 51-item versions using Rasch measurement theory. *J Patient-Rep Outcomes*. 2019;3(1):47. doi:10.1186/s41687-019-0131-4.

27. Stover AM, McLeod LD, Langer MM, Chen W-H, Reeve BB. State of the psychometric methods: patient-reported outcome measure development and refinement using item response theory. *J Patient-Rep Outcomes*. 2019;3(1):50. doi:10.1186/s41687-019-0130-5.

28. Nolte S, Coon C, Hudgens S, Verdam MGE. Psychometric evaluation of the PROMIS® Depression Item Bank: an illustration of classical test theory methods. *J Patient-Rep Outcomes*. 2019;3(1):46. doi:10.1186/s41687-019-0127-0.

29. Bjorner JB. State of the psychometric methods: comments on the ISOQOL SIG psychometric papers. *J Patient-Rep Outcomes*. 2019;3(1):49. doi:10.1186/s41687-019-0134-1.

30. Cella D, Yount S, Rothrock N, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): Progress of an NIH Roadmap Cooperative Group During its First Two Years. *Med Care*. 2007;45(5):S3-S11. doi:10.1097/01.mlr.0000258615.42478.55.

Supplemental Material

Sheet 1 Unidimensional graded response model fit.

RMSEA: root mean square error of approximation; SRMR: standardized root mean square residual; TLI: Tucker-Lewis index; CFI: comparative fit index.

Scale	X2 p value	RMSEA	SRMR	TLI	CFI
Face	0,301	0,014	0,055	0,998	0,999
Nose	< 0.001	0,087	0,045	0,939	0,956
Nostrils	< 0.001	0,174	0,038	0,948	0,969
Lips	< 0.001	0,100	0,034	0,898	0,949
Scar	< 0.001	0,104	0,026	0,982	0,988
Teeth	0,164	0,027	0,033	0,993	0,998
Jaw	< 0.001	0,115	0,031	0,977	0,985

Sheet 2.1 Exploratory factor analysis item loadings.

These values are standardized pattern coefficients (factor loadings) from the exploratory factor analysis, for each item, by each factor.

Darker green cells represent higher factor loadings. Items with factor loadings > 0.3 in two factors (for example, Face 8) were reviewed and considered to cross-load where this made clinical sense.

Items	Factors						
	Scar	**Nose**	**Teeth**	**Jaw**	**Lips**	**Face**	**Nostrils**
Face 1	0,137	0,097	0,068	0,093	0,127	0,539	0,028
Face 2	0,102	0,105	0,070	0,103	0,074	0,550	0,060
Face 3	0,045	0,109	0,003	0,215	0,152	0,386	0,076
Face 4	0,044	0,155	0,038	0,071	0,076	0,586	0,071
Face 5	0,077	0,271	0,044	0,027	0,068	0,362	0,193
Face 6	0,019	0,018	0,270	0,135	0,074	0,528	0,112
Face 7	0,018	0,103	0,257	0,064	0,116	0,532	0,069
Face 8	0,089	0,482	0,021	0,174	0,222	0,323	0,107
Face 9	0,071	0,283	0,069	0,021	0,150	0,441	0,083
Nose 1	0,083	0,836	0,101	0,082	0,007	0,031	0,069
Nose 2	0,115	0,675	0,014	0,036	0,033	0,154	0,092
Nose 3	0,013	0,763	0,052	0,019	0,046	0,015	0,113
Nose 4	0,017	0,858	0,023	0,016	0,060	0,019	0,031
Nose 5	0,084	0,657	0,032	0,014	0,030	0,154	0,182
Nose 6	0,041	0,646	0,002	0,024	0,009	0,074	0,172
Nose 7	0,037	0,764	0,042	0,013	0,081	0,055	0,148
Nose 8	0,108	0,654	0,008	0,072	0,010	0,159	0,185
Nose 9	0,012	0,839	0,075	0,050	0,046	0,027	0,010
Nose 10	0,075	0,750	0,047	0,001	0,011	0,090	0,092
Nose 11	0,005	0,852	0,012	0,056	0,029	0,015	0,005
Nose 12	0,076	0,471	0,025	0,083	0,011	0,058	0,417
Nostrils 1	0,059	0,083	0,043	0,004	0,043	0,051	0,770
Nostrils 2	0,047	0,026	0,031	0,007	0,030	0,063	0,844
Nostrils 3	0,014	0,063	0,015	0,030	0,035	0,059	0,867
Nostrils 4	0,002	0,088	0,027	0,046	0,058	0,103	0,826
Nostrils 5	0,036	0,086	0,011	0,050	0,001	0,076	0,933
Nostrils 6	0,040	0,079	0,005	0,007	0,033	0,008	0,859
Jaw 1	0,006	0,067	0,017	0,938	0,040	0,039	0,043
Jaw 2	0,025	0,003	0,028	0,926	0,019	0,007	0,007
Jaw 3	0,047	0,072	0,016	0,914	0,025	0,021	0,020
Jaw 4	0,076	0,049	0,013	0,895	0,048	0,057	0,050
Jaw 5	0,013	0,035	0,017	0,948	0,034	0,048	0,014
Jaw 6	0,015	0,017	0,036	0,892	0,012	0,017	0,037
Jaw 7	0,021	0,028	0,005	0,919	0,009	0,040	0,017
Teeth 1	0,039	0,079	0,667	0,012	0,118	0,016	0,030
Teeth 2	0,075	0,078	0,818	0,027	0,051	0,076	0,035
Teeth 3	0,043	0,035	0,879	0,030	0,012	0,068	0,033

Items	Factors						
Teeth 4	0,030	0,040	0,772	0,109	0,049	0,050	0,036
Teeth 5	0,064	0,026	0,894	0,043	0,058	0,074	0,023
Teeth 6	0,050	0,056	0,797	0,066	0,078	0,074	0,025
Teeth 7	0,017	0,024	0,893	0,059	0,006	0,059	0,026
Teeth 8	0,042	0,050	0,775	0,149	0,019	0,030	0,015
Lips 1	0,064	0,070	0,106	0,002	0,740	0,111	0,088
Lips 2	0,114	0,116	0,039	0,001	0,924	0,061	0,004
Lips 3	0,070	0,050	0,008	0,012	0,887	0,057	0,031
Lips 4	0,082	0,024	0,154	0,019	0,673	0,115	0,068
Lips 5	0,148	0,033	0,007	0,005	0,794	0,083	0,016
Lips 6	0,095	0,003	0,058	0,013	0,857	0,035	0,026
Lips 7	0,024	0,015	0,016	0,024	0,898	0,032	0,060
Lips 8	0,091	0,061	0,013	0,083	0,910	0,080	0,008
Lips 9	0,150	0,048	0,013	0,032	0,753	0,034	0,035
Scar 1	0,851	0,017	0,019	0,039	0,038	0,026	0,018
Scar 2	0,866	0,037	0,078	0,015	0,097	0,029	0,043
Scar 3	0,949	0,010	0,003	0,054	0,032	0,024	0,013
Scar 4	0,913	0,038	0,032	0,052	0,018	0,003	0,006
Scar 5	0,900	0,012	0,052	0,032	0,063	0,015	0,028
Scar 6	0,888	0,008	0,072	0,018	0,009	0,014	0,041
Scar 7	0,835	0,125	0,004	0,021	0,003	0,059	0,043

Sheet 2.2 Exploratory factor analysis factor correlations.

These values are factor correlations identified in the exploratory factor analysis. The two largest correlations (the nose and nostrils factors, and the lips and scar factors) were considered clinically plausible.

	Scar	Nose	Teeth	Jaw	Lips	Face	Nostrils
Scar	1	0,439	0,492	0,589	0,602	0,388	0,48
Nose	0,439	1	0,375	0,388	0,522	0,522	0,836
Teeth	0,492	0,375	1	0,575	0,478	0,388	0,365
Jaw	0,589	0,388	0,575	1	0,488	0,359	0,356
Lips	0,602	0,522	0,478	0,488	1	0,484	0,566
Face	0,388	0,522	0,388	0,359	0,484	1	0,475
Nostrils	0,48	0,836	0,365	0,356	0,566	0,475	1

Sheet 3.1 Multidimensional graded response model parameters (items). These values are item parameters for the multidimensional graded response model. Parameters a1-a6 represent discrimination in each factor. Parameters d1-d3 relate to the item response threshold (for unidimensional models, d divided by -a produces the traditional IRT 'difficulty' parameter).

Items	Parameters									
	a1 (scar factor)	a2 (nose factor)	a3 (teeth factor)	a4 (jaw factor)	a5 (lips factor)	a6 (face factor)	a7 (nostrils factor)	d1	d2	d3
Face 1	2,648	0	0	0	0	0	0	5,429	2,289	-1,006
Face 2	2,546	0	0	0	0	0	0	5,286	2,259	-1,011
Face 3	1,807	0	0	0	0	0	0	3,744	1,528	-1,119
Face 4	2,718	0	0	0	0	0	0	3,596	0,444	-2,436
Face 5	2,122	0	0	0	0	0	0	2,803	0,195	-1,982
Face 6	2,168	0	0	0	1,283	0	0	3,674	0,658	-2,129
Face 7	2,335	0	0	0	1,095	0	0	3,307	0,497	-2,39
Face 8	1,205	0,835	0	0	0	0	0	2,017	-0,114	-2,384
Face 9	2,713	0	0	0	0	0	0	2,96	0,111	-2,9
Nose 1	0	2,122	0	0	0	0	0	3,281	0,811	-2,44
Nose 2	0	2,359	0	0	0	0	0	3,251	0,521	-2,961
Nose 3	0	2,455	0	0	0	0	0	3,444	0,142	-2,997
Nose 4	0	1,774	0	0	0	0	0	2,726	0,381	-2,018
Nose 5	0	2,438	0	0	0	0	0	2,832	-0,005	-3,216
Nose 6	0	1,79	0	0	0	0	0	2,262	-0,298	-2,718
Nose 7	0	2,115	0	0	0	0	0	2,337	-0,256	-2,802
Nose 8	0	2,407	0	0	0	0	0	2,754	-0,464	-3,679
Nose 9	0	2,535	0	0	0	0	0	2,818	-0,44	-3,685
Nose 10	0	2,282	0	0	0	0	0	2,311	-0,382	-3,254
Nose 11	0	2,386	0	0	0	0	0	2,144	-0,629	-3,424
Nose 12	0	1,112	1,454	0	0	0	0	1,497	-0,988	-3,235
Nostrils 1	0	0	4,809	0	0	0	0	3,284	-0,184	-4,515
Nostrils 2	0	0	5,175	0	0	0	0	3,183	-0,624	-4,96
Nostrils 3	0	0	4,031	0	0	0	0	2,602	-0,56	-4,078
Nostrils 4	0	0	3,894	0	0	0	0	2,453	-0,8	-4,005
Nostrils 5	0	0	4,438	0	0	0	0	2,491	-0,938	-4,506
Nostrils 6	0	0	4,236	0	0	0	0	2,173	-1,172	-4,321
Jaw 1	0	0	0	4,682	0	0	0	7,795	4,444	-0,793

Items	Parameters								
	a1 (scar factor)	a2 (nose factor)	a3 (teeth factor)	a4 (jaw factor)	a5 (lips factor)	a6 (face factor)	a7 (nostrils factor)	d1	d2	d3
Jaw 2	0	0	0	4,472	0	0	0	7,383	3,866	-0,66
Jaw 3	0	0	0	4,256	0	0	0	6,818	3,87	-0,747
Jaw 4	0	0	0	3,936	0	0	0	6,362	3,631	-0,591
Jaw 5	0	0	0	3,819	0	0	0	6,254	3,473	-0,509
Jaw 6	0	0	0	4,119	0	0	0	6,406	3,472	-0,691
Jaw 7	0	0	0	3,843	0	0	0	5,496	3,032	-0,831
Teeth 1	0	0	0	0	2,267	0	0	3,658	1,338	-1,455
Teeth 2	0	0	0	0	2,687	0	0	3,12	0,983	-1,682
Teeth 3	0	0	0	0	4,077	0	0	3,057	0,319	-2,64
Teeth 4	0	0	0	0	3,746	0	0	2,987	0,236	-2,861
Teeth 5	0	0	0	0	2,759	0	0	2,488	0,386	-2,061
Teeth 6	0	0	0	0	3,251	0	0	1,868	-0,244	-2,684
Teeth 7	0	0	0	0	4,055	0	0	2,816	-0,351	-3,832
Teeth 8	0	0	0	0	2,682	0	0	1,802	-0,132	-2,26
Lips 1	0	0	0	0	0	2,745	0	5,16	2,047	-1,451
Lips 2	0	0	0	0	0	2,635	0	4,905	2,036	-1,46
Lips 3	0	0	0	0	0	3,248	0	5,603	2,061	-1,989
Lips 4	0	0	0	0	0	3,021	0	5,217	2,089	-1,851
Lips 5	0	0	0	0	0	2,938	0	5,315	1,812	-2,302
Lips 6	0	0	0	0	0	2,474	0	4,153	1,431	-1,579
Lips 7	0	0	0	0	0	3,339	0	5,465	1,72	-2,158
Lips 8	0	0	0	0	0	2,497	0	4,167	1,331	-1,666
Lips 9	0	0	0	0	0	2,754	0	4,464	1,055	-2,603
Scar 1	0	0	0	0	0	0	4,169	4,48	1,662	-1,106
Scar 2	0	0	0	0	0	0	5,669	5,387	1,84	-1,872
Scar 3	0	0	0	0	0	0	6,21	5,922	2,356	-1,714
Scar 4	0	0	0	0	0	0	6,397	6,092	2,028	-1,821
Scar 5	0	0	0	0	0	0	5,124	4,412	1,572	-1,755
Scar 6	0	0	0	0	0	0	5,852	5,329	1,618	-2,172
Scar 7	0	0	0	0	0	0	5,642	4,923	1,318	-2,341

Sheet 3.2 Multidimensional graded response model parameters (factor correlations). These values are the freely estimated factor correlations in the multidimensional graded response model.

	Scar	Nose	Teeth	Jaw	Lips	Face	Nostrils
Scar	1	0	0	0	0	0	0
Nose	0	1	0,614	0	0	0	0
Teeth	0	0,614	1	0	0	0	0
Jaw	0	0	0	1	0	0	0
Lips	0	0	0	0	1	0	0
Face	0	0	0	0	0	1	0,656
Nostrils	0	0	0	0	0	0,656	1

Sheet 4.1 Monte Carlo simulation results (simulated data).

All comparisons are made with unidimensional, linear Rasch assessment scores (transformed into 0-100 format). Each table relates to a different CAT assessment.

For example, Table S4.1 relates to the Face scale CAT using a Rasch model. Mean absolute error, root mean squared error, correlation, and limits of agreement are calculated through comparison between CAT scores and linear Rasch scores.

CAT: computerized adaptive test; SEM: standard error of measurement; MAE: mean absolute error; RMSE: root mean square error; correlation: Pearson's correlation coefficient; LA: limits of agreement.

Table S4.1 Face scale CAT (Rasch).

Number of items	Median SEM	MAE	RMSE	Correlation	95% LA (upper boundary)	95% LA (lower boundary)	95% LA range
1	1,14	11,1	13,7	0,76	27,1	-26,6	53,7
2	0,91	8,6	10,9	0,85	21,6	-21,1	42,6
3	0,78	7,3	9,2	0,90	18,5	-17,7	36,2
4	0,69	6,3	8,1	0,92	16,1	-15,4	31,5
5	0,63	5,8	7,4	0,94	14,8	-14,3	29,1
6	0,58	5,4	6,9	0,95	13,7	-13,2	26,9
7	0,55	5,0	6,5	0,95	12,9	-12,6	25,5
8	0,52	4,9	6,2	0,95	12,1	-12,4	24,5
9	0,50	4,5	5,9	0,96	11,7	-11,6	23,3

Table S4.2 Face scale CAT (unidimensional GRM).

Number of items	Median SEM	MAE	RMSE	Correlation	95% LA (upper boundary)	95% LA (lower boundary)	95% LA range
1	0,57	10,6	13,2	0,78	25,5	-26,4	51,9
2	0,46	8,7	10,8	0,86	20,5	-21,9	42,4
3	0,39	7,5	9,4	0,90	17,5	-19,2	36,7
4	0,35	6,9	8,6	0,91	15,9	-17,8	33,6
5	0,32	6,1	7,8	0,93	14,3	-16,2	30,4
6	0,30	5,7	7,3	0,94	13,2	-15,3	28,4
7	0,28	5,4	6,9	0,95	12,3	-14,6	26,9
8	0,27	5,1	6,6	0,95	11,6	-13,8	25,3
9	0,26	4,9	6,3	0,96	11,1	-13,2	24,3

Table S4.3 Face scale CAT (multidimensional GRM).

Total item limit for all scales combined	Mean face scale items	Median SEM	MAE	RMSE	Correlation	95% LA (upper boundary)	95% LA (lower boundary)	95% LA range
10	1,01	0,56	13,9	17,7	0,74	34,5	-35,0	69,5
20	2,79	0,39	8,7	10,6	0,87	22,5	-18,2	40,7
30	4,19	0,35	7,5	9,3	0,90	20,2	-14,9	35,1
40	5,75	0,31	7,4	9,2	0,92	20,2	-11,5	31,7
50	6,86	0,29	6,9	8,6	0,93	18,9	-10,4	29,3
58	9,00	0,28	6,4	8,1	0,94	17,8	-9,3	27,0

Table S4.4 Nose scale CAT (Rasch).

Number of items	Median SEM	MAE	RMSE	Correlation	95% LA (upper boundary)	95% LA (lower boundary)	95% LA range
1	1,27	12,3	14,9	0,80	28,7	-29,5	58,3
2	0,98	8,7	10,9	0,90	21,0	-21,6	42,6
3	0,82	7,5	9,3	0,93	17,9	-18,5	36,4
4	0,72	6,7	8,4	0,94	16,2	-16,6	32,8
5	0,65	6,0	7,6	0,95	14,6	-15,0	29,6
6	0,60	5,5	6,9	0,96	13,4	-13,7	27,1
7	0,55	5,0	6,4	0,97	12,4	-12,9	25,2
8	0,53	4,6	5,9	0,97	11,5	-11,7	23,2
9	0,50	4,4	5,7	0,97	11,1	-11,3	22,4
10	0,47	4,2	5,5	0,98	10,8	-10,8	21,6
11	0,46	4,1	5,4	0,98	10,3	-10,7	21,0
12	0,44	3,9	5,2	0,98	10,2	-10,2	20,4

Table S4.5 Nose scale CAT (unidimensional GRM).

Number of items	Median SEM	MAE	RMSE	Correlation	95% LA (upper boundary)	95% LA (lower boundary)	95% LA range
1	0,48	11,4	14,3	0,82	28,1	-27,9	56,0
2	0,39	9,1	11,3	0,89	21,7	-22,4	44,1
3	0,32	7,5	9,4	0,93	17,8	-19,1	36,9
4	0,29	6,9	8,6	0,94	16,2	-17,5	33,7
5	0,26	6,2	7,9	0,95	14,5	-16,3	30,8
6	0,24	5,7	7,4	0,96	13,5	-15,2	28,7
7	0,22	5,4	7,0	0,96	12,8	-14,4	27,2
8	0,21	5,2	6,7	0,96	12,3	-13,9	26,2
9	0,20	4,9	6,4	0,97	11,6	-13,4	25,0
10	0,20	4,7	6,2	0,97	11,2	-12,8	24,0
11	0,19	4,5	6,0	0,97	10,9	-12,5	23,4
12	0,18	4,3	5,7	0,98	10,4	-11,9	22,3

Table S4.6 Nose scale CAT (multidimensional GRM).

Total item limit for all scales combined	Mean nose scale items	Median SEM	MAE	RMSE	Correlation	95% LA (upper boundary)	95% LA (lower boundary)	95% LA range
10	1,00	0,55	13,8	17,3	0,76	34,9	-32,9	67,8
20	2,84	0,41	8,4	10,6	0,90	20,7	-21,0	41,7
30	4,21	0,35	7,0	8,8	0,93	16,4	-18,1	34,4
40	5,92	0,31	6,2	7,6	0,95	13,9	-15,7	29,6
50	8,30	0,27	5,5	6,8	0,96	11,5	-14,6	26,1
58	12,00	0,25	4,8	6,1	0,97	11,1	-12,6	23,6

Table S4.7 Nostrils scale CAT (Rasch).

Number of items	Median SEM	MAE	RMSE	Correlation	95% LA (upper boundary)	95% LA (lower boundary)	95% LA range
1	1,47	11,9	14,7	0,84	28,3	-29,4	57,8
2	1,12	9,2	11,8	0,90	22,5	-23,6	46,2
3	0,94	7,9	10,0	0,93	19,1	-20,0	39,1
4	0,83	6,9	8,8	0,95	16,8	-17,6	34,4
5	0,75	6,1	7,9	0,96	15,1	-15,8	30,9
6	0,69	5,6	7,4	0,96	14,4	-14,5	28,9

Table S4.8 Nostrils scale CAT (unidimensional GRM).

Number of items	Median SEM	MAE	RMSE	Correlation	95% LA (upper boundary)	95% LA (lower boundary)	95% LA range
1	0,40	12,1	15,3	0,83	29,9	-29,9	59,8
2	0,31	9,6	12,3	0,90	24,3	-23,8	48,1
3	0,26	8,3	10,6	0,92	20,9	-20,7	41,6
4	0,23	7,4	9,6	0,94	18,7	-18,8	37,5
5	0,21	6,9	8,9	0,95	17,5	-17,3	34,9
6	0,20	6,4	8,4	0,96	16,6	-16,2	32,9

Table S4.9 Nostrils scale CAT (multidimensional GRM).

Total item limit for all scales combined	Mean nostrils scale items	Median SEM	MAE	RMSE	Correlation	95% LA (upper boundary)	95% LA (lower boundary)	95% LA range
10	1,84	0,28	9,7	12,5	0,90	24,3	-24,7	49,0
20	2,94	0,24	8,3	10,6	0,93	20,8	-20,6	41,5
30	4,16	0,21	7,9	10,2	0,95	18,6	-21,0	39,6
40	5,72	0,19	7,4	9,5	0,96	17,6	-19,4	37,0
50	5,98	0,19	7,8	10,1	0,96	18,6	-20,6	39,2
58	6,00	0,18	8,5	10,9	0,96	20,1	-22,5	42,5

Table S4.10 Lips scale CAT (Rasch).

Number of items	Median SEM	MAE	RMSE	Correlation	95% LA (upper boundary)	95% LA (lower boundary)	95% LA range
1	1,46	11,1	13,9	0,85	28,0	-26,6	54,6
2	1,11	8,8	10,9	0,91	22,2	-20,5	42,7
3	0,94	7,2	9,2	0,94	18,6	-17,3	36,0
4	0,83	6,4	8,1	0,95	16,1	-15,5	31,6
5	0,75	5,9	7,4	0,96	14,8	-14,2	28,9
6	0,70	5,4	6,8	0,97	13,6	-13,1	26,6
7	0,65	5,0	6,3	0,97	12,6	-12,3	24,9
8	0,62	4,7	6,0	0,97	12,0	-11,5	23,5
9	0,59	4,3	5,7	0,98	11,3	-11,0	22,3

Table S4.11 Lips scale CAT (unidimensional GRM).

Number of items	Median SEM	MAE	RMSE	Correlation	95% LA (upper boundary)	95% LA (lower boundary)	95% LA range
1	0,44	11,6	14,7	0,83	27,3	-30,1	57,4
2	0,34	9,2	11,7	0,90	20,9	-24,4	45,3
3	0,30	7,8	9,9	0,93	17,7	-20,8	38,5
4	0,27	7,1	9,0	0,95	15,7	-18,9	34,6
5	0,24	6,7	8,5	0,95	14,4	-18,1	32,6
6	0,23	6,4	8,1	0,96	13,4	-17,5	30,8
7	0,21	6,1	7,8	0,96	13,0	-16,6	29,6
8	0,20	5,7	7,3	0,97	12,2	-15,7	27,8
9	0,19	5,5	7,0	0,97	11,8	-15,0	26,9

Table S4.12 Lips scale CAT (multidimensional GRM).

Total item limit for all scales combined	Mean lips scale items	Median SEM	MAE	RMSE	Correlation	95% LA (upper boundary)	95% LA (lower boundary)	95% LA range
10	1,01	0,48	13,3	16,8	0,80	29,4	-35,5	64,9
20	2,47	0,37	9,4	11,9	0,90	20,5	-25,3	45,8
30	4,02	0,30	8,0	9,9	0,94	15,6	-21,5	37,2
40	5,67	0,27	6,5	8,2	0,95	15,7	-16,5	32,2
50	8,46	0,23	5,4	6,9	0,97	13,8	-13,3	27,2
58	9,00	0,23	5,3	6,8	0,97	13,7	-12,8	26,6

Table S4.13 Scar scale CAT (Rasch).

Number of items	Median SEM	MAE	RMSE	Correlation	95% LA (upper boundary)	95% LA (lower boundary)	95% LA range
1	1,48	12,3	15,0	0,87	29,6	-29,1	58,7
2	1,11	9,2	11,4	0,92	22,9	-22,0	44,8
3	0,93	7,7	9,7	0,95	19,6	-18,6	38,1
4	0,82	6,6	8,4	0,96	16,7	-16,2	32,8
5	0,74	5,8	7,5	0,97	14,9	-14,5	29,4
6	0,68	5,4	7,0	0,97	14,0	-13,6	27,5
7	0,63	4,9	6,7	0,97	13,2	-12,9	26,1

Table S4.14 Scar scale CAT (unidimensional GRM).

Number of items	Median SEM	MAE	RMSE	Correlation	95% LA (upper boundary)	95% LA (lower boundary)	95% LA range
1	0,37	12,9	16,1	0,84	29,4	-33,3	62,6
2	0,28	9,7	12,3	0,92	22,6	-25,3	47,9
3	0,23	8,3	10,8	0,94	19,8	-22,4	42,2
4	0,21	7,7	10,0	0,95	18,3	-20,7	39,0
5	0,19	7,0	9,4	0,96	16,8	-19,5	36,3
6	0,18	6,7	8,9	0,96	15,8	-18,7	34,5
7	0,17	6,5	8,7	0,96	15,4	-18,3	33,7

Table S4.15 Scar scale CAT (multidimensional GRM).

Total item limit for all scales combined	Mean scar scale items	Median SEM	MAE	RMSE	Correlation	95% LA (upper boundary)	95% LA (lower boundary)	95% LA range
10	1,94	0,23	10,3	13,1	0,91	21,3	-28,2	49,4
20	2,98	0,19	9,8	12,1	0,94	19,3	-26,1	45,4
30	4,71	0,16	8,9	10,9	0,96	18,5	-23,3	41,8
40	5,77	0,15	8,3	10,3	0,96	19,1	-21,1	40,2
50	6,84	0,14	8,6	10,6	0,97	19,8	-21,7	41,5
58	7,00	0,14	8,7	10,8	0,97	20,0	-22,0	42,0

Table S4.16 Teeth scale CAT (Rasch).

Number of items	Median SEM	MAE	RMSE	Correlation	95% LA (upper boundary)	95% LA (lower boundary)	95% LA range
1	1,23	12,2	15,0	0,81	29,6	-29,2	58,8
2	0,96	9,2	11,5	0,89	22,6	-22,4	45,1
3	0,83	7,8	9,7	0,93	18,9	-19,0	37,9
4	0,72	6,8	8,6	0,94	16,8	-17,0	33,9
5	0,66	6,0	7,6	0,95	15,1	-14,9	29,9
6	0,60	5,5	7,1	0,96	13,9	-13,9	27,8
7	0,56	5,2	6,8	0,96	13,4	-13,1	26,5
8	0,54	4,8	6,4	0,97	12,8	-12,5	25,3

Table S4.17 Teeth scale CAT (unidimensional GRM).

Number of items	Median SEM	MAE	RMSE	Correlation	95% LA (upper boundary)	95% LA (lower boundary)	95% LA range
1	0,43	12,3	15,0	0,81	27,9	-30,8	58,7
2	0,35	9,5	11,9	0,89	22,0	-24,2	46,3
3	0,31	8,1	10,0	0,92	18,5	-20,6	39,1
4	0,27	7,0	8,9	0,94	16,4	-18,3	34,7
5	0,25	6,5	8,3	0,95	15,5	-17,0	32,5
6	0,24	6,0	7,8	0,95	14,3	-16,0	30,3
7	0,23	5,7	7,4	0,96	13,5	-15,3	28,8
8	0,22	5,4	7,1	0,96	12,9	-14,8	27,7

Table S4.18 Teeth scale CAT (multidimensional GRM).

Total item limit for all scales combined	Mean teeth scale items	Median SEM	MAE	RMSE	Correlation	95% LA (upper boundary)	95% LA (lower boundary)	95% LA range
10	1,67	0,40	11,5	15,2	0,86	31,8	-27,3	59,1
20	3,00	0,28	8,3	10,5	0,92	15,5	-23,1	38,5
30	4,01	0,25	7,2	9,2	0,94	15,2	-19,7	34,9
40	5,00	0,24	6,9	8,8	0,94	14,8	-18,7	33,5
50	6,58	0,22	6,3	8,0	0,95	14,3	-16,8	31,0
58	8,00	0,21	5,8	7,5	0,96	15,7	-13,1	28,8

Table S4.19 Jaw scale CAT (Rasch).

Number of items	Median SEM	MAE	RMSE	Correlation	95% LA (upper boundary)	95% LA (lower boundary)	95% LA range
1	1,54	13,7	16,7	0,82	33,0	-32,6	65,6
2	1,18	10,5	12,9	0,90	25,5	-25,0	50,5
3	0,99	8,7	11,0	0,93	21,8	-21,2	43,1
4	0,88	7,5	9,6	0,95	19,2	-18,6	37,8
5	0,81	6,3	8,4	0,96	16,6	-16,5	33,1
6	0,75	5,7	7,8	0,97	15,3	-15,1	30,4
7	0,70	5,1	7,3	0,97	14,4	-14,3	28,7

Table S4.20 Jaw scale CAT (unidimensional GRM).

Number of items	Median SEM	MAE	RMSE	Correlation	95% LA (upper boundary)	95% LA (lower boundary)	95% LA range
1	0,42	13,4	16,7	0,84	27,2	-36,1	63,3
2	0,33	11,1	13,8	0,90	21,2	-30,0	51,2
3	0,27	9,5	11,9	0,93	18,2	-26,0	44,2
4	0,24	8,6	11,0	0,95	16,5	-24,1	40,5
5	0,22	8,1	10,3	0,96	15,3	-22,6	37,9
6	0,20	7,9	10,2	0,96	15,1	-22,3	37,4
7	0,19	7,6	9,9	0,96	14,3	-21,8	36,1

Table S4.21 Jaw scale CAT (multidimensional GRM).

Total item limit for all scales combined	Mean jaw scale items	Median SEM	MAE	RMSE	Correlation	95% LA (upper boundary)	95% LA (lower boundary)	95% LA range
10	1,54	0,37	12,6	16,3	0,86	23,3	-35,7	58,9
20	2,98	0,28	8,8	12,2	0,93	15,2	-26,8	42,0
30	4,71	0,23	7,9	10,4	0,95	12,5	-22,9	35,4
40	6,17	0,21	7,5	9,9	0,96	11,9	-21,7	33,7
50	6,99	0,20	7,5	9,7	0,96	12,0	-21,3	33,3
58	7,00	0,20	7,5	9,7	0,96	12,0	-21,3	33,3

Sheet 4.2 Monte Carlo simulation results (real patient data). All comparisons are made with unidimensional, linear Rasch assessment scores (transformed into 0-100 format). CAT: computerized adaptive test; SEM: standard error of measurement; MAE: mean absolute error; RMSE: root mean square error; correlation: Pearson's correlation coefficient; LA: limits of agreement.

Table S4.22 Face scale CAT (Rasch).

Number of items	Median SEM	MAE	RMSE	Correlation	95% LA (upper boundary)	95% LA (lower boundary)	95% LA range
1	1,14	10,1	12,6	0,81	26,4	-22,6	49,0
2	0,91	7,2	9,0	0,91	17,8	-17,4	35,2
3	0,78	5,6	6,9	0,95	13,3	-13,8	27,2
4	0,71	5,0	6,2	0,96	11,8	-12,3	24,1
5	0,64	3,8	4,9	0,97	8,5	-10,3	18,8
6	0,59	3,3	4,1	0,98	7,1	-8,6	15,7
7	0,56	2,4	3,2	0,99	5,4	-6,8	12,2
8	0,53	1,3	1,8	1,00	2,7	-3,9	6,6
9	0,51	0,0	0,0	1,00	0,0	0,0	0,0

Table S4.23 Face scale CAT (unidimensional GRM).

