Clin Orthop Relat Res (2010) 468:815–826
DOI 10.1007/s11999-009-1036-y

SURVEY (SYSTEMATIC REVIEW)

Which Are the Most Frequently Used Outcome Instruments in Studies on Total Ankle Arthroplasty?

Florian D. Naal MD, Franco M. Impellizzeri PhD, Pascal F. Rippstein MD

Received: 17 November 2008 / Accepted: 28 July 2009 / Published online: 12 August 2009

Abstract The number of studies reporting on outcomes after total ankle arthroplasty is continuously increasing. As the use of valid outcome measures represents the cornerstone for successful clinical research, we aimed to identify the most frequently used outcome instruments in ankle arthroplasty studies and to analyze the evidence to support their use in terms of different quality criteria. A systematic review of the literature identified 15 outcome instruments reported in 79 original studies. The most commonly used measures were the American Orthopaedic Foot and Ankle Society hindfoot score (n = 41), the Kofoed ankle score (n = 21), a visual analog scale assessing pain (n = 15), and the generic SF-36 (n = 6). Eight additional instruments were used only once or twice. The American Orthopaedic Foot and Ankle Society and Kofoed instruments include a clinical examination and score up to 100 points. Evidence to support their use in terms of validity, reliability, responsiveness, and interpretability is limited, raising the question whether their use is justified. Self-reported questionnaires related to ankle osteoarthritis or arthroplasty are rather disregarded in the current literature, and only the Foot Function Index is associated with evidence in terms of the above-mentioned quality criteria. Future research is warranted to improve the outcome assessment after total ankle arthroplasty.

Each author certifies that he or she has no commercial associations (eg, consultancies, stock ownership, equity interest, patent/licensing arrangements, etc) that might pose a conflict of interest in connection with the submitted article.

F. D. Naal (✉), P. F. Rippstein
Department of Orthopaedic Surgery, Foot and Ankle Center, Schulthess Clinic, Lengghalde 2, 8008 Zurich, Switzerland
e-mail: florian.naal@gmail.com

F. M. Impellizzeri
Department of Research and Development, Schulthess Clinic, Zurich, Switzerland

Introduction

Total ankle arthroplasty (TAA) has evolved during the past decades. High failure rates and discouraging clinical outcomes of first-generation implants in the 1970s resulted in restricted use but also led to the development of modern three-component implants [15, 17]. These designs allow for flexion and extension and for rotational and sliding movements, resulting in improved congruency, reduced shear forces at the bone-implant interface, and less bone removal during implantation [32]. As a consequence, clinical studies suggest improved outcomes and likely as a result apparently increased interest in this procedure [15, 17]. Considering this evolution, the number of outcome reports will increase and quality outcome measures should be used to reflect high-quality research.

Numerous instruments, clinician-generated and self-reported, are available to assess outcomes after foot and ankle surgery. In 2004, Button and Pinney [6] identified 49 rating scales of which 18 were used more than once, but the authors stated none of these measures had demonstrable reliability, validity, and responsiveness in patients with various foot and ankle disorders. Similarly, Parker et al. [33] noted the fundamental problem of existing instruments used to evaluate foot and ankle surgery is their limited exploration and evaluation of what patients perceive to be most important in their outcomes. They concluded none of the existing measures could claim to be valid for patient perceptions of outcome [33]. More recently, Martin and Irrgang [27] comprehensively surveyed self-reported outcome instruments for various foot and ankle disorders and

identified 14 different instruments. Five of these measures had some evidence to support their use relative to content validity, construct validity, reliability, and responsiveness [27]. These reports, however, focused on available scores to assess the outcomes of foot and ankle surgery in general or focused only on self-reported questionnaires and did not comprehensively consider outcome instruments in TAA.

Our systematic review, therefore, addressed the following questions: (1) Which are the most frequently used outcome instruments in studies reporting on TAA? (2) Does the literature provide evidence to support their use in terms of validity, reliability, responsiveness, and interpretability?

Search Strategies and Criteria

Initially, we wrote a protocol defining the objectives, search terms, inclusion and exclusion criteria, and the methods of documentation based on the method described by Wright et al. [45]. According to the protocol, the electronic databases MEDLINE through PubMed, EMBASE, and Cochrane (all until May 2009) were searched using the following search terms: "ankle arthroplasty" OR "ankle replacement" AND "outcome" OR "results" OR "score" OR "questionnaire". The reference lists of all included articles and recent reviews were checked manually for additional relevant studies. We included all articles that (1) reported on TAA; (2) used a specifically defined outcome instrument (either clinician-generated or self-reported, eg, American Orthopaedic Foot and Ankle Society [AOFAS] hindfoot score [21] or Foot Function Index [FFI] [4]); and (3) were published in English, German, French, Italian, or Spanish. We excluded articles that (1) reported on ankle fusion or conversion of TAA to fusion; (2) single or a few cases; (3) used no specific outcome instrument (eg, simple outcome rating as good, fair, or poor); and (4) were published in a language other than one of those previously mentioned.

After removing duplicates, we obtained 763 citations from the searches (Fig. 1). Two independent reviewers (FDN, FMI) then screened titles, abstracts, and full texts against the inclusion and exclusion criteria. Disagreements during each step of the review process were discussed and resolved. Four hundred twenty-six citations appeared to be irrelevant by title, and 221 citations by abstract. The full texts of the remaining 116 references were further analyzed. Forty-five studies did not match the selection criteria and were removed. Eight studies were added after reviewing reference lists and recent reviews. Finally, we included 79 original articles referring to 15 different instruments (Fig. 1). We did not intend to calculate cumulative score values as part of a meta-analysis.

