Development and reliability of a standard rating system for outcome measurement of foot and ankle disorders II: interclinician and intraclinician reliability and validity of the newly established standard rating scales and Japanese Orthopaedic Association rating scale

Hisateru Niki, Haruhito Aoki, Suguru Inokuchi, Satoru Ozeki, Mitsuo Kinoshita, Hideji Kura, Yasuhito Tanaka, Masahiko Noguchi, Shigeharu Nomura, Masahito Hatori, and Shinobu Tatsunami

1 Department of Orthopaedic Surgery, St. Marianna University School of Medicine, 2-16-1 Sugao, Miyamae-ku, Kanagawa 216-8511, Japan
2 Department of Orthopaedic Surgery, Keio University, Tokyo, Japan
3 Department of Orthopaedic Surgery, Koshigaya Hospital, Dokkyo University School of Medicine, Saitama, Japan
4 Department of Orthopaedic Surgery, Osaka Medical College, Osaka, Japan
5 Department of Orthopaedic Surgery, Nara Medical University, Nara, Japan
6 Department of Orthopaedic Surgery, Yamanouchi Medical University, Yamanouchi, Japan
7 Department of Orthopaedic Surgery, Nara Medical University, Nara, Japan
8 Nomura Seikeigeka Ganka Clinic, Yamaguchi, Japan
9 Department of Orthopaedic Surgery, Tohoku University, Miyagi, Japan
10 Unit of Medical Statistics, Faculty of Education and Culture, St. Marianna University School of Medicine, Kanagawa, Japan

Abstract

Background. This study evaluated the validity and inter- and intraclinician reliability of (1) the Japanese Society of Surgery of the Foot (JSSF) standard rating system for four sites [ankle-hindfoot (AH), midfoot (MF), hallux (HL), and lesser toe (LT)] and the rheumatoid arthritis (RA) foot and ankle scale and (2) the Japanese Orthopaedic Association’s foot rating scale (JOA scale).

Methods. Clinicians from the same institute independently evaluated participating patients from their institute by two evaluations at a 1- to 4-week interval. Statistical evaluation was as follows. (1) The intraclass correlation coefficient (ICC) was calculated from data collected from at least two examinations of each patient by at least two evaluating clinicians (Data A). (2) Total scores for the two evaluations were determined from the distribution of differences in data between the two evaluations (Data B); each item was evaluated by determining Cohen’s coefficient of agreement. (3) The relation between patient satisfaction and total score was investigated only for patients who underwent surgery (Data C). Spearman’s rank correlation coefficient was obtained.

Results. Participants were 65 clinicians and 610 patients, including those with disorders of the AH (313), MF (47), HL (153), and LT (50) and those with RA (47). From Data A, the ICC was high for AH and HL by JSSF scales and for AH, MF, and LT by the JOA scale. From Data B, the coefficient showed high validity for both scales for AH, with almost no difference between the two scales; the validity for HL was higher with the JOA scale than with the JSSF scale. From Data C, correlations were significant between patient satisfaction and outcome for AH and HL by the JSSF scales and for AH, HL, and LT by the JOA scale.

Conclusions. The validity of both scales was high. Clinical evaluation of the therapeutic results using these scales would be highly reliable.

Introduction

Recently, therapeutic options have been selected quite often on the basis of evidence-based medicine (EBM). Thus, we are beginning to appreciate the importance of a standard rating system to evaluate such evidence. Such a rating system demands reliability in rating as well as appropriate coverage of the diseases concerned and methods for their therapy. In this context, in orthopedic surgery, several standard rating systems have undergone a number of examinations for reliability.1-8 Unfortunately, however, in the field of foot and ankle joints the validity and reliability of the Japanese Orthopaedic Association (JOA) scale have not been verified.9,10 Moreover, although the American Orthopaedic Foot and Ankle Society (AOFAS) clinical rating system11 could now be called a global standard, it has not been verified as to its validity and reliability.