Number of items	Median SEM	MAE	RMSE	Correlation	95% LA (upper boundary)	95% LA (lower boundary)	95% LA range
1	0,57	10,1	13,0	0,79	26,7	-23,8	50,5
2	0,47	6,8	8,4	0,92	16,0	-17,0	33,0
3	0,39	5,7	7,2	0,94	13,9	-14,4	28,3
4	0,35	4,4	5,7	0,96	11,0	-11,3	22,3
5	0,32	3,6	4,7	0,97	8,8	-9,6	18,4
6	0,30	2,9	4,0	0,98	7,6	-8,0	15,6
7	0,28	2,2	3,2	0,99	5,8	-6,8	12,6
8	0,27	1,8	2,4	0,99	3,9	-5,3	9,2
9	0,27	1,3	1,9	1,00	2,8	-4,1	6,9

Table S4.24 Jaw scale CAT (Rasch).

Number of items	Median SEM	MAE	RMSE	Correlation	95% LA (upper boundary)	95% LA (lower boundary)	95% LA range
1	1,54	11,0	12,7	0,92	26,4	-22,9	49,3
2	1,18	7,4	9,0	0,96	18,3	-17,0	35,3
3	0,99	5,3	6,5	0,98	13,0	-12,4	25,4
4	0,88	3,6	4,7	0,99	9,7	-8,8	18,4
5	0,80	2,6	3,4	0,99	6,8	-6,6	13,4
6	0,74	1,6	2,2	1,00	4,3	-4,4	8,7
7	0,70	0,0	0,0	1,00	0,0	0,0	0,0

Table S4.25 Jaw scale CAT (unidimensional GRM).

Number of items	Median SEM	MAE	RMSE	Correlation	95% LA (upper boundary)	95% LA (lower boundary)	95% LA range
1	0,42	10,8	13,1	0,90	22,5	-27,7	50,2
2	0,32	7,7	9,0	0,96	13,8	-19,7	33,5
3	0,27	6,4	7,9	0,97	11,5	-17,4	28,9
4	0,24	5,6	6,9	0,99	8,6	-15,1	23,7
5	0,22	5,0	6,1	0,99	6,4	-13,2	19,7
6	0,20	4,7	5,7	0,99	4,9	-12,3	17,2
7	0,19	4,2	5,4	1,00	3,6	-11,3	14,9

Table S4.26 Teeth scale CAT (Rasch).

Number of items	Median SEM	MAE	RMSE	Correlation	95% LA (upper boundary)	95% LA (lower boundary)	95% LA range
1	1,23	10,6	13,1	0,82	21,8	-28,0	49,8
2	0,93	7,8	9,9	0,90	17,3	-21,0	38,2
3	0,80	5,7	7,2	0,95	13,1	-15,1	28,2
4	0,71	4,7	5,9	0,97	10,6	-12,2	22,8
5	0,63	3,5	4,4	0,98	8,3	-9,0	17,3
6	0,58	2,6	3,3	0,99	6,1	-6,8	12,9
7	0,54	1,5	2,1	1,00	4,3	-3,9	8,2
8	0,52	0,0	0,0	1,00	0,0	0,0	0,0

Table S4.27 Teeth scale CAT (unidimensional GRM).

Number of items	Median SEM	MAE	RMSE	Correlation	95% LA (upper boundary)	95% LA (lower boundary)	95% LA range
1	0,43	9,9	13,0	0,82	22,2	-27,6	49,7
2	0,34	7,8	9,8	0,91	15,3	-21,3	36,7
3	0,29	6,3	8,2	0,94	11,8	-17,9	29,7
4	0,26	5,6	7,0	0,96	9,0	-15,4	24,4
5	0,24	4,4	5,5	0,97	7,5	-12,1	19,7
6	0,23	3,5	4,4	0,98	5,9	-9,7	15,7
7	0,22	2,5	3,2	0,99	4,2	-7,2	11,4
8	0,22	2,1	2,8	0,99	2,8	-6,1	8,9

PART II

Implementation Challenges

Chapter 6

Barriers and Facilitators to the International Implementation of Standardized Outcome Measures in Clinical Cleft Practice

Apon I, MD, MHS[1]; Rogers-Vizena CR, MD[2]; Koudstaal MJ, MD, DMD, PhD[1]; Allori AC, MD, MPH, PhD[3]; Peterson P, MD[4]; Versnel SL, MD, PhD[5]; Ramirez JP, MPH[6]

[1] Department of Oral and Maxillofacial Surgery, Erasmus University Medical Center, Rotterdam, The Netherlands
[2] Department of Plastic and Oral Surgery, Boston Children's Hospital, Boston, Massachusetts, USA
[3] Division of Plastic, Maxillofacial, and Oral Surgery, Duke University Hospital, Durham, North-Carolina, USA
[4] Department of Plastic and Craniofacial Surgery, Karolinska University Hospital, Stockholm, Sweden
[5] Department of Plastic and Reconstructive Surgery, Erasmus University Medical Center, Rotterdam, The Netherlands
[6] International Consortium for Health Outcomes Measurement (ICHOM), Boston, Massachusetts, USA

Cleft Palate Craniofac J. 2022 Jan;59(1):5-13

Abstract

Objective: To identify barriers and facilitators to implementation of a prospective system for standardized outcomes measurement in cleft care.

Design: Cleft teams that have implemented the ICHOM Standard Set for cleft care were invited to participate in this two-part qualitative study: (1) an exploratory survey among clinicians, health information technology (HIT) professionals, and project coordinators, and (2) semi-structured interviews of project leads. Thematic content analysis was performed, with organization of themes according to the dimensions of the RE-AIM framework: reach, effectiveness, adoption, implementation, and maintenance.

Results: Four cleft teams in Europe and North America participated in this study. Thirteen participants completed exploratory questionnaires and five interviewees participated in follow-up interviews. Survey responses and thematic content analysis revealed common facilitators and barriers to implementation at all sites. Teams reach patients either via e-mail or during the clinic visit to capture patient-reported outcomes. Adopting routine data collection is enhanced by aligning priorities at the organizational and cleft team level. Streamlining workflows and developing an efficient data collection platform is necessary early on, followed by pilot testing or stepwise implementation. Regular meetings and financial resources are crucial for implementing, sustaining, analyzing collected data, and providing feedback to healthcare professionals and patients. Fostering patient-centered care was articulated as a positive outcome, whereas time presented challenges across all RE-AIM dimensions.

Conclusions: Identified themes can inform ongoing implementation efforts. Intentionally investing time to lay a sound foundation early on will benefit every phase of implementation and help overcome barriers such as lack of support or motivation.

Key words: implementation, value-based health care, RE-AIM framework, cleft lip and palate, patient-reported outcomes.

Introduction

The use of various disease-specific outcome measures to capture what truly matters to patients is of increasing importance in daily clinical practice. Outcome measures can be used to enhance patient-centered care and evaluate treatment effects.[1] To facilitate the measurement of cleft-specific outcomes in clinical practice, the International Consortium for Health Outcomes Measurement (ICHOM) convened a Working Group of cleft experts including clinicians from various specialties, patients and parents, and academicians to establish international consensus on the outcomes that should be measured routinely as a standard part of cleft care. Emphasis was placed on including clinical indicators for all relevant disciplines as well as patient-reported outcome measures (PROMs) to incorporate patient and parent perspectives. The result was a holistic, patient-centered Standard Set of measures and guidelines for prospective data collection over the course of care, from birth to young adulthood.[2,3] Within the Standard Set, satisfaction with appearance, speech function, psychosocial function, oral health, breathing, and eating and drinking are assessed by PROMs (CLEFT-Q scales, NOSE and COHIP-OSS questionnaires). Examples of clinical measures are tone-audiometry for the assessment of hearing, Percent Consonants Correct for speech assessment, and screening for velopharyngeal incompetence. Recommended time points for collection of these measures are 5 years (only clinical measures), 8 years, 12 years, and 22 years of age.[3]

The ICHOM Standard Set for the comprehensive appraisal of cleft care (hereafter, “Standard Set”) was designed for broad implementation, internationally and across cultures. Over the past four years, four cleft teams in North America and Europe have implemented Standard Set collection in their routine clinical practice, and implementation is ongoing at multiple other institutions. Collected outcome data are being used toward quality-improvement (QI) efforts, research, and inter-center collaborations to identify and disseminate “best practices”. These endeavors are of special importance in cleft care, since research has shown that treatment protocols and quality of care vary widely throughout the world.[4,5]

The work of these pilot sites is important; however, before meaningful outcome comparisons can be made, widespread adoption and implementation of the Standard Set by more cleft centers is needed. While many teams are keen to adopt the Standard Set, implementation is not easy. Many cleft teams are cautious about the myriad of challenges and obstacles that they will face. Factors hampering implementation efforts include lacking a defined strategy or a clear understanding of conditions that promote or hinder routine outcome measurement.[6]

Following their collaboration with ICHOM to develop the Standard Set, four cleft teams including Boston Children's Hospital, Duke Children's Hospital, Erasmus University Medical Center, and Karolinska University Hospital served as pilots for implementing the Standard Set in clinical practice.[7,8] Their experiences can help inform ongoing implementation endeavors of other cleft teams. The comprehensive evaluation framework RE-AIM is often used to evaluate implementation of an intervention or healthcare program focusing on five dimensions: reach, effectiveness, adoption, implementation, and maintenance.[9,10] The purpose of this investigation is to use the RE-AIM framework to identify facilitators and barriers for implementing the ICHOM Standard Set for cleft care in routine clinical practice, based on the experiences of four pilot centers.

Methods

This study was conducted in two phases, beginning with an exploratory survey followed by in-depth interviews to understand the different centers' experiences implementing the Standard Set. In this study, implementation was defined as "the continuous process of actively measuring, collecting, and analyzing outcomes according to the Standard Set in clinical practice". Participants were recruited through purposive sampling from the authors' personal networks, thus ensuring a diversity of stakeholders who could provide rich context and details regarding the implementation of the Standard Set by their teams. These stakeholders included clinicians, team coordinators, administrative personnel, IT professionals, project coordinators, and managers. Because the aim of this study was to provide an overview of facilitators and barriers from a healthcare provider's perspective, we decided not to recruit patients and families. The pilot sites invited to participate included Boston Children's Hospital (USA), Duke Children's Hospital (USA), Erasmus University Medical Center (The Netherlands), and Karolinska University Hospital (Sweden). Informed consent prior to the survey or interview was provided by all participants. This qualitative analysis of the facilitators and barriers to implementation was designated by the Institutional Review Board as exempt research (MEC-2020-0343).

Surveys and interviews

A preliminary exploratory survey was constructed based upon the dimensions present in the RE-AIM framework: reach, effectiveness, adoption, implementation, and maintenance (**Table 1** and **Supplemental Material - Appendix A)**. The comprehensive evaluation framework RE-AIM is often used to evaluate implementation of an intervention or health care program.[9,10] Open-ended questions allowed each participant to expound on the implementation process and corresponding facilitators and barriers. The survey also included questions regarding numerical data such as response rates for the dimensions of reach and adoption. The survey was sent via email to all eligible participants followed by two reminders at biweekly intervals. Data collection for the exploratory survey took place between March 2, 2020 and April 6, 2020. Responses were transcoded according to overarching themes and tallied to discover what participants deemed the most important facilitators and barriers. Survey responses were described in frequencies of verbalization (n). Because survey respondents were able to name multiple factors in one answer, the total number of verbalizations could outraise the number of participants. The responses were used to elucidate relevant topics to include in subsequent in-depth interviews.

Following the completion of their exploratory surveys, the cleft team leaders or coordinators from each site were invited for in-depth, semi-structured interviews to further explore various dimensions of implementation. Two researchers (I.A. and J.P.R.) conducted the interviews between April 3, 2020 and April 8, 2020. The researchers performing the interviews were not involved in the implementation process.

An interview guide (**Supplemental Material - Appendix B**) ensured the same questions were asked uniformly of all interviewees, but interviewees were allowed to follow their train of thought and bring up any issues that came to mind. Interviews were conducted in English, and all interviewers and interviewees were fluent in English; however, since the native language of some participants was different from English, they were offered the opportunity to add specific words or sentences in their own language to more accurately express feelings and perspectives. If needed, these parts could be separately translated by two additional objective researchers (Dutch and Swedish native speakers) with a good understanding of the English language.

Interviews were audio-recorded and transcribed verbatim in English using NVivo 12 Pro Software for Windows.[11] Thematic content analysis was performed by a main coder (I.A.), then reviewed by a second coder (J.P.R) who checked that transcripts were accurate and appropriately coded, and that no sections were missed during analysis. All coded themes were then grouped according to the RE-AIM dimensions, and appropriately sub-coded.

Dimension	Original definition (Glasgow et al.[9])	Definition in current study
Reach	Proportion of the target population that participated in the program	Methods to reach participants (e.g. patients and parents)
Effectiveness	Outcome effects of implementing the program as planned	Positive and negative effects of the implementation of the Standard Set in clinical practice
Adoption	Proportion of practices and individuals that adopted the program	Facilitators and barriers to reach adoption of the Standard Set among individuals involved in cleft care (e.g. clinicians, organization, leadership, patients)
Implementation	Extent to which the program is implemented as intended	Facilitators and barriers in implementing the Standard Set in clinical cleft care as planned
Maintenance	Extent to which the program is sustained over time	Activities executed to sustain the (implementation of the) Standard Set over time

Table 1 Original definitions of the RE-AIM framework dimensions adapted for this study from Glasgow *et al*.[9]

Results

Twenty participants were invited to complete the exploratory survey; 15 from Erasmus University Medical Center, 1 from Duke Children's Hospital, 1 from Boston Children's Hospital, and 3 from Karolinska University Hospital. Completion rate was 65% (*n = 13*). Five respondents were eligible for in-depth interviews. Every pilot center was represented by at least one interviewee, and one interviewee provided feedback on behalf of two centers, since he has been the implementation lead at both centers at different points in time. Interview duration ranged between 47 and 122 minutes. Survey respondent and interviewee characteristics are described in **Table 2**.

Characteristics	**Survey respondents** Count (%) Total *n = 13*	**Interviewees** Count Total *n = 5*
Sex		
Male	5 (38)	2
Female	8 (62)	3
Age		
30-39	3 (23)	0
40-49	7 (54)	5
50-59	3 (23)	0
Institution		
Erasmus University Medical Center	10 (76)	2 *
Boston Children's Hospital	1 (8)	1
Duke Children's Hospital	1 (8)	1
Karolinska University Hospital	1 (8)	2 *
Main job function		
Clinician	11 (84)	5
Surgeon	6 (46)	5
Other	5 (38)	0
HIT	1 (8)	0
Management	1 (8)	0
	Mean (range)	Mean (range)
Years of working experience in cleft care	8.7 (0-19)	10.2 (7.5-12.5)

Table 2 Survey respondent and interviewee characteristics. HIT = health information technology. Interviewees representing multiple institutions are indicated by *.

Findings from the survey and in-depth interviews are discussed per RE-AIM dimension below (**Table 3**).

	Methods	**Response rates**
Reach	Pen and paper	Labor-intensive, not utilized in the included institutions
	Electronically via clinic	Response rate 85 to 99%
	Electronically via e-mail	Response rate 75 to 85%
	Positive outcomes	**Negative outcomes**
Effectiveness	Patient connection	Time
	Teambuilding	
	Awareness parents and patients	
	Focus for discussion	
	Themes	
Adoption	Creating importance and urgency	
	Aligning motivation and priorities through regular meetings	
	Securing resources	
Implementation	Reorganizing the clinical workflows	
	Developing an efficient HIT-system	
	Pilot testing and stepwise implementation	
Maintenance	Analyzing and utilizing collected data	

Table 3 Overview of themes and most important findings per RE-AIM dimension.

Reach

To engage patients in providing PROMs, three different approaches were used. One center started by sending paper questionnaires with appointment letters to patients' homes. Due to the amount of work (mailing questionnaires, sorting them, entering data in a digital system, storing paper forms), they switched to inviting patients to complete questionnaires on an iPad while waiting for their clinic appointment. Teams using the in-clinic iPad approach reached response rates of 85 to 99%. However, interviewees articulated that it was sometimes noted by clinicians that some patients and parents felt uncomfortable thinking about their appearance while surrounded by others in a waiting room. These concerns made one team change to a third approach of sending questionnaires, including information on how answers will be used for clinical care, by e-mail a few days before the clinic visit so patients could answer in a quiet, private environment. Teams using the latter approach reached 75 to 85% of patients; some could not be reached due to incorrect or missing e-mail addresses, encountered most often for 22-year-olds (as a result of moving from their family home, large time gap since last visit, and switching to their own e-mail address, from that of their parents).

The latter was verbalized 8 times by survey respondents as a barrier in reaching patients for PROM collection. Notably, interviewees mentioned that some patients and parents shared negative reactions about unsolicited e-mails with the team members. At the end of implementation, two teams used e-mailed invitations to complete PROMs at home, and two used an iPad to complete PROMs during the clinic visit.

Effectiveness

Interviewees mentioned the ultimate goal of comparing outcomes is not yet possible as individual centers need to reach more robust levels of data first. However, other effects of implementing the Standard Set in routine clinical practice became visible.

Positive effects

Survey respondents most frequently answered that the ability to quickly assess patient's well-being (n=8) is the main positive outcome of using the Standard Set. Interviewees and survey respondents (n=5) added that using PROMs enables them to plan ahead of the clinic visit (n=5) and provides a launching point for more focused and intentional discussions (n=5). As a result, interviewees felt that using PROMs routinely fosters connection between patient and team. Additionally, interviewees mentioned that the use of PROMs gives the parents an opportunity to better prepare for the visit together with their child.

Two illustrative quotes about the positive effects of using the Standard Set are below:

Interviewee # 2: "So, I think it [use of PROMs] is great…for making the parents…more aware of what the concerns are that the children might have, and it makes our work much easier because we can focus on those [concerns] and not miss out on them."

Interviewee # 3: "The psychologists say, 'Why haven't we done this [collecting outcome data] before? This is so useful and we're now reaching families, and parents who are struggling, and kids who are struggling, and we never asked these questions [CLEFT-Q psychosocial scales], they never raised it until it was too late.' So, I think there is a true benefit of just using the set."

Furthermore, interviewees reported that introduction of the Standard Set has helped foster team solidarity by generating a common goal and giving the team an opportunity to self-evaluate.

Negative effects

Survey respondents listed time (n=7) and extra work (n=4) as negative effects of using the Standard Set. Interviewees were more nuanced about these limiting factors:

Interviewee # 4: "In the beginning we had some people argue that [collecting outcomes] costs a lot of extra time but once you have everything up and running, and you're used to it, ..., it really fits. You find a way that it fits in the workflow of your team, we don't experience it as a burden or extra administration or those kinds of things."

Adoption

Survey respondents emphasized that hospital leadership (n=3) and cleft team coordinators (n=3) are crucial stakeholders in successful adoption of the Standard Set. Motivation (n=4) was most frequently mentioned by survey respondents as a facilitator for adoption, and time as a barrier (n=4). Three themes were identified.

Theme 1: Creating importance and urgency

The hospital boards of all four centers were supportive of the initiative, and interviewees felt that lack of leadership support would hinder widespread implementation. To enhance adoption, interviewees advised teams to get on the hospital board's agenda and explain the value of implementing the Standard Set, for example to improve quality of care by having your own local outcome registry or positioning cleft teams to benchmark (inter) nationally. An additional advice of interviewees was to use cases from the literature and the experiences of pilot institutions to support this process. Interviewees also advised starting with a simplified implementation collecting only specific parts of the Standard Set, to show the feasibility, benefit, and value in expanding data collection. Demonstrating importance was not only found useful to garner commitment from leadership but also to convince other members of the cleft team, another key stakeholder, to adopt the Standard Set:

Interviewee # 5: "I think the main person who we're really talking about is the main team director, but it could also be the chair of a department or something like that. In any case, that person needs to convey to the team that this [measuring outcomes] is important, that this is creating a new sense of normal ..., a new standard operating procedure. That this is not really voluntary, but this is what we as a team want to do, it fulfills our mission... So, you have to create a sense of urgency."

Informing patients and parents about the importance of the project varied by institution. When data collection was wrapped into a broader research program, patients underwent informed consent at the beginning of their clinic visit. If collection of Standard Set data was integrated into routine clinical practice for the purpose of quality improvement (QI), this advancement was announced to patients through newsletters, informational meetings, and on cleft team and/or scientific society websites.

An interviewee articulated how prioritizing patient engagement in decision-making increased adoption:

Interviewee # 3: "I think the fact that we ask questions from them [patients] and that we do something with these questions, increases the connection between the patient and the team, knowing that we look into it, that we care, that we listen to what they're saying, and try to do something with it."

Theme 2: Aligning motivation and priorities through regular meetings

Interviewees reported that due to the multidisciplinary nature of cleft care, it is essential to ensure every specialty buys into measuring outcomes routinely. In addition, interviewees stated that interdisciplinary friction points should be discussed and incentives stated clearly, so the project will not be jeopardized later on because of competing priorities. All four participating centers held regular meetings to discuss feelings, visions, thoughts, challenges, and organizational matters regarding implementation of the Standard Set to keep everyone engaged. The most frequently discussed topics were how to organize different data collection workflows in clinical practice, what impact PROM questions might have on the child and how to deal with the answers, and what will ultimately happen with the data. Regular meetings also provided opportunities to build an overarching implementation strategy:

Interviewee # 4: "They [cleft team members] were all taken along with what we [the implementation team] would do. We had regular meetings, to discuss what was the plan, what would be the next step, and everyone could have a say in that, what they thought about it. Then we did something and had a new meeting or evaluation. So, it was done as a team."

Theme 3: Securing resources

Interviewees articulated that teams that want to implement the Standard Set but lack financial resources and time will face barriers implementing health information technology (HIT) solutions and will more likely succeed by starting outcomes collection on paper. Two of the four hospitals partially financed their implementation projects through grants. Interviewees mentioned that Duke and Erasmus are now providing open-access platforms in collaborative networks, to decrease the startup time for teams wanting to adopt and implement the Standard Set.

Implementation

The Standard Set was implemented as planned at all four centers, but the initial implementation period was longer than anticipated (ranging between 6-24 months). Survey respondents most frequently (n=7) answered that approximately 10 to 15 people were involved in the implementation team. One respondent reported a number of over 40 people. Interviewees articulated that a small core implementation team was preferred over a larger group because communication problems and staff turnover could disrupt the process. Crucial members of the implementation team included the clinical lead (n=8), a HIT lead (n=5), and a clinic coordinator (n=9) or specialized nurse (n=5). These members did not differ by teams. The participants felt that the implementation lead could come from any specialty, as long as they are enthusiastic and dedicated, familiar with workflows, and able to build good relationships. Representatives of every specialty could be invited to the team and HIT personnel were mostly included by consultation. Three unique themes were identified.

Theme 1: Reorganizing the clinical workflows

Interviewees frequently mentioned that evaluating and transforming workflows and clinical visits are important aspects of the implementation phase. Teams started by evaluating how data collection would best fit their current workflow, ensuring all outcomes are collected. The four centers already worked as a multi- or interdisciplinary team, making it easier for them to streamline workflows of the various specialties involved. Developing flowcharts of treatment protocols including designated Standard Set outcome time points and measurements was very useful to gain insights on how to seamlessly integrate data collection into the existing workflow. Awareness of the extra time needed for speech and language therapists to perform additional testing, and of possible increase in patient volume for the psychologist was necessary. Furthermore, assessing patient’s answers, providing feedback to them, and recording clinical outcomes resulted in an additional 5 minutes on average per clinical visit per patient.

Three teams reported that each specialty records their own clinical outcome measures in their HIT-system for tracking outcomes, which interfaces directly with the patient's electronic medical record (EMR) in two of the four teams. One team mentioned having a dedicated person who collects all outcomes from the clinicians in a standardized form, and then registers them in the system. At all four teams, after the completion of PROMs by the patient, both at home as in clinic, the answers were directly stored in the HIT-system without the intervention of a person. Scoring algorithms for each PROM were programmed within the HIT-systems, and access to both PROMs and clinical outcomes was the same.

Interviewee # 1: "I don't think one [way of collecting data] is right or wrong, but there are some pros and cons to each. … By doing it in a specialty-specific way, you guarantee that the data quality is pretty good … The downside is that in many cases you might get incomplete data because people forget or they get busy in clinic, whereas the benefit of having a research person who is … always available is always making sure data is collected. … The downside is if they don't have a clinical background, you might have some incorrect data in there."

Theme 2: Developing an efficient HIT-system

All five interviewees and 7 survey respondents agreed that a HIT platform was an important facilitator that will save time and increase ease of data management while reducing risk of data loss as compared to tracking outcomes using pen and paper. The most frequently mentioned system requirements were easy access, allowing concurrent users of the database system, dealing with versioning, and keeping permanent records of changes made, with easy data extraction for use in QI projects. There were no teams that had HIT systems that automatically extracted outcomes from the EMR. Interviewees advised making the HIT-system as compatible as possible with other systems to aid in future data exchange. Furthermore, interviewees found it helpful to get advice from someone who has dealt with this process before to prevent mistakes that can later create barriers.

Theme 3: Pilot testing and stepwise implementation

One team started with pilot testing the complete Standard Set for 3 to 4 patients with different ages and cleft diagnoses per clinic day. This enabled them to explore time requirements per visit and gave them the opportunity to adjust workflows accordingly to solve errors early on, before measuring outcomes for all patients. Another team started with the complete set, but scaled implementation up from one patient per week to all patients to ease into it, improving the process of data collection gradually, and working out friction points. The other two teams preferred stepwise implementation, starting with implementing PROMs followed by clinical measures. This allowed them to spend more time developing their HIT-system.

Interviewee # 3: "We decided to go for a pilot phase. I know that different hospitals in the world have chosen different routes, so some have said 'Okay, we are just going to do only the 5-year-olds', for example, or 'We are only going to do the cleft lips for a while'. That's one approach, a choice you need to make.... We then said, 'We are going to do the whole set, we want to have all the patients from the beginning'. So, the HIT-system was built for all diagnoses and for all aspects of the set."

Maintenance

Theme 1: Analyzing and utilizing collected data

In order to maintain momentum, most survey respondents and all interviewees felt that it is important to analyze and use locally collected data early in the process (n=9). For example, QI projects like analyzing data completion or complication rates facilitated opportunities for improvement and sustain motivation. Also, interviewees articulated that research on outcomes data and measurement instruments can provide insights to improve future iterations of the Standard Set. Most importantly, it was found that sharing early wins with the entire team is a good way to maintain engagement, since decreasing commitment levels over time was recognized as a barrier.

Discussion

This study applied qualitative methods and the RE-AIM framework in the evaluation of facilitators and barriers to implementation of the ICHOM Standard Set for the comprehensive appraisal of cleft care. Major themes identified included creating importance and urgency, aligning motivation and priorities through regular meetings, and securing resources. The dimension of implementation was characterized by reorganizing clinical workflows and developing efficient HIT-systems, followed by pilot testing and stepwise implementation. While implementing the Standard Set requires extra time and effort, especially in the beginning, interviewees experienced advancements in patient-centered care as a positive outcome. Analyzing and utilizing the data collected in practice could help sustain implementation over time.

Three methods were identified to reach patients and collect PROM data: paper surveys mailed to the patient's home; e-mail surveys prior to clinic visit; and real-time data collection using an iPad during clinical visits. Only the two electronic approaches are now used by the pilot centers because paper forms were too labor-intensive with higher risk of losing information. Several studies have shown that patients are more receptive towards electronic collection systems compared to pen and paper for the collection of PROMs; however, no clear comparisons have been made between completing surveys at home or during the clinical visit and how this is viewed by the pediatric patient population.[12-15] Cultural or societal differences might play a role, since the North-American institutions chose the in-clinic iPad approach, while the European centers incorporated e-mailed invites to complete PROMs at home. Other factors that could influence choice of data collection method is the payment model of the health care system and a patient's travel time to the clinic. Patients will not come to clinic when they do not experience problems or concerns if they have to pay extra for each visit or travel long distances. Missing out on collecting data for these patients could potentially jeopardize a center's outcomes. These factors should be taken into consideration when deciding on the best way to reach patients. Including patient advocacy groups in this decision could be valuable.

Unfortunately, investing in electronic systems for data collection might not always be feasible, due to limited financial or technological resources, or differing organizational priorities.[6,16,17] Middle- and low-income countries might especially face these challenges. Currently, two large initiatives offer support in these circumstances. The European Reference Network for rare and/or complex craniofacial anomalies and ear-nose-throat disorders (ERN CRANIO) aims to pool disease-specific expertise, knowledge, and resources from across Europe to improve quality of care. The network is currently developing a registry for the collection of outcome measurement data for cleft care. This registry will be accessible to all participating centers for the primary purpose of quality control, and outcomes research in the future.[18] Similarly, the ACCQUIREnet collaborative, led by Duke University, makes its REDcap-based implementation available to member institutions that join the network.

Creating importance, and aligning motivation and priorities among team members and leadership is a crucial and universal part of implementing an outcomes measurement framework in clinical practice. This is consistent with recent literature on understanding and overcoming barriers to change, which states that it is important for health care professionals to understand the benefits of changing practices.[19] Across various health care settings, implementation was boosted when collection of outcome data is supportive of patient-centered care at an individual patient level, instead of at an aggregated level.[20-22]

The current study identified a common belief among cleft professionals that implementation of the Standard Set had a positive effect on their team and on patient-centered care. Previous literature reported that patient-clinician communication, clinician's awareness of symptoms, and patient satisfaction can be improved by the use of PROMs, and by reviewing the results with the patient.[15,23] In addition, a recent study showed that over 80% of children completing the CLEFT-Q scales, representing 9 of the 12 PROMs in the Set, liked answering the questions, and felt it made them understand their condition and feelings better.[24] The fact that the children get something in return (insight in their own well-being, more individualized care) could be a reason for obtaining relatively high response rates in contrast to the reported email survey response rates of 20-40% among adults in literature.[25-27]

Implementation efforts were most constrained by time. Time, as part of resources, was articulated to have an overarching and continuing influence on all dimensions of the RE-AIM framework, especially on adoption and implementation. In general, approximately five extra minutes per patient were necessary during clinical visits for the discussion of the PROM results with the patient, registration of clinical outcome data, and in some cases extra speech or audiometry screening. For the latter, planning extra time for speech therapists and audiological consultants might be necessary and coordination with the specific departments regarding other obligations is of considerable importance when implementing the Standard Set. When barriers are not properly addressed due to time constraints, teams might struggle with problems later on, experience setbacks or jeopardize the project. Therefore, intentionally investing time to set the parameters for implementation will benefit every phase and help overcome barriers such as lack of support or motivation.

Limitations and future directions

A strength of this study is inclusion of four cleft centers with different implementation methods from various countries, representing unique cultures and societal habits. However, all four centers are located in high-income countries, limiting the generalizability of these findings to low- and middle- income countries.[28] It is likely that factors influencing change management will not differ profoundly, while differences in financial and technological resources will be more prominent.

Another important factor limiting the generalizability and interpretability of our findings, is the fact that there was a sizeable disparity in the number of people per cleft team approached for participation in this study, and that all interviewees represented one discipline, instead of a variety in stakeholders. The first disbalance is caused by a high turnover of personnel involved in the clinical implementation of the Standard Set, resulting in a limited number of eligible patients at three sites for completing the exploratory survey. The loss of continuity in personnel was mentioned by these sites as a barrier in implementation resulting in slowing down the process. The second disbalance of interviewing only surgeons, is caused by the fact that the implementation efforts were all led by surgeons as project coordinators, and because a relatively large proportion of clinicians within a cleft team has a surgical expertise. Also, healthcare management and coordinating tasks are often employed by clinicians, since they are familiar with the clinical workflows.

Furthermore, centers who are currently implementing or have abandoned implementation due to problems, were outside the scope of this study. Anecdotally, some of these centers experienced a lack of institutional and financial support. The findings of this study can help teams experiencing challenges in their implementation efforts to move forward, as well as serve as starting point for future research by centers struggling with implementation, and by centers in low- and middle-income countries.

