Inter-rater reliability of WHO ‘disability’ grading

WIM BRANDSMA*++, MEREDITH LARSEN**, CHARLES RICHARD* & MANNAM EBENEZER*
*International Nepal Fellowship, Green Pastures Hospital and Rehabilitation Centre, PO Box 5, Pokhara, Nepal
**St Olaf’s College, St Paul, Minnesota, USA
++Shieffelin Leprosy Research & Training Centre, Karigiri, 632 106, Tamil Nadu, India

Accepted for publication 26 February 2004

Summary The World Health Organization ‘disability’ grading system was introduced in 1960. It is mainly used as an indicator for early diagnosis or reporting. Disability grades are usually aggregated at national levels. Comparison of data with previous years or comparison of data between programmes may show that patients report earlier for treatment, alternatively, are diagnosed earlier, that is without, or with fewer ‘disabilities’. Despite its long and universal use as an epidemiological parameter, the WHO disability grading has not been the subject of reliability studies. In this study, three testers unfamiliar with the grading prior to the study each graded 65 (former) leprosy affected persons. The weighted kappa ranged from 0.89 to 0.89 (95% CI 0.73–1.00) for the highest score and from 0.96 to 0.96 (95% CI 0.90–0.99) for the EHF (Eye, Hand, Foot) score, indicating excellent reliability. The study shows that with limited training and little experience a high degree of reliability in grading ‘disabilities’ between testers is attainable.

Introduction

The main purpose of the WHO ‘disability’ grading is to have an indicator for early case detection. Disability grading was first recommended in 1960.¹ The last report of the WHO Expert Committee also discussed the grading and stated some objectives for its use.²³ More recently, the WHO grading has also been used as a change indicator of impairments for patients while on treatment. The individual scores for eyes, hands, and feet were added to get the so-called EHF sum score.⁴ The EHF score has been used in some other studies also.⁵⁶

In a Prevention of Disability (POD) programme the score could be used, for example, to find out how many leprosy affected persons have loss of protective sensation of their feet only

Correspondence to: W. Brandsma, International Nepal Fellowship, Green Pastures Hospital and Rehabilitation Centre, PO Box 5, Pokhara, Nepal (e-mail: wbrandsma@inf.org.np)
(grade 1), and how many have a more severe foot problem (grade 2). Following an intervention programme, one could then monitor how many persons/feet improved, remained the same or deteriorated. Or, as another example, a programme could assess and monitor in more detail if the number (or percentage) of patients diagnosed with leprosy report with fewer or less severe disabilities. In spite of its universal use and its importance as a relevant indicator for early case detection and its use as a POD indicator (sum score) the WHO disability grading has not been systematically subjected to reliability testing. Thus far, only Saunderson reported on the reliability of EHF grading. This study reports on the inter-tester reliability of the WHO disability grading and the EHF score.

Materials and methods

Three physiotherapy trainees who did not have any prior experience or training in the WHO ‘disability’ grading were introduced to the grading. They were introduced to the grading in one theoretical session followed by a supervised practice session. Questions were answered and differences in grading between the trainees were discussed.

Individuals with and without leprosy impairments both on treatment and released from treatment comprised the study population. Subjects were referred from the outpatient departments from the hospital. Also in-patients were assessed. The main coordinator of the study made sure that not too many leprosy affected persons were included with obvious grade 2 disabilities of many sites. For the grading the operational definitions of WHO grading were used as recently published. All three testers graded all leprosy subjects ($n = 65$) and 18 non-leprosy affected persons. Testing was done independently in a quiet environment. The testers did not have access to each other’s scores.

Sensation was tested with a ballpoint pen if there were no obvious visible impairments (grade 2). A visible impairment automatically qualifies the site for a grade 2. For the purpose of this study, sensation was then not tested. Six sites were tested on each hand (pulp and proximal phalanx little finger, pulp index and middle fingers and hypothenar and thenar eminence) and five on each foot (big toe, 1st, 2nd and 3rd metatarsal head and midlateral border). In the absence of a grade 2 impairment grade 1 was recorded if the stimulus was not felt on more than one site.

