EVALUATIONS of POTENTIAL APBD PRIMARY ENDPOINTS

By Larry Schwartz, Econometrician and Survey Specialist, February 2017

INTRODUCTION

The FDA says that a primary endpoint for a clinical trial should measure detectable benefits to patients (when possible). And detectable benefits include lengthened survival, improved symptoms, enhanced functional capacity, and decreased chances of developing a disease complication.[1]

Further, the FDA defines two acceptable types of primary endpoints: (1) Objective measures: survival, disease exacerbation, clinical event; and (2) Subjective measures: Symptom score and a "health related quality of life" survey. [2] Both types of primary endpoints need to be valid and reliable as well as sensitive to patient baseline differences and condition changes.[3]

Investigators have focused on the following four measures for APBD clinical trials thus far:

(1) The Six-minute walk (6MW)---an objective measure;
(2) The 13-item Spastic Paraplegia Rating Scale (SPRS)---subjective measure;
(3) The Rand 36-item Health Survey (36HS)----a subjective measure; and
(4) The 10-item Amyotrophic Lateral Sclerosis Functional Rating Scale (ALSFRS)--a subjective measure.

From 2007-2015, the 6MW was used as the primary endpoint with both the C7 Phase 1 and Phase 2 trials for APBD, while the 36HS was used as a secondary endpoint for the C7 Phase 2 trial for this disease. The SPRS has been discussed as a possible primary endpoint in the upcoming Guaiacol trial for APBD. Similarly, someone outside the APBD family recently proposed the ALSFRS for an APBD clinical trial.

The author reviews published statistical evaluations of these four measures to answer the following question: Which measures are the most promising for an APBD primary endpoint?

---

[1] **Sullivan, E.J., Clinical Trial Endpoints, Course for the FDA, 2013. However, validated surrogate endpoints may substitute for primary endpoints when direct measures are not practical medically or prohibitively expensive. This paper does not address the Cures Act of December 2016 that seems to allow observational data arising from routine clinical use in place of prospectively collected data from randomized clinical trials. But the FDA's regulations for the Cures Act are pending.**

[2] **According to the FDA, a primary endpoint does not necessarily have to be a single measure. It is acceptable to combine primary endpoint measures in clinical trials. In 2016, Clinicaltrials.gov shows more than 7,000 combined primary endpoints, about three percent of the total clinical trials.**

[3] **For the six-minute walk, validity refers to the degree to which this measure is in agreement with other measures of aerobic exercise capacity, such as a treadmill; reliability to demonstrate reproducible results; and sensitivity to track the progression of the disease**.

The other three measures are survey instruments. In this paper, we focus on construct validity of the survey instrument to determine if it measures what is intended; at a minimum, such validity is determined by pre-testing the survey instrument with a handful of would-be respondents to make sure that the questions are understood, clear, relevant, and complete. Reliability is the extent to which related survey questions are internally consistent with one another, while sensitivity is concerned with how well survey results over time track disease progression from onset to death (or cure).

The paper is organized as follows:

Section 1: Summary of Findings

Table 1 summarizes the findings of the four measures as potential primary endpoints in APBD clinical trials. It is based upon the details of these measures for meeting the desirable properties of validity, reliability and sensitivity (Sections 3-6).

Statistical correlations are used to determine the strength of these properties for each of the measures. The author uses the following correlation standards to categorize degrees of strength;

"Weak" is for a correlation less than .70;
"Moderate" between 0.70 and 0.85; and
"Strong" greater than 0.85.

Appendix A explains the different correlations.

Table 1: Strength of Measures for Meeting Desirable Properties

PROPERTY	***M E A S U R E***			
	6MW	**SPRS**	**36HS**	**ALSFRS**
Validity	**Moderate**	**Strong**	**Moderate**	**Moderate**
Reliability	**Moderate**	**Strong**	**Moderate**	**Moderate**
Sensitivity	**Weak**	**Weak**	**Weak**	**Moderate**

Broadly speaking, Table 1 shows that a measure may have good theoretical properties for capturing what is intended (validity) and reproducing those intentions (reliability), but it may or may not predict real world disease changes very well (sensitivity).