The JOA attempted to provide an internationally accepted standard rating system that incorporated not only objective evaluation by orthopedists but also sub-
jective evaluation by patients. The JOA thus delegated tasks to each member association to adjust and modify standard rating systems and verify their validity and reliability. In responding to this request, the Japanese Society of Surgery of the Foot (JSSF) organized the Committee on Rating Standards for Foot Disease in June 2000. After many discussions they created the JSSF standard rating system composed of five new scales, four of which were set up for four respective sites by modifying the AOFAS clinical rating systems; the remaining scale was for the rheumatoid arthritis (RA) foot and ankle joint by modifying the conventional JOA scale (part I of this study, which appears in this issue). Moreover, each scale included an explanation as well as rating scores for each item so the individual items to be evaluated could be understood (part I of this study). Our current four site-specific scales are a completely novel and original Japanese version and are far from a duplicate of the AOFAS clinical rating system, as we modified the expressions and content to suit Japanese people. We also added interpretation criteria for each item and rating criteria, such as a pain scale, which were lacking in the AOFAS scale. This is why the Committee on Rating Standards for Foot Disease of the JSSF grouped together the five scales, comprised of four site-specific scales and the RA foot and ankle scale and termed it the JSSF standard rating system. From the year 2001 on, actual patients were evaluated to collect data employing the JSSF standard rating system in multiple institutes.

In part II (described herein) we report the results of studies performed on a multiple-institution scale on the validity and inter- and intraclinician reliability of the evaluation items with regard to the JSSF standard rating system composed of these five scales as well as the conventional JOA scale.

Materials and methods

Selection of clinicians as evaluators
The subjects were orthopedists at nine institutions to which the authors belonged. Because it was thought that clinical experience would influence the reliability of the evaluation, the clinicians were selected according to the following three levels of experience: (1) much experience (specialist with at least 2 years’ experience in foot surgery); (2) moderate experience (generalist with approximately 6–7 years’ experience in an orthopedics department); and (3) little experience (recently (within 1–2 years) graduated resident from a medical university). In most cases two orthopedists representing each level of experience were selected from each institute.

Selection of patients as evaluators
Patients with diseases of the foot and ankle who met the following criteria were included: (1) symptomatically stable for at least 1 month prior to the study; (2) symptomatically stable for at least 1 month after the first evaluation; (3) consented to participate in the study; and (4) had no underlying diseases or complications that might interfere with the results of the evaluation.

Study design
A clinician from the same institution independently evaluated all the patients selected from that institution (first evaluation). Attempts were made to conduct the evaluation within 1 day, but when it was not possible it was extended into the second day. No other evaluating clinicians were present during this first evaluation. The evaluating clinician explained to the patients that simple answers to the questions were expected. When possible, the same evaluating clinician performed both the first and second evaluations. The second evaluation was conducted within 1–4 weeks of the first evaluation. As for the first examination, the second was conducted on the same day if possible. The results were recorded immediately after the evaluation, and subsequent corrections were prohibited. The results of the first evaluation were concealed at the time of the second evaluation. Patients were evaluated according to the order of the items on the instrument being evaluated. The evaluation of the items in both the JSSF standard rating system and the JOA scale were conducted on the same day as far as possible. The results were sent to the server at each institution using the Web system established for data collection in the present study and stored until tabulation.

Statistical methods
1. To determine interclinician agreement in terms of the total scores (validity), the intraclass correlation coefficient (ICC) was calculated from the evaluation data, which was collected from at least two patients who underwent the same evaluation by at least two clinicians from the same institution if all relevant data from those institutions were available (Analyzed Subject Data A). To establish the multinstitutional overall scale for interclinician reliability, the ICC was calculated by the random effect model using data obtained for patients with diseases of the ankle-hindfoot. Sufficient data for other sites were not available from all of the institutions, but sufficient data for this site was available from five institutions.
2. To determine intraclinician agreement (validity), the total scores from the first and second evaluations, respectively, were determined from the distribution of differences in the data between the two evaluations for each institution that provided sufficient data (Analyzing Subject Data B). Each item was evaluated by determining Cohen’s coefficient of agreement (κ) and the rate of complete agreement (RC) between the first and second evaluations.

3. To determine the relation between the scores in each scale and patient satisfaction, the relation between patient satisfaction and outcome (total score) was investigated using the evaluations of only those patients who had undergone surgery (Analyzing Subject Data C). The degree of satisfaction was evaluated as “very satisfactory,” “satisfactory,” “noncomputable,” “slightly unsatisfactory,” and “very unsatisfactory.” The total score for each degree of satisfaction was 0–50, 60–69, 70–79, 80–89, and 90–100 points ranked as 0, 1, 2, 3, and 4, respectively. Spearman’s rank correlation coefficient (ρ) was then obtained.

