Use of the Multiple Sclerosis Functional Composite as an Outcome Measure in a Phase 3 Clinical Trial

Jeffrey A. Cohen, MD; Gary R. Cutter, PhD; Jill S. Fischer, PhD; Andrew D. Goodman, MD; Fedor R. Heidenreich, MD; Amy J. Jak, MA; Judith E. Kniker, MA; Mariska F. Kooijmans, MD, PhD; Julia M. Lull, BA; Alfred W. Sandrock, MD, PhD; Jack H. Simon, MD; Nancy A. Simonian, MD; John N. Whitaker, MD; for the IMPACT Investigators

Background: The Multiple Sclerosis Functional Composite (MSFC) is a multidimensional clinical outcome measure that includes quantitative tests of leg function/ambulation (Timed 25-Foot Walk), arm function (9-Hole Peg Test), and cognitive function (Paced Auditory Serial Addition Test). The MSFC is the primary outcome measure in the ongoing multinational phase 3 trial of interferon beta-1a (Avonex) in patients with secondary progressive MS.

Objective: To assess the practice effects, reliability, and validity of the MSFC clinical outcome measure.

Design: Examining technicians underwent formal training using standardized materials. The MSFC was performed according to a standardized protocol. The 436 patients enrolled in the International Multiple Sclerosis Secondary Progressive Avonex Controlled Trial underwent 3 prebaseline MSFC testing sessions before randomization.

Results: Practice effects were evident initially for the MSFC but stabilized by the fourth administration. The Paced Auditory Serial Addition Test demonstrated the most prominent practice effects. The reliability of the MSFC was excellent, with an intraclass correlation coefficient for session 3 (final prebaseline session) vs session 4 (baseline) of 0.90. The MSFC at baseline correlated moderately strongly with the Kurtzke Expanded Disability Status Scale. Among the MSFC components, the Timed 25-Foot Walk correlated most closely. Correlations among the 3 MSFC components were weak, suggesting they assess distinct aspects of neurologic function in patients with MS.

Conclusions: The MSFC demonstrated excellent intrarater reliability in this multinational phase 3 trial. Three prebaseline testing sessions were sufficient to compensate for practice effects. The pattern of correlations among the MSFC, its components, and the Kurtzke Expanded Disability Status Scale supported the validity of the MSFC.

Arch Neurol. 2001;58:961-967
PATIENTS AND METHODS

PATIENT ENROLLMENT

Informed consent for participation in IMPACT was obtained after the potential risks and benefits of the study were reviewed with potential subjects.

TECHNICIAN TRAINING

Examining technicians at 42 study sites in the United States, Canada, and Europe, most without previous experience with the MSFC, were trained to administer the MSFC at a pre-study investigators' meeting. Formal training included a standardized didactic description of the MSFC background and testing procedures, review of the MSFC administration and scoring manual, viewing of a training videotape, and an interactive MSFC administration practice session with feedback. Examiners' knowledge and performance were assessed at the end of the training session to confirm familiarity with the MSFC procedures. Examiner training required approximately 4 hours.

MSFC TESTING PROTOCOL

The MSFC was administered to patients using a standardized protocol. The order of testing in a session was as follows: (1) T25FW (trial 1 and trial 2), (2) 9HPT (dominant hand: trial 1 and trial 2; and nondominant hand: trial 1 and trial 2), and (3) PASAT (3-second interstimulus interval and 2-second interstimulus interval).

Patient instructions for the T25FW, 9HPT, and PASAT were available in English, French, German, Dutch, Greek, and Hebrew. Two alternate forms for the PASAT were used. The order in which the 2 forms was used varied among patients and among sessions for an individual patient. The same examining technician at each study site administered the MSFC to a given patient at each study visit. To compensate for practice effects and to achieve a stable baseline before randomization, the IMPACT protocol incorporated 3 prebaseline testing sessions over 28 days before randomization. Examining technicians did not have access to the results of previous testing sessions once completed.

STATISTICAL ANALYSES

The MSFC score was derived from 3 components: (1) T25FW (the mean of the scores of the 2 T25FW trials), (2) 9HPT (the 2 trials for each hand were averaged, then converted to reciprocals, and the 2 reciprocals were then averaged), and (3) PASAT3 (the number correct on the PASAT with a 3-second interstimulus interval). The MSFC score was calculated as the mean of the z scores of the 3 components (T25FW, 9HPT, and PASAT3). A z score is a standardized score representing the number of SD units a given value is from a population mean and is calculated by subtracting the mean of the reference population from the test result, then dividing by the SD of the reference population. z Scores for the T25FW, 9HPT, and PASAT3 were calculated with reference to the pooled data set derived from the IMPACT baseline visit (session 4). For the T25FW and 9HPT, a higher raw score (time to complete the task in seconds) represents deterioration, whereas for the PASAT3, a lower raw score (number correct) represents deterioration. Consequently, in combining the 3 components into a single z score, the sign of the T25FW z score was reversed, and the z score of the inverse of the 9HPT time was used so that the direction of change was consistent across components. Thus, a decrease in the z scores of the 3 components and of the MSFC z score all represent deterioration in neurologic function.

