Scoring in forefoot surgery

A statistical evaluation of single variables and rating systems

Wolfgang Schneider and Karl Knahr

We assessed 13 scores for forefoot surgery and their single parameters in 200 cases of hallux surgery. Outcome expressed in descriptive terms from excellent to poor, as well as calculated as a numerical value differed to a great extent. The rank correlation between different scores showed poor conformity, in some combinations even slight negative correlations. The rank correlation between single parameters and the overall outcome gave good values for some variables included in most scores, e.g., walking distance, general activity, problems on uneven surfaces, shoe wear, cosmetics and pain. Little, if any, correlation was found for joint instability, range of motion of the MTP- and IP-joints, the use of walking aids and medication. We find a need for a generally accepted score, with parameters of high clinical relevance.

Orthopaedisches Spital Wien-Speising, Speisinger Strasse 109, AT-1134 Vienna, Austria. Tel +43 1 80182-0. Fax -285.
e-mail: wsmail@ping.at
Submitted 97-08-28. Accepted 98-05-12

Bonney and Macnab (1952) were the first to use a simple numerical system for separate anatomical and functional grading, to describe the outcome of hallux surgery. In the following decades, this system was used as a basis for the development of further rating systems. Vallier et al. (1991) upgraded the Bonney/Macnab classification to a numerical overall score by dividing the sum of the three partial outcomes into excellent, good, fair or poor. Gainor et al. (1988) calculated a numerical overall outcome, with subsequent rating into poor to excellent.

In the following years, rating systems gained more interest, leading to a variety of scores. There are two basic concepts: numerical scores expressing an overall outcome as a total of single variables (Sherman et al. 1984, Gainor et al. 1988, Geissele and Stanton 1990, Kitaoka and Holiday 1991, Kitaoka et al. 1991, Shankar et al. 1991, Vallier et al. 1991, Moeckel et al. 1992, Sammarco et al. 1993, Kitaoka et al. 1994) and descriptive scores, directly describing the outcome as excellent, good, fair or poor (Steinböck and Leder 1988, Anderl et al. 1991, Coughlin 1991). Most of the numerical scores (Gainor et al. 1988, Geissele and Stanton 1990, Kitaoka and Holiday 1991, Kitaoka et al. 1991, Shankar et al. 1991, Vallier et al. 1991, Moeckel et al. 1992, Sammarco et al. 1993) include a subsequent descriptive rating (Table 1).

We applied 13 different outcome scores simultaneously to a series of 200 hallux operations to assess the suitability of each score for clinical use.

Material and methods

The pre- and postoperative data about 200 operations (resection arthroplasties (Keller-Brandes), distal metatarsal osteotomies (chevron), distal soft tissue procedures, cheilectomies, arthrodeses, proximal metatarsal osteotomies) for hallux deformities, with a minimum follow-up period of 5 years were used to calculate all 13 scores in this study. The data were collected using a questionnaire, which was in plain colloquial language. To keep the influence of the observer as low as possible, the patient was requested to answer all questions alone and the completed questionnaire was discussed once again with the examiner. Missing answers were completed and ambiguous answers were adjusted, when necessary. 30 clinical and radiographic variables used in the 13 rating systems (Table 1) were included and pre- and postoperative results were calculated for each patient and score.

10 of 13 scoring systems (Bonney and Macnab 1952, Sherman et al. 1984, Steinböck and Leder 1988, Geissele and Stanton 1990, Anderl et al. 1991, Coughlin 1991, Kitaoka and Holiday 1991, Kitaoka et al. 1991, Moeckel et al. 1992, Sammarco et al. 1993) contain at least one variable, consisting of several sub-variables (for example, the parameter "function" consisting of the assessment of "shoe wear" + "walking ability"). The value of a compound variable like this was calculated according to the worst subvariable, as recommended by Steinböck and Leder (1988). The remaining 3 scores (Gainor et al. 1988, Shankar et al. 1991, Kitaoka et al. 1994) allowed cor-

Table 1. Variables collected to calculate scores and their distribution in 13 rating systems

