Viewpoints on Setting Clinical Trial Futility Criteria

Vivian H. Shih, AstraZeneca LP
Paul Gallo, Novartis Pharmaceuticals

BASS XXI
November 3, 2014

Reference

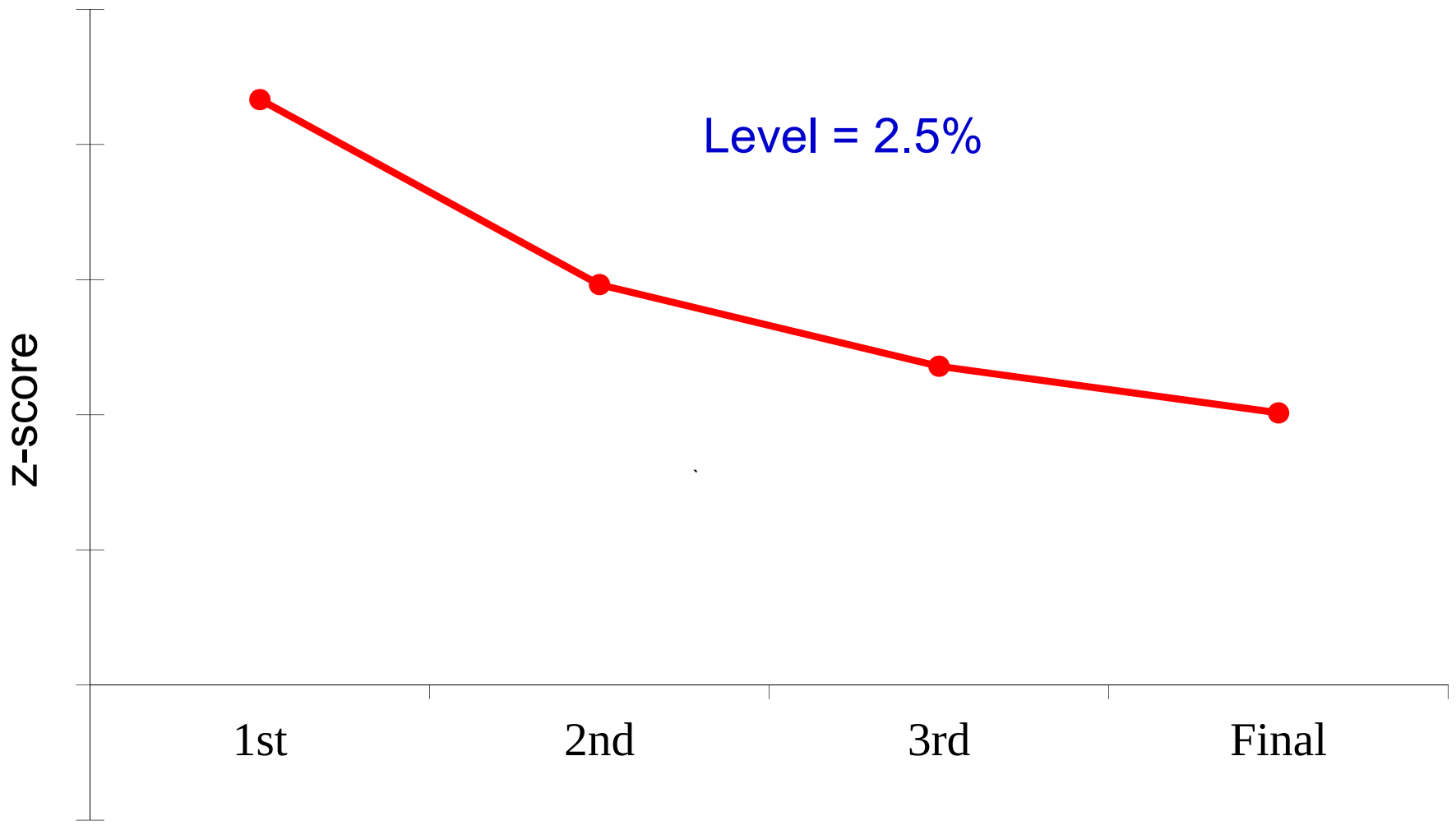
- Based on:

Gallo P, Mao L, Shih VH (2014). Alternative views on setting clinical trial futility criteria. *Journal of Biopharmaceutical Statistics*, 24(5):976-993.

