Vascular Medicine 2007; **12**: 97–104

Optimal assessment of baseline treadmill walking performance in claudication clinical trials

Eric P Brass[a], Jenny Jiao[b] and William Hiatt[c]

Abstract: Treadmill testing is frequently used to assess the functional capacity of patients with claudication, but the optimal application of treadmill testing in the setting of multicenter clinical trials remains uncertain. The current study used data from a recent clinical trial of the drug NM-702, which employed three baseline assessments of peak walking time (PWT) using a graded treadmill. These data were used to describe the different methods of defining the baseline peak treadmill performance with respect to reproducibility, stability over time and detection of treatment effect. A series of baseline definitions (first test only, last test only, highest PWT of the three tests, arithmetic mean of the three tests, mean of the first two tests, median of the three tests and a reproducibility-based criterion) were used to calculate the population ($n = 386$) variability in baseline testing, the placebo response over the 24 weeks of treatment, and the effect size of NM-702. Placebo responses and NM-702 effect sizes were not substantively affected by the method used to calculate baseline PWT. Changes in PWT on placebo were less than 25% for all methods of baseline quantitation. No method yielded an NM-702 effect size quantitatively greater than that obtained using only the first baseline test in the analysis for either PWT or claudication onset time. The graded treadmill test quantifies PWT with high reproducibility and stability over time. These characteristics may obviate the need for multiple treadmill tests, potentially saving study costs and improving patient acceptance of trial participation.

Key words: claudication; clinical trials; exercise testing; peripheral arterial disease

Introduction

Exercise impairment due to peripheral arterial disease (PAD) and its associated symptom of claudication remain major causes of disability in patients with cardiovascular disease.[1,2] This has motivated the development of new therapeutic modalities to relieve symptoms and improve walking ability. Change in exercise performance as measured using treadmill-based walking protocols provides a valuable and objective tool for defining the efficacy of therapeutic interventions.[3] However, use of treadmill-based performance measures is challenging for many frail patients with PAD and its application in multicenter trials is problematic. Numerous site methodological issues in performing treadmill tests have been identified.[4] This is particularly problematic since changes in peak exercise performance often serve as the primary endpoint in studies designed to support new drug approvals. Therefore, optimization of treadmill testing is critical to allow the detection and quantitation of performance changes associated with a potential new therapy.

One major limitation with treadmill-based assessments is an apparent time-dependent improvement in performance independent of the intervention.[3] This phenomenon is manifest as a large increase in treadmill exercise time in the placebo arm of trials over 6 months or less, despite no changes in background therapy. Additionally, the change in exercise time over the course of the study in the placebo cohort frequently has a large variance. These characteristics of the placebo group increase the number of patients required to define treatment effects, particularly for interventions with a small effect size.

The observed improvement in the placebo arm of trials is contrary to expectations based on the natural history of the disease which is characterized by clinical stability or decrements in performance.[5,6] This suggests that the observed increases in treadmill walking

[a]Department of Medicine, Harbor-UCLA Medical Center, Torrance, CA, USA; [b]Catalyst Pharmaceutical Research, Inc., Pasadena, CA, USA; [c]University of Colorado, Divisions of Geriatrics and Cardiology, and the Colorado Prevention Center, Denver, CO, USA

Address for correspondence: Eric P Brass, Harbor-UCLA Center for Clinical Pharmacology, 1124 W Carson Street, Torrance, CA 90502, USA. Tel: +1 310 222 4050; Fax: +1 310 533 0627; E-mail: ebrass@ucla.edu

Dr Mary McGrae McDermott was the Guest Editor for this manuscript.

© 2007 SAGE Publications

10.1177/1358863x07078602

performance during trials may reflect an artifact of the assessment modality, which if eliminated would improve the utility of treadmill testing in assessing intervention-associated changes in performance. Several factors may contribute to an otherwise stable patient improving his/her performance on a treadmill. Patients may require multiple treadmill test experiences before they become familiarized and comfortable walking on the device, and thus their gait may adapt to walking on a treadmill which would manifest as a progressive increase in walking time. Additionally, as patients are frequently asked to walk to maximal claudication pain, their sense of security with continued effort despite pain, and other motivating factors, may also change after several experiences.

The current study was undertaken to characterize and describe the optimal methods for defining the baseline peak walking time (PWT) and whether these methods would enhance the stability of performance in patients treated with placebo and improve the ability to detect the impact of an efficacious intervention.

