Futility Analyses in Confirmatory Clinical Trials – Methods and Procedures

Paul Gallo
Novartis Pharmaceutical Corporation

Shanghai Biostatistics Forum
April 22, 2019

1
Introduction:
Futility Analysis in Clinical Trials

In the news!

Larg
Hun
anal

CREST-E
terminat
no realis
that cre

Thi
Inh
E

#HEALTH NEWS FEBRUARY 14, 2017 / 6:12 PM / 6 MONTHS AGO

Fo
dr
Lil
lar
sai
tha
the

Lil
pr
stu
to

ALZFORUM
NETWORKING FOR A CURE

End of the RoAD for Gantenerumab? Roche Declares Prodromal Alzheimer's Trial Futile

19 Dec 2014 Roche today announced that it is stopping the Phase 3 SCarlet RoAD trial of its investigational antibody gantenerumab in prodromal Alzheimer's disease. According to a company press release, the decision follows the recommendation its independent data monitoring committee made based on a preplanned futility analysis during the trial. The press release noted that no new safety signal arose, implying to experts in the field that the decision to end the trial was made because of either efficacy or target engagement.

Futility

Our topic / concept / scope:

- Addressing the possibility of terminating trials prior to the planned maximum {duration / amount of information}, because results on the main endpoint are weak / disappointing

- {Terminology is not entirely consistent throughout the literature (e.g., “inefficacy”)}

Excluded

Our setting does *not* include stopping a trial for operational reasons that make it infeasible to conduct the trial as planned, e.g.:

- difficulty to enroll
- new emerging external information that undermines the trial's rationale, or makes it unethical or unnecessary
- serious safety risks that make it inappropriate to continue

More on our scope

- What's the objective of a trial? What's *"success"?*
- Most typically, we think in terms of statistical significance
- The reality might not be so clear-cut
  - might we desire to obtain a certain magnitude of estimated effect?
  - can a trial that doesn't quite reach significance still provide meaningful information to help quantify a treatment's effects, or contribute to the further development strategy?
    - *Even a trial that shows weak effects more conclusively can still contribute relevant knowledge to the clinical community!*
  - what about other important endpoints on which the trial might provide meaningful information? Subgroups? etc.

Our scope

- In this course, we'll mainly proceed as if statistical significance of a single main endpoint is the objective of the study
- Most concepts and results that we'll discuss readily extend to broader settings

2
Group Sequential Designs

Background: Group Sequential methods

- Group sequential designs that allow stopping a trial with a claim of having achieved its objectives are quite common, and well-understood
  - can save time / resources / patient exposure if interim results are convincing; can make effective treatments available sooner
- Usually, criteria are set to control type I error level across interim and final analyses
  - by taking into account the *correlation* between tests induced by the common data
- Most frequently, *alpha spending functions* (Lan-DeMets 1983) are used; these allow flexibility in the number and timing of looks, reflecting realities of clinical trials

“Classical” group sequential procedures

- Assume 4 equally-spaced analyses, 1-sided $\alpha = 0.025$

	O’Brien-Fleming		Pocock	
Analysis	z_i	Nominal α	z_i	Nominal α
1	4.049	<0.001	2.361	0.009
2	2.863	0.002	2.361	0.009
3	2.338	0.010	2.361	0.009
4	2.024	0.021	2.361	0.009

– require increased sample size to achieve the same power

Spending function design scheme

- α-spending function is a monotone function on [0, 1], going from 0 to α
- A trial is designed to have desired operating characteristics, under some general expectation of how many analyses will take place, and when (*information*)
- The spending function reflects the cumulative type I error rate up to the point of each analysis
- Commonly used *alpha spending functions*
  - $\alpha(t) = 2 - 2\,\Phi\,(z_{\alpha/2} / \sqrt{t})$: O'Brien-Fleming
  - $\alpha(t) = \alpha \log \{ 1 + (e-1)\, t \}$: Pocock

Spending function implementation

- Interim analyses add a type of inefficiency to the design
  - e.g., for fixed α, β the most efficient procedure would analyze the data a single time
  - Allowing multiple points where success may be achieved decreases the power or requires more information
- A spending function induces a desired *philosophy* as to what types of outcomes could justify early stopping
  - *caution* versus *aggressiveness* ?
  - *maximum SS* versus *expected SS* ?
- *Expected sample size*

$$\text{Exp(SS)} = \sum_i (\text{SS at look i}) \times \text{P(stop at look i)}$$

  - depends on an assumed effect size

Sample size implications

- The trade-off:
  - Conservative boundaries (e.g. O'Brien-Fleming-type) tend to have *smaller impact on maximum sample size*, because most alpha is maintained to the final analysis, but there tends to be *lesser chance of stopping early*
  - Boundaries that are easier to reach early yield a *larger maximum SS*, but are *more likely to result in early stopping*, and thus may have a smaller *expected sample size* under strong alternatives

Implementation

- Information targets are rarely hit exactly
- The spending function is used to determine criteria based on when the analyses actually occur
- Level is protected by definition; power is usually minimally impacted by deviations from the planned analysis timing
- Criteria define a *necessary* condition to stop with a positive claim while protecting α
  - but not necessarily *sufficient*, stopping rules can be over-ridden

3
General Futility Considerations

Shifting gears: What about poor effect?

- Should a trial continue if we're able to judge (*don't worry yet about how*) that it won't meet its objective?
- Various motivations, often quite obvious:
  - Savings: cost, resources, patients (#'s and exposure)
    - perhaps resources can be allocated to more promising endeavors
  - re-evaluating or modifying a program based on what's been learned
  - Ethics: can we continue to commit patients to participation in a trial that will clearly not meet its objectives, or to investigational treatments that will not be viable?
    - for experimental treatments, perhaps we're exposing patients to *as-yet-unknown* safety risks

Counter-motivations

- Some situations have attributes that argue *against* allowing stopping for futility:
  - Particularly for trials involving available / marketed therapies: is the medical community owed a more definitive and precise answer, by continuing
    - as opposed to an ambiguous *"the trial will not show statistical significance"* ?
  - Short-term intervention period, most/all patients enrolled, longer-term follow-up
    - e.g., vaccine trials, surgical interventions
  - Possible negative impact on conduct of ongoing related trials
    - perhaps, same compound in more favorable settings, e.g., indication, population

Not **a good counter-motivation**

"We don't expect to see interim results that weak"

- Futility criteria usually DO correspond to outcomes we didn't expect!
- *And we can't control the laws of probability*
- Don't underestimate the ability of interim data to show surprising signals

Balancing errors

- Keep in mind as we proceed: in considering futility, we can't guarantee avoiding incorrect decisions

	Trial outcome	
Interim decision	Success	Failure
Stop for futility	(Incorrect)	(Correct)
Continue	Correct	Incorrect

- The general goal would be to *control / minimize* the chances that we're on the diagonal

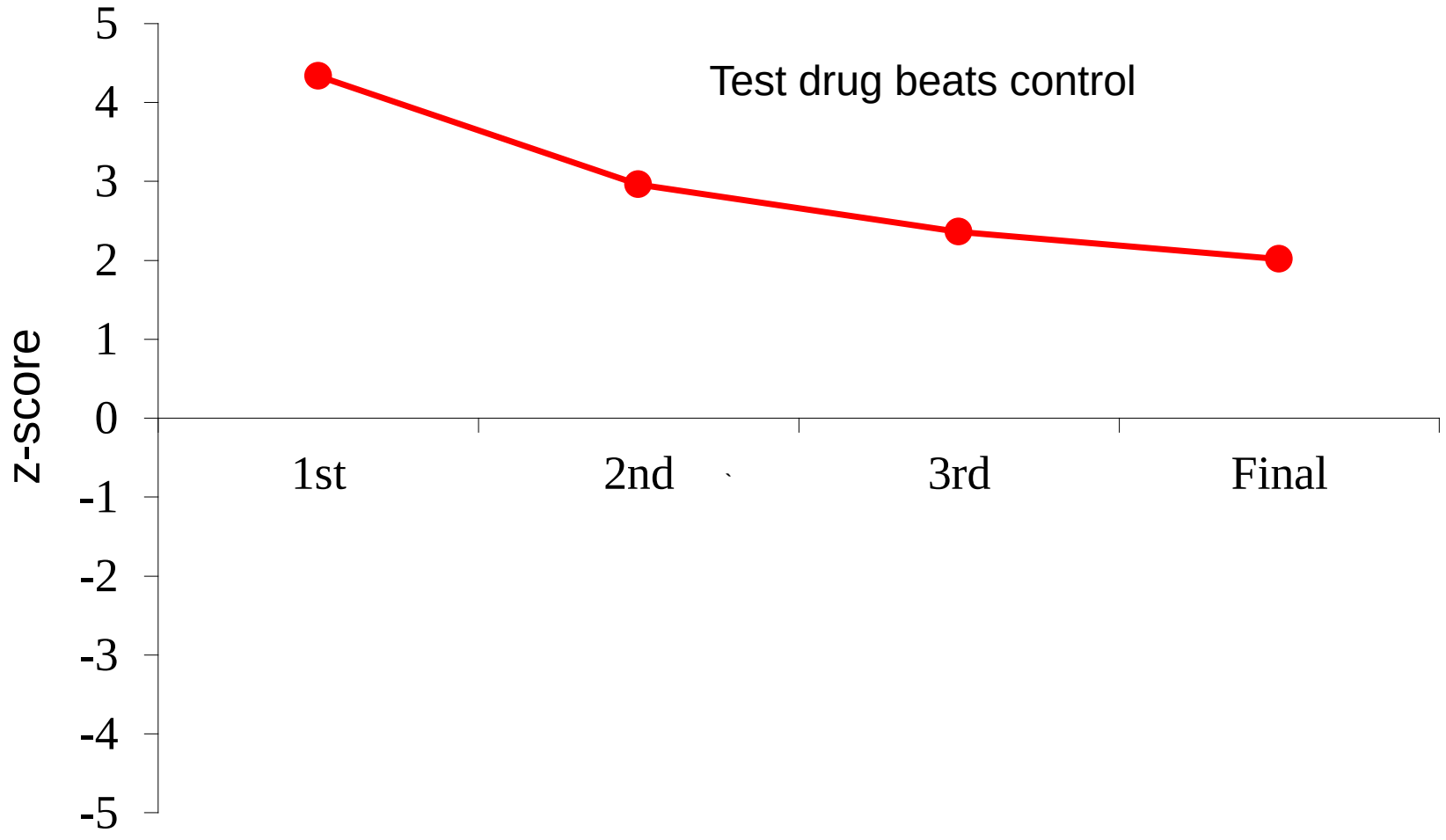
Trade-offs

- At any given analysis point, incorrect decision chances are always *in conflict* with each other:
  - If we set criteria to make one of the error rates *smaller*, then we necessarily make the other one *larger*
  - With interim data we can't expect, to control errors, for example, nearly as well as α, β in the full design
  - This can require careful consideration as to how to make the trade-off, to best meet the needs of a situation

