Published in final edited form as:

Contemp Clin Trials. 2015 November ; 45(0 0): 69–75. doi:10.1016/j.cct.2015.06.013.

The futility study—progress over the last decade

Bruce Levin^{a,*}

^aDepartment of Biostatistics, Columbia University, New York, NY 10032, United States

Abstract

We review the futility clinical trial design (also known as the non-superiority design) with respect to its emergence and methodologic developments over the last decade, especially in regard to its application to clinical trials for neurological disorders. We discuss the design's strengths as a programmatic screening device to weed out unpromising new treatments, its limitations and pitfalls, and a recent critique of the logic of the method.

Keywords

Futility design; Non-superiority design; Phase II trials; Screening program

1. Introduction

As early as the 1960s, oncology clinical trialists were introducing phase II trial designs whose purpose was, in some manner, to screen out unpromising treatments from further, confirmatory testing. Designs by, e.g., Gehan [1], Herson [2], Fleming [3], and Simon [4], allowed for course changes in a treatment evaluation program as a result of sufficiently unpromising efficacy results as they emerged. Whether at an interim or terminal analysis, the primary goal was to determine whether the treatment showed “promise” or a lack thereof (“futility”) in deciding what would be the next stage of the testing program. Such phase II trial designs had four hallmark features. (1) Attention focused (entirely appropriately) on whether or not a treatment had such lackluster performance that enthusiasm for continuing with more definitive testing evaporated, as opposed to the more traditional (but inappropriate) focus on demonstrating superiority with the too-small sample sizes typically used in phase II studies. (2) Clinician researchers knew the natural history of the tumor or disease phenotype sufficiently well to render the use of single-arm studies with no concurrent placebo controls not unreasonable. (3) Cytotoxic agents were available but not enough patients and resources were to test all those agents in phase III trials. (4) Because these experiments by themselves were not intended for regulatory approval, investigators felt at liberty to choose levels of statistical significance and sidedness other than the traditional 5%, two-tailed level if so desired. These hallmark features allowed “futility” tests to enroll many fewer patients, typically one-fourth or less, than those required by a conventional phase III two-arm superiority design with traditional control of error rates, and thereby to reach “go/no-go” decisions that much more quickly.

*Corresponding author at: Department of Biostatistics, Mailman School of Public Health, Columbia University Medical Center, 722 West 168th Street, Room 647, New York, NY 10032, United States. Bruce.Levin@columbia.edu.

The legacy of these early developments continues today not only in modern phase II designs but also in the form of “futility” stopping boundaries for interim data monitoring during conventional phase III trials. The goals, however, are very different in those two settings. Futility stopping boundaries in phase III interim monitoring are intended to save time, resources, and patient burden in a trial where there is already compelling evidence that the trial will, with high probability, not reach its primary goal of determining superiority of one therapy over its comparator. The trial’s primary result is essentially already in hand (although secondary aims may be adversely impacted by stopping for futility). By contrast, the phase II futility design is a screening tool, intended to guide a development program away from unpromising treatments before any definitive answers are in hand. These two senses of “futility” testing are so fundamentally different that we shall limit our focus to phase II testing and give no further consideration here to futility stopping in the context of phase III interim monitoring plans. See [5] for a discussion of the latter topic.

About a decade ago a resurgence of interest in the phase II single-arm futility design emerged due to the influential work of statisticians Barbara Tilley, Yuko Palesch, and their group at the Medical University of South Carolina, together with clinical colleagues Karl Kiebert and others at the University of Rochester, working in collaboration with program officers John Marler, Bernard Ravina, Wendy Galpern, Claudia Moy, and others at the National Institute of Neurological Disorders and Stroke (NINDS) in the series of trials known collectively as NET-PD (NIH Exploratory Trials in Parkinsons Disease). A decades-long search for neuroprotective agents that could reverse, halt, or merely retard the neurodegeneration that takes place in Parkinson’s disease was coming up empty-handed. After more than 20 years and over 5,000 patients enrolled in dozens of phase III trials, none had proven efficacious [6]–[9]. Indeed, the only therapeutic agent that has demonstrated any neuroprotection in any domain is recombinant tissue plasminogen activator (rtPA) which was shown effective in acute stroke patients if treated within three hours of onset [10]. Even today there are still no unambiguously effective neuroprotective agents for neurodegenerative diseases such as Parkinson’s, Amyotrophic Lateral Sclerosis (ALS), Huntington’s, or Alzheimer’s disease, notwithstanding the many animal studies which continue suggesting candidate treatments; see, e.g., [11]. The search had to go on, but putting every candidate through phase III testing was unsustainable. Adopting Herson and Carter’s use of “calibration controls” for single-arm studies in oncology [12], the NET-PD researchers established a screening program whose goal was to weed out unpromising treatments, allowing those which did not evince statistically significant “futility” to proceed to confirmatory, phase III trials. In a series of publications, the NET-PD researchers established such screening trials using the futility design for Parkinson’s disease [13]–[16] and they and other authors recommended its use in other domains including stroke [17]–[19], ALS [20]–[23], Huntington’s disease and other movement disorders [24]–[25], and Alzheimer’s disease [26].

These studies generated a lot of commentary [27]–[31], much, though not all, of it enthusiastic, and the agents which were not rejected as futile have gone on to phase III trials. Though effective neuroprotective agents have not yet been found, these trials have been instructive and continue to increase our methodologic toolkit [32]. Futility trials are still being implemented in various disease domains with the single-arm, calibration control

design (see, e.g., [24] and [33]–[34]) and with two-arm, concurrent controls [35]–[37]. Explications of the design and discussions of its potential and promise continue to appear (see, e.g., [27], [38]–[39], and [40]–[42]).