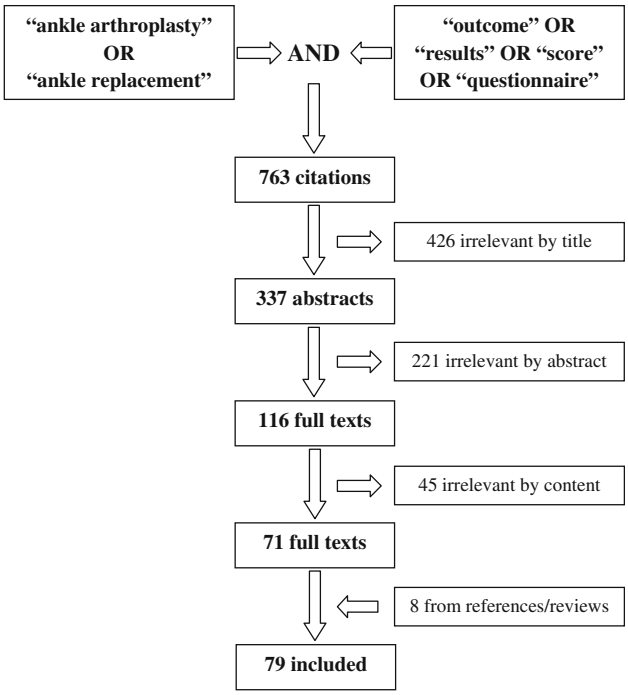

Fig. 1 This flowchart shows the steps of the review process from the search terms to the number of finally included articles.

In the next step, the reference lists of the included articles were checked for references of studies that developed or first described the identified outcome instrument or that investigated its psychometric properties in patients with ankle disorders. Additionally, we searched the above-mentioned electronic databases for the following search terms: "name of identified instrument" AND "ankle" AND "validity" OR "validation" OR "reliability" OR "responsiveness" OR "interpretability". The reference lists of review articles also were analyzed. According to established quality criteria to examine the measurement properties of outcome instruments [40–42], we retrieved information on reproducibility, internal consistency, content validity, criterion validity, construct validity, responsiveness, and interpretability of the identified region- and disease-specific outcome instruments.

To investigate the methodologic quality and results of these clinimetric studies, we used the checklist described by van der Leeden et al. [41] to assign different levels of evidence (Levels 1 to 3) for the following criteria. Reproducibility refers to the degree to which repeated measurements in clinically stable individuals (test-retest) provide similar results. A distinction can be made between reliability and agreement [8]. The intraclass correlation coefficient (ICC) and kappa statistics are adequate measures to assess reliability [8]. Similar to van der Leeden et al. [41], we assigned a Level 1 rating if an ICC or kappa value was reported for a sample size of at least 50 patients.

A positive rating was assigned if these values were greater than 0.70, and a negative for lower values. We assigned a Level 2 rating if a Pearson correlation coefficient was reported for a sample size of also at least 50 patients. Coefficients greater than 0.80 were rated positively, and lower values negatively. A Level 3 evidence was assigned if an ICC, a kappa value, or a Pearson coefficient was reported for a sample size of less than 50 patients. ICC or kappa values greater than 0.80 or Pearson coefficients greater than 0.90 were assigned a positive rating, and lower values a negative rating [41]. Agreement refers to the absolute measurement error of an instrument and therewith describes the precision of this instrument [8]. The standard error of the measurement (SEM), Bland and Altman's limits of agreement, or the smallest detectable change (SDC) are adequate measures to assess agreement [8]. The absolute measurement error should be smaller than the minimal clinically important difference (MCID) of score values in individuals with time [40]. Therefore, the MCID of an instrument or subscale should be defined. We assigned a Level 1 rating for agreement if the limits of agreement, SEMs, or SDCs were determined in a sample size of at least 50 patients, and if the MCID was reported. A Level 2 rating was assigned if the sample size was less than 50 patients. Values for the limits of agreement, SEM, or SDC below the MCID were rated positively, and values above the MCID were rated negatively. A Level 3 rating was assigned if the MCID was not defined [41].

Internal consistency describes the homogeneity of an instrument or its (sub)scales. It is an important quality criterion for instruments that intend to measure one concept or construct [40]. A Level 1 evidence was designated if factor analysis was performed in a sample size of at least seven times the number of items and in a minimum of 100 patients. Additionally, the Cronbach's alpha (CA) had to be reported for each of the instruments' (sub)scales. A Level 1 rating also was assigned if Rasch analysis was used and the methodology was completely defined. A Level 2 evidence was assigned if the descriptive information on the Rasch methodology was incomplete. A Level 2 rating also was assigned if factor analysis was used in a sample size of at least four patients per item and greater than 50 patients in total, and a Level 3 rating needed four patients per item and a total sample size less than 50 patients. A CA greater than 0.70 relates to a positive rating, and lower CA values to a negative rating [41].

Content validity examines the extent to which the concept of interest is measured or represented by the items of an instrument [14]. In the present context, items must reflect areas of importance of patients with ankle osteoarthritis. A positive rating therefore was assigned if patients were involved during the item generation and selection [41].

Criterion validity examines the extent to which an instrument is related to a gold standard [40]. In the present context, we are not aware of any instrument that really can be considered a gold standard related to ankle osteoarthritis and ankle arthroplasty. Considering gold standards used in other scientific areas (eg, doubly labeled water or total oxygen uptake for the quantification of physical activity), we defined no specific instrument (such as the SF-36) as the gold standard in this review, but physical performance tests closely related to ankle function (eg, single heel lifts). A Level 1 rating was assigned if hypotheses concerning expected relationships between the instrument and the gold standard were specified in advance, and these relationships were investigated in a sample of at least 50 patients. A Level 2 rating was assigned if no specific hypotheses were specified beforehand. A Level 3 evidence was designated if plausible relations were found in a sample of less than 50 patients. A positive rating needed a correlation with the gold standard greater than 0.70, and weaker correlations received a negative rating [41].