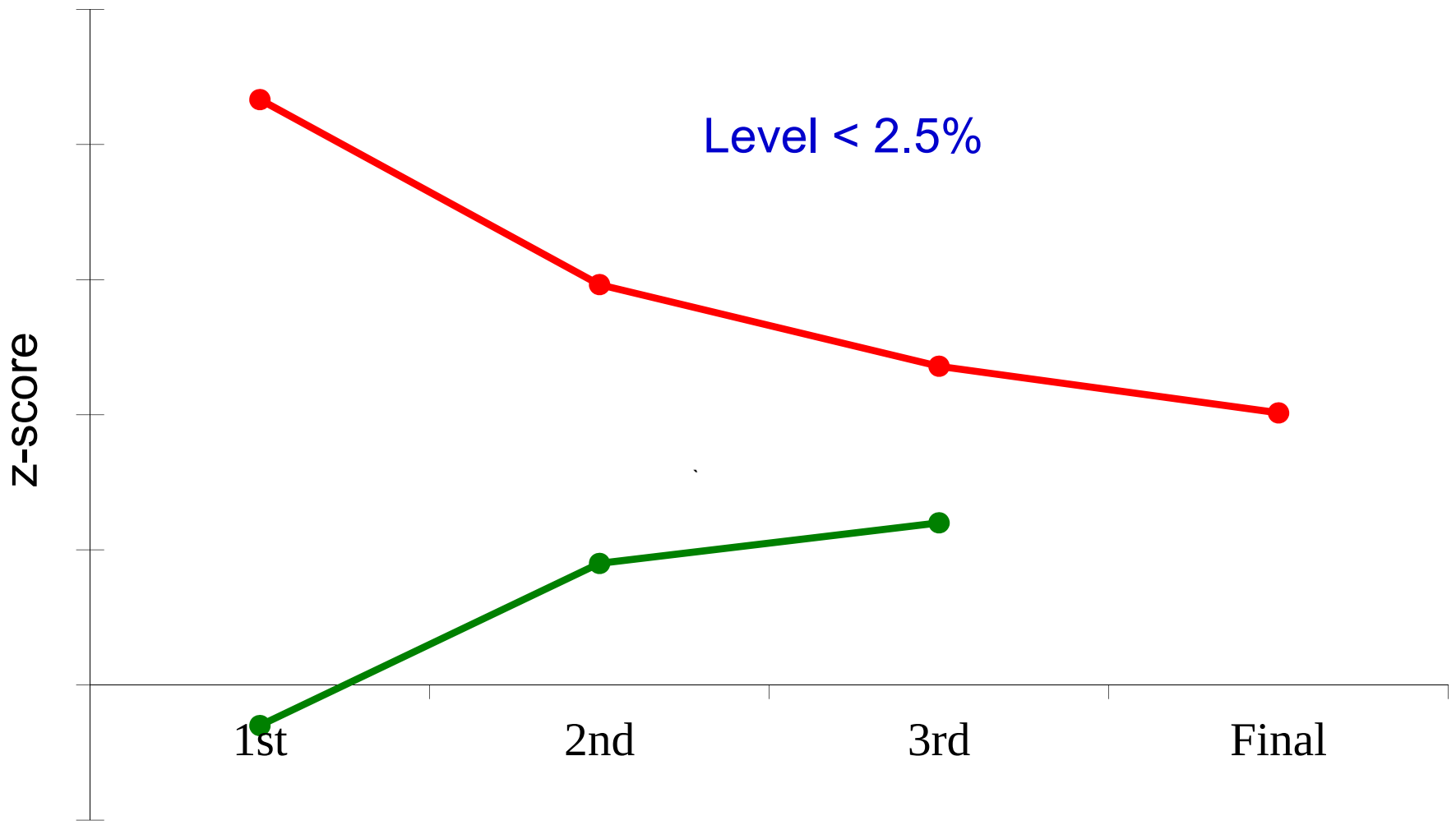
Stopping Trials for Lack of Effect

- *Futility:* based on interim results, a trial seems unlikely to achieve its objectives
- Specific motivations for allowing the possibility of early stopping are situation-dependent, but generally obvious
  - Time
  - Cost
  - Ethics
  - Resource reallocation

