NEUROLOGY 2006;66:626–627

Editorial

CME

Futility studies

Spending a little to save a lot

Steven R. Schwid, MD; and Gary R. Cutter, PhD

Phase III studies of disease-modifying agents for neurodegenerative disorders generally require hundreds of patients to be followed for at least 2 years at a cost of millions of dollars. No matter how promising an agent may appear in preclinical studies and animal models, preliminary evidence that the treatment is likely to be helpful in patients (proof-of-concept) and that it is reasonably safe is essential before asking patients to accept risk and making the necessary investments in human resources, time, and money. In this issue of *Neurology*, three articles discuss futility studies, a clever method of dealing with the trade-off between investment risk and clinical promise.[1-3] The requirements to definitively demonstrate clinically meaningful benefits and long-term safety demand long duration, large sample size Phase III studies. Proof-of-concept studies have more flexibility in their design and endpoints that may allow for smaller and more cost effective studies.

In multiple sclerosis (MS) research, the need for proof-of-concept studies has often led to the use of MRI measures of disease activity, especially gadolinium enhancing lesion incidence, as primary endpoints. Serial assessment of gadolinium enhancing lesions can demonstrate that an agent reduces inflammatory disease activity with as few as 20 patients followed for 3 to 6 months. Similar biomarkers are actively being sought for other neurologic diseases with variable degrees of success. It is not sufficient to identify biomarkers that correlate with disability and are reliably quantified. For such biomarkers to be advantageous, they also need to be more sensitive to therapeutic effects than the clinical measures they replace.[4] Moreover, they need to accurately predict clinical outcomes from definitive studies. This appears to be the case for gadolinium enhancing lesions reflecting MS inflammatory activity, but other biomarkers await this type of validation. More sensitive clinical endpoints, such as those based on quantitative measures of function, may also provide similar advantages compared to traditional ordinal scales.[5]

Another way to reduce the sample size needed for a preliminary study is to focus on futility, designing a study to identify which agents are least likely to demonstrate benefits rather than the more typical goal of identifying the most promising agents. Put simply, most studies focus on efficacy, with a null hypothesis that treatments are equivalent and rejection of the null hypothesis if one treatment is likely to be more effective than the other. On the other hand, the null hypothesis in a futility study is that the treatment has promise and will therefore produce results exceeding a meaningful threshold. If the threshold is not met, the null hypothesis is rejected and further study of the treatment is considered futile. One major conceptual difference exists between these designs: agents passing an efficacy criterion are winners, but agents passing a futility criterion are merely non-losers. As a result, non-futile agents are less likely to show benefit in Phase III trials than agents demonstrating preliminary evidence of efficacy, unless futility thresholds are set stringently.

Thresholds are based on estimated results from a control group, which may be included as one of the treatment arms in the trial (concurrent control) or derived from a previous trial (historical control). For example, in Levy et al.,[3] the investigators chose a threshold requiring a decline in an amyotrophic lateral sclerosis rating scale at least 20% less than the concurrent control rate. Similarly, both the NET-PD study[1] and Tilley et al.[2] focused on a threshold requiring a change in a Parkinson disease (PD) rating scale at least 30% less than a historic control rate. In each of these studies, patients had to progress at rates 20 to 30% less than control rates before considering the agent non-futile. In effect, the investigators decided that agents producing rates closer to placebo were not worth studying further, even though they

See also pages 628, 660, and 664

From the Department of Neurology (S.R.S.), University of Rochester, NY; and Birmingham Department of Biostatistics (G.R.C.), University of Alabama.
Disclosure: The authors report no conflicts of interest.
Address correspondence and reprint requests to Dr. Steven R. Schwid, University of Rochester, Department of Neurology–Neuroimmunology, Box 605, 601 Elmwood Avenue, Rochester, NY 14642; e-mail: steven_schwid@urmc.rochester.edu

might still beat placebo in head-to-head efficacy studies. Depending on the availability of other treatments, seriousness of the disease, toxicity of the treatment, and resources available to perform definitive studies, futility thresholds might be set more or less stringently to adjust the possibility of passing agents that will ultimately prove ineffective and failing agents that would have proven successful. The study sample size required decreases as the futility threshold is set further from the expected control rate.