Construct validity refers to the extent an instrument correlates with other measures in a manner consistent with theoretically derived hypotheses concerning the concepts being measured [20]. Predefined hypotheses regarding these relationships should be specified as precisely as possible. Construct validity can be determined by convergent and divergent (or discriminant) validity. Evidence of convergent validity is provided by moderate to high correlations with other instruments measuring the same construct. In contrast, there should be no or only weak associations with instruments measuring different constructs (divergent validity) [20]. A Level 1 rating was assigned if hypotheses concerning expected correlations with other instruments were specified in advance, and these relationships were investigated in a sample of at least 50 patients. A positive rating needed confirmation of at least 75% of these hypotheses. A Level 2 rating was assigned if no specific hypotheses were specified beforehand and the sample size was greater than 50. A Level 3 evidence was designated if plausible relations with other measures were found in a sample of less than 50 patients [41].

Responsiveness refers to the ability of an instrument to detect clinical changes with time. It can be considered an aspect of longitudinal validity [13]. The effect size (ES), standardized response mean (SRM), and area under the receiver operating characteristics curve (AUC) are adequate measures to assess responsiveness [40]. As responsiveness refers to longitudinal validity, predefined hypotheses also should be specified [40]. The levels of evidence, therefore, were assigned similarly to construct validity [41]. A positive rating was given if the ES or SRM was greater than 0.8 (high responsiveness) or the AUC was greater than 0.70 [36].

Floor and ceiling effects occur when patients score the lowest or highest score on an instrument, respectively. As a consequence, clinical deterioration or improvement cannot be assessed, and patients scoring lowest or highest possible cannot be distinguished from each other. Floor and ceiling effects were considered present if greater than 15% of patients achieved the lowest or highest score, respectively [40]. A positive rating was assigned for the absence of floor or ceiling effects [41].

Interpretability refers to the degree to which one can assign qualitative meaning to quantitative scores [40]. Interpretability, therefore, is related to the MCID, which should be determined in a sample size of at least 50 patients to receive a positive rating [41].

Results

We identified 15 distinct outcome instruments used to determine the clinical outcome of TAA in 79 studies (Fig. 2). Eight of these measures were region specific, one was disease specific, three were generic, two were related to physical activity, and one could not be classified. The most commonly used instruments were the AOFAS hindfoot score (n = 41), the Kofoed ankle score (n = 21) [22, 23], the visual analog scale (VAS) assessing pain (n = 15), and the generic SF-36 (n = 6) [43]. Eight additional instruments were used only once or twice.

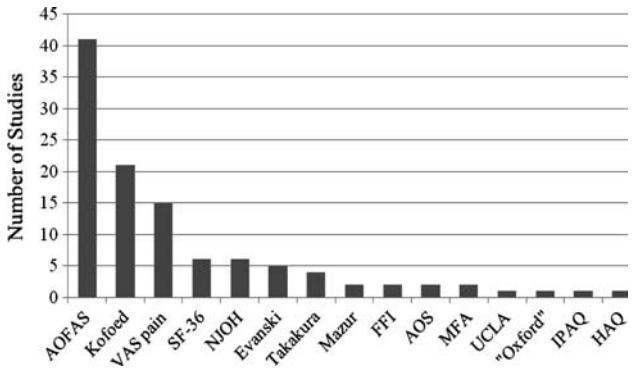

Fig. 2 This graph shows the distribution of the outcome instruments used among the TAA studies in this review. AOFAS = American Orthopaedic Foot and Ankle Society hindfoot score (region-specific) [21]; VAS pain = visual analog scale pain; SF-36 = Short Form 36 (generic) [43]; NJOH = New Jersey Orthopaedic Hospital ankle score (region-specific) [5]; Evanski = outcome instrument of Evanski and Waugh [11]; Takakura = outcome instrument of Takakura et al. [39]; Mazur = outcome instrument of Mazur et al. [29]; FFI = Foot Function Index (region-specific) [3]; AOS = Ankle Osteoarthritis Scale (disease-specific) [9]; MFA = Musculoskeletal Functional Assessment (generic) [10]; UCLA = University of California at Los Angeles activity scale (physical activity assessing) [31]; "Oxford" = questionnaire (region-specific) developed by the authors [18], modeled to the original Oxford Hip Score [7]; IPAQ = International Physical Activity Questionnaire (physical activity assessing) [31]; HAQ = Health Assessment Questionnaire (generic) [12].

We identified 13 articles providing information on quality criteria for the previously identified region- and disease-specific questionnaires (Tables 1, 2). Except for the AOFAS hindfoot score, we found no information on quality criteria for all other clinician-based instruments. Five studies provided information on properties of the AOFAS hindfoot score, eight studies on the FFI, and one study on the Ankle Osteoarthritis Scale (AOS) [9] (Tables 1, 2). Evidence in terms of the different quality criteria was low for the AOFAS hindfoot score and the AOS, and we could assign moderate ratings for the FFI (Table 3). None of the instruments provided evidence in terms of interpretability attributable to unknown MCIDs.

Discussion

TAA has evolved during the past decades and the improvements in survivorship and clinical outcomes have led to the development of various new implants [15, 17]. Assessment of outcomes using appropriate outcome measures is the cornerstone of successful clinical research and allows for comparisons of patients' function and different treatment modalities or implants. Considering the growing popularity of TAA and therewith the increasing need for quality outcome research, this review addressed the following questions: (1) Which are the most frequently used outcome instruments in studies reporting on TAA? (2) Does the literature provide evidence to support their use in terms of validity, reliability, responsiveness, and interpretability?