SECTION 2: CONCLUSION and RECOMMENDATION

CONCLUSION: The ALSFRS is the best measure of the four reviewed. It is the only one that tracks disease progression fairly well (moderate sensitivity). And this measure has acceptable strengths for validity and reliability, as the others do.

The ALSFRS is not only used in clinical trials for ALS but also for Multiple Sclerosis, Muscular Dystrophy, spinal diseases, and other nervous system diseases. Moreover, there is support to self-administer the ALSFRS online for both clinical trials and managing the care of ALS patients.

RECOMMENDATION: APBD investigators should seriously consider using the ALSFRS as a primary endpoint in future APBD clinical trials.

SECTION 3: SIX-MINUTE WALK (6MW) [4]

The 6MW measures the distances that a patient can walk on a flat, hard surface in a period of six minutes. It does not take into account the quality or continuity of the walk. And does not necessarily represent all the other aspects of APBD: Neurogenic bladder, energy deficiency, and possible cognitive problems.

On the average, the studies reviewed on the 6MW were moderately valid and reliable but weakly sensitive.

On validity for example, in a 2008 study Pulz found that that the six minute was moderately valid with patients suffering from chronic heart failure; correlation of 0.76 when compared with the clinical measure of peak oxygen consumption. [5] And on reliability in a 1998 study, Rikki and others concluded that "the 6-min walk can be used to obtain reasonably reliable measures of physical endurance in older adults…."[6]

But on sensitivity, a 2011 longitudinal study of 106 young ambulant boys with Muscular Dystrophy that received steroid treatment showed the following with the six-minute walk: Weak response or sensitivity to condition changes as measured by the clinical North Star Ambulatory Assessment (Spearman correlation was 0.52). [7]

Moreover, a 2013 study of outpatients with chronic heart failure showed that the six minute walk test was weakly sensitive to performing daily activities (correlation 0.59)[8] Further, a 2005 study on the 6-minute walk for elderly patients (60 years and older) with chronic heart disease found "No relationship between baseline symptom severity and 6-MWT performance."[9] Finally in a 2011 study of 826 patients with idiopathic Pulmonary Fibrosis, changes in the 6MW compared were weakly sensitive to changes in the clinical measures of physiologic function, dyspnea, and HRQL(correlation coefficients less than 0.50). [10]

---

[4] **As of 13 January, 2017, Clinical trials.gov shows 1,073 trials involving the six-minute walk; covers heart conditions, metabolic diseases, and Cystic Fibrosis, to name a few.**

[5] **Pulz, C. and others, *Incremental Shuttle and Six Minute in Chronic Heart Failure Walking Tests in the Assessment of Functional Capacity,* Can J Cardiology, 2008.**

[6] **Rikki, R.E. and Jones, C.J., the *Reliability of a 6-Minute Walk Test as a Measure of Physical Endurance in Older Adults,* Journal of Aging and Physical Activity, 1998.**

[7] **Mazzone, E. and others*, Functional Changes in Duchenne Muscular Dystrophy: A 12-month Longitudinal Cohort Study,* Neurology, 2011**.

[8] **Shoemaker, M.J. and others, *Clinically Meaningful Change Estimates for the Six-Minute Walk Test and Daily Activity in Individuals with Chronic Heart Failure*, Cardiopulm Phys Ther J., 2013.**

[9] **Ingel, L. and others, *The Reproducibility and Sensitivity of the 6-Minute Walk Test in Elderly Patients with Chronic Heart Failure,* The European Society of Cardiology, 2005.**

[10] **Du Bois, R.M. and others, *Six-Minute-Walk Test in Idiopathic Pulmonary Fibrosis*, AJRCCM, 2011**.