Results

Evaluating clinicians and patients

A total of 65 clinicians evaluated the patients. The distribution of clinicians according to experience level was 21.5% specialists, 30.8% generalists, and 47.7% residents. There were 610 patients, representing 313 diseases of the ankle-hindfoot, 50 diseases of the midfoot, 153 diseases of the hallux, 50 diseases of the lesser toe, and 47 with RA. Evaluation by the JOA scale was conducted simultaneously with that by JSSF scales in 501 of the 610 patients.

Results of statistical analysis

1. For Data A, the number of patients and the number of evaluating clinicians varied among the institutions. With the lower limit of the 95% confidence interval (CI) of the ICC calculated as an indication of inter-clinician agreement being 0.41, a value of >0.41 was observed for the ankle-hindfoot and hallux by the JSSF scales and for the ankle-hindfoot, midfoot, and lesser toe by the JOA scale (P < 0.05; ICC > 0.4 in testing) (Table 1). As for patients with diseases of the ankle-hindfoot, the overall ICC calculated from the data for the five institutions was 0.93 for the JSSF scale compared with 0.91 for the JOA scale.

2. For Data B, the percentages of values for each site evaluated by the JSSF scales relative to that evaluated by the JOA scale were as follows: 83 to 83 for the ankle-hindfoot, 10 to 4 for the midfoot, 45 to 56 for the hallux, 6 to 4 for the lesser toe, and 21 to 21 for RA.

a. Distribution of differences in total scores.

1) Regardless of the experience level, the difference in total scores between the first and second evaluation was within the range of ±1 in 43.4% and 42.3% of the data evaluated by the JSSF and JOA scales, respectively, for the ankle-hindfoot, indicating almost no difference between the two. These frequencies were higher than those for other sites, and the difference was within ±5 in approximately 70% of data evaluated by the two scales for the ankle-hindfoot. The difference was within a range of ±1 in 31.1% and 37.5% of the data evaluated by the JSSF scales and the JOA scale, respectively, for the hallux. The corresponding frequencies in RA patients were 19.5% and 19.0% of data evaluated by the JSSF and JOA scales, respectively; differences within the range of ±5 were observed in approximately 60% of the data evaluated by the two scales. It was difficult to evaluate the midfoot and lesser toe because of the small number of patients with diseases at these sites (Table 2).

2) The influence of experience level was observed when the difference in the total scores between the first and second evaluations was within a range of ±1; a tendency toward the presence of influence of the experience level was observed in data evaluated by the JSSF scale for the ankle-hindfoot and in data evaluated by both scales for the hallux and RA. When the difference was within the range of ±5, however, there was almost no difference in the results depending on the experience level. It was difficult to evaluate the midfoot and lesser toe because of the small number of patients with diseases at these sites (Table 2).

b. Evaluation of each item.

1) For the first and second evaluations, Cohen’s coefficient of agreement (κ) was high for all items for the ankle-hindfoot evaluated by the JSSF scale and low for sagittal motion, muscle strength, and sensory disturbance (paresthesia) of the hindfoot evaluated by the JOA scale (Table 4). The coefficient (κ) was low for all items other than sagittal motion of the metatarsophalangeal (MTP) joint of the hallux evaluated by the JSSF scale, and high for most of the items evaluated by the JOA scale. It was difficult to evaluate data for the midfoot, lesser toe, and RA because of the
| Institute | No. of patients | No. of clinicians | ICC          | P  |
|-----------|----------------|------------------|--------------|----|
| **JSSF scale** |                |                  |              |    |
| Ankle-hindfoot |            |                  |              |    |
| A         | 6             | 4                | 0.7246 (0.3788–0.9472) | 0.03 |
| B         | 3             | 4                | 0.8526 (0.45–1.0)     | 0.016 |
| C         | 3             | 3                | 0.9318 (0.5284–0.9982) | 0.015 |
| D         | 3             | 2                | 0.2753 (−1.5–0.98)   | 0.5508 |
| E         | 2             | 4                | −0.6691 (−0.7248–0.6935) | 0.9009 |
| **Midfoot** |            |                  |              |    |
| C         | 2             | 3                | 0.6975 (−0.15–1.0)   | 0.1896 |
| E         | 2             | 2                | 0.8324 (−0.02–1.0)   | 0.1984 |
| **Hallux** |            |                  |              |    |
| A         | 6             | 6                | 0.5429 (0.21–0.89)   | 0.1862 |
| C         | 2             | 2                | 0.0 (−1.0–1.0)       | 0.5904 |
| D         | 7             | 2                | 0.971 (0.84–1.0)     | 0.0004 |
| **Lesser toe** |            |                  |              |    |
| A         | 2             | 6                | 0.7298 (0.22–1.0)    | 0.0861 |
| C         | 2             | 3                | 0.569 (−0.18–1.0)    | 0.2666 |
| D         | 4             | 2                | 0.441 (−0.85–0.95)   | 0.4583 |
| **RA**    |                |                  |              |    |
| A         | 2             | 4                | 0.08 (−0.014–0.99)   | 0.7107 |
| B         | 2             | 6                | 0.4162 (−0.04–1.0)   | 0.3289 |
| **JOA scale** |            |                  |              |    |
| A         | 5             | 5                | 0.8721 (0.64–0.98)   | 0.0004 |
| B         | 3             | 4                | 0.4647 (−0.05–0.98)  | 0.3406 |
| C         | 2             | 4                | 0.3018 (−0.31–1.0)   | 0.4241 |