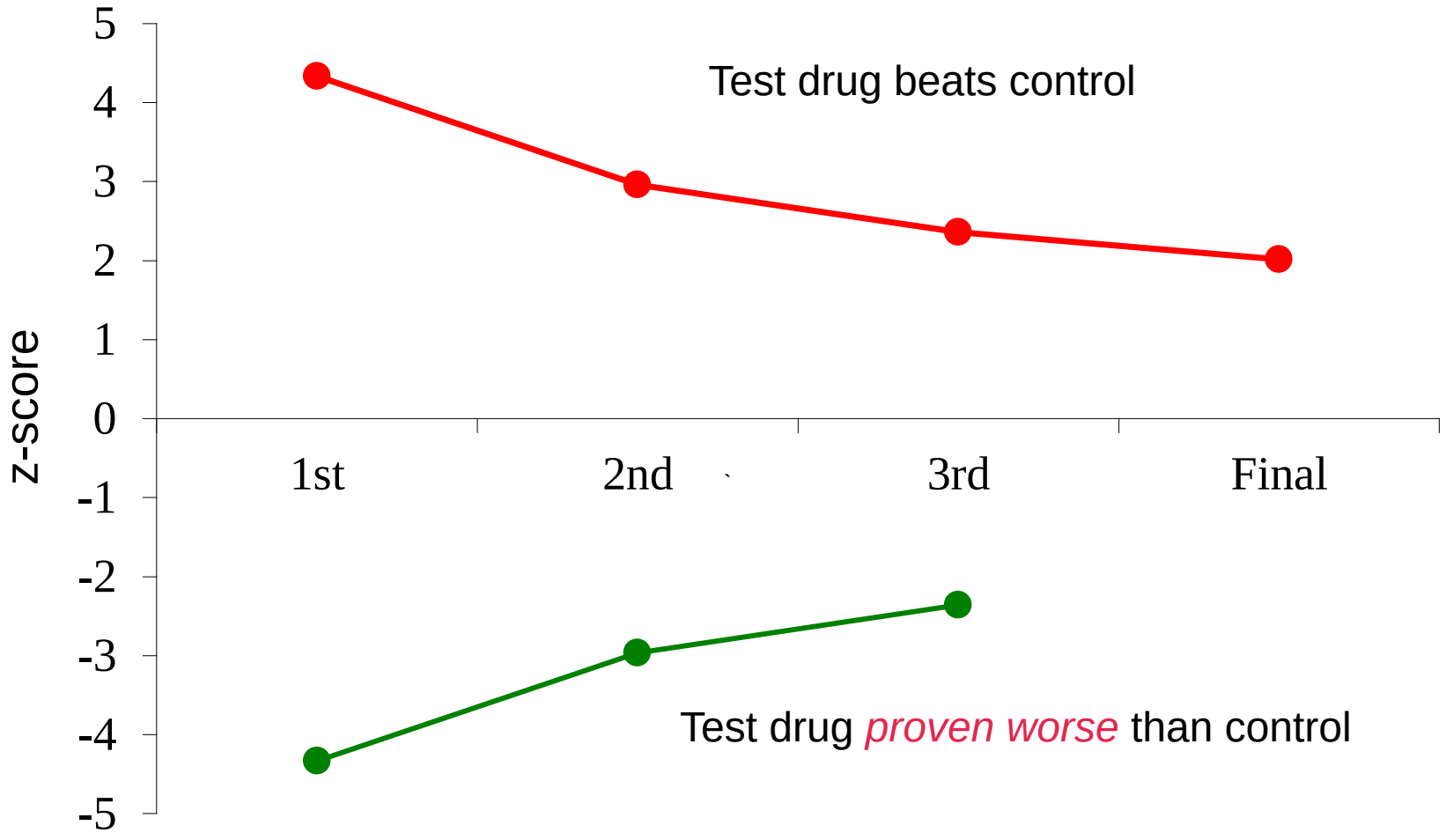
Example: imprecision of interim results

- Conventional 2.5% level, 90% power design for Δ
  - final significance reached if we observe ~ 0.6 Δ
  - with 50% information, length of a 95% CI is ~ 1.7 Δ
    - at 25% information, it would be 2.4 Δ
- Let's say that with half information, we observe 0.35 Δ
  - *disappointing, right?*
  - but a 95% CI ≈ (-0.5 Δ, 1.2 Δ)
  - we haven't come *close* to ruling out either H_0 or H_A
  - plausible effect sizes include some for which success is quite possible, and some where the trial is highly futile
- *So what's a sensible decision?*

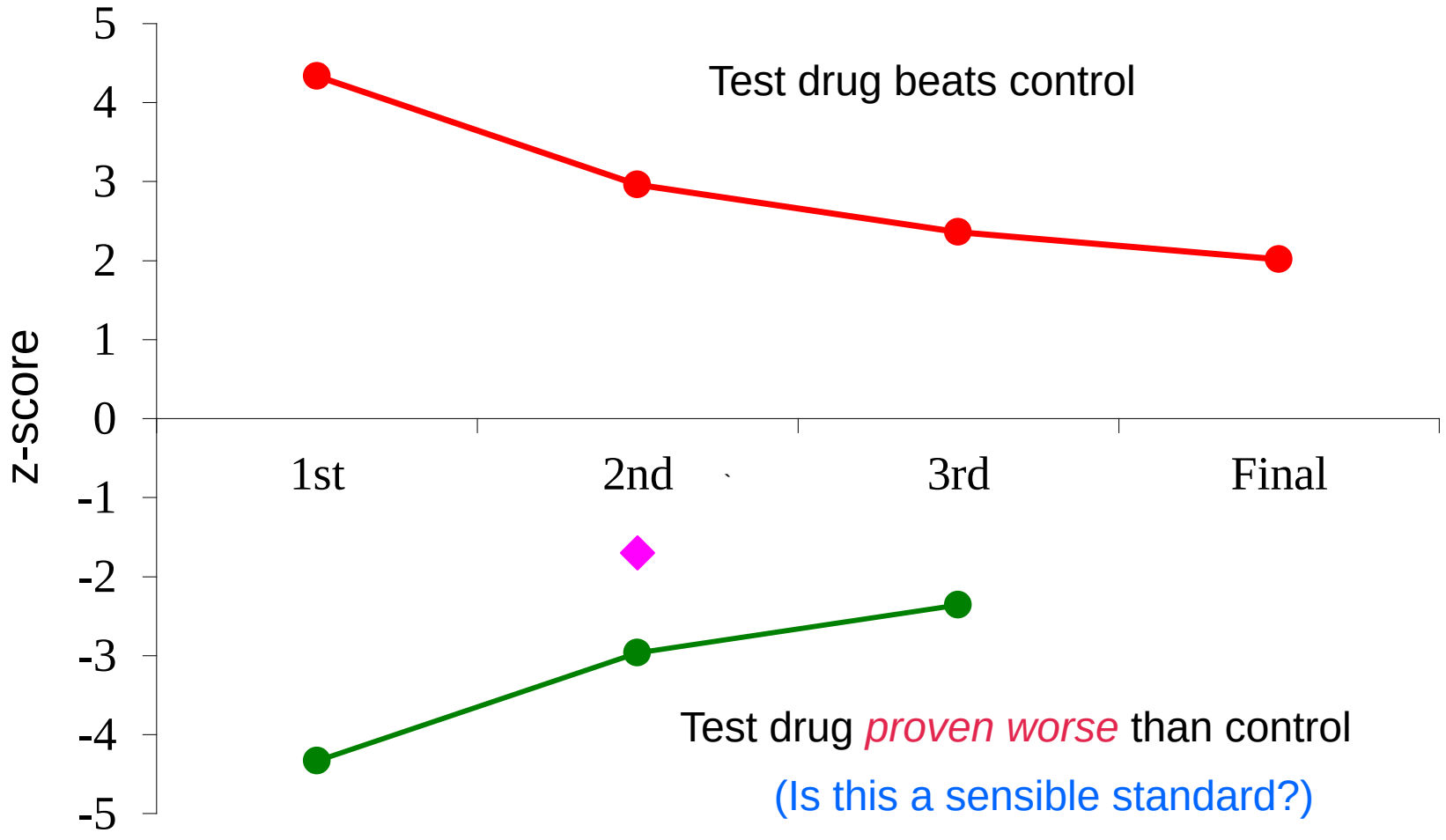
Standard 4-look O-F scheme



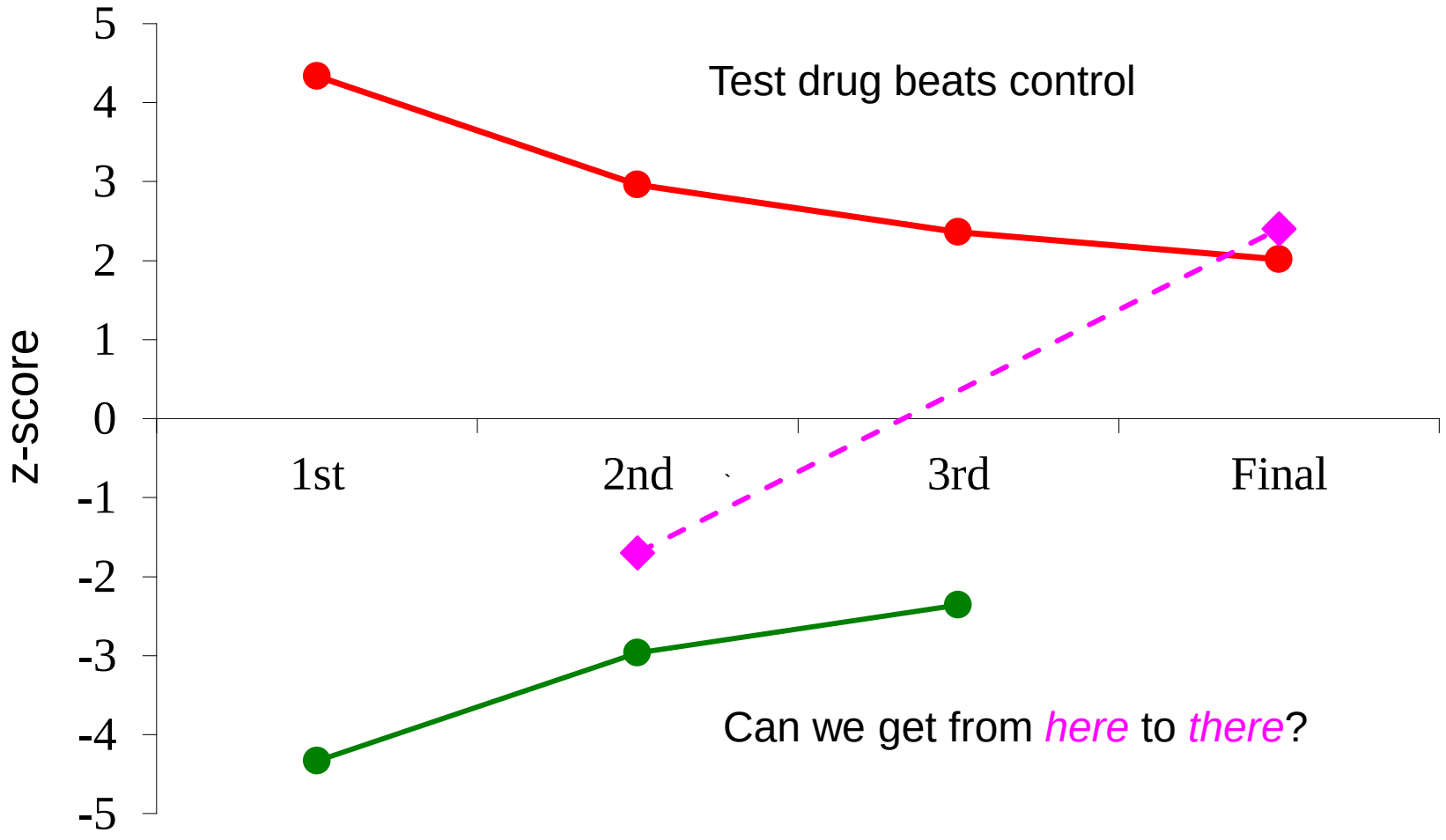
One possibility for “lack of effect”



