Disability Measures in Stroke
Relationship Among the Barthel Index, the Functional Independence Measure, and the Modified Rankin Scale

Sooyeon Kwon, MS; Abraham G. Hartzema, PhD; Pamela W. Duncan, PhD; Sue-Min Lai, PhD

Background and Purpose—Residual disability after stroke presents a major economic and humanistic burden. To quantify disability in patients, activities of daily living (ADL; Barthel Index [BI], and motor component of Functional Independence Measure [M-FIM]) and categorical disability measures (Modified Rankin Scale [MRS]) are used. The purpose of this study is to examine the predicting ability of ADL measures to global disability scale.

Methods—Kansas City Stroke Study data were used for the present study. Correlation coefficient, Kruskal-Wallis test, and polytomous logistic regression analysis were applied to examine the relationship between the ADL measure and global disability scale. Model fit statistics were examined to verify logistic regression appropriateness. A categorization scheme, which minimized the false-positive response rate, was selected as the optimal categorizing system.

Results—The 3 measures were highly correlated. Both BI and M-FIM differentiated disability better in lower than higher disability. In logistic regression, BI differentiated 4 disability levels; M-FIM differentiated 3 levels in MRS. However, on the basis of results of the Kruskal-Wallis and multiple comparison tests, we suspect that M-FIM may have the potential to predict MRS categories better with a different model.

Conclusions—The proposed categorization scheme can serve as a translation between measures. However, because of the ceiling effect of BI and M-FIM, the translation could not be completed for all 6 levels of MRS. No apparent variation over time in the categorization scheme was observed. Further research needs to be conducted to develop better prediction models explaining the relationship between M-FIM and MRS. (Stroke. 2004;35:918-923.)

Key Words: activities of daily living ■ disability evaluation ■ outcome assessment ■ stroke

A pproximately 700,000 strokes occur each year in the United States, leaving 500,000 stroke survivors with disability, and economic loss resulting from stroke approaches an estimated $51.2 billion annually.1 The ratio of indirect to direct costs is approximately 1.3, which indicates that indirect costs in stroke are higher than direct medical costs.2 Indirect costs result mostly from compromised physical functioning and caregiver involvement. The high indirect cost of stroke makes the reduction of disability in poststroke patients a major interest of healthcare providers, researchers, and policy makers. Improving the independence of stroke survivors is the primary objective of poststroke treatment.

Accurate and precise assessment of activities of daily living (ADL) in poststroke patients is important for quality care and for measuring the outcomes of stroke treatment. The Agency for Health Care Policy and Research Post-Stroke Rehabilitation Panel suggests using well-validated and standardized instruments for reliable documentation of poststroke disabilities and prognoses over time.3 The panel recommended 2 instruments—the Barthel Index (BI) and the motor component of the Functional Independence Measures (M-FIM)—for measuring poststroke disability.4,5

Although validated and used globally, the BI and M-FIM scales have limitations in their application and evaluation. First, because the BI and M-FIM scale yield ordinal values, researchers or practitioners may have difficulties in interpreting the clinical meaning of the scores or changes in scores. Interpretation of the summary score of the instrument is limited to numeric increases or decreases in total score.

Understanding and interpreting the clinical significance of changes in the summary score for these measures prove important not only for evaluating individual patient outcomes but also for evaluating population-level outcomes. Epidemiological or health economic research, which often evaluates population-level outcomes, typically defines clinically distinct stages in the disease process to measure outcomes. This approach provides more interpretable information on population outcomes.

Second, it is difficult to compare scores measured in the BI and M-FIM. This is important considering that patients who
experience stroke often receive care from different facilities over time, depending on their prognoses, and the facilities do not always use the same instrument to document patient disability. Even though recorded outcome measures are complete, the patient’s disability outcomes may not be understood thoroughly in the continuum of care. A system for translating and comparing scores would reduce inefficiencies related to multiple instrument use.

Translation of scores across instruments also provides important information for population-level research. Studies using different instruments to evaluate the same program or intervention could be combined and interpreted together. Such integrated information, eg, meta-analyses, supports conclusions more strongly than a list of individual study results.

