# Commentaries

## Gold Standards

**It is generally expected that any new** treatment, test battery or approach to a problem should be compared to a "gold standard." Gold standards die hard, which may or may not be a good thing, depending on how the standard became "gold" to begin with. I have two examples from my own areas of professional interest. One is the standardized assessment of motor function in **Parkinson's disease (PD).** The other is the clinical research paradigm of "intention to treat."

There have been a few batteries of testing used to rate PD severity. At a meeting of experts in the 1980s, a consensus was achieved which rated behavior, activities of daily life, and observed motor function. Later, sections were added to rate treatment complications, and an overall staging. This became our gold standard, and has been used in almost every clinical trial over the last 10-15 years. It has been revised due to some obvious shortfalls, which is, I think, a highly meritorious track record. Periodic reviews of "gold standards" are a good thing.

One of the interesting and troubling aspects of our **Unified Parkinson's Disease Rating Scale (UPDRS)**, however, was a teaching videotape that was made using the panel of experts who developed the test battery, and published in our specialty journal that provided videotape examples of the actual scoring for the motor tasks the patients perform. As a colleague remarked, "When I reviewed this article and videotape for the journal, I recommended that it be published, but that the videotape examples should not be considered the gold standard for evaluating PD patients." Unfortunately, it has become that gold standard. At multiple initiation meetings for drug trials of new medications for PD the UPDRS is reviewed and the investigators are instructed to study the teaching videotape and use it as the gold standard although, it is acknowledged, the tests are not all performed correctly, and are not even performed in a uniform manner. Thus we are told that our gold standard has feet of clay. It is a problem for our field, not rectified after 20 years of discussion.

In a related field, movement disorder side-effects of psychiatric drugs, there are three standard test batteries: the Abnormal Involuntary Movement Scale, to measure tardive dyskinesia; the Barnes Akathisia Rating Scale to measure akathisia (restlessness); and the **Simpson Angus Scale (SAS)** to measure parkinsonism. The SAS, unfortunately is a seriously flawed test that has never, and will never, be used by experts in parkinsonism because it excessively scores some abnormalities, underscores others and gives instructions on measurement that are often impossible to employ. Nevertheless, I doubt that any psychiatric journal would accept a paper on antipsychotic medication that did not include this scale of measurements of parkinsonism. It is, after all, the gold standard.

My final gripe is the Intention to Treat approach in analyzing treatment trials. The ITT approach was designed to obtain complete data on all randomized subjects to reduce bias induced by early terminations. The analytic plan assesses outcome based on treatment assignment, whether or not the subject actually gets treated. For example, a protocol to assess shunting to treat normal pressure hydrocephalus randomly assigns a treatment plan to each subject. Once that assignment is made, the outcome is analyzed based on that assignment, *whether or not the subject is shunted*. Thus, if half the subjects assigned to shunting change their mind, and their gait fails to improve, the analysis may show a failure of shunting, even if the majority of those *actually shunted* improved. Of course, a secondary subgroup analysis can be built into the analytic plan, to reveal that those actually shunted improved, but the "bottom line" of the study, the newspaper headline, will be a failure of treatment.

An article published in this journal *(MHRI)* last year looking at quality assessment of medical care, focused on the treatment of atrial fibrillation as one condition to be used for analysis. I learned from that article that warfarin was ineffective at preventing stroke when the data were analyzed by correlating prescriptions for the drug and outcome. This occurred, of course, because of the high non-compliance rate. When measures of prothrombin time and outcome were measured, the benefit was clear. Thus a study with a significant non-compliance rate may confound the beneficial effect of a drug. Yet, unless reviewers for grant proposals, or journals are statistically savvy (and many are not), or willing to consider non-standard approaches, a perfectly good study may be criticized precisely because it was innovative, recognizing pitfalls of the "gold standard" approach.

Rosalyn Yalow, PhD, at her Nobel Prize acceptance speech, included slides of the letters of rejection she had received from journals for her innovative approach to studying insulin release. Using contemporary "gold standard" paradigms, her work was not acceptable because it challenged those beliefs.

I believe in "gold standards." Unlike real gold, however, they should not be considered noble or immutable, but templates that may require modifications over time.

The Diagnostic and Statistical Manual, defining every psychiatric disorder, is a good example, whether or not you agree with its methodology or choices. It is the gold standard for its field, but it is reassessed and altered every few years.

**– Joseph H. Friedman, MD**

### Disclosure of Financial Interests

Joseph Friedman, MD, Consultant: Acadia Pharmacy, Ovation, Transoral; Grant Research Support: Cephalon, Teva, Novartis, Boehringer-Ingelheim, Sepracor, Glaxo; Speakers' Bureau: Astra Zeneca, Teva, Novartis, Boehringer-Ingelheim, GlaxoAcadia, Sepracor, Glaxo Smith Kline, Neurogen, and EMD Serono.