Methods

Overview of study design

The current analyses were performed on data collected during a clinical trial to assess the safety and efficacy of the phosphodiesterase inhibitor NM-702 in patients with claudication. The details of this trial have previously been published.[7] Briefly, the study was a three armed, placebo-controlled, double-blinded, multicenter randomized trial with a 24-week treatment period. Patients with documented PAD and symptomatic claudication were eligible for enrollment. Claudication-limited exercise performance was quantified using a graded treadmill exercise test, and change in PWT was the primary outcome measure. Prior to baseline assessments, all participants underwent a familiarization session with the treadmill, during which they were taught how to start the test, experienced the approximate speed and grade of the formal testing, and had an opportunity to practice walking under test conditions. At time zero the patients stepped onto the treadmill which was moving at 2 mph (3.2 km/h) at a 0% grade. At 2-minute intervals the grade was increased by 2% as per Gardner et al.[8] Claudication onset time (COT) was defined as the time after initiation of exercise when the patient first experienced symptoms of claudication. The patient continued on the treadmill until he/she could walk no further due to severe claudication pain, and the time at which this occurred was designated as the PWT. Patients underwent three baseline graded treadmill tests to determine PWT. Each baseline test was conducted on a separate day, with each test separated by at least three, but no more than 10 days. All baseline tests were completed during a maximum 42-day screening period. Sites were given specific training on how to properly conduct a treadmill test, and the quality of the testing was monitored during the conduct of the trial (site quality assurance by Colorado Prevention Center, Denver, CO, USA). Patients were eligible for randomization if their median PWT was greater than or equal to 90 seconds, and less than or equal to 600 seconds.

The protocol-specified method for defining the baseline PWT was the median of the three baseline tests. Patients with a baseline PWT greater than or equal to 90 seconds but less than or equal to 600 seconds were eligible for randomization to either placebo, 4 mg NM-702 or 8 mg NM-702, each given twice daily. Follow-up treadmill testing was performed 12 and 24 weeks post-randomization (single assessment at each time point). The randomized, intention-to-treat population formed the basis for the current analyses.

Analyses

All study personnel were aware that for the protocol-specified analyses PWT would be defined as the median of the three baseline tests. Only the authors, none of whom were site investigators, knew of the intent to probe the baseline performance definition based on other approaches.

To evaluate the reproducibility and retest correlations of the treadmill assessment of PWT, baseline PWT data were examined by scatter plot and intra-subject coefficient of variation (CV) over repeated measures. Participants were separated into quartiles based on their baseline median PWT and the intra-subject CV assessed in each quartile to determine whether the baseline PWT affected variability.

A series of alternative definitions for baseline PWT were developed based on those used in other trials and their potential utility (see Table 1). Use of the first treadmill PWT was evaluated because of its potential to simplify trial conduct. Use of the third treadmill test only was evaluated as this might represent the best estimate after the previous tests had served to acclimatize the patient. Use of the highest PWT obtained during the three tests was evaluated as it would provide the best estimate of the patient's true pathophysiologic limitation. The mean of the three tests was assessed as the best way to use all of the obtained data to define the point estimate of PWT. The median of the three tests effectively removes extremes (high or low) from an individual's performance estimate under the assumption that the extremes are not reflective of actual patient limitations. Many trials impose a reproducibility criterion to improve the assessment of performance. To simulate the impact of this had it been employed in the original trial, a reproducibility measure was developed that used the mean of the first two tests if the difference between them was less than 25% of the mean of the first and second tests. If this criterion failed, the same standard was applied to the second and third tests. If neither test replicated, then the patient

Table 1 Definitions and methods used to calculate baseline treadmill peak walking time (PWT).

Method	Definition	Rationale
First only	PWT of the first baseline treadmill test	Only one test required if test highly reproducible
Last only	PWT of the third baseline treadmill test	Most accurate if learning effect substantial over three tests
Highest	Highest PWT of the three baseline treadmill tests	Most reflective of true pathophysiologic limitation
Mean of three	Mean PWT of the three baseline treadmill tests	Utilizes data from all three tests as the best point estimate
Median of three	Median PWT of the three baseline treadmill tests	Removes the influence of outlier test
Mean of two	Mean PWT of the first and second baseline treadmill tests	Removes requirement of third test if test reproducible
Reproducible	If ((treadmill 1 – treadmill 2)/((treadmill 1 + treadmill 2)/2) <0.25 then mean of the first and second baseline treadmills; if not and ((treadmill 2 – treadmill 3)/((treadmill 2 + treadmill 3)/2) <0.25 then mean of the second and third baseline treadmills;if neither, then patient is not reproducible and excluded from analysis	Decreases impact of a bad test and excludes patients who can not perform reproducible test

was considered to have screen failed and the patient was excluded from the analyses based on this baseline definition, as would be done in trials utilizing this approach. This criterion was applied only for those analyses employing the simulated reproducibility baseline definition.