Because there is no systematic report on how many disability levels could be meaningfully categorized in the BI and M-FIM instruments and the cutoffs for those levels in each measure in respect to global disability scale, researchers have operationalized this in various ways.

The purpose of this study is to examine the ability of the BI and the M-FIM scores to differentiate clinically distinct categories of disability and to identify those levels. The modified Rankin Scale (MRS), which has been used to define clinically discrete patient disability categories, was used as reference.

**Subjects and Methods**

**Patient Population and Data Collection**

This study uses data from the Kansas City Stroke Study (KCSS), a prospective cohort study. Data collection started in October 1995 and was completed by 1999. Eligible stroke patients from the KCSS were identified by (1) a review of daily admission records; (2) referrals from physicians, clinical nurse specialists, and therapists on medical, neurology, and rehabilitation units; and (3) review of referrals from physicians, clinical nurse specialists, and therapists on medical, neurology, and rehabilitation units. The World Health Organization (WHO) definition of stroke was used.

**Table 1. Missing Data in Each Month**

<table>
<thead>
<tr>
<th>Month</th>
<th>Valid Sample Size, n</th>
<th>Frequency, MRS-BI, Barthel Data, %</th>
<th>Missing Data, %</th>
<th>Frequency, MRS–M-FIM, n</th>
<th>Missing Data, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>459</td>
<td>459</td>
<td>0</td>
<td>457</td>
<td>0.44</td>
</tr>
<tr>
<td>1</td>
<td>440</td>
<td>439</td>
<td>0.23</td>
<td>439</td>
<td>0.23</td>
</tr>
<tr>
<td>3</td>
<td>404</td>
<td>402</td>
<td>0.5</td>
<td>403</td>
<td>0.25</td>
</tr>
<tr>
<td>6</td>
<td>377</td>
<td>375</td>
<td>0.53</td>
<td>376</td>
<td>0.27</td>
</tr>
<tr>
<td>Total</td>
<td>1680</td>
<td>1675</td>
<td>0.24</td>
<td>1675</td>
<td>0.24</td>
</tr>
</tbody>
</table>

The BI is composed of 10 items with varying weights. Two items regarding personal toilet (wash face, comb hair, shave, and clean teeth) and bathing are evaluated with a 2-score scale (0 and 5 points); 6 items regarding feeding, getting onto and off the toilet, ascending and descending stairs, dressing, controlling bowels, and controlling bladder are evaluated with a 3-score scale (0, 5, and 10 points); and 2 items regarding moving from wheelchair to bed and returning, and walking on a level surface are evaluated with a 4-score scale (0, 5, 10, and 15 points). The BI is a cumulative score calculated by summing each item score. The BI scores are multiples of 5 with a range of 0 (completely dependent) to 100 (independent in basic ADL). Higher scores represent a higher degree of independence.

The M-FIM consists of 13 items: eating, grooming, bathing, dressing upper body, dressing lower body, toileting, managing bladder, managing bowel, transferring to bed/chair/wheelchair, transferring to toilet, transferring to tub/shower, locomotion by walk/wheelchair, and locomotion on stairs. Each item is rated with a score from 1 to 7 (1=complete assistance to perform basic ADL, 2=maximal assistance, 3=moderate assistance, 4=minimal assistance, 5=supervision, 6=modified independence, and 7=complete independence in performing basic ADL).

The MRS defines 6 levels of disability and 1 for death: 0=no symptom at all; 1=no significant disability despite symptoms, able to carry out all usual duties and activities; 2=slight disability, unable to carry out all previous activities but able to look after own affairs without assistance; 3=moderate disability, requires some help but able to walk without assistance; 4=moderately severe disability, unable to walk without assistance and unable to attend to own bodily needs without assistance; 5=severe disability, bedridden, incontinent, and requiring constant nursing care and attention; and 6=dead.

The present study focuses on the disability outcomes among stroke survivors. Individual scores in the MRS describe clinically distinct functional states of the patients.