	Bon	She	Gai	Ste	Gei	And	Cou	KHo	K91	Sha	Moe	Sam	K94
Descriptive	–	–	–	+	–	+	+	–	–	–	–	–	–
Numerical and decriptive	+	+	+	–	+	–	–	+	+	+	+	–	–
Numerical	+	+	–	–	–	–	–	–	–	–	–	+	+
Pain	+	+	+	+	+	+	+	+	+	+	+	+	+
Alignment (cosmesis)	–	–	+	+	+	+	–	+	+	+	+	+	+
Dorsiflexion MTP	+	+	–	+	+	+	–	+	+	+	–	+	+
Plantarflexion MTP	+	+	–	–	+	+	–	+	+	+	–	+	+
Footwear	–	–	+	+	–	–	–	+	+	+	+	+	+
Overall activity limitations	+	+	–	–	+	–	+	+	+	–	–	+	+
Walking distance	–	–	+	–	–	+	+	–	+	+	+	+	–
MTP-angle	+	+	–	+	+	–	–	–	–	+	–	–	–
Support, walking aid	–	–	+	–	–	–	–	+	+	–	–	+	–
Tender callus	–	–	–	–	–	–	–	+	–	–	+	–	+
Medication	–	–	–	–	–	–	–	–	–	–	+	+	–
Overall satisfation	–	–	–	+	–	–	+	–	–	–	–	–	–
Bunion	+	+	–	–	–	–	–	–	–	–	–	–	–
Bursitis	+	+	–	–	–	–	–	–	–	–	–	–	–
Limp	–	–	–	–	–	–	–	–	+	–	–	+	–
Plantarflexion IP	–	+	–	–	–	–	–	–	–	–	–	–	+
IM-angle	–	–	–	–	+	–	–	–	–	+	–	–	–
Terrain problems	–	–	–	–	–	–	–	–	–	–	–	+	–
Stairs	–	–	–	–	–	–	–	–	–	–	–	+	–
Stability, giving way	–	–	–	–	–	–	–	–	–	–	–	+	–
Midfoot and rearfoot ROM	–	–	–	–	–	–	–	–	–	–	–	+	–
Metatarsalgia	–	–	–	–	–	–	–	–	–	+	–	–	–
Revision surgery	–	–	–	–	–	–	–	–	+	–	–	–	–
IP-joint degeneration	–	–	–	–	–	–	–	–	–	+	–	–	–
MTP–IP stability	–	–	–	–	–	–	–	–	–	–	–	–	+

Bon Bonney and Macnab 1952
She Sherman et al. 1984
Gai Gainor et al. 1988
Ste Steinböck and Leder 1988
Gei Geissele and Stanton 1990
And Anderl et al. 1991
Cou Coughlin 1991
KHo Kitaoka and Holiday 1991
K91 Kitaoka et al. 1991
Sha Shankar et al. 1991
Moe Moeckel et al. 1992
Sam Sammarco et al. 1993
K94 Kitaoka et al. 1994

rect calculation of all parameters.

In cases where solitary parameters were impossible to obtain for preoperative scoring (like "satisfaction with the results of surgery"), the calculation was performed with omission of this parameter. Only Shankar's score proved to be completely unsuitable for preoperative use, because the single parameters were oriented strictly for postoperative evaluation.

For comparison of numerical scores, the result was standardized as a percentage of their maximum value. Each single numerical result was also calculated in terms of excellent to poor, where intended by the author. For comparison of descriptive scores, a grading from excellent to poor was used, according to the specification of the author.

To assess the agreement between the numerical results of different scores, the limit of agreement was calculated according to Bland and Altman (1986). We used Spearman's rank correlation index to calculate the correlation between single parameters and an overall result and between the results of different scores.

Results

There were enormous differences in use of descriptive terms ("excellent, good, fair, poor") (Figure 1). The preoperative scores ranged from only 7% poor with Gainor's score to 100% poor, according to the score of Steinböck. For the postoperative assessment, Gainor's score showed a high rate of 83% excellent outcome, whereas Steinböck's score resulted in only 17% excellent outcome.

