Designing Clinical Trials of Interventions for Mobility Disability: Results From the Lifestyle Interventions and Independence for Elders Pilot (LIFE-P) Trial

Mark A. Espeland,1 Thomas M. Gill,2 Jack Guralnik,3 Michael E. Miller,1 Roger Fielding,4 Anne B. Newman,5 and Marco Pahor,6 for the Lifestyle Interventions and Independence for Elders Study Group

1Department of Biostatistical Sciences, Wake Forest University School of Medicine, Winston-Salem, North Carolina.
2Department of Medicine, Yale University School of Medicine, New Haven, Connecticut.
3Epidemiology and Demography Section, National Institute on Aging, Bethesda, Maryland.
4Nutrition, Exercise, Physiology, and Sarcopenia Laboratory, Jean Mayer USDA Human Nutrition Research Center on Aging, Boston, Massachusetts.
5Departments of Epidemiology and Medicine, Graduate School of Public Health, University of Pittsburgh, Pennsylvania.
6Department of Aging and Geriatric Research, University of Florida, Gainesville.

Background. Clinical trials to assess interventions for mobility disability are critically needed; however, data for efficiently designing such trials are lacking.

Methods. Results are described from a pilot clinical trial in which 424 volunteers aged 70–89 years were randomly assigned to one of two interventions—physical activity or a healthy aging education program—and followed for a planned minimum of 12 months. We evaluated the longitudinal distributions of four standardized outcomes to contrast how they may serve as primary outcomes of future clinical trials: ability to walk 400 meters, ability to walk 4 meters in ≤10 seconds, a physical performance battery, and a questionnaire focused on physical function.

Results. Changes in all four outcomes were interrelated over time. The ability to walk 400 meters as a dichotomous outcome provided the smallest sample size projections (i.e., appeared to be the most efficient outcome). It loaded most heavily on the underlying latent variable in structural equation modeling with a weight of 80%. A 4-year trial based on the outcome of the 400-meter walk is projected to require \( N = 962–2234 \) to detect an intervention effect of 30%–20% with 90% power.

Conclusions. Future clinical trials of interventions designed to influence mobility disability may have greater efficiency if they adopt the ability to complete a 400-meter walk as their primary outcome.

As the life expectancy of older Americans increases, prevention of age-associated physical function decline and disabilities has emerged as a major clinical and public health priority (1). A critical factor in an older person’s ability to function independently is mobility, the ability to move without assistance (2). Older people who lose mobility are less likely to remain in the community, have higher rates of morbidity and mortality, have more hospitalizations, and experience a poorer quality of life (3).

Clinical trials are necessary to establish interventions for improving mobility. For major investments to be made, it is important that these trials are designed efficiently, which includes informed decisions on which outcomes have the best measurement characteristics, what effect sizes should be targeted, and how many participants are required. We describe how data collected by the Lifestyle Interventions and Independence for Elders Pilot (LIFE-P) trial have been used to address these design issues.

Methods. The study design, eligibility criteria, and recruitment procedures of LIFE-P were described previously (4). Participants were aged 70–89 years and able to complete a 400-meter walk (W400) in 15 minutes. Major exclusion criteria included presence of severe heart failure, uncontrolled angina, and other severe illnesses that might interfere with physical activity. All participants completed a 1-week behavioral assessment prior to random assignment, with equal probability, to either a physical activity intervention group or a health education control group. Written informed consent was obtained; the National Institutes of Health (NIH) and Institutional Review Boards for all participating institutions approved the protocol and consent forms. Between May 2004 and February 2005, 424 participants were enrolled. At baseline, 6 months, and 12 months, comprehensive standard assessments were conducted by trained research staff who were masked to
intervention assignment. W400 was assessed using a standard walking course. Participants were permitted to stop during the walk, were not allowed to sit or receive help from others (cane use was permitted during follow-up assessments), and were required to complete the course in 15 minutes. Central adjudication was used to classify individuals who did not complete the walk (4). The ability to walk 4 meters in ≤ 10 seconds (W4), corresponding to a gait speed ≥ 0.4 m/s, was assessed by asking the participants to walk at their usual pace (5). The Short Physical Performance Battery (SPPB) is a brief performance-based test that includes W4, repeated chair stands, and a balance test (6, 7). Its three components are each scored 0–4, with 4 indicating the highest level of performance, and were summed to yield an overall score. To be eligible for LIFE-P, participants must have had an SPPB score ≥ 9 at baseline. The Self-Report Disability Scale (SRDS) has been used in clinical trials of LIFESTYLE Interventions and Independence for Elders Pilot trial; SD = standard deviation.

