Therapeutic Innovation & Regulatory Science (2023) 57:515–520
https://doi.org/10.1007/s43441-023-00499-0

DIA

REVIEW

Challenges in Implementing Futility Schemes, with Reference to Aducanumab

Paul Gallo[1] · **Satrajit Roychoudhury**[2]

Received: 2 November 2022 / Accepted: 24 January 2023 / Published online: 3 February 2023

Abstract

Stopping an ongoing clinical trial based on an interim analysis that shows poor outcomes, often referred to as a judgment of "futility", is a familiar feature in current clinical trials practice. Interim data can be misleading, and the implications of prematurely terminating a trial that should not stop are severe. It is thus critical that designs allowing futility stopping be planned and implemented carefully and cautiously. A recent Phase III development program for aducanumab in Alzheimer's disease was halted based on a pre-defined futility guideline, yet based upon updated data and closer examination, the terminated studies became the basis for a regulatory submission. Not surprisingly, this situation generated much controversy and discussion. It provides a good basis for illustrating important principles governing the planning and implementation of futility schemes.

Keywords Futility analysis · Interim analysis · Conditional power · Alzheimer's disease

Introduction

Futility analyses, that is, examinations of interim data with potential to allow stopping a clinical trial prior to its planned end if it seems it will not meet its efficacy objectives, are a common feature in current clinical trial practice. These are a specific application of *interim analysis*, that is, analysis of data obtained thus far while a trial is underway. Other motivations for interim analyses include stopping for positive efficacy if a standard of proof has been met, or implementing a change to some aspect of the study, such as its sample size, according to a pre-specified adaptation plan. In this paper we focus specifically on futility analyses. A recent highly publicized case involves the aducanumab program in Alzheimer's disease (AD) [1], in which two Phase III outcome studies were running concurrently, and it was announced that these were being stopped for futility; subsequent data obtained during shutdown were more favorable, and in fact became the main basis for a regulatory submission. Opinions were subsequently expressed that the program should not have been stopped [2]. Additionally, viewpoints were expressed, with a focus on AD trials and this program in particular, to the effect that futility analyses themselves are often inappropriate and should be avoided [3]; in a public forum, futility analyses were referred to as a "statistical misadventure" [4].

On the contrary, we feel strongly that futility analyses, properly applied, play an important and necessary role in clinical development. It would in many circumstances be highly unethical to disallow them and doing so runs unnecessary risks of harm to patients. Additionally, they often enhance the efficiency of clinical development programs. It is critically important that stopping guidelines be defined carefully to limit the chances of incorrect decisions, and that plans are implemented cautiously, with decision makers considering the totality of information available to them when making a judgment. This should generally allow flexibility in decision making relative to pre-stated stopping guidelines. Clinical and strategic considerations are fundamental, with statistical methods and computations playing an important supportive role. The aducanumab situation provides an excellent basis to illustrate a number of relevant principles.

✉ Satrajit Roychoudhury
satrajit.roychoudhury@pfizer.com

[1] Independent Consultant, East Hanover, NJ, USA

[2] Pfizer Inc, New York, NY, USA

Aducanumab in AD Background

Two concurrent and essentially identical long-term Phase III outcome studies of aducanumab for prevention of progression of Alzheimer's Disease, ENGAGE and EMERGE, were initiated in 2015 [1]. The trials enrolled a total of 3285 patients across 20 countries; patients were randomized to either placebo or one of two dose regimens of aducanumab. The studies' primary objective was to evaluate the efficacy of aducanumab in reducing clinical decline, and the primary outcome was change from baseline in the *Clinical Dementia Rating Scale – Sum of Boxes* (CDR-SB) at Week 78.

The clinical context is clearly of very high current importance; we note that treatments in the same therapeutic class had not previously been able to demonstrate convincing evidence of effect. A futility analysis was planned to be performed after about half the participants in each study had reached the primary endpoint assessment point. *Conditional power* (CP) was to be computed separately for each study under an assumption that the effect governing the remainder of the trial was equal to a pooled estimate obtained from combined interim data from both studies. According to the defined threshold, a study could be considered futile if this CP value was below 20%.