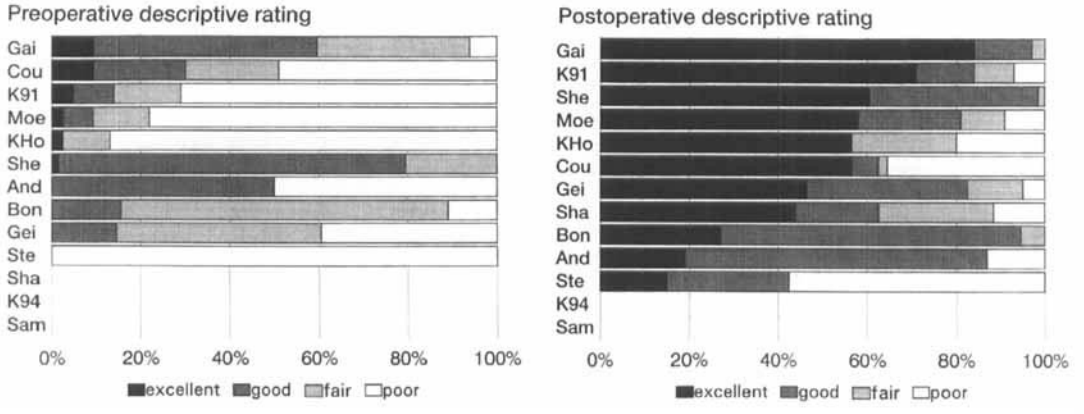

Figure 1. Descriptive ratings. Score abbreviations, see Table 1.

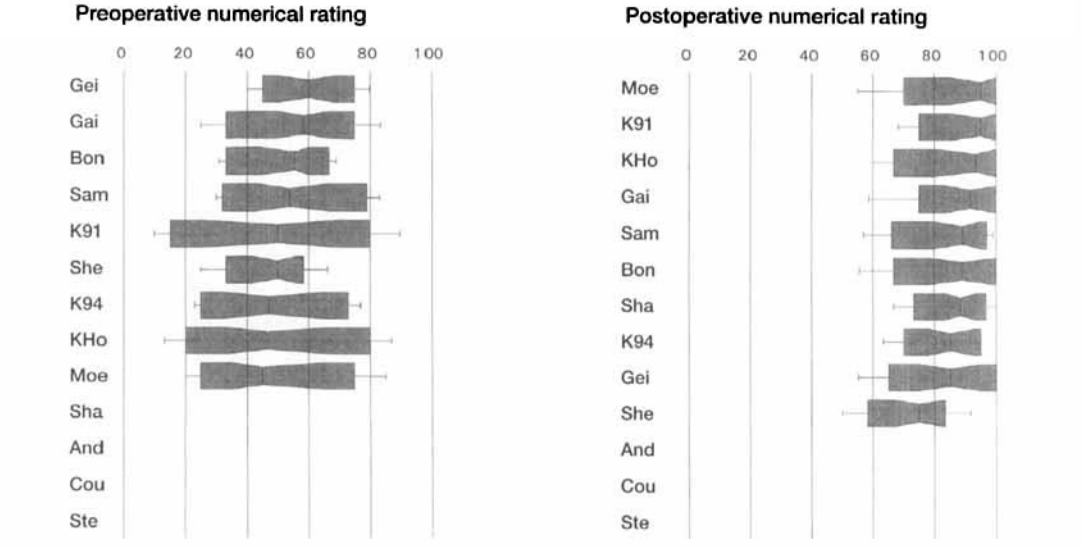

Figure 2. Box and Whisker diagram showing 50% percentile, 25% and 75% percentiles (tapering), 10% and 90% percentiles (box) and 5% and 95% percentiles (whiskers). Score abbreviations, see Table 1.

The score of the American Orthopaedic Foot and Ankle Society (Kitaoka et al. 1994) and the Maryland Footscoring Profile (Sammarco et al. 1993) had to be omitted from this evaluation, because the authors intentionally did not assign numerical values to descriptive terms. Shankar's score was calculated only for postoperative parameters, because it is unsuitable for preoperative use.