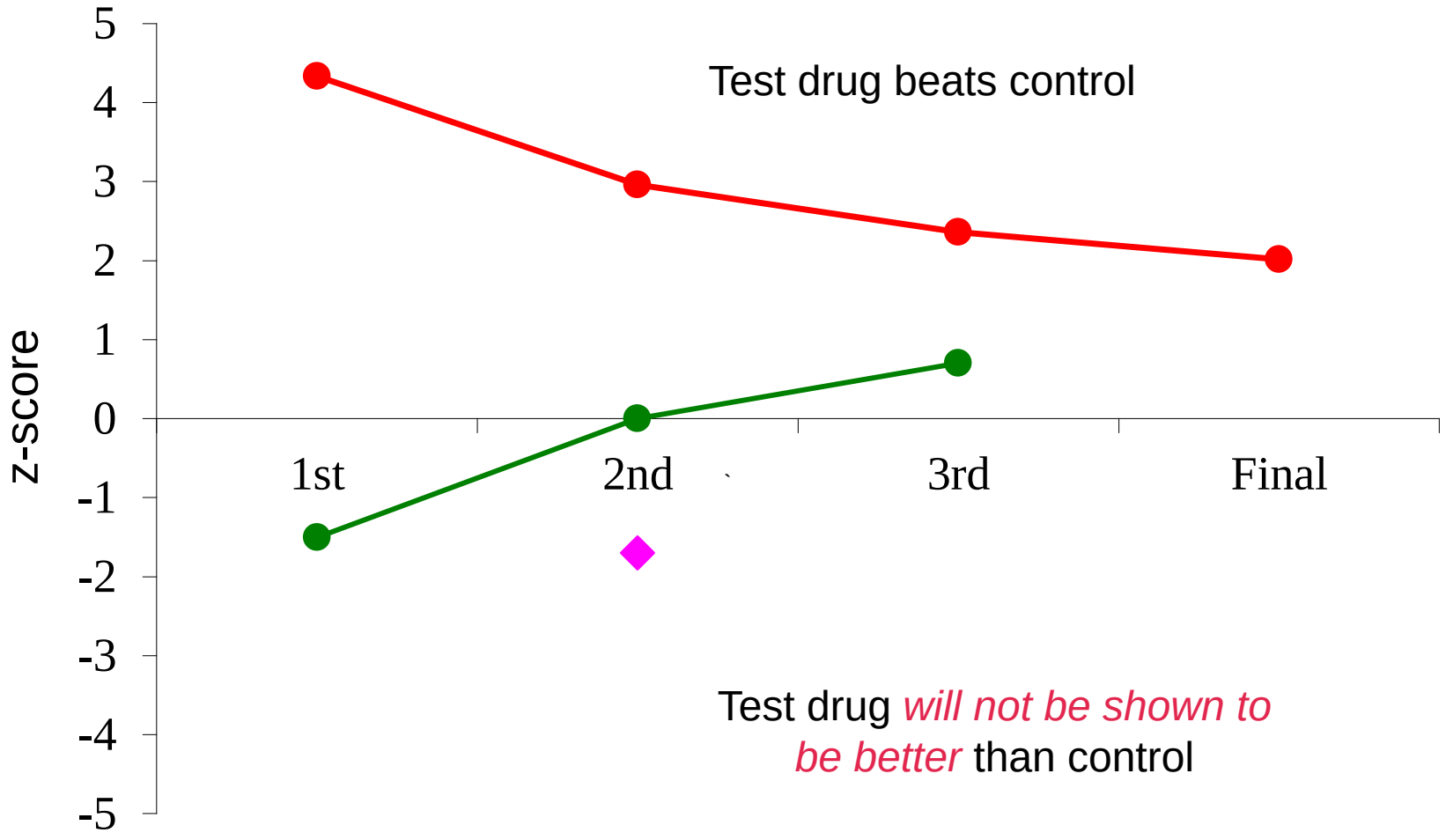
One possibility for “lack of effect”



What do we hope to achieve?



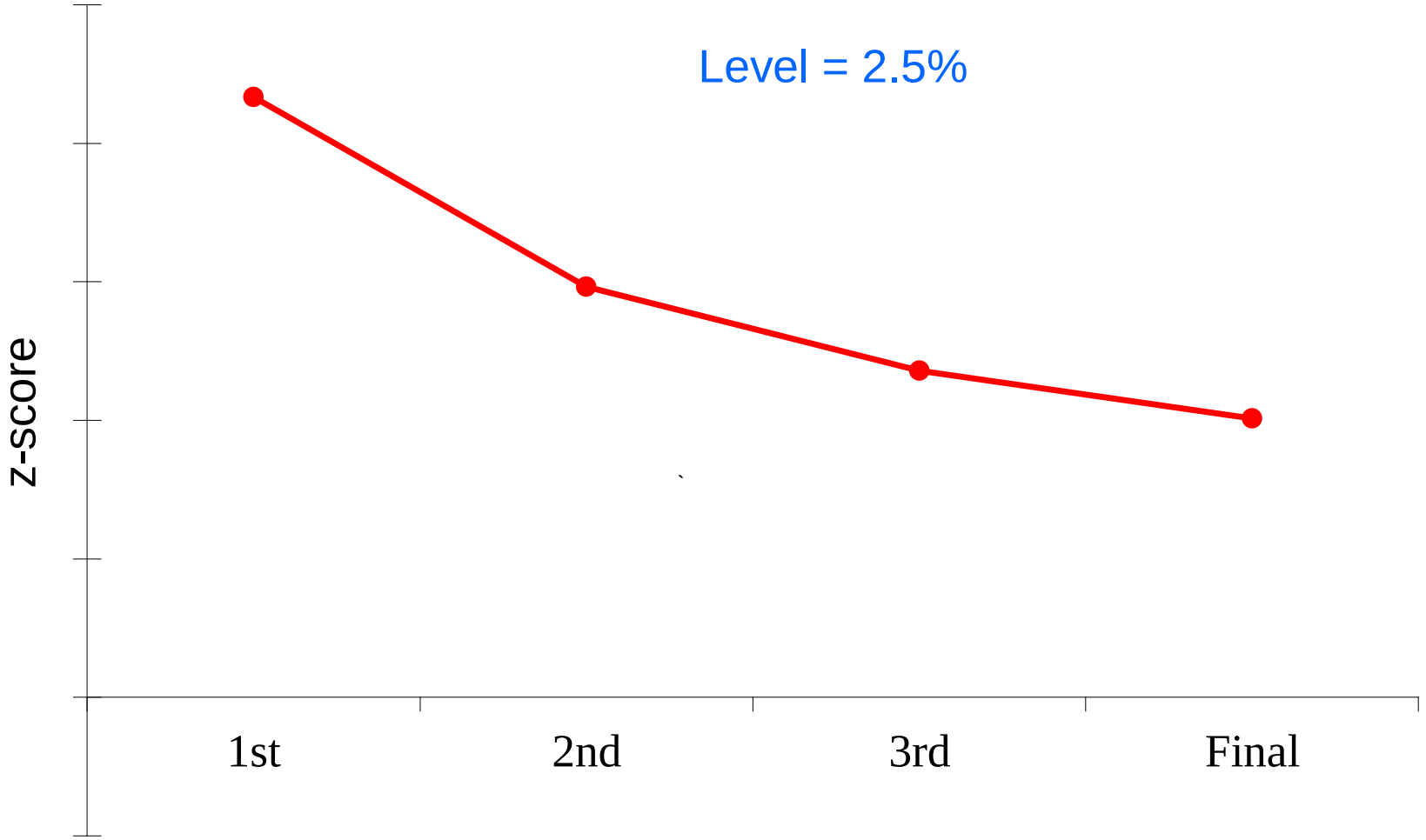
Does this make more sense?