JSSF, Japanese Society of Surgery of the Foot; JOA, Japanese Orthopaedic Association; ICC, intraclass correlation coefficient; RA, rheumatoid arthritis

Boldface type indicates that ICC > 0.4 (P < 0.05)
small number of patients in the respective categories.

2) The mean RCs for each item evaluated by the JSSF and JOA scales were 81.2% and 84.3%, respectively, for the ankle-hindfoot; 70% and 57.1%, respectively, for the midfoot; 75.6% and 78.5%, respectively, for the hallux; 83.3% and 82.1%, respectively, for the lesser toe; and 76.2% and 77.5%, respectively, for RA. According to the items, the intraclass RC was high for all items of the ankle-hindfoot by the JSSF scale, whereas the rate was low for instability of the ankle-hindfoot by the JOA scale. The rate was low for alignment of the hallux by the JSSF scale and for pain, deformed forefoot, hindfoot sagittal motion, and walking on tiptoe by the JOA scale. The rate was low for a deformed lesser toe of the forefoot, deformed hindfoot, and ability to walk when evaluated by the JSSF scale in RA patients and for pain, deformed forefoot, hindfoot sagittal motion, and ability to walk when evaluated by the JOA scale.

3. For Data C, the ratios of the total score for each site as evaluated by the JSSF scales to those as evaluated by the JOA scale were as follows: 169:161 for the ankle-hindfoot, 14:14 for the midfoot, 99:105 for the hallux, 34:33 for the lesser toe, and 24:24 for RA.

a. There was a significant correlation between patient satisfaction and the total score (outcome) for the hindfoot and hallux by the JSSF standard rating system and for the ankle-hindfoot, hallux, and lesser toe by the JOA scale (Table 5).

### Table 2. Distribution of difference in data between first and second evaluations (regardless of experience level)

| Site and scale       | % Difference in range of ±1 to ±5 |
|----------------------|-----------------------------------|
|                      | ±1      | ±3      | ±5      |
| Ankle-hindfoot       |         |         |         |
| JSSF                 | 83      | 43.4    | 61.4    | 68.7    |
| JOA                  | 83      | 42.3    | 56.6    | 75.9    |
| Midfoot              |         |         |         |
| JSSF                 | 10      | 50.0    | 50.0    | 70.0    |
| JOA                  | 4       | 25.0    | 25.0    | 25.0    |
| Hallux               |         |         |         |
| JSSF                 | 45      | 31.1    | 40.0    | 55.6    |
| JOA                  | 56      | 37.5    | 53.8    | 62.5    |
| Lesser toe           |         |         |         |
| JSSF                 | 6       | 33.3    | 66.7    | 83.3    |
| JOA                  | 4       | 75.0    | 100     | 100     |
| RA                   |         |         |         |
| JSSF                 | 21      | 19.5    | 47.6    | 61.9    |
| JOA                  | 21      | 19.0    | 33.3    | 52.4    |