Analyzing the results as a percentage of the maximum numerical value (Figure 2 shows the data expressed as percentiles), a better agreement among the various scores was found. The preoperative mean values differed from 60% of the possible maximum of Geissele and Stanton's score and 46% according to Sherman et al. (1984). Postoperatively, the best mean result was found for Gainor's score with 91%, contrasting with the 72% according to Sherman et al. (1984). Calculating the postoperative improvement, the best increase was reported for Kitaoka et al. (1991) with 41%, contrasting with an improvement of only 24% calculating the data according to Geissele and Stanton. Using the Bland and Altman technique to assess the agreement between two methods (Table 2), the best values for the limits of agreement were found comparing the postoperative outcomes according to Kitaoka and Holiday (1991) and Kitaoka et al. (1991) with –11% and 8%. The worst measuring agreement was calculated by comparing the preoperative results of Sherman et al. (1984) and Kitaoka et al. (1991) with –53% and 48%.

For each single parameter, a correlation to a pre- and postoperative overall outcome was calculated. To

Table 2. Numerical rating—limits of agreement according to Bland and Altman in percentages

Post-op.	Preoperative									
	Bon	She	Gai	Gei	KHo	K91	Sha	Moe	Sam	K94
Bon		-7 19	-39 28	-30 15	-35 41	-43 49	...	-33 39	-31 27	-30 37
She	-3 29		-49 26	-39 12	-46 40	-53 48	...	-44 38	-42 26	-40 35
Gai	-31 21	-47 11		-33 29	-17 34	-21 38	...	-14 31	-19 25	-13 31
Gei	-19 23	-35 12	-16 29		-32 52	-38 60	...	-29 50	-28 38	-25 47
KHo	-25 18	-44 11	-17 20	-31 21		-19 20	...	-18 18	-25 15	-17 18
K91	-26 16	-44 8	-16 16	-32 18	-11 8		...	-25 25	-30 20	-19 19
Sha	-21 18	-35 6	-17 24	-24 17	-23 26	-19 25		...	...	...
Moe	-29 23	-48 15	-16 20	-33 24	-13 14	-13 17	-27 25		-24 14	-20 21
Sam	-23 23	-41 15	-13 24	-27 24	-12 19	-10 20	-21 25	-11 18		-12 23
K94	-16 20	-34 11	-11 24	-23 22	-9 18	-6 18	-16 23	-11 20	-14 16	

Score abbreviations, see Table 1.

Table 3. Correlation between parameters and overall result —preoperative

Parameter	Correlation
Walking distance	61
Overall activity limitations	56
Tender callus	54
Terrain problems	54
Alignment (cosmesis)	43
Stairs	43
Footwear	43
Limp	41
Bursitits	40
Pain	38
Other deformities	36
IM angle	31
MTP angle	29
Orthotics	28
Bunion	27
Stability, giving way	25
Medication	22
IP plantarflexion	11
Support, walking aid	11
MTP plantar flexion	8
MTP arthrosis	6
MTP dorsiflexion	1
MTP–IP stability	0

Table 4. Correlation between parameters and overall result —postoperative

Parameter	Correlation
Footwear	50
Other deformities	48
Terrain problems	46
Pain	45
Alignment (cosmesis)	43
Walking distance	43
Overall activity limitations	42
MTP angle	36
Orthotics	33
MTP arthrosis	31
Stairs	30
Support, walking aid	27
Stability, giving way	27
Tender callus	25
Limp	25
Bursitits	25
IM angle	24
Bunion	21
MTP dorsiflexion	20
IP plantarflexion	19
MTP plantar flexion	17
Medication	14
MTP–IP stability	6
Parameters only available postoperatively	
Overall satisfaction	43
Midfoot adduction	34
Midfoot abduction	33
Subtalar inversion	26
Subtalar eversion	24
Tibiotalar plantarflexion	22
Tibiotalar dorsiflexion	22

obtain this overall result, all single parameters were added, with equal emphasis on each parameter. Tables 3 and 4 shows the single parameters, ranked according their correlation to the overall result.

The correlation between scores using Spearman's rank correlation coefficients was calculated separately for descriptive and numerical results. A higher correlation was found mostly with the results expressed numerically, especially when calculated for the postoperative situation. For single combinations, even slightly negative correlations were found (Tables 5 and 6).

