

## Test-Retest Reproducibility of the Exercise Treadmill Examination in Lumbar Spinal Stenosis

H. GORDON DEEN, JR, MD; RICHARD S. ZIMMERMAN, MD; MARK K. LYONS, MD;  
MALCOLM C. MCPHEE, MD; JOSEPH L. VERHELJDE, PhD; AND SUSAN M. LEMENS, PA-C

• **Objective:** To provide further validation of the treadmill test by assessing its "test-retest" reproducibility.

• **Patients and Methods:** In this prospective study, 28 patients with severe lumbar spinal stenosis underwent exercise treadmill testing, first at a walking speed of 1.2 mph and then at the patient's preferred walking speed. All patients had a second treadmill examination or "retest." No treatment intervention was performed between the initial test and the retest. Time to first symptoms (TFS) and total ambulation time (TAT) were measured. Differences between the baseline examination and the retest examination were assessed by using the concordance correlation coefficient (CCC) as well as graphically.

• **Results:** There was good reproducibility between baseline test and retest results for all 4 end points: 1.2 mph,

TFS (CCC = 0.90); 1.2 mph, TAT (CCC = 0.89); preferred walking speed, TFS (CCC = 0.98); and preferred walking speed, TAT (CCC = 0.96). The median difference between trials was not significantly different from zero for any of the 4 outcomes.

• **Conclusions:** Exercise treadmill testing has good test-retest reproducibility. There was no learning phenomenon associated with the test procedure. The study further validates the clinical utility of exercise treadmill testing in patients with lumbar spinal stenosis and neurogenic claudication.

*Mayo Clin Proc.* 2000;75:1002-1007

CCC = concordance correlation coefficient; TAT = total ambulation time; TFS = time to first symptoms

Degenerative lumbar spinal stenosis is a common condition in elderly persons. Lumbar decompressive laminectomy is often considered when the degree of stenosis is severe and symptoms are debilitating. Contemporary imaging tests, such as magnetic resonance imaging, provide a wealth of anatomical information and are able to define the presence and severity of lumbar spinal stenosis accurately; however, they provide little insight into the functional status of the patient, either before or after surgery.

In an effort to define baseline functional status and response to surgical intervention more objectively, protocols for exercise stress testing on a treadmill have been developed.<sup>1-4</sup> Treadmill testing of patients with symptomatic lumbar spinal stenosis has been shown to be safe, easily administered, inexpensive, and quantifiable. By using this technique, most patients have been shown to have objective improvement after lumbar decompressive laminectomy. Based on a functional grading system, 88% of 60 patients improved at least 1 functional grade after surgery.<sup>3</sup> A general belief is that the postoperative im-

provement is the direct result of decompressive surgery. However, we thought it was necessary to consider the possibility that the measured postoperative improvement might not be due to surgical intervention, but rather to a "learning phenomenon" derived from an increased familiarity and comfort with the treadmill examination format. This prospective study was undertaken to provide further validation of the treadmill test by evaluating its reproducibility and assessing whether there is any learning phenomenon by which patients could improve their treadmill test performance simply by practicing the test procedure.

### PATIENTS AND METHODS

Twenty-eight patients (17 men and 11 women) with a mean age of 73.8 years (range, 57-91 years) were prospectively studied. All had the clinical diagnosis of intractable neurogenic claudication, defined as leg pain or paresthesias, precipitated by standing and walking and relieved by sitting or lying down. All patients had severe lumbar spinal stenosis, confirmed by magnetic resonance imaging or computed tomography and myelography. Patients with peripheral vascular disease were excluded based on a history of vascular claudication, which is leg pain, usually in the calf, precipitated by walking and relieved by standing still. Peripheral pulses were examined in all patients, and noninvasive vascular studies were performed in selected

From the Department of Neurologic Surgery (H.G.D., R.S.Z., M.K.L., S.M.L.) and Department of Physical Medicine and Rehabilitation (M.C.M., J.L.V.), Mayo Clinic, Scottsdale, Ariz.