Correlations among the MSFC, its components, and the EDSS were analyzed using Spearman rank correlations. The intraclass correlation coefficient was used to measure the session-to-session intrarater reliability of the MSFC. The MSFC score for session 3 (the last prebaseline session) was compared with that of session 4 (the baseline session) as the primary measure of reliability. All statistical analyses were performed using SAS statistical software (SAS Institute Inc, Cary, NC).

RESULTS

BASELINE PATIENT CHARACTERISTICS

A total of 436 subjects with SP MS and an EDSS score of 3.5 to 6.5 were enrolled in IMPACT. Baseline demographic and clinical characteristics are as follows:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female-male ratio</td>
<td>279:157</td>
</tr>
<tr>
<td>Age, mean±SD, y</td>
<td>47.6±9.0</td>
</tr>
<tr>
<td>Duration of MS, mean±SD, y*</td>
<td>16.5±9.0</td>
</tr>
<tr>
<td>EDSS score, mean±SD</td>
<td>5.2±1.1</td>
</tr>
</tbody>
</table>

*Since the onset of symptoms.

Figure 1 illustrates the distribution of patients across the range of baseline EDSS scores. As expected, there was an overrepresentation of patients with EDSS scores of 4.0, 6.0, and 6.5, with a relative underrepresentation of patients with EDSS scores of 4.5 to 5.5, as has been seen in most population distributions for the EDSS.15,16
COMPARISON WITH THE NMSS TASK FORCE DATA SET

The z scores for the MSFC components were calculated with reference to the group means and SDs from the IMPACT baseline visit (session 4). The sample size estimate for IMPACT was based on an analysis of a cohort of 326 patients in the NMSS Task Force data set with SP MS and an EDSS score of 3.5 to 6.5. The values at baseline in IMPACT were comparable to those in the NMSS Task Force data set, with some slightly greater mean and SD of the T25FW score in the IMPACT baseline data set.

PRACTICE EFFECTS

Practice effects on the MSFC, manifested as improved performance and decreased session-to-session variability, were evident in the first 3 prebaseline sessions but stabilized by the fourth session (Table 2). At the time of this analysis, pooled data blind to treatment group from the first postenrollment assessment at month 3 were available for 426 patients. The mean MSFC z score decreased slightly from 0.00 at baseline to −0.12 at month 3.

RELIABILITY

The MSFC demonstrated excellent intrarater reliability in IMPACT. The intraclass correlation coefficient for the MSFC for session 3 (the last prebaseline session) vs session 4 (the baseline visit) was 0.90. The intraclass correlation coefficient for the 4 sessions taken together was 0.87, despite the increased variability in the early testing sessions due to practice effects.

MSFC-EDSS CORRELATIONS

Table 3 provides the Spearman rank correlations among the MSFC, its components, and the EDSS at baseline. Similar to the analysis of the NMSS Task Force data set, correlation between the MSFC and the EDSS was moderately strong, even over the more restricted EDSS score range.
The MSFC developed by the NMSS Task Force is anticipated to have better reliability, advantageous psychometric properties, and greater sensitivity to change compared with the EDSS and other available MS clinical outcome measures. The results of previous studies and those presented herein demonstrate that the MSFC meets many of these goals.

It was anticipated that most clinicians and most patients with MS would be unfamiliar with the MSFC. Therefore, in preparation for IMPACT, standardized examiner training materials and a patient testing protocol were developed. A pilot study confirmed the effectiveness of the examiner training procedures and the feasibility of the testing protocol in terms of ease of administration, the time required, and patient acceptance. The MSFC demonstrated excellent intrarater and interrater reliability in this small pilot study carried out at a single site. The MSFC again demonstrated excellent reliability in IMPACT, a multinational study with 42 participating sites, virtually none of which had prior experience with the measure. The reliability of the MSFC in IMPACT was substantially better than that previously reported for the EDSS in other trials.