Discussion

Although not every problem can be described using a single score, it has proved to be a good basis for discussion. A score simplifies the interpretation of data and facilitates comparisons between methods and centers. When results are reported, a certain quantification has to be made, and the more complex the data become, the more important it is to use a score to summarize the outcome.

While several attempts have been made to quantify the outcome of forefoot surgery, the application of these rating systems remained restricted mostly to the original author. Most of these scores contain parameters according to a special interest of the author or a certain operating technique. Therefore, it is difficult to select the variables representative of an overall

Table 5. Spearman's rank correlation coefficient (x100)—preoperative results. Respective higher values are printed bold

Numerical results	Descriptive results												
	Bon	She	Gai	Ste	Gei	And	Cou	KHo	K91	Sha	Moe	Sam	K94
Bon		58	**42**	0	36	32	42	**52**	**50**	–	**42**	–	–
She	**79**		10	0	29	52	3	**23**	13	–	10	–	–
Gai	32	**13**		0	**46**	-9	57	48	62	–	53	–	–
Ste	–	–	–		0	0	0	0	0	–	0	–	–
Gei	**50**	**36**	45	–		28	17	27	29	–	21	–	–
And	–	–	–	–	–		-5	11	0	–	-2	–	–
Cou	–	–	–	–	–	–		54	76	–	71	–	–
KHo	41	21	**80**	–	**30**	–	–		72	–	75	–	–
K91	40	**19**	**82**	–	29	–	–	**93**		–	81	–	–
Sha	–	–	–	–	–	–	–	–	–		–	–	–
Moe	34	**14**	**83**	–	**24**	–	–	**92**	**90**	–		–	–
Sam	45	26	78	–	34	–	–	90	92	–	87		–
K94	43	26	82	–	37	–	–	92	94	–	88	91	

Score abbreviations, see Table 1.

Table 6. Spearman's rank correlation coefficient (x100)—postoperative results. Respective higher values are printed bold

Numerical results	Descriptive results												
	Bon	She	Gai	Ste	Gei	And	Cou	KHo	K91	Sha	Moe	Sam	K94
Bon		55	32	72	53	60	31	57	46	55	53	–	–
She	**76**		23	48	44	46	26	37	44	48	30	–	–
Gai	**54**	**33**		36	48	33	40	48	62	**46**	51	–	–
Ste	–	–	–		45	63	38	65	44	58	67	–	–
Gei	**65**	**54**	**63**	–		55	41	41	49	64	37	–	–
And	–	–	–	–	–		48	43	36	59	42	–	–
Cou	–	–	–	–	–	–		44	48	36	44	–	–
KHo	**70**	**45**	**77**	–	**53**	–	–		74	42	**95**	–	–
K91	**71**	**48**	**81**	–	**56**	–	–	**96**		47	66	–	–
Sha	**66**	**60**	44	–	**66**	–	–	**51**	**52**		**43**	–	–
Moe	**61**	**33**	**80**	–	**50**	–	–	86	**81**	40		–	–
Sam	55	36	70	–	52	–	–	71	71	48	73		–
K94	72	51	75	–	58	–	–	84	82	53	88	75	

Score abbreviations, see Table 1.

result and applicable as a basis for scientific comparison.

For foot surgery, a commonly accepted score or a comparison of existing scores is not available. Bonney and Macnab (1952) introduced a rating system for hallux surgery, consisting of three categories of assessment, without intending to calculate an overall outcome (this was done by Vallier in 1991 who used this score and simply added the numerical partial results assigning these overall values to poor to excellent). Gainor et al. (1988) were the first to use a straightforward numerical score consisting of subjective parameters. In the following years, many scoring systems were published with increasing detail, culminating in Shankar's score (1991) consisting of 13 variables, including radiographic parameters and even pedobarographic assessment. Recently the scores have been simplified again, with deliberate omission of sophisticated equipment and even radiographic data (Kitaoka et al. 1994).