Address reprint requests and correspondence to H. Gordon Deen, Jr, MD, at his current address: Department of Neurosurgery, Mayo Clinic, 4500 San Pablo Rd, Jacksonville, FL 32224.

patients. Lumbar decompressive laminectomy of the stenotic areas was performed.

Medical comorbidities were typical for patients in this age group. Nine patients had symptomatic degenerative joint disease of the hip, knee, or foot; 7 had coronary artery disease or valvular heart disease; and 2 had Parkinson disease. Two other patients smoked cigarettes and had chronic obstructive pulmonary disease. One patient each had diabetes mellitus, carcinoma of the breast, and carcinoma of the prostate. The mean body mass index was 28.2 kg/m<sup>2</sup>, indicating that the typical patient was moderately overweight, but not obese.

## OUTCOME MEASURES

### Treadmill Examination

All patients had preoperative and postoperative exercise treadmill testing according to a standard protocol in which time to first symptoms (TFS) and total ambulation time (TAT) were measured, first at a walking speed of 1.2 mph and then at the patient's preferred walking speed. A ramp incline of 0° was used for all examinations. A time of zero was recorded when symptoms were present at the beginning of the test. The examination was stopped after 15 minutes or at the onset of severe symptoms, defined as the level of discomfort that would cause patients to stop walking in usual life situations. Patient-selected walking speeds ranged from 0.4 to 2.2 mph.

Patients were instructed to tell the examiner when symptoms first appeared. This time was recorded as TFS. The examination was stopped when symptoms were severe. This time was recorded as TAT. Patients were instructed to walk with an upright posture and to avoid using the front or side handrails. The test protocol has been described in more detail elsewhere.<sup>1</sup> All examinations were conducted during normal working hours, between 8 AM and 5 PM. There is no evidence that patient performance is affected by the time of day that the test is conducted, although this has not been studied formally.

To assess the short-term reproducibility of the procedure, all patients had a second treadmill examination or "retest" after the initial test. Nine patients were retested on the same day, 16 within 1 day, and 3 within 2 to 4 days. No surgery or any other treatment was performed during the time interval between the initial test and the retest. When both components were done the same day, the patient was given sufficient time to rest after the initial test before being retested.

Eighteen patients had a postoperative treadmill examination with test and retest components. This approach was used to study the reproducibility of the treadmill procedure after surgical treatment of the spinal stenosis.

## Statistical Methods

The reproducibility between the baseline examination and the retest examination was evaluated by using the concordance correlation coefficient (CCC).<sup>5</sup> This index evaluates the agreement between pairs of trials by measuring the variation from the line of equality through the origin (the concordance line) when the data are graphed as trial 2 vs trial 1. Like the standard correlation coefficients, the CCC ranges from -1 to 1, with 1 indicating perfect concordance. Although the Pearson correlation coefficient measures a linear relationship, it fails to detect any departures from the line of equality and therefore fails to detect nonreproducibility. Thus, the CCC statistic was used. The learning effect between trials was evaluated by assessing whether the median difference between trials was significantly different from zero by using the Wilcoxon signed rank test. This significance test based on the ranks of the data was used because the examinations were arbitrarily stopped at 15 minutes.

## RESULTS

The baseline trial results are summarized in Table 1. One half of the patients assessed preoperatively developed neurogenic claudication symptoms as soon as they started walking (median TFS, 0 minutes). This was true for both the 1.2 mph trial and the preferred walking speed trial. The median TAT of 5.3 minutes in the 1.2 mph preoperative group corresponds to a walking distance of 186 yards. In the postoperative trials, most patients could walk symptom-free for 15 minutes at both 1.2 mph and preferred walking speed.

Test-retest results are outlined in Table 2. At both walking speeds, more than one half of patients in the preoperative trial developed neurogenic claudication symptoms as soon as they started walking. In the preoperative trial, there was good reproducibility between baseline test and retest results for all 4 end points: 1.2 mph, TFS (CCC = 0.90); 1.2 mph, TAT (CCC = 0.89); preferred walking speed, TFS (CCC = 0.98); and preferred walking speed, TAT (CCC = 0.96). In the postoperative trial, most patients completed a full 15-minute examination, and there was little variability.