The futility study has had some successes in weeding out unpromising treatments with a relatively small sample size. A notable case from 2006 was the QALS trial for testing futility of high-dose co-enzyme Q10 in ALS [36]–[37]. The anti-oxidant appeared promising in animal models and there was great interest in it for both Parkinson's disease and ALS. The QALS trial incorporated an adaptive selection procedure in a first stage to determine which of two doses appeared preferable, and then continued randomizing patients to the selected dose and concurrent placebo arms in the second stage. The data from both stages were used in a two-arm futility test with a statistical adjustment to account for selection bias. The trial enrolled its total sample size of 185 on time and within budget (notwithstanding the fact that the trial used concurrent placebo controls with patients who had a life-threatening disease). The authors commented that with a conventional three-arm, phase III design using a one-way analysis of variance to test the null hypothesis of no difference among the two dose and placebo groups, the trial would have needed 852 patients to achieve comparable power and type I error rates. Regrettably, the QALS trial found insufficient evidence to justify continuing to phase III, but the savings in time, resources, and patient burden to get this sobering result was considerable.

The single-arm futility design has not been an unmitigated success in neurology, however, due in large part to the inherent difficulties with the use of historical controls. After discussing the formal logic of the design in the next section we will review its limitations and some criticisms.

2. The logic of the futility design

The essential features of futility tests can be discussed by considering the simplest case of a normally distributed outcome variable with known variance. Practical examples with unknown variance and other endpoint distributions are discussed in [17], [20], [30], and [38]. The reader is referred to [38] for a detailed discussion of futility testing.

We start with the case of a two-arm study randomizing subjects to an experimental agent or a concurrent placebo control, wherein data from the placebo control subjects will be used in the statistical analysis. Because controls are randomized concurrently, this case possesses greater internal validity than the single-arm trial discussed below, though the former has been used less often than the latter due to the larger sample sizes required. We note that conventional nomenclature names a hypothesis test after the alternative hypothesis space, so the futility test would perhaps be more properly termed a “one-sided non-superiority” test. One-sided non-superiority tests are familiar as one component of an *equivalence* test of two drugs or two pharmaceutical formulations, the other component being a non-inferiority test. The futility test discussed here is formally identical to the one-sided non-superiority test used in equivalence testing (except that equivalence tests would ordinarily compare two active treatments), so for the sake of parsimony we shall use “futility” as synonymous with “non-superiority”.

The essential first step is to designate a *margin of superiority* which forms the boundary between the *null hypothesis of superiority* and the *alternative hypothesis of futility (non-superiority)*. Let μ_1 denote the mean of the primary endpoint in the experimental group and let μ_0 denote that in the control group. We suppose that larger values of the mean response correspond to better patient outcomes, and we further suppose it is reasonable to define superiority in terms of additive shifts in the mean response. See [38] for tests in which superiority is defined in terms of multiplicative shifts. Let $\Delta = \mu_1 - \mu_0$, and let Δ_0 denote a *pre-specified* value of Δ called the *margin of superiority*. Then the null and alternative hypotheses are:

$$H_0: \Delta \geq \Delta_0 \quad \text{versus} \quad H_1: \Delta < \Delta_0.$$

Given sample mean responses $\bar{Y}_1$ and $\bar{Y}_0$ from experimental and control groups, respectively, based on samples of size n in each group, we reject H_0 when

$\bar{Y}_1 - \bar{Y}_0 \leq \Delta_0 - z_\alpha \sigma \sqrt{2/n}$, where σ is the standard deviation of the individual responses, assumed known and equal in the two groups, and where z_α is the standard normal quantile cutting off probability α in the upper tail.

In the predominant approach to the design of a futility trial, the margin of superiority Δ_0 is interpreted, importantly, as the *minimal worthwhile improvement* in mean response under the experimental treatment compared to placebo, below which there would be no clinical interest or cost-effectiveness in taking the treatment to confirmatory phase III testing. Thus, unlike the conventional test of the null hypothesis of no difference in means, the futility test *assumes* under H_0 that the experimental treatment offers at least a minimally worthwhile improvement in true mean response, but if the study observations warrant it, they can be used to reject that hypothesis. Because Δ_0 is the *minimal* worthwhile improvement, a truly non-superior treatment would be of no interest to bring forward for further testing and, in that sense, if we reject H_0 it would be “futile” to continue to phase III *even if it were the case that in truth* $\Delta > 0$, i.e., the experimental treatment were truly somewhat better than placebo.

There are some important consequences to this formulation of hypotheses.

1. The data may only be used to reject H_0 , not to confirm it. This is no different from any hypothesis test, but the assertion matters here because of the temptation to accept H_0 and conclude superiority. If we fail to reject H_0 , we *cannot* conclude superiority of the experimental treatment over placebo. We can only conclude that the data *do not rule out superiority*, and that it is therefore logical to bring the experimental treatment forward for confirmatory testing. This is a relatively strong position to be in. Contrast this to the case of a conventional test of the null hypothesis of no efficacy difference, where perhaps due to inadequate sample sizes we do not reject the conventional null. Typically one would have to say that one could not rule out “chance” as the explanation for any observed differences and enthusiasm for further testing would be sharply dampened. In the futility test, however, failure to reject the null hypothesis means that the observations are either promising enough on the face of it to continue testing, or they are at least

insufficient to disqualify the experimental treatment from further examination. There is a palpable strategic advantage in being able to state, on the basis of a well-planned, *a priori* statistical design, that “We haven’t proven superiority, of course, but with 90% confidence we cannot rule out that we have a truly superior treatment,” as compared to being forced to say, “We haven’t found a significant difference so we can’t rule out the possibility of no true difference in efficacy” and then have to justify *post-hoc* why a non-significant result looked interesting enough to continue testing. This advantage is especially relevant insofar as it is not uncommon for phase II studies to be designed as if they were phase III studies, though underpowered.