Some limitations and aspects must be considered before interpreting our results. First, we did not include all electronic databases in the systematic search. Therefore, some instruments used in studies reporting on TAA outcomes may have been missed. Second, the definitions of criterion validity and construct validity have been used rather confusingly in the different studies. Budiman-Mak et al. [3, 4], for example, defined the "50 feet walking time test" as a measure for criterion validity in their first paper on the FFI [4] but as a measure for construct validity in their last paper on the FFI [3]. The SF-36 and the WOMAC [2] also have been used to measure criterion validity [9], whereas the SF-36 was used to evaluate construct validity in another study [30]. We doubt if any generic tool, such as the SF-36, Musculoskeletal Function Assessment (MFA) [10], or Quality Adjusted Life Year (QUALY) [35], could really be considered a gold standard for patients undergoing TAA. Similarly, the WOMAC might be considered the gold standard instrument for osteoarthritis, but only for patients having hip and knee arthroplasties. In this review, we,

Table 1. Reproducibility, internal consistency, and floor and ceiling effects*

Instrument/study	Reproducibility					Internal consistency				Floor and ceiling effects†
	Reliability	Sample size	Agreement	Sample size	MCID	Factor analysis	Cronbach's alpha	Rasch analysis	Sample size	
AOFAS hindfoot score										
Ibrahim et al. [19] (2007)‡	Test-retest mean scores 45 and 49 points ($p = 0.27$)	35								
Pena et al. [34] (2007)										No floor effects, mild ceiling effects§
SooHoo et al. [37] (2003)										No floor or ceiling effects
FFI										
Budiman-Mak et al. [4] (1991) (FFI)	ICC 0.87 (total)	39				Yes	0.96 (total)		86	
	ICC 0.70 (pain)	39					0.93 (pain)			
	ICC 0.81 (limit)	40					0.73 (limit)			
	ICC 0.84 (disab)	40					0.95 (disab)			
Budiman-Mak et al. [3] (2006) (FFI-R)	0.96 (person-alpha)	92					0.95	Yes	92	
	0.93 (item-alpha)									
Agel et al. [1] (2005) (FFI)	68.8% responded within 1 point between 1st and 2nd questionnaire administration	54								Floor effects: 0%–10% (pain) 0%–14% (limit) 0%–16% (disab) Ceiling effects: 6%–40% (pain) 15%–82% (limit) 4%–41% (disab)
Kuyvenhoven et al. [24] (2002) (FFI-5pt, Dutch)	ICC 0.81 (total)	206				Yes	0.93 (total)		206	
	ICC 0.70 (pain)						0.88 (pain)			
	ICC 0.83 (disab)						0.92 (disab)			
Naal et al. [30] (2008) (FFI-D, German)	ICC 0.98 (total)	20	-0.2 ± 2.1‖(total)	20		No	0.97 (total)		53	
	ICC 0.97 (pain)		0.3 ± 2.5‖(pain)				0.90 (pain)			
	ICC 0.99 (disab)		-0.6 ± 2.9‖(disab)				0.95 (disab)			

Table 1. continued

Instrument/study	Reproducibility					Internal consistency				Floor and ceiling effects†
	Reliability	Sample size	Agreement	Sample size	MCID	Factor analysis	Cronbach's alpha	Rasch analysis	Sample size	
Wu et al. [46] (2008) (FFI, Taiwan Chinese)	ICC 0.82 (total) ICC > 0.80 (limit) ICC 0.70–0.80 (pain, disab)	24					0.94 (total) 0.91 (pain) 0.75 (limit) 0.95 (disab)		79	Floor effects: 0%–22% (pain) 0%–10% (limit) 0%–21% (disab) Ceiling effects: 0%–34% (pain) 10%–98% (limit) 4%–30% (disab)
AOS										
Domsic and Saltzman [9] (1998)	ICC 0.97 (total) ICC 0.95 (pain) ICC 0.94 (disab)	28								

* No studies reported reproducibility, internal consistency, and floor and ceiling effects for the following instruments: Kofoed [22, 23], Evanski and Waugh [11], New Jersey Orthopaedic Hospital [5], Takakura et al. [39], Mazur et al. [29], and "Oxford" [18]; †floor effects defined as clinically worst scores and ceiling effects vice versa; ‡only subjective part; §not quantified; ‖values expressed as mean ± standard deviation; MCID = minimal clinically important difference; AOFAS = American Orthopaedic Foot and Ankle Society [21]; FFI = Foot Function Index [4]; AOS = Ankle Osteoarthritis Scale [9]; ICC = intraclass correlation coefficient; total = total score; pain = pain subscale; limit = limitations subscale; disab = disability subscale.

Table 2. Content validity, criterion validity, construct validity, and responsiveness*