Figures 1 through 4 depict the preoperative test and retest trials for each patient. The solid line represents the line of equality. Pairs (test, retest) of trials that were within 1 minute fall within the dashed lines. Pairs of trials that were within 2 minutes fall within the dotted lines. If a learning effect were to occur, a greater proportion of patients would have retest values that exceeded their initial test values (ie, differences greater than zero); however, this was not observed. Based on the Wilcoxon signed rank test, the median difference between the test and retest trials was

Table 1. Summary of Baseline Trial Results\*

Period	Speed (mph)	Measure	Degree of performance				Median (IQR), min
			Unable to perform	0 seconds	>0 min	15 min	
Preoperative (n=28)	1.2	TFS	4 (14.3)	13 (46.4)	10 (35.7)	1 (3.6)	0 (0, 1.6)
	1.2	TAT	4 (14.3)	0 (0)	21 (75.0)	3 (10.7)	5.3 (2.0, 9.3)
	Preferred	TFS	0 (0)	18 (64.3)	9 (32.1)	1 (3.6)	0 (0, 1.3)
	Preferred	TAT	0 (0)	0 (0)	25 (89.3)	3 (10.7)	3.5 (1.7, 8.5)
Postoperative (n=18)	1.2	TFS	3 (16.7)	2 (11.1)	2 (11.1)	11 (61.1)	15 (10.8, 15)
	1.2	TAT	3 (16.7)	0 (0)	1 (5.6)	14 (77.8)	15 (15, 15)
	Preferred	TFS	0 (0)	4 (22.2)	3 (16.7)	11 (61.1)	15 (1.5, 15)
	Preferred	TAT	0 (0)	0 (0)	5 (27.8)	13 (72.2)	5 (10.1, 15)

\*Values are number (%) of patients unless indicated otherwise. IQR = interquartile range, 25th and 75th percentiles; TAT = total ambulation time; TFS = time to first symptoms.

not significantly different from zero for each of the 4 end points ( $P > .05$ ).

## DISCUSSION

Traditional clinical outcome measures and indicators are deficient in many areas of medicine.<sup>6</sup> Criteria for evaluating patients with back-related symptoms have been especially subjective and variable, leading to uncertainty about diagnostic criteria, indications for surgery, and surgical outcome.<sup>7</sup> A wide range of outcome indicators has been used to study such patients, including symptoms, findings on neurologic and musculoskeletal examinations, imaging test results, psychologic testing, employment status, disability status, and cost-utilization of medical services. Each has major limitations in its ability to assess baseline status and response to treatment. Currently available outcome

studies have been compromised by a focus on process of care and technical measures of success, rather than on measures of patient function and quality of life.<sup>8</sup> Recently, the trend has been toward emphasizing functional status as a key outcome indicator for patients with back-related symptoms. To improve the evaluation process, various patient-centered function and symptom rating systems, including questionnaires and functional tests, have been developed.

Standardized exercise protocols to assess cardiopulmonary function have been in use since 1929.<sup>9</sup> Exercise stress testing has been widely used for many years in patients with coronary artery disease.<sup>10,11</sup> The Bruce protocol is perhaps the most widely used. Other cardiovascular exercise tests include those of Naughton, Cornell, Balke-Ware, and Weber.<sup>12</sup>

Table 2. Summary of Test-Retest Results\*

Period	Speed (mph)	Measure	No. of patients	No. of patients with no change between trials		Paired difference (retest-test, min)		Concordance correlation coefficient
				0 seconds	15 min	Median (IQR)	Mean (SD)	
Preoperative (n=28)	1.2	TFS	24†	12	1	0.0 (0, 0.04)	-0.42 (1.6)	0.90
	1.2	TAT	24†	0	3	-0.09 (-2.5, 0.3)	-0.72 (2.1)	0.89
	Preferred	TFS	27‡	16	1	0.0 (0, 0)	0.04 (0.8)	0.98
	Preferred	TAT	27‡	0	3	-0.04 (-0.7, 1.0)	-0.08 (1.4)	0.96
Postoperative (n=18)	1.2	TFS	15§	1	11	...	...	...
	1.2	TAT	15§	0	13	...	...	...
	Preferred	TFS	18	3	11	...	...	...
	Preferred	TAT	18	0	13	...	...	...