SECTION 4: The Spastic Paraplegia Rating Scale (SPRS)[11]

Developed in the 1970s, the SPRS contains the following 13 factors:[12]

1. Walking distance without pause
2. Gait quality
3. Maximum gait speed
4. Climbing stairs
5. Speed of chair climbing
6. Arising from chair
7. Spasticity, hip abductor muscles
8. Spasticity, knee extension
9. Weakness, hip abduction
10. Weakness, dorsiflexion
11. Contractures of lower limbs
12. Pain due to SP-related symptoms
13. Bladder and bowel function.

Note that SPRS contains much of what is relevant to APBD, especially aspects of walking and bladder and bowel function. However, the cognitive factor and energy level are not included.

Clinicians and patients together rate health conditions with the 13 SPRS items. Each factor is rated with a Likert scale from 1 to 4, best result to worse.

On walking distance without pause, for example, the SPRS rating scale is:

1=able to walk>500 meters without walking aid;
2=able to walk > 500 meters with walking aid;
3=able to walk <500 meters with walking aid; or
4=not able to walk.

It normally takes 15 minutes to complete the SPRS, and it can be done on an outpatient basis without any special equipment.

On the average, the evaluation studies showed that the SPRS is strongly reliable and valid but weakly sensitive to disease progression.

In a 2006 exhaustive study of spastic paraplegia, Shule shows that the SPSS has strong internal consistency or reliability with a Cronbach Alpa statistic of 0.91. And he shows that its validity is equally high according to the correlation with clinical measures.

But on sensitivity in a 2016 study of spastic paraplegia, it was found that SPRS followed clinical measures with a 0.64 correlation (weak sensitivity).[13]

---

[11] **As of 13 January 2017, Clinicaltrials.gov shows only four trials involving the SPRS, three cases for spastic paraplegia and one for spin cerebella.**

[12] **Shule, R. and others, *The Spastic Paraplegia Rating Scale (SPRS): A Reliable and Valid Measure of Disease Severity,* Neurology, 2006**

[13] **Martinuzzi, A. and others*, Clinical and Paraclinical Indicators of Motor System Impairment in Hereditary Spastic Paraplegia: A Pilot Study,* PLoS One, 2016.**

Also in a 2010 study of motor and functionality for spastic paraplegia and related conditions, it was found that the "SPRS was not a useful indicator of disease progression."[14]

Most importantly as of May 1, 2010, the American Academy of Neurology states that the SPRS is not very sensitive to disease progression. This is based upon the 2006 Shule study that analyzed clinical observations of 63 spastic paraplegia patients at ages 9 to 80 and compared them to independently derived measures of the 13 factors in the SPRS. It found a Spearman correlation coefficient of 0.40 between changes in the SPRS scores and the duration of spastic paraplegia (weak sensitivity).[15]

Shule and others think that the heterogeneity of the spastic paraplegia condition is the reason why the SPRS does not have strong sensitivity. They further think that a longitudinal study may help define distinct subtypes of this disease, thereby reducing variability for each subtype and increasing its sensitivity.

If the Foundation uses the SPRS today, it probably would face the same problem of weak sensitivity because APBD also is a heterogeneous condition, much like any other rare disease.[16] But it too may be able to breakdown APBD into subtypes with the collection of extensive natural history data.

---

[14] **Graciani, Z. and others*, Motor and Functional Evaluation of Patients with Spastic Paraplegia, Optic Atrophy, and Neuropathy,* Arq Neuropsiquiatr, 2010.**

[15] **Same source as footnote 12.**

[16] ***NIH and FDA, Workshop on Natural History Studies of Rare Diseases: Summary*, NIH Campus, May 2012.**

SECTION 5: RAND 36-ITEM Health Survey (36HS) [17]

The 36HS is an outgrowth of a much larger survey, the 116-Item Medical Outcome Study.

Its 36 questions can be summarized by the following umbrella categories:[18]

General Health---questions 1, 2, and 33-36;
Health limiting physical activities---questions 3-12, including three questions on walking;
Health creating problems at work---questions 13-16;
Emotional problems---questions 17-19, 25, 28;

Physical/emotional problems interfering with social activities---questions 20 and 32;
Bodily pain---questions 21 and 22;
Energy level---questions 23, 27, 29, and 31;
Nervousness—question 24; and
Calm, peaceful and happy—questions 26 and 30.