Statistical Methods
For the dichotomous measures (W400, W4), we described rates and transition probabilities over time using log-linear models (13). For the measures analyzed as continuous data (SPPB, SRDS), we examined mean changes over 6-month time intervals. To project the required sample sizes for continuous measures, we simulated data that had the same variances and longitudinal correlations observed in LIFE-P, but longer (i.e., 4 years) follow-up. For the control group, the mean changes over 6-month intervals were set to equal those observed in LIFE-P, with those from 6–12 months repeated to extend throughout 4 years. To introduce 20%–30% intervention effects, we increased any beneficial mean changes and decreased any nonbeneficial mean changes by these amounts to simulate data for the intervention group. In a parallel manner, we projected sample sizes for categorical measures by simulating data that had transition rates based on LIFE-P data. For the control group, the 6-month transition rates were set to equal those observed in LIFE-P, with those from 6 to 12 months repeated to extend throughout 4 years. To introduce 20%–30% intervention effects, we increased the rates of beneficial transitions and decreased the rates of nonbeneficial transitions by these amounts to simulate data for the intervention group. Missing data were randomly introduced at accumulating rates of 4% per 6 months (corresponding to what LIFE-P observed) in both groups. We used changes from baseline over follow-up (in repeated-measures models) as the outcome for continuous data. For dichotomous measures, we examined two potential outcomes: times until the first failure during follow-up and the occurrence of two successive failures. Data were simulated (100,000 sequences) and analyzed with general linear models or survival analyses, depending on the outcome, to project power.

We fitted a structural equation to examine how the four outcomes tracked over time against an underlying construct or “commonality,” and fitted this model using Gibbs sampling (14, 15) (see Appendix). Standardized weights estimated from this model express how strongly changes in each outcome were related to changes in the underlying commonality.

We examined whether changes in outcomes were related to changes in health-related quality of life, a factor expected to be associated with mobility disability. To do this, we examined the correlations that 1-year changes in QWBS had with 1-year changes in SPPB and SRDS, after covariate adjustment for intervention assignment. We also examined mean changes in QWBS for participants who did and did not complete the W400 and W4 during follow-up.

RESULTS
Table 1 provides a description of the 424 LIFE-P participants at enrollment. The mean (standard deviation) SPPB and SRDS scores were 7.52 (1.41) and 1.36 (0.36), respectively. All participants successfully completed the W400 as part of eligibility criteria, with average walk speeds of 0.85 m/s (0.18 m/s); all but 9 (2.1%) participants completed the W4 in ≤ 10 seconds, with average walk speeds of 0.74 m/s (0.16 m/s). Mean QWBS score was 0.64
Successful aging and physical activity interventions. During the first year of follow-up, 20 of the 213 participants assigned to the physical activity intervention and 24 of the 211 assigned to the successful aging intervention were unable to complete the 400 m walk at the 6-month and/or 12-month assessment. Of participants who were successful at completing a 4 m walk at baseline, 11 of the 206 who were assigned to the physical activity intervention and 13 of the 202 assigned to successful aging were unable to complete the 4 m walk at the 6-month and/or 12-month examination.

Table 2 describes estimated rates (standard errors) at 6 months and 12 months for the SPPB and SRDS. Data-collection rates at 6 months were 94.8% (W400), 93.9% (W4), 94.6% (SPPB), and 96.2% (SRDS). At 12 months, these rates were 92.2% (W400), 93.9% (W4), 93.9% (SPPB), and 94.6% (SRDS). One participant in each intervention condition died during each 6-month time period (i.e., four total deaths).

Table 2. Transition Rates and Standard Errors for 400-Meter Walk and 4-Meter Walk Outcomes

<table>
<thead>
<tr>
<th>Successful Aging</th>
<th>Between 0 and 6 Months</th>
<th>Between 6 and 12 Months</th>
</tr>
</thead>
<tbody>
<tr>
<td>400 m walk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Success to failure</td>
<td>6.9% (1.9%)</td>
<td>7.9 (2.1%)</td>
</tr>
<tr>
<td>Failure to success</td>
<td>—</td>
<td>33.3% (13.4%)</td>
</tr>
<tr>
<td>4 m walk</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Success to failure</td>
<td>4.7% (1.5%)</td>
<td>2.3% (1.2%)</td>
</tr>
<tr>
<td>Failure to success</td>
<td>—</td>
<td>66.7% (16.0%)</td>
</tr>
</tbody>
</table>

| Physical Activity |                        |                         |
|-------------------|                        |                         |
| 400 m walk        |                        |                         |
| Success to failure | 7.8% (1.9%)          | 3.6% (1.5%)             |
| Failure to success | —                     | 50.0% (17.4%)          |
| 4 m walk          |                        |                         |
| Success to failure | 2.6% (1.1%)          | 3.3% (1.3%)             |
| Failure to success | —                     | 40.0% (22.0%)          |