\*Ellipses indicate that because of the lack of variability in the postoperative trial, ie, almost all patients completed a full 15-minute examination, these summary statistics were not estimated. IQR = interquartile range, 25th and 75th percentiles; TAT = total ambulation time; TFS = time to first symptoms.

†Four patients could not perform at 1.2 mph, in either trial.

‡One patient could not perform at a preferred speed in the second trial.

§Three patients could not perform at 1.2 mph, in either trial.

Historically, exercise testing has been infrequently used in patients with spinal disorders. First reported by van Gelderen<sup>13</sup> in 1948, functional status has been assessed by using a bicycle,<sup>14</sup> a horizontal treadmill or other level surface,<sup>15</sup> and a treadmill with a downhill slope.<sup>16</sup> Dong and Porter<sup>17</sup> reported a study comparing walking and bicycling tolerance. Treadmill testing has also been reported to have utility in screening for lumbar instability.<sup>18</sup> Fritz et al<sup>19</sup> reported a series of patients with lumbar spinal stenosis who were assessed with the exercise treadmill. None of these studies used treadmill testing to measure surgical outcome.

Expanding on these studies, functional exercise testing has been extended to the preoperative and postoperative assessment of patients with degenerative lumbar spinal stenosis. Treadmill testing has the advantage of directly assessing neurogenic claudication, the predominant symptom for which the patient has sought medical attention. Patients are screened for medical conditions that would make exercise testing unsafe.<sup>13</sup> All analgesic medications are withdrawn for 24 hours before the procedure. With a few simple instructions and a brief period of adjusting to the treadmill apparatus, virtually all patients can perform the examination. The procedure is done initially as a preoperative baseline assessment of functional status and repeated 3 months after surgery. Waiting 3 months before performing the postoperative assessment is important because in some patients symptoms do not immediately resolve but improve gradually. Patients are assigned a functional grade ranging from 1 (best category) to 4 (worst category) based on treadmill ambulation times as follows: grade 1, 15 minutes symptom-free ambulation; grade 2, 15 minutes with symptoms; grade 3, 5 to 15 minutes with symptoms; and grade 4, less than 5 minutes with symptoms.

Functional rating systems can be considered based on their practicality, comprehensiveness, validity, reproducibility, and responsiveness.<sup>20</sup> Treadmill testing has proved to be practical. Ongoing experience with more than 150 patients over a 3-year period indicates that the examination is fast, easy to administer, inexpensive, quantifiable, and readily done on a serial basis, eg, before and after treatment. In contrast to a generic questionnaire, such as the Short-Form Health Survey, which is applicable to a wide range of health concerns, treadmill testing in patients with spinal disorders has been used only for neurogenic claudication and would therefore be considered a focused, rather than a comprehensive, testing instrument. The procedure has validity in that it measures what it claims to measure, ie, functional capacity of patients with neurogenic claudication due to lumbar spinal stenosis. Test results correlate with other measures of disease severity, such as degree of stenosis demonstrated on imaging studies.