Intra-subject coefficient of variation was calculated for each individual as the standard deviation of the three baseline tests divided by the mean of the three tests presented on a percent scale. The intra-subject coefficient of variation for the population at baseline was summarized by the mean and range for all individuals. The coefficient of variation (expressed as a percentage) for the change in treadmill performance over time in the placebo group was calculated as the standard deviation of the change in the group divided by the mean change in the group, multiplied by 100.

Each baseline definition was used to calculate trial outcomes. First, the placebo group was analyzed for stability based on each baseline PWT definition, with stability defined as the change in PWT from baseline to 6 months. The intent-to-treat population was utilized with last observation carried forward to replace missing data, as was done for the original trial's primary analyses. The mean and standard deviation of change in PWT over the 24-week study was calculated.

Additional analyses were performed to define the influence of the baseline method utilized on the treatment effect of NM-702 when compared with placebo. For these analyses each baseline definition was used to calculate the change over the 24-week treatment period for each patient, with last observation carried forward as previously detailed.[7] The change in performance was calculated as either the percent change from baseline, for ease of interpretation, or the natural logarithm (ln) ratio of the 24-week and baseline treadmill PWTs to correspond with the statistical methods employed for group comparisons. Logarithmic transformation was used to minimize the impact of extreme values and to meet the statistical assumptions for formal hypothesis testing, as pre-specified in the original study design. Effect sizes as defined by Cohen[9] were used for comparing the treatment effect based upon each baseline method. The effect size and corresponding *p*-values of NM-702 versus placebo utilizing each baseline definition were calculated using the variance estimates for population (pooled variances) from the ANCOVA model on ln-PWT ratios. Variables included in the model were treatment, smoking status and baseline ln-PWT as previously described.[7] These analyses were repeated for COT.

As no prospectively defined hypothesis testing was done, all statistical analyses should be considered descriptive.

The authors had full access to the data and take responsibility for its integrity. All authors have read and agree to the manuscript as written.

Results

Baseline PWT was calculated using each of the proposed definitions (Table 1) for 386 patients randomized and included in the clinical trial's intention-to-treat population (Table 2). All definitions yielded very similar estimates for the population's baseline performance and distribution, except for the use of the highest of three baseline tests, which was 10–15% higher than other methods.

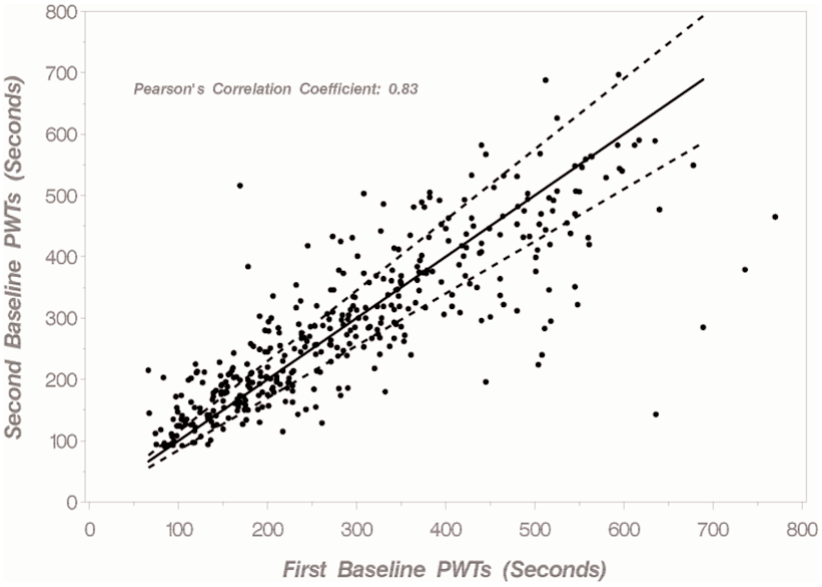
Systematic bias in the performance of the three baseline tests was explored using the results of each participant's first baseline test plotted against his/her second (Figure 1A) or third baseline test (Figure 1B). Most individuals demonstrated excellent reproducibility.