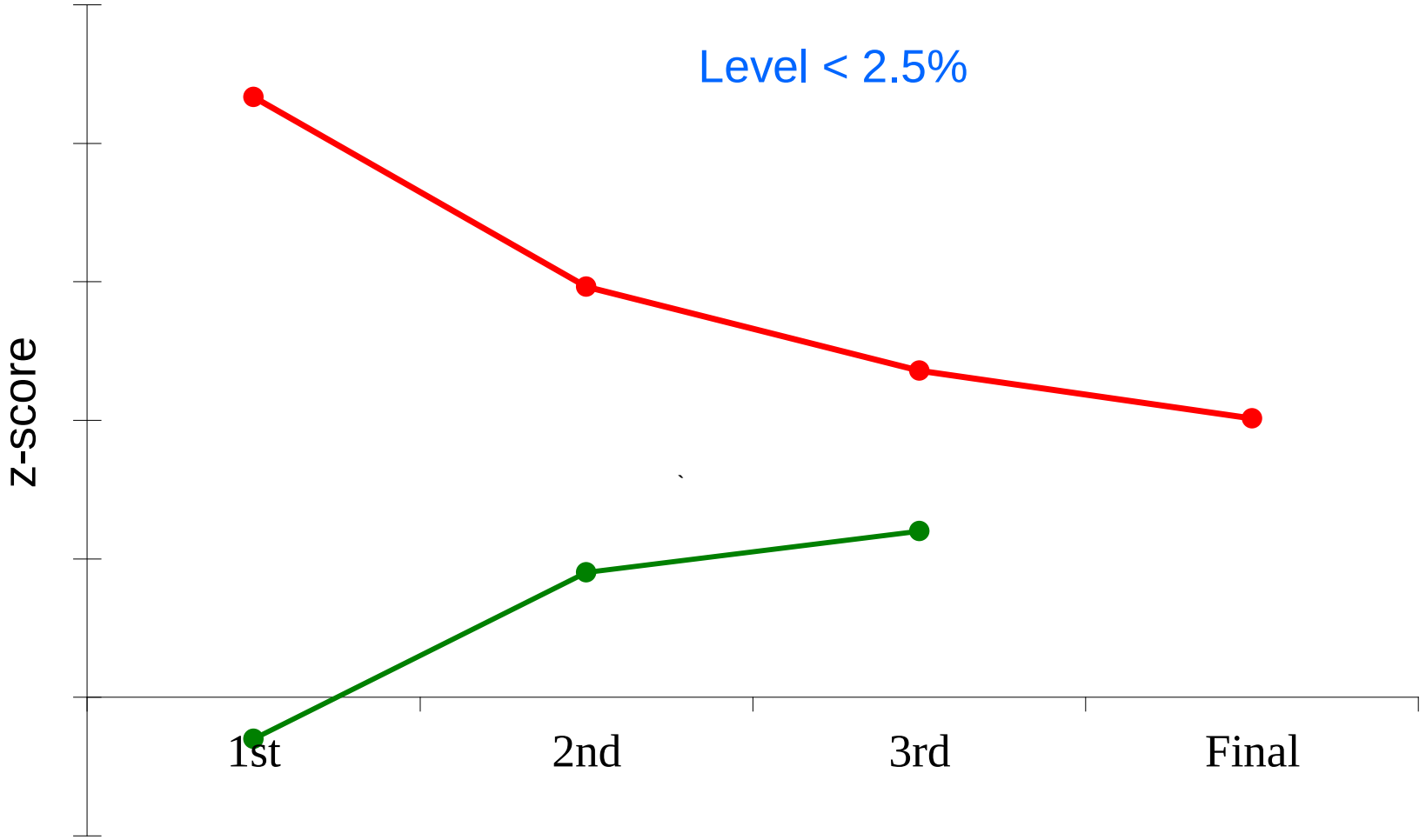
Adjusting success criteria?

- A futility scheme decreases both the level and power that the design would have in its absence
- Can we modify the success criteria to *regain the lost α?*
  - e.g., final critical z could be *below* 1.96?
- This would imply that to claim α control, reaching a futility criterion *requires* stopping
- But rigidly following an algorithm is not generally viewed as the nature of how futility should be addressed in practice

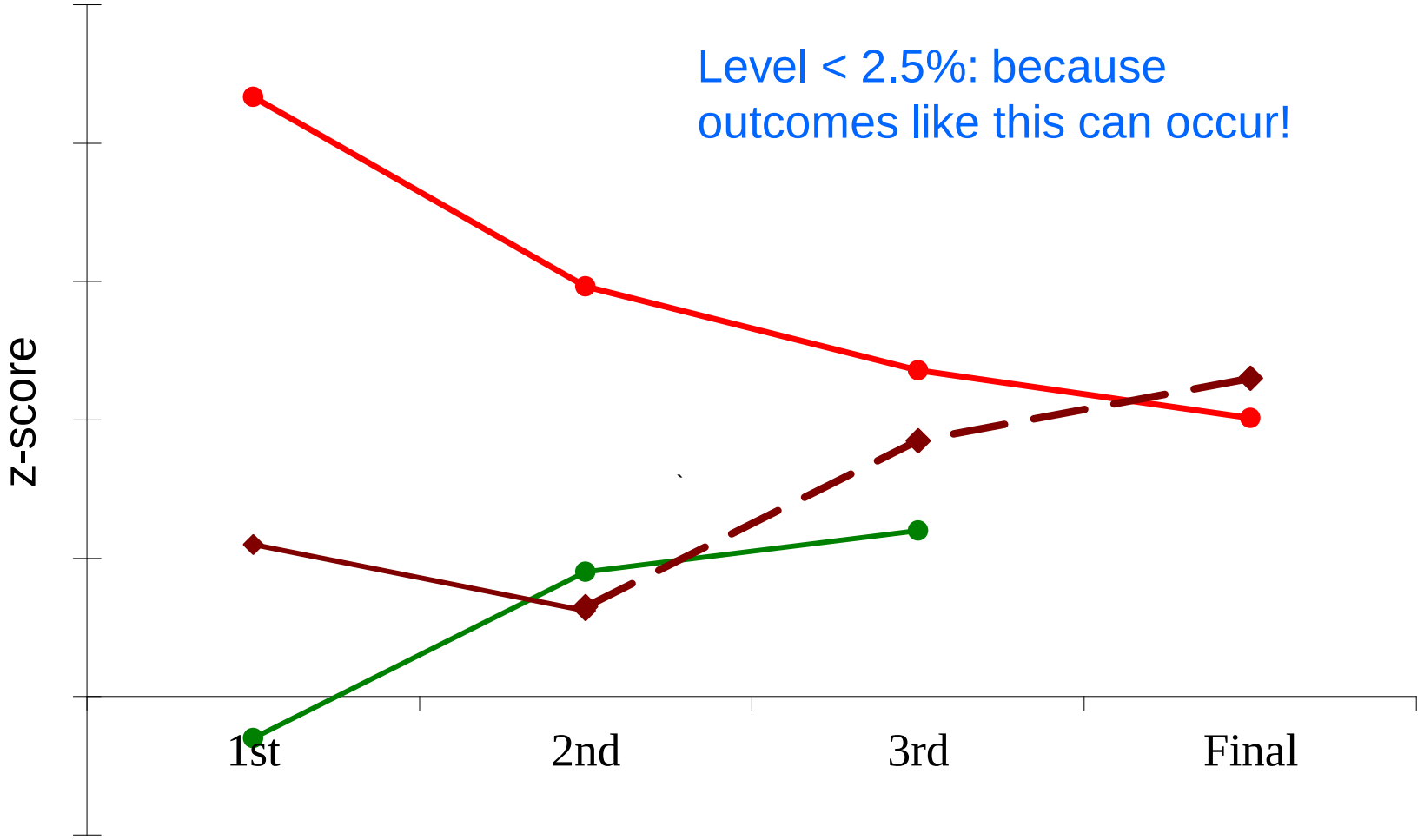
Typical efficacy scheme



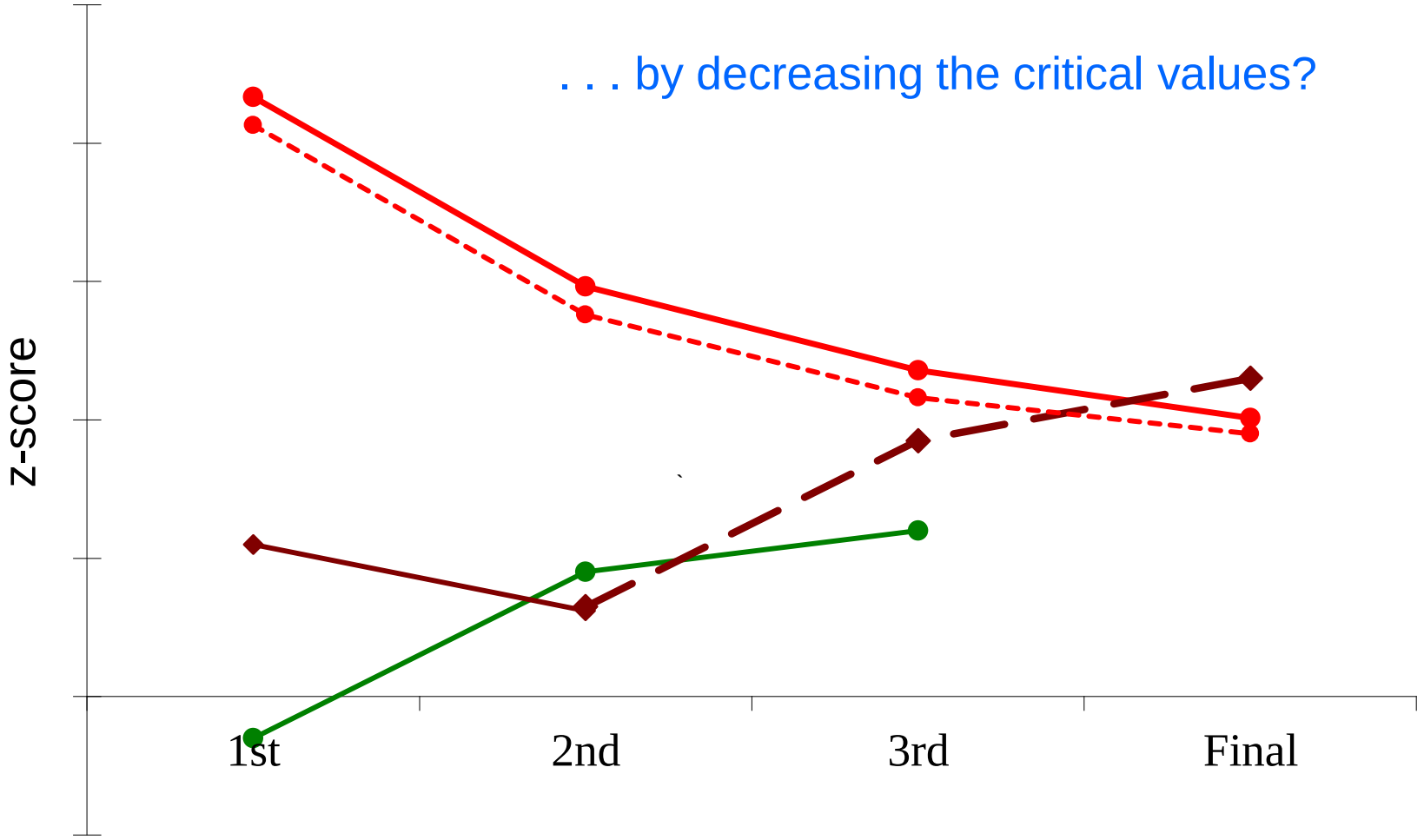
Impose a futility boundary



Level is decreased



Can we fix this . . .



Other relevant considerations

- Futility stopping is often a complex judgment involving the totality of information available
  - are there other outcomes / early markers that suggest that the results may eventually trend differently?
  - are there pending or unadjudicated outcomes with potential to change the current interpretation? A long “pipeline” of data that will eventually be obtained even if the study stops?
  - does the data suggest a time trend, e.g., increasing benefit with increased exposure, or investigator experience?
    - *non-proportional hazards!* – a big challenge (more on this later)
  - are there data ambiguities, e.g., suggestion of meaningful benefit in a subgroup, that justify getting more data to resolve?