2. The above logic makes sense for a screening program. In that context, we *want* to reject the null hypothesis of superiority if the data warrant doing so. We should be clear about conflicting goals. Everyone *hopes* an experimental treatment will be successful, so it may seem counterproductive to *want* to reject the hypothesis H_0 :
 0. Naturally sponsors and investigators are ultimately hoping to *prove* H_0 . But in the context of phase II testing with smaller sample sizes than are required for confirmatory phase III trials, we will be unable to confirm the truth of H_0 —that goal will have to wait. Furthermore, because we have acknowledged that bringing every candidate treatment directly to phase III is wasteful, especially when there is a low *a priori* probability of success for any given candidate treatment, it is appropriate to use the futility design to screen out unpromising treatments, i.e., treatments with a poor likelihood of success.
3. Carrying the analogy with screening a step further, we consider the meaning of type I and type II errors and the corresponding sensitivity and specificity of the screening program. The following discussion draws heavily from [38]. In a futility design, a type I error occurs when a truly superior treatment by chance produces sufficiently unpromising results as to cause a declaration of futility. We adopt the attitude that this would be a serious error whose rate of occurrence is to be controlled by specifying a reasonably low alpha level at the superiority boundary, α . A common choice of α in this context is 0.10. A type II error occurs when we fail to declare a truly non-superior treatment as futile. The power of the test is then naturally of interest at the “design alternative” of no efficacy difference, $\delta = 0$.

Now suppose we define *sensitivity* as the probability that we declare a truly superior treatment “non-futile” and *specificity* as the probability that we declare a truly non-superior treatment “futile”. Then sensitivity equals the probability of failing to reject the null hypothesis of superiority with a truly superior treatment, i.e., $1 - \alpha$, at the criterion for superiority (or greater if the treatment is even better), while specificity equals the power of the test at the true δ . [In the traditional design, by comparison, sensitivity would correspond to the power of the test (the probability of rejecting the null hypothesis of no benefit with a superior treatment at a given level of efficacy) while specificity would correspond to the probability of failing to reject the null hypothesis of no benefit given that the efficacy of the treatment is the same as that of placebo, or $1 - \alpha$.] Insofar as it is common to set the type I error probability α lower than the type II error probability β in a traditional

trial, it follows that sensitivity will be greater than specificity for the futility design compared to the traditional design. For example, if $\alpha = 0.05$ and $\beta = 0.20$ (for 80% power) at the design alternative, the futility design will have 95% sensitivity and 80% specificity, whereas the traditional design would have 80% sensitivity and 95% specificity. Suppose we now interpret “futility” as a negative outcome and “non-futility” as a positive outcome (or at least a non-negative outcome). Then the negative predictive odds of a futility outcome is given by the prior odds on a non-superior treatment times the likelihood ratio of specificity over one minus sensitivity, or $(1 - \beta)/\alpha = .80/.05 = 16$. This likelihood ratio means that a futility outcome is at least 16 times more likely under the non-superiority hypothesis at the design alternative of no benefit than under the null hypothesis of criterion superiority. On the other hand, the positive predictive odds of a non-futile outcome is given by the prior odds on a superior treatment times the likelihood ratio of sensitivity over one minus specificity, or $(1 - \alpha)/\beta = .95/.20 = 4.75$ (meaning a non-futile outcome is 4.75 more likely under the superiority hypothesis than under the design alternative of no benefit). Thus a futility outcome multiplies the prior odds on non-superiority (which for neuroprotective agents must be quite high, given the past record of failure) by a factor of 16 or more, yielding a posterior odds on non-superiority yet an order of magnitude greater than the prior odds; whereas failure to declare futility increases the prior odds on superiority (which must be quite small) by a factor of only 4.75. (We note here that these likelihood ratios consider only the evidence of having declared a treatment futile or nonfutile, nothing more specific or quantitative. Much more informative likelihood ratios can generally be constructed using the observed data from the experiment which can strengthen the predictive values.)

For example, if the prior odds on non-superiority were 10 to 1 (corresponding to a prior probability of superiority of 1/11), then increasing the prior odds by a factor of 10 would yield posterior odds on non-superiority of 100 to 1 (corresponding to a posterior probability of superiority of $1/101 \approx 0.01$). On the other hand, increasing the prior odds on superiority of 1 to 10 by a factor of 4.75 would yield posterior odds on superiority of $4.75/10 = 0.475$ (corresponding to a posterior probability of superiority of only $0.475/(1 + 0.475) = 0.322$). The odds would still not be in favor of a success in phase III testing.

Consequently, the futility design does a reasonable job of producing negative weight of evidence for unpromising therapies. If a therapy is not screened out as futile, it still must undergo subsequent definitive phase III testing before it can be considered efficacious, as noted above. Note that for a conventional one-sided superiority design, the likelihood ratios are reversed. Thus the superiority design has greater positive predictive value and smaller negative predictive value than does the futility design.