On the average, the 36Hs has moderate reliability and validity but weak sensitivity.

On reliability, in a 1996 general health study, Jenkinson and others found moderate internal consistency for the items of the SPRS, with a Cronbach's Alpha averaging 0.82.[19] And on construct validity, a 2005 Malaysian asthmatics study found a moderate Spearman correlation of 0.70 with objective measures for nine physical functioning items in the SPRS.[20]

And more generally on validity and reliability, in a 1996 Australian SF-36 application to 90 1-year stroke survivors (mean age, 72 years), it was found that "satisfactory internal consistency [reliability] and a valid measure of physical and mental health after stroke… but not [for] social-functioning."[21] And in 2012, 460 Chinese Chronic Hepatitis patients aged 18-76 completed SF-36v2 [the Chinese version of the SF-36], it found: "both reliability and validity demonstrated to be strongly satisfactory….but more work should be carried out to evaluate the sensitivity…." [22]

But on the sensitivity in a 1997 study of post-operative patient changes, changes in the 36Hs survey correlated weakly with changes in various clinical parameters: correlation of 0.47 for hip replacement; 0.6 averages for lung cancer; and somewhat above 0.5 for abdominal aortic aneurysm repair. [23]

---

[17] **As of 13 January 2017, Clinicaltrials.gov shows 769 studies involving 36HS; covering everything from degenerative disk disease and heart bypass surgery to depression, neck pain, and diabetes.**
[18] **Rand Corporation**.
[19] **Jenkinson, C. and others*, Evidence for the Sensitivity of the SF-36 Health Status Measure to Inequalities in Health, Journal of Epidemiology & Community Health, June 1996.***
[20] **Sararaks, S. and others, *Validity and Reliability of the SF-36: the Malaysian Context*, Med J Malaysia, June 2005.**

[21] **Anderson, C. and others, *Validation of the Short Form 36 (SF-36) Health Survey Questionnaire among Stroke Patients, Stroke*, October 1996.**
[22] **Zhou, K.N. and others, Reliability, Validity, and Sensitivity of the Chinese (Simple) Short Form 36 Health Survey (SF-36v2) in Patients with Chronic Hepatitis B, TOC, April 2013.**
[23] **Mangione, C.M. and others, *Health-Related Quality of Life after Elective Surgery: Measurement of Longitudinal Changes*, J Gen Intern Med, November 1997.**

And for tracking conditions after orthopedic surgery, it was found that "SF-36 subscales have low sensitivity to individual changes and so we caution against using SF-36 to monitor the health status of individual patients undergoing orthopedic surgery."[24]

Others reported certain problems areas with 36HS sensitivity. In 1996 Jenkinson and others state "The short form 36…. may not be appropriate for…. assessing health gains because of the low responsiveness (sensitivity to change) …"[25] And in a 2014 article on Rheumatoid Arthritis: "Low responsiveness precluded estimation of valid MCIIs for many SF-36 scales in patients with RA, particularly the scales assessing mental health."[26]

[24] Busija, L. and others, Magnit*ude and Meaningfulness in SF-36 Scores in Four Types of Orthopedic Surgery,* Health Qual Life Outcomes, July 2008.
[25] Same source as footnote 19.

[26] Ward, M.M. and others, *Clinically Important Changes in Short Form 36 Health Survey Scales for Use in Rheumatoid Arthritis Clinical Trials: The Impact of Low Responsiveness*, Arthritis Care & Research, December 2014.

SECTION 7: The Amyotrophic Lateral Sclerosis Functional Rating Scale (ALSFRS) [27]

After beginning development in 1991, the current version of ALSFRS consists of 10 items:

Speech
Salivation
Swallowing

Handwriting
Cutting food and handling utensils
Dressing and hygiene

Turning in bed and adjusting bed clothes
Walking
Climbing stairs
Breathing

The Likert scale for each of its 10 items ranges from one to four, with four indicating the best result and one the worse. For example, the ALSFRS scale for walking is as follows:

Walking Normal--4
Early ambulation difficulties---3
Walks with assistance---2
Non-ambulatory functional movement only---1

On the average, the ALSFRS studies showed moderate reliability, validity and sensitivity.