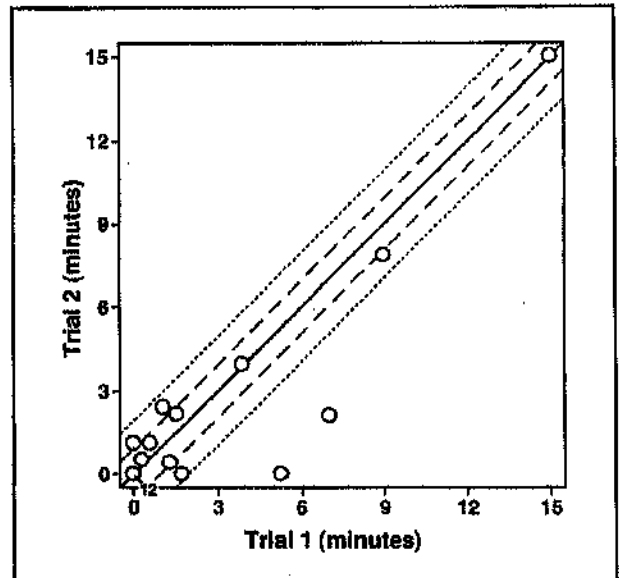


Figure 1. Preoperative treadmill test-retest: time to first symptoms, 1.2 mph trial. In 12 patients, symptoms first occurred at time zero during both trials. The solid line represents the line of equality. Pairs (test, retest) of trials that were within 1 minute fall within the dashed lines. Pairs of trials that were within 2 minutes fall within the dotted lines.

The key finding in the current study is that treadmill testing has a high degree of short-term reproducibility. Serial treadmill test results showed very little variation

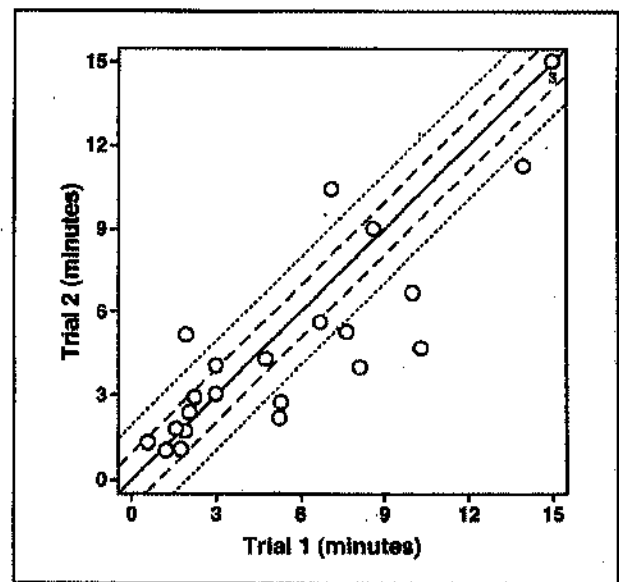


Figure 2. Preoperative treadmill test-retest: total ambulation time, 1.2 mph trial. For more information, see legend to Figure 1.

R),

3)

5)

5)

)

5)

5)

les;

process of

r than on

Recently,

I status as

ck-related

a, various

systems.

have been

liopulmo-

cise stress

n patients

rotocol is

ular exer-

lke-Ware,

ncordance

relation

efficient

0.90

0.89

0.98

0.96

...

...

...

...

15-minute

ambulation

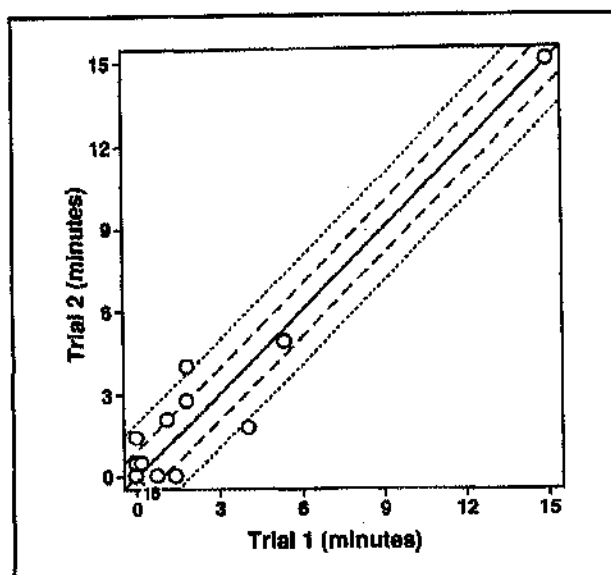


Figure 3. Preoperative treadmill test-retest: time to first symptoms, preferred walking speed. For more information, see legend to Figure 1.