4. As for any trial design, careful attention must be given to control of the type I error rate, sample size, the power curve, standard errors, and—especially for futility designs—the implications of the critical region demarcating the boundary between rejecting and not rejecting the null hypothesis. Once the superiority criterion α_0 has

been chosen along with the type I error rate α and the power $1 - \beta$ at the design alternative, the sample size n per group is determined. It is important to elicit consensus among the trial leadership concerning the degree of enthusiasm for proceeding to phase III given a marginally non-significant finding of non-futility. This is because, clearly, if the data reveal that $\bar{Y}_1 - \bar{Y}_0$ is close to albeit slightly greater than $\Delta_0 - z_\alpha \sigma \sqrt{2/n}$, the *sample* results will definitely *fall short* of the superiority criterion. This is, of course, the necessary price to pay for statistical uncertainty, but in order to avoid the awkwardness of declaring non-futility when nobody is excited about the trial results—and thus to improve the correspondence between statistical inferences and actual actions—the boundary of the critical region should be located at a value where one would actually be just willing to move forward, other things being equal. Stating this another way, for symmetrical distributions like the normal, there will be power of 50% to declare futility if the true Δ is non-superior and located at the value corresponding to the boundary of the critical region. If that happens to be the true state of nature, then in half of all such cases the futility test will declare non-futility; one should be willing to allow the phase III trial to settle the matter in those cases. The sample size thus should be adequately large not only for control of power at the design alternative but from this logistical perspective as well. An example of very poor planning would be to choose n so small that the boundary of the critical region falls *below* zero. In that case one could be in the extremely awkward position of failing to declaring futility with a sample result in which the experimental treatment fared *worse* than the placebo.

The above considerations suggest an alternative design approach which could have appeal when there is difficulty eliciting or achieving consensus on the “minimal worthwhile improvement” criterion Δ_0 . In such cases it might be easier to identify the boundary of the critical region itself, interpreted as “the *sample* mean efficacy difference at which we would be just indifferent to moving ahead to phase III or not.” Suppose we can elicit such a value; call it Δ^* . To control the type I error at α and achieve power $1 - \beta$ at $\Delta = 0$, we require

$$\Pr[\bar{Y}_1 - \bar{Y}_0 \leq \Delta^* = \Delta_0 - z_\alpha \sigma \sqrt{2/n}] = \Pr[Z \leq (\Delta_0 / \sigma \sqrt{2/n}) - z_\alpha],$$

which implies both that $\Delta_0 = (z_\alpha + z_\beta) \sigma \sqrt{2/n}$ and that $\Delta^* = \Delta_0 - z_\alpha \sigma \sqrt{2/n} = z_\beta \sigma \sqrt{2/n}$. It follows that we require $n = 2z_\beta^2 \sigma^2 / \Delta^{*2}$ patients per group and that the criterion value of superiority should be

$$\Delta_0 = \{1 + (z_\alpha / z_\beta)\} \Delta^*.$$

For example, in the symmetrical case where we choose equal type I and type II error rates of $\alpha = \beta = 0.10$, then the criterion value of superiority should be twice the value of the consensus boundary of the critical region, and the sample size should be $n =$

$2(1.282)^2/(\delta^*/\sigma)^2 = 3.287/(\delta^*/\sigma)^2$, where $(\delta^*/\sigma)^2$ is the standardized critical mean difference. For example, if the consensus value for the go/no-go critical value were one-quarter of a standard deviation, meaning that we would be essentially indifferent to proceeding to phase III with such results, then 53 subjects per group would be required for $\alpha = 0.10$ and 90% power at the design alternative $\delta^* = 0$.

We note that the sample size formula returns the *same* sample size as a conventional two-arm superiority trial with the same structure (one-sided test for given α with equal group sizes and given β at the design alternative $\delta^* = 0$). Thus it is a conceptual error to believe that futility designs always require smaller sample sizes...it is important to add, compared to what? As just mentioned, compared to a one-sided superiority test with the same α and β (albeit with different meanings), there is no reduction in required sample size. The main advantages in this case are those discussed in consequences (1) to (3) above. The sample size savings accrue when comparing a futility test with a conventional phase III superiority design due to one-tailed versus two-tailed testing and the use of larger values of α than $\alpha = 0.05$. The major savings in sample size, however, comes from abandoning the concurrent placebo control arm. We turn to this case next.

3. The single-arm futility design and criticisms of futility testing

The single-arm, historical control futility design with enrollment of calibration controls has been advocated to gain further reductions of the required sample size (see, e.g., [12], [14], [15], and [30]). Here one decides on a value of the true mean response for the experimental treatment, say μ^* , that would represent the minimally worthwhile improvement over standard treatment or placebo. In the NET-PD studies this value was derived from the historical control data of patients receiving either placebo or α -tocopherol in the Deprenyl and Tocopherol Antioxidative Therapy of Parkinsonism (DATATOP) trial [43]. That patient group numbered $n=401$, and the mean increase in the primary endpoint—the total Unified Parkinson's Disease Rating Scale or UPDRS score—was 10.65 units with a standard deviation of 10.4 units. The definition of superiority was then taken as a 30% improvement in the DATATOP increase in UPDRS between baseline and either the time at which there was sufficient disability to warrant symptomatic therapy for Parkinson's disease or 12 months, whichever came first. Thus, μ^* was taken as a mean increase of 0.7×10.65 or $\mu^* = 7.455$ (note that larger values of the UPDRS indicate worse patient outcomes). Then, experimental treatments whose observed mean change in UPDRS was greater than or equal to $\mu^* + t_{n-1;\alpha} s / \sqrt{n}$ were deemed futile, where s^2 is the unbiased sample variance estimator from the experimental treatment and $t_{n-1;\alpha}$ denotes the critical value of Student's t distribution cutting off probability α in the upper tail. The major reduction in sample size comes about because the analogous sample size formula, namely $n = z_{\beta}^2 \sigma^2 / \mu^*$, lacks the factor 2 in the numerator of the two-arm formula since there is only one source of sampling variability, and, obviously, there is only a single group contributing to the sample size. Thus the sample size in the single-arm design is one-fourth that required in the analogous two-arm design (not counting the calibration controls).