Instrument/ study	Content validity	Criterion validity			Construct validity			Responsiveness		
	Item generation and instrument rationale	Hypotheses	Results	Sample size	Hypotheses	Results	Sample size	Hypotheses	Results	Sample size
AOFAS hindfoot score										
Ibrahim et al. [19] (2007)†					No	Moderate correlation with FFI ($r = -0.68$)	45			
Malviya et al. [26] (2007)					No	Weak to moderate correlations (coefficient of determination) with QUALY scores ($r^2 = 0.47$ preoperatively, $r^2 = 0.33$ after 6 months, $r^2 = 0.22$ after 12 months)	40			
Pena et al. [34] (2007)					No	Weak and inconsistent association between alignment component and MFA domains (up to $r = -0.32$); moderate but inconsistent associations between pain and function components and MFA domains (up to $r = -0.65$)	84–154	No	6 months after TAA: ES 2.15‡ 24 months after TAA: ES 2.39‡	109 84
SooHoo et al. [37] (2003)					Yes	Weak to moderate correlations with SF-36 domains and summary scales (up to $r = 0.58$)	48			
SooHoo et al. [38] (2006)								Yes	6 months after surgery§: ES 1.12, SRM 1.10	25
FFI										
Budiman-Mak et al. [4] (1991) (FFI)	Expert panel developed items	Yes	Moderate correlation with 50 feet walking time ($r = 0.48$) and painful foot joint count ($r = 0.53$); lower correlation with painful hand joint count ($r = 0.33$) and grip strength ($r = -0.47$)	57–87				No	Moderate correlation with change in painful foot joint count 6 months later	42
Budiman-Mak et al. [3] (2006) (FFI-R)	Expert panel revised items after literature search and patient interviews	Yes	Moderate correlation with 50 feet walking time ($r = 0.31$)	92						

Table 2. continued

Instrument/ study	Content validity	Criterion validity			Construct validity			Responsiveness		
	Item generation and instrument rationale	Hypotheses	Results	Sample size	Hypotheses	Results	Sample size	Hypotheses	Results	Sample size
Kuyvenhoven et al. [24] (2002) (FFI-5pt, Dutch)	Revision based on patient responses				No	Significant differences between patients with high or low SF-36 score values	206	No	Low to moderate by analyzing score changes after 8 weeks in patients subjectively improved, unchanged, or deteriorated	206
Naal et al. [30] (2008) (FFI-D, German)	Revision based on patient perceptions				Yes	Moderate to high correlations with SF-36 domains and summary scales (up to r = −0.80, total; up to r = −0.75, pain; up to r = −0.76, disab); high correlations with VAS pain and VAS function (up to r = 0.81); moderate correlations with UCLA activity scale (up to r = −0.56)	53			
SooHoo et al. [36] (2006) (FFI)					Yes	Moderate correlations with SF-36 domains and summary scales (up to r = −0.61, pain; up to r = −0.64, limit; up to r = −0.67, disab)	96			
SooHoo et al. [38] (2006) (FFI)								Yes	6 months after surgery: ES −0.86 (pain) ES −0.55 (limit) ES −0.75 (disab) SRM −0.83 (pain) SRM −0.39 (limit) SRM −0.68 (disab)	25
Wu et al. [46] (2008) (FFI, Taiwan Chinese)					No	Moderate correlations with SF-36 domains and summary scales (up to r = −0.69, total; up to r = −0.51, pain; up to r = −0.68, limit; up to r = −0.63, disab)	79			

Table 2. continued

Instrument/study	Content validity	Criterion validity			Construct validity			Responsiveness		
	Item generation and instrument rationale	Hypotheses	Results	Sample size	Hypotheses	Results	Sample size	Hypotheses	Results	Sample size
AOS										
Domsic and Saltzman [9] (1998)	Scale modified by the authors from the FFI to adapt for patients with ankle osteoarthritis	No	Moderate to high correlations with single heel lifts (r = 0.88, total; r = 0.63, pain; r = 0.90 disab)	10	No	Moderate to high correlations with WOMAC (r = 0.79, pain; r = 0.65, disab) and SF-36 domains and subscales (up to r = −0.66)	15			

* No studies reported content validity, criterion validity, construct validity, and responsiveness for the following instruments: Kofoed [22, 23], Evanski and Waugh [11], New Jersey Orthopaedic Hospital [5], Takakura et al. [39], Mazur et al. [29], and "Oxford" [18]; †only subjective part; ‡calculated from the data available; §included AOFAS fore- and midfoot scores for different types of surgery; AOFAS = American Orthopaedic Foot and Ankle Society [21]; FFI = Foot Function Index [4]; QUALY = Quality Adjusted Life Year [35]; AOS = Ankle Osteoarthritis Scale [9]; MFA = Musculoskeletal Functional Assessment [10]; VAS = visual analog scale; UCLA = University of California at Los Angeles activity scale [31]; total = total score; pain = pain subscale; limit = limitations subscale; disab = disability subscale; TAA = total ankle arthroplasty; ES = effect size; SRM = standardized response mean.

therefore, considered these different instruments and questionnaires as being related to construct validity, and we defined physical performance tests closely linked to ankle function as being related to criterion validity. The VAS pain was the third most frequently used instrument in original studies reporting on TAA, but it is neither foot nor ankle specific and represents only a single-item construct. The SF-36 [43], MFA [10], and Health Assessment Questionnaire [12] are generic tools assessing general health or health-related quality of life. These measures are self-reported, but they cannot be considered to adequately determine the region-specific health state of patients with ankle osteoarthritis or after TAA. It is beyond question these measures, similar to the physical activity-assessing University of California at Los Angeles activity scale or International Physical Activity Questionnaire [31], offer important additional information about the patients' health state. Nevertheless, the following discussion focuses on the identified region- and disease-specific outcome instruments.

Our review suggests the AOFAS hindfoot score [21] and the Kofoed ankle score [22, 23] are the two most frequently used outcome instruments in studies reporting on TAA. Both are clinician-based, region-specific 100-point scores, with 100 points reflecting the best clinical state. Their structure slightly differs in that the AOFAS attributes 40 points to the pain component, 50 points to the function component (including 16 points for hindfoot motion), and 10 points to hindfoot alignment. The Kofoed, in contrast, attributes 50 points to the pain component, only 30 points to the function component, and 20 points to range of motion (ROM). These differences illustrate, despite being 100-point scores, the absolute values cannot simply be compared between studies. The less frequently used instruments, ie, that of Evanski and Waugh [11] (pain 40, function 50, ROM 10), the New Jersey Orthopaedic Hospital ankle score [5] (pain 40, function 40, ROM 15, deformity 5), that of Takakura et al. [39] (40 pain, 40 function, 20 ROM), and that of Mazur et al. [29] (pain 50, function 40, ROM 10), are similarly clinician-based, region-specific 100-point rating systems, also with differences in pain, function, and ROM weightings.

