The Utility of Futility

Bruce Levin, PhD

In this issue, Palesch et al discuss the single-arm, phase II futility study design and illustrate how its use might have avoided 3 large (and costly) but negative phase III therapeutic trials for ischemic stroke patients. The authors offer strong arguments to support their conclusion that use of this design as a strategy in phase II development "could permit the testing of a wider array of promising treatments at a fraction of the cost of taking all treatments directly to phase III trials." In a nutshell, they argue that there is utility in futility testing.

Although common in early-phase oncology trials, the futility study (single- or double-armed) may be less familiar to readers of this journal, and careful scrutiny of the design, especially of the formulation of null and alternative hypotheses, is worthwhile. Briefly, in a futility study, the null hypothesis states that the experimental therapy is sufficiently promising to warrant definitive, phase III testing, whereas the alternative hypothesis states that the experimental therapy lacks the prespecified superiority. Thus, the futility design reverses the logical status of null and alternative hypotheses as most often formulated in the traditional efficacy design. Whereas in the latter design, sufficient evidence is required to declare a therapeutic effect statistically significant, in the futility design, there is a presumption of benefit, and sufficient evidence is required to declare a significant shortfall from that benefit, such that it would be futile to proceed to large-scale testing with the given therapy. The authors argue that this formulation, with its null presumption of benefit, is appropriate in phase II research on the grounds that of the 2 types of error that can be committed—declaring a truly superior therapy futile or declaring a truly nonsuperior therapy worthy of continued testing—the former is the more important. One should therefore view it as a type I error with appropriate control of the error rate (α). This the futility design accomplishes.

I suspect the main attraction of the authors' proposal will be the relatively small number of patients required to conduct the experiment in comparison to the sample sizes required for the traditional phase III randomized, controlled trial. Indeed, the authors illustrate 5-fold, 10-fold, and even larger potential reductions in sample size. Three key elements serve to achieve this efficiency. The first is the one-sided nature of the hypotheses (the null states there is a real therapeutic benefit; the alternative denies this directional superiority). The second key element is the use of somewhat more liberal values of alpha, eg, 0.10 (1-tailed), compared with the conventional 0.05 level (2-tailed, or 0.025 1-tailed). This is arguably appropriate for phase II testing. The third key element, which is the most influential and perhaps the most controversial, is the use of only a single (experimental) arm. For this, one needs a clear clinical notion of benefit, based on historical–control data, to quantify the "minimally worthwhile improvement" required for the single-arm futility design.

The authors focus on the single-arm design, but there is no intrinsic reason why the futility design cannot be applied with a concurrent control arm. Indeed, the ongoing NINDS-funded QALS study of high-dose coenzyme Q10 in patients with amyotrophic lateral sclerosis uses a 2-arm futility design. Although this is not the appropriate forum to debate the pros and cons of single-arm studies, the following points may help to avoid some pitfalls of the single-arm futility design:

1. Sample size plays an important role in statistical power, as always, but here statistical power refers to the probability that a therapy that truly does not achieve the minimally worthwhile improvement (eg, something no better than a placebo) is actually deemed futile. Because a therapy will be deemed futile only if its outcomes are significantly worse than the prespecified minimally worthwhile improvement, care should be taken to use sample sizes sufficiently large so that the point at which significance (futility) is declared is no worse than the (historical control) placebo effect. If this were not the case, one could find oneself in the situation of failing to declare a therapy futile (and therefore worthy of continued testing) whose success rate is actually worse than the (historical control) placebo success rate. This embarrassment can be most easily avoided by requiring statistical power of 50% or more to declare a placebo-rate therapy futile.

2. If the minimally worthwhile improvement is set too high, truly beneficial therapies may be deemed futile—an obvious point, perhaps, but worth keeping in mind. In the same vein, the authors point out that when designing a phase II futility study, investigators should choose a value of the minimally worthwhile improvement as close as possible to the "clinically meaningful effect size" they would use in the future phase III trial to provide a reasonable test of the futility hypothesis. This is sage advice, because if the minimally worthwhile improvement is set too low, a larger sample size...
would be required to keep the same rejection region of futility, or, if the sample size remained the same, as mentioned above, the chances increase of failing to reject the null hypothesis with a therapy that would otherwise have been ruled out as futile.