Some placebo patients were enrolled in these studies as a “calibration control” group which served two purposes: to allow the study to be conducted in double-blind fashion and to

inform the investigators concerning the reasonableness of the historical control-based criterion of superiority. Difficulties arose because the concurrent placebo controls fared much better than expected based on the historical control; indeed, the concurrent placebo arm would have been deemed non-futile. As Olanow, Wunderle, and Kieburtz [25] pointed out in 2011:

Perhaps the most significant concern with futility studies is their use of historical data to establish the futility threshold. When clinical practices change, such historical data may no longer offer an appropriate threshold. A study for treatment of PD [44] [Parkinson's disease] testing the antioxidant coenzyme Q₁₀ and the neuroimmunophilin ligand GPI-1485 used data from placebo patients in the DATATOP study [43] to set a threshold for the natural change that occurs in UPDRS scores. Although the therapies were both nonfutile based on the preestablished threshold, a validation placebo group run in the same study indicated that the rate of progression of UPDRS scores in a more modern placebo group was less than that in the DATATOP study. Using a recalculated threshold based on the validation group and other more recent clinical trials would have caused both interventions to be rejected as futile [44]. The choice of data used to establish a threshold must therefore be carefully considered, and concurrent validation groups can help to establish the soundness of the threshold. Comparison with a concurrent placebo group is also possible in futility studies, but the sample-size saving is then much less.

Perhaps so, but it may be resources well-spent. Writing in 2013, Yeatts [39] finds the increase in sample size worthwhile and mentions that the NINDS-funded phase II trial of deferoxamine mesylate in intracerebral hemorrhage (Hi-Def in ICH) employs a concurrent control group [33].

Other cautions in the use of futility studies have been raised by several authors. The authors cited above [25] point out that some of the single-arm futility design's strength (including small samples, short time frames, one-sided testing, and alpha set at 0.1) are also weaknesses. Short-term changes may be difficult to identify in progressive movement disorders; short time scales may incorrectly reject therapies that are effective over a longer time scale; and smaller sample sizes may not adequately reveal safety issues.

In discussing the single-arm futility design for Alzheimer's disease, Cummings, Gould, and Zhong [26] caution that placebo groups have often varied substantially from trial to trial "so that careful matching between the selection criteria of the historical controls and the selection criteria for the futility study is critical to insure that accurate inferences can be drawn." Schwid and Cutter [29] write,

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. In neurologic diseases, such as MS [multiple sclerosis], where the disease definition itself has been changed twice in the past 6 years and trial endpoints are subjective, the use of historical controls is problematic.

Cutter and Kappos [42], citing [29], identify several limitations of the single-arm futility design:

They require a knowledge base from which to estimate the design parameters, and there is limited ability to stop the trial before a decent proportion of the planned phase II trial has been observed. Further, shortening the time and limiting the sample size may diminish the safety information that is essential to moving therapy forward in development. Treatments with delayed effects may be missed and this may be a major drawback for neuroprotection. Historic controls are weak, especially in an evolving disease where we have seen the definition of disease altered several times over the past decade, declines in annualized relapse rates in trials, and continuing evolution of treatments. Finally, because the conclusions from these shortened and smaller futility studies may not be sufficient to declare futility, they may increase the costs of the phase III endeavors because of the need for increased sample sizes to detect feasible, but weaker, treatments.

We suspect that the suggestion made above in regard to the two-arm futility study—to base the superiority threshold on a consensus value of the observed performance required to move an experimental treatment forward, instead of on an estimate of current placebo performance based on historical data—would be useful even in the single-arm design. While it wouldn't eliminate the historical control problem altogether because of the partial reliance of the consensus value on historical data, it may help nevertheless to alleviate much angst over “incorrect” choices of the superiority threshold. After all, expert judgment, informed not only by historical data but also recent clinical experience, may well be able to identify a level of performance that would encapsulate and quantify the meaning of “promising”. If the boundary of the critical region is set a little too lax in terms of good performance, the screening program simply becomes less stringent, while if it is set a little too strict, the screening program becomes more stringent. But the goals of the screening program will still be largely met.

4. A misplaced criticism

A criticism of the futility design by Rogatko and Piantadosi has recently appeared [45]. These authors assert that because the null hypothesis of superiority is what investigators “really” want to prove, the futility “design carries significant flaws that investigators should know before implementing.” What flaws? The flaw they allege is that the true efficacy of a treatment declared as non-futile is unknown, so one proceeds to phase III with uncertainty. In a somewhat apples-to-oranges comparison they contrast this with a conventional superiority test, wherein if one rejects the null hypothesis of no efficacy, one at least knows the probability of having committed a type I error. This is a specious argument. Essentially Rogatko and Piantadosi posit that investigators really “should” want to “prove” superiority, thus they should use a conventional superiority design. Well, yes, if you want to go east then don't head west. But up until the very end of their critique, which we discuss below, their arguments ignore the primary purpose of the futility design as a programmatic screening tool. After reviewing the basics of the futility design, they write:

We can now see a pitfall of reversing the null and alternative hypotheses. Suppose the null hypothesis states “the effect of treatment X exceeds standard therapy by 30% or more” and the alternative states “the effect of treatment X is less than a 30% improvement.” If results cause us to reject the null, treatment X will not be developed further and we know exactly the chance that we have discarded a useful therapy: the chance that treatment X exceeded our threshold but was discarded is exactly the α -level of our hypothesis test. In contrast, if we do not reject the null hypothesis, we continue to develop treatment X and are unsure of the true probability that it actually performs below our intended threshold.