At the point of the futility analysis, using data available as of December 2018, the primary outcome effect estimate for the aducanumab high dose was notably higher in the EMERGE study, with an 18% difference favoring aducanumab, than in ENGAGE, which in fact showed a negative signal, a 15% disadvantage versus placebo. CP values computed as pre-defined (that is, assuming the combined estimate from the 2 studies) were 0% for ENGAGE and 12% for EMERGE. Thus, both studies reached the pre-defined threshold that potentially allowed stopping. The studies' Data Monitoring Committee (DMC) recommended the trials be terminated; sponsor leadership accepted the recommendation and termination was announced on March 19, 2019.

Subsequently, the sponsor analyzed a dataset augmented with additional data available up to the time of the March 2019 announcement. Results in both studies had improved somewhat from the December 2018 results (22% advantage in EMERGE, only a 2% detriment in ENGAGE). After further examination, the sponsor felt that these studies could in fact be a basis for a regulatory submission for approval of aducanumab, along with other data from the clinical program. Factoring into the decision was a rationale for the apparent difference between the results in the two studies. This included dosing regimen changes reflected in protocol amendments which allowed more aducanumab patients to titrate to a higher dose; this might have a tendency to make later results more favorable, and these changes affected more patients in EMERGE than in ENGAGE. Ultimately, the sponsor and FDA jointly decided that the early termination of the studies "did not compromise the ability to interpret the results" [2]. FDA requested that CP be re-computed for each study assuming study-specific effect estimates rather than the pooled estimate; results were 0% for ENGAGE and 59% for EMERGE. Thus, eventual success in ENGAGE, had it continued, seemed extremely unlikely, while a positive outcome in EMERGE was quite plausible. At a Type C meeting in June 2019, FDA stated that "it would have been more appropriate if futility had not been declared for those studies" [2]. Ultimately, FDA granted approval of aducanumab in June 2021, though both the sponsor characterization of the degree of evidence and the approval decision met with controversy in certain quarters [5, 6].

It is not our intent to comment on the submission and its outcome; our focus is on the futility issue, in particular on the setting of futility boundaries and the decision process. Nor are we in position to make a definitive judgement, with hindsight, on the futility decision itself; we don't have the full perspectives of the DMC and sponsor program experts, nor knowledge of all information available to them. Nevertheless, we feel the situation provides an excellent basis to raise several points related to proper determination and implementation of futility plans.

Futility Statistical Background

We briefly describe statistical approaches commonly employed to assist in futility evaluations; these generally involve pre-specified boundary thresholds for a study's main outcome. As the complexity of a treatment's effects may not be adequately captured in a single measure, it's usually desirable that decision makers consider all information available at a decision point, and not base decisions solely and rigidly on the basis of a single computed quantity, as we discuss later. Nevertheless, pre-specified boundaries play an important role, describing outcomes where futility stopping may well be indicated, pending a deeper look at available information.

Commonly used tools include *conditional power* and *predictive probability*. Conditional power for an endpoint describes the chance that its final results will reach a defined success threshold, based on the data observed thus far, and some specific effect size assumed to govern the remainder of the data to be collected if the trial continues to completion. Two values commonly assumed are the effect that was hypothesized for trial design, and an effect equal to the interim estimate thus far. These should *not* be casually interpreted as chances of trial success, as of course the true effect

size is unknown. When conditioning on the study hypothesis, in a situation where the estimate is weak so that futility might be warranted, we could view such an assumption as quite optimistic, because the data thus far is suggesting that the hypothesized benefit is *not* true. Conditioning on the interim estimate may seem more realistic in this sense, but high variability of interim estimates also limits interpretability of this quantity as a chance of success.