Notes: Estimates from log-linear models were applied to participants enrolled in the successful aging and physical activity interventions. During the first year of follow-up, 20 of the 213 participants assigned to the physical activity intervention and 24 of the 211 assigned to the successful aging intervention were unable to complete the 400 m walk at the 6-month and/or 12-month assessment. Of participants who were successful at completing a 4 m walk at baseline, 11 of the 206 who were assigned to the physical activity intervention and 13 of the 202 assigned to successful aging were unable to complete the 4 m walk at the 6-month and/or 12-month examination.

SE = standard error.

(0.10). Data-collection rates at 6 months were 94.8% (W400), 93.9% (W4), 94.6% (SPPB), and 96.2% (SRDS). At 12 months, these rates were 92.2% (W400), 93.9% (W4), 93.9% (SPPB), and 94.6% (SRDS). One participant in each intervention condition died during each 6-month time period (i.e., four total deaths).

Table 2 describes estimated rates (standard errors) at 6 months and 12 months for the SPPB and SRDS. Data-collection rates at 6 months were 94.8% (W400), 93.9% (W4), 94.6% (SPPB), and 96.2% (SRDS). At 12 months, these rates were 92.2% (W400), 93.9% (W4), 93.9% (SPPB), and 94.6% (SRDS). One participant in each intervention condition died during each 6-month time period (i.e., four total deaths).

Table 3 lists the numbers of participants required to detect intervention effects of 20%, 25%, and 30%, respectively, for each outcome. We project that 1414 participants would be required for 90% statistical power to detect a 25% effect size for the outcome of time until the first failure of the W400; 2839 participants would be required for 90% power to detect this effect if the outcome is time until two successive failures. The projected sample sizes for the other outcomes were uniformly larger.

The same (percent) effects sizes are used for each of the outcome measures in Table 4. It is likely, however, that these measures vary in their fidelity to mobility disability. Some may more closely represent its changes than others, so similar effect sizes may correspond to different changes in mobility disability. Table 5 provides results from an analysis aimed at describing how changes in the outcomes portray underlying changes. The fitted weights indicate the relative strength of the relationship between changes in each outcome and changes in the underlying commonality (reported in standard deviation units). Weights from each outcome are bounded away from zero (changes in measures are inter-related), and relationships are in the expected directions. The W400 expressed changes in the underlying construct most directly: An intervention producing a 1 standard deviation change in this outcome was estimated to yield a 0.80 standard deviation change in the commonality (e.g., “mobility disability”). The SPPB, which is negatively correlated with the other outcomes due to how its scoring was ordered, was the second most direct outcome, with a 0.32 standard deviation change. The dichotomous W4 outcome was projected to receive a 0.22 unit change. The SRDS was relatively inefficient, receiving only a 0.13-unit change in the underlying construct.

QWBS scores showed little overall change from baseline to 1 year among participants assigned to both the physical activity and successful aging interventions. Mean (standard deviation) changes were 0.01 (0.10) and 0.01 (0.09), respectively. After covariate adjustment for intervention assignment, the correlations between 12-month changes in QWBS and 12-month changes in SPPB and SRDS were r = 0.13 (p = .01) and r = −0.31 (p < .001), respectively. Mean

Table 3. Estimated Mean Changes, Standard Errors (SEs), and Covariance Structure for the Short Physical Performance Battery (SPPB) and Self-Reported Disability Scale (SRDS) for Participants Assigned to the Two Intervention Programs