Using an extensive open-ended survey as well as the fact that experts were recruited through purposive sampling could have influenced answers, because participants could assume specific information or opinions are already common knowledge for the researchers. The use of open-ended questions was chosen to gather as many different opinions and feelings as possible, since a qualitative study towards implementation of such a specific outcomes set has not yet been performed. Therefore, it was deemed a preliminary exploratory survey was necessary to explore the main themes and directions for the interviews. A possible lack of in-depth information on the survey was addressed by the follow-up semi-structured interviews with clinical leads and coordinators.

Conclusion

The themes identified in this qualitative study may be helpful to other cleft teams that are considering adopting and implementing the Standard Set. Specifically, each team should strive to adequately communicate to all stakeholders the reason for adopting the standard set, seek to align motivation and priorities, and provide frequent communication during the initial phases of implementation. At the organizational level, proper attention must be given to setting up the HIT-platform, the implementation effects on workflow and provider burden, and securing resources for sustaining the endeavor. Multi-site collaboratives may assist in facilitating implementation.

Declaration of conflicting interests

The authors declare that there is no conflict of interest.

References

1. Desomer A, Van den Heede K, Triemstra M, et al. Het gebruik van patiëntuitkomsten en -ervaringen (PROMs/PREMs) voor klinische en beleidsdoeleinden – Synthese. In: Health Services Research (HSR) Brussel: Belgian Health Care Knowledge Centre (KCE); 2018.
2. Allori AC, Kelley T, Meara JG, et al. A Standard Set of Outcome Measures for the Comprehensive Appraisal of Cleft Care. *Cleft Palate Craniofac J.* 2017;54(5):540-554.
3. International Consortium of Health Outcomes Measurement (ICHOM). Data collection reference guide. https://ichom.org/files/medical-conditions/cleft-lip-palate/cleft-lip-palate-reference-guide.pdf. Published 2020. Accessed July 1, 2020.
4. Russell K, Long RE, Jr., Hathaway R, et al. The Americleft study: an inter-center study of treatment outcomes for patients with unilateral cleft lip and palate part 5. General discussion and conclusions. *Cleft Palate Craniofac J.* 2011;48(3):265-270.
5. Shaw WC, Semb G, Nelson P, et al. The Eurocleft project 1996-2000: overview. *J Craniomaxillofac Surg.* 2001;29(3):131-140; discussion 141-132.
6. Foster A, Croot L, Brazier J, Harris J, O'Cathain A. The facilitators and barriers to implementing patient reported outcome measures in organisations delivering health related services: a systematic review of reviews. *J Patient Rep Outcomes.* 2018;2:46.
7. Arora JH, M. Implementing ICHOM's Standard Sets of Outcomes: Cleft Lip and Palate at Erasmus University Medical Centre in the Netherlands. *London, UK: International Consortium for Health Outcomes Measurement (ICHOM) (available at www.ichom.org).* 2016.
8. Bittar PG, Carlson AR, Mabie-DeRuyter A, Marcus JR, Allori AC. Implementation of a standardized data-collection system for comprehensive appraisal of cleft care. *Cleft Palate Craniofac J.* 2018;55(10):1382-1390.
9. Glasgow RE, Harden SM, Gaglio B, et al. RE-AIM Planning and Evaluation Framework: Adapting to New Science and Practice With a 20-Year Review. *Front Public Health.* 2019;7:64.
10. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. *Am J Public Health.* 1999;89(9):1322-1327.
11. QSR International. NVivo qualitative data analysis software, Version 12 Pro. https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home. Published 2020. Accessed.
12. Richter JG, Becker A, Koch T, et al. Self-assessments of patients via Tablet PC in routine patient care: comparison with standardised paper questionnaires. *Ann Rheum Dis.* 2008;67(12):1739-1741.
13. Salaffi F, Di Carlo M, Carotti M, Farah S, Gutierrez M. The Psoriatic Arthritis Impact of Disease 12-item questionnaire: equivalence, reliability, validity, and feasibility of the touch-screen administration versus the paper-and-pencil version. *Ther Clin Risk Manag.* 2016;12:631-642.
14. Salaffi F, Gasparini S, Ciapetti A, Gutierrez M, Grassi W. Usability of an innovative and interactive electronic system for collection of patient-reported data in axial spondyloarthritis: comparison with the traditional paper-administered format. *Rheumatology (Oxford).* 2013;52(11):2062-2070.
15. Recinos PF, Dunphy CJ, Thompson N, Schuschu J, Urchek JL, 3rd, Katzan IL. Patient Satisfaction with Collection of Patient-Reported Outcome Measures in Routine Care. *Adv Ther.* 2017;34(2):452-465.
16. Bausewein C, Simon ST, Benalia H, et al. Implementing patient reported outcome measures (PROMs) in palliative care--users' cry for help. *Health Qual Life Outcomes.* 2011;9:27.

17. Malhotra K, Buraimoh O, Thornton J, Cullen N, Singh D, Goldberg AJ. Electronic capture of patient-reported and clinician-reported outcome measures in an elective orthopaedic setting: a retrospective cohort analysis. *BMJ Open.* 2016;6(6):e011975.
18. ERN CRANIO. Network activities. https://ern-cranio.eu/network-activities/. Published 2020. Accessed June 29th, 2020.
19. National Centre for Health Excellence. How to change practice. https://www.nice.org.uk/media/default/about/what-we-do/into-practice/support-for-service-improvement-and-audit/how-to-change-practice-barriers-to-change.pdf. Published 2007. Accessed.
20. Boyce MB, Browne JP, Greenhalgh J. The experiences of professionals with using information from patient-reported outcome measures to improve the quality of healthcare: a systematic review of qualitative research. *BMJ Qual Saf.* 2014;23(6):508-518.
21. Greenhalgh J, Pawson R, Wright J, et al. Functionality and feedback: a protocol for a realist synthesis of the collation, interpretation and utilisation of PROMs data to improve patient care. *BMJ Open.* 2014;4(7):e005601.
22. Howell D, Molloy S, Wilkinson K, et al. Patient-reported outcomes in routine cancer clinical practice: a scoping review of use, impact on health outcomes, and implementation factors. *Ann Oncol.* 2015;26(9):1846-1858.
23. Basch E, Barbera L, Kerrigan CL, Velikova G. Implementation of Patient-Reported Outcomes in Routine Medical Care. *Am Soc Clin Oncol Educ Book.* 2018;38:122-134.
24. Klassen AF, Dalton L, Goodacre TEE, et al. Impact of Completing CLEFT-Q Scales That Ask About Appearance on Children and Young Adults: An International Study. *Cleft Palate Craniofac J.* 2020;57(7):840-848.
25. Fowler FJ, Jr., Cosenza C, Cripps LA, Edgman-Levitan S, Cleary PD. The effect of administration mode on CAHPS survey response rates and results: A comparison of mail and web-based approaches. *Health Serv Res.* 2019;54(3):714-721.
26. Rodriguez HP, von Glahn T, Rogers WH, Chang H, Fanjiang G, Safran DG. Evaluating patients' experiences with individual physicians: a randomized trial of mail, internet, and interactive voice response telephone administration of surveys. *Med Care.* 2006;44(2):167-174.
27. Toomey SL, Elliott MN, Zaslavsky AM, et al. Improving Response Rates and Representation of Hard-to-Reach Groups in Family Experience Surveys. *Acad Pediatr.* 2019;19(4):446-453.
28. The World Bank. World Bank Country and Lending Groups. https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups. Published 2020. Accessed August 1st, 2020.

Supplemental Material

Appendix A - Exploratory Survey

Individual and organizational information

What is your name?

What is your age?

What is the best email address to reach you?

What is the name of your institution / hospital?

What is your job function?

How many years have you been employed by the hospital where implementation of the CL/P Standard Set is taking place?

How many years of working experience do you have in cleft care?

What is the size of your CL/P team (please include all clinical, non-clinical, and administrative professionals)? *1 – 5, 6 – 10, 11 – 15, 16 – 20, 21 – 25, 26 – 30, 31 or more.*

How many patients does your clinic or center serve on average each year? *Less than 50, 51 – 100, 101 – 150, 151 – 200, 201 – 250, 251 – 300, more than 300.*

Program information

What is the stage of the implementation process right now? *Complete; fully incorporated in clinical practice, almost complete; not yet completely incorporated in clinical practice, implementation is going on, in the starting phase of implementing, not started, other, please specify.*

What was / is your role in the implementation process?

When did implementation begin?

Could you describe how the ICHOM CL/P Standard Set is used in your hospital right now?

Reach

Employees, hospital and team members

How many people were involved in implementing the CL/P Standard Set at your institution?

Which members do you consider key in implementing the CL/P Standard Set?

What were the most important facilitators in generating commitment from these stakeholders?

What were the most important barriers in generating commitment from these stakeholders?

When the implementation process started, were you in agreement with implementing the CL/P Standard Set? If yes, please explain why and how agreement was reached. If not, please explain why not and what should have been done to get your agreement.

Patients

How many patients have you reached with the CL/P Standard Set (average response rate)? *0-25%, 26-50%, 51-75%, 76-100%.*
How do you reach your patients?
What went well or what do you consider important facilitators in reaching patients?
What went less well or do you consider important barriers in reaching patients?

Effectiveness

What do you consider as positive effects or outcomes of implementing the ICHOM CL/P Standard Set?
What do you consider as negative effects or outcomes of implementing the ICHOM CL/P Standard Set?
What do you expect the future impact of implementing the ICHOM CL/P Standard Set will be?

Adoption

Which people do you consider key stakeholders to facilitate the adoption of the program?
What are the most important facilitators for these collaborations? Please specify per collaboration.
What are the most important barriers for these collaborations? Please specify per collaboration.
What percentage of your team (established in question 8 under Individual and Organizational Information) is actively using the CL/P Standard Set? *0-25%, 26-50%, 51-75%, 76-100%.*

Implementation

Could you describe the implementation plan?
Did you implement the program based on other examples of CL/P teams, ICHOM or other partner organizations doing similar work? Please clarify.
If implementation has been completed; how long did it take to implement?
What were the most important facilitators for the implementation process?
What were the most important barriers for the implementation process?

Maintenance

Which actions have been taken to ensure the local CL/P Standard Set remains current?
Within your team, how do you capture learnings and feedback from implementing the CL/P Standard Set?
What are the most important facilitators to maintain implementation over time?
What are the most important barriers to maintain implementation over time?

Other

Do you have any other information that you find relevant for this research which has not been addressed by previous questions?

Appendix B – Interview guide

Introduction

What is your motivation for sharing your experience with others who are wanting to implement the CL/P Standard Set?
Can you describe your day to day role in your cleft center?
If your role changed as implementation of the CL/P Standard Set progressed, could you please describe how and why?
How would you compare a typical patient visit before and after full implementation of the CL/P Standard Set?
Did any of these changes spread to the rest of your organization?

General facilitators and barriers

Please describe how support from leadership influenced any of the RE-AIM components?
What would you advise to teams who are wanting to implement the CL/P Standard Set but lack commitment from their organization and/or leadership?
Please describe how IT resources influenced any of the RE-AIM components?
Why did you choose your specific system for data collection over another? If the system changed along the way, what changed and why?
The CL/P Standard Set does not specify who should collect which data. What would be your overarching recommendation or rule of thumb about who collects the data?
Please describe how time influenced any of the RE-AIM components during your implementation experience?
Please describe how finances influenced any of the RE-AIM components during your implementation experience?
What advice would you give teams wanting to implement the CL/P Standard Set, regarding the influence of time and financing on implementation efforts?

Reach

Who do you consider part of the cleft team? And who do you consider part of the implementation team (roles)?

Which crucial roles would you recommend cleft centers to start out with if they can't have a full team?

Can you describe the types of resistance you or the team faced during implementation, and how you overcame them?

Can you give an example of ways in which you motivated the team?

How did you inform patients about the "transformation" that was happening?

What methods worked well to keep patients committed to their care?

Effectiveness

At your cleft center, how would you describe the positive and negative outcomes of implementing the CL/P Standard Set

Overall, do the positive outcomes outweigh the negative benefits of implementing the CL/P Standard Set?

Adoption

What advice would you give teams wanting to implement the CL/P Standard Set regarding transformation or syncing of workflows to aid implementation efforts?

Would you say that the barriers and facilitators are similar or different across cultures and health systems? How might your experiences and findings extrapolate to other centers in low- and middle-income countries?

Implementation

If you had to describe your implementation plan in phases, what would those phases be and can you elaborate a bit on each phase?

Were there changes to your implementation plan or strategy that surprised you or that you did not anticipate? Did this have a positive or negative impact?

With regards to the timepoints for collection of outcomes specified by the CL/P Standard Set, are they realistic in practice? If not, how is this mitigated by your cleft team? If yes, does your cleft team ensure consistency?

You discussed in the survey the possibility of the set promoting patient-centered care, can you describe in more detail how your center is achieving this?

What influences whether your team uses outcomes collected, as a platform for QI or as a repository of local or national outcomes?
What would you say are the implementation milestones, cleft teams need to reach before moving into intentional benchmarking?
Now that the CL/P Standard Set has been available to teams for a few years, what would you say are the main factors that promote or inhibit benchmarking at this stage?

Maintenance

How do you maintain data collection up to date? What are facilitators and barriers?
What are the best practices to maintain stakeholder motivation to continue implementing the CL/P Standard Set?
Do you have regular meetings to discuss the outcomes at your center? If yes, how is that going? If no, why not? What is needed to get there?
What is one area where ICHOM can support the growing cleft community?
Is there something you want to change in near future?

Closing questions

What do you wish you knew, before starting to implement the CL/P Standard Set?
Where do you see this network in 5 years? Are there changes we need to make in the next year or so to get there?
Are there any other topics, views or perspectives regarding facilitators and barriers to implementation of the CL/P Standard Set that you wish to share?

Chapter 7

Whitepaper: Implementation of the ICHOM Standard Set for Cleft Lip and/or Palate Across Four Centers

Ramirez JP, EPHM, IMPH[1]; Rogers-Vizena CR, MD[2]; Koudstaal MJ, MD, DMD, PhD[3,4];
Allori AC, MD, PhD[5]; Peterson P, MD[4]; Versnel SL, MD, PhD[6]; **Apon I, MD, MHS[3]**

[1] International Consortium for Health Outcomes Measurement (ICHOM), Boston, Massachusetts, USA
[2] Boston Children's Hospital, Boston, MA, USA and Harvard Medical School, Boston, MA, USA
[3] Department of Oral and Maxillofacial Surgery, Erasmus University Medical Center, Rotterdam, The Netherlands
[4] Department of Craniofacial Surgery, Karolinska University Hospital, Stockholm, Sweden
[5] Duke Cleft & Craniofacial Center, Duke Children's Hospital; Durham, NC, USA; and Duke University School of Medicine; Durham, NC, USA
[6] Department of Plastic and Reconstructive Surgery, Erasmus University Medical Center, Rotterdam, The Netherlands

Published online by ICHOM
https://conference.ichom.org/wp-content/uploads/2021/09/24915-CLP-Cleft-Lip-whitepaper.pdf

Introduction

Cleft lip and/or palate (CL/P) is a common congenital anomaly that involves malformation of the lip, dental arch, palate, facial skeleton, and nose, resulting in functional problems related to speech, hearing, eating, and breathing. Children with CL/P require specialty care delivered in stages ranging from birth through young adulthood. To meet these complex needs, comprehensive care is best provided by a multidisciplinary team.

For a cleft team to provide the best care possible — or to identify areas for improvement — it first must have a way to appraise its performance. Explicit measurement of holistic health outcomes enables health systems to prioritize resources on the outcomes that matter most. It is desirable that the outcome measures used by each team are standardized, such that a team may compare its performance relative to that of peer institutions, providing a better frame of reference for what the results mean.

In 2014, the International Consortium for Health Outcomes Measurement (ICHOM) convened a Working Group of 28 multi-disciplinary clinicians, academicians, and patient/family representatives from 8 countries, in order to create a set of standardized outcome measures for cleft teams. Over the course of the next year, the working group developed the ICHOM Standard Set for Cleft Lip/Palate Care[1], which provides guidelines for the outcome domains, specific outcome measures, phenotypic and demographic variables, and timepoints that should be used in the comprehensive assessment of cleft care (**Figure 1**). The Standard Set was designed with an emphasis on what matters most to patients and includes patient-reported outcome measures (PROMs). It was designed to be practical for implementation (fitting into routine clinical workflows), sustainable in the long-term, and adaptable to meet future needs.

After preparation of the Standard Set, four cleft teams piloted implementation:[2]

- Boston Children’s Hospital (BCH) Cleft and Craniofacial Center (Boston, MA), where Dr. Carolyn Rogers-Vizena leads the implementation efforts under the continued guidance of Dr. John Meara.

- Erasmus University Medical Center (EMC) Cleft Team at Sophia Children’s Hospital (Rotterdam, The Netherlands), where Dr. Maarten Koudstaal, previous team lead, with Dr. Sarah Versnel as the current team lead, implemented the Standard Set in routine clinical practice.[3]

- Duke Cleft and Craniofacial Center at Duke Children's Hospital (Durham, NC), where Dr. Alexander Allori leads the implementation efforts.[4]

- Stockholm Craniofacial Team at Karolinska University Hospital (Stockholm, Sweden), under the leadership of Dr. Petra Peterson with support from Dr. Koudstaal.

The purpose of this whitepaper is to share what these four teams experienced during their implementation processes. Key lessons for successful implementation relate to organizational "culture change", health information technology, and adaptation of clinical practice and workflow.

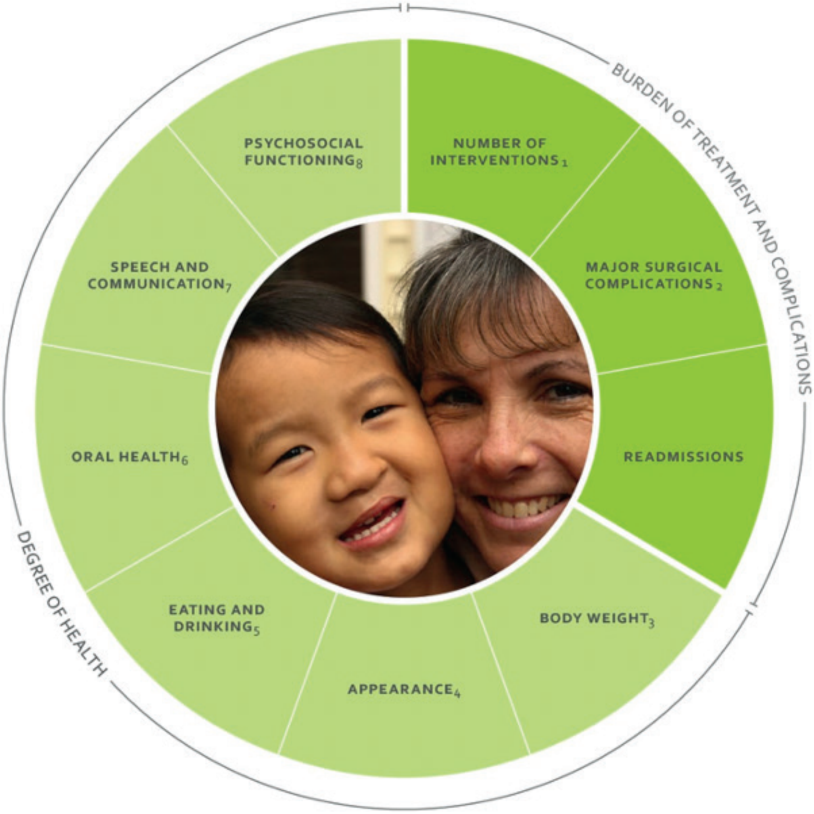
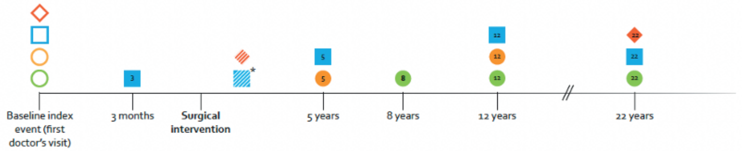

BURDEN OF TREATMENT AND COMPLICATIONS
NUMBER OF INTERVENTIONS 1
MAJOR SURGICAL COMPLICATIONS 2
READMISSIONS
BODY WEIGHT 3
APPEARANCE 4
EATING AND DRINKING 5
ORAL HEALTH 6
SPEECH AND COMMUNICATION 7
PSYCHOSOCIAL FUNCTIONING 8
DEGREE OF HEALTH

Figure 1 ICHOM CL/P Wheel. This wheel captures ICHOM's CL/P recommended minimum set of outcomes.



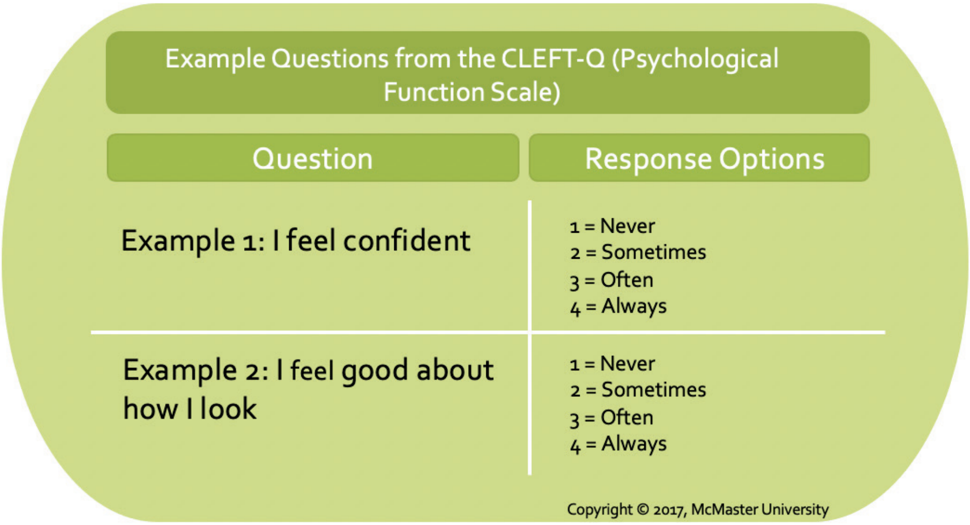
- ○ Baseline Patient-Reported Form
- ○ Baseline Parent-Reported Form
- □ Baseline Clinical Form
- ◇ Baseline Administrative Form
- ▨ Post-Operative Clinical Form
- ◆ Post-Operative Administrative Form
- ■ 3 Month Clinical Form
- ● 5 Year Parent-Reported Form
- ■ 5 Year Clinical Form
- ● 8 Year Patient-Reported Form
- ● 12 Year Patient-Reported Form
- ● 12 Year Parent-Reported Form
- ■ 12 Year Clinical Form
- ● 22 Year Patient-Reported form
- ■ 22 Year Clinical Form
- ◆ 22 Year Administrative Form

Timings

We encourage measurement according to the time points included in the above table. Recognizing that different countries have established practices and may take time to transition to these time points, we have agreed age categories for measurement that fit around the recommended specific ages:

3 months: +/- 2 weeks
5 years: 4-6 years
8 years: 8-9 years
12 years: 10 - 12.5 years
22 years: 22 years or end of treatment, whichever is soonest.
Post-operative period: 30 days following the date of the operation

Figure 2 ICHOM CL/P Timepoints. This figure captures the suggested timepoints for the set's PROMs and CROMs.



Example Questions from the CLEFT-Q (Psychological Function Scale)

Question	Response Options
Example 1: I feel confident	1 = Never 2 = Sometimes 3 = Often 4 = Always
Example 2: I feel good about how I look	1 = Never 2 = Sometimes 3 = Often 4 = Always

Figure 3 Example questions from the CLEFT-Q Psychological function scale.

Organizational transformation

Change management at all organizational levels is crucial for successful implementation. Specifically, this includes:

- Securing support from institutional leadership;
- Aligning and motivating all stakeholders involved in the project;
- Securing on-going financial and technical support.

Securing support from institutional leadership

Endorsement of outcomes measurement by top leadership is key for securing the necessary financial and technical support for implementation. It also helps foster alignment across all parties affected by the transformation (e.g., clinicians, administrators, and patients).

In all four institutions featured here, departmental and hospital leadership prioritized the creation of the ICHOM Standard Set from the start with the aim towards implementation once the Standard Set was finalized. The support from administrators that believed in the project was crucial in providing motivation and financial and administrative support needed for this work.

For example, in 2017, the board of Karolinska Hospital made implementation of all available ICHOM Standard Sets a key initiative. Similarly, at Erasmus University Medical Center, implementation of the Standard Set was part of a five-year strategic plan to transform the institution into a center for innovation in value-based healthcare (VBHC). Dr. John Meara, plastic surgeon-in-chief at Boston Children’s Hospital served as ICHOM’s Working Group Chair for the development of the Standard Set and was keen to see it implemented locally. He engaged the support of Boston Children’s Chief Executive Officer, Chief Medical Officer, and Information Technology leadership from the beginning of the implementation effort. Similarly, Dr. Alexander Allori, who served as an ICHOM Research Fellow and co-director of the Working Group during development of the Standard Set, was able to convince the director of the Duke Cleft & Craniofacial Center that Duke should be one of the pilot sites for implementation.

Aligning and motivating all stakeholders involved in the project

The implementation of outcomes measurement affects many aspects of clinical care. In particular, it requires changing clinic workflows, as is discussed below. Cleft care requires multiple interventions and long-term follow-up by an interdisciplinary team. So, garnering the endorsement and support of clinical leaders such as department chairs across all of these disciplines is critical for success. These leaders should be engaged in the project from the very beginning so that they can develop a sense of personal investment and ownership in the project. Support from hospital leadership can also facilitate this.

"The time investment at the beginning is critical for making sure that all stakeholders not only endorse the project, but also feel a sense of ownership of the project. That's the only way for the project to be accepted, cherished, and sustained over the course of the years".
Dr. Allori, Duke University Hospital

Early on, along with garnering the support from institutional leadership, the four teams communicated their vision to staff and others affected by the implementation process. The team leaders explained that the project mission is to better understand their clinical performance and discussed how collecting PROMs provides important data for understanding outcomes that matter most to patients, while focusing their discussions as a care team. They also highlighted how outcomes measurement could position centers to succeed in new performance-based reimbursement models as well as create a culture of increased ownership of results and satisfaction within the team.

Each team had the following members in their interdisciplinary cleft teams:

- Speech pathologist
- Maxillofacial surgeon
- Plastic Surgeon
- Otolaryngologists
- Psychologists
- Orthodontist
- Specialized Nurse
- Clinic Coordinator
- Pediatric Dentist
- Audiologist
- Clinical Genderists
- Pediatrician
- Obstetrician/Gynecologist

Moreover, each implementation team had a core team that was composed of an IT specialist, Project Lead, Academic Researcher, Cleft Surgeon, Clinic Coordinator, and Administrator. In some cases, one person fulfilled multiple roles, for example a cleft surgeon also acted as the project lead.

Managing change takes time and requires sustained levels of motivation and commitment from all stakeholders. As John Kotter recommends in his book *Leading Change*[5], significant changes in culture and process requires frequent reminders of the purpose for the change as well as celebration of early wins. Early on, the Project Leads invested significant time in change management – especially allaying fears of how the new processes might affect daily work, such as creating more documentation burden that would slow down the clinical workflow during a busy clinic day. Also important was explaining how every member of the team could get involved in utilizing the new data to improve team care. The Duke Team was responsive to early stakeholder feedback as the process unfolded, clarifying and improving the user interface of the data-collection platform whenever necessary. When the team coordinator asked if the outcomes data-collection system could also keep track of appointments and attendance, the implementation team built an extension to the project that could facilitate team administration. This allowed the team coordinator to use the system to identify risk factors for "no shows" (missed appointments) and to implement a rapid quality-improvement project that remedied the situation. These kinds of successes helped reinforce the value of measuring outcomes and sustain engagement across the cleft team.

Securing on-going financial and technical support

"If you have to use pen and paper, the risk of losing data is of course higher and it will be more difficult to do follow-up. Still, it is better to start collecting data and then ask for funding to analyze it later on."
Dr. Peterson, Karolinska University Hospital

While implementing outcomes measurement projects, teams will be faced with questions about who funds and delivers each component of the implementation process, especially since funding for outcomes measurement in cleft care is limited. Currently, the four cleft centers that are represented here fund their work through a combination of private grants, government funding, and institutional budgets. Although outcomes can be collected using pen and paper, to have a "real-time" picture of performance for use in quality improvement initiatives and for future benchmarking initiatives, an early investment in IT is necessary.

In addition to the initial investments in IT infrastructure required for outcomes measurement, it is also important to consider the cost of employees' time on the project.

This includes everything from the time that the core implementation team spends on the project to the additional time administrators and clinicians spend with patients to collect and review the outcomes data. Opportunity costs can arise when reallocating staff time to support the project or decreasing the number of patients seen per clinic visit to allow time for the collection and review of outcomes data. This is where securing commitment from organizational leadership, cleft team members, and supporting staff is central to ensuring the long-term success of the project.

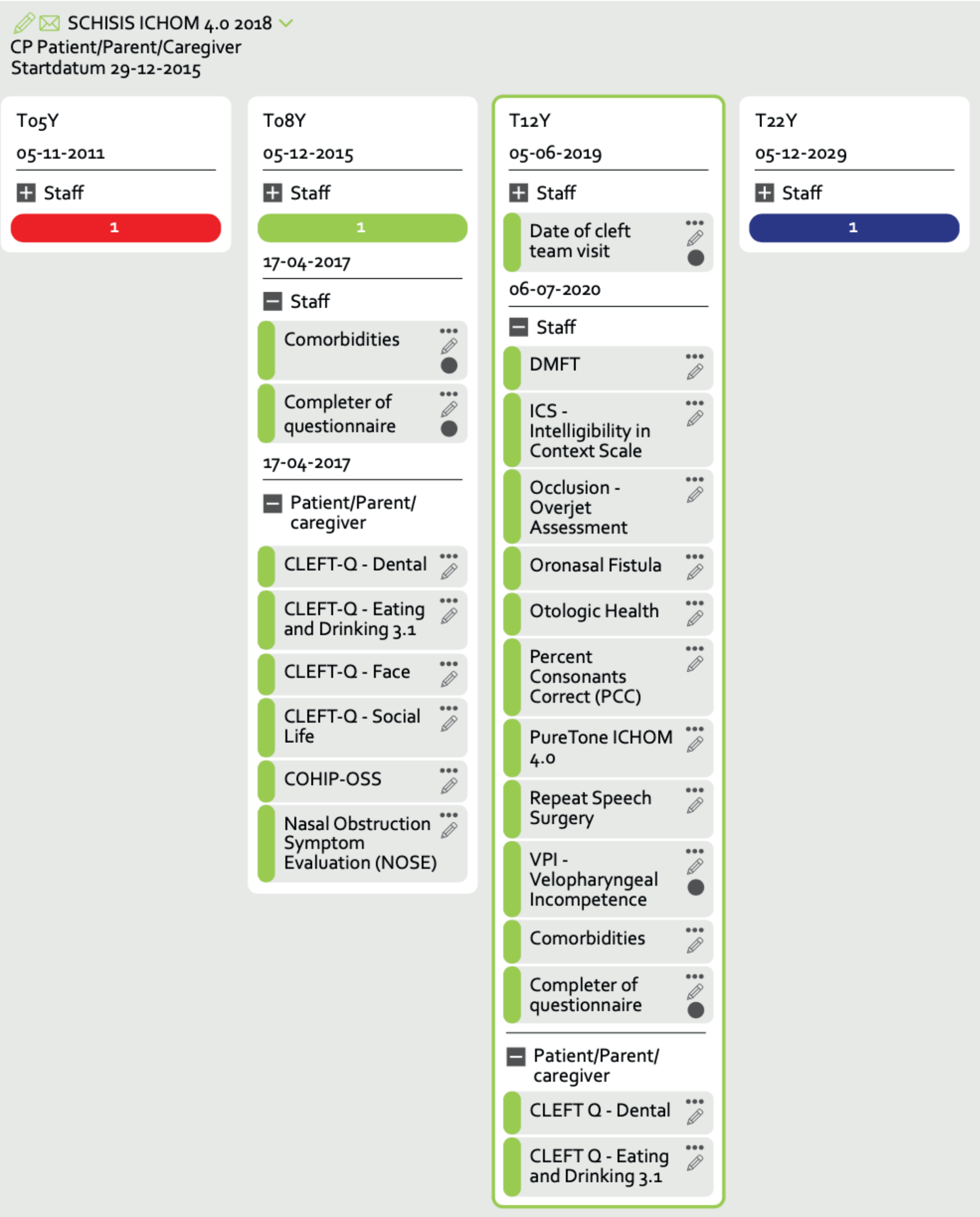
"In general, expect the implementation to take more time and resources then initially contemplated. If finances, or not exceeding a certain budget, will preclude the team form collecting the data, perhaps opt for a lower tech solution."
Dr. Rogers-Vizena, Boston Children's Hospital

Transformation in health-information technology (HIT)

What platform will be used to collect PROMs? Should outcomes data be stored in the EHR or separately? Should data platforms be built in-house or purchased from an external vendor?