3. Using a single-arm design risks a version of type I error that is not encompassed in the type I error rate (α level) of the futility design. If the placebo success rate in the current study population would be much lower than the historical control value used to determine the minimally worthwhile improvement for the futility study—because, say, the current population is at higher risk—then it is entirely possible for a treatment that would be truly worthwhile for further study in this population to be deemed futile in the single-arm futility study because the historical control rate is too high. In the opposite direction, use of a historical control success rate that is too low compared with what would be true in the current study population—because, say, the historical control studies are out of date and patients generally do better today than they used to—could lead to inefficacious treatments (for this population) being brought to phase III testing, thus thwarting the purpose of the phase II futility design. Some confidence that the historical control data apply to the current study population would thus appear necessary to avoid ambiguities in the interpretation of the study.

This provocative article raises larger issues. One knotty problem is this: should researchers and funding agencies devote the time and resources to screen out futile therapies at all or should we move directly from early-phase research to more definitive phase III testing? In other words, are we wasting precious time in our search for effective therapies to conduct futility studies? If the patient pool and funding resources are each adequate, there may be some merit to proceeding directly to phase III testing. However, in disease domains where patients are rare, when funding is limited, or both, some organized screening program such as the authors suggest may be the more prudent and cost-effective approach.

Viewing the futility design as a screening tool leads us to ask, in the language of the screening paradigm, what are the positive and negative predictive values of futility testing? If we define a “negative” result as a finding of futility and a “positive” result as a finding of “nonfutility” (more precisely, a failure to reject the null hypothesis of superiority), then negative predictive value refers to the proportion of all therapies deemed futile that are truly less efficacious than the minimally worthwhile improvement, and positive predictive value refers to the proportion of all therapies deemed “nonfutile” that truly exceed the minimally worthwhile improvement. Interestingly, the authors’ collection of examples shows a high negative predictive value (of a finding of futility) but only a 1-in-3 positive predictive value (of a finding of nonfutility). Insofar as there is a high prior likelihood of futility (for neuroprotective agents at this point in time), a not-so-high sensitivity of 1−α=0.90 may still yield a high negative predictive value. However, positive predictive value may remain low. For example, an alpha of 0.10 and power (1−β=specificity) of 0.85 implies that a result of “nonfutility” may increase the prior odds on a worthwhile improvement only 6-fold, because (1−α)/β=0.90/0.15=6. If the prior odds on a worthwhile improvement are less than one to 6, the odds of bringing a truly worthwhile therapy into the phase III trial will still be less than 50/50. That may or may not be deemed acceptable, but my point is that, as attractive as the futility design is, it does not assure the identification of minimally worthwhile therapies.

The futility design is set squarely within the paradigm of statistical hypothesis testing, yet it can be argued that its purpose is also to select which of several therapies to bring forward to phase III testing. Indeed, the NINDS Parkinson Disease Network is currently conducting just such a selection process using the futility testing approach. When viewed this way, another paradigm comes to mind, to wit, the “statistical selection” paradigm. At least 2 ongoing NINDS-funded phase II studies are using such techniques to select among several therapeutic doses, the previously mentioned QALS trial and the Phase 2B Study of Tenecteplase (TNK) in Acute Stroke (TNK-S2B). Statistical selection procedures also have much to offer in terms of reduced sample sizes, because they are less concerned with testing null hypotheses under tight control of type I error and more concerned with selecting the best of several competing therapies with a high probability of correct selection when there is a truly best therapy. The textbook by Bechhofer et al is a good source of information on such techniques.

It is gratifying that the field of statistics continues to provide methods as innovative as that of Palesch et al. The authors are to be congratulated for their stimulating contribution.

References

The Utility of Futility
Bruce Levin

Stroke. 2005;36:2331-2332; originally published online October 13, 2005;
doi: 10.1161/01.STR.0000185722.99167.56
Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2005 American Heart Association, Inc. All rights reserved.
Print ISSN: 0039-2499. Online ISSN: 1524-4628

The online version of this article, along with updated information and services, is located on the
World Wide Web at:
http://stroke.ahajournals.org/content/36/11/2331

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published
in Stroke can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office.
Once the online version of the published article for which permission is being requested is located, click
Request Permissions in the middle column of the Web page under Services. Further information about this
process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Stroke is online at:
http://stroke.ahajournals.org/subscriptions/