<table>
<thead>
<tr>
<th>Measure</th>
<th>0–6 Months</th>
<th>6–12 Months</th>
<th>0–6 Month Change σ₁²</th>
<th>6–12 Month Change σ₂²</th>
<th>Correlation of Changes ρ₁₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>Successful aging</td>
<td>N = 203</td>
<td>N = 195</td>
<td>0.589 (0.139)</td>
<td>−0.168 (0.115)</td>
<td>3.885</td>
</tr>
<tr>
<td>SPPB</td>
<td>0.021 (0.022)</td>
<td>0.017 (0.022)</td>
<td>0.102</td>
<td>0.097</td>
<td>−0.42</td>
</tr>
<tr>
<td>SRDS</td>
<td>N = 204</td>
<td>N = 177</td>
<td>1.241 (0.126)</td>
<td>−0.258 (0.115)</td>
<td>3.154</td>
</tr>
<tr>
<td>Physical activity</td>
<td>N = 204</td>
<td>N = 177</td>
<td>−0.027 (0.024)</td>
<td>0.024 (0.024)</td>
<td>0.119</td>
</tr>
</tbody>
</table>
We evaluated four distinct mobility-related outcomes. Three outcomes were objectively measured (W400, W4, and SPPB), whereas the fourth (SRDS) was based on self-report. The W400 and W4 assess mobility directly, with inability to complete the latter task representing the more severe form of mobility disability (19). Although not considered pure measures of mobility, the SPPB includes three elements (gait, transfers, and balance) that serve as the building blocks for mobility, and the SRDS includes several mobility items, including walking across a small room, walking one block, walking a quarter of a mile, walking several blocks, and climbing a flight of stairs. The SPPB and SRDS have served as primary outcome measures in previous intervention trials (20,21). The physical activity intervention in LIFE-P led to relative improvements in SPPB scores that were statistically significant (22). Increasing evidence supports the reliability and validity of the W400 as a measure of major mobility disability (21,23,24). In trials for which it is impossible to mask participants to intervention assignment, objective measures may be preferred to those based on subjective self-report (17).

LIFE-P data suggest that outcomes may vary in their ability to express underlying changes in mobility disability. Outcomes that most clearly express these changes are preferable, and targeted effects should be of sufficient magnitude to produce meaningful changes in the underlying construct. Although the structural equation model we fitted indicated that the W400 is most strongly intercorrelated with the other measures under consideration, the model certainly falls short of capturing the full complexity of interrelationships and our underlying construct may not correspond directly to mobility disability.

Sample Size Requirements
Sample size targets are based on assumptions regarding the magnitude of the intervention effect and the distribution of outcomes. A pilot trial does not have sufficient size to estimate accurately an intervention effect; this is the purpose of a full-scale trial. Although results presented in Tables 2 and 3 by intervention assignment are encouraging that the physical activity intervention may be beneficial, more definitive assessment is required. LIFE-P provided transition rates for categorical outcomes and means and covariances of changes for continuous outcomes, which we have used in our calculations.

If an intervention produces effects of 20%–30% in each of the four measures, our projections indicate that the W400 is the most efficient primary outcome. Fewer than half as many participants would be required as for the W4 or SPPB, and considerably fewer than for the SRDS. In general, categorizing continuous outcomes reduces power; however,
a categorical outcome may be a more efficient expression of an underlying continuous commonality than continuous outcomes that are less directly related.

Among participants who failed the W400 at 6 months, those in the physical activity group were more likely to successfully complete the test at 12 months. Nonetheless, our results indicate that nearly 50% more participants would be required for a primary outcome based on two successful failures to complete the W400. Thus, although the outcome of persistent disability may be important clinically, it would likely require much larger clinical trials.

Associations With Health-Related Quality of Life

We found that 1-year changes in SPPB, SRDS, and W400 were associated with relatively better changes in a commonly used index of health-related quality of life. No such association was found for changes in W4; our power to detect such an association, however, may have been low. To our knowledge, these four mobility-related outcomes have not been previously evaluated in a single study. Nonetheless, each has been previously linked to poorer quality of life and subsequent adverse outcomes, including loss of independence, institutionalization, and mortality (25–32).

Limitations

The 1-year experience from LIFE-P may not allow us to accurately project what would occur over the course of a longer trial; it is possible, for example, that the covariance, transitions, and trajectories in later years will differ from those in the pilot phase. Participants in LIFE-P, although recruited across four centers in a manner to enhance its diversity, may not broadly represent physically impaired older persons. As for any trial, the generalizability of our results may be limited by inclusion criteria and the characteristics of participants. In particular, LIFE-P participants were required to complete a W400 at baseline, hence our results may not extend to persons who have more limited mobility. Our simulations incorporated the same rates of missing data for all four outcomes. Although high rates of data collection were observed for each outcome during the LIFE-P trial, the SRDS was successfully collected on 1%–2% more participants than the other outcomes at both time points. We have not factored in either cost or participant burden considerations into our recommendation on outcomes; some advantages may accrue according to these considerations for questionnaire-based outcomes.

We have not explicitly addressed differences among clinical sites in the distributions of outcomes. These depend, in part, on how sites are selected and monitored. If not accounted for in analyses, variance may be inflated and generalizability may be limited (33). The LIFE-P analysis plan anticipated this eventuality by including site as a covariate when evaluating the impact of interventions on outcomes (4,22). In our results, these differences contribute to the overall variability we describe. If future trials anticipate different levels of intra-site correlations, some adjustment to our projected sample size may be required. There is also the potential that the effect sizes of interventions may vary among sites; if information on the likely variance of these differences is available, further adjustments may be made. The LIFE-P intervention featured walking as the primary means to enhance physical activity, thus the intercorrelations among outcomes may reflect a common sensitivity to increased walking.