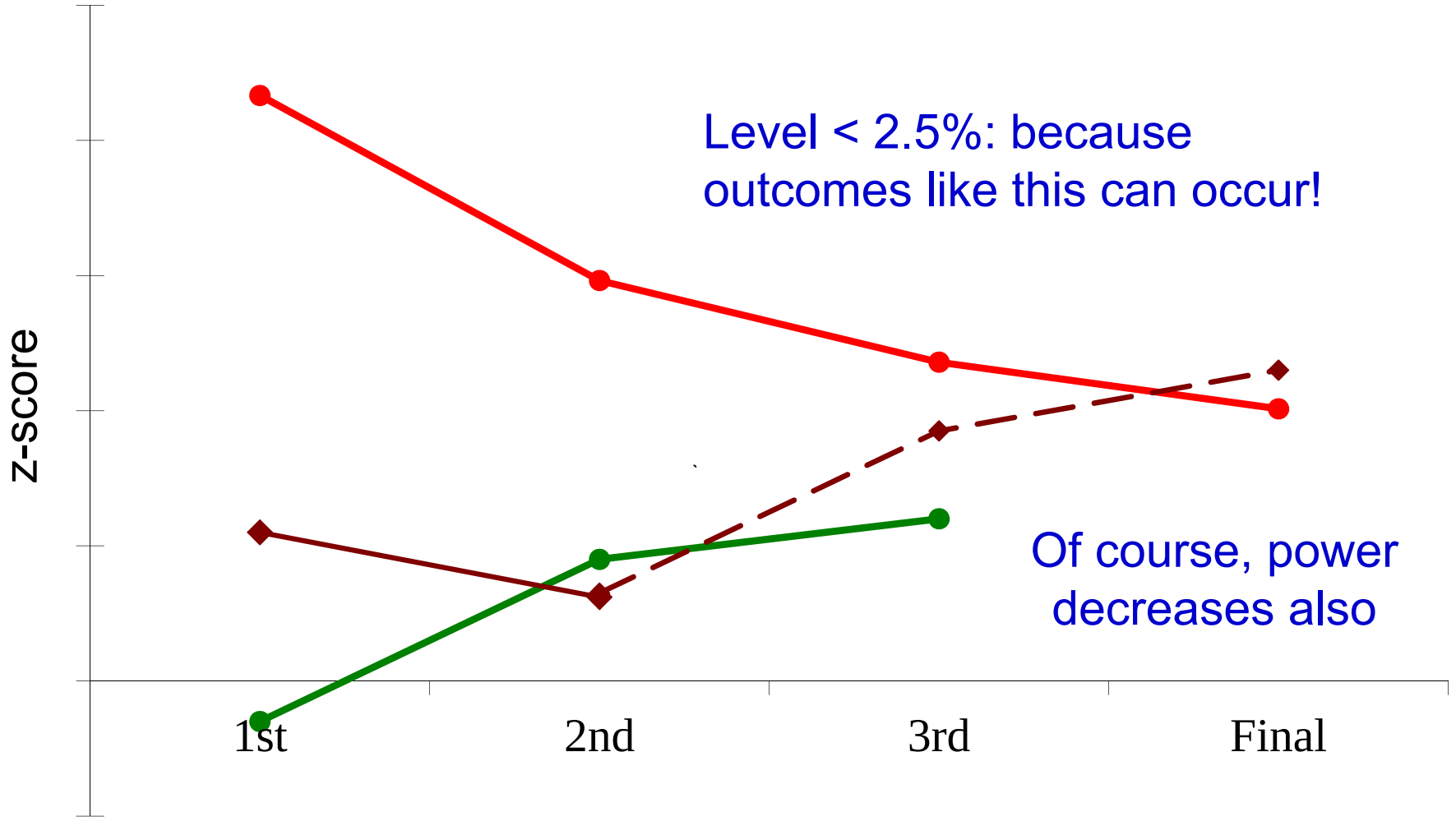
Typical Efficacy Scheme



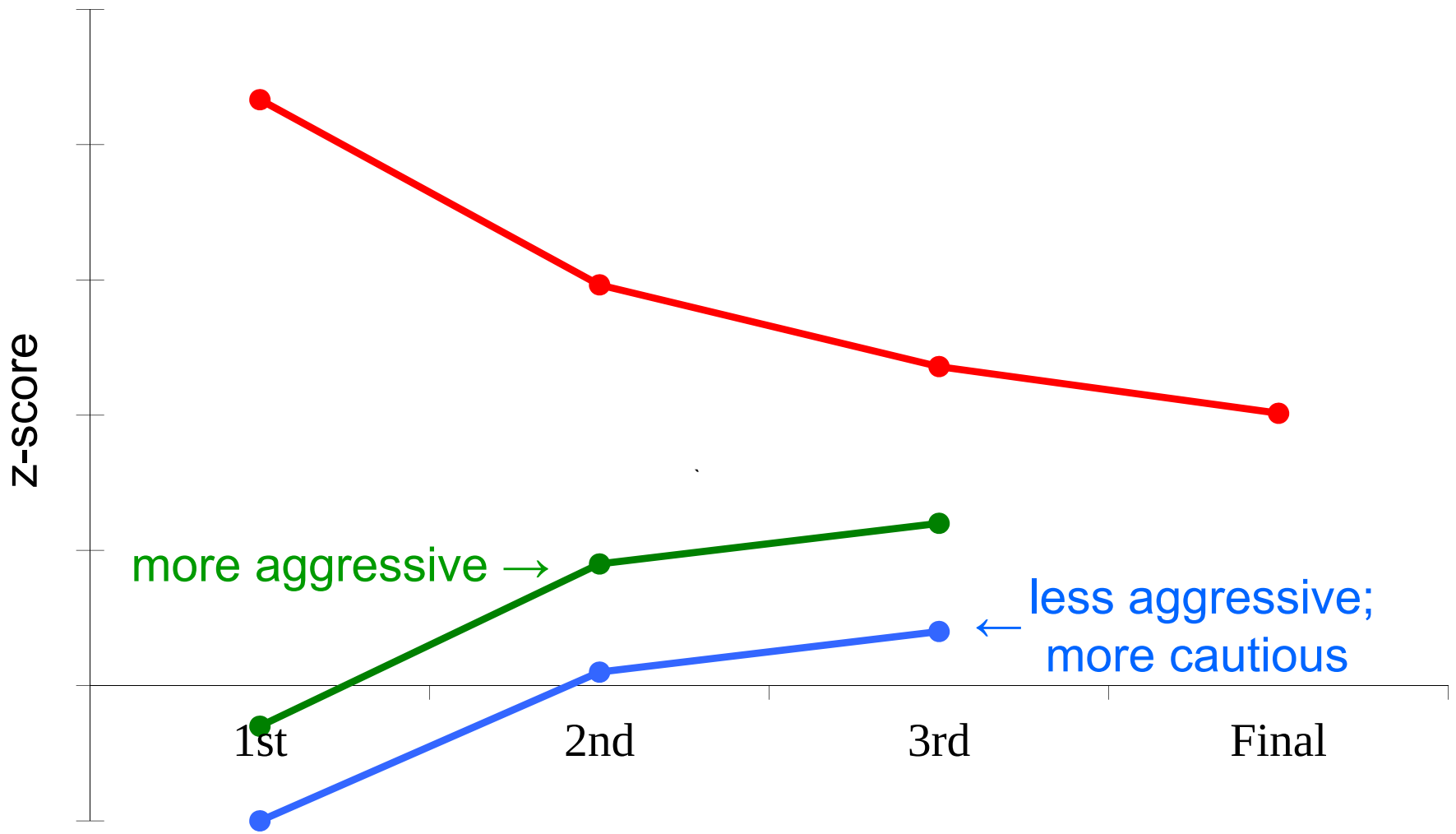
Impose a Futility Boundary



Level is Decreased



Terminology – "*Aggressiveness*"