Predictive probability provides a measure that can be viewed much more broadly as providing a valid probability statement. Simplistically, this might be described as a weighted average of conditional probabilities over effect sizes weighted according to their degree of consistency with the interim estimate. Again though, we may keep in mind the variability of the estimate—we are usually averaging over a broad range of effect sizes, some much better or poorer than the estimate, only one among which is true. This approach can be extended by identifying a prior distribution of effect sizes, which is then updated based upon the interim data.

Beta spending functions are also sometimes used. These are an analog of alpha spending functions that are sometimes used for early stopping for success. Beta spending can be viewed as allocating a study's false negative rate across interim and final analyses.

We now extend to considering futility in two ongoing trials, sufficiently similar in endpoint, population, and conduct, so that there is no advance expectation that treatment effects should meaningfully differ, such as in the aducanumab program. Deng et al. [7] described an extension of a CP approach that uses combined data from both trials to determine a single estimate assumed to govern the future data in both trials. When trials are expected to reflect similar effects, this is sensible because the combined estimate is much more precise than single trial estimates, and this approach thus adds efficiency.

When using such an approach, it may be kept in mind that sometimes trials expected to reflect a common effect have shown differences, and the reasons may not initially be understood. Approaches using a Bayesian hierarchical model can be considered to limit the influence of "borrowing" results across trials. This can increase efficiency by borrowing information when indicated, but also provide a greater degree of robustness by accommodating possible discrepancy across trials. These have been described and illustrated in [8].

Regardless of the statistical approach used, it is of course important to consider how to set criteria. For any method, this is not a "one size fits all" process. One can, for example, always increase the chance of stopping trials that will not succeed by defining criteria that are more *aggressive* (that is, corresponding to a stronger effect estimate, thus making futility stopping easier); however, this also tends to increase the chance of stopping trials for futility that might succeed if continued, thus resulting in a loss of power. It is important to consider the tradeoffs between the types of errors that can be made, appropriate to the practical realities of particular studies. Considerations for doing so within a single trial were described in [9].

General Futility Considerations

Stopping an ongoing clinical trial, whether because it has reached its efficacy objectives, or will not, or due to safety concerns, or because of other issues that make it infeasible to continue, is of course a major decision, and generally irrevocable. Such decisions should not be taken without an extremely compelling rationale. We fully agree with a statement in a viewpoint article focusing on the aducanumab example that "the cost of making a wrong decision to stop a trial can be staggering" [3]. Decisions should be made on a situation-specific basis, taking into account considerations such as the following: have the questions of importance been answered sufficiently?; would the trial provide meaningful further information to the clinical community by continuing?; do the ethics of the situation allow continuation?; etc.

Advance specification of outcomes for which futility stopping may be justified is an important aspect of trial planning, and in fact, a fundamental aspect of trial design. Planning should include careful development of the quantitative basis on which stopping a trial may be justified. We emphasize two aspects: (1) the importance of setting futility thresholds *cautiously,* i.e., boundaries that are not unduly aggressive; and (2) pre-specified thresholds should be interpreted as defining outcomes for which stopping *might be* appropriate, but the totality of information available to decision makers can be quite complex, and could lead in some cases to over-riding a purely algorithmic decision. Performing thorough simulations at the design stage, to consider the possible impact of violations of assumptions that are made, can be very helpful both in setting criteria, and in evaluating actions at decision points.

Interim effect estimates are inherently and unavoidably imprecise, and do not necessarily accurately convey a treatment's true effect, nor predict well a study's final results. As a simple illustrative example, consider a two-arm trial with a normally distributed endpoint, designed conventionally to have 90% power to detect an effect of 10 units (a mean difference, for example). If trial assumptions hold, then when half of the planned amount of data have been obtained, a conventional 95% confidence interval has a length of about 17 units. Thus, for example, if the interim estimate was half as large as the design assumption, that is, 5 units, this interval would be (− 3.5, 13.5), which would rule out neither a zero effect nor the design hypothesis of 10. In general, a weakly positive interim result by no

means excludes the chance that the true effect is as large as hypothesized or that a study might yet succeed. In predefining futility thresholds, this imprecision must be taken into account to appropriately limit the chance of stopping a study that might yet show meaningful results. Speaking very broadly, sensible futility bounds usually correspond to poor interim outcomes, sometimes even favoring the control during the very early part of a trial. Stopping trials where there is a mildly positive interim estimate, even if disappointing relative to a study's hypothesis, often risks substantial loss of power.