***Guidelines*, not rigid rules**

- In most cases, futility rules are viewed as defining outcomes where stopping may be seriously considered, and *may* be implemented, pending a thorough review of all information available
  - success criteria should NOT be modified to “buy back” lost α
- Thus, futility boundaries will result in lost power
  - we could, if desired, increase SS to recover power
  - in this session, we will for the most part focus on *power loss* as a metric for comparing schemes, but we could if desired flip this around, i.e., fix power and vary SS

Terminology

- *Binding criteria*: if futility thresholds are reached, the understanding is that the study will stop; success criteria may be loosened accordingly
- *Non-binding criteria*: success criteria are NOT modified by the presence of futility rules (implicitly suggesting and allowing that they can be over-ridden when justified)
- Non-binding criteria are mainly used in current practice
  – we'll assume this during this session

4
Statistical Methodology

Methodological approaches

- Let's start to put some structure around quantifying the concept: allowing stopping if interim results are poor so that the chance of success seems small
  - but how do we determine the "chance of success"?
  - and what's *"small"?*
- What does this mean / how do we quantify this / what are the available tools?

Our notation

- Interim analyses with statistical information, respectively, t_1, t_2, . . . , final analysis at $t_F = 1$
- Asymptotically normal test statistics Z_1, Z_2, . . . , Z_F
- Study designed to have level α, power 1-β for an effect Δ (we'll mainly use $\alpha = 0.025$, $\beta = 0.10$)
  - e.g., mean difference for continuous data; for time-to-event data, Δ = ln(hazard ratio)
- For simplicity, we'll mostly assume there's no group sequential efficacy scheme, i.e., trial success occurs only if $Z_F > z_\alpha$

Tools

- A number of methods / approaches are used in current practice to address futility
- We're not yet addressing the setting of *criteria*; for the moment, only reviewing some available methodologic approaches

Tools for addressing futility

- Methodologies commonly used in setting futility criteria include:
  - β-spending functions
    - similar in concept to alpha spending: describes cumulative Type II error across the interim and final looks
  - Conditional power
    - what effect size is assumed to govern the rest of the trial?
    - often conditions on the original study alternative, or point estimate; other quantities may also be used
  - Predictive probability
    - Bayesian framework

β-spending functions

- Analog of α spending
- β(t) is an increasing function of information with β(1) = β
- Criteria are set so that at information time t, the cumulative false negative rate (i.e., stop for futility given Δ) across all analyses so far is β(t)

β-spending functions *(continued)*

- O'Brien-Fleming and Pocock-type spending functions have been defined
- Parametric families allow extension to schemes allowing different (intermediate) behavior, e.g.
- gamma family

$$\gamma(t) = \beta \{1 - \exp(-\gamma t)\} / \{1 - \exp(-\gamma)\}$$

- rho family

$$\rho(t) = \beta t^{\rho} , \rho > 0$$

Examples: β-spending

- α = 0.025, 90% power, single look at 50% information

Boundary	Z-score	Power (%)
O'Brien-Fleming	0.287	89.4
Pocock	0.955	85.4
Gamma(-1)	0.589	88.3

Conditional Power (CP)

- At a single interim analysis, what's the chance that the trial will be successful if it continues?
- Define an effect Δ^* that will be assumed to govern the remainder of the data

$$CP(\Delta^*) = P(Z_F > z_\alpha \mid Z_i, \Delta^*)$$

- Can be solved by de-composing Z_F into its fixed and stochastic parts

Conditional Power calculation

- 2-look scheme, interim done at information t
- Z_1, Z_F represent interim and final test statistics, z_F is the critical value for final analysis
- *Decomposition*: $Z_F = \sqrt{t}\, Z_1 + \sqrt{(1-t)}\, Z_2^*$
- where Z_2^* is the increment of data obtained between the interim and final analyses

Conditional Power calculation

- Assume that Δ^* is the effect size governing the remainder of the trial

- $CP(\Delta^*) = P(Z_F > z_F \mid Z_1, \Delta^*)$

  $= P(Z_2 > (z_F - \sqrt{t}\, Z_1) / \sqrt{(1-t)} \mid \Delta^*)$

- In multiple-look schemes, we can incorporate the chance of success at *any* later interim analysis
  - but this often yields results very similar to a calculation such as shown above that focuses on success just at the final analysis

Predictive Probability (PP)

- Bayesian framework
- Can be viewed as averaging CP across a distribution of effect sizes induced by a prior and the data so far

$$\text{PP}=\int CP(\Delta^*)P(\Delta^*|interim\ data)d\Delta^*$$

- A valid probability statement much more broadly (i.e., than CP)
  - though only one value within the predictive distribution is true
  - sometimes referred to as *probability of success*
- Most commonly a non-informative prior is used, but certainly informative priors can be considered

Analytical forms

$$CP\ (d) = \Phi\left(-\frac{1}{\sqrt{1-t}}\left(z_\alpha - \frac{Z_I}{\sqrt{t}} \right) \right)$$

- Non-informative prior

$$PP = \Phi\left(-\frac{1}{\sqrt{1-t}}\left(\sqrt{t}\, z_\alpha - Z_I \right) \right)$$

- For a trial designed with level α, power 1-β

$$CP(\Delta) = \Phi\left(-\frac{1}{\sqrt{1-t}}\left(t z_\alpha - \sqrt{t}\, Z_I - (1-t)\, z_\beta \right) \right)$$

$$\beta\ spent = \Phi\left((Z_I - \sqrt{t}\,(z_\alpha + z_\beta) \right)$$

5
Practical Implementation Considerations

Which approach to use?

- Discussions of the relative merits of the different approaches often seem to focus on *philosophical* grounds
  - e.g. the assumptions seemingly being made
  - the degree to which quantities might be interpreted as chances of success
    - *but are they really?*
  - Spiegelhalter *et al* (CCT 1986): conditional approaches *"fall short of being a rational aid"*
    - i.e., compared to predictive approaches

Examples

- Consider 3 different outcome scenarios where action in an ongoing trial is planned / taken, based on interim results:

1. stop for futility because the interim effect is weak, and CP conditioning on that value is low (< 20%)
2. positive signal but GS boundary not reached, CP based on design Δ is 74%, and more favorable effect sizes are very plausible based on the data, so no action taken
3. positive signal seen, noticeably less than hypothesized, but an SSR scheme (say, Cui-Hung-Wang) was pre-specified and leads to a substantial SS increase

You might have guessed . . .

- The 3 scenarios and outcomes are *identical*
- Conventional study design, 2.5% level, 90% power for an effect Δ
- Interim analysis at information time t=0.4:
  - point estimate: d = 0.4
  - 95% CI for Δ = (-0.56 d, 1.36 d)
- So - how do we decide what's a good action, and what role the interim results and calculated CPs should play?

CP: condition on *what?*

- Most commonly:
- the originally hypothesized design effect Δ?
  - in some sense, extends the study design
  - but now we have *data* – shouldn't we use it?
  - but in a neighborhood where stopping could make sense, this is much more optimistic than the data is suggesting
  - so we're conditioning on an effect that's somewhat *contradicted* by the data?
- the point estimate (say, d)
  - as a *best guess*?
  - but certainly not a *good guess,* as we've seen

Other choices

- An effect between the initial hypothesis and the estimate?
  - perhaps determined based upon a pre-specified prior? other data?
- A pre-specified confidence limit on the point estimate?
  - how *optimistic* or *pessimistic* am I willing to be regarding the effect that was observed?
- Certainly multiple versions / quantities can be presented to a Data Monitoring Committee (DMC)
  - though usually one of these will have been pre-stated to be the quantity of main initial focus

PP interpretation issues

- Averages CP over a distribution of effect sizes induced by the interim data
- Only one of these is true (we don't know which one)
- Much more broadly a valid probability statement
- How *wide* is the predictive distribution?

Setting criteria

- In a given situation, in addition to picking a *method*, we need to decide on the *details* of its implementation
  - e.g., what specific CP or PP threshold, or what parameter value for a spending function family?

- It's easy to find literature examples where CP(Δ) has been described as a basis for a futility plan
  - 10%, 20%, 30% are thresholds specifically mentioned

Are these concepts *intuitive?*

- Two actual proposals / consultations for futility criteria that I've experienced:
  1. With 20% of data available, conditional power assuming the original Δ must be at least 5%
  2. At ⅔ information, the conditional power computed assuming that the observed effect is the true effect is at least 70%
- *More on these later . . .*

Relationship between criteria

- At a given time point, a futility rule expressed on *any particular scale* can be transformed to *any other*
- For example, in a 2.5% level, 90% power trial, single look at t = 50%, say we set a criterion of PP = 20%
- The same rule can be expressed as:
  - CP(Δ) = 62%
  - CP(d) = 12%
  - 'Beta spent' = 6.7%
- Question: is the *scale* on which we express a futility criterion *really that important?*

Analytical forms *(as before)*

$$CP\ (d) = \Phi\left(-\frac{1}{\sqrt{1-t}}\left(z_\alpha - \frac{Z_I}{\sqrt{t}} \right) \right)$$

- Non-informative prior

$$PP = \Phi\left(-\frac{1}{\sqrt{1-t}}\left(\sqrt{t}\, z_\alpha - Z_I \right) \right)$$

- For a trial designed with level α, power 1-β

$$CP(\Delta) = \Phi\left(-\frac{1}{\sqrt{1-t}}\left(t z_\alpha - \sqrt{t}\, Z_I - (1-t)\, z_\beta \right) \right)$$

$$\beta\ spent = \Phi\left((Z_I - \sqrt{t}\,(z_\alpha + z_\beta) \right)$$

Inter-relationships

- Note that there are *structural similarities* among the quantities we've described, leading to straightforward conversions, for example:

$$CP(\Delta) = \Phi(\sqrt{t}\,\Phi^{-1}(PP) + \sqrt{\{1-t\}}\,z_\beta)$$

$$CP(d) = \Phi(\Phi^{-1}(PP)/\sqrt{t})$$

$$\beta\ spent = \Phi(\sqrt{\{1-t\}}\ \Phi^{-1}(PP) - \sqrt{t}\ z_\beta)$$

Logic?

- The literature contains suggestions that one should start by defining a *"chance of success"*-type threshold, and *then* decide which quantity should be evaluated relative to this threshold
  - stop if the chance is <20%; *but should I use CP(Δ), CP(d), PP, etc.?*
- Proposition: this is inherently *illogical* and *counter-productive*
  - *backwards?*
- How can I set a sound threshold without adequately understanding the statistical behavior of the specific quantity that will be the basis for the decision?

Example: 90% power, t = 0.50

How much of *this* am I willing to spend . . .

. . . to get how much of *this*?