Assumptions

- *Non-binding* futility boundary
  - i.e., we don't modify success criteria to *buy back* lost α
  - consistent with an understanding that futility is a "soft" decision (*guidelines*, not *rules*)
- We'll compare schemes in terms of *power loss*
  - Another option: *increase SS* to regain lost power
- No early stopping for efficacy
- Notation:
  - Δ = hypothesized design effect, d = point estimate
  - I = information time, z_I = corresponding test stat

Tools for Addressing Futility

- Conditional power (CP) calculations
  - usually conditions on the original study alternative
  - sometimes on other quantities (e.g. point estimate)
- Predictive probability (PP)
  - usually non-informative prior
- Beta-spending functions
  - describes cumulative Type II error across the interim and final looks
- Others (B-value, stochastic curtailment, reject H_A)

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*
    - *are they really?*
- What's the real issue?
  - Emerson *et al* (2005): operating characteristics

Consultation Examples

- Two actual proposals / consultations for futility criteria:
  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 . . .

Possible Scenarios

	Trial outcome / True state of nature	
Interim decision	Success	Failure
Stop for futility	(Incorrect)	(Correct)
Continue	Correct	Incorrect

- Generally, we'd like "*small*" chances of outcomes on the diagonal
  - but of course decreasing one increases the other . . .