However, for all these clinician-generated measures, except the AOFAS, the literature provides no evidence of validity, reliability, responsiveness, or interpretability of these scores. Also, no details on development strategies or the rationale of their structures could be identified, either in the original reports or in following studies. The lack of a clear theoretical framework behind these instruments also makes it difficult to interpret any evidence of content or construct validity as it is not clear what these instruments were supposed to measure. The World Health Organization, with its International Classification of Functioning,

Table 3. Levels of evidence regarding quality criteria for the region- and disease-specific instruments used in TAA studies

Instrument/study	Reliability	Agreement	Internal consistency	MCID	Content validity	Criterion validity	Construct validity	Floor and ceiling effects	Responsiveness
AOFAS hindfoot score [21]	0*	0	0	0	0	0	2	+	2+
FFI [4]	1+	3	1+	0	+	1−	1+	†	3±‡
AOS [9]	3+	0	0	0	0	3+	3	0	0
Kofoed [22, 23]	0	0	0	0	0	0	0	0	0
Evanski and Waugh [11]	0	0	0	0	0	0	0	0	0
NJOH [5]	0	0	0	0	0	0	0	0	0
Takakura et al. [39]	0	0	0	0	0	0	0	0	0
Mazur et al. [29]	0	0	0	0	0	0	0	0	0
"Oxford" [18]	0	0	0	0	0	0	0	0	0

* Reliability investigated, but methods insufficient to assign an evidence level; †severe ceiling effects for the limitation subscale, moderate floor and ceiling effects for the pain and disability subscales; ‡positive rating for the pain subscale, negative ratings for the limitation and disability subscales; 0 = no information available; 1 = Level 1 rating; 2 = Level 2 rating; 3 = Level 3 rating; + = positive rating; − = negative rating; MCID = minimal clinically important difference; AOFAS = American Orthopaedic Foot and Ankle Society; FFI = foot function index; AOS = Ankle Osteoarthritis Scale; NJOH = New Jersey Orthopaedic Hospital ankle score; "Oxford" = questionnaire developed by the authors, modeled to the validated Oxford Hip Score [7].

Disability and Health (ICF), proposed a conceptual model according to which items of a measure can be categorized [44]. The ICF identified three levels of human functioning: (1) body or body part, (2) whole person, and (3) whole person in a social context. Disability involves dysfunctioning at one or more of these three levels: impairments (problems in body function or structures), activity limitation (difficulties in executing activities), and participation restriction (problems in involvement in a life situation) [44]. Symptoms and clinical signs (ie, pain, ROM, alignment, etc) are related to the impairment domain, whereas activities of daily living such as self-care or sports are related to the activity limitation and participation restriction domains. As there is evidence that these different domains are not necessarily dependent on or correlated with each other, combining items of these domains into one score is questionable. This occurs, however, with all of the above-mentioned clinician-based outcome measures but also with the self-reported AOS [9] and FFI [3, 4].

The AOFAS hindfoot score has been the subject of concern before. Guyton [16] described several conceptual limitations of the AOFAS using Monte Carlo modeling. He pointed out, in addition to other drawbacks, the small number of answer categories in several subscales of the score is a major confounding factor leading to skewed data. He concluded the AOFAS cannot produce reliable data and score values obtained by parametric statistics must be interpreted with care [16]. SooHoo et al. [37] correlated the AOFAS score with the SF-36 and found only weak associations, suggesting poor construct validity of this instrument. Two other studies investigating the association between the AOFAS and the generic questionnaires QUALY and MFA also found only low correlations between these instruments [26, 34]. A greater association was found between the subjective part of the AOFAS and the FFI [19]. There are two other possible weaknesses of the AOFAS, as with the other clinician-based outcome measures. First, including a clinical examination in a score always introduces a possible confounder as different examiners might measure different things. Intrarater or interrater relations have so far not been determined for the AOFAS or the other clinician-determined instruments. Second, these rating systems might be not specific enough to measure TAA outcomes. The inclusion of objectively measured ROM represents a problem for patients who have, for example, an additional subtalar arthrodesis. Such patients may be completely satisfied with a well-functioning TAA, but they lose 8 points on the AOFAS owing to their fused subtalar joint. However, because of its wide use in the literature, AOFAS score values still offer the best comparison between different studies.

Self-reported questionnaires, in contrast, might more adequately reflect the patients' perspective. What we have learned from outcome research in fields other than foot and ankle surgery is that self-reported outcome instruments allow for a more complete estimation of the patients' health status and of issues relevant to the patients. The only self-reported region- or disease-specific measures used in studies on TAA, however, were the FFI [4], AOS [9], and "Oxford" [18]. The "Oxford" is put in quotation marks because this instrument is a not validated, self-developed questionnaire modeling to the original Oxford Hip Score [7]. The original FFI also was a result of an expert panel, initially developed for patients with rheumatoid arthritis [4]. Although, in the meantime, numerous studies broadened its use to the entire spectrum of foot and ankle

disorders, included patient perceptions, and adapted the instrument for use in different languages [1, 24, 30, 36, 38, 46], several limitations of this instrument have been highlighted, resulting in a recently performed extensive revision of this questionnaire based on Rasch analysis [3]. Although we found reasonable ratings in terms of the different quality criteria for the FFI, the above-mentioned studies resulted in at least five different FFI versions (FFI original, FFI-R long, FFI-R short, FFI-D, FFI-5pt) and its use in patients having TAA, therefore, can be recommended only cautiously. Considering the AOS is based on the original FFI [9], it can be concluded no quality region- or disease-specific tools have been used regularly in TAA studies until now. Recognizing the recent literature, it is interesting that very well-developed self-reported instruments, such as the Foot and Ankle Ability Measure (FAAM), have not yet been used in studies reporting on TAA outcomes [28]. The FAAM has shown evidence of validity, reliability, and responsiveness in patients with a broad spectrum of foot complaints, including ankle osteoarthritis [28].