Because of previous clinical experience with quantitative measures and analysis of the NMSS Task Force pooled data set, practice effects on the MSFC were expected to occur. That is, performance and variability were anticipated to improve as subjects (and examiners) became familiar with the tasks. In the pilot study carried out before IMPACT, practice effects appeared to stabilize by the fourth testing session. Therefore, the IMPACT protocol included 3 prebaseline testing sessions over 28 days before randomization. The experience in IMPACT confirmed that the MSFC is susceptible to practice effects. The substantial variability in the change from session to session before baseline suggested that the magnitude and time course of practice effects differed among patients. This potential variability could obscure underlying biological change and impede the detection of a treatment effect in the context of a clinical trial. Three prebaseline sessions appeared to be sufficient to compensate for practice effects on the MSFC, which stabilized by the fourth session. The pooled MSFC data from month 3 demonstrated neither substantial improvement, which would suggest further practice effects, nor substantial worsening, which would suggest “forgetting.”

With the small sample size in the pilot study, it was not possible to discern the relative contributions of the 3 MSFC components to the practice effects seen for the overall measure. With the larger sample size in IMPACT, it was clear that the primary source of practice effects on the MSFC was the PASAT3. Strong practice effects on the PASAT have been reported previously in normal and neurologically impaired subjects and in subjects with MS. In IMPACT, modest practice effects on the 9HPT also were evident. We are unaware of any previous systematic study of variability over time or practice effects on the 9HPT in patients with impaired upper extremity function, including those with MS. No practice effects were seen on the T25FW in IMPACT. A previous study of 63 subjects with MS undergoing T25FW testing on 5 consecutive days also failed to detect any systematic trend suggesting practice effects (Steven Schwind, MD, written communication, March 1999). Although the main source of MSFC practice effects is the PASAT3, we recommend that prebaseline testing include all 3 components to fa-

---

**Table 3. Correlation of the MSFC and EDSS at Baseline**

<table>
<thead>
<tr>
<th>Test</th>
<th>9HPT</th>
<th>PASAT3</th>
<th>MSFC</th>
<th>EDSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>T25FW</td>
<td>0.51</td>
<td>0.26</td>
<td>0.66</td>
<td>-0.78</td>
</tr>
<tr>
<td>9HPT</td>
<td>0.35</td>
<td>0.84</td>
<td>-0.47</td>
<td></td>
</tr>
<tr>
<td>PASAT3</td>
<td>0.70</td>
<td>-0.23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSFC</td>
<td>-0.56</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*S* Spearman rank correlations (r values) are listed among the EDSS and z scores for the MSFC and its components. All correlations were statistically significant (P < .001 [n = 436]). MSFC indicates Multiple Sclerosis Functional Composite; EDSS, Kurtzke Expanded Disability Status Scale; 9HPT, 9-Hole Peg Test; PASAT3, Paced Auditory Serial Addition Test with a 3-second interstimulus interval; and T25FW, Timed 25-Foot Walk.

---

Quantitative assessment of neurologic impairment and disability in patients with MS is difficult because of the marked clinical heterogeneity exhibited by the disease, between patients and within individual patients over time. There is variability in the manifestations that are present, their severity, and the time course over which they develop. Also, some of the common manifestations of MS that can have substantial impact on quality of life, such as cognitive dysfunction, are difficult to quantify using the standard neurologic examination and the rating scales that can have substantial impact on quality of life, such as cognitive dysfunction, are difficult to quantify using the standard neurologic examination and the rating scales that can have substantial impact on quality of life, such as cognitive dysfunction, are difficult to quantify using the standard neurologic examination and the rating scales that can have substantial impact on quality of life, such as cognitive dysfunction, are difficult to quantify using the standard neurologic examination and the rating scales that can have substantial impact on quality of life, such as cognitive dysfunction, are difficult to quantify using the standard neurologic examination and the rating scales.
miliarize subjects and examiners with the standardized MSFC testing protocol.

Two remaining key issues with the MSFC are its validity as a measure of disability in patients with MS and its utility as an outcome measure in clinical trials. Preliminary studies demonstrated good correlation between magnetic resonance imaging measures of cranial lesion burden and atrophy and the MSFC. Three previous studies have shown that the MSFC correlated with the EDSS, the most widely used measure of neurologic disability in MS clinical trials. The MSFC was shown to be worse in patients with SP MS compared with those with relapsing-remitting disease. Finally, the MSFC correlated well with patient self-report measures of symptoms and quality of life.