Table 2 Effect of baseline peak walking time (PWT) definition on stability of PWT during placebo treatment for 24 weeks. (See Table 1 for details of baseline PWT definitions.)

Baseline PWT definition	Baseline PWT – all patients	Baseline PWT – placebo group	Percent change in PWT over 24 weeks – placebo group	Coefficient of variation for percent change in PWT – placebo group
First only	292 ± 146	277 ± 140	18.5 ± 51.6	280%
Last only	283 ± 131	266 ± 130	23.2 ± 64.4	280%
Highest	329 ± 148	312 ± 146	1.5 ± 39.5	2700%
Mean of three	288 ± 128	272 ± 126	15.9 ± 46.8	290%
Median of three	287 ± 129	272 ± 127	17.1 ± 49.0	290%
Mean of two	286 ± 127	270 ± 126	17.8 ± 51.7	290%
Reproducibility	290 ± 130	275 ± 128	14.6 ± 43.2	300%

Values are mean ± SD and are expressed in seconds.
$n = 386$ for all patients and $n = 130$ for the placebo only group; except for the reproducibility criterion where $n = 358$ for all patients and $n = 117$ for the placebo only group.

A



B

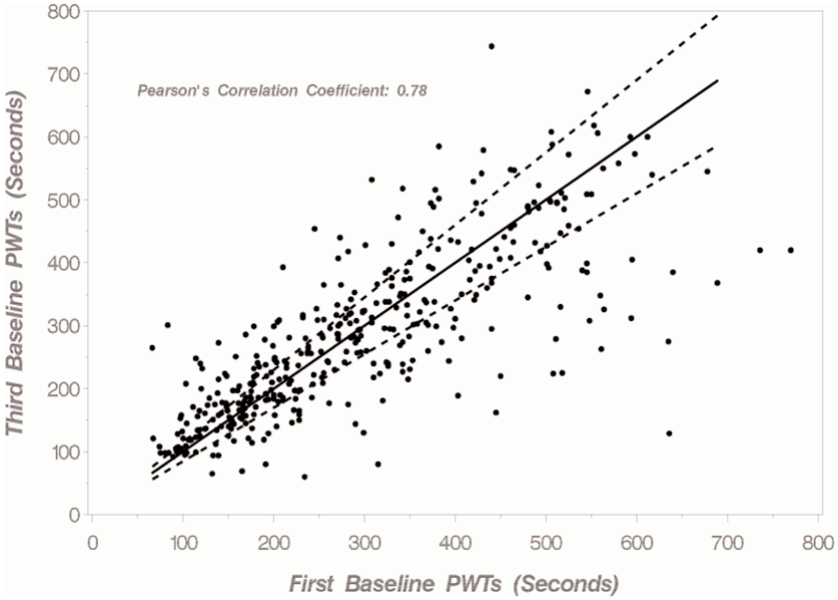


Figure 1 Relationship between peak walking time on first and second (A), and first and third (B) baseline treadmills. Each point represents an individual patient. The line of identity is shown, and lines representing 15% above and 15% below the identity line are also added.

Some had substantially longer PWT's on the first test as compared with the second and third test, but only rarely did an individual perform dramatically better on the second or third tests. The mean change in PWT from the first to second baseline test was –3 seconds (95% CI of –12 to +5 seconds), and between the first and third tests was –9 seconds (95% CI of –18 to +1 seconds). The intra-subject coefficient of variation for PWT at baseline averaged 15%, with a range amongst participants of 1% to 95%.

The effect of baseline PWT on the intra-subject coefficient of variation based on the three baseline tests was explored by dividing the cohort into quartiles based on their median baseline PWT (lowest quartile: $n = 97$, PWT 90–177 seconds; second quartile: $n = 97$, PWT 177–270 seconds; third quartile: $n = 96$, PWT 270–376 seconds; fourth quartile: $n = 96$, PWT 376–600 seconds). The intra-subject coefficient of variation based on the three baseline tests was independent of median baseline PWT (16%, 17% 14% and 14% in the lowest through fourth quartiles respectively).