Analyzing our comparison, the following points have to be discussed:

1) The score must be suitable for pre- and postoperative assessment. With 2 scoring systems it was impossible to obtain reasonable preoperative values. The parameters of Shankar's numerical score were selected only for postoperative evaluation. Steinböck's descriptive score for all 200 preoperative cases resulted in "poor", due to an extremely rigorous assessment of single variables.

However, it is of importance to present meaningful preoperative data, since it is the improvement of the

preoperative score which can be used to compare methods and centers.

2) Collection of data should not require sophisticated equipment. If special equipment (for example pedobarography in Shankar's score) should be used for quantification, these data should be described and analyzed apart from an overall score. As a consequence, the American Orthopaedic Foot and Ankle Society (Kitaoka et al. 1994) does not recommend the inclusion of even simple measurements on plain dorsoplantar radiographs. This may be controversial, since it should be possible to obtain standardized plain radiographs.

3) The score must not be complicated by using many parameters. It is impossible to create a system that fits requirements concerning all imaginable medical problems. Special circumstances will always need additional descriptive data or measurements.

A score does not automatically become more conclusive with an increasing number of variables: Shankar's or Sammarco's scores showed no advantage, compared to other scores consisting of fewer parameters. Time-consuming data collection reduces the general acceptance of a score. The scoring system has to minimize the number of parameters and include only variables with high clinical relevance.

4) The score should not include compound parameters. 10 of the 13 analyzed scores use compound parameters consisting of subvariables. Only Steinböck and Leder (1988) give instructions how to handle a situation where the answers to subvariables are divergent and the examiner has to decide, whether to take the best of all possibilities, the worst or a mean value. For correct comparison, the rating of a compound parameter was deduced from the worst subvariable, as recommended by Peters et al. (1997) in his comparison of knee instability scores.

5) Parameters should be representative of the overall outcome. Many scores include parameters with a low correlation to the overall outcome, both pre- and postoperatively. Only parameters with a high correlation to the overall aspects of hallux surgery should be used. Parameters essential for the description of a specific problem, but with low correlation to the overall outcome, have to be classified and mentioned separately but not within a score. The best correlation was found for walking distance, general activity, problems on uneven surfaces, shoe wear, cosmetics and pain. The worst correlation resulted for instability of the IP- or MTP-joint, surprisingly for range of motion of the MTP-joint, use of walking aids and range of motion of the IP-joint. The parameters with the best correlation to the overall outcome are, indeed, part of most scoring systems.

6) Results should be expressed as a numerical value. A numerical value gives a much more differentiated result compared to a descriptive score. When our results were calculated as a descriptive rating, the correlation between different scores tended to be lower, especially for postoperative assessment. This indicates that descriptive ratings are too blunt and ratings like "poor" or "excellent" should be derived from a numerical score and mentioned separately.

7) Comparing results requires the use of identical scores. To compare one's own results with previous publications or to list results reviewing the literature is common in scientific papers. These comparisons will not be relevant, if different scores have been used.

In our analysis, the best correlation of 96% was found for the postoperative evaluation according to Kitaoka and Holiday (1991) and Kitaoka et al. (1991). Comparing these two scores and calculating their limits of agreement, as introduced by Bland and Altmann, the result of Kitaoka and Holiday (1991) may be 11 points below or 8 points above Kitaoka et al. (1991). This satisfactory measuring agreement indicates that these two methods may be used interchangeably in clinical interpretation. On the other hand, the wide range of correlations extended up to a slight negative(!) correlation of –9% when comparing Gainor's and Anderl's score. Using Bland and Altman's method, the worst measuring agreement showed that Sherman may be 53 points below or 48 points above Kitaoka (1991), which is unacceptable for the comparison of clinical data. This lack of agreement is by no means obvious in Figure 2 that shows exactly the same 50th percentile for these methods. This extremely poor correlation between methods designed to measure similar clinical outcome is unexpected, as comparisons of scores for knee ligament instability (Peters et al. 1997) showed much better results with the least correlation of 61%. A comparison of hip scores (Bryant et al. 1993) in the worst case showed a correlation between two scores of as much as 44%.