Z-score	d / Δ	CP(Δ)	CP(d)	PP	β spent	Power loss	Stop under H_0
No stopping	-	-	-	-	-	0	0
0	0	32%	<1%	3%	1.1%	0.2%	50%
0.25	0.11	41%	1%	5%	2.1%	0.6%	60%
0.50	0.22	51%	4%	11%	3.7%	1.3%	69%
0.75	0.33	61%	10%	18%	6.2%	2.7%	77%
1.00	0.44	70%	22%	29%	9.8%	5.1%	84%

scales for expressing futility rule

behavior

Choice of metric

- A futility threshold should be chosen as an integral part of the study design to induce desirable behavior and operating characteristics onto the design
- The particular *scale* or metric we use to express a sound rule certainly is not totally without interpretive value, but often it's really more of a *device* or *convenience*, rather than the *driver*
  - it may facilitate modification of criteria when interim information targets are not exactly achieved

Aggressiveness / caution

- *{We need not focus only on H_0, H_A; other measures of weak effect, likely success, etc. could be considered and evaluated}*
- When setting criteria, how much

  risk of stopping when we shouldn't
- are we willing to pay to buy a desired amount of

  chance of stopping when we should ?
- A number of factors can be relevant to decide on the degree of caution / aggressiveness that might be appropriate in a particular situation

Caution / aggressiveness factors

1. How much belief do we have in the investigational product?

- Is there strong scientific rationale or extensive empirical evidence (e.g., from phase II) leading us to believe that the drug should work?
  - so if early results were weak, we might be more likely to suspect that this could, at least in part, be chance (i.e., bad luck!)
- Or is there less evidence / rationale for a favorable prior belief?
  - in which case we might be more inclined to *"let the data speak for itself"*

Caution / aggressiveness factors

2. Ethics

- Is the endpoint itself harmful (i.e., not just symptomatic), so that a *signal* of harm could be actionable?
- Are there potential safety risks to which we might be subjecting patients by continuing exposure?

Caution / aggressiveness factors

3. Costs

- Quantify savings in terms of whatever dimensions of savings ($, time, enrolled patients) we're particularly interested in

4. "Upside" / potential net benefit

- Is the investigational treatment one that could have a major medical / societal / commercial impact?
  - which might argue in the direction of more caution

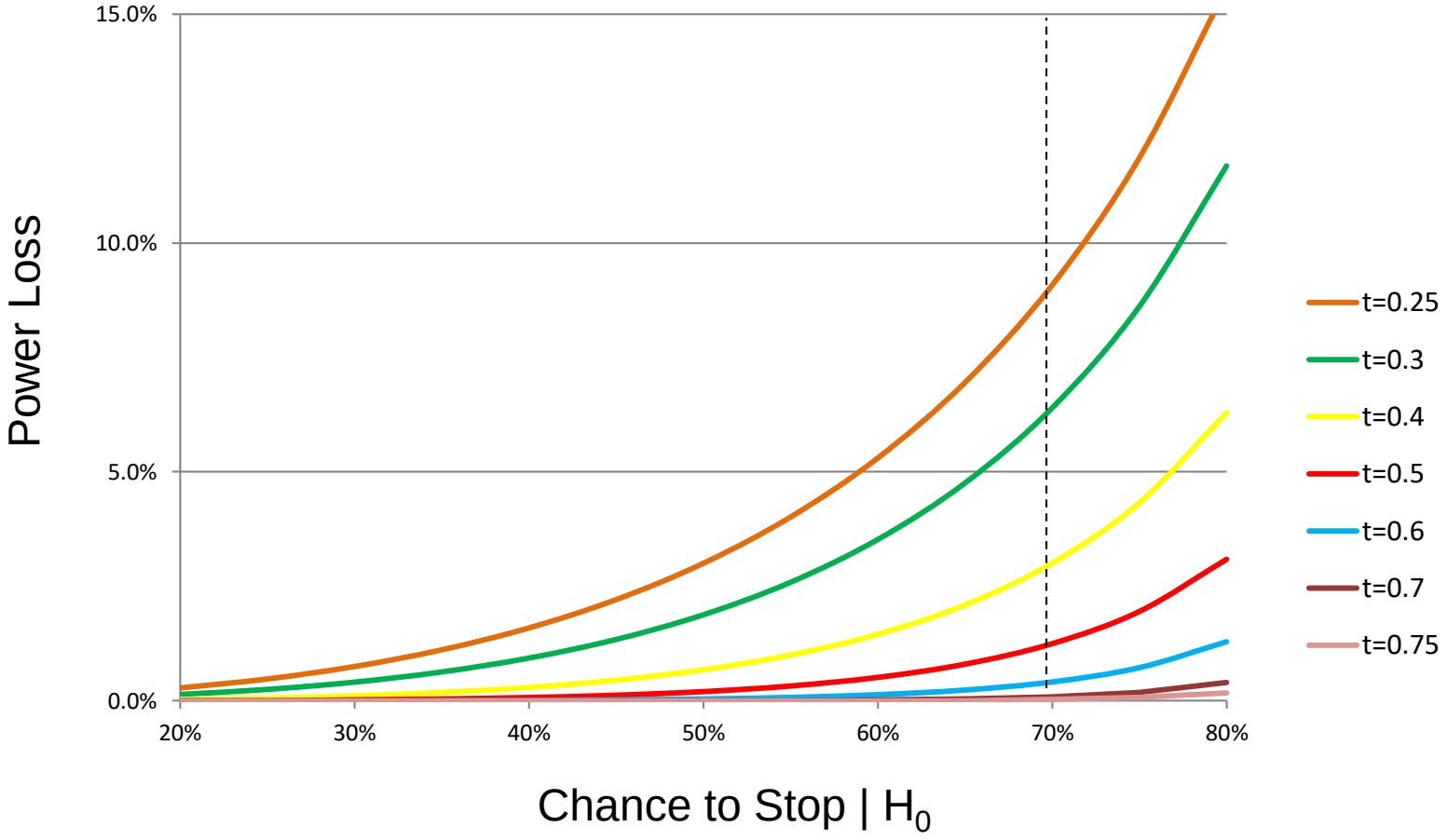
When to evaluate futility?

- Again, a *conflict* :
  - early: allows potentially greater savings
  - later: better statistical properties
    - better ability to distinguish between scenarios which *should / should not* justify continuing
- Futility behavior improves with information in *2 ways*:
  - greater accuracy of inferences from increased data (of course)
  - *less data still to come* that can overturn current trends

Futility timing

- Fixed versus variable costs
  - in some trials, and for some dimensions of cost, a large proportion of costs are incurred or committed early
  - in others, costs are distributed more “proportionately” as the trial proceeds
- Previous example: consider the criteria: z = 0.5
  - at t = ½, we saw that power loss was 1.3%
  - at t = ¼, it’s 9.2%

Error trade-offs for varying time choices



Multiple futility looks

- *Why not?*
  - particularly in long-term studies
  - the motivations are not necessarily specific to a single point
- There are practical limitations (on both ends) to when looks should take place
  - too early, too late: *no point*
- The presence of a later look might impact the choice of criteria at a prior look
  - a decision to continue *doesn't commit to trial completion*, but only to proceed until *a later point where data is more mature*

Choosing the metric, threshold

- Ideally, we might describe a scheme *simply*
  - i.e., constant value on a particular scale (e.g., CP(Δ))?
- *Now the scale matters!*
  - equal criteria across looks on one scale could be very unequal on another scale
- Let's say that at t = ½, we feel that CP(Δ) = 50% is a sensible criterion
  - what if we also used the same criterion at t = ¼, ¾ ?
  - PP across the 3 looks: 1.3%, 10.0%, 23.0%

Constant thresholds?

- If CP(Δ) = 50% works well at t = ½, should it be the metric and threshold at all points of the trial?
  - is there any reason to expect this same standard has comparably good behavior at t = 0.25 ? at t = 0.75 ?
    - hint: *it doesn't . . .*

- CP at time 25% vs 50% vs 75%
  - far different *degrees of precision* in the estimates
  - presumably different degrees of *contradicting the original study hypothesis*
  - different amounts of data yet to come that can *overturn current trends*

Multiple-look principle

- As we determine boundaries in a multi-look scheme, it's important to address whether it behaves well *at all timepoints at which it might be used*
- This applies regardless of the approach, e.g., it's relevant for β-spending functions as well
  - the function "links" criteria at different looks, but does it do this in a desirable way?

Quantifying the error trade-offs

- How to extend to multiple looks?
- The cost of incorrect stopping:
  - how about *"power loss across the whole scheme"?*
  - of course, there are multiple schemes that achieve the same degree of power loss overall
    - perhaps, equal power loss at each analysis?
- The benefit of correct stopping:
  - ASN: average sample proportion under H_0
  - i.e., $\Sigma\ t_k \times P(\text{study stops at } t_k \mid H_0)$

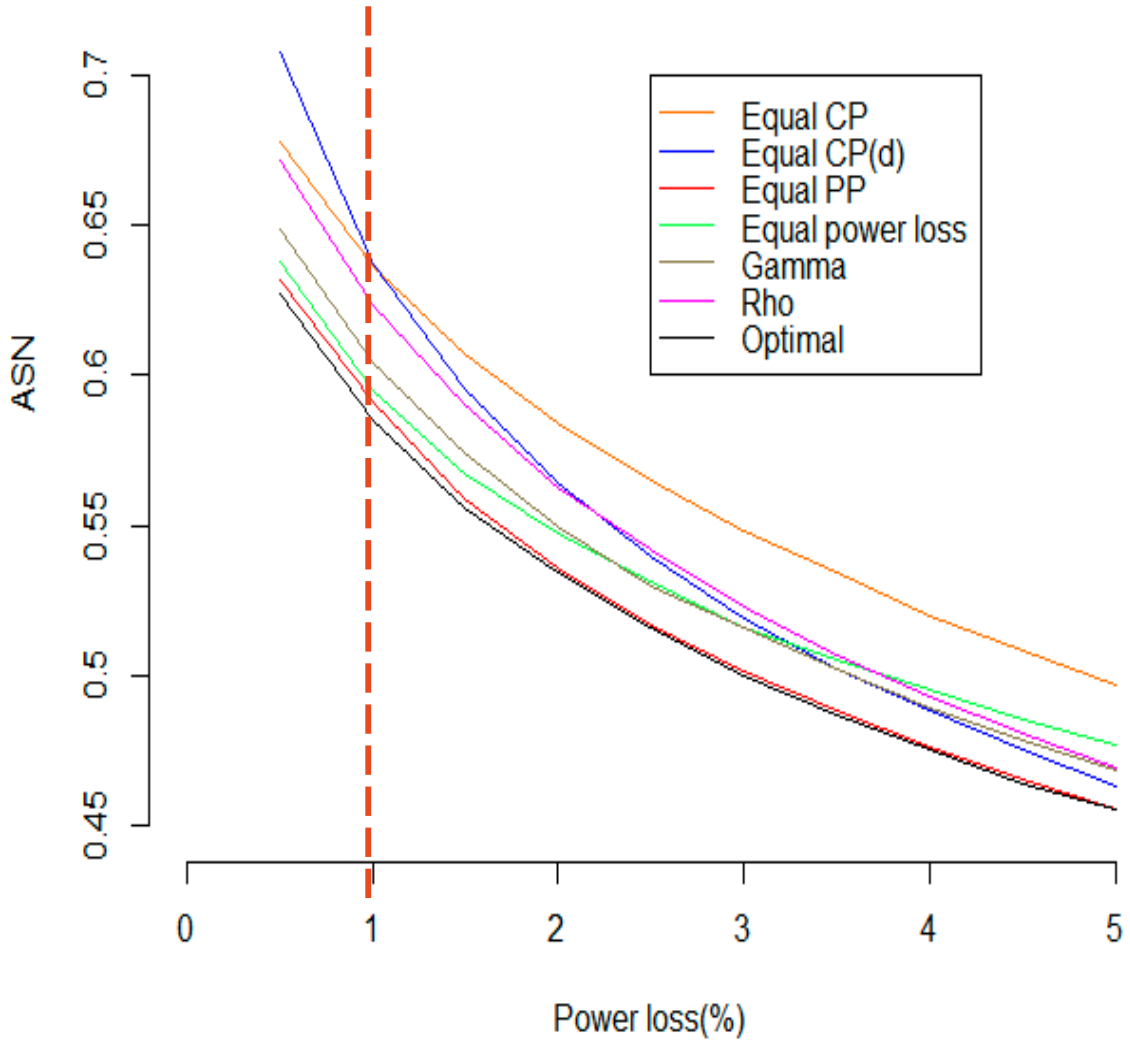
Optimality?