Focusing on the patients randomized to placebo, the baseline method used had little impact on the assessed population mean change in treadmill performance over the 24-week study period, whether expressed as the percent change or the ln ratio of the week 24 and baseline PWTs (Tables 2 and 3). The exception was the use of the highest of the three baseline assessments, which, as might be anticipated, yielded a much smaller percent change in performance over the 24 weeks. Numerically, those definitions which employed more than one of the three baseline treadmills to estimate baseline performance resulted in smaller changes over the 24 weeks, but these differences were small. The group coefficient of variation was very similar using all of the definitions except for the highest-of-three method. In all cases the placebo-associated change in PWT over 24 weeks was less than 25% (Table 2).

The most important attribute of an optimized baseline assessment method will be its ability to detect the effect of an intervention as compared with placebo. This will be dependent on the influence of the baseline assessment on the relative magnitude of the change in treadmill performance over the study period in the placebo and treatment arms, and the population variability in these responses. Using data from the NM-702 trial, all baseline definitions yielded similar estimated magnitudes of increased PWT in response to NM-702, except the use of the highest of the three baseline tests which demonstrated a smaller net increase in PWT from baseline to 24 weeks (Table 3). The effect size for NM-702 was calculated using the Cohen *d* statistic based on data for each baseline definition. Using this method an effect size of 0.2 is considered small, 0.5 is considered medium, and 0.8 large.[9] The baseline assessment method had little impact on the NM-702 effect size and thus on the trial's ability to discern drug effect. Of note, use of the highest of the three determinations as the baseline yielded effect sizes similar to the other methods despite its differential assessment of absolute change over the 24-week study. Calculation of *p*-values using ANCOVA for each baseline definition yielded values between 0.198 and 0.304 for the 4 mg NM-702 arm vs placebo, and between 0.002 and 0.012 for the 8 mg NM-702 vs placebo. No method of baseline assessment was superior to the use of the single first baseline treadmill performed. Analysis of COT supported the conclusion that no definition was superior to the use of the first baseline treadmill assessment (Table 4).

Discussion

Optimizing the baseline assessment of treadmill walking time in clinical trials of claudication therapies is critical to minimizing responses in the placebo-treated

Table 3 Effect of baseline peak walking time (PWT) definition on the ability to detect response to NM-702. (See Table 1 for details of baseline PWT definitions.)

PWT baseline definition	Placebo: ln ratio (24 weeks / baseline)	4 mg NM-702 – percent change in PWT	4 mg NM-702 – ln ratio (24 weeks / baseline)	4 mg NM-702 – effect size vs placebo	8 mg NM-702 – percent change in PWT	8 mg NM-702 – ln ratio (24 weeks / baseline)	8 mg NM-702 – effect size vs placebo
First only	0.086	25.4%	0.130	0.161	32.6%	0.196	0.386
Last only	0.109	26.6%	0.144	0.129	28.9%	0.195	0.314
Highest	–0.059	6.4%	–0.022	0.129	12.7%	0.059	0.380
Mean of three	0.076	21.6%	0.116	0.143	27.2%	0.180	0.350
Median of three	0.079	22.1%	0.122	0.144	28.1%	0.185	0.352
Mean of two	0.080	23.3%	0.126	0.146	27.1%	0.181	0.332
Reproducibility	0.065	22.2%	0.115	0.158	26.6%	0.174	0.358

Change in PWT for the NM-702 treatment arms are expressed as mean change above baseline and the mean ln ratio of the week 24 measurement and the baseline assessment. Values are unitless, and are expressed as the mean. Effect size was calculated using the method of Cohen[9] as the standardized mean difference for the sample. An effect size of 0.2 is considered small, 0.5 is considered medium, and 0.8 large.[9] See Table 2 for the percentage change in the placebo group over the 24-week study period.

Table 4 Effect of baseline COT definition on the ability to detect response to NM-702.