### Table 3. Distribution of difference in data between first and second evaluations (with regard to experience level)

| Experience level     | % Difference, by JSSF scale | % Difference, by JOA scale |
|----------------------|-----------------------------|----------------------------|
|                      | No. | ±1 | ±3 | ±5 | No. | ±1 | ±3 | ±5 |
| Ankle-hindfoot       |     |    |    |    |     |    |    |    |
| Specialist           | 34  | 47.1| 64.7| 76.5| 33  | 45.5| 57.6| 75.8|
| Generalist           | 25  | 52.0| 76.0| 80.0| 26  | 38.5| 61.5| 84.6|
| Resident             | 24  | 29.2| 41.7| 45.8| 24  | 41.7| 50.0| 66.7|
| Midfoot              |     |    |    |    |     |    |    |    |
| Specialist           | 4   | 75.0| 75.0| 100 | 1   | —   | —   | —   |
| Generalist           | 2   | 50.0| 50.0| 50.0| 1   | —   | —   | —   |
| Resident             | 4   | 25.0| 25.0| 50.0| 2   | 0   | 0   | 0   |
| Hallux               |     |    |    |    |     |    |    |    |
| Specialist           | 17  | 41.2| 52.9| 64.7| 22  | 50.0| 54.5| 63.6|
| Generalist           | 13  | 23.1| 23.1| 38.5| 18  | 27.8| 50.0| 55.6|
| Resident             | 15  | 26.7| 40.0| 60.0| 16  | 31.3| 56.3| 68.9|
| Lesser toe           |     |    |    |    |     |    |    |    |
| Specialist           | 3   | 0   | 33.3| 66.7| 2   | 100 | 100 | 100 |
| Generalist           | 2   | 50.0| 100 | 100 | 1   | —   | —   | —   |
| Resident             | 1   | —   | —   | —   | 1   | —   | —   | —   |
| RA                   |     |    |    |    |     |    |    |    |
| Specialist           | 6   | 33.3| 50.0| 66.7| 5   | 20.0| 40.0| 60.0|
| Generalist           | 7   | 14.3| 42.9| 57.1| 8   | 37.5| 37.5| 50.0|
| Resident             | 8   | 12.5| 50.0| 62.5| 8   | 0   | 25.0| 50.0|

—, noncomputable (insufficient sample number)
### Table 4. Rate of complete agreement and Cohen's coefficient of agreement

| Parameter                  | JSSF scale | JOA scale |
|----------------------------|------------|-----------|
| JSSF scale                 | RC (%)     | κ         |
| Ankle-hindfoot (n = 83)    |            |           |
| Pain                       | 79.5       | 0.672     |
| Activity limitations       | 71.1       | 0.568     |
| Maximum walking distance   | 85.5       | 0.604     |
| Walking surfaces           | 83.1       | 0.711     |
| Gait abnormality           | 83.1       | 0.582     |
| Sagittal motion            | 85.5       | 0.625     |
| Hindfoot motion            | 80.7       | 0.573     |
| Stability                  | 86.7       | 0.405     |
| Alignment                  | 79.5       | —         |
| Midfoot (n = 10)           |            |           |
| Pain                       | 60         | 0.492     |
| Activity limitations       | 60         | 0.31      |
| Max. walking distance      | 60         | —         |
| Footwear requirements      | 70         | —         |
| Walking surfaces           | 60         | —         |
| Gait abnormality           | 90         | 0.821     |
| Alignment                  | 90         | —         |
| Hallux (n = 45)            |            |           |
| Pain                       | 66.6       | —         |
| Activity limitations       | 64.4       | —         |
| Footwear requirements      | 73.3       | —         |
| MTP joint motion           | 75.6       | 0.559     |
| IP joint motion            | 97.8       | —         |
| MTP-IP Stability           | 88.9       | 0.237     |
| Callus or clavus           | 80         | 0.281     |
| Alignment                  | 57.8       | 0.282     |
| Lesser toe (n = 6)         |            |           |
| Pain                       | 100        | 1         |
| Activity limitations       | 66.7       | 0.25      |
| Footwear requirements      | 66.7       | —         |
| MTP joint motion           | 66.7       | —         |
| IP joint motion            | 83.3       | —         |
| MTP-IP Stability           | 100        | —         |
| Callus or clavus           | 100        | —         |
| Alignment                  | 83.3       | 0.667     |
| Ankle-hindfoot (n = 83)    |            |           |
| Pain                       | 77.1       | 0.639     |
| Deformity, forefoot        | 89.2       | 0.574     |
| Deformity, hindfoot        | 79.5       | 0.514     |
| MTP/IP joint motion        | 71.1       | 0.358     |
| Hindfoot motion            | 94         | 0.78      |
| Stability                  | 63.9       | —         |
| Walking ability            | 77.1       | —         |
| Muscle strength            | 86.7       | 0.286     |
| Sensory disturbance        | 86.7       | 0.252     |
| Climbing/descending stairs| 96.4       | 0.928     |
| Sitting on heels           | 88         | 0.81      |
| Standing on toes           | 79.5       | 0.548     |
| Footwear                   | 88         | 0.522     |
| Japanese-style toilet      | 73.5       | 0.514     |
| Midfoot (n = 4)            |            |           |
| Pain                       | 50         | —         |
| Deformity, forefoot        | 50         | —         |
| Deformity, hindfoot        | 25         | —         |
| MTP/IP joint motion        | 75         | —         |
| Hindfoot motion            | 50         | 0.333     |
| Stability                  | 75         | 0.5       |
| Walking ability            | 50         | 0.2       |
| Muscle strength            | 25         | —         |
| Sensory disturbance        | 75         | —         |
| Climbing/descending stairs| 75         | —         |
| Sitting on heels           | 75         | —         |
| Standing on toes           | 100        | —         |
| Footwear                   | 25         | —         |
| Japanese-style toilet      | 50         | —         |
| Hallux (n = 56)            |            |           |
| Pain                       | 57.1       | 0.357     |
| Deformity, forefoot        | 64.3       | 0.474     |
| Deformity, hindfoot        | 91.1       | 0.51      |
| MTP/IP joint motion        | 94.6       | 0.024     |
| Hindfoot motion            | 69.6       | 0.526     |
| Stability                  | 78.6       | 0.361     |
| Walking ability            | 73.2       | 0.532     |
| Muscle strength            | 80.4       | —         |
| Sensory disturbance        | 87.5       | 0.162     |
| Climbing/descending stairs| 91.1       | 0.707     |
| Sitting on heels           | 85.7       | 0.439     |
| Standing on toes           | 69.6       | 0.479     |
| Footwear                   | 71.4       | 0.492     |
| Japanese-style toilet      | 75.7       | —         |
| Lesser toe (n = 4)         |            |           |
| Pain                       | 50         | —         |
| Deformity, forefoot        | 100        | —         |
| Deformity, hindfoot        | 100        | —         |
| MTP/IP joint motion        | 50         | —         |
| Hindfoot motion            | 100        | —         |
| Stability                  | 50         | —         |
| Walking ability            | 75         | —         |
| Muscle strength            | 100        | —         |
| Sensory disturbance        | 100        | —         |
| Climbing/descending stairs| 100        | —         |
| Sitting on heels           | 100        | —         |
| Standing on toes           | 75         | 0.5       |
| Footwear                   | 50         | —         |
| Japanese-style toilet      | 100        | —         |
Discussion