Based on our analysis, the suitability of each score to the requirements mentioned above is listed in Table 7. Corresponding to these findings, the scoring systems of Gainor et al. (1988), Kitaoka and Holiday (1991), Kitaoka et al. (1991) and Kitaoka et al. (1994) appear to be applicable in most clinical settings. Among these four scores, the rating system published on behalf of the American Orthopaedic Foot and Ankle Society (Kitaoka et al. 1994) seems to be gaining acceptance as a "consensus score". Nevertheless, the limited clinical relevance of some parameters is a shortcoming that needs to be corrected.

Table 7. Suitability of scores to requirements for a rating system

Requirements	Bon	She	Gai	Ste	Gei	And	Cou	KHo	K91	Sha	Moe	Sam	K94
Pre- and postop. suitability	++	++	++	−	+	++	−	++	++	−−	o	++	++
Simplicity of parameters	++	++	++	++	+	++	++	++	++	−−	++	++	++
Number of parameters	++	++	++	++	++	++	++	++	++	−	++	−−	++
Use of compound parameters	−	−	++	−−	+	−−	−−	+	+	++	−−	o	++
Relevance of parameters	o	−	+	+	o	o	++	+	+	o	+	−	o
Numerical results	−	−	++	−−	++	−−	−−	++	++	++	++	++	++

Score abbreviations, see Table 1.
++ excellent, + good, o fair, − poor, −− very poor

Anderl W, Knahr K, Steinböck G. Langzeitergebnisse der Hallux-Rigidus-Operation nach Keller-Brandes. Z Orthop 1991; 129: 42-7.

Bland J M, Altman D G. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1: 307-10.

Bonney G, Macnab I. Hallux valgus and hallux rigidus. J Bone Joint Surg (Br) 1952; 34: 366-85.

Bryant M J, Kernohan W G, Nixon J R, Mollan R A B. A statistical analysis of hip scores. J Bone Joint Surg (Br) 1993; 75: 705-9.

Coughlin M J. Treatment of bunionette deformity with longitudinal diaphyseal osteotomy with distal soft tissue repair. Foot Ankle 1991; 11: 195-203.

Gainor B J, Epstein R G, Henstorf J E, Olson S. Metatarsal head resection for rheumatoid deformities of the forefoot. Clin Orthop 1988; 230: 207-13.

Geissele A E, Stanton R P. Surgical treatment of adolescent hallux valgus. J Pediatr Orthop 1990; 10: 642-8.

Kitaoka H B, Holiday A D Jr. Metatarsal head resection for bunionette: long-term follow-up. Foot Ankle 1991; 11: 345-9.

Kitaoka H B, Franco M G, Weaver A L, Ilstrup D M. Simple bunionectomy with medial capsulorrhaphy. Foot Ankle 1991; 12: 86-91.

Kitaoka H B, Alexander I J, Adelaar R S, Nunley J A, Myerson M S, Sanders M. Clinical rating systems for the ankle-hindfoot, midfoot, hallux and lesser toes. Foot Ankle 1994; 15: 349-53.

Moeckel B H, Sculco T P, Alexiades M M, Dossick P H, Inglis A E, Ranawat C S. The double-stemmed silicone-rubber implant for rheumatoid arthritis of the first metatarsophalangeal joint. J Bone Joint Surg (Am) 1992; 74: 564-70.

Peters G, Wirth C J, Kohn D. Vergleich von Scores und Bewertungsschemata bei Kniebandinstabilitäten. Z Orthop 1997; 135: 63-69.

Sammarco G J, Brainard B J, Sammarco V J. Bunion correction using proximal chevron osteotomy. Foot Ankle 1993; 14: 8-14.

Shankar N S, Asaad S S, Craxford A D. Hinged silastic implants of the great toe. Clin Orthop 1991; 272: 227-34.

Sherman K P, Douglas D L, Benson M K D'A. Keller's arthroplasty: is distraction useful? J Bone Joint Surg (Br) 1984; 66: 765-9.

Steinböck G, Leder K. Operation des Hallux valgus nach Akin-New. Z Orthop 1988; 126: 420-4.

Vallier G T, Petersen S A, LaGrone M O. The Keller resection arthroplasty: A 13-year experience. Foot Ankle 1991; 11: 187-94.