The big advantage of the futility design is that it allows some agents to be weeded out at the preliminary study stage while requiring a fraction of the patients and resources of more conventional efficacy studies. Levy et al. estimated that a conventional Phase II study would have required 850 to 1,080 patients, while their futility design needs only 185.[3] NET-PD performed a study assessing two different agents with 195 patients rather than the 600 to 800 that would likely be needed for a conventional design.[1] Tilley et al. demonstrated that the classic DATATOP study could have been performed with 400 patients instead of 800 if it had been preceded by a futility study of Vitamin E requiring as few as 84 patients.[2] As in the NET-PD study, advantages for futility studies can be compounded by using one-sided statistical tests and comparing group means to a fixed value (the futility threshold) rather than a control group mean with a distribution of uncertainty around it. However, the benefits of examining futility only make sense in the context of an integrated drug development plan extending across study phases, especially when there are many potential agents to be tested. For example, as isolated observations, knowing that creatine and minocycline are non-futile has limited value.[1] As a prelude to a Phase III study, however, such studies may help in the choice of the agent to study.

So, what is this panacea hiding that must be considered before we begin all trials with futility studies? There are at least four disadvantages. First, reducing the sample size and study duration required in Phase II will limit the ability to identify safety concerns and other time-dependent issues before larger groups of patients are put at risk. Second, agents with delayed therapeutic effects may be inappropriately dismissed (also a risk in conventional studies). Third, relying on historical data, as in the NET-PD design, may provide misleading results. Although the NET-PD group had recent, high-quality historical data, the small control group included in the NET-PD study did not match with the historical control data, raising questions about the appropriateness of the futility threshold used in the study. Clinical trialists in oncology have long used similar strategies, but the quarter century or more of high quality trial data and availability of hard endpoints (e.g., death) makes historical data more useful.[6] In neurologic diseases, such as MS, where the disease definition itself has been changed twice in the past 6 years[7] and trial endpoints are subjective,[8] the use of historical controls is problematic.

Finally, in the context of a drug development program, we must ensure that futility studies and other methods of saving resources during Phase II do not increase costs in Phase III. For example, the NET-PD study showed that minocycline is not futile for delaying progression of PD, but it still does not appear highly promising. Given this borderline result, are huge investments in patients, dollars, and opportunity costs now justified for Phase III studies of minocycline? Perhaps this type of non-futility result demands additional modestly priced Phase II studies before proceeding. In any case, after decades with little to offer patients with neurodegenerative diseases, being forced to choose between non-futile treatments to pursue in additional studies is a good problem to have.

References

1. The NINDS NET-PD Investigators. A randomized, double-blind, futility clinical trial of creatine and minocycline in early Parkinson disease. Neurology 2006;66:664–671.
2. Tilley BC, Palesch YY, Kieburtz K, et al. Optimizing the ongoing search for new treatments for Parkinson disease: using futility designs. Neurology 2006;66:628–633.
3. Levy G, Kaufmann P, Buchsbaum R, et al. A two-stage design for a phase II clinical trial of coenzyme Q10 in ALS. Neurology 2006;66:660–663.
4. Miller DH. Biomarkers and surrogate outcomes in neurodegenerative disease: lessons from multiple sclerosis. NeuroRx 2004;1:284–294.
5. Schwid SR, Goodman AD, Apatoff BR, et al. Are quantitative functional measures more sensitive to worsening MS than traditional measures? Neurology 2000;55:1901–1903.
6. Simon R, Thall PF, Ellenberg SS. New designs for the selection of treatments to be tested in randomized clinical trials. Stat Med 1994;13:417–429.
7. Polman CH, Reingold SC, Edan G, et al. . Diagnostic criteria for multiple sclerosis: 2005 revisions to the "McDonald Criteria." Ann Neurol 2005; 58:840–846.
8. Whitaker JN, McFarland HF, Rudge P, Reingold SC. Outcomes assessment in multiple sclerosis clinical trials: a critical analysis. Mult Scler 1995;1:37–47.