COT baseline definition	Percent change in COT over 24 weeks – placebo	Placebo: ln ratio (24 weeks / baseline)	4 mg NM-702 – percent change in COT	4 mg NM-702 – ln ratio (24 weeks / baseline)	4 mg NM-702 – effect size vs placebo	8 mg NM-702 – percent change in COT	8 mg NM-702 – ln ratio (24 weeks / baseline)	8 mg NM-702 – effect size vs placebo
First only	41.6%	0.183	67.7%	0.335	0.317	73.1%	0.382	0.441
Last only	29.5%	0.114	52.7%	0.216	0.250	48.0%	0.247	0.355
Highest	5.7%	–0.071	17.9%	0.022	0.222	19.0%	0.053	0.316
Mean of three	27.2%	0.115	45.8%	0.228	0.258	46.3%	0.266	0.365
Median of three	30.0%	0.128	48.6%	0.234	0.251	49.1%	0.286	0.378
Mean of two	24.4%	0.097	47.4%	0.210	0.259	43.0%	0.236	0.347
Reproducibility	31.6%	0.125	42.1%	0.202	0.206	47.2%	0.262	0.311

Baseline COT definitions were used that were identical to those used for PWT as defined in Table 1. Change in COT for the NM-702 treatment arms are expressed as mean change above baseline and the mean ln ratio of the week 24 measurement and the baseline assessment. Values are unitless, and are expressed as the mean. Effect size was calculated using the method of Cohen[9] as the standardized mean difference for the sample. An effect size of 0.2 is considered small, 0.5 is considered medium, and 0.8 large.[9] *P*-values for the effect of 4 mg NM-702 were all less than 0.05 except 0.07 for the use of the highest baseline test and 0.107 for the reproducibility criterion. *P*-values for the effect of 8 mg NM-702 were all less than 0.02.

group and to detecting beneficial effects of interventions. Trials performed to date have used a variety of methods to assess baseline performance and to generate a point estimate of baseline performance based on these tests. The current analyses suggest that all of these methods are likely equivalent when using a graded treadmill protocol and that therefore a single, well-conducted treadmill test may be adequate for determining treadmill performance in this population. The blinding of sites to the planned analysis of the three baseline tests ensured that bias in the conduct of these tests was minimized and increases the confidence in the results presented.

The use of a graded treadmill test is intended to raise the work rate progressively until an inherent physiologic (or pathophysiologic) limitation is reached. In the case of claudication, this may reflect the exercise limitation associated with a mismatch between oxygen delivery to the muscle and the increasing energy demands of the muscle. To the degree that the test is able to assess this intrinsic limitation, reproducibility of the test and stability over time would be anticipated. This is confirmed in the current analyses, as strategies to improve the point estimate (mean of two or three tests), to minimize the influence of outliers (the median of three tests), to ensure the attainment of a true maximal test (the use of the highest of the three tests), anticipating a learning effect (use of the third test only) or simulating a reproducibility criterion, yielded similar estimates of the placebo-group response over 24 weeks, and equivalent ability to detect the efficacy of NM-702. In fact, no combination of baseline tests appeared to be superior to the simple use of the first test conducted. This conclusion was supported by analyses using both PWT and COT as endpoints. The excellent reproducibility of the graded treadmill test for quantifying PWT at baseline was confirmed across the full range of performance levels studied (PWTs between 90 and 600 seconds). Importantly, as only graded treadmill testing was used, it is not clear whether these conclusions can be extrapolated to testing with a fixed work rate treadmill.

A definition of baseline PWT that incorporates a reproducibility criterion is conceptually attractive. Demonstrating reproducibility has the potential to increase confidence in the point estimate of baseline performance and identify patients more likely to have stable performance over the course of the trial in the absence of intervention. For these reasons, many trials incorporate such criteria, but large placebo responses are still observed (for example, the 46% increase in the placebo arm of Brevetti et al[10]). The reasons for this are unclear, but such criteria may introduce bias as the investigator is aware of the need to achieve the reproducibility criterion and has an incentive to have the replicate test match the preceding test. The current analysis could not formally challenge this possibility as the investigators were not attempting to meet a reproducibility criterion when the data were collected. Nonetheless, the current analyses simulating the effect of a reproducibility criterion suggest that such criteria are not superior to other definitions which do not risk the introduction of such bias.

The current analysis of baseline treadmill performance addresses the question of the value of repeat testing to optimally define PWT or COT. However, the same questions can be raised with respect to the determination of these parameters at the end of a study. While duplicate exit treadmills are sometimes performed, this is done less frequently than repetition at baseline under the assumption that the measurement instability is greatest during the patient's initial testing. This assumption is untested, and the current analyses suggest little, if any, potential benefit for duplicate assessments of PWT or COT at the end of the treatment period based on the excellent reproducibility at baseline.