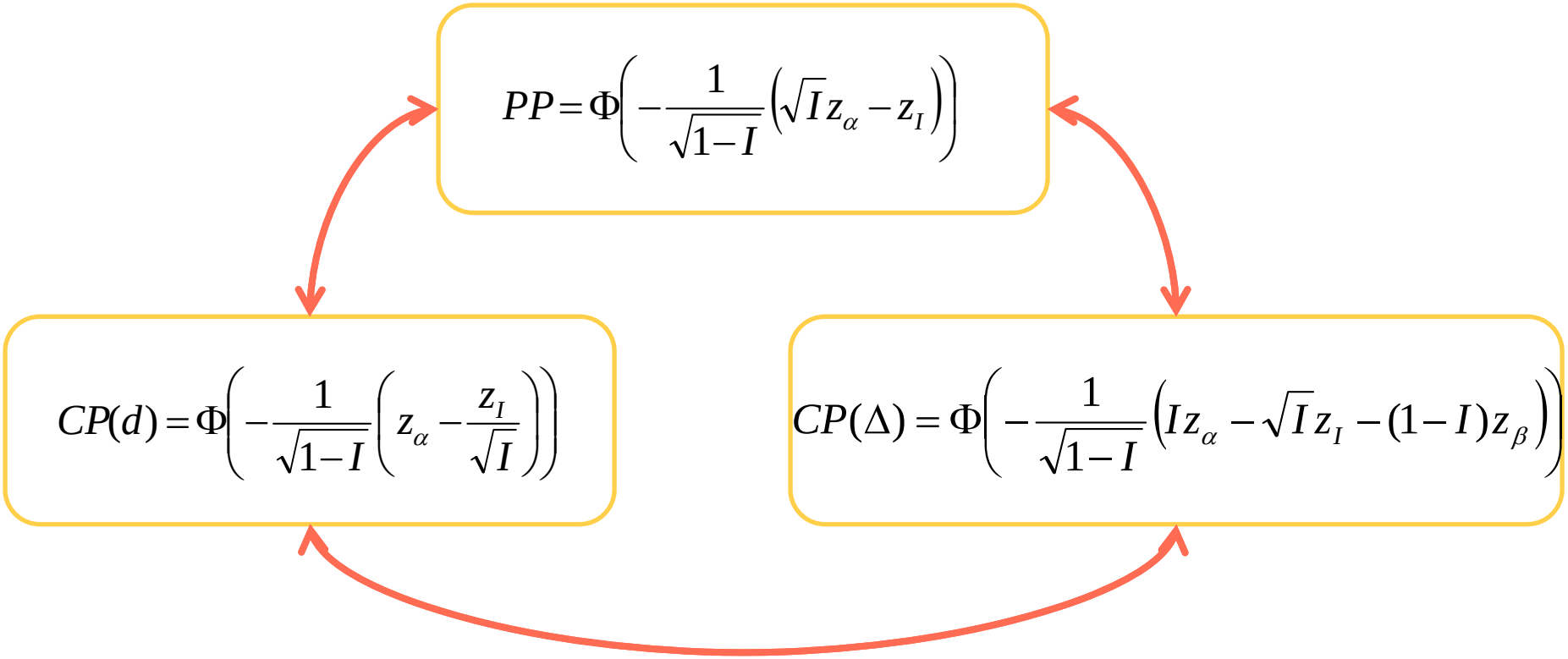
Striking a Balance

- We can't control error rates nearly as well as we typically do for an entire study
- *{Stopping when we should}* versus *{continuing when we should}* are always *in conflict*
- We should aim to strike an appropriate balance while limiting the chance of wrong decisions
- Proposal: usually, the *worse transgression* is *stopping a trial which would have been successful*

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, with a single look at $\mathbf{I}$ = 50%, say we set a criterion of PP = 20%
- The same rule can be expressed as:
  - CP = 62%
  - CP(*d*) = 12%
  - 'Beta spent' = 6.7%
- *Question*: is the *scale* on which we express a futility criterion *really that important*?

Interrelationships



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

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

90% Power for Δ, I = 0.5

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

scales for expressing futility rule

behavior

Aggressiveness / Caution

- {We need not focus only on H_0, H_A; other definitions of *weak effect*, *likely success*, etc. could be considered and evaluated}
- 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* ?
- Incorporate into a loss function?

How Aggressive?

- What are the dimensions of savings of interest?
  - e.g., $, resources, time, patients, etc.?
- What factors affect the trade-offs?
  - *fixed* vs *variable* costs
  - *prior belief*: how much faith? / evidence from related trials
  - *ethics:* unknown safety risks for experimental treatment
  - *upside*: blockbuster, or "*me too*"?

When to Evaluate Futility?

- Again, a *conflict* :
  - stopping earlier yields potentially greater savings; but . . .
  - less ability to distinguish between scenarios which should / should not justify continuing
- Futility behavior improves with information in *2 ways*:
  - added precision from more data
  - less data still to come that can overturn a poor trend
- Previous example, criteria: z = 0.5
  - at $\mathcal{I}$ = ½, we saw that power loss was 1.3%
  - at $\mathcal{I}$ = ¼, it's 9.2%

Multiple Futility Looks

- *Why not?*
  - i.e., in long-term studies
- There are practical limitations (on both ends) to when looks should take place
  - too early, too late: *no point*
- The existence of a *later look* might impact the choice of criteria at a *prior look*
  - because a decision to continue *does not commit to trial completion*, but only to proceed *until a later point where data is more mature*

Quantifying the Trade-offs

- How to extend to multiple looks?
- The cost of incorrect stopping:
  - how about "*power loss across the whole scheme*"?
  - of course, different ways to achieve this.
    - perhaps, equal power loss at each analysis?
- The benefit of correct stopping:
  - ASN: average sample size under H_0