The relationship between the instruments is the focus of the present analysis, so the unit of analysis was defined by each pair of observations on the BI and MRS and the M-FIM and MRS. Ideally, there would be 1836 possible records for the analysis if there were no attrition because 459 subjects were followed up at 4 time points. However, because of attrition, 1680 records were available for analysis. The reasons for attrition included death, withdrawal of family, a move, and refusal. Valid sample sizes for each month are listed in Table 1. Proportions of missing data were calculated on the basis of the available sample sizes. Overall, the missing data are minimal: 0.24% for MRS and BI and for MRS and M-FIM (Table 1).

Even though the loss increases over time, this phenomenon should not be considered a potential bias factor because the focus of this analysis is the relationship between instruments.

**Analysis Methods**

Correlations were examined among the BI, M-FIM, and MRS. The Kruskal-Wallis test was applied to examine the mean difference in BI and M-FIM scores among MRS levels. Because there are 6 levels in MRS, multiple comparison tests (Dwass, Steel, Critchlow-Fligner tests) were used to examine which pairs were significantly different.
The probability distributions of MRS given the BI or M-FIM were derived using a polytomous logistic regression analysis method. The final model is selected by considering model fit statistics. The model is stated as follows:

\[
\text{logit}(\frac{\text{MRS}_i}{\text{MFS}_0}) = \alpha + \beta x, \quad \text{where} \quad i = 0, 1, 2, 3, 4, 5; \quad x = 0 \text{ to } 100 \text{ for BI score or } 13 \text{ to } 91 \text{ for M-FIM score}; \quad \alpha \text{ is intercept; and} \quad \beta \text{ is estimate from polytomous logistic regression analysis.}
\]

Probability is stated as follows:

\[
\text{Probability (MRS}_i \text{)} = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}
\]

SAS version 8.2 (SAS Institute), StatDirect (version 2.3.3., StatDirect Ltd), and Sigma Plot Version 8.02 (SPSS Inc) were used for statistical analysis and graphic presentation.

Results

Frequency Distribution

Figure 1 illustrates the frequency distribution of BI and M-FIM scores compared with each MRS in 1675 observations. It shows the mean, median, interquartile range, 5th and 95th percentiles, and minimum and maximum of the BI and M-FIM scores. In this illustration, ceiling effects are observed for both BI and M-FIM scores compared with the MRS. The BI and M-FIM scores did not differentiate disability well in higher ADL levels, which correspond to MRS0, MRS1, and MRS2.

Distribution Free Correlation Coefficients

Spearman correlation coefficients were examined for the BI, M-FIM, and MRS. They were 0.8856 (P=0.0001) between BI and MRS, 0.9479 (P=0.0001) between BI and M-FIM score, and 0.8894 (P=0.0001) between M-FIM and MRS.

Kruskal-Wallis and Multiple Comparison Procedure, Dwass, Steel, Critchlow-Fligner

Kruskal-Wallis tests were applied to determine whether there was a statistically significant difference (P<0.05) in mean BI and M-FIM scores among MRS levels. Results showed that there were significant differences among levels in BI (\(\chi^2=1338.29, df=5, P<0.0001\)) and M-FIM (\(\chi^2=1338.74, df=5, P<0.0001\)). Kruskal-Wallis mean scores for BI and M-FIM for each level of MRS were expected to be highest in MRS0 and lowest in MRS5: MRS0>MRS1>MRS2>MRS3>MRS4>MRS5. However, for both BI and M-FIM, the expected order showed differently. The actual order of the mean scores was MRS1>MRS0>MRS2>MRS3>MRS4>MRS5 for both BI and M-FIM.