In a 1996 seminal study with 75 ALS patients, Brooks and others found the following with the ALSFRS: [28]

- Internal consistency or reliability was moderate (Cronbach's alpha=0.81);
- Construct validity with objective measures was strong (Spearman correlation coefficients was 0.94); and
- Sensitivity to changes in the ALS condition were moderate (correlations of ALSFRS changes with changes in various objective measures ranged from 0.71 to 0.79).

Brooks concluded that the ALFRS may be used as a "surrogate measure in clinical practice or in ALS clinical trials…." In fact, a 2012 study supported the self-administration of the ALFSRS online within clinical trials and for managing the care of ALS patients.[29]

---

[27] As of 13 January 2017, Clinicaltrials.gov shows 102 trials involving the ALSFRS; mostly ALS but also Muscular Dystrophy, Multiple Sclerosis, and spinal cord diseases, to name a few. Note that physicians have commonly misdiagnosed APBD for Multiple Sclerosis and ALS.

[28] Brooks, B. and others, Assessment of Activities of Daily Living in Patients with Amyotrophic Lateral Sclerosis, Archives of Neurology, February 1996.

[29] Maier, A. and others, *Online Assessment of ALS Functional Rating Scale Compares Well to In-Clinic Evaluation: a Prospective Trial*, Amtotroph. Lateral Scler., 2012

But in a 1999 study of 387 ALS patients evaluated for each of nine months, Cedarbuam and others found lower results for the strength of the measures: Correlation of 0.71 on internal consistency (moderate reliability); 0.71 on construct validity (moderate), and 0.57 when comparing the ALSFRS with the inverse of the Sickness Impact Profile (weak sensitivity).

More recently on sensitivity, however, a 2014 study of ALS patients found that ALSFRS scores estimated the four ALS clinical stages very well. The Spearman correlation coefficient between the ALSFRS scores and the ALS patient stages was 0.92 (strong sensitivity). [30]

[30] **Balendra R. and others, *Estimating Clinical Stage of Amyotrophic Lateral Sclerosis from the ALS Functional Rating Scale,* Journal of Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 2014.**

APPENDIX A: Correlation Statistics

In the references for this investigation, two correlation statistics were repeatedly used to evaluate the different measures. They are the Spearman correlation coefficient and Cronbach's alpha (intra-item correlation).

Each is explained, in turn.

Cronbach's Alpha (intra-class correlation)

Cronbach's Alpha indicates the degree to which related items in a survey form a single construct. In the context of clinical trials, the Cronbach's Alpha is useful for measuring internal consistency (reliability) of related items in a quality of life survey. A Cronbach's Alpha of 0.70 is considered the minimum level for a statistically important construct.

The formula for Cronbach's alpha is as follows:

$$\alpha = \frac{N \cdot \bar{c}}{\bar{v} + (N-1) \cdot \bar{c}}$$

Here N is equal to the number of items, c-bar is the average inter-item covariance and v-bar equals the average variance of the items. And variance is a measure of the scatter for each item taken separately, while covariance is the degree of change between any two items taken at a time.

Spearman Rank Correlation Coefficient

In general terms, two variables that perfectly move together would have a Spearman Rank Correlation of unity. When there is no relationship between the two variables, the Spearman correlation coefficient would be zero. The more common case is somewhere between the two extremes. A Spearman rank correlation of 0.70 is usually considered the minimum level for a statistically important relationship.

Technically to calculate Spearman's correlation (SRC) between two variables, you rank the two of them, calculate their differences, and square those differences or $\Sigma d^{2,}$

And then use the following formula to calculate SRC:

SRC=1-6(Σd^2) /[n(n^2-1)], where n=the number of observations for each of the two variables.

The Spearman Rank Correlation is used for assessing the validity and sensitivity of all the four measures, and for the reliability of the Six-Minute Walk.