over time, in the absence of surgery or other treatment. There was no appreciable learning phenomenon whereby a patient might have achieved a better treadmill performance simply by gaining experience and familiarity with the testing format. This high degree of test-retest reproducibility gives the clinician confidence that an improved postopera-

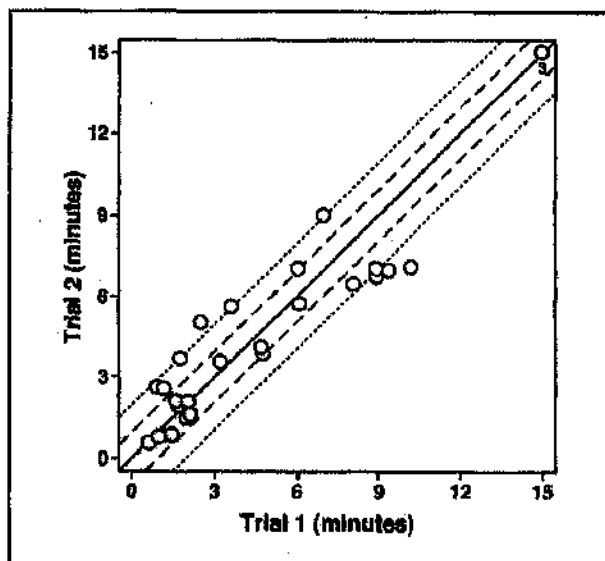


Figure 4. Preoperative treadmill test-retest: total ambulation time, preferred walking speed. For more information, see legend to Figure 1.

tive treadmill test performance is the direct result of surgery and that changes in test performance indicate a real change in the patient's functional status.

A substantial number of patients in this series developed neurogenic claudication as soon as they started to walk. This accounts for the 12 patients clustered at zero in Figure 1. These patients could all ambulate some distance at 1.2 mph, at a slower preferred walking speed, or both. This indicates that TAT is probably a more useful measure than TFS. Patients with neurogenic claudication and lumbar spinal stenosis are rarely confined to a wheelchair. Therefore, treadmill testing should be feasible in almost all circumstances.

A "ceiling effect" is a potential concern on the postoperative treadmill examination, which is arbitrarily stopped at 15 minutes. Many patients are grade 1 ambulators (15 minutes symptom-free ambulation). Our early experience indicated that if a patient could complete 15 minutes on the treadmill, he or she could continue to walk essentially without limitation. Thus, we have not found it helpful for patients to walk for more than 15 minutes. Nevertheless, there may be some functional differences among these patients, which could be established through questionnaire measurement of function.

Responsiveness is the capacity of a test instrument to detect changes over time. Further work is needed for a more accurate definition of the responsiveness of the treadmill examination. Our previous studies confirmed a basic level of responsiveness with documentation of substantial improvement in test performance after lumbar decompressive laminectomy in most cases.<sup>1-3</sup> Additional study is needed to determine whether the treadmill procedure will be able to detect more subtle, but still clinically important, changes in patient function that might occur after conservative treatment interventions. Further investigation is also necessary to define the role of treadmill testing in the long-term assessment of the postoperative patient.

Several questionnaire-based rating scales, including the Short-Form Health Survey, Oswestry Questionnaire, and Roland Questionnaire, have been validated for use in patients with disorders of the lumbar spine.<sup>21</sup> The treadmill examination is the only functional test that has been validated for use in these patients.

## CONCLUSIONS

Exercise treadmill testing can be used to assess the functional status of patients with symptomatic lumbar spinal stenosis. The current study shows that there is no learning phenomenon associated with the procedure and that patients do not experience improved test performance in the absence of surgical intervention. These findings indicate good test-retest reproducibility. Clinicians can have a high

degree of confidence that improved treadmill test performance in the postoperative patient is a direct result of surgery and indicates a real change in the patient's functional status. Treadmill testing appears to be a practical, focused, reproducible, and valid outcome measure for patients with neurogenic claudication due to lumbar spinal stenosis.