Multiple comparison tests (Dwass, Steel, Critchlow-Fligner) were performed as a follow-up to the Kruskal-Wallis test. The BI did not differentiate between MRS1 and MRS0 or between MRS0 and MRS2. On the other hand, M-FIM did not differentiate between MRS 1 and 0; it completely lacked the differentiation ability between these 2 levels (Table 2). The Kruskal-Wallis and follow-up multiple comparison tests are summarized as below; an underline indicates that the pair is not significantly different:

\[
\begin{align*}
\text{BI:} & \quad \underline{\text{MRS1}} > \text{MRS0} > \text{MRS2} > \text{MRS3} > \text{MRS4} > \text{MRS5} \\
\text{M-FIM:} & \quad \underline{\text{MRS1}} > \text{MRS0} > \text{MRS2} > \text{MRS3} > \text{MRS4} > \text{MRS5}
\end{align*}
\]

<table>
<thead>
<tr>
<th>Comparisons</th>
<th>BI</th>
<th>M-FIM</th>
</tr>
</thead>
<tbody>
<tr>
<td>MRS0 vs MRS1</td>
<td>78.35 (0.5136)</td>
<td>-33.85 (-0.9999)</td>
</tr>
<tr>
<td>MRS0 vs MRS2</td>
<td>57.52 (0.5424)</td>
<td>144.91 (0.0039)</td>
</tr>
<tr>
<td>MRS0 vs MRS3</td>
<td>443.88*</td>
<td>530.67*</td>
</tr>
<tr>
<td>MRS0 vs MRS4</td>
<td>953.34*</td>
<td>1029.90*</td>
</tr>
<tr>
<td>MRS0 vs MRS5</td>
<td>1240.76*</td>
<td>1320.89*</td>
</tr>
<tr>
<td>MRS1 vs MRS2</td>
<td>135.88*</td>
<td>178.75*</td>
</tr>
<tr>
<td>MRS1 vs MRS3</td>
<td>522.23*</td>
<td>564.52*</td>
</tr>
<tr>
<td>MRS1 vs MRS4</td>
<td>1031.70*</td>
<td>1063.75*</td>
</tr>
<tr>
<td>MRS1 vs MRS5</td>
<td>1319.11*</td>
<td>1354.74*</td>
</tr>
<tr>
<td>MRS2 vs MRS3</td>
<td>386.35*</td>
<td>385.76*</td>
</tr>
<tr>
<td>MRS2 vs MRS4</td>
<td>895.82*</td>
<td>885.00*</td>
</tr>
<tr>
<td>MRS2 vs MRS5</td>
<td>1183.24*</td>
<td>1175.98*</td>
</tr>
<tr>
<td>MRS3 vs MRS4</td>
<td>599.47*</td>
<td>499.23*</td>
</tr>
<tr>
<td>MRS3 vs MRS5</td>
<td>796.88*</td>
<td>790.22*</td>
</tr>
<tr>
<td>MRS4 vs MRS5</td>
<td>287.42*</td>
<td>290.99*</td>
</tr>
</tbody>
</table>

KW indicates Kruskal Wallis test statistic. 
*P<0.0001.
Polytomous Logistic Regression Model: Model Appropriateness and Probabilistic Distribution

To predict disability level given BI or M-FIM score, polytomous logistic regression method was used. The appropriateness of the model was examined with the diagnostic statistics available in the SAS statistical package. Defining 6 levels for the MRS as dependent variable was not appropriate in the polytomous logistic model for either the BI or M-FIM score. On the basis of model diagnostics, the BI score cannot differentiate between the 6 different levels of MRS but is appropriate for 5 levels in the MRS. These are MRS5, MRS4, MRS3, MRS2, and the collapsed level of MRS0 and MRS1 [MRS(0,1)]. The M-FIM scale, however, was not able to achieve appropriateness until the scale was reduced to 3 levels.

The first graph in Figure 2 shows the BI score distribution for 5 levels. Although a 5-level model was statistically appropriate, MRS(0,1) was completely included in the MRS2 probability distribution line, so no differentiation between MRS2 and MRS(0,1) can be observed. The second graph in Figure 2 illustrates the 4-level MRS-BI model: MRS(0,1,2), MRS3, MRS4, and MRS5. A 4-level MRS-BI model differentiates between individual levels. Figure 3 presents the probabilistic distribution of M-FIM scores in the 3-level MRS–M-FIM model.