With the practice of EBM gaining ground worldwide, many epidemiological surveys and clinical studies are being performed for the purpose of obtaining evidence. An assessment of the results is essential for surveys and studies, and the relative superiority of the efficacy of one treatment or therapeutic effect over another should be evaluated based on the results of such determinations. For objective assessment of the results, a standard rating scale for evaluation should therefore be established. Important requirements for a rating scale are a high degree of validity and reliability. To our knowledge, the intraclinician and interclinician validity and reliability of standard rating systems for evaluating diseases of the foot and ankle, including the AOFAS clinical rating systems, have never been examined by multistitutional studies.

As for the interclinician agreement in terms of the total scores, the ICC was calculated from data obtained from evaluation of at least two of the same patients by multiple clinicians at the same institution. Only institutions from which there were sufficient data for analysis were included. At each institution, the ICC was high for the ankle-hindfoot and hallux by the JSSF scales and high for the ankle-hindfoot, midfoot, and lesser toe by the JOA scale. These results indicate that reliability was high at each institution, although overall multistitutional interclinician reliability could not be evaluated. When following the method employed in the report that evaluated reliability over all participating institutions using the ICC by the random effect model, it is possible that one cannot obtain a correct evaluation in such cases where the experience or knowledge of the examiners or the severity of the disease in patients differs among institutions or where the amount of data is small. Therefore, in principle we calculated each ICC for each institution. To verify our findings, we calculated the ICC from data for the ankle-hindfoot for all five institutions following a similar random effect model and found that the ICC was 0.9 or higher by both the JSSF scale and the JOA scale. Even when the same patient was examined at many institutions, the reliability of the standard rating scale for evaluation of diseases of the ankle-hindfoot was estimated to be high.