For all the instruments identified in our review, no studies have supplied enough information to understand the interpretability of the results; in particular, no MCIDs have been reported yet. Interpretability is defined by the Scientific Advisory Committee as "the degree to which one can assign easily understood meaning to an instrument's quantitative score" [25]. To facilitate interpretability, various kinds of information are needed, eg, norm values, differences between subgroups expected to differ in scores, and MCIDs [40, 42]. Unfortunately, most of this information is not available for the instruments used in TAA studies.

Several different outcome instruments have been used in studies reporting on TAA, with the AOFAS hindfoot score and the Kofoed ankle score being the most common. However, there is no or only limited evidence to support their use in terms of patient relevance, validity, reliability, responsiveness, and interpretability. Self-reported questionnaires to assess TAA outcomes are rather uncommon until now, and considerable research is required to broaden the knowledge regarding the existing measures and to develop and investigate new measures that validly and reliably assess outcomes in this target population.

References

1. Agel J, Beskin JL, Brage M, Guyton GP, Kadel NJ, Saltzman CL, Sands AK, Sangeorzan BJ, SooHoo NF, Stroud CC, Thordarson DB. Reliability of the Foot Function Index: a report of the AOFAS Outcomes Committee. *Foot Ankle Int.* 2005;26:962–967.
2. Bellamy N, Buchanan WW, Goldsmith CH, Campbell J, Stitt LW. Validation study of WOMAC: a health status instrument for measuring clinically important patient relevant outcomes to antirheumatic drug therapy in patients with osteoarthritis of the hip or knee. *J Rheumatol.* 1988;15:1833–1840.
3. Budiman-Mak E, Conrad K, Stuck R, Matters M. Theoretical model and Rasch analysis to develop a revised Foot Function Index. *Foot Ankle Int.* 2006;27:519–527.
4. Budiman-Mak E, Conrad KJ, Roach KE. The Foot Function Index: a measure of foot pain and disability. *J Clin Epidemiol.* 1991;44:561–570.
5. Buechel FF, Pappas MJ, Iorio LJ. New Jersey low contact stress total ankle replacement: biomechanical rationale and review of 23 cementless cases. *Foot Ankle.* 1988;8:279–290.
6. Button G, Pinney S. A meta-analysis of outcome rating scales in foot and ankle surgery: is there a valid, reliable, and responsive system? *Foot Ankle Int.* 2004;25:521–525.
7. Dawson J, Fitzpatrick R, Carr A, Murray D. Questionnaire on the perceptions of patients about total hip replacement. *J Bone Joint Surg Br.* 1996;78:185–190.
8. de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. *J Clin Epidemiol.* 2006;59:1033–1039.
9. Domsic RT, Saltzman CL. Ankle osteoarthritis scale. *Foot Ankle Int.* 1998;19:466–471.
10. Engelberg R, Martin DP, Agel J, Obremsky W, Coronado G, Swiontkowski MF. Musculoskeletal Function Assessment instrument: criterion and construct validity. *J Orthop Res.* 1996;14:182–192.
11. Evanski PH, Waugh TR. Management of arthritis of the ankle: an alternative of arthrodesis. *Clin Orthop Relat Res.* 1977;122:110–115.
12. Fries JF, Spitz P, Kraines RG, Hotman HR. Measurement of patient outcome in arthritis. *Arthritis Rheum.* 1980;23:137–145.
13. Guyatt GH, Deyo RA, Charlson M, Levine MN, Mitchell A. Responsiveness and validity in health status measurement: a clarification. *J Clin Epidemiol.* 1989;42:403–408.
14. Guyatt GH, Feeny DH, Patrick DL. Measuring health-related quality of life. *Ann Intern Med.* 1993;118:622–629.
15. Guyer AJ, Richardson G. Current concepts review: total ankle arthroplasty. *Foot Ankle Int.* 2008;29:256–264.
16. Guyton GP. Theoretical limitations of the AOFAS scoring systems: an analysis using Monte Carlo modeling. *Foot Ankle Int.* 2001;22:779–787.
17. Haddad SL, Coetzee JC, Estok R, Fahrbach K, Banel D, Nalysnyk L. Intermediate and long-term outcomes of total ankle arthroplasty and ankle arthrodesis: a systematic review of the literature. *J Bone Joint Surg Am.* 2007;89:1899–1905.
18. Hosman AH, Mason RB, Hobbs T, Rothwell AG. A New Zealand national joint registry review of 202 total ankle replacements followed for up to 6 years. *Acta Orthop.* 2007;78:584–591.
19. Ibrahim T, Beiri A, Azzabi M, Best AJ, Taylor GJ, Menon DK. Reliability and validity of the subjective component of the American Orthopaedic Foot and Ankle Society clinical rating scales. *J Foot Ankle Surg.* 2007;46:65–74.
20. Kirshner BF, Guyatt GH. A methodological framework for assessing health indices. *J Chronic Dis.* 1985;38:27–36.
21. Kitaoka HB, Alexander IJ, Adelaar RS, Nunley JA, Myerson MS, Sanders M. Clinical rating systems for the ankle-hindfoot, midfoot, hallux, and lesser toes. *Foot Ankle Int.* 1994;15:349–353.
22. Kofoed H. A new total ankle joint prosthesis. In: Kossowsky R, Kossovsky V, eds. *Material Sciences and Implant Orthopedic Surgery.* Dordrecht, The Netherlands: Martinus Nijhoff; 1986:75–84.
23. Kofoed H. Cylindrical cemented ankle arthroplasty: a prospective series with long-term follow-up. *Foot Ankle Int.* 1995;16:474–479.
24. Kuyvenhoven MM, Gorter KJ, Zuithoff P, Budiman-Mak E, Conrad KJ, Post MW. The foot function index with verbal rating