Categorization of BI

From the results of polytomous logistic regression analysis, cutoffs in the BI were determined. Ideal cutoffs were defined as the corresponding scores at intersections of probabilistic distributions for 2 adjacent probability lines, which means that these scores have an equal 50% probability of being located in either of 2 adjacent MRS levels. The ideal cutoff points, however, do not represent the real BI score because they are ideal values generated by the probability density function based on logistic regression estimates. Several scores around the ideal cutoffs were selected as potential cutoffs, and the combinations of potential cutoffs were selected to be mutually exclusive and exhaustive. Among combinations of potential cutoffs, the 1 set minimizing the false-positive response rate was selected for the BI score. The false-positive rate was minimized at 21.6% for this categorization scheme: 0 ≤ MRS5 < 15, 15 ≤ MRS4 < 70, 70 ≤ MRS3 < 95, and 95 ≤ MRS(0,1,2) ≤ 100. With this categorization scheme, 603, 461, 484, and 128 records were categorized to MRS(0,1,2), MRS3, MRS4, and MRS5, respectively, compared with 606, 487, 459, and 127 from the MRS.

Categorization of the M-FIM

According to the model appropriateness test in logistic regression, the 3-level model was found to be appropriate to categorize the M-FIM score: MRS(0,1,2,3), MRS4, and MRS5. The optimal set of cutoffs minimizing the false-positive response rate at 12% was as follows: 13 ≤ MRS5 ≤ 26, 26 ≤ MRS4 ≤ 62, and 62 ≤ MRS(0,1,2,3) ≤ 91. With this categorization scheme, 1103, 448, and 125 records were categorized to MRS(0,1,2,3), MRS4, and MRS5, respectively, compared with 1093, 459, and 127 from the MRS.

Time Influence on Categorization Scheme

The data from 4 waves were collapsed to increase the sample size and to obtain more robust analytical results. To verify whether the categorization schemes behaved differently for each wave, probability density functions were generated and examined for each time point. Figure 4 illustrates the probability distribution of each MRS level given the BI or M-FIM score for baseline, month 1, month 3, and month 6. Even when data from each wave were separated for logistic regression, the distribution presented similar trend and cutoff points.

Discussion

To measure poststroke disability, researchers and practitioners often use basic ADL measures. This study showed the relationship among widely used poststroke ADL measures and a global disability measure. Scales were highly correlated, and the Kruskal-Wallis one-way analysis of variance test and multiple comparison tests showed that the mean scores of BI and M-FIM were not different among levels representing higher functioning. To examine the predictability of the BI and M-FIM scores to MRS level, polytomous logistic regression was performed, and the higher ADL scores cannot efficiently differentiate between the disabilities levels. In this study, we used the MRS as a reference to categorize the BI and M-FIM scores because we are interested primarily in developing a scheme that converts basic ADL measures to a global disability measure that presents clinically distinct disability levels. Definition of these stages can be used in simulating long-term disability outcomes in poststroke population with a Markov model. The MRS is appropriate as a reference for this purpose because the 6 levels are clinically distinct and respond well to patient-reported outcomes.
Our finding about the ceiling effect of BI and M-FIM was consistent with other studies. The ceiling effect of BI and M-FIM compared with MRS is evident; however, it is due mainly to the different nature of the scale itself rather than the inferiority of either instrument. Choice of instrument would depend on the purpose of the study. For example, in longitudinal follow-up studies, researchers usually use the BI or M-FIM; both enable capture of minimal changes in physical functioning. In contrast, MRS is not sensitive enough to detect those small changes.

The schemes were developed to translate the BI or M-FIM to MRS with logistic regression. Our interest was in predicting disability levels from the BI or M-FIM. From this standpoint, logistic regression model with MRS as the dependent variable and BI (or M-FIM) as the independent variable was appropriate, and for the BI, the Kruskal-Wallis and logistic regression results were consistent. However, for the M-FIM, logistic regression and Kruskal-Wallis test results were not consistent: M-FIM was differentiated into 5 categories with the Kruskal-Wallis test, but logistic regression modeled only 3 categories.