When interclinician and intraclinician reliability of the JSSF standard rating system and the JOA scale were investigated merely from the viewpoint of differences in the total scores between the first and second evaluations, the range of validity tended to increase for the hallux and RA compared to that for the ankle-hindfoot, for which the validity was already found to be relatively high. The RC, which was reflected by Cohen’s coefficient of agreement for each item, also showed high validity on the JSSF and JOA scales for evaluation of the

| Parameter | JSSF scale | JOA scale |
|-----------|------------|-----------|
| RA (n = 21) |           |           |
| Pain      | 0.762      | —         |
| Deformity, hallux | 0.762 | 0.608 |
| Deformity, lesser toe | 0.571 | 0.016 |
| Deformity, midfoot | 0.714 | 0.475 |
| Deformity, hindfoot | 0.476 |
| MTP/IP joint motion | 0.712 | 0.632 |
| Hindfoot motion | 0.762 | 0.578 |
| Walking ability | 0.571 |
| Climbing/descending stairs | 0.952 |
| Sitting on heels | 1 |
| Standing on toes | 0.857 |
| Footwear | 0.857 | 0.745 |
| Japanese-style toilet | 0.905 | 0.811 |
| RA (n = 21) |           |           |
| Pain      | —          | 0.408     |
| Deformity, forefoot | 66.7 |
| Deformity, hindfoot | 71.4 |
| MTP/IP joint motion | 57.1 | 0.171 |
| Hindfoot motion | 71.4 | 0.571 |
| Stability | 71.4 |
| Walking ability | 66.7 |
| Muscle strength | 76.2 |
| Sensory disturbance | 95.2 |
| Climbing/descending stairs | 95.2 |
| Sitting on heels | 90.5 |
| Standing on toes | 81 |
| Footwear | 85.7 | 0.725 |
| Japanese-style toilet | 95.2 | 0.905 |

RC, rate of complete agreement; \( \kappa \), Cohen’s coefficient of agreement; —, noncomputable

\( \kappa \) Values: boldface indicates \( \kappa > 0.6 \) and italics indicates \( \kappa > 0.4 \)

Table 5. Relation between patient satisfaction and total score (outcome)

| Parameter | Spearman rank correlation (\( \rho \)) | JSSF scale | JOA scale |
|-----------|--------------------------------------|------------|-----------|
| Ankle-hindfoot | 0.373 (\( P < 0.0001 \)) | 0.341 (\( P < 0.0001 \)) |
| Midfoot | 0.104 | –0.007 |
| Hallux | 0.399 (\( P < 0.0001 \)) | 0.271 (\( P < 0.005 \)) |
| Lesser toe | 0.321 | 0.737 (\( P < 0.0001 \)) |
| RA | — | — |

—, Noncomputable
ankle-hindfoot, with almost no difference observed between the two scales, whereas the validity of the JOA scale for the hallux was higher than that of the JSSF scale. Thus, there was a difference in validity between the two scales for some sites of the foot and ankle. There were also some items for which statistical analysis could not be conducted because of the small number of patients; but the validity of the JSSF standard rating system was evaluated as being high by the assessment of intraclinician agreement because the concept of each scale of the JSSF standard rating system is almost the same.

As for intraclinician agreement assessed according to the level of clinical experience, it is assumed that proficiency in evaluation is necessary to obtain high validity of the evaluation when investigated only from the distribution of differences in the total scores.

“The degree of satisfaction” in the evaluation of treatment is related to psychological aspects on the part of patients and differs from the functional aspects evaluated by clinicians. Therefore, the correlation between the degree of satisfaction on the part of patients and functional assessment by clinicians is not necessarily high, but there was a tendency for the outcome to be correlated with patient satisfaction. Each item in the standard rating system was considered to be a reflection of a subjective evaluation on the part of the patients. Recently, results of findings by instruments on the severity of pain by visual analogue scales (VAS) and questionnaires about the quality of life (QOL) by SF-36 and others, in which QOL is evaluated based on scales that take into account the viewpoint of patients, have been shown to be as reproducible as results based on data from pathophysiologic evaluations by clinicians. In other words, therapeutic results are increasingly determined directly according to the patient’s own evaluation from the viewpoint of EBM because there is much room for bias in evaluations by clinicians; thus, instruments such as the VAS and SF-36 produce highly accurate information. Therefore, each standard rating scale for evaluation that was inspected in this study is assumed to be a reflection to some extent of the subjective evaluation on the part of patients, but a standard rating system that would allow evaluation of the symptomatic improvement and QOL of patients from different viewpoints needs to be established in the future.

The present study was conducted with the aim of evaluating the validity and reliability of the JSSF standard rating system and the JOA scale according to the site of involvement in the foot and ankle. Diagnostic workups of the same patients at multiple institutions are difficult. Therefore, we were obliged to limit our analysis of interclinician reliability to that from data compiled at individual institutions. To analyze interclinician reli-

ability more precisely, a different study design from that employed in the present study may be required.