Where is the pitfall? That we don’t know the true efficacy of a treatment deemed non-futile as we go forward with further confirmatory testing? That is formally identical to the opposite problem in a conventional test when we *fail* to reject the null hypothesis of no efficacy. There we don’t know the error probability in having failed to proceed, because that is the conventional type II error, which depends on the unknown efficacy of the treatment. Is that then not a “significant flaw” too? In the context of a superiority-style phase II trial, that type II error is arguably more serious than the type I error, which is precisely why the futility design takes superiority as the null, so that it can control the rate of calling a good treatment futile. In a futility test, the maximum type I error rate is certainly known, namely α , so that when we reject the null hypothesis of superiority, we know we have limited the probability of a false declaration of futility as an operating characteristic of the procedure. And yes, when we fail to reject that null, we may not have great confidence in the true efficacy of the treatment, which is precisely why we have to (and *want* to) move on to phase III. This makes perfect sense from the screening perspective; it is only illogical if one believes one should be testing for superiority. In short, Rogatko and Piantadosi’s primary criticism boils down to a simple preference for superiority testing because it has greater positive predictive value, as mentioned above, as opposed to futility testing, wherein we are *not* trying to prove efficacy in phase II but rather screen out unpromising treatments with good negative predictive value. This preference is on exhibit in their rhetorical question, “Why do we want to generate strong evidence that a new therapy is inferior to our benchmark? Broadly speaking, it may be inferentially and ethically more appropriate for us to generate strong evidence that a new treatment is superior rather than inferior.” However, they do not adduce any additional arguments to indicate why it is inferentially more appropriate, nor do they adduce any ethical arguments. We are merely left with their assertion that we should want to prove superiority, and if so, futility testing does not accomplish that goal.

At the conclusion of their article, Rogatko and Piantadosi turn to screening. They argue that we do not “need” futility testing and that a screening goal can be accomplished within the conventional superiority framework by an appropriate setting of type I and type II error rates. Here is the example they give:

For example, suppose we test the null hypothesis that a treatment has a 15% response frequency versus the alternative of a 35% response frequency. After evaluating 35 patients, if ten or more responses are observed, we may reject the null hypothesis. The trial has an actual 2.9% type I error and an 83% power (17%

type II error). If we change the cut point from ten to eight responses, using the same sample size, the trial has a 14% type I error and a 95.8% power (4.2% type II error). The sum of error probabilities is about the same for both arrangements, but the second design protects an effective treatment from premature rejection. Moreover, by increasing power from 83 to 95.8%, we have decreased the false-negative error probability by a factor of four.

It may not be immediately apparent, but here the authors re-discover the futility design! In order to achieve desirable operating characteristics for a screening program in terms of false positive and false negative rates, they need to set their α to 0.15 at the null response rate of 15% and their power to 0.95 at the design alternative of a 35% response rate. To achieve those operating characteristics the authors require a sample size of 35 subjects and use a critical value of at least 8 treatment responses. (The precise type I error probability at response rate 15% is then 0.144 and the precise power at the design alternative is 0.958.) Though the design is “conventional” the pair of choices (0.15, 0.95) for alpha and power would likely be viewed as unconventional for a superiority design. *But operationally this design is precisely equivalent to a futility design* in which the null hypothesis of superiority posits a minimum worthwhile response rate of 35% (or greater) and the rejection region consists of 7 or fewer treatment responses: the two tests lead to *exactly* the same accept/reject decisions. The precise type I error probability of the futility test is the same 0.042 as the type II error probability of the conventional test, and the power of the futility test at the futility design alternative of a 15% response rate is the same 0.856 as the complement of the type I error of the conventional test. Choices of 0.05 and 0.85 for type I error and power, respectively, seem so much more natural to this reviewer. Moreover, if the two tests are equivalent because they lead to exactly the same decisions, where are the “significant flaws that investigators should know before implementing”? And if the experimental observations happened to contain exactly 8 responses, would we really be less “unsure of the true probability that the treatment performs below our intended threshold” because the result came from a conventional design as opposed to an *equivalent* futility design? No.

Rogatko and Piantadosi’s concluding example is even more startling.

Reversal of the null and alternative hypotheses is not required either to construct an optimistic pipeline, demonstrate futility/nonsuperiority, or to align type I and II errors with their consequences. For example, we could conventionally construct the null hypothesis to be “the effect of treatment X is less than a 30% improvement over standard,” and the alternative hypothesis to be “the effect of treatment X exceeds standard therapy by 30% or more.” This is a clean nonsuperiority hypothesis. If we want an optimistic pipeline that favors moving therapies forward, we might set the type I error at 10% or even 20%. It could be a very serious error to miss a treatment that actually represented a 50% improvement, in which case we might want the type II error for that alternative to be as small as say 5%. Then weak evidence makes it difficult to advance a treatment, and strong evidence is likely to advance a treatment when it is actually good. Thus the properties that we admire in a “futility” design can be achieved using conventional ideas.