These are the key information technology considerations when implementing outcomes measurement. Each approach offers unique advantages and challenges, which depend on the ultimate goal or motivation for outcomes measurement and the resources available to accomplish the goal. Regardless of the approach taken, it is important that the ultimate solution minimizes the burden of the end users – clinicians and patients.

At Erasmus University Medical Center, implementation of the Standard Set was funded via a grant from two major health-insurance companies in The Netherlands. A provision of the grant was that Erasmus would help other Dutch cleft teams to implement the Standard Set with the aim of developing a national registry for outcomes benchmarking in cleft care. To ensure standardized data collection across these different cleft teams, Erasmus invested in building its own outcomes collection and visualization platform, which was then made available free of charge to other implementing hospitals. In addition, since outcomes measurement is part of Erasmus's strategic mission, it made sense to invest resources in building a platform that could then be used to capture outcomes for all value-based healthcare projects at Erasmus. The platform is called *Zorgmonitor* (meaning 'Health Monitor') and is developed using a combination of Gemstracker and Limesurvey, both which are survey software's. It interfaces directly with Erasmus's EHR and includes clinician-specific and patient-specific dashboards.



SCHISIS ICHOM 4.0 2018
CP Patient/Parent/Caregiver
Startdatum 29-12-2015

T05Y
05-11-2011
Staff
1

T08Y
05-12-2015
Staff
1
17-04-2017
Staff
Comorbidities ●
Completer of questionnaire ●
17-04-2017
Patient/Parent/caregiver
CLEFT-Q - Dental
CLEFT-Q - Eating and Drinking 3.1
CLEFT-Q - Face
CLEFT-Q - Social Life
COHIP-OSS
Nasal Obstruction Symptom Evaluation (NOSE)

T12Y
05-06-2019
Staff
Date of cleft team visit ●
06-07-2020
Staff
DMFT
ICS - Intelligibility in Context Scale
Occlusion - Overjet Assessment
Oronasal Fistula
Otologic Health
Percent Consonants Correct (PCC)
PureTone ICHOM 4.0
Repeat Speech Surgery
VPI - Velopharyngeal Incompetence ●
Comorbidities
Completer of questionnaire ●
Patient/Parent/caregiver
CLEFT Q - Dental
CLEFT Q - Eating and Drinking 3.1

T22Y
05-12-2029
Staff
1

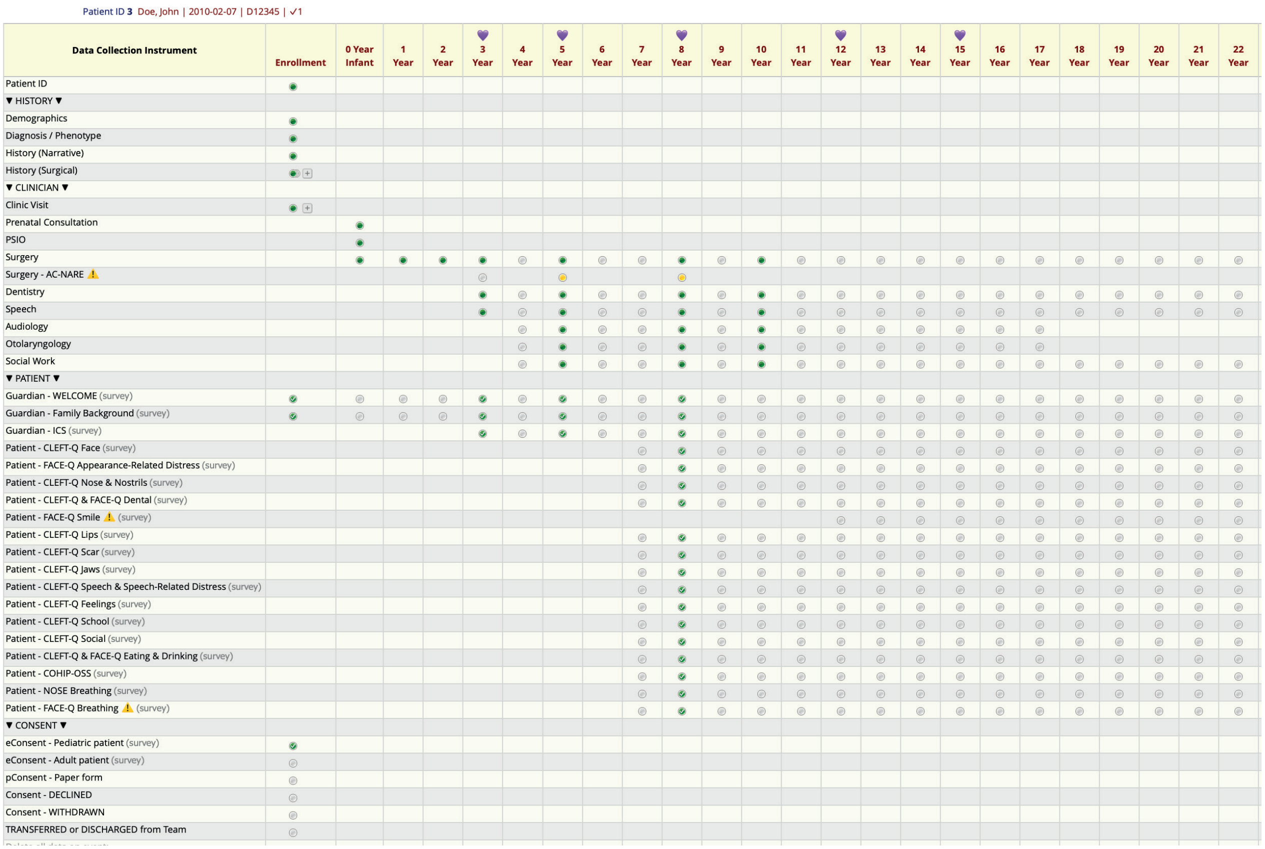
Figure 4 Erasmus University Medical Center's data collection tool. Each column represents a timepoint (0, 5, 8, 12 and 22 years) for the collection of case-mix variables and clinical outcomes (indicated by 'staff') and/or patient-reported outcome measures (indicated 'parents/caregivers'). Green buttons indicate the measurement has been completed and can be viewed, whereas the yellow measurements are still open for completion. Red buttons represent time for completion has expired. Blue buttons will open for data completion in the future. Extra buttons per surgical procedure including post-operative complications can be added when indicated.

Similarly, implementation of outcomes measurement was considered a key priority by top leadership at the Karolinska University Hospital. So, after piloting data collection using pen and paper, the hospital contracted a third-party vendor to build an outcomes collection and visualization platform called *Webformulär*. Like the platform developed at Erasmus, it interfaces directly with Karolinska's EHR.

Boston Children's Hospital used a phased implementation approach. Dr. Rogers-Vizena, who led the initiative, first worked with the team's QI lead to engage the various cleft care specialties and develop a core interdisciplinary team. Next, they focused on implementation of PROMs measurement in clinical care. After facing challenges in trying to integrate an external PROMs measurement solution with the hospital EHR, they chose to capture PROMs using REDCap[6] because it is easy, flexible and cost-effective. In addition to capturing PROMs data in-clinic, the team also developed "red flags" for PROM scores. PROM surveys are administered by a research or QI assistant on an iPad during a clinic visit. If a patient's scores indicate a concern, the "red flag" is triggered and the assistant alerts the team's social worker or other relevant clinician who then evaluates the situation and schedules further consultations when necessary. Once the team felt confident in the PROMs implementation, they turned their focus to collecting the clinician-reported measures.

Implementation at Duke Children's Hospital started a year later than at Boston Children's and Erasmus. As a result, their implementation team had the opportunity to learn from the other teams' experiences. Also, in contrast to Boston Children's and Erasmus, which implemented the Standard Set as quality-improvement projects, Duke chose to implement the framework as a research project. Doing so required more work (informed consent from patients and/or caregivers, continually updated IRB protocols, and data-transfer agreements (DTA)) but would permit easier integration of the new data into existing research activities. Research funds were limited, so technical solutions for data capture analysis needed to be affordable, practical, and easy to maintain. Dr. Allori chose REDCap as the exclusive platform to organize and run the entire project. REDCap offered several advantages: (1) it is open-access and free to use; (2) it is well-known to researchers and regulatory agencies; (3) it has an adequate and flexible feature set for database and questionnaire development; (4) it has very robust versioning and security features – crucial for safeguarding protected health information (PHI); and (5) usage of REDCap lends itself toward agile development practices – by freeing the project from the constraints of the EHR, the team could iterate on the design very quickly. Agility was very important, as ICHOM had worked on a few early revisions in the Standard Set (from version 1.0 to 2.0, 3.0, 3.4, 4.0, and 4.1 between 2015-2018), and with every revision, the project would need to be updated to remain compliant with the standards. Being able to adapt this process directly, eliminated extra time, cost, and even effort and frustration.

The unique motivations and IT constraints of the four organizations shaped the way each team developed technical implementations of the Standard Set and workflows for data collection. Each approach was ultimately successful because it respected the needs of the various clinical disciplines involved in cleft care and was designed to optimize local workflow.



Patient ID 3 Doe, John | 2010-02-07 | D12345 | ✓1

Data Collection Instrument	Enrollment	0 Year Infant	1 Year	2 Year	💜 3 Year	4 Year	💜 5 Year	6 Year	7 Year	💜 8 Year	9 Year	10 Year	11 Year	💜 12 Year	13 Year	14 Year	💜 15 Year	16 Year	17 Year	18 Year	19 Year	20 Year	21 Year	22 Year
Patient ID	●																							
▼ HISTORY ▼																								
Demographics	●																							
Diagnosis / Phenotype	●																							
History (Narrative)	●																							
History (Surgical)	● +																							
▼ CLINICIAN ▼																								
Clinic Visit	● +																							
Prenatal Consultation		●																						
PSIO		●																						
Surgery		●	●	●	●	○	●	○	○	●	○	●	○	○	○	○	○	○	○	○	○	○	○	○
Surgery - AC-NARE ⚠					○		🟡			🟡														
Dentistry					●	○	●	○	○	●	○	●	○	○	○	○	○	○	○	○	○	○	○	○
Speech					●	○	●	○	○	●	○	●	○	○	○	○	○	○	○	○	○	○	○	○
Audiology						○	●	○	○	●	○	●	○	○	○	○	○	○	○					
Otolaryngology						○	●	○	○	●	○	●	○	○	○	○	○	○	○					
Social Work						○	●	○	○	●	○	●	○	○	○	○	○	○	○	○	○	○	○	○
▼ PATIENT ▼																								
Guardian - WELCOME (survey)	✓	○	○	○	✓	○	✓	○	○	✓	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Guardian - Family Background (survey)	✓	○	○	○	✓	○	✓	○	○	✓	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Guardian - ICS (survey)					✓	○	✓	○	○	✓	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Patient - CLEFT-Q Face (survey)									○	✓	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Patient - FACE-Q Appearance-Related Distress (survey)									○	✓	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Patient - CLEFT-Q Nose & Nostrils (survey)									○	✓	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Patient - CLEFT-Q & FACE-Q Dental (survey)									○	✓	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Patient - FACE-Q Smile ⚠ (survey)														○	○	○	○	○	○	○	○	○	○	○
Patient - CLEFT-Q Lips (survey)									○	✓	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Patient - CLEFT-Q Scar (survey)									○	✓	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Patient - CLEFT-Q Jaws (survey)									○	✓	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Patient - CLEFT-Q Speech & Speech-Related Distress (survey)									○	✓	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Patient - CLEFT-Q Feelings (survey)									○	✓	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Patient - CLEFT-Q School (survey)									○	✓	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Patient - CLEFT-Q Social (survey)									○	✓	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Patient - CLEFT-Q & FACE-Q Eating & Drinking (survey)									○	✓	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Patient - COHIP-OSS (survey)									○	✓	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Patient - NOSE Breathing (survey)									○	✓	○	○	○	○	○	○	○	○	○	○	○	○	○	○
Patient - FACE-Q Breathing ⚠ (survey)									○	✓	○	○	○	○	○	○	○	○	○	○	○	○	○	○
▼ CONSENT ▼																								
eConsent - Pediatric patient (survey)	✓																							
eConsent - Adult patient (survey)	○																							
pConsent - Paper form	○																							
Consent - DECLINED	○																							
Consent - WITHDRAWN	○																							
TRANSFERRED or DISCHARGED from Team	○																							

Figure 5 Duke Children's Hospital Dashboard. The radio buttons show available outcomes measures from infancy through 22 years of age. The filled buttons show when data were collected – in this simulation, team and surgical data at years 0, 1 and 2; clinical and caregiver data at 5 years; and clinical, caregiver, and patient-reported data at 8 years. Partial data collection occurred at age 10 years. The "purple heart" ages represent the required data-collection timepoints for ACCQUIREnet (Allied Cleft & Craniofacial Quality-Improvement and Research Network).[7] The team may optionally measure outcomes in other years.

Transformation in workflow and clinical care

Outcome measurement requires significant changes both to how a clinic is run – the workflow – and the role of the patient in their own care. All four teams noted how these changes:

- Effect clinic operations or workflows;
- Serve as a catalyst for increased collaboration between clinical disciplines and clinic staff;
- Highlight the need to educate patients and families about the role of outcomes measurement in their care.

Effects on clinic operations or workflows

The Standard Set details the specific outcomes and case-mix variables that need to be collected (the "what" of data collection) and the timepoints for data collection ("when"). But it does not prescribe the "how" or "who" of data collection, as the Working Group knew that this would vary across organizations and cleft teams. Given the highly multidisciplinary nature of cleft care, the "who" (i.e., who should measure the outcome and record the data) warranted special consideration. For example, a dentist, orthodontist, and oral-maxillofacial surgeon are all appropriate for measuring *Dental Health* using Decayed, Missing, and Filled Teeth (DMFT) index scores; similarly, both a psychologist and social worker may administer CLEFT-Q subscales (social, school and psychological) to measure sociometric and assess *Psychosocial Functioning*. It is important to figure out which of these specialists should be assigned the responsibility to do so for standardized, prospective data collection.

"In the beginning we had some people argue that it costs a lot of extra time but once you have everything up and running and you're used to a new way of working, it really fits."
Dr. Versnel, Erasmus University Medical Center

To craft an approach that would work best for their teams, all four implementation directors started out by mapping current clinic workflows. This provided the necessary information to determine who needed to be involved in data collection and agree on new workflows that allowed for reliable data collection. Each of the implementation teams then piloted the new workflows and noted the additional time needed to allow for outcome measurement. For example, Duke started by collecting outcomes data on paper forms. Cleft team members were instructed to collect the data (filling out the paper forms) on a few patients each week for a short trial period. The bottom of each form had a blank area where each team member could write thoughts, questions, comments, criticisms, suggestions, etc. This early experience was critical to answering the "who" question – who should collect each data element. In Duke's case, this process identified that certain elements assigned to Otolaryngology should be switched to Audiology, and certain elements assigned to Social Work were better worked into the family-reported surveys. After the data-collection workflow had been adequately clarified, the Duke team designed a REDCap-based implementation of this workflow. All team members were trained to use the new system and use it live in clinic, starting first with only one patient per clinic, then two, then four, etc., until they were capturing data on the majority of patients. During this phase, the emphasis was on the process of data collection rather than the data being collected. They used this period to identify friction points, create necessary clarifications or workarounds, and retraining staff as necessary. After working out the kinks, the team announced a "go-live" date, which allowed for a practice run, and then began true prospective data collection.

EMC used a different approach. They started out measuring the full standard set – all diagnoses and all ages. To make this manageable, they selected two to five patients per clinic to test the data-collection software until they had experience with capturing all outcomes at all timepoints. Then, they made any adjustments suggested by this experience and then moved to measuring outcomes on the full patient population.

Collection of speech outcomes data has an influence on clinical practice and workflow

The ICHOM Standard Set requires some speech and audiology measurements that might not traditionally be collected for specific patients at the timepoints outlined by the Standard Set. These cleft disciplines require extra time to collect measurements and adjustments have been made to ensure these outcomes are measured as part of routine care.

For example, the Standard Set measures *articulation* as an outcome of the *speech* outcome domain, using the Percentage of Correct Consonants (PCC) instrument. At Duke, speech therapists use connected speech rather than isolated speech for clinical evaluation therefore clinicians need to do a different speech evaluation that's not part of their standard clinical practice. This speech consult takes about 5-10 minutes longer as compared to non-data collection patients. To support speech therapists in ensuring this outcome is measured, the Duke cleft team established a norm that these counts had to be included in the patient's record by noon the next day. This allowed more time for speech therapists to complete their counts, measure what matters to patients, and help maintaining the fidelity of the Standard Set variables.

Outcomes measurement as a catalyst for interdisciplinary collaboration

All four cleft teams found that having patients complete PROMs as part of the clinic visit greatly enhanced the quality of their discussions with patients and catalyzed discussions between patients and their families, and between the clinical disciplines represented in the care team. Some cleft teams, such as the team at EMC, noted that measuring outcomes as defined by the Standard Set extended each patient's clinical encounter by a few minutes, which means fewer patients are seen per clinic. However, the benefit of improved communication offset this concern.

For some cleft centers, patients and family members complete outcomes questionnaires in advance of the clinic visit. For example, at EMC and Karolinska, an e-mail invitation to complete PROMs on-line is sent out to patients two weeks prior to their visit. Once patient responses are received, they are visible in the EHR. A nurse specialist generates an overview of each patient's responses. The cleft team then meets right before cleft clinic to discuss each patient, including reviewing their patient-reported outcome measures (PROMs). As a result, the cleft clinicians are able to focus their consultation on the patient's main concerns. The cleft centers that used this approach found that, to ensure completion of the questionnaires at home, patients needed to be educated through discussions with their providers about the importance of this data in their care. This also served as a good way for parents to prepare their children before the clinical visit, where they could explain, discuss and answer sometimes difficult questions in the safe home-environment.

Putting outcomes measurement to work for patients

Collecting PROMs creates an expectation for patients that clinicians will follow up and address the concerns the patient expresses. At EMC, patient reported outcome scores are always discussed during the outpatient clinic visit by a specialized nurse. The nurse gives feedback on the scores and asks more specific questions if an answer raises concern. For example, if a patient's self-reported scores for psychosocial well-being are low, the nurse will discuss whether the patient wishes to have a consult with the psychologist or social worker. Before measuring outcomes based on the ICHOM Standard Set, psychological needs sometimes went undetected. Now, they are intentionally screened and discussed.

"The fact that the child doesn't bring it up, doesn't necessarily mean that it's not an issue. If you don't raise it [as a clinician], you don't know it and you can't help them in a timely

fashion. It's important to ask questions, but it's more important to train your team to start the conversation and make sure that they're not afraid of the answer…having the psychologist explaining this to our team, but also if the parents raised the issue, we were able to explain during clinic this is why we're doing it. From research and experience, we know it is better to ask the question than avoid it."
Dr. Koudstaal, Erasmus University Medical Center

One of the CLEFT-Q speech scales in the Standard Set measures speech-related distress and questions "how do you feel about speaking?".[8] If a patient shows low marks on this scale, clinicians at Duke use this as a starting point for a discussion with a question like, "I noticed that you answered these questions 'always'. It seems to bother you a lot. Can you tell me more about that?" This invites patients into a conversation so that their needs can truly be addressed.

Although these cleft centers review and discuss outcomes with patients, they are not yet using the data directly in clinical decision making. This is largely due to the lack of data on normative values and cut-off scores for the PROMs included in this Standard Set. EMC is conducting research on how these values should be presented to patients at different ages (young children vs. 22-year-old patients) and whether or not results should be shown against earlier outcome scores, against normative data, or against outcome scores of other cleft populations.

Future direction: towards collaborative networks and benchmarking

Comparison of outcomes and benchmarking requires the collection of large, robust datasets that are accurate, complete, and provide a cross-representation of different ages and measurement timepoints. Even more robust datasets are required for risk adjustment across phenotypes, syndromic conditions, etc. The nature of CL/P care presents significant challenges for developing such a dataset. ICHOM's CL/P Standard Set is designed to capture all these outcomes over a large span of time, often with wide time intervals (3 years or more) in between measurements. This presents a challenge for longitudinal data collection. For example, it is not uncommon for patients to transfer their care to a different institution at some point. Therefore, it takes considerable time and collaborative data sharing to develop robust outcomes datasets.

The four teams identified the following challenges to outcomes comparisons and benchmarking:

- Navigating privacy laws that create a barrier to accessing and sharing data, which can exclude teams from benchmarking efforts. This has highlighted the need for pooled analysis and on-site analysis;
- The need to develop protocols to ensure that data is extracted in a uniform format/coding for running pooled analysis across centers;
- The need to develop risk-adjustment models for outcomes benchmarking and best practices or guidelines for performing cohort analyses.

As a result of these challenges, the implementation efforts discussed here have not yet resulted in outcomes comparisons or benchmarking between organizations. However, both Duke Children's Hospital and Erasmus Medical Center are leading regional collaboratives.

Duke Children's Hospital founded ACCQUIREnet - the Allied Cleft & Craniofacial Quality-Improvement and Research Network. It is a multi-site collaborative network dedicated to implementation of the Standard Set as well as multi-site aggregation, benchmarking, and comparison of outcomes. The project is under Dr. Allori's direction. Duke serves as the coordinating center for ACCQUIREnet, as well as the statistical support center for data analysis. The project is registered as an observational study on clinicaltrial.gov (NCT02702869). Presently, six additional American centers have joined ACCQUIREnet, agreeing to collect the ICHOM Standard Set data using the REDCap-based system developed by Dr. Allori. ACCQUIREnet is open for membership to all North American cleft teams.

Similarly, Erasmus University Medical Center is currently the coordinating center for the European Reference Network (ERN) for rare and/or complex Craniofacial Anomalies and ENT disorders (ERN CRANIO).[9] The network has 29 member hospitals across 11 EU member states. The ERN CRANIO working group for cleft lip and/or palate agreed to adopt the ICHOM Standard Set as minimal dataset for registering outcomes at ERN CRANIO sites. The ERN registry is under development and will enable data collection across Europe, for the primary purpose of evaluation of quality of treatment, and outcomes research in the future. In order to optimize the set for outcomes research in the different domains, additions and adjustments to the Standard Set are being examined. The goal is to make the database compatible with the ACCQUIREnet database in order to facilitate future collaboration in outcomes research and possibly benchmarking, while respecting privacy laws. Currently, six other Dutch cleft centers are also working to implement the

ICHOM Standard Set in their clinical practice. The cleft team at Erasmus University Medical Center is collaborating with these six teams and Dutch Hospital Data (DHD) to develop a national benchmark with uniform collected outcome data. DHD is a foundation that collects, manages, and processes data from hospitals to provide information for decision making management. DHD is developing a secure, on-line dashboard for presenting the aggregated outcomes data from the various Dutch cleft teams. The plan is to hold regular meetings with representatives of each cleft team to compare outcome results, discuss differences and learn from each other. It will also be possible to use this dashboard for quality improvement projects within one cleft team.

It is noteworthy that since both ACCQUIREnet and ERN CRANIO have implemented the ICHOM Standard Set, data collected by sites in these two networks is interchangeable. Already, the cleft teams from Duke, BCH, Erasmus, and Karolinska have rich research collaborations, particularly focused at the moment on optimizing the ICHOM Standard Set. Their observations and recommendations will be shared with the ICHOM Stewardship Committee and Scientific Advisory Council for consideration for future iterative improvements to the Standard Set.

Conclusion

The implementation experiences of these four cleft centers illustrate the different approaches that can be taken to successfully implement outcomes measurement in routine clinical practice as well as some of the common challenges and barriers. Their experiences all highlight the benefits of outcomes data collection for improved communication between patients and clinicians. The hope is that the experiences shared here will inform and encourage others to implement outcomes measurement, laying the groundwork for outcomes comparisons and benchmarking over time.

Acknowledgements

The authors thank all team members from Duke Cleft & Craniofacial Center and its research team, Erasmus University Medical Center, Boston Children's Hospital Cleft & Craniofacial Center and Karolinska Stockholm Craniofacial Team. Full list of names at the end of the paper. The authors would also like to thank the ICHOM team, Dr Stephanie Wissig and Andria Joseph, for their help in finalizing the paper.

Duke Cleft & Craniofacial Center

Jeffrey R. Marcus
Carlee Jones
Alexander C. Allori
Susan E. Anderson
Erica Brecher
Jeffrey Cheng
Sindhura Citineni
Ana Maria Fernandez
Dylan Hamilton
Melissa Hill
Christine Holmes
Martha Ann Keels
Jenny Kern
Dan King
Blessine McPherson
Jillian Nyswonger
David B. Powers
Eileen Raynor
Krista Roper
Pedro E. Santiago
A. Barron Suárez
Yvette Thompson
Karma Tockman
Ashanti Ballard
Mary Atkinson
Cathy Frye

Erasmus University Medical Center

Mona Haj
Mariska van Veen
Eppo Wolvius
Jet de Gier
Irene Mathijssen
Jan Hazelzet
Roel Faber
Piet-Hein van Twisk
Nicole Posch
Mieke Pleumeekers
Henriette Poldermans
Lindsay Heijkoop
Jolanda Okkerse
Francien Meertens
Laura Kind
Nicoline van der Kaaij
Ellis van der Voort
Gladys Mijnals

Boston Children's Hospital

Susan Flath Sporn
Michael Doyle
Karl Sanchez
Jonathan Cho
Tamia Hargrove
Mariana Nava
Alan Dupre
Abiola Faniyan
Chitra Dwarka
Prerna Kahlon
Geralyn Woodnorth
Roger Nuss
Lynn Schwartz
Liza Catallozzi
Roseanne Clark
Ellyn Zitzer
Elizabeth Ross
Richard Bruun
Joan Stoler
Catherine Nowak
Lauren Mednick
Bonnie Padwa
John Meara

Samantha Hall
Elske Strabbing
Olivia Oppel
Aimee Madden

Karolinska University Hospital
Katrin Stabel-Svensson
Kristina Jansson
Erik Neovius
Jill Nyberg
Liisi Raud-Westberg
Emilie Hagberg
Agneta Karsten
Marie Pegelow
Mathias Lemberger
Malin Vesterbacka
Kirsi Inola-Valck
Sara Bagher

References

1. Allori AC, Kelley T, Meara JG, et al. A Standard Set of Outcome Measures for the Comprehensive Appraisal of Cleft Care. *Cleft Palate Craniofac J.* 2017;54(5):540-554.
2. Apon I, Rogers-Vizena CR, Koudstaal MJ, et al. Barriers and Facilitators to the International Implementation of Standardized Outcome Measures in Clinical Cleft Practice. *Cleft Palate Craniofac J.* 2021:1055665621997668.
3. Arora J, Haj M. Implementing ICHOM's Standard Sets of Outcomes: Cleft Lip and Palate at Erasmus University Medical Centre in the Netherlands. International Consortium for Health Outcomes Measurement (ICHOM). www.ichom.org. Published 2016. Accessed December 1, 2020.
4. Bittar PG, Carlson AR, Mabie-DeRuyter A, Marcus JR, Allori AC. Implementation of a standardized data-collection system for comprehensive appraisal of cleft care. *Cleft Palate Craniofac J.* 2018;55(10):1382-1390.
5. Kotter J. *Leading Change.* Boston: Harvard Business School Press; 1996.
6. Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: Building an international community of software platform partners. *J Biomed Inform.* 2019;95:103208.
7. ACCQUIREnet. https://surgery.duke.edu/divisions/plastic-maxillofacial-and-oral-surgery/research/clinical-research/datalab-clinical-care-and-population-health/accquirenet. Accessed January 14, 2023.
8. Klassen AF, Riff KWW, Longmire NM, et al. Psychometric findings and normative values for the CLEFT-Q based on 2434 children and young adult patients with cleft lip and/or palate from 12 countries. *CMAJ.* 2018;190(15):E455-E462.
9. Cranio ERN. https://ern-cranio.eu/network-activities/activities/. Accessed January 14, 2023.

Chapter 8

Healthcare Use and Direct Medical Costs in a Cleft Lip and Palate Population: An Analysis of Observed and Protocolized Care and Costs

Apon I, MD, MHS[1]; van Leeuwen N, PhD[2]; Polinder S, PhD[3]; Versnel SL, MD, PhD[4]; Wolvius EB, MD, DMD, PhD[1]; Koudstaal MJ, MD, DMD, PhD[1]

[1] Department of Oral and Maxillofacial Surgery, Erasmus University Medical Center, Rotterdam, The Netherlands
[2] Department of Public Health, Section Medical Decision-Making, Erasmus University Medical Center, Rotterdam, The Netherlands
[3] Department of Public Health, Section Health Technology Assessment and Implementation, Erasmus University Medical Center, Rotterdam, The Netherlands
[4] Department of Plastic and Reconstructive Surgery, Erasmus University Medical Center, Rotterdam, The Netherlands

Manuscript accepted for publication in IJOMS

Abstract

Costs in cleft care have been scarcely investigated. This study describes observed healthcare utilization and medical costs for patients with a cleft, compares this to expected costs based on treatment protocol, and explore additional costs of implementing the International Consortium for Health Outcomes Measurement (ICHOM) Standard Set for Cleft Lip and Palate (CL/P). Forty patients with unilateral CL/P between 0 and 24 years, treated between 2012 and 2019 at Erasmus University Medical Center were included. Healthcare services (consultations, diagnostic and surgical procedures) were counted and costs were calculated. Expected costs based on treatment "protocol" were calculated by multiplying healthcare products by its product prices. Correspondingly, additional expected costs after implementing the ICHOM Standard Set ("protocol+ICHOM") were calculated. Observed costs were compared with "protocol" costs, and additional expected "protocol+ICHOM" costs were described. Total mean costs were highest the first year after birth (€5,596,) mainly due to surgeries. Mean observed total costs (€40,859) for the complete treatment (0-24 years) were 1.6 times the expected "protocol" costs (€25,198) due to optional, non-protocolized procedures. Hospital admissions including surgery were main cost drivers accounting for 42% of observed costs and 70% of expected "protocol" costs. Implementing the ICHOM Standard Set increased protocol-based costs by 7%.

Key words: cleft lip, cleft palate, health care costs, patient-reported outcome measures, practice patterns, physicians'.

Introduction

In 2006, the publication of the book “Redefining Health Care” by professor Porter and Teisberg has initiated a paradigm shift in healthcare; achieving high value for patients became the overarching goal of healthcare delivery in which value is defined as health outcomes per dollar spent.[1,2] As one of the value-based healthcare (VBHC) pilot centers in the Netherlands, the cleft team of the Erasmus University Medical Center started measuring outcomes in patients with cleft lip and palate as part of their routine care. In 2016, the Standard Set for Cleft Lip and Palate (Standard Set), developed by the International Consortium for Health Outcomes Measurement (ICHOM)[3-5], was implemented for this purpose. The Standard Set includes clinician-reported outcomes, patient characteristics, and incorporates the patient’s perspective on health by multiple patient-reported outcome measures (PROMs) related to appearance, speech, and facial and psychosocial functioning.[3,5,6] These domains are of special interest for patients with a cleft because this congenital facial disorder causes feeding difficulties and impairs facial growth, articulation, dentition, and psychosocial health.[7-9] Besides the addition of various PROMs to the treatment protocol, the Standard Set introduced an extra cleft team visit at 22 years, and additional audiological and speech examinations.[3,5,6]

So far, research on value-based healthcare initiatives in cleft has mainly focused on the patient’s outcomes and how to measure and improve them. Medical costs, as part of the value-based healthcare equation, have been scarcely investigated. Most studies only focused on specific parts of the cleft treatment with a relatively short follow-up period, such as surgical interventions, or described costs at a highly aggregated health-insurance level.[10-15] Also, research has shown that treatment protocols for cleft vary widely, both nationally and internationally[16,17], but there are no studies exploring to what extent treatment protocols correspond with actual care provided. Furthermore, the Standard Set was developed to be implemented in routine care globally, but the adoption is hindered by the belief that implementation will increase medical costs considerably[18] even though this assumption has not been investigated yet.