Application of treadmill testing to multi-site clinical trials is challenging as sites may not be able to conduct the treadmill protocol in a manner that ensures optimal test characteristics.[4] Importantly, the current analyses can only be generalized in the context of the overall elements of the original trial. For example, the study incorporated an extensive site quality assurance program to improve treadmill testing. This program included a review of site procedures, remediation efforts directed at problems identified, on-line continuing education tools to review key aspects of test conduct, follow-up assessments of site test conduct and monitoring of site performance by the study's steering committee. Sites were also taught to familiarize patients with the treadmill and procedures during the screening visit prior to the baseline assessments. Thus, the first formal treadmill test was not the participant's first encounter with the treadmill. Patients who could not walk 2.0 mph (3.2 km/h) on the treadmill or who were otherwise uncomfortable with the testing procedure were excluded prior to the first treadmill assessment of peak walking time. The trial's incorporation of an intensive quality assurance program likely contributed to the utility of the single, first baseline treadmill test. The findings from this study should be evaluated in the context of other claudication clinical trials to ensure the robust nature of the findings.

The current analyses, if confirmed in future trials, may have significant implications for future trials of claudication therapies. Optimizing all aspects of the treadmill assessment procedures is critical for minimizing variability and responses in the placebo group, and thus allowing for detection of potential therapeutic signals. This optimization will allow interventions with small effect sizes to be assessed, or require fewer study participants when the intervention is associated with a larger effect size. The use of a single baseline treadmill test has the potential to decrease study costs and increase patient (and investigator) acceptability of the study by decreasing the number of visits a study requires. It can be estimated that removing two treadmill tests and associated visits from a trial could decrease direct site costs by 10–15% per patient without sacrificing the robustness of the study.

While the current analyses are based on a study in patients with claudication, treadmill or ergometer-based assessments are used in other chronic disease populations, including chronic obstructive pulmonary disease, pulmonary hypertension and congestive heart failure. The degree to which the current conclusions can be extrapolated to these other diseases is unknown, but the implications for trial design merit similar considerations for testing in these populations.

Acknowledgements

The authors thank Nissan Chemical and Catalyst Pharmaceutical Research for access to the NM-702 study database for the purpose of the current analyses.

Funding sources

The original clinical trial from which the current data was obtained was funded by Nissan Chemical, which also provided partial support for the analyses in the current manuscript.

Disclosures

EPB and WRH received financial support from Nissan Chemical and Catalyst Pharmaceutical Research to serve on the study's Steering Committee. JJ is an employee of Catalyst Pharmaceutical Research.

References

1 Hiatt WR. Medical treatment of peripheral arterial disease and claudication. *N Engl J Med* 2001; **344**: 1608–21.
2 McDermott MM, Mehta S, Liu K et al. Leg symptoms, the ankle–brachial index, and walking ability in patients with peripheral arterial disease. *J Gen Intern Med* 1999; **14**: 173–81.
3 Hiatt WR, Hirsch AT, Regensteiner JG, Brass EP. Clinical trials for claudication. Assessment of exercise performance, functional status, and clinical end points. *Circulation* 1995; **92**: 614–21.
4 Hiatt WR, Cox L, Greenwalt M, Griffin A, Schechter C. Quality of the assessment of primary and secondary endpoints in claudication and critical leg ischemia trials. *Vasc Med* 2005; **10**: 207–13.
5 McDermott MM, Liu K, Greenland P et al. Functional decline in peripheral arterial disease: associations with the ankle brachial index and leg symptoms. *JAMA* 2004; **292**: 453–61.
6 Imparato AM, Kim GE, Davidson T, Crowley JG. Intermittent claudication: its natural course. *Surgery* 1975; **78**: 795–99.
7 Brass EP, Anthony R, Cobb FR, Koda I, Jiao J, Hiatt WR. The novel phosphodiesterase inhibitor NM-702 improves claudication-limited exercise performance in patients with peripheral arterial disease. *J Am Coll Cardiol* 2006; **48**: 2539–45.
8 Gardner AW, Skinner JS, Cantwell BW, Smith LK. Progressive vs single-stage treadmill tests for evaluation of claudication. *Med Sci Sports Exerc* 1991; **23**: 402–08.
9 Cohen J. Statistical power analysis for the behavioral sciences. Hillsdale, NJ: Lawrence Erlbaum Associates; 1988.
10 Brevetti G, Diehm C, Lambert D. European multicenter study on propionyl-L-carnitine in intermittent claudication. *J Am Coll Cardiol* 1999; **34**: 1618–24.