Beyond the setting of thresholds, a number of aspects should be carefully considered in deciding whether a trial that has reached a threshold should indeed be stopped. The use of *non-binding* futility boundaries, quite common in current practice, allows this. The term "non-binding" applies to both its semantic interpretation and its statistical definition: reaching such a threshold does not rigidly mandate stopping, and final efficacy thresholds are not loosened so that a decision to continue does not inflate a trial's allowed false positive rate. While futility rules are most commonly defined based on a study's primary outcome, a DMC or other decision maker will usually be looking at far more information, including many other outcomes. Issues to be considered, beyond simply the inherent imprecision of interim estimates, might include the following:

- Might there be some "drift" in the trial population as a study proceeds, so that early data might not be fully reflective of later data? (This can arise in a number of ways, for example, more severe patients or more experienced investigators might enroll earlier, distribution of patients across regions might change, investigators might gain experience in administering a complex therapy or reacting to certain adverse events, etc. The dosing regimen change in the aducanumab studies also falls into this category).
- Are there inconsistencies across outcomes, for example, more favorable signals in secondary endpoints or early markers or assessments at other timepoints, that suggest that the primary outcome results might yet improve?
- Are there treatment group imbalances for important prognostic factors that were not accounted for in the main analysis?
- Are there signals of different effects in subsets of the trial population, perhaps stronger or weaker in certain risk groups or regions, suggesting that more data are needed to understand better? Could there be meaningful benefit in some important subgroup? Should more data be obtained in lightly represented subgroups so that such questions can be better addressed?
- Particularly for time-to-event outcomes, are there outcome patterns over time or clinical rationale that allow possible eventual improvement, or at least justify obtaining more data?

An important point to keep in mind is that in many situations, continuing a trial beyond the point where it has reached a futility threshold need not necessarily mean that it must continue to its initially planned end. Uncertainties such as mentioned above might justify continuing to a later re-evaluation point where data are more mature, allowing more certainty in the futility interpretation, or better resolution of ambiguities, and it is advisable that this be allowed in the plan where applicable.

As a simple illustration motivated by the aducanumab program, we contrast different futility schemes using 20% CP thresholds assuming an effect estimated from data at a single interim analysis performed when half the planned data are available. We consider trials designed in a conventional manner with 90% power to detect a specified meaningful effect size for a normally distributed response with constant effect throughout. If we consider a single trial and condition on its interim estimate, then even if the true treatment effect is as large as hoped (that is, equal to the design hypothesis), the 20% CP boundary would be reached about 9.2% of the time, risking an erroneous decision to stop for futility. We would characterize such a threshold as aggressive, introducing a meaningful loss in power, and would not often recommend such a boundary.

Next, we extend to two identical concurrent trials, both with true effect equal to the hypothesized value, and as proposed in [7] we compute CP in each assuming the pooled estimate (this was essentially the design plan in the aducanumab program). This use of combined data increases the efficiency as desired: in each trial the chance of misleadingly reaching the futility threshold has been cut in half, to 4.6%, yielding a more tolerable risk strategy.

But now consider a situation where in one trial the underlying true effect is equal to the hypothesized effect, and in the other, the null hypothesis is true, that is, zero effect (perhaps an unusual degree of disparity, but we note that this is fully consistent with what was observed in the aducanumab program). If we condition on the pooled estimate, the chance that the trial in which the true effect is strong would reach the 20% CP threshold would now nearly double to 17%, a much riskier strategy. A strong suggestion of meaningful violation of a key assumption – namely, common effect in the two trials—would likely be apparent to decision makers, raising questions as to whether the pre-defined threshold should be applied in the manner originally planned.