Summary

Our analyses of data from LIFE-P suggest that its planned full-scale trial may have greater efficiency if, among the outcomes it considered, it adopts the ability to complete a W400 as its primary outcome. It is likely, however, that no universally optimum outcome exists for trials of mobility disability. Primary outcomes should not be selected solely on the basis of their interrelationships with alternative outcomes and statistical efficiency, but also on the basis of their associations with other health markers, participant burden and safety, and their ability to influence clinical practice (34,35). Thus other outcomes may prove to be more useful for trials in different populations and contexts, and the relative advantages of competing outcomes will be better understood as experience accrues. To this end, it may be profitable to include more than one measure of mobility disability in future study designs.

Acknowledgments

The Lifestyle Interventions and Independence for Elders (LIFE) Pilot Study was funded by a National Institutes of Health/National Institute on Aging (NIH/NIA) Cooperative Agreement (U01AG22376) and was sponsored in part by the Intramural Research Program, NIA, NIH. The Wake Forest University School of Medicine, Winston-Salem, NC 27157. E-mail: mespelan@wfu.edu

Address correspondence to Mark A. Espeland, PhD, Department of Biostatistical Sciences, Wake Forest University School of Medicine, Medical Center Boulevard, Winston-Salem, NC 27157. E-mail: mespelan@wfu.edu
REFERENCES


Accepted March 9, 2007
Decision Editor: Luigi Ferrucci, MD, PhD

APPENDIX

In our structural equation, we labeled the underlying construct X. X expresses the greatest ‘‘commonality,’’ that is, underlying intercorrelation of changes in outcomes. Loosely, we consider and interpret it as an expression of the degree of mobility disability. We assumed that each individual ‘‘i’’ ages along a trajectory, which over a relatively short range of 1 year could be described by a linear model:

\[ X_i(t_k) = a_i + b_i t_k \]

in which \( a_i \) and \( b_i \) are the intercept and rate of decline for participant \( i \), \( k \) denotes occasion, and \( t_k \) denotes time. We assumed that \( a_i \) and \( b_i \) were random effects that followed Gaussian distributions. If mobility disability, \( X \), were observable, it would be the focus of clinical trials. Because \( X \) is not observable, we assumed that changes in the \( W_{400} \) (‘‘unable’’ coded as \( W_{400} = 1 \) and ‘‘able’’ coded as \( W_{400} = 0 \)), \( W_4 \) (‘‘unable’’ coded as \( W_4 = 1 \) and ‘‘able’’ coded as \( W_4 = 0 \)), SPPB (labeled ‘‘S’’), and SRDS (labeled ‘‘D’’).
each were driven in some part by changes in X. Critically, we viewed interventions to have been designed to have a beneficial effect on how X changes over time. How interventions do this is expressed indirectly through their effects on other outcomes. Thus,

$$\log \left( \text{odds} | W_{400}(t_k) \right) = w_0 + w_1(X_{ik}(t_k)),$$

$$\log \left( \text{odds} | W_{4}(t_k) \right) = m_0 + m_1(X_{ik}(t_k)),$$

$$E[S_{ik}(t_k)] = s_0 + s_1(X_{ik}(t_k)),$$

$$E[D_{ik}(t_k)] = d_0 + d_1(X_{ik}(t_k)),$$

where “E” refers to the expected value and intervention effects are modeled as influencing the mean of $X_{ik}(t_k)$. The slopes ($w_1, m_1, s_1,$ and $d_1$) in these relationships are reported in Table 4 in standard deviation units to express, using a common yardstick, the degree to which changes in individual outcomes are associated with changes in X.

Many statistical algorithms may be applied to fit this model. We used the Bayesian algorithm of Gibbs sampling (14,15), which provided flexibility in addressing missing data. This approach required us to define prior distributions for parameters, which we chose to be noninformative. Potential values for parameters were iteratively sampled and accumulated to produce a posterior distribution for each. We report the medians of these posterior distributions as point estimates and 95% equal-tail credible intervals, which are roughly analogous of 95% confidence intervals in frequentist approaches. A “burn-in” phase (during which the first 50,000 samples were discarded) was used to remove the influence of initial starting values. Analyses were rerun from a range of starting points; the congruence of these results and graphical inspection of the sequential samples allowed us to conclude that the estimation process was stationary and that 20,000 samples were sufficient to provide stable estimates.