Here the suggestion seems to be that testing the null hypothesis of no more than a 30% improvement over standard treatment versus the alternative hypothesis of more than a 30% improvement can or should be conducted in lieu of a futility test (where now 30% is the minimally worthwhile improvement and the pre-specified definition of superiority). But their suggestion has an obvious consequence—the data would have to demonstrate significantly *better* than a 30% improvement in order for the conventional null hypothesis to be rejected and for development to go forward (at whatever α level they choose). This is in contrast to the futility test criterion, for which the treatment only needs to produce data that are not significantly *worse* than a 30% improvement. This difference in required performance seems rather important.

5. Conclusion

There is no compelling reason to limit the arsenal of developmental trial designs to the conventional superiority test. The futility design has a useful role to play in an institutional screening program to weed out unpromising treatment in an environment where patients and resources are precious and testing every candidate treatment in the pipeline with a phase III trial is unsustainable. Yes, great care must be taken to heed the limitations of the method and avoid the pitfalls, but this is surely true of any experimental undertaking.

Acknowledgments

This paper was supported in part by NIH Grant P30-MH43520 to the HIV Center for Clinical and Behavioral Studies at Columbia University and the NY State Psychiatric Institute.

References

1. Gehan EA. The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. *J Chronic Dis.* 1961; 13:346–53. [PubMed: 13704181]
2. Herson J. Predictive probability early termination plans for phase II clinical trials. *Biometrics.* 1979; 35:775–83. [PubMed: 526523]
3. Fleming TR. One-sample multiple testing procedure for phase II clinical trials. *Biometrics.* 1982; 38:143–51. [PubMed: 7082756]
4. Simon R. Optimal two-stage designs for phase II clinical trials. *Control Clin Trials.* 1989; 10:1–10. [PubMed: 2702835]
5. Betensky R. Alternative derivations of a rule for early stopping in favor of H_0 . *The American Statistician.* 2000; 54:35–9.
6. Ravina BM, Fagan S, Hart RG, Murphy D, Marler JR. Neuroprotective agents for clinical trials in Parkinson's disease: a systematic assessment. *Neurology.* 2003; 60:1234–40. [PubMed: 12707423]
7. Biglan KM, Ravina B. Neuroprotection in Parkinson's disease: an elusive goal. *Semin Neurol.* 2007; 27:106–112.
8. Voss T, Ravina B. Neuroprotection in Parkinson's disease: myth or reality? *Curr Neurol Neurosci Rep.* 2008; 8(4):304–9. [PubMed: 18590614]
9. Hart RG, Pearce LA, Ravina BM, Yalcho TC, Marler JR. Neuroprotection trials in Parkinson's disease: systematic review. *Mov Disord.* 2009; 24(5):647–54. [PubMed: 19117366]
10. The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. . Tissue plasminogen activator for acute ischemic stroke. *N Engl J Med.* 1995; 333:1581–8. [PubMed: 7477192]
11. Beal MF. Neuroprotective effects of creatine. *Amino Acids.* 2011; 40(5):1305–13. [PubMed: 21448659]