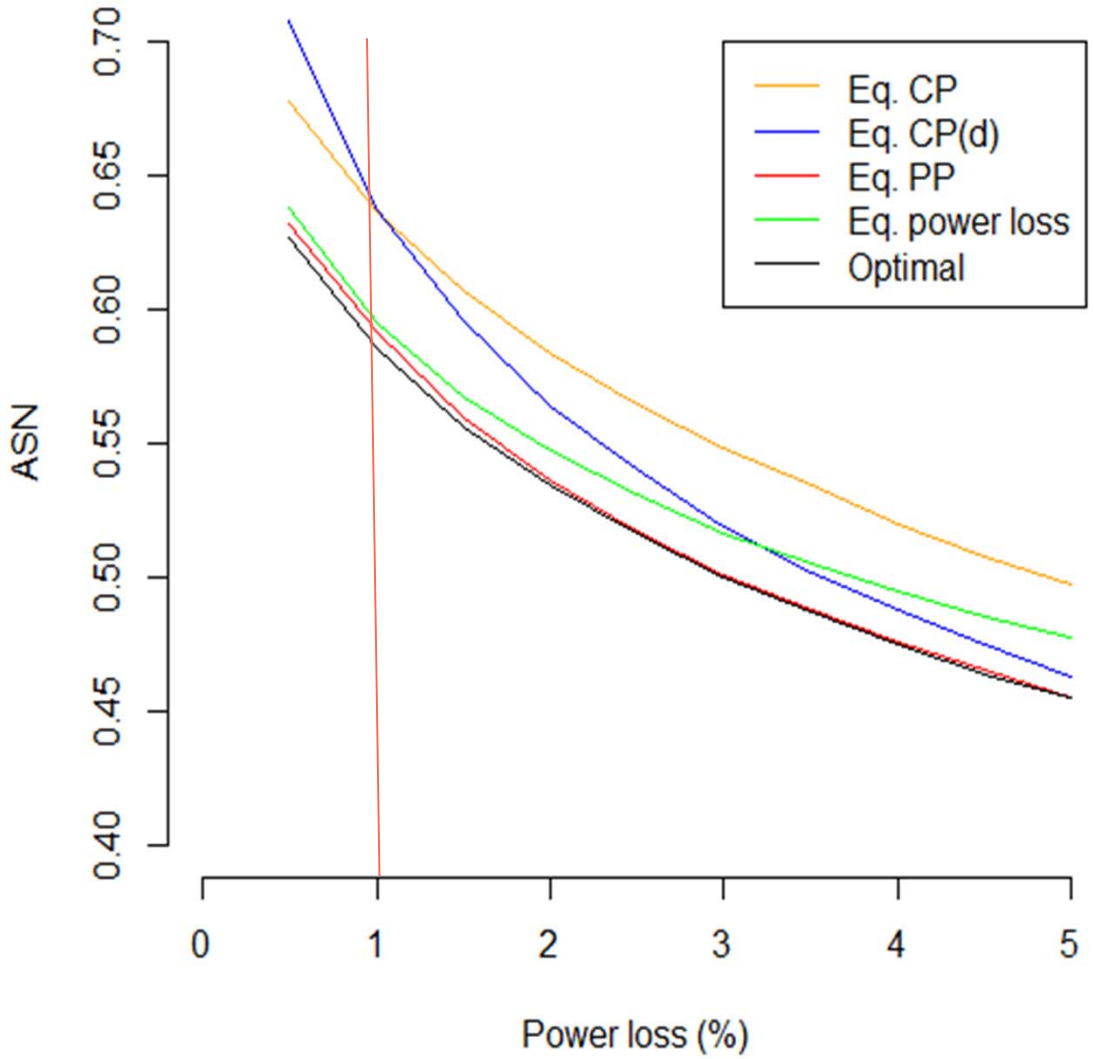
Multiple-look Considerations

- Ideally, we could describe a scheme *simply*
- *Now the scale matters!*
  - equal criteria across looks on one scale could be very unequal on another scale
- Example: say that at $I = ½$, we judge CP = 50% to be a sensible criterion
  - What if we also used the same rule at $I = ¼, ¾$?
  - PP across the 3 looks: 1.3%, 10.0%, 23.0%
  - But is there any reason to expect that the *same CP threshold behaves well* at the other timepoints?
    - *hint*: it doesn't . . .