Two explanations for this phenomenon are possible. First, test results are very likely to be dependent on sample sizes because the Kruskal-Wallis and multiple comparison methods examined whether there was statistical significance in mean BI (or M-FIM) scores among MRS levels. In fact, for MRS0, there were only 28 observations, and the small sample size likely affected the Kruskal-Wallis test results. Second, it is possible that M-FIM scores may not be relevant for a linear model. Using nonlinear regression models and/or having a large sample size might improve the M-FIM prediction of MRS categories. Future research is needed to look into this issue.

As shown in this study, intervals in the BI or M-FIM cannot be treated equally, so arithmetic cutoffs of BI or M-FIM scores without appropriate validation should be avoided. Often, the BI score of 60 has been used to identify large sample size likely affected the Kruskal-Wallis test results. Second, it is possible that M-FIM scores may not be relevant for a linear model. Using nonlinear regression models and/or having a large sample size might improve the M-FIM prediction of MRS categories. Future research is needed to look into this issue.

As shown in this study, intervals in the BI or M-FIM cannot be treated equally, so arithmetic cutoffs of BI or M-FIM scores without appropriate validation should be avoided. Often, the BI score of 60 has been used to identify different degrees of disability in patients, but this approach should be considered carefully because it is not clear that this dichotomization creates clinical significant groups.

Both BI and M-FIM are valid and reliable measures of ADL and widely used instruments for longitudinal follow-up in poststroke outcomes in patients. Translating schemes to global disability level from those ADL measures would facilitate better understanding of different degrees of disability from a population perspective.

Recently, a few studies used a modified version of the BI in stroke outcomes research. It would be more informative if we had measured the modified BI and compared results, but unfortunately, this instrument was not available when this study was performed. Future research is needed to investigate the predicting ability of the modified BI of the disability categories.

To obtain a robust categorization scheme, we aggregated data for baseline and 1, 3, and 6 months. We also performed logistic regression analysis for each time point, and as shown in Figure 4, the time effect on the categorization scheme was found to be minimal. The distribution and cutoffs for different disability levels had similar trends for different time points. Even though the overall distribution and categorization schemes were similar, variation in disability was reduced as time passed after stroke. Because of this reduction and the relatively small sample size compared with the aggregated one, the logistic model for individual time points was not consistently stable in terms of logistic regression model fit statistics. If the sample size were larger, this instability of the model would be reduced. Further research is needed on the generalizability of the categorization scheme.

Acknowledgments

This study was conducted with KCSS data. We acknowledge the 12 participating hospitals in the greater Kansas City area: Baptist Hospital, Department of Veterans Affairs Medical Centers at Kansas City and Leavenworth, Liberty Hospital, Medical Center of Independence, Mid-American Rehabilitation Hospital, Rehabilitation Institute, Research Medical Center, St Luke’s Hospital, St Joseph Health Center, Trinity Lutheran Hospital, and the University of Kansas Medical Center. We also acknowledge Samuel Wu, PhD, for comments on the statistical analysis and Brian Sauer and Jessica DeLeon, PhD, for comments on the manuscript.

References

Disability Measures in Stroke: Relationship Among the Barthel Index, the Functional Independence Measure, and the Modified Rankin Scale

Sooyeon Kwon, Abraham G. Hartzema, Pamela W. Duncan and Sue Min-Lai

*Stroke.* 2004;35:918-923; originally published online February 19, 2004;
doi: 10.1161/01.STR.0000119385.56094.32

*Stroke* is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2004 American Heart Association, Inc. All rights reserved.
Print ISSN: 0039-2499. Online ISSN: 1524-4628

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://stroke.ahajournals.org/content/35/4/918

**Permissions:** Requests for permissions to reproduce figures, tables, or portions of articles originally published in *Stroke* can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

**Reprints:** Information about reprints can be found online at:
http://www.lww.com/reprints

**Subscriptions:** Information about subscribing to *Stroke* is online at:
http://stroke.ahajournals.org/subscriptions/