Considerations in the Aducanumab Program

We return now more specifically to the aducanumab program. As mentioned previously, the trials were near their midpoint on the primary assessment, CP values assuming the pooled estimate were 0% and 12% for ENGAGE and EMERGE, respectively, and 0% and 59% when assuming the trial-specific estimates. Thus, the trials showed very different trends. We note in particular that if CP assuming an effect equal to a study's interim estimate exceeds 50%, this almost always means that *the study will be successful if the current trend is maintained through the end of the trial*. While there is no guarantee that this will be the case, this provides perspective on how strongly the EMERGE results were trending towards success at the time the time the study was stopped as 'futile'.

At that point, what might one have expected if the trials continued? The weak effect seen in ENGAGE could legitimately lead one to surmise that the strong estimate in EMERGE might possibly be an over-estimate, somewhat of a random-high (and conversely, the ENGAGE estimate a random-low). But what if this pattern reflected to some degree a true difference between the studies in some aspect of trial conduct not yet understood? Many clinical programs show quantitatively different results across trials, and sometimes the reasons are not initially understood. Might there have been some as-yet-unrealized difference between the trials leading to the divergent signals, and might the strong result in EMERGE be indicative of a meaningful effect of the treatment? This possibility seems exactly to have been acknowledged by the sponsor in the evaluation of the final data just a few months after the futility shutdown. Such a pattern might be viewed as precisely the type of ambiguity we alluded to earlier as warranting caution in a decision or more data to better understand.

To the extent that the interim results reflected a true difference between the studies, this would be a potentially impactful violation of a key assumption of the pre-defined methodology, namely, similarity of study effects. This could make the pre-specified threshold dangerously aggressive for EMERGE, as per the example described in the previous section. And the dosing regimen change mentioned previously could make the future data less reflective of early data, which could be viewed as another violation of a key assumption: a CP calculation that assumes the interim effect governs the remainder of the trial implicitly seems to assume constancy of effect throughout the trial. A shift towards stronger outcomes would not be adequately reflected in the calculation as defined. Thus, there would have been reasons to question the behavior of the pre-defined methodology, which might behave very differently than intended at the design stage.

Discussion and Recommendations

Formal futility analyses certainly are not needed in all trials. Trial aspects which might argue against their use include short trial duration; symptomatic endpoint with minimal safety concerns; short-term treatment administration with longer-term follow-up to endpoint (vaccine trials, for example). Aisen and Raman [3] present arguments why futility analyses may be particularly problematic in AD trials.

With regard to opinions that futility analyses are inherently flawed or risky, and generally to be avoided, we strongly disagree. We would counter that any trial design or conduct decision entails risks, and decisions should be made in a manner that understands and takes into account the risk aspects, and the tradeoffs among different strategies. Like any tool in the trial practitioner's arsenal, futility analyses can be applied well or not well. While limitations of interim data are apparent, this doesn't mean that there aren't situations where it's clear that it is in no parties' interests to continue. Very prominently, this can involve ethics: for an investigational product, if interim data make definitively clear that a treatment is not viable, then how can we ask patients to continue contributing their participation, simply because we did not sufficiently realize at the design stage that lack of viability? Perhaps for novel therapies we may be subjecting subjects to long-term safety risks that are *as yet unknown*. Additionally, stopping a futile trial may lead to advantageous changes to a clinical program based on what has been learned. Futility schemes may allow sponsors to undertake long-term resource-intensive trials for potential breakthrough therapies, in the presence of some uncertainties, because the commitment is not necessarily to run the trial to completion. Saying that futility analyses should not be performed in a given trial implies that there is no possible efficacy pattern, no matter how poor, that could ever justify discontinuation.