scales (FFI-5pt): a clinimetric evaluation and comparison with the original FFI. *J Rheumatol.* 2002;29:1023–1028.
25. Lohr KN, Aaronson NK, Alonso J, Burnam MA, Patrick DL, Perrin EB, Roberts JS. Evaluating quality-of-life and health status instruments: development of scientific review criteria. *Clin Ther.* 1996;18:979–992.
26. Malviya A, Makwana N, Laing P. Correlation of the AOFAS scores with a generic health QUALY score in foot and ankle surgery. *Foot Ankle Int.* 2007;28:494–498.
27. Martin RL, Irrgang JJ. A survey of self-reported outcome instruments for the foot and ankle. *J Orthop Sports Phys Ther.* 2007;37:72–84.
28. Martin RL, Irrgang JJ, Burdett RG, Conti SF, Van Swearingen JM. Evidence of validity for the Foot and Ankle Ability Measure (FAAM). *Foot Ankle Int.* 2005;26:968–983.
29. Mazur JM, Schwartz E, Simon SR. Ankle arthrodesis: long-term follow-up with gait analysis. *J Bone Joint Surg Am.* 1979;61: 964–975.
30. Naal FD, Impellizzeri FM, Huber M, Rippstein PF. Cross-cultural adaptation and validation of the Foot Function Index for use in German-speaking patients with foot complaints. *Foot Ankle Int.* 2008;29:1222–1228.
31. Naal FD, Impellizzeri FM, Loibl M, Huber M, Rippstein PF. Habitual physical activity and sports participation after total ankle arthroplasty. *Am J Sports Med.* 2009;37:95–102.
32. Neufeld SK, Lee TH. Total ankle arthroplasty: indications, results, and biomechanical rationale. *Am J Orthop.* 2000;29: 593–602.
33. Parker J, Nester CJ, Long AF, Barrie J. The problem with measuring patient perceptions of outcome with existing outcome measures in foot and ankle surgery. *Foot Ankle Int.* 2003;24: 56–60.
34. Pena F, Agel J, Coetzee JC. Comparison of the MFA to the AOFAS outcome tool in a population undergoing total ankle replacement. *Foot Ankle Int.* 2007;28:788–793.
35. Radford PJ. General outcomes measures. In: Pynsent PB, Fairbank JC, Carr A, eds. *Outcome Measures in Orthopaedics.* London, UK: Butterworth-Heinemann; 1993:59–80.
36. SooHoo NF, Samimi DB, Vyas RM, Botzler T. Evaluation of the validity of the Foot Function Index in measuring outcomes in patients with foot and ankle disorders. *Foot Ankle Int.* 2006; 27:38–42.
37. SooHoo NF, Shuler M, Fleming LL; American Orthopaedic Foot and Ankle Society. Evaluation of the validity of the AOFAS Clinical Rating Systems by correlation to the SF-36. *Foot Ankle Int.* 2003;24:50–55.
38. SooHoo NF, Vyas R, Samimi D. Responsiveness of the foot function index, AOFAS clinical rating systems, and SF-36 after foot and ankle surgery. *Foot Ankle Int.* 2006;27:930–934.
39. Takakura Y, Tanaka Y, Sugimoto K, Tamai S, Masuhara K. Ankle arthroplasty: a comparative study of cemented metal and uncemented ceramic prostheses. *Clin Orthop Relat Res.* 1990; 252:209–216.
40. Terwee CB, Bot SD, de Boer MR, van der Windt DA, Knol DL, Dekker J, Bouter LM, de Vet HC. Quality criteria were proposed for measurement properties of health status questionnaires. *J Clin Epidemiol.* 2007;60:34–42.
41. van der Leeden M, Steultjens MP, Terwee CB, Rosenbaum D, Turner D, Woodburn J, Dekker J. A systematic review of instruments measuring foot function, foot pain, and foot-related disability in patients with rheumatoid arthritis. *Arthritis Rheum.* 2008;59:1257–1269.
42. Veenhof C, Bijlsma JW, van den Ende CH, van Dijk GM, Pisters MF, Dekker J. Psychometric evaluation of osteoarthritis questionnaires: a systematic review of the literature. *Arthritis Rheum.* 2006;55:480–492.
43. Ware JE Jr, Sherbourne CD. The MOS 36-item short-form health survey (SF-36). I. Conceptual framework and item selection. *Med Care.* 1992;30:473–483.
44. World Health Organization. International Classification of Functioning, Disability and Health (ICF). Geneva, Switzerland: World Health Organization; 2001. Available at: http://www.who.int/classifications/icf/en/. Accessed June 20, 2009.
45. Wright RW, Brand RA, Dunn W, Spindler KP. How to write a systematic review. *Clin Orthop Relat Res.* 2007;455:23–29.
46. Wu SH, Liang HW, Hou WH. Reliability and validity of the Taiwan Chinese version of the Foot Function Index. *J Formos Med Assoc.* 2008;107:111–118.