Based on intraclinician reliability and the results of analysis of the relation between patient satisfaction and outcome, however, the validity of the JSSF standard rating system and the JOA scale was high for the items evaluated. It can be considered that clinical evaluation of therapeutic results using these scales would be highly reliable.

References

1. Sidor ML, Zuckerman JD, Lyon T, Koval K, Cuomo F, Schoeneng N. The Neer classification system for proximal humeral fractures; an assessment of interobserver reliability and intraobserver reproducibility. J Bone Joint Surg Am 1993;75:1745–50.
2. Siebenrock KA, Gerber C. The reproducibility of classification of fractures of the proximal end of the humerus. J Bone Joint Surg Am 1993;75:1751–5.
3. Rome K, Cowieson F. A reliability study of the universal goniometer, fluid goniometer, and electrogoniometer for the measurement of ankle dorsiflexion. Foot Ankle Int 1996;17:28–32.
4. Cummings RJ, Loveless EA, Campbell J, Samelson S, Mazur JM. Interobserver reliability and intraobserver reproducibility of the system of King et al. for the classification of adolescent idiopathic scoliosis. J Bone Joint Surg Am 1998;80:1107–11.
5. Irgang JJ, Snyder-Mackler L, Wainner RS, Fu FH, Harner CD. Development of a patient-reported measure of function of the knee. J Bone Joint Surg Am 1998;80:1132–45.
6. Lenke LG, Bets RR, Bridwell KH, Clements DH, Harms J, Lowe TG, et al. Intraobserver and interobserver reliability of the classification of thoracic adolescent idiopathic scoliosis. J Bone Joint Surg Am 1998;80:1097–106.
7. Yonenobu K, Abumi K, Nagata K, Taketomi E, Ueyama K. Inter- and intra-observer reliability of the Japanese Orthopaedic Association scoring system for evaluation of cervical myelopathy. Rinshou Seikeigeka (Clinical Orthopaedic Surgery) 2001;36:423–8 (in Japanese).
8. Greenfield MLVH, Kuhn JE, Wojtys EM. A statistic primer; validity and reliability. Am J Sports Med 1998;26:483–5.
9. Japanese Orthopaedic Association. Assessment criteria for foot disorders of the Japanese Orthopaedic Association. J Jpn Orthop Assoc 1991;65:680 (in Japanese).
10. Hisateru N, Nango A. Clinical rating systems for ankle disorders. In: Murota K, Yabe Y, Sano S, editors. Manual of orthopaedic clinical rating systems. Tokyo: Zen Nihonbyoin Shuppan Kai; 1995. p.117–35 (in Japanese).
11. Kitaoka HB, Alexander JJ, Adelaar RS, Nunley JA, Myerson MS, Sanders M. Clinical rating systems for the ankle-hindfoot, midfoot, hallux, and lesser toes. Foot Ankle Int 1994;15:349–53.
12. Fukuhara S, Suzugamo Y. Manual of SF-36v2 Japanese version. Kyoto: Institute for Health Outcomes & Process Evaluation Research; 2004 (in Japanese).
13. Toolan BC, Wright Quinones VJ, Cunningham BJ, Brage ME. An evaluation of the use of retrospectively acquired preoperative AOFAS clinical rating scores to assess surgical outcome after elective foot and ankle surgery. Foot Ankle Int 2001;22:775–8.
14. Thordarson DB, Rudicel SA, Ebramzadeh E, Gill LH. Outcome study of hallux valgus surgery: an AOFAS multi-center study. Foot Ankle Int 2001;22:956–9. Erratum in: Foot Ankle Int 2002;23:96.
15. Hunsaker FG, Cioffi DA, Amadio PC, Wright JT, Caughlin B. The American Academy of Orthopaedic Surgeons outcomes
16. SooHoo NF, Shuler M, Fleming LL. Evaluation of the validity of the AOFAS clinical rating systems by correlation to the SF-36. Foot Ankle Int 2003;24:50–5.
17. Johanson NA, Liang MH, Daltroy L, Rudicel S, Richmond J. American Academy of Orthopaedic Surgeons lower limb outcomes assessment instruments: reliability, validity, and sensitivity to change. J Bone Joint Surg Am 2004;86:902–9.
18. Thordarson D, Ebramzadeh E, Moorthy M, Lee J, Rudicel S. Correlation of hallux valgus surgical outcome with AOFAS forefoot score and radiological parameters. Foot Ankle Int 2005;26:122–7.