Aducanumab was an unusual situation that illustrates a number of points relevant to implementing futility analyses. The disparity between the trials' interim results might seem exactly the type of anomaly that could justify continuing beyond the point where a futility boundary had been reached—if not to the end of the trial, then at least somewhat further, to see whether the pattern was maintained. If in fact the EMERGE interim results did reflect that aducanumab was an important therapy, and if there were extenuating circumstances explaining the weaker results in ENGAGE, then the negative implications of stopping would be severe. As previously mentioned, this seemed to be the way the sponsor viewed the situation just a short time later. By then it was too late to continue the studies, which might well have provided more convincing answers to important open questions.

The futility decision that was made seems quite unique in our experience. At risk of oversimplifying, we might describe the usual futility paradigm as follows: stopping may be indicated if interim results are very weak, and there is no plausible reason to believe that they will substantially improve. In this instance, that paradigm was, in effect, reversed: EMERGE showed favorable results, to an extent that the study was trending well towards stand-alone success. Nevertheless, this study was halted based on a judgment of futility.

Implications of incorrectly stopping a trial are so serious as to argue in the direction of extreme caution. This applies firstly to defining criteria, which should be set conservatively. But secondly, criteria should be viewed as 'flags' indicating that stopping might well be indicated, but decisions should be based on careful evaluation by decision makers of all relevant available information, and whether there are extenuating circumstances that justify continuing even if thresholds are reached. We re-emphasize that decisions need not be bound by pre-stated thresholds if full details of a situation suggest otherwise.

Criticisms of futility analysis broadly based on the aducanumab situation, and questions as to whether they should be used less frequently, might better focus on the specific implementation and decision process here; that is, whether the thresholds were set reasonably, and whether potentially meaningful ambiguities in the data were sufficiently considered before the decision was taken. We note that key assumptions of the methodology seemed violated, possibly rendering the initially defined threshold inappropriate. We do not accept the view that this situation raises questions about the merit of futility analyses in general. If a situation such as the aducanumab program serves to highlight the need to be more cautious in setting criteria and making decisions, and that pre-specified criteria can be over-ridden for sound reasons, that seems to us a good thing. But if such a case were to widely discourage the usage of futility analyses, we believe that would be to the detriment of clinical trial practice and would put patients at risk unnecessarily.

Data availability

Data sharing is not applicable to this article as no datasets were generated or analysed in this article.

References

1. Budd Haeberlein S, Aisen P, Barkhof F, et al. Two randomized Phase 3 studies of aducanumab in early Alzheimer's Disease. J Prev Alzheimers Dis. 2022;9(2):197–210.
2. Meeting of the Peripheral and Central Nervous System Drugs Advisory Committee Meeting, 2020. https://www.fda.gov/advisory-committees/advisory-committee-calendar/november-6-2020-meeting-peripheral-and-central-nervous-system-drugs-advisory-committee-meeting.
3. Aisen PS, Raman R. Viewpoint: futility analyses in Alzheimer's disease (AD) clinical trials: a risky business. J Prev Alzheimers Dis. 2020;7(3):195–6.
4. Aisen PS. And now what? Where are we headed in AD Drug development? Keynote presentation at clinical trials on Alzheimer's disease conference, San Diego CA, December 2019. https://www.ctad-alzheimer.com/files/files/Paul%20Aisen%20CTAD%202019.pdf
5. Knopman DS, Jones DT, Greicius MD. Failure to demonstrate efficacy of aducanumab: an analysis of the EMERGE and ENGAGE trials as reported by Biogen, December 2019. Alzheimers Dement. 2021;17(4):696–701.
6. Schneider LS. Editorial: aducanumab trials EMERGE but don't ENGAGE. J Prev Alzheimer's Dis. 2022;9(2):193–6.
7. Deng Q, Zhang Y-Y, Roy D, et al. Superiority of combining two independent trials in interim futility analysis. Stat Methods Med Res. 2020;29(2):522–40.
8. Neuenschwander B, Roychoudhury S, Schmidli H. On the use of co-data in clinical trials. Stat Biopharm Res. 2016;8(3):345–54.
9. Gallo P, Mao L, Shih VH. Alternative views on setting clinical trial futility criteria. J Biopharm Stat. 2014;24(5):976–93.