- Optimal boundaries: For a given schedule of analyses, and a specified amount of power loss, we can define boundaries that minimize ASN
- In what follows, we'll compare various boundary approaches:
  - equal CP(Δ)
  - equal CP(d)
  - equal PP
  - equal power loss
  - *optimal* (as above)
  - gamma family
  - rho family

Comparative investigation

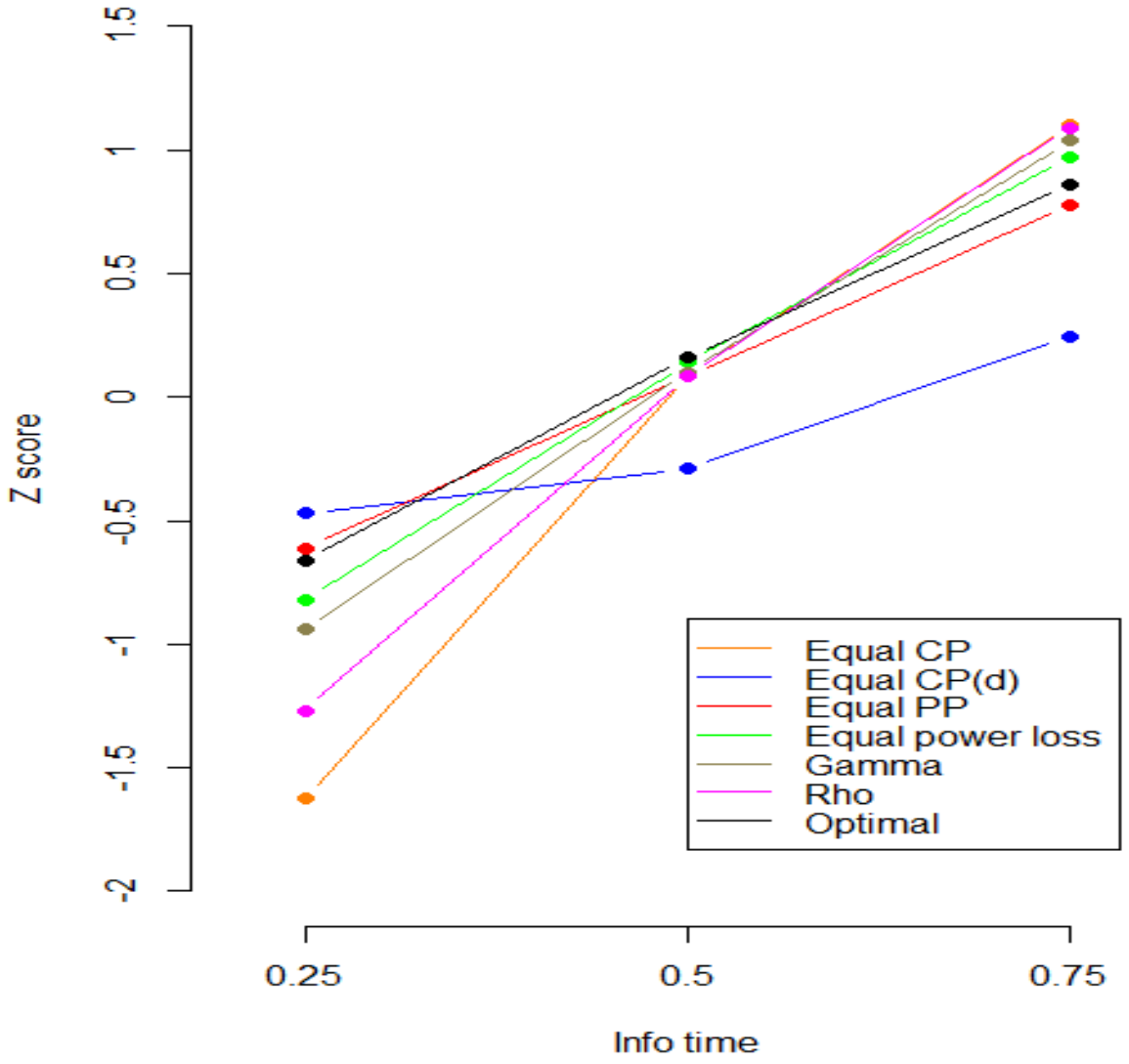
- 3 interim looks at t = ¼, ½, ¾
- For each method, fix degrees of power loss and determine the corresponding boundary values, e.g.:
  - what common CP value at the 3 looks achieves that value of power loss?
  - what gamma family parameter achieves that degree of power loss?
  - etc.
- Determine ASN for the resulting boundaries
- Plot ASN versus power loss

ASN versus power loss



- Equal PP at the 3 looks is *quite close to optimal*
- Equal CP(Δ) fares *particularly poorly*
- Suggests that PP provides an efficient futility framework?

1% power loss comparison

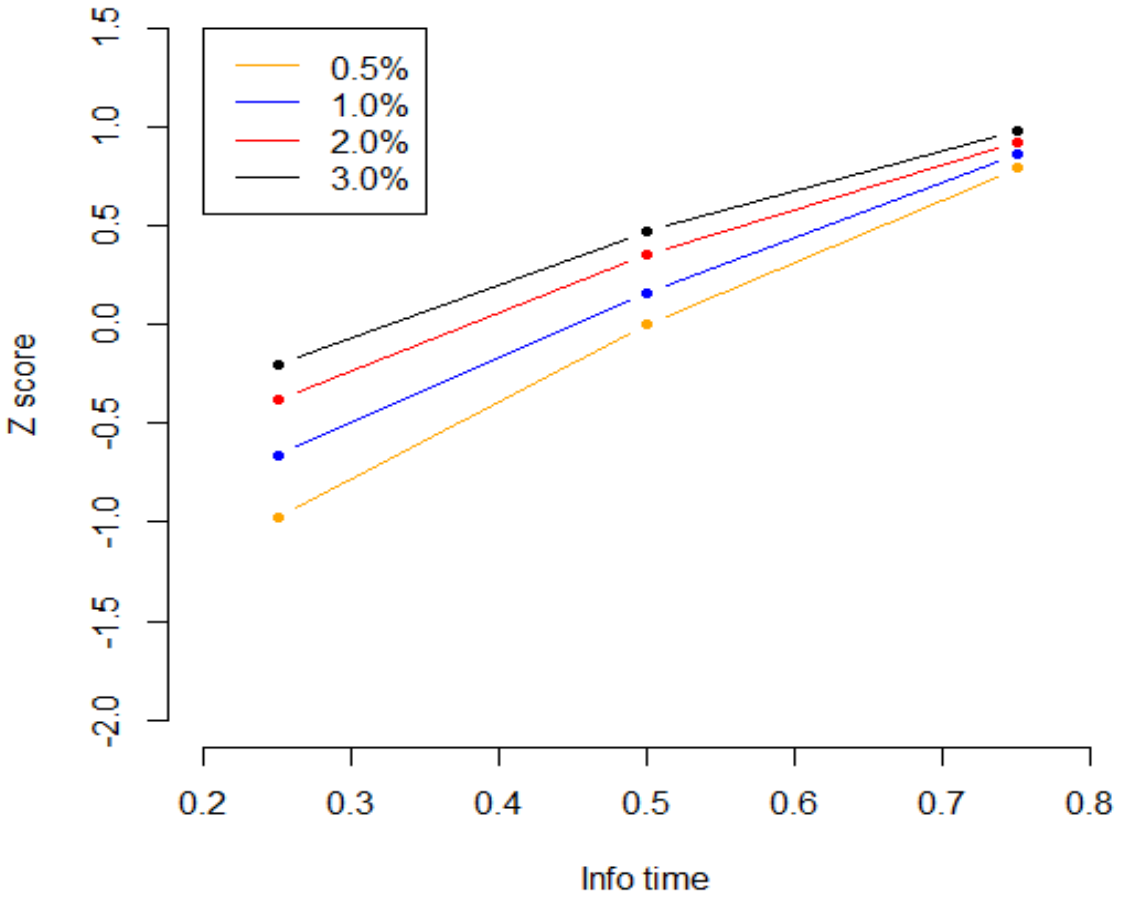


1% power loss boundaries

			Futility boundary on z-scale		
Boundary type	**Common value**	**ASN**	**1st look**	**2nd look**	**3rd look**
Equal CP(Δ)	0.347	0.636	-1.622	0.087	1.101
Equal CP(*d*)	0.0004	0.637	-0.472	-0.291	0.245
Equal PP	0.033	0.590	-0.612	0.086	0.780
Equal power loss	0.0033	0.595	-0.819	0.138	0.972
Gamma	(γ = -3.362)	0.604	-0.941	0.101	1.037
Rho	(ρ = 2.917)	0.623	-1.269	0.085	1.091
Optimal	-	0.585	-0.660	0.160	0.860

What do “good” boundaries look like?

- Optimal boundaries for varying amounts of power loss



What do “good” boundaries look like?

- *Interim results should not be expected to predict well the final study results !!*

- *Personal viewpoints:*
  - in many situations: power loss 1 – 2% ?
  - early in a study, boundaries correspond to negative outcomes
  - they cross into positive territory somewhere towards the middle of the trial
  - never correspond to highly favorable outcomes

“Messaging”

- Might trial personnel be *encouraged* by seeing that a study proceeded beyond a futility analysis - *“we’ve made it past a futility analysis, so there’s a good chance the trial will be successful”* - and then *disappointed*?
- The proper interpretation of continuation beyond a futility evaluation is:
  - *not* that the trial is *likely* to succeed
  - but rather, that it *has a chance* to succeed
  - *or else we would stop too many trials that turn out to be successful*

Sports analogies!