12. Herson J, Carter SK. Calibrated phase II clinical trials in oncology. *Statistics in Medicine*. 1986; 5:441–447. [PubMed: 3786998]
13. Elm JJ, Goetz CG, Ravina B, Shannon K, Wooten GF, Tanner CM, et al. for the NET-PD Investigators. A responsive outcome for Parkinson's disease neuroprotection futility studies. *Ann Neurol*. 2005; 57(2):197–203. [PubMed: 15668964]
14. Tilley BC, Palesch YY, Kieburtz K, Ravina B, Huang P, Elm JJ, et al. for the NET-PD Investigators. Optimizing the ongoing search for new treatments for Parkinson disease: using futility designs. *Neurology*. 2006; 66(5):628–33. [PubMed: 16534099]
15. NINDS NET-PD Investigators. A randomized, double-blind, futility clinical trial of creatine and minocycline in early Parkinson disease. *Neurology*. 2006; 66(5):664–71. [PubMed: 16481597]
16. NINDS NET-PD Investigators. A pilot clinical trial of creatine and minocycline in early Parkinson disease: 18-month results. *Clin Neuropharmacol*. 2008; 31(3):141–50. [PubMed: 18520981]
17. Palesch YY, Tilley BC, Sackett DL, Johnston KC, Woolson R. Applying a phase II futility study design to therapeutic stroke trials. *Stroke*. 2005; 36(11):2410–4. [PubMed: 16224086]
18. Pavlakis SG, Sacco R, Levine SR, Meschia JF, Palesch Y, Tilley BC, et al. Lessons from adult stroke trials. *Pediatr Neurol*. 2006; 34(6):446–9. [PubMed: 16765822]
19. Tilley BC, Galpern WR. Screening potential therapies: lessons learned from new paradigms used in Parkinson disease. *Stroke*. 2007; 38(2 Suppl):800–3. [PubMed: 17261742]
20. Palesch YY, Tilley BC. An efficient multi-stage, single-arm Phase II futility design for ALS. *Amyotroph Lateral Scler & Other Motor Neuron Disord*. 2004; 5 (Suppl 1):55–6.
21. Czaplinski A, Haverkamp LJ, Yen AA, Simpson EP, Lai EC, Appel SH. The value of database controls in pilot or futility studies in ALS. *Neurology*. 2006; 67(10):1827–32. [PubMed: 17130417]
22. Schoenfeld DA, Cudkowicz M. Design of phase II ALS clinical trials. *Amyotroph Lateral Scler*. 2008; 9(1):16–23. [PubMed: 18273715]
23. Chiò A, Logroscino G, Hardiman O, Swinger R, Mitchell D, Beghi E, Traynor BG. Eurals Consortium. Prognostic factors in ALS: A critical review. *Amyotroph Lateral Scler*. 2009; 10(5–6):310–23. [PubMed: 19922118]
24. The Huntington Study Group DOMINO Investigators. A futility study of minocycline in Huntington's disease. *Mov Disord*. 2010; 25(13):2219–24. [PubMed: 20721920]
25. Olanow CW, Wunderle KB, Kieburtz K. Milestones in movement disorders clinical trials: advances and landmark studies. *Mov Disord*. 2011; 26(6):1003–14. [PubMed: 21626545]
26. Cummings J, Gould H, Zhong K. Advances in designs for Alzheimer's disease clinical trials. *Am J Neurodegener Dis*. 2012; 1(3):205–16. [PubMed: 23383393]
27. Levin B. The utility of futility. *Stroke*. 2005; 36:2331–2. [PubMed: 16224096]
28. Kieburtz K. Issues in neuroprotection clinical trials in Parkinson's disease. *Neurology*. 2006; 66(10 Suppl 4):S50–7. [PubMed: 16717252]
29. Schwid SR, Cutter GR. Futility studies: spending a little to save a lot. *Neurology*. 2006; 66(5):626–7. [PubMed: 16534098]
30. Ravina, B.; Palesch, Y. *Progress in Neurotherapeutics and Neuropsychopharmacology*. Vol. 2. Cambridge Univ Press; 2007. The Phase II Futility Clinical Trial Design; p. 27–38.
31. Hung AY, Schwarzschild MA. Clinical trials for neuroprotection in Parkinson's disease: overcoming angst and futility? *Curr Opin Neurol*. 2007; 20(4):477–83. [PubMed: 17620885]
32. Elm JJ. for the NINDS NET-PD Investigators. Design innovations and baseline findings in a long-term Parkinson's trial: the National Institute of Neurological Disorders and Stroke Exploratory Trials in Parkinson's Disease Long-Term Study-1. *Mov Disord*. 2012; 27(12):1513–21. [PubMed: 23079770]
33. Yeatts SD, Palesch YY, Moy CS, Selim M. High dose deferoxamine in intracerebral hemorrhage (HI-DEF) trial: rationale, design, and methods. *Neurocritical Care*. 2013; 19(2):257–66. [PubMed: 23943316]
34. Lewis RA, McDermott MP, Herrmann DN, Hoke A, Clawson LL, Siskind C, et al. High-dosage ascorbic acid treatment in Charcot-Marie-Tooth disease type 1A: Results of a randomized, double-masked, controlled trial. *JAMA Neurology*. 2013; 70(8):981–7. [PubMed: 23797954]

35. Moore CG, Schenkman M, Kohrt WM, Delitto A, Hall DA, Corcos D. Study in Parkinson disease of exercise (SPARX): translating high-intensity exercise from animals to humans. *Contemp Clin Trials*. 2013; 36(1):90–8. [PubMed: 23770108]
36. Levy G, Kaufmann P, Buchsbaum R, Montes J, Barsdorf A, Arbing R, et al. A two-stage design for a phase II clinical trial of coenzyme Q10 in ALS. *Neurology*. 2006; 66:660–3. [PubMed: 16534103]
37. Kaufmann P, Thompson JLP, Levy G, Buchsbaum R, Shefner J, Krivickas LS, et al. for the QALS Study Group. Phase II trial of CoQ10 for ALS finds insufficient evidence to justify phase III. *Annals of Neurology*. 2009; 66:235–44. [PubMed: 19743457]
38. Levin, B. Selection and futility designs. In: Ravina, B.; Cummings, J.; McDermott, MP.; Poole, M., editors. Chapter 8 in: *Clinical Trials in Neurology*. Cambridge: Cambridge University Press; 2012.
39. Yeatts SD. Novel methodologic approaches to phase I, II, and III trials. *Stroke*. 2013; 44(6 Suppl 1):S116–8. [PubMed: 23709704]
40. Sackett DL. Clinician-trialist rounds: 18. Should young (and old!) clinician-trialists perform single-arm Phase II futility trials? *Clinical Trials*. 2013; 10(6):987–9. [PubMed: 24085773]
41. Koch MW, Korngut L, Patry DG, Agha-Khani Y, White C, Sarna JR, et al. The promise of futility trials in neurological diseases. *Nat Rev Neurol*. 2015 epub ahead of print. 10.1038/nrneurol.2015.34
42. Cutter G, Kappos L. Clinical trials in multiple sclerosis. Chap 20 in: *Handbook of Clinical Neurology*. 2014; 122:445–53. 3rd series.
43. The Parkinson Study Group. Effects of tocopherol and deprenyl on the progression of disability in early Parkinson's disease. *New Engl J Med*. 1993; 328:176–83. [PubMed: 8417384]
44. NINDS NET-PD Investigators. A randomized clinical trial of coenzyme Q10 and GPI-1485 in early Parkinson disease. *Neurology*. 2006; 66:664–71. [PubMed: 16481597]
45. Rogatko A, Piantadosi S. Problems with constructing tests to accept the null hypothesis. Chap 4 in: *Interdisciplinary Bayesian Statistics*, Springer Proc in Math & Stat. 2015; 118:49–54.