A better understanding of healthcare utilization patterns and medical costs during the complex and long treatment period for cleft lip and palate is essential for the following reasons: 1) to adapt care pathways adequately and efficiently, 2) to determine the ‘value’ of cleft lip and palate care based on the “VBHC-equation”[1,2], and 3) to lead negotiations between health-insurance companies and hospitals towards fair pricings for future payment strategy transformations, such as bundled payments.[2]

Therefore, the aim of this study was twofold. Firstly, to describe the total healthcare use and direct medical costs of care for patients from 0 to 24 years old with a unilateral cleft lip and palate, and compare this to the expected costs based on the treatment protocol. Secondly, to explore the additional protocol-based costs after the Standard Set implementation.

Methods

This retrospective cohort study was conducted from a Dutch academic hospital's perspective (Erasmus University Medical Center, Rotterdam). Collected data was registered as part of routine care and extracted from the patient's electronic health record (EHR) or the institution's information systems. Research ethics approval was granted by the Institutional Review Board of the Erasmus University Medical Center, Rotterdam, The Netherlands (MEC-2016-156).

Study population

Patients with a unilateral cleft lip and palate (UCLAP) between 0 and 24 years, treated by the Erasmus University Medical Center's cleft team between January 1, 2012 and December 31, 2019, were eligible. The UCLAP phenotype was chosen because this is the most complex and severe entity of cleft, and the unilateral variant is more common than the bilateral form.[7] There were no exclusion criteria since our aim was to obtain a patient population representing real practice. Eligible patients were identified through the '*Zorgmonitor'* (English: *Healthcare Monitor*), a secured platform linked to the patient's EHR, for the collection of outcome data within the Erasmus University Medical Center.[4] From all identified patients with UCLAP, a group of 40 patients was randomly sampled to match the real patient population as close as possible. This number was chosen for feasibility reasons, due to the time-consuming and labor-intensive nature of collecting and sorting all data.

Study parameters

First, the volume of cleft-related healthcare services delivered to the patients was counted. Healthcare services included medical consultations, diagnostics, and surgical procedures with hospital admissions. A detailed overview of the collected parameters is presented in **Table 1**. Due to privacy legislation, requesting any type of information on externally performed cleft-related treatments, such as speech and language therapy or dental and orthodontic care, was not allowed.

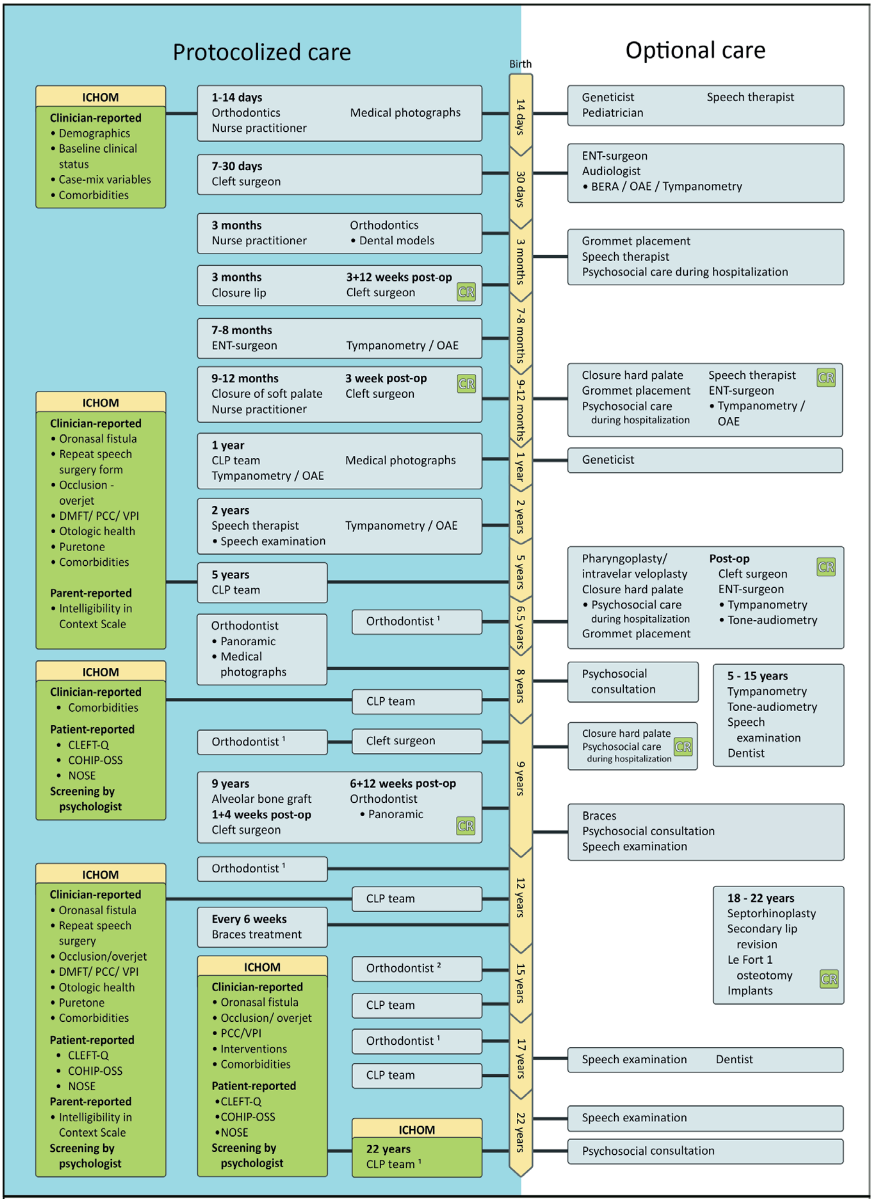
Second, all observed direct medical costs were calculated using the formula of 'costs = volume of healthcare service *x* price of the healthcare product'.[19] Prices of healthcare products were collected from the hospital's financial information systems in Euros and were based on the 2019 price allocations (**Table 1**). Prices for the years of 2012 - 2018 were adjusted for inflation according to the Dutch price index percentages (**Supplemental Material - Table 1**).[20]

Healthcare services			**Mean price**
Medical consultations	Protocolized consultations	Cleft surgeon	€89
		Ear, nose, throat specialist	
		Orthodontist	
		Dentist (per 5 minutes)	
		Speech therapist	
		Specialized nurse	
	Optional, non-protocolized consultations	Social care worker	€165
		Psychologist	
		Screening CLEFT-Q (by psychologist)	
		Anesthesiologist	
		Geneticist	
		Pediatrician	
		Psychosocial care during hospitalization	
Diagnostic procedures	Medical imaging	Medical photographs	€30
		Dental models	
		Skull-profile X-ray photograph	
		Panoramic photograph	
	Audiological testing	Tympanometry	€37
		Oto-acoustic emission	
		Tone audiometry (including PureTone)	
	Other procedures	Psychological examination	€195
		Speech/language examination	
		Percent Consonants Correct (PCC, by speech therapist)	
		Naso-endoscopy	
Surgical procedures	Primary procedures	Closure of cleft lip	€3038
		Closure of soft palate	
		Closure of hard palate	
		Alveolar bone grafting	
	Secondary or optional, non-protocolized procedures	Pharyngoplasty	€1415
		Grommet placement (per side)	
		Lip revision	
		Septorhinoplasty	
		Le Fort 1 osteotomy	
		Implants	
Hospitalization		One day hospital admission	€795

Table 1 Overview of healthcare services collected, and related mean, rounded price allocations of 2019 as used in this study.

Third, since cleft care is highly protocolized, two treatment protocols, followed by the Erasmus University Medical Center's cleft team, were outlined to estimate care use and medical costs in case a patient solely follows one of the protocols: 1) the treatment protocol applicable before 2016, hereafter named "protocol", and 2) the treatment protocol expanded by the implementation of the outcome measures of the Standard Set, hereafter named "protocol+ICHOM" (2016 - ongoing). Important additions to the "protocol" by the Standard Set included various psychosocial PROMs and extra audiological and speech testing (e.g. PureTone, Percent Consonants Correct) around the age of 8, 12 and 22 years, and an extra cleft team visit at the age of 22.[3,5,6] The volume of care and costs based on the protocols do not include optional, non-protocolized surgeries and treatments due to complications. Details of both treatment protocols are presented in **Figure 1.**

Observed healthcare use and cost data was linked to the year in which the service was delivered, or costs were made. General information on sex, age, adoption status and presence of a genetic syndrome was also collected.



🟩: ICHOM addition CR: Clinician reported outcomes including complications and burden of treatment forms

Prior to surgery, a visit to an anesthesiologist takes place. The CLP team consists of a cleft surgeon (specialized plastic or maxillofacial surgeon), orthodontist, ENT-surgeon, speech therapist, dentist and nurse practitioner. ENT: Ear, nose, throat; PCC: Percent Consonants Correct; COHIP-OSS: Child's Oral Health Impact Profile - Oral Symptom Scale; NOSE: Nasal Obstruction Symptom Evaluation; DMFT: Decayed, Missed, Filled Teeth; VPI: Velopharyngeal incompetence assessment; BERA: Brainstem Evoked Response Audiometry; OAE: Otoacoustic Emission; 1: Diagnostics performed are medical photographs, panoramic, lateral cephalogram and dental models; 2: Same diagnostics as ¹, but without dental models.

Figure 1 Overview of treatment protocols of the Erasmus Medical Center for the treatment of UCLAP. Light blue boxes in blue field indicate "protocol", green boxes indicate the additions by the Standard Set, which together with light blue boxes makes the "protocol+ICHOM". The white field presents optional, non-protocolized treatments offered after diagnosing specific problems or needs.

Statistical analysis

The observed healthcare use (including optional surgeries and treatments due to complications) was counted and related medical costs were calculated. Because it was not feasible to measure a 24-year care trajectory for each patient, patients were followed up to 8 years within the study period. Observed mean care use and costs per person year for each year of the treatment trajectory were calculated, and totaled to obtain overall healthcare use and costs of a full treatment trajectory from 0 to 24 years of age.

Consequently, the care and costs of the full treatment trajectory were broken down into six important phases based on the Standard Set time points for outcome measurements: 0-4 years (no additional outcome measurements), 5-7 years, 8-10 years, 11-13 years, 14-21 years (no additional outcome measurements), and 22-24 years of age.[6]

Subsequently, mean observed costs were compared to expected costs based on the “protocol” to treat patients with UCLAP. Further, the expected additional costs of “protocol+ICHOM” were described. All statistical analyses were performed using IBM SPSS Statistics version 24.0.[21] An overview of terminology used in this paper is presented in **Table 2**.

Terminology	**Description**
Healthcare services / use (cleft-related)	Medical consultations, diagnostic procedures, surgical procedures, and hospital admissions as registered by the cleft team of the Erasmus University Medical Center, see **Table 1** for more details
Observed costs	Calculated with the formula of costs = volume of observed healthcare service utilization *x* price of the healthcare product
Expected costs	Calculated with the formula of costs = volume of expected healthcare services based on the treatment protocol *x* price of the healthcare product
“Protocol”	The treatment protocol for cleft lip and palate that was employed before 2016
“Protocol+ICHOM”	The treatment protocol for cleft lip and palate employed from 2016 and onwards, including additional consultation and diagnostics introduced by the local implementation of the ICHOM Standard Set for Cleft Lip and Palate

Table 2 Overview of terminology with its descriptions as used in this paper.

Results

In total, 40 patients with UCLAP contributed 301 observed person years. Twenty-seven (67%) patients were male, 5 (13%) were adopted and 4 (10%) were diagnosed with a genetic syndrome.

Healthcare use

Table 3 highlights observed healthcare per age phase. Highest mean number of medical consultations was provided to patients of 8-10 years (n=7), and 11-13 years (n=8). The mean number of diagnostic procedures was highest at the age groups of 5-7 (n=4), 8-10 (n=4) and 11-13 (n=4). The highest mean number of surgical procedures performed was during the age of 0-4 (n=1) and 8-10 years (n=1), with the highest number of surgeries in the first year after birth (n=2) and a mean total of 10 surgical interventions over the course of 24 years (**Supplemental Material – Figure 1**). An overview of observed counts per person year for medical consultations and diagnostic procedures can be found in **Supplemental Material – Figure 2** and **Supplemental Material – Figure 3**, respectively.

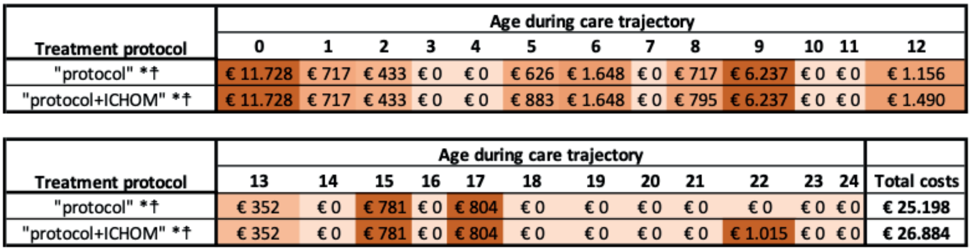
	0 - 4 years	5 - 7 years	8 - 10 years	11 - 13 years	14 - 21 years	22 - 24 years
Medical consultations	4	3	7	8	6	4
Diagnostic procedures	4	4	4	4	3	2
Surgical procedures	1	0	1	0	0	0

Table 3 Counts per person year of medical consultations, diagnostical and surgical procedures for the various age groups.

Observed costs and comparison with expected costs

The mean observed total costs for the treatment period from birth until 24 years (including optional, non-protocolized treatments due to complications) were 1.6 times higher (€40,859) compared to expected costs based on the "protocol" (€25,198). Mean total costs for observed care per patient based on a maximum of 8 years follow-up was €11,809 (range €2,616 – €33,323), with 50% of patients within the interquartile range (€6,513 – €14,831). This distribution was similar for the adopted and syndromic patients together. Observed mean costs per person year were €1,702. Highest mean observed costs were made in the first year after birth, and at the age of 11 years (€5,596 and €3,454 (ratio 1:0.6), respectively). Clustering the data into age groups of 0-4 and 11-13 years, observed costs were €2,681 and €2,383 (ratio 1:0.8), respectively (**Supplemental Material – Figure 4)**. Surgeries including hospital admissions accounted for 42% of total observed costs.

Based on the “protocol”, expected mean costs per person year were €1,050, and highest costs were expected in the first year after birth (€11,728) and at 9 years of age (€6,236) (ratio 1:0.5), while no costs were expected at 11 years of age (**Figure 2**). Surgical procedures including hospital admission accounted for 70% of total expected costs.



Treatment protocol	Age during care trajectory												
	0	1	2	3	4	5	6	7	8	9	10	11	12
"protocol" *†	€ 11.728	€ 717	€ 433	€ 0	€ 0	€ 626	€ 1.648	€ 0	€ 717	€ 6.237	€ 0	€ 0	€ 1.156
"protocol+ICHOM" *†	€ 11.728	€ 717	€ 433	€ 0	€ 0	€ 883	€ 1.648	€ 0	€ 795	€ 6.237	€ 0	€ 0	€ 1.490

Treatment protocol	Age during care trajectory												
	13	14	15	16	17	18	19	20	21	22	23	24	Total costs
"protocol" *†	€ 352	€ 0	€ 781	€ 0	€ 804	€ 0	€ 0	€ 0	€ 0	€ 0	€ 0	€ 0	€ 25.198
"protocol+ICHOM" *†	€ 352	€ 0	€ 781	€ 0	€ 804	€ 0	€ 0	€ 0	€ 0	€ 1.015	€ 0	€ 0	€ 26.884

Figure 2 Costs based on the “protocol” and “protocol+ICHOM” per time point of care. *Costs are based on 2019 prices only. †Costs of closure of hard palate is divided over 0, 6 and 9 years due to varying surgical timings.

The addition of the Standard Set to the treatment protocol resulted in an increase of €1.686 (7%) on total expected costs (€26,884). The additional team visit at 22 years accounted for 3% point of this increase.

Discussion

This study described healthcare utilization and costs of a long and complex treatment period for UCLAP and assessed the impact of the implementation of the ICHOM Standard Set on medical costs. In general, large variations between patients, between costs based on observed healthcare use and costs based on treatment protocols for cleft along 24 years of care were assessed; costs for the full treatment period from birth until young adulthood were 1.6 times as high as expected costs based on the treatment protocols. This is mainly due to many diagnostic and surgical procedures, such as speech therapy, grommet placement or septorhinoplasty, that are not routinely performed but are offered after diagnosing specific needs or wishes. Even though it is known that procedures such as orthognatic surgery, dental implant placement or septorhinoplasty benefit the countenance of patients as perceived by laypersons[22,23], the need to perform these procedures and its timing primarily depends on the patient’s feelings, concerns and wishes. Therefore, these procedures are defined as ‘optional’ in our local treatment procotol. Currently, the use of PROMs in clinical practice helps to identify areas of concern and act as a conversation starter during routine medical consultations to discuss the patient’s worries and wishes and related possible interventions.[18]

First year of life was most expensive

The high number of medical consultations and surgical procedures, including closure of the lip and soft palate, performed during the first year after birth are probable causes for these high expenditures. Previous micro-costing research by Abbott *et al.* presented similar results with costs ranging from $35,826 to $56,611 for various subtypes of cleft lip and palate for the first 18 months in life, and costs ranging from $10,426 to $16,115 in patients with cleft lip.[12,24] The majority of these first year of life costs stemmed from inpatient care, i.e. surgical procedures and hospital admissions.[12,24] Additionally, Boulet *et al.* reported that mean expenditures for children with a cleft in the USA were decreasing with increasing age; starting from $95,819 for infants to approximately $5,054 at 7 to 8 years of age.[25] Noteworthy, the costs of cleft treatments described by Boulet *et al*. and Abbott *et al.* were much higher compared to our results.[12,24] These differences may be explained by the fact that 1) data on costs were collected from (private) health insurance companies, 2) costs were not limited to cleft-related healthcare, 3) treatment protocols differ between hospitals, and 4) healthcare costs in general might be higher in the USA than in The Netherlands due to differences in healthcare organization and insurance strategies.

An unanticipated finding was that observed costs of the first year after birth were lower than expected based on the protocol, in contrast to the subsequent years in which costs were higher than expected. Interventions protocolized within the first year of life are sometimes spread over a longer period, due to planning difficulties, or late referrals. A similar pattern was seen at a later stage of care; higher expenditures were expected at 9 years of age based on the "protocol", due to the alveolar bone grafting procedure and orthodontic treatments. In practice, the alveolar bone grafting timing is dependent on the dental development status of the child, resulting in higher costs between 10 and 12 years.[26,27] Evaluating costs of a long period of complex care might benefit from clustering multiple years, since the start of an intervention may vary and treatments, such as orthodontics or speech therapy, may continue multiple months, or even years.

The effect of the Standard Set on medical costs

With the implementation of the Standard Set, additional speech, audiological and psychosocial screening, and an extra cleft team visit were introduced.[3,6,28] Aside these clinical encounters, patients of 8, 12 and 22 years were asked to complete PROMs at home prior their visit.[5,6] The use of PROMs could provide insight in a patient's perspective on their functioning and well-being, and detect concerns or problems that otherwise remain undiscussed. Tackling healthcare problems and improving quality of life early on, could potentially reduce complication rates and treatment costs in the long run. Even if the Standard Set implementation would lead to increasing expenditures, by improving patient's outcomes, an increase in value and concomitantly cost-effectiveness could still be reached. Because of the complex character of cleft lip and palate and the need for long-term care into young adulthood, patience is asked from clinicians, researchers and policymakers before the cost-effectiveness and potential value-improvement can be reliably examined. Meanwhile, measuring outcomes could be utilized to improve patient-centered care, shared decision-making, and local quality improvement endeavors.[18,29]

Heterogeneity complicates the understanding of healthcare use patterns

This research highlights the heterogeneity in healthcare use and medical costs for patients with a cleft lip and palate, in which care consumption and costs varied widely with 50% of patients outside the interquartile range. Surgical procedures were expected to be responsible for 70% of medical care costs based on the protocol, however, actual surgical costs were found to be much lower (42%), suggesting that additional consultations and diagnostics are more often needed than expected. Consequently, solely relying on cost estimates based on a cleft treatment protocol to reform payment strategies or to lead negotiations between hospitals and health-insurance companies should be done with caution. Further, understanding the patterns of healthcare use aids determining most efficient treatment pathways. For example, knowledge on clinician's consultation burden could guide reorganization of the cleft protocol and team; it might be needed for a specialist to be more (or less) often available for consultations, or at altered time points during the treatment trajectory. Further research is needed to specify predictors for variability in healthcare consumption, such as cleft type, family circumstances, and socio-economic status, to target individuals in need of more extensive care and enabling risk stratification.[30] The methodology described in this paper can be a useful first step in mapping and gaining insights in healthcare use and medical costs on a local level.

Strengths and limitations

A unique point of this study is the evaluation of healthcare utilization and costs of a challenging, complex treatment trajectory for cleft lip and palate with a long follow-up time of 8 years to reconstruct a full treatment trajectory of 24 years. In addition, exploring the additional costs due to the implementation of the Standard Set for cleft care has not been done before. In contrast to previous cost studies focusing on aggregated data, we presented healthcare use and costs per individual. The latter was possible because our sample size was relatively small, which at the same time, limited the possibility to adjust for potential confounders, such as adoption status or presence of a genetic syndrome.

For the calculation of costs, in-hospital pricing for services and interventions was used. These prices depend on both the total volume of care delivered, and on a department's own preferences on how to attribute costs to healthcare items. For example, costs for administrative personnel, or utilization of rooms for outpatient clinic visits, can be attributed to a general overarching cost item within a department, or to one specific cost item such as a medical consultation. This approach results in price differences between years, between departments within one hospital, but potentially also between hospitals and countries. As a result, costs should be interpreted as estimates rather than exact numbers, and extrapolation of costs to other cleft care practices should be done with caution.

Further, this study only included healthcare use and direct costs from an academic healthcare provider's perspective. We were unable to include costs such as out-of-pocket expenses by patients, medication costs, costs of out-of-hospital treatments such as speech and language therapy, psychosocial care or dental and orthodontic care, travel costs, loss in work productivity of parents, and costs of administrative personnel.[14,19] Also, costs for the implementation of the Standard Set itself were not incorporated. Therefore, costs described in this paper are most likely an underestimation of the true economic burden of cleft care.

In addition, this study was conducted in a specialized cleft center where various medical specialists work together in an integrated practice unit ('cleft team'). However, in some geographic areas, cleft care is not provided by such a coordinated team but rather by individual clinicians, limiting the generalizability of our findings and hampering payment reform strategies.[31]

Performing a cost-effectiveness evaluation was hampered, because outcome measures were not routinely collected in clinical practice before the implementation of the Standard Set. Consequently, we cannot draw any conclusions whether extra costs wage against the effects of the implementation, for example in terms of patient satisfaction with treatment or better patient outcomes and quality of care. This issue deserves further study and could be a promising opportunity for centers who are planning to implement the Standard Set in their clinical practice.

In conclusion, there is a large variety in healthcare use and medical costs between patients with a UCLAP and throughout the cleft treatment trajectory, with highest costs in the first year of life. Observed costs for the treatment from birth until young adulthood were 1.6 times as high as costs based on protocols, due to a wide range of secondary diagnostics and surgeries performed. Surgical procedures were found to be main cost drivers, while the increase of medical costs due to the implementation of additional assessments, as defined by the ICHOM Standard Set for Cleft Lip and Palate, was 7%.

Acknowledgements

Authors would like to thank Job J. de Ridder for his help in data acquisition and for developing the graphics for this manuscript.

Financial disclosures

Implementation of the ICHOM Standard Set for Cleft Care at the Erasmus University Medical Center was financially supported by Stichting Theia and CZ Health Insurance.

Conflicts of interest

None.

References

1. Porter ME, Teisberg, E. *Redefining healthcare; creating value-based competition on results.* Harvard Business Review Press; 2006.
2. Porter ME. What is value in health care? *N Engl J Med.* 2010;363(26):2477-2481.
3. Allori AC, Kelley T, Meara JG, et al. A Standard Set of Outcome Measures for the Comprehensive Appraisal of Cleft Care. *Cleft Palate Craniofac J.* 2017;54(5):540-554.
4. Arora J, Haj M. Implementing ICHOM's Standard Sets of Outcomes: Cleft Lip and Palate at Erasmus University Medical Centre in the Netherlands. International Consortium for Health Outcomes Measurement (ICHOM). www.ichom.org. Published 2016. Accessed December 1, 2020.
5. Haj M, de Gier HHW, van Veen-van der Hoek M, et al. [Improving care for cleft lip and palate patients: uniform and patient-orientated outcome measures]. *Ned Tijdschr Tandheelkd.* 2018;125(2):70-75.
6. International Consortium for Health Outcomes Measurement (ICHOM). Data collection reference guide. https://ichom.org/files/medical-conditions/cleft-lip-palate/cleft-lip-palate-reference-guide.pdf. Published 2018. Accessed December 1, 2020.
7. International Perinatal Database of Typical Oral Clefts Working Group. Prevalence at birth of cleft lip with or without cleft palate: data from the International Perinatal Database of Typical Oral Clefts. *Cleft Palate Craniofac J.* 2011;48(1):66-81.
8. Kirschner RE, LaRossa D. Cleft lip and palate. *Otolaryngol Clin North Am.* 2000;33(6):1191-1215, v-vi.
9. Fadeyibi IO, Coker OA, Zacchariah MP, Fasawe A, Ademiluyi SA. Psychosocial effects of cleft lip and palate on Nigerians: the Ikeja-Lagos experience. *J Plast Surg Hand Surg.* 2012;46(1):13-18.
10. Mahboubi H, Truong A, Pham NS. Prevalence, demographics, and complications of cleft palate surgery. *Int J Pediatr Otorhinolaryngol.* 2015;79(6):803-807.
11. Ligh CA, Fox JP, Swanson J, Yu JW, Taylor JA. Not all clefts are created equal: patterns of hospital-based care use among children with cleft lip and palate within 4 years of initial surgery. *Plast Reconstr Surg.* 2016;137(6):990e-998e.
12. Abbott MM, Meara JG. A microcosting approach for isolated, unilateral cleft lip care in the first year of life. *Plast Reconstr Surg.* 2011;127(1):333-339.
13. Holzmer S, Davila A, Martin MC. Cost utility analysis of staged versus single-stage cleft lip and palate repair. *Ann Plast Surg.* 2020;84(5S Suppl 4):S300-S306.
14. Wehby GL, Cassell CH. The impact of orofacial clefts on quality of life and healthcare use and costs. *Oral Dis.* 2010;16(1):3-10.
15. Nguyen C, Hernandez-Boussard T, Davies SM, Bhattacharya J, Khosla RK, Curtin CM. Cleft palate surgery: an evaluation of length of stay, complications, and costs by hospital type. *Cleft Palate Craniofac J.* 2014;51(4):412-419.
16. Shaw WC, Brattstrom V, Molsted K, Prahl-Andersen B, Roberts CT, Semb G. The Eurocleft study: intercenter study of treatment outcome in patients with complete cleft lip and palate. Part 5: discussion and conclusions. *Cleft Palate Craniofac J.* 2005;42(1):93-98.
17. Shaw WC, Semb G, Nelson P, et al. The Eurocleft project 1996-2000: overview. *J Craniomaxillofac Surg.* 2001;29(3):131-140; discussion 141-132.
18. Apon I, Rogers-Vizena CR, Koudstaal MJ, et al. Barriers and Facilitators to the International Implementation of Standardized Outcome Measures in Clinical Cleft Practice. *Cleft Palate Craniofac J.* 2021:1055665621997668.

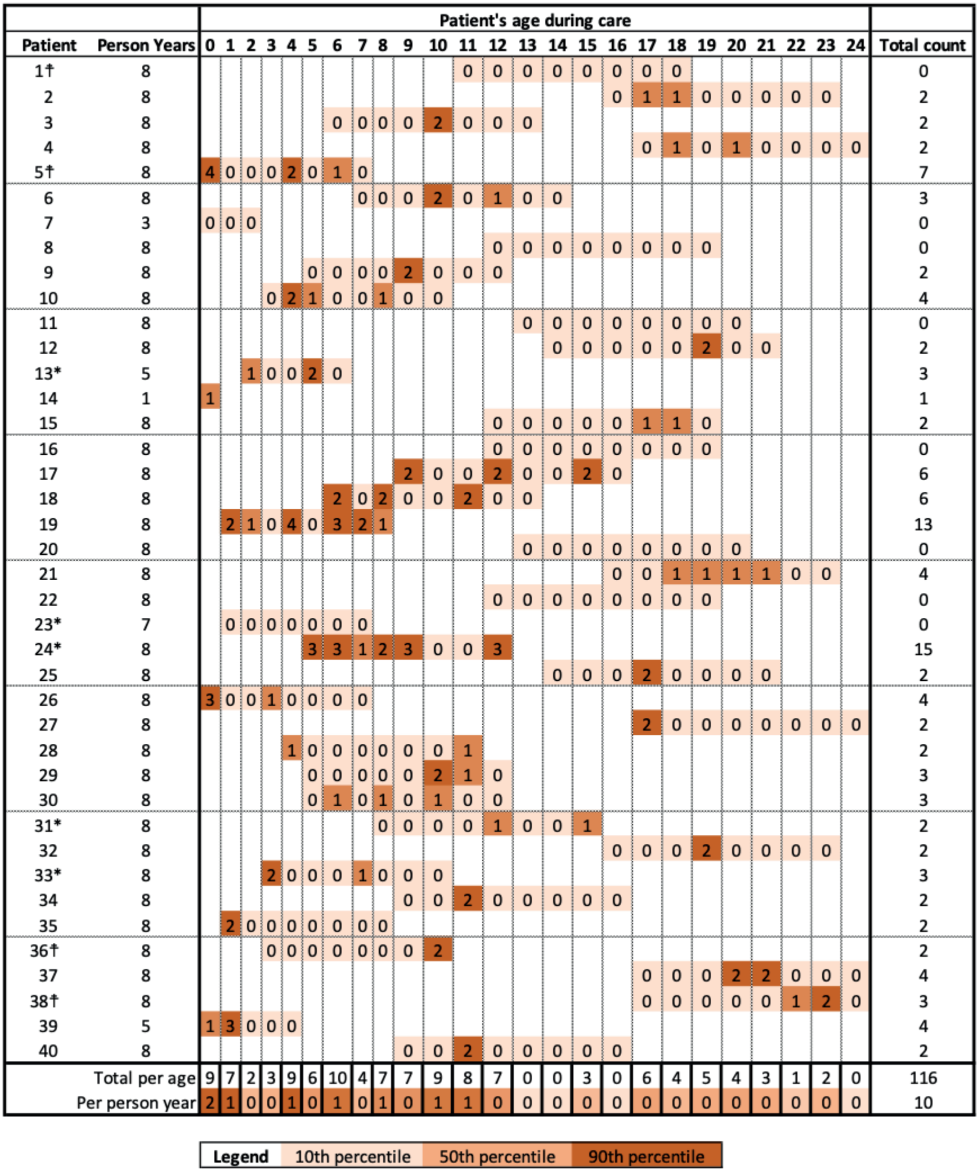
19. Zorginstituut Nederland. Richtlijn voor het uitvoeren van economische evaluaties in de gezondheidszorg. https://www.zorginstituutnederland.nl/over-ons/publicaties/publicatie/2016/02/29/richtlijn-voor-het-uitvoeren-van-economische-evaluaties-in-de-gezondheidszorg. Published 2016. Accessed April 1, 2020.
20. Nederlandse Zorgautoriteiten. Kostenbedragen van DBC zorgproducten. https://www.nza.nl/documenten/vragen-en-antwoorden/wat-is-het-indexcijfer-voor-kostenbedragen-van-dbczorgproducten. Published 2020. Accessed December 1, 2020.
21. *IBM Corp. IBM SPSS Statistics for Windows, Version 24.0* [computer program]. Armonk, NY: IBM Corp.; 2016.
22. Lin LO, Zhang RS, Mazzaferro DM, et al. Influence of Repaired Cleft Lip and Palate on Layperson Perception following Orthognathic Surgery. *Plast Reconstr Surg.* 2018;142(4):1012-1022.
23. Posnick JC, Susarla SM, Kinard BE. Reconstruction of residual cleft nasal deformities in adolescents: Effects on social perceptions. *J Craniomaxillofac Surg.* 2019;47(9):1414-1419.
24. Abbott MM, Rosen H, Kupfer P, Meara JG. Measuring value at the provider level in the management of cleft lip and palate patients. *Ann Plast Surg.* 2014;72(3):312-317.
25. Boulet SL, Grosse SD, Honein MA, Correa-Villasenor A. Children with orofacial clefts: health-care use and costs among a privately insured population. *Public Health Rep.* 2009;124(3):447-453.
26. Mink van der Molen AB, van Breugel JMM, Janssen NG, et al. Clinical Practice Guidelines on the Treatment of Patients with Cleft Lip, Alveolus, and Palate: An Executive Summary. *J Clin Med.* 2021;10(21).
27. Nederlandse Vereniging voor Plastische Chirurgie (NVPC). Behandeling van patienten met een schisis. https://richtlijnendatabase.nl/?query=Behandeling+van+patienten+met+een+schisis&specialism=. Published 2018. Accessed December 1, 2020.
28. Bittar PG, Carlson AR, Mabie-DeRuyter A, Marcus JR, Allori AC. Implementation of a standardized data-collection system for comprehensive appraisal of cleft care. *Cleft Palate Craniofac J.* 2018;55(10):1382-1390.
29. Ramirez JP, Rogers-Vizena CR, Koudstaal MJ, et al. Whitepaper: Implementation of the ICHOM Standard Set for Cleft Lip and/or Palate Across Four Centers. 2021. https://conference.ichom.org/wp-content/uploads/2021/09/24915-CLP-Cleft-Lip-whitepaper.pdf. Accessed January 31, 2022.
30. Agarwal R, Liao JM, Gupta A, Navathe AS. The Impact Of Bundled Payment On Health Care Spending, Utilization, And Quality: A Systematic Review. *Health Aff (Millwood).* 2020;39(1):50-57.
31. Ahsanuddin S SF, Lu J, Sanati-Mehrizy P, Taub PJ. Considerations for Payment Bundling in Cleft Care. *Plast Reconstr Surg.* 2021;47(4):927-932.