The results reported herein for IMPACT provide further support for the validity of the MSFC from analysis of a fourth independent data set. First, the population values for the T25FW, 9HPT, and PASAT3 at baseline in IMPACT were comparable to those calculated for the cohort of patients with SP MS and an EDSS score of 3.5 to 6.5 in the NMSS Task Force data set used to estimate sample size for the trial. The somewhat greater mean and SD shown for the T25FW at baseline in IMPACT likely was due to the higher proportion of patients with EDSS scores of 6.0 and 6.5.

More important, correlations among the MSFC, its components, and the EDSS, which were collected prospectively in IMPACT, were similar to those previously reported by the NMSS Task Force based on retrospective analysis of existing data sets. Correlations among the 3 components of the MSFC were modest. Thus, although patients with MS tend to demonstrate parallel impairment in several domains, the T25FW, 9HPT, and PASAT3 clearly measure independent clinical dimensions of MS. As expected, the correlation was strongest between the T25FW and 9HPT, but neither correlated well with the PASAT3. The strengths of the correlations between the MSFC and its 3 components were roughly comparable, suggesting that all 3 contribute information to the MSFC for the patient population as a whole. The MSFC correlated moderately with the EDSS, supporting the convergent validity of the MSFC (correlation with another measure of neurologic disability) and its divergent validity (measurement of aspects of MS not covered by the EDSS). As expected, among the components of the MSFC, the T25FW correlated best with the EDSS. For an EDSS score between 3.5 and 6.5, the range used as an enrollment criterion for IMPACT, the EDSS is primarily an ambulation scale with the score determined solely by how far a patient can walk and the type of assistive device required. It was not surprising that walking speed and distance would be related. Correlation between the 9HPT and EDSS was less strong. Patients with motor impairment affecting ambulation also tend to have motor impairment in the

![Figure 3. Distributions of baseline z scores (given as mean±SD) for the Multiple Sclerosis Functional Composite (MSFC) (A), the Timed 25-Foot Walk (T25FW) (B), the 9-Hole Peg Test (9HPT) (C), and the Paced Auditory Serial Addition Test with a 3-second interstimulus interval (PASAT3) (D) as a function of the Kurtzke Expanded Disability Status Scale (EDSS) score.](http://jamanetwork.com)
arms. However, arm function does not affect EDSS scoring in the range studied in IMPACT. Correlation between the PASAT3 and the EDSS was weak. The MSFC assesses aspects of neurologic function not measured by the EDSS, suggesting that it will be more sensitive to the effects on the MSFC. The baseline IMPACT results corroborated the results of previous studies correlating the MSFC and the EDSS. The pattern of correlations among the MSFC, its components, and the EDSS supported the validity of the MSFC. These results again indicated that the MSFC assesses aspects of neurologic function not measured by the EDSS, suggesting that it will be more sen-
sitive to detect change over time and better able to demonstrate a therapeutic effect when one exists.

Accepted for publication November 27, 2000.

From the Mellen Center for Multiple Sclerosis Treatment and Research and the Department of Neurology, The Cleveland Clinic Foundation, Cleveland, Ohio (Drs Cohen and Fischer and Mss Jak and Kniker); the Center for Research Methodology and Biometrics, AMC Cancer Center, Lakewood, Colo (Dr Cutter); the Departments of Neurology, University of Rochester, Rochester, NY (Dr Goodman), Hannover Medical School, Hannover, Germany (Dr Heidenreich), and the University of Alabama at Birmingham (Dr Whitaker); Biogen, Inc, Cambridge, Mass (Drs Kooijmans, Sandrock, and Simonian and Ms Lu!); and the Department of Radiology, University of Colorado, Denver (Dr Simon). Drs Kooijmans, Sandrock, and Simonian and Ms Lu! are full-time employees of Biogen, Inc. None of the other authors has a personal financial investment, ownership, equity, or other financial holdings with Biogen, Inc. Dr Fischer and Mss Jak and Kniker supervised the training of examining technicians and were reimbursed through a contract with Biogen, Inc, which was paid to The Cleveland Clinic. Drs Cohen, Cutter, Goodman, Heidenreich, Simon, and Whitaker have served as consultants for, received honoraria from, or received research support from Biogen, Inc.

This study and the development of the MSFC manual (a manual describing MSFC testing and scoring procedures, which is available through the National MS Society) were supported by Biogen, Inc, Cambridge, Mass.

Presented at the annual meeting of the American Academy of Neurology, Toronto, Ontario, April 22, 1999.

We thank Richard A. Rudick, MD, for reviewing the manuscript.

Corresponding author and reprints: Jeffrey A. Cohen, MD, The Mellen Center-U10, The Cleveland Clinic Foundation, 9500 Euclid Ave, Cleveland, OH 44195 (e-mail: cohenj@ccf.org).

REFERENCES