For reliability testing, the weighted kappa statistic was used. This statistic corrects for ‘chance’ agreement between testers.

Results

Table 1 gives the results of the three possible combinations with the three testers. Reliability of grading between the pairs is consistently high both for the WHO ‘highest’ grade and the sum score. According to the scale for grading reliability with the kappa statistic, the results can be classified as very good (Table 2).

Discussion

For this study, the operational definitions were used as recently published. WHO in their documentation about the WHO disability grading never gave operational definitions for the
grading. For example grade 1, anaesthesia: which instrument should be used, how many sites should be tested on which parts of the hand or foot, and on the basis of how many sites ‘not felt’ should the hand or foot be given a grade one? In like manner, one could argue and disagree about the interpretation of visible impairment. These, and possibly other reasons, may be the reason that in many programmes only grade 0 and grade 2 are recorded. Only these grades are then used to calculate the number (percentage) of new patients reporting with disability.

With the introduction of the EHF sum score as a change indicator in POD programmes and when following individual patients or a cohort of individuals affected by leprosy, there is a definite need for operational definitions of the grades.

The score still remains a crude impairment indicator and should not be the only indicator for impairment monitoring for individual patients. It can be a useful indicator for specific cohorts when for example monitoring a cohort of individuals with ‘chronic’ foot problems. In that case, a F(eet) sum score only could be used.

In this study, 18 controls, non-leprosy affected persons, were included. This was done to see if there were differences in recording between the testers for grade 0 and 1. This was not the case. For some of these individuals, grade 2 was recorded. This could be explained by the fact that the testers were not informed to not grade non-leprosy related impairments. Having this information would defeat the purpose of having non-leprosy controls included. These cases were not included in the analysis.

Vision was not tested with a specific test, nor was corneal sensation tested. In actual practice, the eye was graded 0 when nothing was obviously wrong and 2 when there was no vision or lagophthalmos present.

We tested for sensation with a ballpoint pen on selected sites only with a clear definition of ‘anaesthesia’: more than one point not felt. Using different filaments for hands or feet on different or more sites could have an influence on the number of individuals or hands or feet (sum score) given a grade 1. This could have an effect on the prevalence of grade 1 in a cohort.

### Table 1. Inter-rater reliability of WHO disability ‘grading’

<table>
<thead>
<tr>
<th>Tester pair</th>
<th>Highest grade</th>
<th>Sum score (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB</td>
<td>89 (0.76–1)</td>
<td>96 (0.93–0.99)</td>
</tr>
<tr>
<td>BC</td>
<td>87 (0.73–1)</td>
<td>95 (0.93–0.99)</td>
</tr>
<tr>
<td>AC</td>
<td>89 (0.75–1)</td>
<td>90 (0.90–0.98)</td>
</tr>
</tbody>
</table>

### Table 2. Classification of results (from Altman 8)

<table>
<thead>
<tr>
<th>Value of kappa</th>
<th>Strength of agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;0.20</td>
<td>Poor</td>
</tr>
<tr>
<td>0.21–0.40</td>
<td>Fair</td>
</tr>
<tr>
<td>0.41–0.60</td>
<td>Moderate</td>
</tr>
<tr>
<td>0.61–0.80</td>
<td>Good</td>
</tr>
<tr>
<td>0.81–1.00</td>
<td>Very good</td>
</tr>
</tbody>
</table>
and could affect the reliability for this grade if sensory grading with different filaments for hand and feet shows different reliability than testing with a ballpoint pen for sensory loss. A relative weakness of this study was that the study was done in a ‘controlled’ manner. The testers were introduced to the training, they knew what the purpose of the study was and there was a training session prior to the study. The strength of the study is that it shows that with little practice, staff unfamiliar with the grading, obtained high reliability. Studies should be undertaken in which reliability is assessed in field situations, grading ‘disabilities’ as routinely done.

Acknowledgement

We would like to thank the SLR&TC physiotherapy for facilitating the study. In particular we would like to thank Sr Leena Panjikaram, Sr Jesmy Paul, and Sr Nithya for assessing the patients.

References