Supplemental Material

Table 1 Dutch price index percentages for healthcare products[1].

Year	Index percentages
2019	3,75%
2018	2,87%
2017	1,92%
2016	0,26%
2015	1,15%
2014	3,14%
2013	1,93%
2012	3,16%

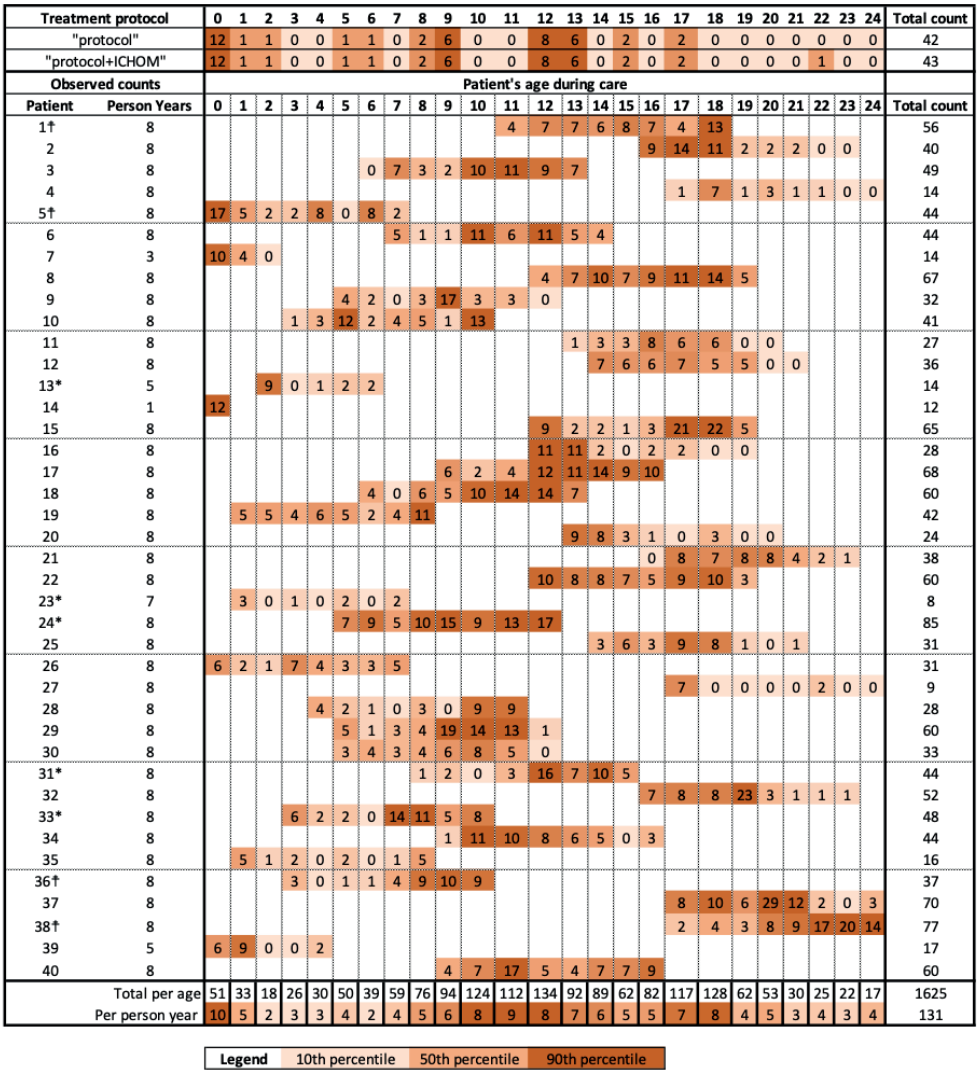
1. Nederlandse Zorgautoriteiten. Kostenbedragen van DBC zorgproducten. https://www.nza.nl/documenten/vragen-en-antwoorden/wat-is-het-indexcijfer-voor-kostenbedragen-van-dbczorgproducten. Published 2020. Accessed December 1, 2020.



		Patient's age during care																									
Patient	Person Years	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	Total count
1†	8												0	0	0	0	0	0	0	0							0
2	8																	0	1	1	0	0	0	0	0		2
3	8							0	0	0	0	2	0	0	0												2
4	8																		0	1	0	1	0	0	0	0	2
5†	8	4	0	0	0	2	0	1	0																		7
6	8								0	0	0	2	0	1	0	0											3
7	3	0	0	0																							0
8	8													0	0	0	0	0	0	0	0						0
9	8						0	0	0	0	2	0	0	0													2
10	8				0	2	1	0	0	1	0	0															4
11	8														0	0	0	0	0	0	0	0					0
12	8															0	0	0	0	0	2	0	0				2
13*	5			1	0	0	2	0																			3
14	1	1																									1
15	8													0	0	0	0	0	1	1	0						2
16	8													0	0	0	0	0	0	0	0						0
17	8										2	0	0	2	0	0	2	0									6
18	8							2	0	2	0	0	2	0	0												6
19	8		2	1	0	4	0	3	2	1																	13
20	8														0	0	0	0	0	0	0	0					0
21	8																	0	0	1	1	1	1	0	0		4
22	8													0	0	0	0	0	0	0	0						0
23*	7		0	0	0	0	0	0	0																		0
24*	8						3	3	1	2	3	0	0	3													15
25	8															0	0	0	2	0	0	0	0				2
26	8	3	0	0	1	0	0	0	0																		4
27	8																		2	0	0	0	0	0	0	0	2
28	8					1	0	0	0	0	0	0	1														2
29	8						0	0	0	0	0	2	1	0													3
30	8						0	1	0	1	0	1	0	0													3
31*	8									0	0	0	0	1	0	0	1										2
32	8																	0	0	0	2	0	0	0	0		2
33*	8				2	0	0	0	1	0	0	0															3
34	8										0	0	2	0	0	0	0	0									2
35	8		2	0	0	0	0	0	0	0																	2
36†	8				0	0	0	0	0	0	0	2															2
37	8																		0	0	0	2	2	0	0	0	4
38†	8																		0	0	0	0	0	1	2	0	3
39	5	1	3	0	0	0																					4
40	8										0	0	2	0	0	0	0	0									2
	Total per age	9	7	2	3	9	6	10	4	7	7	9	8	7	0	0	3	0	6	4	5	4	3	1	2	0	116
	Per person year	2	1	0	0	1	0	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	10

Legend	10th percentile	50th percentile	90th percentile

Figure 1 Observed **surgical procedures** for each patient, in total and per person year. In total, 4 surgical procedures were counted for both "protocol" and "protocol+ICHOM"; all other surgical procedures are optional. Multiple procedures could be combined in one operative setting, e.g. closure of soft palate combined with bilateral grommet placement (counts for 3 procedures in 1 setting). *Adopted child. †Presence of genetic syndrome.



Treatment protocol		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	Total count
"protocol"		12	1	1	0	0	1	1	0	2	6	0	0	8	6	0	2	0	2	0	0	0	0	0	0	0	42
"protocol+ICHOM"		12	1	1	0	0	1	1	0	2	6	0	0	8	6	0	2	0	2	0	0	0	0	1	0	0	43
Observed counts		Patient's age during care																									
Patient	Person Years	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	Total count
1†	8												4	7	7	6	8	7	4	13							56
2	8																	9	14	11	2	2	2	0	0		40
3	8							0	7	3	2	10	11	9	7												49
4	8																		1	7	1	3	1	1	0	0	14
5†	8	17	5	2	2	8	0	8	2																		44
6	8								5	1	1	11	6	11	5	4											44
7	3	10	4	0																							14
8	8													4	7	10	7	9	11	14	5						67
9	8						4	2	0	3	17	3	3	0													32
10	8				1	3	12	2	4	5	1	13															41
11	8														1	3	3	8	6	6	0	0					27
12	8															7	6	6	7	5	5	0	0				36
13*	5			9	0	1	2	2																			14
14	1	12																									12
15	8													9	2	2	1	3	21	22	5						65
16	8													11	11	2	0	2	2	0	0						28
17	8										6	2	4	12	11	14	9	10									68
18	8							4	0	6	5	10	14	14	7												60
19	8		5	5	4	6	5	2	4	11																	42
20	8														9	8	3	1	0	3	0	0					24
21	8																	0	8	7	8	8	4	2	1		38
22	8													10	8	8	7	5	9	10	3						60
23*	7		3	0	1	0	2	0	2																		8
24*	8						7	9	5	10	15	9	13	17													85
25	8															3	6	3	9	8	1	0	1				31
26	8	6	2	1	7	4	3	3	5																		31
27	8																		7	0	0	0	0	2	0	0	9
28	8					4	2	1	0	3	0	9	9														28
29	8						5	1	3	4	19	14	13	1													60
30	8						3	4	3	4	6	8	5	0													33
31*	8									1	2	0	3	16	7	10	5										44
32	8																	7	8	8	23	3	1	1	1		52
33*	8				6	2	2	0	14	11	5	8															48
34	8										1	11	10	8	6	5	0	3									44
35	8		5	1	2	0	2	0	1	5																	16
36†	8				3	0	1	1	4	9	10	9															37
37	8																		8	10	6	29	12	2	0	3	70
38†	8																		2	4	3	8	9	17	20	14	77
39	5	6	9	0	0	2																					17
40	8										4	7	17	5	4	7	7	9									60
Total per age		51	33	18	26	30	50	39	59	76	94	124	112	134	92	89	62	82	117	128	62	53	30	25	22	17	1625
Per person year		10	5	2	3	3	4	2	4	5	6	8	9	8	7	6	5	5	7	8	4	5	3	4	3	4	131

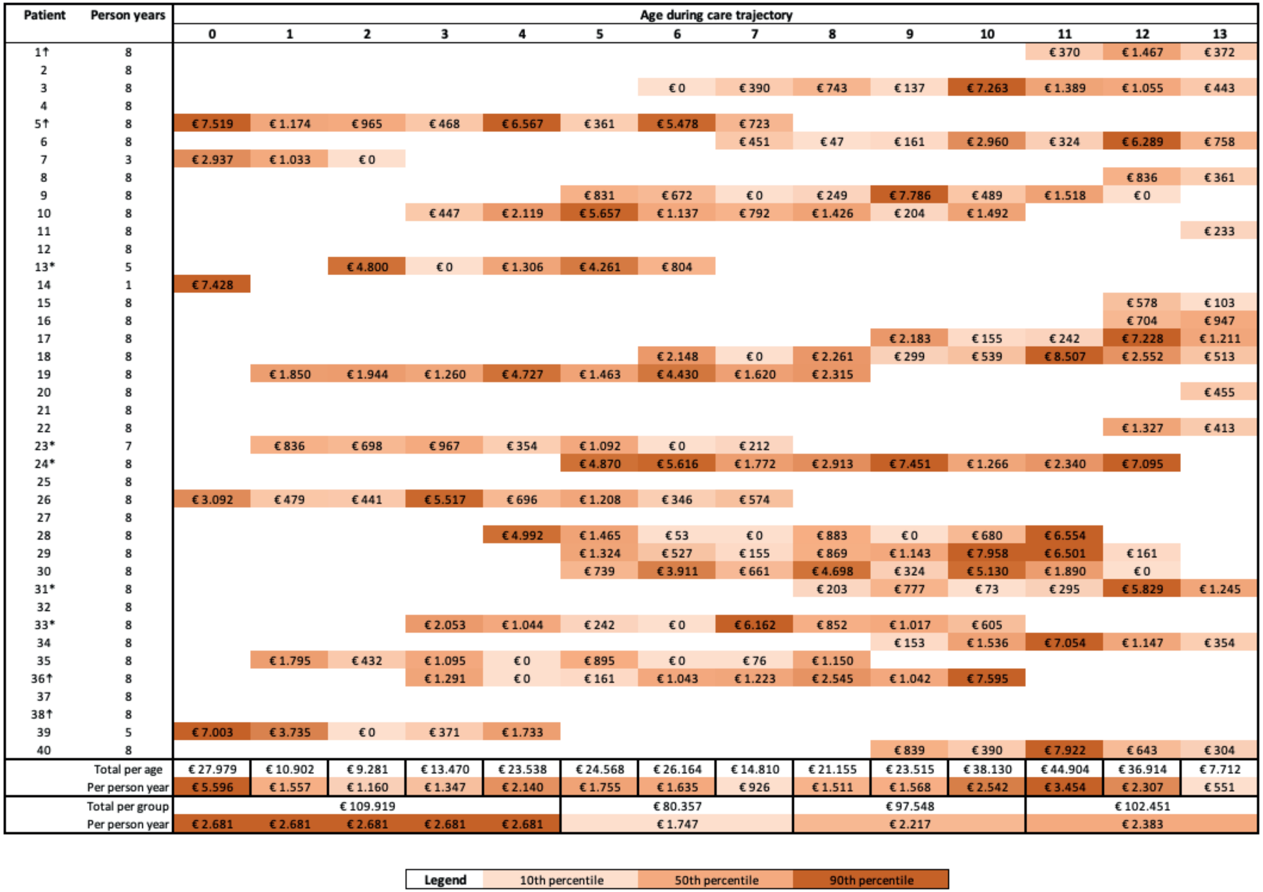
Legend	10th percentile	50th percentile	90th percentile

Figure 2 Observed **medical consultations** and utilization based on the protocols for each patient, in total and per person year. *Adopted child. †Presence of genetic syndrome.

Treatment protocol		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	Total count
"protocol"		4	3	3	0	0	0	4	0	1	5	0	0	4	0	0	3	0	4	0	0	0	0	0	0	0	31
"protocol+ICHOM"		4	3	3	0	0	2	4	0	1	5	0	0	7	0	0	3	0	4	0	0	0	0	2	0	0	38
Observed counts		Patient's age during care																									
Patient	Person Years	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	Total count
1†	8												6	0	0	1	0	0	3	14							24
2	8																	5	4	5	0	1	6	0	0		21
3	8							0	1	3	1	2	6	1	1												15
4	8																		1	1	0	0	0	0	0	0	2
5†	8	5	6	4	4	2	1	10	4																		36
6	8								2	0	4	6	0	8	4	0											24
7	3	2	2	0																							4
8	8													3	0	3	1	1	3	1	2						14
9	8						11	2	0	3	4	0	8	0													28
10	8				3	6	8	5	7	5	3	4															41
11	8														0	4	2	4	0	6	0	0					16
12	8															0	3	0	5	1	0	0	1				10
13*	5			1	0	4	1	4																			10
14	1	5																									5
15	8													5	0	4	0	6	3	7	5						30
16	8													1	5	0	0	4	0	0	0						10
17	8										8	1	1	9	8	5	1	3									36
18	8							2	0	6	1	0	11	6	2												28
19	8		7	7	5	3	8	3	3	10																	46
20	8														0	4	0	4	0	4	0	0					12
21	8																	0	5	6	1	0	2	0	1		15
22	8													1	0	4	2	0	5	1	0						13
23*	7		3	2	2	1	4	0	2																		14
24*	8						12	8	7	8	10	8	9	7													69
25	8															0	4	4	1	6	0	0	5				20
26	8	6	3	3	5	3	5	5	4																		34
27	8																		5	0	0	0	0	6	0	0	11
28	8					2	6	0	0	5	0	3	9														25
29	8						9	5	0	5	4	6	8	1													38
30	8						4	6	3	9	0	2	15	0													39
31*	8									4	6	2	4	6	4	2	4										32
32	8																	4	5	3	14	3	0	5	0		34
33*	8				9	4	4	0	8	5	6	1															37
34	8										4	3	4	6	1	4	0	4									26
35	8		4	3	5	0	4	0	0	6																	22
36†	8				4	0	4	2	4	5	3	2															24
37	8																		4	0	4	18	9	1	0	0	36
38†	8																		4	0	0	2	9	10	5	2	32
39	5	2	8	0	1	4																					15
40	8										5	1	6	2	3	1	0	5									23
	Total per age	20	33	20	38	29	81	52	45	74	59	41	87	56	28	32	17	44	48	55	26	24	32	22	6	2	971
	Per person year	4	5	3	4	3	6	3	3	5	4	3	7	4	2	2	1	3	3	3	2	2	4	3	1	1	79

Legend	10th percentile	50th percentile	90th percentile

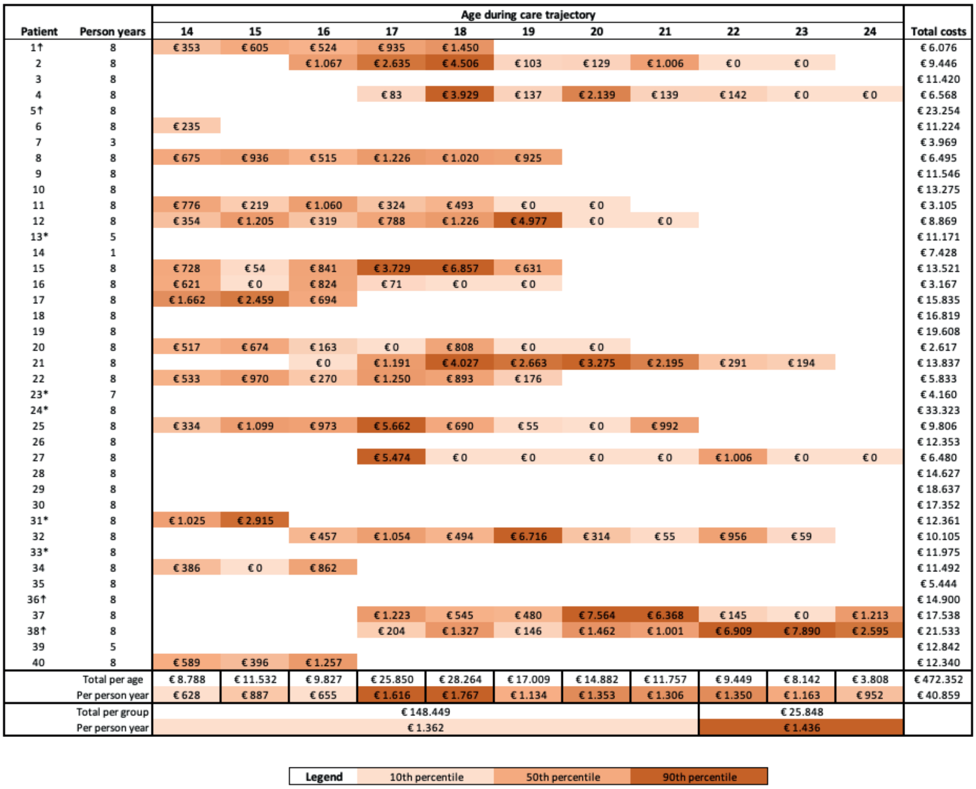
Figure 3 Observed **diagnostic procedures** and utilization based on the protocols for each patient**,** in total and per person year. *Adopted child. ǂPresence of genetic syndrome.



Patient	Person years	Age during care trajectory													
		0	1	2	3	4	5	6	7	8	9	10	11	12	13
1†	8												€ 370	€ 1.467	€ 372
2	8														
3	8							€ 0	€ 390	€ 743	€ 137	€ 7.263	€ 1.389	€ 1.055	€ 443
4	8														
5†	8	€ 7.519	€ 1.174	€ 965	€ 468	€ 6.567	€ 361	€ 5.478	€ 723						
6	8								€ 451	€ 47	€ 161	€ 2.960	€ 324	€ 6.289	€ 758
7	3	€ 2.937	€ 1.033	€ 0											
8	8													€ 836	€ 361
9	8						€ 831	€ 672	€ 0	€ 249	€ 7.786	€ 489	€ 1.518	€ 0	
10	8				€ 447	€ 2.119	€ 5.657	€ 1.137	€ 792	€ 1.426	€ 204	€ 1.492			
11	8														€ 233
12	8														
13*	5			€ 4.800	€ 0	€ 1.306	€ 4.261	€ 804							
14	1	€ 7.428													
15	8													€ 578	€ 103
16	8													€ 704	€ 947
17	8										€ 2.183	€ 155	€ 242	€ 7.228	€ 1.211
18	8							€ 2.148	€ 0	€ 2.261	€ 299	€ 539	€ 8.507	€ 2.552	€ 513
19	8		€ 1.850	€ 1.944	€ 1.260	€ 4.727	€ 1.463	€ 4.430	€ 1.620	€ 2.315					
20	8														€ 455
21	8														
22	8													€ 1.327	€ 413
23*	7		€ 836	€ 698	€ 967	€ 354	€ 1.092	€ 0	€ 212						
24*	8						€ 4.870	€ 5.616	€ 1.772	€ 2.913	€ 7.451	€ 1.266	€ 2.340	€ 7.095	
25	8														
26	8	€ 3.092	€ 479	€ 441	€ 5.517	€ 696	€ 1.208	€ 346	€ 574						
27	8														
28	8					€ 4.992	€ 1.465	€ 53	€ 0	€ 883	€ 0	€ 680	€ 6.554		
29	8						€ 1.324	€ 527	€ 155	€ 869	€ 1.143	€ 7.958	€ 6.501	€ 161	
30	8						€ 739	€ 3.911	€ 661	€ 4.698	€ 324	€ 5.130	€ 1.890	€ 0	
31*	8									€ 203	€ 777	€ 73	€ 295	€ 5.829	€ 1.245
32	8														
33*	8				€ 2.053	€ 1.044	€ 242	€ 0	€ 6.162	€ 852	€ 1.017	€ 605			
34	8										€ 153	€ 1.536	€ 7.054	€ 1.147	€ 354
35	8		€ 1.795	€ 432	€ 1.095	€ 0	€ 895	€ 0	€ 76	€ 1.150					
36†	8				€ 1.291	€ 0	€ 161	€ 1.043	€ 1.223	€ 2.545	€ 1.042	€ 7.595			
37	8														
38†	8														
39	5	€ 7.003	€ 3.735	€ 0	€ 371	€ 1.733									
40	8										€ 839	€ 390	€ 7.922	€ 643	€ 304
	Total per age	€ 27.979	€ 10.902	€ 9.281	€ 13.470	€ 23.538	€ 24.568	€ 26.164	€ 14.810	€ 21.155	€ 23.515	€ 38.130	€ 44.904	€ 36.914	€ 7.712
	Per person year	€ 5.596	€ 1.557	€ 1.160	€ 1.347	€ 2.140	€ 1.755	€ 1.635	€ 926	€ 1.511	€ 1.568	€ 2.542	€ 3.454	€ 2.307	€ 551
	Total per group	€ 109.919					€ 80.357			€ 97.548			€ 102.451		
	Per person year	€ 2.681	€ 2.681	€ 2.681	€ 2.681	€ 2.681	€ 1.747			€ 2.217			€ 2.383		

Legend 10th percentile 50th percentile 90th percentile

Figure 4 Observed costs for each patient, in total, per person year and per phase. *Adopted child. ‡Presence of genetic syndrome. *(Figure continues on next page).*



		Age during care trajectory											
Patient	Person years	14	15	16	17	18	19	20	21	22	23	24	Total costs
1†	8	€ 353	€ 605	€ 524	€ 935	€ 1.450							€ 6.076
2	8			€ 1.067	€ 2.635	€ 4.506	€ 103	€ 129	€ 1.006	€ 0	€ 0		€ 9.446
3	8												€ 11.420
4	8				€ 83	€ 3.929	€ 137	€ 2.139	€ 139	€ 142	€ 0	€ 0	€ 6.568
5†	8												€ 23.254
6	8	€ 235											€ 11.224
7	3												€ 3.969
8	8	€ 675	€ 936	€ 515	€ 1.226	€ 1.020	€ 925						€ 6.495
9	8												€ 11.546
10	8												€ 13.275
11	8	€ 776	€ 219	€ 1.060	€ 324	€ 493	€ 0	€ 0					€ 3.105
12	8	€ 354	€ 1.205	€ 319	€ 788	€ 1.226	€ 4.977	€ 0	€ 0				€ 8.869
13*	5												€ 11.171
14	1												€ 7.428
15	8	€ 728	€ 54	€ 841	€ 3.729	€ 6.857	€ 631						€ 13.521
16	8	€ 621	€ 0	€ 824	€ 71	€ 0	€ 0						€ 3.167
17	8	€ 1.662	€ 2.459	€ 694									€ 15.835
18	8												€ 16.819
19	8												€ 19.608
20	8	€ 517	€ 674	€ 163	€ 0	€ 808	€ 0	€ 0					€ 2.617
21	8			€ 0	€ 1.191	€ 4.027	€ 2.663	€ 3.275	€ 2.195	€ 291	€ 194		€ 13.837
22	8	€ 533	€ 970	€ 270	€ 1.250	€ 893	€ 176						€ 5.833
23*	7												€ 4.160
24*	8												€ 33.323
25	8	€ 334	€ 1.099	€ 973	€ 5.662	€ 690	€ 55	€ 0	€ 992				€ 9.806
26	8												€ 12.353
27	8				€ 5.474	€ 0	€ 0	€ 0	€ 0	€ 1.006	€ 0	€ 0	€ 6.480
28	8												€ 14.627
29	8												€ 18.637
30	8												€ 17.352
31*	8	€ 1.025	€ 2.915										€ 12.361
32	8			€ 457	€ 1.054	€ 494	€ 6.716	€ 314	€ 55	€ 956	€ 59		€ 10.105
33*	8												€ 11.975
34	8	€ 386	€ 0	€ 862									€ 11.492
35	8												€ 5.444
36†	8												€ 14.900
37	8				€ 1.223	€ 545	€ 480	€ 7.564	€ 6.368	€ 145	€ 0	€ 1.213	€ 17.538
38†	8				€ 204	€ 1.327	€ 146	€ 1.462	€ 1.001	€ 6.909	€ 7.890	€ 2.595	€ 21.533
39	5												€ 12.842
40	8	€ 589	€ 396	€ 1.257									€ 12.340
	Total per age	€ 8.788	€ 11.532	€ 9.827	€ 25.850	€ 28.264	€ 17.009	€ 14.882	€ 11.757	€ 9.449	€ 8.142	€ 3.808	€ 472.352
	Per person year	€ 628	€ 887	€ 655	€ 1.616	€ 1.767	€ 1.134	€ 1.353	€ 1.306	€ 1.350	€ 1.163	€ 952	€ 40.859
	Total per group	€ 148.449								€ 25.848			
	Per person year	€ 1.362								€ 1.436			

Legend	10th percentile	50th percentile	90th percentile

Figure 4 ***Continued*** Observed costs for each patient, in total, per person year and per phase. *Adopted child. †Presence of genetic syndrome.

Chapter 9

General Discussion

General discussion

The overall aim of this thesis was to explore how to optimize the measurement and implementation of patient-reported outcomes in clinical cleft practice. As a designated outcomes framework, the ICHOM Standard Set for Cleft Lip and Palate was chosen. Various research methodologies were applied to answer the following research questions:

1. How can we optimize the measurement of patient-reported outcomes in the ICHOM Standard Set for Cleft Lip and Palate?
    a. How is the psychometric performance and concept coverage of the patient-reported outcome measures of the ICHOM Standard Set for Cleft Lip and Palate?
    b. How can we maximize information while reducing burden when measuring psychosocial function within the ICHOM Standard Set for Cleft Lip and Palate?
    c. What is the external validity of the CLEFT-Q Computerized Adaptive Test (CAT) in patients with cleft lip and palate?

2. How can we optimize the implementation of the ICHOM Standard Set for Cleft Lip and Palate in clinical cleft care?
    a. What are facilitators and barriers to the implementation of the ICHOM Standard Set for Cleft Lip and Palate in clinical practice?
    b. What are the healthcare use and medical costs patterns of clinical cleft care and how is this influenced by the use of the ICHOM Standard Set for Cleft lip and Palate?

This chapter reflects on the key findings of the included studies, which are also summarized in **Table 1**. The implications of the findings, recommendations and future research ideas (**Table 2 and 3**) are discussed for both the measurement studies as for the implementation studies. In the last paragraph, the conclusions of this thesis are presented.

Research questions	Key findings
Part I Measurement challenges	
How is the psychometric performance and concept coverage of the patient-reported outcome measures of the ICHOM Standard Set for Cleft Lip and Palate?	The psychometric parameters of 9 CLEFT-Q scales, NOSE and COHIP-OSS instruments and Intelligibility in Context Scale administered to 714 patients with CL/P were analyzed using Rasch measurement theory. The patient- and parent-reported components within the facial appearance, psychosocial function, and speech domains are valid measures; however, the facial function and oral health domains are not sufficiently covered by the CLEFT-Q eating and drinking, NOSE, and COHIP-OSS instruments.
How can we maximize information while reducing burden when measuring psychosocial function within the ICHOM Standard Set for Cleft Lip and Palate?	Correlational and regression analyses were performed on prospectively collected data from 3,067 patients treated at 3 specialized cleft centers or participating in a large CLEFT-Q validation study, categorized into 5 time-points of measurement: 8-9, 10-13, 14-16, 17-19, and 20-22 years. As the CLEFT-Q social function showed strong correlations with both school and psychological function, its additional value for measuring psychosocial function within the Standard Set is limited.
What is the external validity of the Computerized Adaptive Testing (CAT) version of the CLEFT-Q scales in the ICHOM Standard Set for Cleft Lip and Palate?	CATs were calibrated and validated with Rasch measurement theory, using full-length responses of 8 CLEFT-Q scales collected in cross-sectional studies between October 2014 and April 2019 (2970 patients in total). The user interface of the CAT platform was prospectively piloted and interviews were conducted to explore end-user experiences. The CAT assessments reduced full-length CLEFT-Q scores accurately from 76 to 59 items. Partial credit Rasch models and graded response models produced very similar CAT scores. The platform was perceived to improve clinical communication and facilitate shared decision-making.
Part II Implementation challenges	
What are facilitators and barriers to the implementation of the ICHOM Standard Set for Cleft Lip and Palate in clinical practice?	Thematic content analyses of exploratory surveys and in depth-interviews revealed common facilitators and barriers to implementation at all sites. Teams reach patients either via email or during the clinic visit to capture patient-reported outcomes. Adopting routine data collection is enhanced by aligning priorities at the organizational and cleft team level. Streamlining workflows and developing an efficient data collection platform are necessary early on, followed by pilot testing or stepwise implementation. Regular meetings and financial resources are crucial for implementing, sustaining, analyzing collected data, and providing feedback to healthcare professionals and patients. Fostering patient-centered care was articulated as a positive outcome, whereas time presented challenges across all dimensions.
What are the healthcare use and medical costs patterns of clinical cleft care and how is this influenced by the use of the ICHOM Standard Set for Cleft lip and Palate?	Healthcare services, including medical consultations, diagnostic and surgical procedures, of 40 patients with unilateral CL/P were counted and related costs were calculated. Expected treatment protocol costs and additional expected costs after implementing the ICHOM Standard Set were calculated and compared. Mean observed total costs (€40,859) for the complete treatment (0-24 years) were 1.6 times the expected costs due to optional, non-protocolized procedures, with highest costs first year after birth. Hospital admissions including surgery accounted for 42% of observed costs and 70% of expected protocol-based costs. Implementing the ICHOM Standard Set increased protocol-based costs by 7%.