#### ACKNOWLEDGMENT

We thank Amy L. Weaver, MS, for preparation of the statistical analysis in this article.

#### REFERENCES

1. Deen HG Jr, Zimmerman RS, Lyons MK, McPhee MC, Verheijde JL, Lemens SM. Measurement of exercise tolerance on the treadmill in patients with symptomatic lumbar spinal stenosis: a useful indicator of functional status and surgical outcome. *J Neurosurg*. 1995;83:27-30.
2. Deen HG, Zimmerman RS, Lyons MK, McPhee MC, Verheijde JL, Lemens SM. Use of the exercise treadmill to measure baseline functional status and surgical outcome in patients with severe lumbar spinal stenosis. *Spine*. 1998;23:244-248.
3. Deen HG. Exercise treadmill testing in patients with lumbar spinal stenosis. *Perspect Neurol Surg*. 1998;9:11-22.
4. Gupta P, Tenhula J, Lenke L, Bridwell K, Chapman M, Marsicano J. Evaluation/functional outcome of neurogenic claudication using treadmill-bicycle testing [abstract]. In: Proceedings of the North American Spine Society Meeting; October 23-26, 1996; Vancouver, British Columbia. 1996;11:138-139.
5. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*. 1989;45:255-268.
6. Walters BC. Clinical practice parameter development in neurosurgery. *Concepts Neurosurg*. 1998;9:99-111.
7. Turner JA, Ersek M, Herron L, Deyo R. Surgery for lumbar spinal stenosis: attempted meta-analysis of the literature. *Spine*. 1992;17:1-8.
8. Keller RB, Rudicel SA, Liang MH. Outcomes research in orthopaedics. *J Bone Joint Surg Am*. 1993;75:1562-1574.
9. Master AM, Oppenheimer ET. Simple exercise tolerance test for circulatory efficiency with standard tables for normal individuals. *Am J Med Sci*. 1929;177:223-243.
10. Ellestad MH. *Stress Testing: Principles and Practice*. 3rd ed. Philadelphia, Pa: FA Davis; 1986.
11. Froelicher VF. *Exercise and the Heart: Clinical Concepts*. 2nd ed. Chicago, Ill: Year Book Medical Publishers; 1987.
12. Chaitman B. Exercise stress testing. In: Braunwald E, ed. *Heart Disease: A Textbook of Cardiovascular Medicine*. 4th ed. Philadelphia, Pa: WB Saunders Co; 1992:164.
13. van Gelderen C. Ein orthotisches (jordotisches) Kaudas Syndrom. *Acta Psychiatr Neurol*. 1948;23:57-68.
14. Dyck P, Doyle JB Jr. "Bicycle test" of van Gelderen in diagnosis of intermittent cauda equina compression syndrome: case report. *J Neurosurg*. 1977;46:667-670.
15. Johnsson K-E, Willner S, Pettersson H. Analysis of operated cases with lumbar spinal stenosis. *Acta Orthop Scand*. 1981;52:427-433.
16. Jensen OH, Schmidt-Olsen S. A new functional test in the diagnostic evaluation of neurogenic intermittent claudication. *Clin Rheumatol*. 1989;8:363-367.
17. Dong G, Porter RW. Walking and cycling tests in neurogenic and intermittent claudication. *Spine*. 1989;14:965-969.
18. Tokuhashi Y, Matsuzaki H, Sano S. Evaluation of clinical lumbar instability using the treadmill. *Spine*. 1993;18:2321-2324.
19. Fritz JM, Erhard RE, Delitto A, Welch WC, Nowakowski PE. Preliminary results of the use of a two-stage treadmill test as a clinical diagnostic tool in the differential diagnosis of lumbar spinal stenosis. *J Spinal Disord*. 1997;10:410-416.
20. Deyo RA. Measuring the functional status of patients with low back pain. *Arch Phys Med Rehabil*. 1988;69:1044-1053.
21. Deen HG Jr. Use of patient-centered function and symptom rating systems in spinal disorders. *Mayo Clin Proc*. 1999;74:40-44.