- Thought exercise: what type of *deficit* in a sporting event might correspond to a level of futility that would justify stopping a clinical trial?
- Major league baseball
  - 4^{th} inning? 7^{th} inning?
- World Cup soccer
- NFL football
  - midway through 3^{rd} quarter?
  - *What if the score was 28 - 3?*

Predictive probability?
Super Bowl 51

Non-constant effect / non-proportional hazards

- There's a number of reasons constant effect might not hold
  - early-enrolling patients at initially-opening sites might be systematically different from patients enrolled later
  - or there might otherwise be some *"drift"* in the patient population over time
  - for chronic treatment, benefit might emerge slowly based upon cumulative amount of therapy received
  - investigators may gain experience in optimally administering a complex therapy
- A particularly important case: non-proportional hazards
  - not at all uncommon
  - later data will more strongly reflect longer follow-up

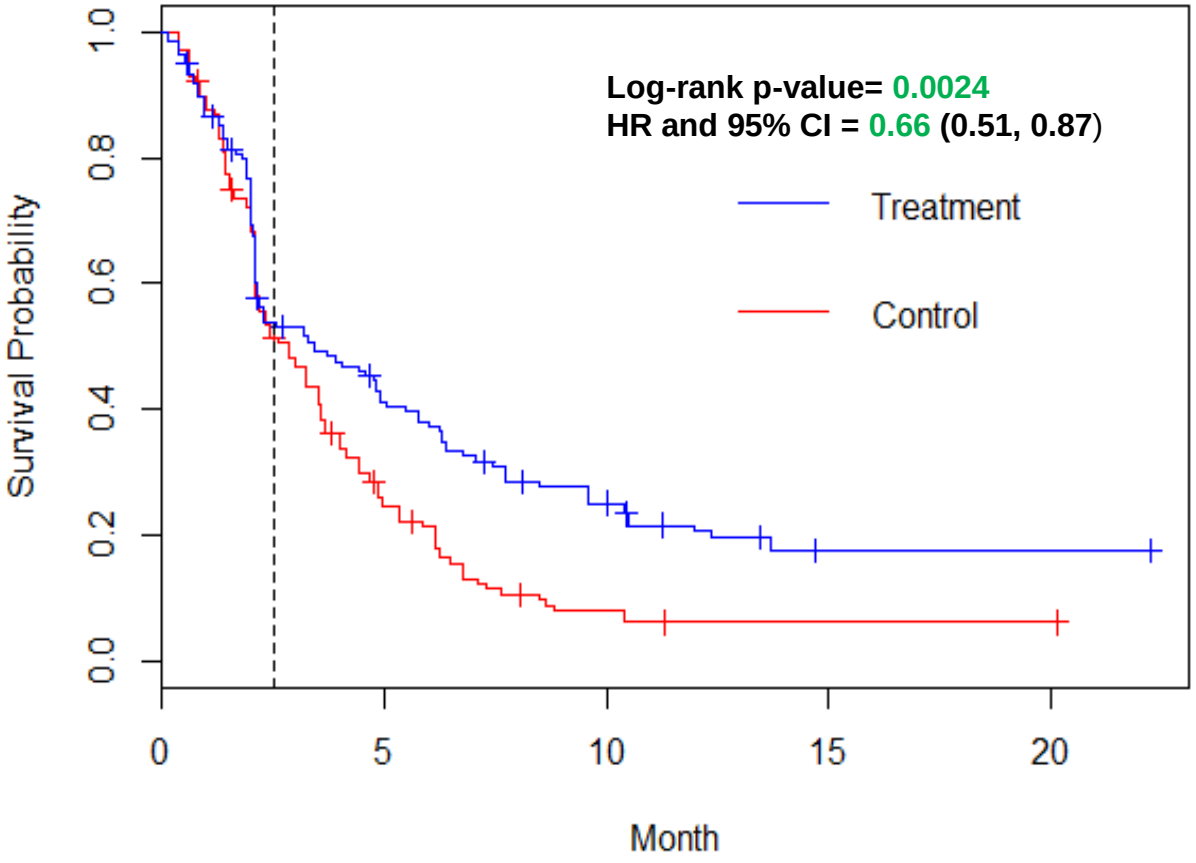
Non-constant effect (contd.)

- Much of what we've discussed implicitly suggests that treatment effect remains constant within a trial
  - *though not necessarily* – we could compute CP conditioning on *any* effect we might hypothesize to govern the remainder of the data
- but quantities like CP(d) or PP with a non-informative prior would be hard to rationalize if we weren't assuming constant effect

Non-constant effect (contd.)

- Acknowledging this possibility might lead us to set criteria *even more cautiously* than suggested elsewhere in this course
- A pre-specified rationale for a possible direction or specific nature of non-constancy can be very helpful
- For non-PH, we might consider hypothesizing a model describing a potentially changing effect, and applying it to account for the different mixture of follow-up times in different subsets of trial data

Example: Delayed treatment effect



- Treatment effect emerges later in the trial; example adopted from an Immuno-oncology trial

Example: Delayed treatment effect *(continued)*

Information Fraction	No. of Events	Time (month)	HR	95% CI
22%	49	1.4	0.906	(0.52, 1.59)
49%	110	2.1	0.933	(0.64, 1.36)
52%	118	2.2	0.971	(0.68, 1.39)
62%	140	3.2	0.843	(0.60, 1.18)
81%	183	5.4	0.702	(0.52, 0.94)
96%	218	10.0	0.651	(0.50, 0.85)
100%	228	22.3	0.664	(0.51, 0.87)

Overall follow-up is low even with 80% events

Treatment effect emerges late in the trial

Futility with non-constant effect

- Conservative futility boundaries at higher information fraction seems sensible
  - e.g. futility threshold with HR > 1 at ≥50% information
- Model-based approaches
  - calculating PP considering “promising effect” in future
  - requires assumption: needs justification and understanding of operating characteristics
- Different “non-constant” treatment effect scenarios exist
  - *one size does NOT fit all*

Special topic: non-inferiority

- N-I trial: objective is to exclude a negative outcome for an investigational treatment relative to an active control, defined by a pre-specified non-inferiority margin (say, δ)
- Almost everything we've discussed translates directly to N-I trials, accounting for the shifted hypotheses:

$$H_0: \Delta < -\delta \quad \text{vs} \quad H_A: \Delta \geq 0$$

- Thus, for example, statistical properties of a futility rule that corresponds to an estimate d in a superiority trial will be exactly the same as those of an estimate $\hat{\theta} - \delta$ in an identically-powered N-I trial

Non-inferiority (contd.)

- However, the practical issues, in terms of the ethics – strategic balance, may play out differently, resulting in different choices
- In superiority trials, sound futility rules usually correspond to outcomes where the efficacy signal is *comparable* between the treatments
- For N-I, the signal corresponding to a rule with the same properties can be quite negative, and allows the possibility that the investigational group is *harmful* (relative to control)

Non-inferiority (contd.)

- This might lead us, on an ethical basis, to set more aggressive rules?
  – and accept larger power loss?

- Ultimately, the principle remains as before: case-by-case quantification and balancing of the various trade-offs should lead to decisions as to how aggressive or cautious we should be

Case 1

- *"When 20% of the data is available, continue the trial as long as the conditional power (assuming the original Δ), is at least 5%"*
- This would correspond to z = - 4.6
- Basically impossible to reach even under H_0
- A substantial signal of harm

Case 2

- *"When the endpoint has been assessed on ⅔ of the patients, continue the study only if the conditional chance of success, computed assuming that the observed effect is the true effect, is at least 70%"*
- This is quite late in the trial; enrollment may be complete, and it's likely that a very high proportion of resources has been expended
- Yet the "failure" threshold clearly corresponds to an observed effect greater than the value that would be significant at the end of the trial!

Conclusions

- A futility scheme should be implemented with careful consideration of its motivation and objectives, and quantification of relative costs and trade-offs
- Familiar expression scales can be a useful device for describing criteria, but are not a substitute for sound investigation of operating characteristics
- Sensible futility criteria often correspond to quite poor observed outcomes, and it is important that trial personnel understand this
- Effective implementation of a sound monitoring plan, including communication with DMC, plays an important role

References

1. Chang WH, Chuang-Stein C (2004). Type I error and power in trials with one interim futility analysis. *Pharm Stat* 3:51-59.
2. DeMets DL (2006). Futility approaches to interim monitoring by data monitoring committees. *Clin Trials* 3:522-9.
3. Ellenberg S, Fleming TR, DeMets DL (2002). *Data Monitoring Committees in Clinical Trials*. Chichester: Wiley.
4. Emerson SE, Kittelson JM, Gillen DL (2005). On the use of stochastic curtailment in group sequential clinical trials. University of Washington Biostatistics Working Paper Series #243: http://www.emersonstatistics.com/dnlds/bep243.pdf.
5. Emerson SS, Kittelson JM, Gillen DL (2007). Frequentist evaluation of group sequential clinical trial designs. *Stat Med* 26:5047-80.
6. FDA CDER (2010). Adaptive Design Clinical Trials for Drugs and Biologics. link.

References *(continued)*

7. Friedlin B, Korn EL, Gray R (2010). A general inefficacy interim monitoring rule for randomized clinical trials. *Clin Trials* 7:197-208.
8. Gallo P, Mao L, Shih VH (2014). Alternative views on setting clinical trial futility criteria. *J Biopharm Stat* 24:976-93.
9. Jennison C, Turnbull BW (2000). *Group Sequential Methods with Applications to Clinical Trials*. London: Chapman & Hall/CRC.
10. Lachin JM (2005). A review of methods for futility stopping based on conditional power. *Stat Med* 24:2747-64.
11. Lachin JM (2009). Futility interim monitoring with control of type I and type II error probabilities using the interim Z-value or confidence limit. *Clin Trials* 6:565-73.
12. Lan KKG, Simon R, Halperin M (1982). Stochastically curtailed tests in long-term clinical trials. *Sequential Analysis* 1:207-19.
13. Lan KKG, DeMets DL (1983). Discrete sequential boundaries for clinical trials. *Biometrika* 70:659-63.
14. Lan KKG, Wittes J (1988). The B-value: A tool for monitoring data. *Biometrics* 44, 579–585.

References *(continued)*

15. Ohrn CF (2011). Group sequential and adaptive methods – topics with applications to clinical trials. Ph.D. dissertation, University of Bath.

16. Pampallona S, Tsiatis AA, Kim K (2001). Interim monitoring of group sequential trials using spending functions for the type I and type II error probabilities. *Drug Inf J* 35:1113-21.

17. Snapinn S, Chen MG, Jiang Q, Koutsoukos T (2006). Assessment of futility in clinical trials. *Pharm Stat* 5:273-81.

18. Spiegelhalter DJ, Freedman LS, Blackburn PR (1986). Monitoring clinical trials: conditional or predictive power? *Cont Clin Trials* 7:8-17.

20. Whitehead J, Stratton I (1983). Group sequential clinical trials with triangular continuation regions. *Biometrics* 39: 227–236.

21. Zhang Y, Clarke WR (2010). A flexible futility monitoring method with time-varying conditional power boundary. *Clin Trials* 7:209-18.