Table 1 Key findings per research question.

Part I Measurement challenges

Part I of this thesis has focused on the patient-reported outcome measures in the ICHOM Standard Set for the comprehensive appraisal of cleft care. The Standard Set for Cleft Lip and Palate includes nine CLEFT-Q scales, the COHIP-OSS, the NOSE instrument and the Intelligibility in Context Scale to measure the patient's perspective on health. These various instruments cover the core concepts of facial appearance, psychosocial function, speech, facial function (including eating and drinking, and breathing), and oral health. In 2016, these outcome measures were implemented in routine clinical practice for the first time at the Erasmus MC, followed by several other hospitals in the United States and Sweden.

As we know that an instrument's performance can alter after implementation in a new setting and patient population, it is necessary to verify that each of the instruments of the Standard Set remains robust enough to accurately and reliably inform the corresponding outcome domain in the local situations. In addition, to ensure the feasibility and sustainment of the implementation of the Standard Set, any unnecessary outcome measures resulting in registration burden for both patients and healthcare providers should be limited.

Concept coverage

The rationale of the first study conducted and described in **Chapter 2** on the psychometric performance of the patient- and parent-reported outcome measures in the ICHOM Standard Set for Cleft Lip and Palate was to identify ways for improving concept coverage. While the majority of the scales proved to be valid measures with high reliability after Rasch analysis, specific problems with the CLEFT-Q eating and drinking, NOSE, and COHIP-OSS instruments reflecting facial function and oral health were noted. The analysis revealed low reliability values and disordered thresholds for multiple items within the three previously mentioned questionnaires. Disordered thresholds can occur as a consequence of unclear definitions, too many response options, or underutilization of a response option.[1] When combining this finding with the probability of a patient choosing a specific response option, it was found that the middle response options were hardly ever used and thus no more than two groups could be discriminated with the instruments, namely groups at both ends of the continuum. Altogether, this suggests that these three instruments were not robust enough for outcome comparisons in their concepts for patients with CL/P in this setting, and work like a checklist rather than a measurement scale.

In a checklist, every single item (i.e. question) may be appraised as an independent entity measuring a specific dimension within a construct with a separate score. This is called multidimensionality. As a consequence, no overall sum score could be calculated. This is in contrast to a truly valid scale, where all items measure the same construct and the sum score could inform patients and healthcare professionals on the overall well-being of the specific construct measured by that scale. Instruments with multidimensionality are less suitable for outcome comparisons, such as comparing treatment techniques, protocols, or centers, as their sum scores are not interpretable and cannot be related to one specific construct measured.[2,3] As one of the main goals of collecting patient-reported outcomes in the cleft population is outcome comparison, these instruments seem to be less valuable in this setting than previously thought.

Therefore, the instruments covering the concepts of facial function and oral health in the ICHOM Standard Set should be reconsidered and a search for alternatives to close the concept coverage gap is desirable. The NOSE-questionnaire will be used as an example for demonstrating the various options to improve concept coverage with appropriate and patient-fitted outcome measures. A first step is to go back to the development stage of the outcome set and redefine the goal of measurement, *what* to measure (construct), *how* to measure, and *when* to measure.[4] For example, during the development of the ICHOM Standard Set it was found important to measure the construct of 'breathing'. Carefully reviewing the NOSE questionnaire, shows that the construct of 'breathing' might be too broad and that 'nasal breathing' or 'nasal obstruction' are perhaps a better fit. As a result, it can be considered to alter the construct measured, but performance of the instrument remains the same. However, a checklist can still be relevant for clinical-decision making, because individual elements can be intervened upon. Another corroborating instrument or quantitative measurement, for example nasometry, can be applied when a patient scores poorly on the NOSE checklist.

Going back to the essence of the measurement instrument and finding out why and how it was developed, can provide more information on the applicability of the instrument. The NOSE-questionnaire was originally developed for the evaluation of septorhinoplasty procedures in adult patients.[5] The development and validation studies did not include patients with CL/P.[5] Our research showed that the majority of children with CL/P reported no problems of nasal breathing at the ages of 8 and 12 years. Since surgical interventions, such as a correction of septal deviation, are held back until complete skeletal development, the inclusion of this instrument at young ages seems redundant. In addition, the NOSE instrument might be more useful as an optional tool for the evaluation of septorhinoplasty at a later stage in the treatment trajectory.

Another possibility is trying to improve the performance of the instrument by adding items, or changing response options. In our study, we found that merging the middle response options of the NOSE and COHIP-OSS instruments resulted in better threshold ordering and a slightly higher reliability and discriminative value. A recent study modified the NOSE questionnaire by merging the five response options to three and adding mildly verbal alterations of questions to investigate the prevalence of nasal obstruction symptoms in children with CL/P. The study confirmed the increased discriminative value by finding differences in nasal obstruction severity between patients with unilateral and bilateral CL/P.[6] Additionally, the verbal alterations showed that there is room for improvement of the comprehensibility of the scale. In practice, young children have been observed to have difficulty interpreting the NOSE questions, and parents were often asked to explain terms such as “nasal blockage or obstruction”. To improve compliance and inclusiveness, a low literacy institution, for example the Dutch institution Pharos[7], or specialized patient associations, could be consulted to advice on how to improve the comprehensibility of instruments for young children and illiterate patients, for example by adding pictograms or simplifying wording.[7] As 2.5 million people of 16 years and above in The Netherlands are illiterate[7,8], these types of institutions could be a promising stakeholder in the improvement of standardized outcome measures.

If measuring the construct of ‘breathing’ at a young age remains desirable, it can be decided that some instruments of the Set are optional and will not be used for outcome comparisons, or perhaps choosing a different instrument would be a better option. For this purpose, the freely available database of systematic reviews of outcome measurement instruments, developed by COSMIN, could be a useful tool.[9] The COSMIN (COnsensus-based Standards for the selection of health Measurement INstruments) initiative is an international, multidisciplinary group of researchers with expertise in epidemiology, psychometrics, healthcare and qualitative research, aiming to improve the selection of measurement instruments and developing guidelines and tools to support the selection of the most suitable outcome measurement instrument.[2,10-12]

Since CL/P is often accompanied by other craniofacial differences, for example maxillary retrognathia and backward displacement of the tongue in case of a Robin sequence diagnosis, a more generic outcome measure that can be used across various disease groups could be considered. The recently published FACE-Q Kids Craniofacial Module, developed specifically for patients between 8 and 29 years old with a visible and functional craniofacial difference could be an interesting alternative.[13] The module includes 27 independently functioning instruments on facial appearance, facial function, health-related quality of life, and adverse effects (checklists).[13-15] The health-related quality of life scales are applicable to all patients with a craniofacial difference and the rest of the scales can be administered optionally, depending on a patient's entity.[14,15] A recent systematic literature review has identified the FACE-Q as an instrument with high potential for its use in rhinoplasty outcome evaluation.[16] In addition, one of the facial function scales is the FACE-Q breathing scale. The FACE-Q breathing contains questions related to nasal obstruction, as well as to upper airway obstruction, and thus covers the concept of breathing more widely than the NOSE questionnaire. Moreover, the use of outcome measures in other clinical practices, including various pediatric and craniofacial settings, is increasing. By using a more generic measure, the patient's well-being can be assessed by one and the same outcome measure and interpreted by the involved clinicians related to the various disease entities. This will help keeping the response burden for a patient as low as possible and will increase the possibility of comparing outcomes across various disorders. So far, the validity and reliability of the FACE-Q breathing scale has not yet been investigated in the CL/P population. These topics should be addressed first in future research projects, before a final decision can be made to include new outcome measures in the ICHOM Standard Set.

Burden reduction

As we have seen that the majority of scales provide good concept coverage with valid and reliable measures, further exploration of the concepts is in place to make sure that each scale provides unique and clinically useful information. In **Chapter 3,** an evaluation of the informative value of the instruments covering the concept of psychosocial function in the ICHOM Standard Set is described. Measuring psychosocial function is of importance because impairment of this area has been commonly reported in patients with CL/P. Key contributing factors include being teased or bullied, dissatisfaction with appearance, and dissatisfaction with speech.[17-22] Thus, timely identification of psychosocial problems is needed to provide appropriate care within the cleft team or, when necessary, by referring the patient to a psychologist for further evaluation and support.

The psychosocial function outcome instruments defined by the ICHOM Standard Set are the CLEFT-Q psychological, CLEFT-Q social, and CLEFT-Q school function scales.[23] Strong correlations were observed between the social scale and the psychological and school scales, where the correlation between the school and psychological scale was considerably lower. As higher correlations indicate the measurement of more similar constructs[24], these findings suggest that the additional value of the social scale in the ICHOM Standard Set is limited and the inclusion of the scale as a mandatory measure should be reconsidered. In addition, our study found a negative trend of decreasing outcome score of psychological function over time with higher chances of being referred to psychosocial care. This implies that the CLEFT-Q psychological scale could become of great importance for both longitudinal outcome comparisons and clinical decision-making.

To take outcome measurement to the next level in terms of clinical decision-making, more knowledge on the interpretation of scores is necessary. So far, research on how to contextualize the outcome scores of the ICHOM Standard Set and translate this into clinical practice have been scarce. The availability of reference scores, normative scores and minimal clinically important differences from the patient perspective could be useful in early detection of health problems and taking jointly decisions on care management. In addition, it can stimulate the uptake and compliance of PROMs in clinical practice.[25] Research has shown that PROMs are more valued by healthcare professionals when they are influential on the clinical decision-making process.[26]

In **Chapter 4,** a Computer Adaptive Test version of the CLEFT-Q instruments for burden reduction and enhancing compliance in patient-reported outcome measurement is proposed. External validity testing showed that the CLEFT-Q CAT was able to reduce the length of eight CLEFT-Q scales from 76 to 59 items, while maintaining high accuracy. In practice, when a patient changes one entity on an item in a full-length scale, for example answering “a little bit” instead of “not at all”, the 0-100 score could change up to 4 points.[27] With the CAT version, scores varied 2 to 5 points out of 100 from the full-length assessment scores. Even though there is a knowledge gap regarding clinical and minimal important differences of the CLEFT-Q scale scores, the findings of the CAT validity study suggests that score interpretation will be similar.

The CLEFT-Q CAT version has been developed according to Rasch Measurement Theory (RMT)[28], whereas other high profile CAT initiatives, such as the Patient-Reported Outcomes Measurement Information System (PROMIS)[29,30], are developed using Item Response Theory (IRT).

IRT models appear to have greater measurement reliability with lower standard error of measurement (SEM). However, a lower SEM does not have to guarantee a more accurate reproduction of linear assessment scores. In fact, the study described in **Chapter 5** showed that even though the unidimensional graded response model (GRM) CAT algorithms (IRT) achieved lower SEM than the Rasch equivalents, the reproduced linear assessment scores were in close agreement and were reproduced with similar accuracy.

This is in accordance with previous studies published by the International Society for Quality-of-Life Research (ISOQOL) Psychometric Special Interest Group. They created and evaluated scales from the PROMIS depression item bank using RMT and IRT and achieved similar measurement results.[31-33] This suggests an interchangeability of GRM and RMT techniques and this finding is likely generalizable to RMT-developed PROMs in other clinical settings.

Together with the development of the CLEFT-Q CAT version, a digital platform was created to facilitate the uptake of the CAT in clinical practice and to visualize outcome scores to benefit (clinical) interpretation.[25,34] The study undertaken in **Chapter 4** also included a qualitative part with semi-structured interviews and pilot-testing of the CAT web-application in 'Concerto'.[35] The use of the CAT platform was perceived to improve clinical communication and facilitate shared decision-making by both patients as health care professionals. The CAT was not felt to cause excessive response burden to patients or healthcare professionals. The platform compares a person's score to scores obtained from people with similar clinical and demographic characteristics. Concordantly, a recently published systematic review on visualization formats of PROM data found that patients preferred bar charts and line graphs and that scores were mostly compared with patients' own previous outcomes.[36] For further clinical interpretation, scores were compared to norm population scores.[36] However, this study excluded the pediatric population, while children could have other visualization preferences. A national project, named Beslist Samen 2.0, has explored the use of various 'score visualization dashboards' amongst young adults with CL/P. Their findings were in line with previously discussed literature[36], but also emphasized that dashboards should be dynamic to match the wishes of each patient. Especially in cleft care, as we deal with patients with a wide variety of ages. Unfortunately, the focus groups of this national project did not include the very young patient, thus the way of visualizing outcome scores to this group still needs further exploration.

Part II Implementation challenges

Part II of this thesis focuses on the implementation of the Standard Set in clinical cleft practice. The cleft teams of the Erasmus University Medical Center (NL), Boston children's Hospital (USA), Duke Children's Hospital (USA), and Karolinska University Hospital (SE) have successfully implemented the Standard Set in their routine clinical practice. However, at multiple other institutions, both nationally and internationally, implementation efforts are ongoing and often challenged by the lack of a defined strategy or clear understanding of conditions that promote or hinder routine outcome measurement[37], and a general feeling among (non-adopting) teams that with the implementation of routine outcome measurement, medical care and costs are likely to increase.[38] For the purpose of value-based healthcare, including learning from best practices and outcomes comparisons, it is essential that outcomes measurement is widely implemented. Knowledge on promoting and hindering factors can be used to develop implementation strategies that focus on the perceived strengths of implementing outcome measurement, while overcoming its perceived barriers.[39]

Facilitators and barriers

To support other cleft teams during their implementation endeavors, **Chapters 6 and 7** described the experiences of the first four cleft centers who successfully managed the implementation process of the ICHOM Standard Set for CL/P. Building support for routine outcome measurement implementation is an essential first step in change management.[39-41] The majority of the interviewees felt that the routine use of PROMs fostered the connection between the patient and the team of healthcare providers. As strongest facilitators for adopting outcome measurement in routine practice, motivation and importance were mentioned repeatedly by all stakeholders. A comparable quantitative analysis by Weidler *et al.*[39] in the USA and UK found that cleft care professionals believe that standardized outcome measurement assists them in identifying areas where a child could benefit from further treatment. This view was shared by the caregivers. Also, the belief amongst care providers that outcome measurement will help them to compare results across techniques, protocols, providers and/or teams was perceived as a strong motivation for adoption of routine outcome measurement.[39] Internal enthusiasm for the implementation may be further enhanced by repeatedly communicating the benefits of the program to all stakeholders during team meetings, by making results of outcome measurement available during clinical evaluations, by providing training on how to interpret these outcomes, by making the registration system easily accessible, and by performing evaluation cycles to identify the effectiveness of current strategies and ways to improve them.

Even though there seems to be a supportive attitude towards the implementation of outcome measurement in practice, this support is somewhat tempered by the perceived barriers to implementation. Our study found that implementation efforts were most constrained by time and the health information technology. Time, as part of resources, was articulated to have an overarching and continuing influence on all dimensions of the framework, especially on adoption and implementation. These findings are, to a large extent, in accordance with findings from previous studies in other clinical fields.[38,42,43] Therefore, intentionally investing time to lay a sound foundation is important to foster each phase of the implementation process.

As IT-infrastructure can be a large hindering factor, teams facing problems with implementation technology are advised to collaborate with other teams and international initiatives for capturing patient outcomes. Within the Netherlands, the NFU (Dutch Federations of University Hospital) expert committee has been founded to accelerate the implementation of outcome measurement at the Dutch cleft teams and to support future comparison projects. The teams share their knowledge on a regular basis and helped each other further by providing access to already developed HIT-systems. On a European level, the European Reference Network (ERN) for rare craniofacial anomalies provides a network with access to a central patient registry for research and quality improvement.[44] On the other side of the ocean, ACCQUIRENet is an American collaborative initiative sharing their registry management system if you join their network.[45,46] In addition, the CLEFT-Q CAT platform will be made available free of charge by the Oxford Research Group.[35]

A theme that was not directly highlighted in our study, but described in a systematic review by Duncan *et al.*[38] was patient considerations. Their review reported various concerns from clinicians about the patient's ability to complete outcome measures. Beliefs were that PROMs might be too complicated to complete independently, that the instruments required a high language proficiency, that ethnic and cultural sensitivity issues might be present, or that patients might become discouraged if their progress turns out to be less than others.[38,47-49] Various studies on the measurement of HRQOL by children showed that self-report was possible from the age of 8 years old.[50,51] One study even demonstrated that children as young as 5 years can reliably and validly self-report their HRQOL with the use of an age-appropriate instrument.[52] During outpatient visits at the Erasmus MC, it was noted that the children regularly completed PROMs together with their parents or caregivers, even though most of the scales are deemed suitable for young children. This phenomenon might potentially lead to biased results, as literature showed evidence that the child-parent agreement rate on outcome measures varies considerably, and is often low.[50,51] A disagreement is more likely a result of each individual views on the child's health, rather than that someone is right or wrong.[53,54] In order to obtain a richer understanding of the paediatric HRQOL, or in case of very young children or developmentally delayed children, the inclusion of proxy- or parent-reported outcome measures could be useful.[51,53]

In response to our facilitators and barriers study, Harrison *et al.*[55] presented concerns about the negative wording of the speech-related CLEFT-Q scales.[55] An international study evaluating the impact of answering the CLEFT-Q scales reported that 88% of the participants liked answering the questions and most of them did not feel unhappy or upset afterwards.[56] Also, the instruments contain small pictograms to help understand the questions. Nonetheless, the authors did stress that CLEFT-Q scores should be examined as soon as possible after its completion in order to identify patients in possible need for additional care.[56] This point was also emphasized by the interviewees in our qualitative study of **Chapter 6**. Together with the previously described patients' preferences for comparing outcome scores with their own previous scores[36], the belief that patients might become discouraged if their progress turns out to be less than others could be overcome.

In addition to the qualitative analysis, the narrated experiences of the four cleft teams have been collected in **Chapter 7**. As these sites are collecting outcomes for a few years now, their teams are at a stage where they face new challenges in proceeding towards collaborative outcome comparison projects and sustaining the implementation endeavours.

First, sharing individual patient data between centers is hindered by navigating through privacy laws and the GDPR (General Data Protection Regulation) framework, which is time-consuming and exclude teams from benchmarking efforts.[57] To overcome these issues, a recent study in the field of hand surgery proposed an alternative approach where data is analysed in each center locally, and only summary statistics are transferred and further analysed at a multi-center level using meta-analyses.[58] In order to make this work smoothly, there is a need for protocols to ensure that data is extracted in a uniform format to correctly run the analysis script.[58] In the current case of cleft care, the ICHOM Standard Set for CL/P already meets these needs to a large extent, but should perhaps become more detailed when an analysis script is available. Future research should test this promising way of comparing outcomes without sharing raw data to determine what information and knowledge is still missing to enable benchmarking on a global scale in cleft care.

The second challenge mentioned was the need to develop risk-adjustment models for outcome comparison projects. Our study in **Chapter 3** showed that patients with a genetic syndrome are more likely to score lower on the CLEFT-Q psychological function and were more often referred to psychosocial care. This finding suggests that case-mix variables are important factors for meaningful outcome comparisons and adjustments, and for the prediction of patients in need of more attention. Especially in terms of international benchmarking initiatives, since heterogeneity in patient population between centers exists.[59,60] Only a small proportion of observed differences is due to an actual difference in quality of care. The rest is mainly a result of random variation, unexplained differences and registration bias.[59,60] Outcomes can be negatively influenced in centers where more severe patients are treated, while this might have nothing to do with the quality of care that the healthcare professionals are offering. A recent publication by Oemrawsingh *et al.*[59] on case-mix adjustment in ischemic stroke care, concluded that variables, such as psychological or social factors, should be considered as potential case-mix variables for PROMS.[59] Educational achievements[61,62], being teased or bullied[63], and perceived parent-child relationships[64,65] could be some of those potential factors. The Dutch socio-economic status scores, as used in our research, cannot serve the purpose of international comparison. Education, income and profession of parents are other variables that could represent this characteristic.[66] Finding the appropriate case-mix variables for the population with CL/P is challenging and requires further research.[17]

Healthcare utilization

Aside of the focus on patient's outcomes and how to compare and learn from these outcomes, another important aspect of value-based healthcare is represented by the denominator of the equation, namely costs of care. When increasing outcomes are accompanied by an incremental increase of costs, the added value will be limited. On the other hand, when costs decrease as care is better targeted, and outcomes improve, the value improvement will be much more.[67-69] For this, we need to know the current state of costs and how healthcare services are used. Therefore, in **Chapter 8** we described healthcare utilization and medical costs for patients with unilateral CL/P, and found that the mean total costs observed for a complete treatment trajectory were almost 41,000 euros, and these costs were 1.6 times the expected costs based on the protocol. The large amount of optional, non-protocolized procedures were main drivers of this difference. Hospital admissions including surgery accounted for 42 percent of the observed costs, while 70 percent of total expected protocol-based costs were dedicated to hospital admissions. This finding suggests that treatment protocols within cleft care are suboptimal predictors of actual healthcare utilization since a lot of care is unprotocolized. In our study, we have only researched one academic institution, but within the large inter-center Eurocleft study including 201 cleft centers, a total of 194 different treatment protocols were found for the treatment of UCLAP.[70] All these varieties are probably due to the heterogeneity of the cleft population, the multidisciplinary and long character of care, and the beliefs and experiences of healthcare professionals.[70] These aspects also complicate the analysis of healthcare utilization patterns and challenge us to find valid research solutions.

Especially in cleft care, specific treatments can start at varying timepoints depending on development stage, such as orthognathic surgery, and treatments can continue over longer time periods, such as orthodontics and speech therapy. Therefore, evaluating costs might benefit from clustering multiple treatment years, for example 10-12 years can become one cluster. The ICHOM Standard Set can be used to guide this categorization. In addition, orthodontic care and speech therapy are largely delivered outside of the hospital in local and specialized clinics. Since we were unable to review an extended range of costs, including but not limited to out-of-pocket expenses by patients, medication costs, travel costs, and societal costs due to absence from work[71], the cost estimations in our study are likely an underestimation of real expenses and care delivery reflecting only the hospital perspective. To obtain a broader, societal perspective, this data could be collected as part of a cost-effectiveness study, which is seen as one of the most thorough and valid study designs to investigate the impact of a new treatment modality on health outcomes and costs.[72] These studies are traditionally performed to substantiate the effectiveness of a new drug or therapeutic intervention, in comparison to the gold standard, to request reimbursement at the health insurance companies.[71] As the theory of value-based healthcare claims that value for the patient will increase by improving outcomes and decreasing costs, it is striking that such large care transformations, as implementing outcome registration frameworks, are rolled out before a thorough testing phase or cost-effectiveness analysis has been done.

Unfortunately, in the specific case of cleft care, valid effectiveness measures were not routinely measured before the implementation of the ICHOM Standard Set. As a consequence, pre- and post-implementation comparisons of effectiveness and (out of hospital) costs are hindered. Still, it would be valuable to gain more insight in the actual mechanism of costs and the influence on quality of care, even if it’s only perceived quality of care by patients. Therefore, teams that want to implement outcome measurement in their future routine practice should consider finding partner-teams to collaborate with and who are willing to implement the intervention through a stepped-wedge design. A stepped-wedge design is a type of cluster-randomized trial where the intervention is gradually introduced to all study centers, rather than all at once.[73,74] This design allows for the collection of both pre- and post-intervention data and can be used to study the effectiveness and implementation of the intervention over time.[73,74] It will also provide direct insights in the facilitators and barriers of various stages of implementation. It should be noted that a general measure of quality of care, such as clinical outcome or a generic PROM, should be chosen as effectiveness measure to compare both care delivery systems. Even though a stepped-wedge implementation study is time-consuming, at the end of the study all participating centers have implemented the same outcomes framework enabling between-center comparisons in the future.[74]

Analyzing healthcare utilization patterns can provide a variety of benefits, where identifying areas of overuse or underuse of healthcare resources is one of them. By understanding how and where healthcare services are being used, healthcare providers can identify areas where resources are being wasted or where they are needed but not being provided. The methodology used in **Chapter 8** can be a good first step in mapping and gaining insights in local healthcare use. At the moment of conducting the cost analysis study, the Standard Set had been implemented for four years. To evaluate the long-term costs and to get a more reliable image of the healthcare use among cleft patients, repeating this study with a larger sample size could provide interesting insights.

However, during the execution of this study, one main challenge was the extraction of utilized healthcare service data and prices of these services from the local information systems: utilized codes varied per specialism or between specialist, and codes changed with irregular intervals. This made the collection of reliable data extremely labor-intensive and time-consuming, limiting us from including a larger patient group in the analysis. Moreover, non-uniform coding and various extraction strategies might be hindering similar endeavours in future, especially when such an undertaking will be deployed among multiple centres for between-center comparisons. This observation is in accordance with barriers faced by a large national project, initialized by the Ministry of Healthcare and Sports, in which 9 regional initiatives of overarching healthcare systems (Dutch: ‘Proeftuinen’) were cooperating.[75] The goal was to reduce the costs of care, improve the quality of care and the health of the population. All initiatives faced the limits of the current information and knowledge infrastructure in which information could not be technically extracted from the system. It turned out to be especially difficult to map healthcare costs.[75] As a result, intended payment transformations as part of the projects were being delayed or completely held back.[75] Meanwhile, national project groups have started to create more clarity at the registration level, such as ‘Registratie aan de Bron’[76] and ‘Programma Uitkomstgerichte Zorg’.[77]

Strengths, limitations and generalizability

The data used to conduct the outcome measurement studies consisted of a large international sample of patients with CL/P. To make this possible, a collaboration between Boston Children's Hospital (USA), Duke Children's Hospital (USA), McMaster University (CA) and Erasmus University Medical Center (NL) was initiated. These studies are unique in its large number of patients, as within CL/P research sample sizes are generally smaller due to the heterogeneity of the disorder and a relatively low prevalence in comparison to more common disorders as diabetes or cardiovascular diseases. However, patients from middle- to low-income countries were underrepresented in this research sample. So far, only very few centers in middle-income countries have started collecting patient-reported outcomes. This limitation was also present in the cross-sectional, qualitative analysis described in **Chapters 6 and 7**, which provided views from four cleft centers from high-income countries with different implementation methods, representing unique cultures and societal habits. It is likely that factors influencing change management will not differ profoundly, but perhaps differences in financial and technological resources will be more prominent in low-resource countries. The interviewees mentioned that starting out with pen and paper versions might then be the way to go, as long as you start collecting outcomes. However, several papers have described that this method is perceived as labor-intensive[43,55], which could result in a decreasing motivation for outcomes collection.

Also, to increase the uptake and maintain reliability of outcome measures in middle- to low-income countries, instruments should be translated into various languages followed by cross-cultural testing.[24] Currently, the CLEFT-Q is available in 32 languages and ready for international implementation.[27] However, the other patient-reported instruments of the ICHOM Standard Set are in need of expanding translations to increase the uptake; a recent PubMed search showed that the NOSE questionnaire is validated in 9 languages, the COHIP-OSS in 6 languages, and the ICS in 8 languages. In addition to translations, centers in middle- to low-income countries need resources to implement the instruments in their clinical practices.

With regards to the implementation experiences, there may have been memory bias among the interviewees. Successful implementation and the passage of time can blur the memories of negative experiences and obstacles encountered along the way. This could shine a more positive light on their experiences. Therefore, it would be interesting to repeat the study on facilitators and barriers among groups that have not started implementation, or where implementation is ongoing but not yet finalized. Together with our results, this will lead to a broader spectrum of facilitators and barriers and provides the opportunity to develop implementation strategies for each implementation phase, whether you are a starter or at the sustainment level.

A unique point of the study described in **Chapter 8** is the evaluation of healthcare utilization and costs of a challenging and complex treatment trajectory for CL/P with a long total follow-up time of 8 years to compose a full treatment trajectory of 24 years. Nonetheless, as this study include data from only one academic hospital, its generalizability to other centers is limited, as patient populations might differ. Further research is needed to specify predictors for variability in healthcare use, such as cleft type, family circumstances, and socio-economic status, to target individuals in need of more extensive care enabling risk stratification and risk-adjustment.

Conclusion

As the value-based healthcare transformation takes place in cleft care, and implementation of standardized outcome measurement is increasing, we aimed to expand on the challenges around patient-reported outcome measurements and related implementation efforts. Developing and sustaining an outcomes framework is an iterative process of evaluating, adjusting and (re-) implementing the chosen outcome instruments in real practice. The use of a Computerized Adaptive Test version of scales could help reduce registration burden for both patients and healthcare professionals and could stimulate the uptake and compliance of measuring outcomes. As implementation efforts are often constrained by time and health information technology, it is important for teams to collaborate with international initiatives to accelerate the implementation by sharing their knowledge on a regular basis and providing access to already developed HIT-systems or registries. Locally analyzing healthcare utilization patterns can provide an understanding on how and where healthcare services are used or needed. In short, transforming care comes with great challenges, but together we can face and overcome these hurdles to ultimately provide the best possible care to our patients with a cleft.

Part I: Measurement challenges
The patient-reported outcome instruments representing the concepts of facial function and oral health in the ICHOM Standard Set for cleft should be reconsidered.
A standardized outcomes set should include core outcomes to measure and compare, but teams should retain the freedom for additional or optional measures according to their specific wishes or patient population.
Include low literacy or specialized patient institutions to improve comprehensibility of outcome measurements.
Start implementing the CLEFT-Q Computer Adaptive Test in practice to reduce registration burden.
In order to obtain a richer understanding of the paediatric HRQOL, or in case of very young children or developmentally delayed children, the inclusion of proxy- or parent-reported outcome measures could be useful.
The use of the ICHOM Standard Set in a clinical setting should be iteratively evaluated, adjusted and (re-) implemented.
Part II: Implementation challenges
Intentionally investing time to lay a sound foundation is important to foster each phase of the implementation process.
As IT-infrastructure can be a large hindering factor, teams facing problems with implementation technology should collaborate with other teams and international initiatives for capturing patient outcomes.
Outcome scores should be examined as soon as possible after its completion in order to identify patients in possible need for additional care.
To increase the uptake of outcome measures in middle- to low-income countries, instruments should be translated into various languages followed by cross-cultural testing.
Teams that want to implement outcome measurement in their future routine practice should consider finding partner-teams to collaborate with and who are willing to implement the intervention through a stepped-wedge design.
Exploring healthcare utilization patterns can provide insight in current care use and provide ideas for more efficiently arranging workflows.
Extraction of healthcare service data from the local information systems should become easier and more transparent.

Table 2 Overview of key recommendations.

Part I: Measurement challenges
Explore the validity and reliability of the FACE-Q breathing scale in a population with CL/P at various timepoints.
Contextualize the outcome scores of the ICHOM Standard Set by further researching reference scores and minimal clinically important differences from the patient's perspective.
Investigate the visualization preferences for patient-reported outcome scores by young children with cleft lip and palate.
Part II: Implementation challenges
Explore statistical ways for comparing outcomes without sharing raw data and determine missing information to enable global benchmarking.
Research case-mix variables and develop risk-adjustment models for outcome comparison projects.
Perform a multi-center stepped-wedge implementation to study the effectiveness of the intervention over time and provide direct insights in the facilitators and barriers of the various stages of implementation.
Repeat the cost analysis study with a larger group over a longer time period to get a clearer view on the influence of the ICHOM Standard Set on healthcare use and costs.

Table 3 Future research ideas.

References

1. Pallant JF, Tennant A. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). *Br J Clin Psychol.* 2007;46(Pt 1):1-18.
2. Prinsen CAC, Mokkink LB, Bouter LM, et al. COSMIN guideline for systematic reviews of patient-reported outcome measures. *Qual Life Res.* 2018;27(5):1147-1157.
3. Carle AC, Weech-Maldonado R. Validly interpreting patients' reports: using bifactor and multidimensional models to determine whether surveys and scales measure one or more constructs. *Med Care.* 2012;50(9 Suppl 2):S42-48.
4. Boers M, Kirwan JR, Tugwell P, et al. *The OMERACT Handbook.* 2018.
5. Stewart MG, Witsell DL, Smith TL, Weaver EM, Yueh B, Hannley MT. Development and validation of the Nasal Obstruction Symptom Evaluation (NOSE) scale. *Otolaryngol Head Neck Surg.* 2004;130(2):157-163.
6. Sobol DL, Allori AC, Carlson AR, et al. Nasal Airway Dysfunction in Children with Cleft Lip and Cleft Palate: Results of a Cross-Sectional Population-Based Study, with Anatomical and Surgical Considerations. *Plast Reconstr Surg.* 2016;138(6):1275-1285.
7. Expertisecentrum P. https://www.pharos.nl/thema/laaggeletterdheid-gezondheidsvaardigheden. Accessed December 12, 2022.
8. Rijksoverheid. Aanpak laaggeletterdheid. https://www.rijksoverheid.nl/onderwerpen/laaggeletterdheid/aanpak-laaggeletterdheid. Accessed.
9. COSMIN. https://database.cosmin.nl/. Accessed December 12, 2022.
10. Gagnier JJ, Lai J, Mokkink LB, Terwee CB. COSMIN reporting guideline for studies on measurement properties of patient-reported outcome measures. *Qual Life Res.* 2021;30(8):2197-2218.
11. Prinsen CA, Vohra S, Rose MR, et al. How to select outcome measurement instruments for outcomes included in a "Core Outcome Set" - a practical guideline. *Trials.* 2016;17(1):449.
12. 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.
13. Longmire NM, Wong Riff KWY, O'Hara JL, et al. Development of a New Module of the FACE-Q for Children and Young Adults with Diverse Conditions Associated with Visible and/or Functional Facial Differences. *Facial Plast Surg.* 2017;33(5):499-508.
14. Klassen AF, Rae C, Riff W, et al. FACE-Q craniofacial module: Part 2 Psychometric properties of newly developed scales for children and young adults with facial conditions. *J Plast Reconstr Aesthet Surg.* 2021;74(9):2330-2340.
15. Klassen AF, Rae C, Wong Riff KW, et al. FACE-Q Craniofacial Module: Part 1 validation of CLEFT-Q scales for use in children and young adults with facial conditions. *J Plast Reconstr Aesthet Surg.* 2021;74(9):2319-2329.
16. van Zijl F, Mokkink LB, Haagsma JA, Datema FR. Evaluation of Measurement Properties of Patient-Reported Outcome Measures After Rhinoplasty: A Systematic Review. *JAMA Facial Plast Surg.* 2019;21(2):152-162.
17. Stock NM, Feragen KB. Psychological adjustment to cleft lip and/or palate: A narrative review of the literature. *Psychol Health.* 2016;31(7):777-813.
18. Feragen KB, Borge AI, Rumsey N. Social experience in 10-year-old children born with a cleft: exploring psychosocial resilience. *Cleft Palate Craniofac J.* 2009;46(1):65-74.

19. Feragen KB, Saervold TK, Aukner R, Stock NM. Speech, Language, and Reading in 10-Year-Olds With Cleft: Associations With Teasing, Satisfaction With Speech, and Psychological Adjustment. *Cleft Palate Craniofac J.* 2017;54(2):153-165.

20. Feragen KB, Stock NM. Risk and Protective Factors at Age 10: Psychological Adjustment in Children With a Cleft Lip and/or Palate. *Cleft Palate Craniofac J.* 2016;53(2):161-179.

21. Hoek IH, Kraaimaat FW, Admiraal RJ, Kuijpers-Jagtman AM, Verhaak CM. [Psychosocial adjustment in children with a cleft lip and/or palate]. *Ned Tijdschr Geneeskd.* 2009;153:B352.

22. Hunt O, Burden D, Hepper P, Stevenson M, Johnston C. Parent reports of the psychosocial functioning of children with cleft lip and/or palate. *Cleft Palate Craniofac J.* 2007;44(3):304-311.

23. International Consortium for Health Outcomes Measurement (ICHOM). Data collection reference guide. https://ichom.org/files/medical-conditions/cleft-lip-palate/cleft-lip-palate-reference-guide.pdf. Published 2018. Accessed December 1, 2020.

24. de Vet HC, Terwee CB, Mokkink LB, Knol DL. *Measurement in medicine: a practical guide.* Cambridge University Press; 2011.

25. Porter I, Goncalves-Bradley D, Ricci-Cabello I, et al. Framework and guidance for implementing patient-reported outcomes in clinical practice: evidence, challenges and opportunities. *J Comp Eff Res.* 2016;5(5):507-519.

26. Boyce MB, Browne JP, Greenhalgh J. The experiences of professionals with using information from patient-reported outcome measures to improve the quality of healthcare: a systematic review of qualitative research. *BMJ Qual Saf.* 2014;23(6):508-518.

27. QPortfolio. https://qportfolio.org/wp-content/uploads/2022/11/CLEFT-Q-USERS-GUIDE.pdf. Accessed December 12, 2022.

28. Harrison CJ, Geerards D, Ottenhof MJ, et al. Computerised adaptive testing accurately predicts CLEFT-Q scores by selecting fewer, more patient-focused questions. *J Plast Reconstr Aesthet Surg.* 2019;72(11):1819-1824.

29. Hung M, Baumhauer JF, Latt LD, et al. Validation of PROMIS (R) Physical Function computerized adaptive tests for orthopaedic foot and ankle outcome research. *Clin Orthop Relat Res.* 2013;471(11):3466-3474.

30. Cella D, Yount S, Rothrock N, et al. The Patient-Reported Outcomes Measurement Information System (PROMIS): progress of an NIH Roadmap cooperative group during its first two years. *Med Care.* 2007;45(5 Suppl 1):S3-S11.

31. Bjorner JB. State of the psychometric methods: comments on the ISOQOL SIG psychometric papers. *J Patient Rep Outcomes.* 2019;3(1):49.

32. Cleanthous S, Barbic SP, Smith S, Regnault A. Psychometric performance of the PROMIS(R) depression item bank: a comparison of the 28- and 51-item versions using Rasch measurement theory. *J Patient Rep Outcomes.* 2019;3(1):47.

33. Stover AM, McLeod LD, Langer MM, Chen WH, Reeve BB. State of the psychometric methods: patient-reported outcome measure development and refinement using item response theory. *J Patient Rep Outcomes.* 2019;3(1):50.

34. Evans JP, Gibbons C, Toms AD, Valderas JM. Use of computerised adaptive testing to reduce the number of items in patient-reported hip and knee outcome scores: an analysis of the NHS England National Patient-Reported Outcome Measures programme. *BMJ Open.* 2022;12(7):e059415.

35. Harrison C, Loe BS, Lis P, Sidey-Gibbons C. Maximizing the Potential of Patient-Reported Assessments by Using the Open-Source Concerto Platform With Computerized Adaptive Testing and Machine Learning. *J Med Internet Res.* 2020;22(10):e20950.

36. Albers EAC, Fraterman I, Walraven I, et al. Visualization formats of patient-reported outcome measures in clinical practice: a systematic review about preferences and interpretation accuracy. *J Patient Rep Outcomes.* 2022;6(1):18.

37. Basch E, Barbera L, Kerrigan CL, Velikova G. Implementation of Patient-Reported Outcomes in Routine Medical Care. *Am Soc Clin Oncol Educ Book.* 2018;38:122-134.

38. Duncan EA, Murray J. The barriers and facilitators to routine outcome measurement by allied health professionals in practice: a systematic review. *BMC Health Serv Res.* 2012;12:96.

39. Weidler EM, Britto MT, Sitzman TJ. Facilitators and Barriers to Implementing Standardized Outcome Measurement for Children With Cleft Lip and Palate. *Cleft Palate Craniofac J.* 2021;58(1):7-18.

40. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. *Implement Sci.* 2009;4:50.

41. Bradley EH, Holmboe ES, Mattera JA, Roumanis SA, Radford MJ, Krumholz HM. Data feedback efforts in quality improvement: lessons learned from US hospitals. *Qual Saf Health Care.* 2004;13(1):26-31.

42. Philpot LM, Barnes SA, Brown RM, et al. Barriers and Benefits to the Use of Patient-Reported Outcome Measures in Routine Clinical Care: A Qualitative Study. *Am J Med Qual.* 2018;33(4):359-364.

43. Amini M, Oemrawsingh A, Verweij LM, et al. Facilitators and barriers for implementing patient-reported outcome measures in clinical care: An academic center’s initial experience. *Health Policy.* 2021;125(9):1247-1255.

44. Cranio ERN. https://ern-cranio.eu/network-activities/activities/. Accessed January 14, 2023.

45. ACCQUIREnet. https://surgery.duke.edu/divisions/plastic-maxillofacial-and-oral-surgery/research/clinical-research/datalab-clinical-care-and-population-health/accquirenet. Accessed January 14, 2023.

46. Bittar PG, Carlson AR, Mabie-DeRuyter A, Marcus JR, Allori AC. Implementation of a standardized data-collection system for comprehensive appraisal of cleft care. *Cleft Palate Craniofac J.* 2018;55(10):1382-1390.

47. Jette DU, Halbert J, Iverson C, Miceli E, Shah P. Use of standardized outcome measures in physical therapist practice: perceptions and applications. *Phys Ther.* 2009;89(2):125-135.

48. Copeland JM, Taylor WJ, Dean SG. Factors influencing the use of outcome measures for patients with low back pain: a survey of New Zealand physical therapists. *Phys Ther.* 2008;88(12):1492-1505.

49. Simmons-Mackie N, Threats TT, Kagan A. Outcome assessment in aphasia: a survey. *J Commun Disord.* 2005;38(1):1-27.

50. Matza LS, Swensen AR, Flood EM, Secnik K, Leidy NK. Assessment of health-related quality of life in children: a review of conceptual, methodological, and regulatory issues. *Value Health.* 2004;7(1):79-92.

51. Bevans KB, Riley AW, Moon J, Forrest CB. Conceptual and methodological advances in child-reported outcomes measurement. *Expert Rev Pharmacoecon Outcomes Res.* 2010;10(4):385-396.

52. Varni JW, Limbers CA, Burwinkle TM. How young can children reliably and validly self-report their health-related quality of life?: an analysis of 8,591 children across age subgroups with the PedsQL 4.0 Generic Core Scales. *Health Qual Life Outcomes.* 2007;5:1.

53. Upton P, Lawford J, Eiser C. Parent-child agreement across child health-related quality of life instruments: a review of the literature. *Qual Life Res.* 2008;17(6):895-913.

54. Parsons SK, Barlow SE, Levy SL, Supran SE, Kaplan SH. Health-related quality of life in pediatric bone marrow transplant survivors: according to whom? *Int J Cancer Suppl.* 1999;12:46-51.
55. Harrison CJ, Rodrigues JN, Furniss D, Swan MC. Response to Barriers and Facilitators to the International Implementation of Standardized Outcome Measures in Clinical Cleft Practice. *Cleft Palate Craniofac J.* 2022;59(5):669-670.
56. Klassen AF, Dalton L, Goodacre TEE, et al. Impact of Completing CLEFT-Q Scales That Ask About Appearance on Children and Young Adults: An International Study. *Cleft Palate Craniofac J.* 2020;57(7):840-848.
57. Sheehan M, Friesen P, Balmer A, et al. Trust, trustworthiness and sharing patient data for research. *J Med Ethics.* 2020.
58. Duraku LS, Hoogendam L, Hundepool CA, et al. Collaborative hand surgery clinical research without sharing individual patient data; proof of principle study. *J Plast Reconstr Aesthet Surg.* 2022;75(7):2242-2250.
59. Oemrawsingh A, van Leeuwen N, Venema E, et al. Value-based healthcare in ischemic stroke care: case-mix adjustment models for clinical and patient-reported outcomes. *BMC Med Res Methodol.* 2019;19(1):229.
60. van Dishoeck AM, Lingsma HF, Mackenbach JP, Steyerberg EW. Random variation and rankability of hospitals using outcome indicators. *BMJ Qual Saf.* 2011;20(10):869-874.
61. Knight J, Cassell CH, Meyer RE, Strauss RP. Academic outcomes of children with isolated orofacial clefts compared with children without a major birth defect. *Cleft Palate Craniofac J.* 2015;52(3):259-268.
62. Wehby GL, Collet B, Barron S, Romitti PA, Ansley TN, Speltz M. Academic achievement of children and adolescents with oral clefts. *Pediatrics.* 2014;133(5):785-792.
63. Hunt O, Burden D, Hepper P, Stevenson M, Johnston C. Self-reports of psychosocial functioning among children and young adults with cleft lip and palate. *Cleft Palate Craniofac J.* 2006;43(5):598-605.
64. Tyler MC, Wehby GL, Robbins JM, Damiano PC. Separation anxiety in children ages 4 through 9 with oral clefts. *Cleft Palate Craniofac J.* 2013;50(5):520-527.
65. Murray L, Arteche A, Bingley C, et al. The effect of cleft lip on socio-emotional functioning in school-aged children. *J Child Psychol Psychiatry.* 2010;51(1):94-103.
66. Shavers VL. Measurement of socioeconomic status in health disparities research. *J Natl Med Assoc.* 2007;99(9):1013-1023.
67. Porter ME. What is value in health care? *N Engl J Med.* 2010;363(26):2477-2481.
68. Porter ME. Value-based health care delivery. *Ann Surg.* 2008;248(4):503-509.
69. Porter ME. A strategy for health care reform--toward a value-based system. *N Engl J Med.* 2009;361(2):109-112.
70. Shaw WC, Semb G, Nelson P, et al. The Eurocleft project 1996-2000: overview. *J Craniomaxillofac Surg.* 2001;29(3):131-140; discussion 141-132.
71. Zorginstituut Nederland. Richtlijn voor het uitvoeren van economische evaluaties in de gezondheidszorg. https://www.zorginstituutnederland.nl/over-ons/publicaties/publicatie/2016/02/29/richtlijn-voor-het-uitvoeren-van-economische-evaluaties-in-de-gezondheidszorg. Published 2016. Accessed April 1, 2020.
72. Rudmik L, Drummond M. Health economic evaluation: important principles and methodology. *Laryngoscope.* 2013;123(6):1341-1347.

73. Ouyang Y, Karim ME, Gustafson P, Field TS, Wong H. Explaining the variation in the attained power of a stepped-wedge trial with unequal cluster sizes. *BMC Med Res Methodol.* 2020;20(1):166.
74. Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. *BMC Med Res Methodol.* 2006;6:54.
75. Rijksinstituut voor Volksgezondheid en Milieu (RIVM). *Regio's in beweging naar een toekomstbestendig gezondheidssysteem. Landelijke Monitor Proeftuinen – reflectie op 5 jaar proeftuinen.* 2018.
76. Nictiz. *Dáárom, Registratie aan de bron! Eindrapportage 2016-2019.*
77. Ministry of Healthcare Welfare and Sport. *Programma Uitkomstgerichte Zorg.* 2019.

Chapter 10

Summary

Samenvatting

Summary

Chapter 1, the general introduction, provides an overview of the background and aims of this thesis.

In 2016, the ICHOM Standard Set for Cleft Lip and Palate was implemented in routine clinical practice for the first time at the Erasmus MC, followed by several other hospitals in the United States and Sweden. The Standard Set includes, aside of clinical indicators and case-mix variables, nine CLEFT-Q scales, the COHIP-OSS, the NOSE instrument and the Intelligibility in Context Scale to measure the patient's perspective on health. Since the use of outcome instruments in cleft practice is a relatively new and spreading phenomenon, it is in need of an evaluation to determine what it delivers and where and how to improve.

Therefore, this thesis covers multiple aspects concerning the measurement challenges of patient-reported outcomes in patients with a cleft and the implementation challenges of these outcome measures in clinical cleft practice. Specific research questions discussed in this thesis were:

1. How can we optimize the measurement of patient-reported outcomes in the ICHOM Standard Set for Cleft Lip and Palate?
    a. How is the psychometric performance and concept coverage of the patient-reported outcome measures of the ICHOM Standard Set for Cleft Lip and Palate?
    b. How can we maximize information while reducing burden when measuring psychosocial function within the ICHOM Standard Set for Cleft Lip and Palate?
    c. What is the external validity of the CLEFT-Q Computerized Adaptive Test (CAT) in patients with cleft lip and palate?
2. How can we optimize the implementation of the ICHOM Standard Set for Cleft Lip and Palate in clinical cleft care?
    a. What are facilitators and barriers to the implementation of the ICHOM Standard Set for Cleft Lip and Palate in clinical practice?
    b. What are the healthcare use and medical costs patterns of clinical cleft care and how is this influenced by the use of the ICHOM Standard Set for Cleft lip and Palate?

Chapter 2 described the Rasch analysis of the patient-reported outcome measures in the ICHOM Standard Set for CL/P to identify potential gaps in concept coverage. Patient-reported outcomes data from 714 patients with CL/P, aged 8 to 9, 10 to 12.5, and 22 years were collected in four-year time. The Rasch analysis showed that the scales relating to the concepts of facial appearance, speech function, and psychosocial function worked properly with high reliability parameters. However, specific problems with individual items within the CLEFT-Q eating and drinking, NOSE and COHIP-OSS questionnaires were noted, such as poor item fit to the Rasch model and disordered thresholds. Measurement precision was lowest for the COHIP-OSS and NOSE questionnaire. These observations suggest that the facial function and oral health domains are not sufficiently covered by the CLEFT-Q eating and drinking, NOSE, and COHIP-OSS, and these questionnaires may not be accurate enough to stratify cleft-related outcomes.

In **Chapter 3**, the psychosocial function outcome instruments as defined by the ICHOM Standard, i.e the CLEFT-Q psychological, CLEFT-Q social, and CLEFT-Q school function scales, were extensively explored using correlational analyses. Prospectively collected data from 3,067 patients with CL/P were categorized into five time points of measurement: 8-9 years, 10-13 years, 14-16 years, 17-19 years and 20-22 years were included. Strong correlations were observed between social and psychological and school function scales. Correlation between school and psychological function was lower, suggesting that the CLEFT-Q social scale might be redundant in measuring the psychosocial concept. The presence of a genetic syndrome was a significant predictor for referral to psychosocial care. Linear regression revealed a negative significant association between time points and outcome scores of the psychological function scale; a higher age group was associated with lower scores.

Chapter 4 and **Chapter 5** investigated the use of the CAT version of the CLEFT-Q to reduce burden for both patient and healthcare provider. The length of all eight CLEFT-Q scales in the ICHOM Standard Set combined was reduced from 76 to 59 items. CAT assessments reproduced full-length CLEFT-Q scores accurately. Linear assessment scores generated by Rasch models and unidimensional graded response models showed close agreement in a simulated dataset and were closely reproduced in the real patient dataset. The Concerto platform was perceived to improve clinical communication and facilitate shared decision making.

To identify barriers and facilitators to the international implementation of the ICHOM Standard Set for cleft, a two-part qualitative study has been conducted and described in **Chapter 6 and Chapter 7**. The study consisted of an exploratory survey among clinicians, health information technology professionals, and project coordinators, and semi-structured interviews of project leads from 4 cleft centers from The Netherlands, USA and Sweden. Thematic content analysis was performed, with organization of themes according to the dimensions of the reach, effectiveness, adoption, implementation and maintenance (RE-AIM) framework. Results showed that teams reach patients either via email or during the clinic visit to capture patient-reported outcomes. Adopting routine data collection is enhanced by aligning priorities at the organizational and cleft team level. Streamlining workflows and developing an efficient data collection platform are necessary early on, followed by pilot testing or stepwise implementation. Regular meetings and financial resources are crucial for implementing, sustaining, analyzing collected data, and providing feedback to healthcare professionals and patients. Fostering patient-centered care was articulated as a positive outcome in effectiveness, whereas time presented challenges across all RE-AIM dimensions. The identified themes can inform ongoing implementation efforts. Multisite collaboratives may assist in facilitating implementation. Intentionally investing time to lay a sound foundation early on will benefit every phase of implementation and help overcome barriers such as lack of support or motivation.

Chapter 8 described a retrospective cohort study of 40 patients with unilateral cleft lip and palate, aged 0 to 24, treated between 2012 and 2019 at Erasmus University Medical Center. Healthcare services, including consultations, diagnostic and surgical procedures, were counted and costs were calculated. Expected costs based on treatment protocol were calculated by multiplying healthcare products by product prices. Mean observed total costs (€40,859) for the complete treatment (0-24 years) were 1.6 times the expected protocol-based costs (€25,198) due to optional, non-protocolized procedures. Hospital admissions including surgery were main cost drivers. Implementing the ICHOM Standard Set increased protocol-based costs by 7%.

Chapter 9 discussed the study results with its implications and provided recommendations to move forward. The first half of this thesis, focused on the challenges around outcome measurement in cleft care. Developing and sustaining an outcomes framework is an iterative process of evaluating, adjusting and (re-) implementing the chosen outcome instruments in real practice. The use of a Computerized Adaptive Test version of scales could help reduce registration burden for both patients and healthcare professionals and could stimulate the uptake and compliance of measuring outcomes.

The second half of this thesis reviewed the implementation of the ICHOM Standard Set for cleft care in clinical practice. As implementation efforts are often constrained by time and health information technology, it is important for teams to collaborate with international initiatives to accelerate the implementation by sharing their knowledge on a regular basis and providing access to already developed HIT-systems or registries. Locally analyzing healthcare utilization patterns can provide an understanding on how and where healthcare services are used or needed. This thesis concludes that transforming care comes with great challenges, but together we can face and overcome these hurdles to ultimately provide the best possible care to our patients with a cleft.

Samenvatting

Hoofdstuk 1, de algemene inleiding, geeft een overzicht van de achtergrond en doelstellingen van dit proefschrift. In 2016 werd de ICHOM Standard Set voor patiënten met schisis voor het eerst geïmplementeerd in de kliniek in het Erasmus MC, gevolgd door verschillende andere ziekenhuizen in de Verenigde Staten en Zweden. De Standard Set bevat, naast klinische indicatoren en case-mixvariabelen, negen CLEFT-Q-schalen, de COHIP-OSS, de NOSE-vragenlijst en de Intelligibility in Context Scale om het perspectief van de patiënt op zijn gezondheid te meten. Aangezien het gebruik van uitkomstinstrumenten in de schisispraktijk een relatief nieuw en toenemend gebruik is, is een evaluatie nodig om te bepalen wat het oplevert, waar en hoe het kan worden verbeterd.

Dit proefschrift behandelt verschillende aspecten met betrekking tot de uitdagingen rondom het meten van patiënt-gerapporteerde uitkomsten bij patiënten met een schisis en de uitdagingen rondom de implementatie van de uitkomstmaten in de klinische schisiszorg. De specifieke onderzoeksvragen in dit proefschrift zijn:

1. Hoe kunnen we het meten van patiënt-gerapporteerde uitkomsten in de ICHOM Standard Set voor schisis optimaliseren?
    a. Hoe is de psychometrische prestatie en vertegenwoordiging van concepten van de patiënt-gerapporteerde uitkomstmaten van de ICHOM Standard Set voor schisis?
    b. Hoe kunnen we informatie maximaliseren en tegelijkertijd de registratielast verminderen bij het meten van de psychosociale functie met de ICHOM Standard Set voor schisis?
    c. Wat is de externe validiteit van de CLEFT-Q Computerized Adaptive Test (CAT) bij patiënten met een schisis?
2. Hoe kunnen we de implementatie van de ICHOM Standard Set voor schisis in de klinische praktijk optimaliseren?
    a. Wat zijn motiverende en belemmerende factoren voor de implementatie van de ICHOM Standard Set voor schisis in de klinische praktijk?
    b. Wat zijn de zorggebruik- en medische kostenpatronen van klinische schisiszorg en hoe wordt dit beïnvloed door het gebruik van de ICHOM Standard Set voor schisis?

Hoofdstuk 2 beschrijft de Rasch-analyse van de patiënt-gerapporteerde uitkomstmaten in de ICHOM Standard Set voor schisis om mogelijke hiaten in de concepten te identificeren. Patiënt-gerapporteerde uitkomsten van 714 patiënten met schisis in de leeftijd van 8 tot 9, 10 tot 12,5 en 22 jaar werden verzameld in een tijdsbestek van vier jaar. De Rasch-analyse toonde aan dat de schalen met betrekking tot de concepten 'facial appearance', 'speech function' en 'psychosocial function' goed functioneren met hoge betrouwbaarheidsparameters. Er werden echter specifieke problemen met individuele items binnen de CLEFT-Q eating & drinking, NOSE en COHIP-OSS opgemerkt, zoals een slechte 'item-fit' met het Rasch-model en 'disordered thresholds'. De meetnauwkeurigheid was het laagst voor de COHIP-OSS en de NOSE-vragenlijst. Deze waarnemingen suggereren dat de concepten 'facial function' en 'oral health' onvoldoende worden gedekt door de CLEFT-Q eating & drinking, NOSE en COHIP-OSS instrumenten, en deze vragenlijsten zijn mogelijk niet nauwkeurig genoeg om schisis-gerelateerde uitkomsten te stratificeren.

In **Hoofdstuk 3** worden de uitkomstinstrumenten voor psychosociale functie zoals gedefinieerd door de ICHOM Standard Set, dat wil zeggen de CLEFT-Q psychological function, CLEFT-Q social function en CLEFT-Q school function, uitgebreid onderzocht met behulp van correlatieanalyses. Prospectief verzamelde gegevens van 3.067 patiënten met schisis werden onderverdeeld in vijf meetmomenten: 8-9 jaar, 10-13 jaar, 14-16 jaar, 17-19 jaar en 20-22 jaar. Er werden sterke correlaties waargenomen tussen de social en de psychological en school function (schalen. De correlatie tussen school en psychological function was lager wat suggereert dat de CLEFT-Q social function overbodig zou zijn bij het meten van het psychosociale concept. Het hebben van een genetisch syndroom was een significante voorspeller voor verwijzing naar psychosociale zorg. Lineaire regressie onthulde een negatief significant verband tussen tijdstippen en uitkomstscores van de psychologische functieschaal; een hogere leeftijdsgroep ging gepaard met lagere scores.

Hoofdstuk 4 en **Hoofdstuk 5** onderzochten het gebruik van de CAT-versie van de CLEFT-Q om de registratielast voor zowel patiënt als zorgverlener te verminderen. De lengte van alle acht CLEFT-Q schalen in de ICHOM Standard Set samen werd teruggebracht van 76 naar 59 items. CAT-beoordelingen reproduceerden nauwkeurige CLEFT-Q-scores over de volledige lengte. De lineaire beoordelingsscores gegenereerd door het Rasch-model en het unidimensionale graded respons model vertoonden een grote overeenkomst in de simulatiedata en werden nauwkeurig gereproduceerd in de echte patiënten-dataset. Het Concerto-platform zou de klinische communicatie verbeteren en de gedeelde besluitvorming faciliteren.

Om motiverende en belemmerende factoren te identificeren die de internationale implementatie van de ICHOM Standard Set voor schisis in de weg staan, is een tweedelig kwalitatief onderzoek uitgevoerd en beschreven in **Hoofdstuk 6** en **Hoofdstuk 7**. Het onderzoek bestond uit een verkennend onderzoek onder clinici, ICT-specialisten, en projectcoördinatoren, en semi-gestructureerde interviews met projectleiders van 4 schisiscentra uit Nederland, de VS en Zweden. Er werd een thematische analyse uitgevoerd, met een organisatie van thema's volgens de dimensies van 'reach, effectiveness, adoption, implementation en maintenance' (RE-AIM). De resultaten toonden aan dat teams patiënten via e-mail of tijdens het bezoek aan de kliniek benaderen voor het vastleggen van de patiënt-gerapporteerde resultaten. Routinematige gegevensverzameling wordt verbeterd door prioriteiten op organisatie- en schisisteam niveau op elkaar af te stemmen. Het stroomlijnen van workflows en het ontwikkelen van een efficiënt platform voor gegevensverzameling zijn in een vroeg stadium noodzakelijk, gevolgd door pilot-testen of stapsgewijze implementatie. Regelmatige vergaderingen en financiële middelen zijn cruciaal voor het implementeren, onderhouden, analyseren van de verzamelde gegevens en het geven van feedback aan zorgprofessionals en patiënten. Het bevorderen van patiëntgerichte zorg werd benadrukt als een positief resultaat in effectiviteit, terwijl de tijd voor uitdagingen zorgde in alle RE-AIM-dimensies. De geïdentificeerde thema's kunnen de lopende implementatie inspanningen ondersteunen. Samenwerkingen tussen verschillende centra kunnen helpen bij het vergemakkelijken van de implementatie. Opzettelijk tijd investeren om in een vroeg stadium een solide basis te leggen, zal elke fase van de implementatie ten goede komen en barrières zoals een gebrek aan ondersteuning of motivatie helpen overwinnen.

Hoofdstuk 8 beschrijft een retrospectieve cohortstudie van 40 patiënten met een eenzijdige lip-, kaak- en gehemeltespleet in de leeftijd van 0 tot 24 jaar, die tussen 2012 en 2019 werden behandeld in het Erasmus Universitair Medisch Centrum. Medische consultaties, diagnostische en chirurgische procedures werden geteld en de kosten werden berekend. Verwachte kosten op basis van het behandelprotocol werden berekend door zorgproducten te vermenigvuldigen met productprijzen. De gemiddelde geobserveerde totale kosten (€ 40.859) voor de volledige behandeling (0-24 jaar) waren 1,6 keer de verwachte kosten op basis van het protocol (€ 25.198) vanwege optionele, niet-geprotocolleerde procedures. Ziekenhuisopnames inclusief operaties waren de belangrijkste kostenposten. Het implementeren van de ICHOM Standard Set verhoogde de protocol-gebaseerde kosten met 7%.

Hoofdstuk 9 bespreekt de onderzoeksresultaten met de implicaties en aanbevelingen voor de toekomst. De eerste helft van dit proefschrift was gericht op de uitdagingen rond het meten van uitkomsten in de schisiszorg. Het ontwikkelen en onderhouden van een uitkomstenset is een iteratief proces van evalueren, aanpassen en (her) implementeren van de gekozen uitkomstinstrumenten in de praktijk. Het gebruik van een CAT-versie van meetinstrumenten kan helpen de registratielast voor zowel patiënten als zorgprofessionals te verminderen en kan de acceptatie van het gebruik van meetresultaten stimuleren.

De tweede helft van dit proefschrift besprak de implementatie van de ICHOM Standard Set voor schisiszorg in de klinische praktijk. Aangezien implementatie pogingen vaak worden beperkt door tijd en ICT, is het belangrijk voor teams om samen te werken met internationale initiatieven om de implementatie te versnellen door hun kennis regelmatig te delen en toegang te bieden tot reeds ontwikkelde informatiesystemen of registers. Het lokaal analyseren van gebruikspatronen van de zorg kan inzicht geven in hoe en waar specifieke zorg wordt gebruikt of nodig is. Tot slot besluit dit proefschrift dat het transformeren van zorg gepaard gaat met grote uitdagingen, maar dat we door samen te werken deze hindernissen kunnen overwinnen om uiteindelijk de best mogelijke zorg te bieden aan onze patiënten met schisis.

Propositions

1. Developing an outcomes set is an iterative process of evaluating, adjusting and (re-) implementing the chosen outcome instruments in real practice. *This thesis*

2. In addition to the standardized outcomes framework, cleft teams should be given the freedom to implement additional outcome measures according to their specific wishes or patient population. *This thesis*

3. The use of a Computerized Adaptive Test (CAT) can be helpful in reducing registration burden for both patients and healthcare professionals. *This thesis*

4. Teams should collaborate with other teams and international initiatives to overcome IT-related implementation barriers together. *This thesis*

5. Treatment protocols within cleft care are suboptimal predictors of actual healthcare utilization. *This thesis*

6. Clinicians need to ensure that rating scales are fit for purpose, and maximising the scientific rigour of rating scales improves the chances of coming to the correct conclusion about the efficacy of a treatment. *Hobart JC et al., Lancet Neurology, 2007*

7. Research on the link between initial conditions and outcomes is essential, not a distraction, because it informs the factors that affect the success of care and reveals avenues for learning and innovation. *Porter ME, Annals of Surgery, 2008*

8. In order to prepare our [orthopedic] trainees to survive in a value-based healthcare environment, we must expose them to and educate them about value-based programs. *Murrey DB et al., The Journal of Bone and Joint Surgery, 2021*

9. Doctors aren’t burned out from overwork. We are demoralized by our health care system. *Reinhart E, The New York Times, 2023*

10. Our attention has never been as overwhelmed as it is today and we’ve never been so busy while accomplishing so little. *Chris Bailey, Hyperfocus, 2018*

11. If you are always trying to be normal, you will never know how amazing you can be. *Maya Angelou*