Optimality

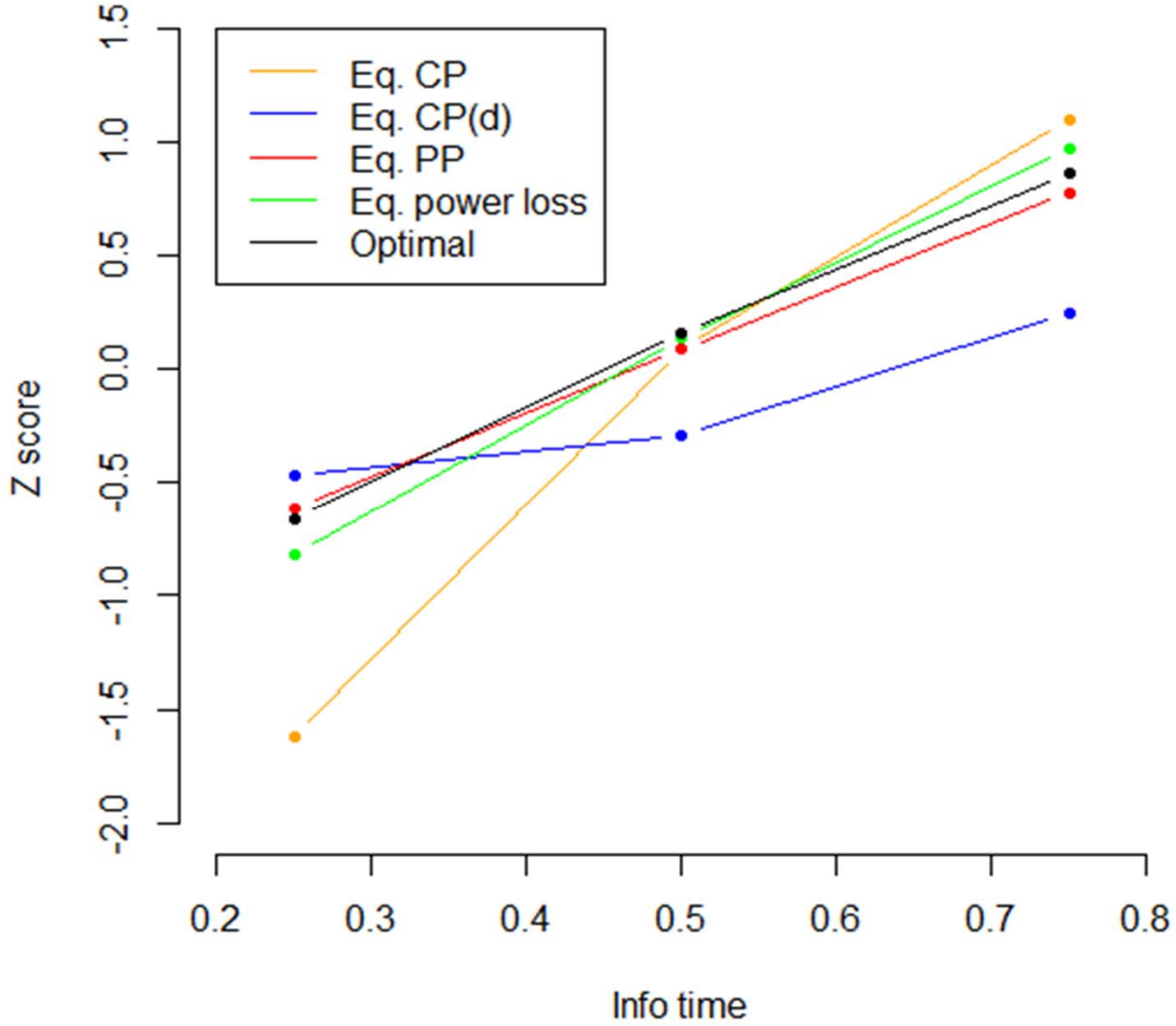
- Optimal boundaries: For a given schedule of analyses, and a specified amount of power loss, we can define boundaries that minimize ASN
  - optimization done by grid search
- In what follows, we'll assume 3 looks at $I = 0.25$, 0.50, 0.75, and describe various boundaries:
  - equal CP
  - equal CP(*d*)
  - equal PP
  - equal power loss
  - optimal (as above)

ASN vs Power Loss



- Equal PP boundaries at the 3 looks is quite close to optimal.
- Equal CP fares particularly poorly.

Comparing Boundaries: 1% Power Loss

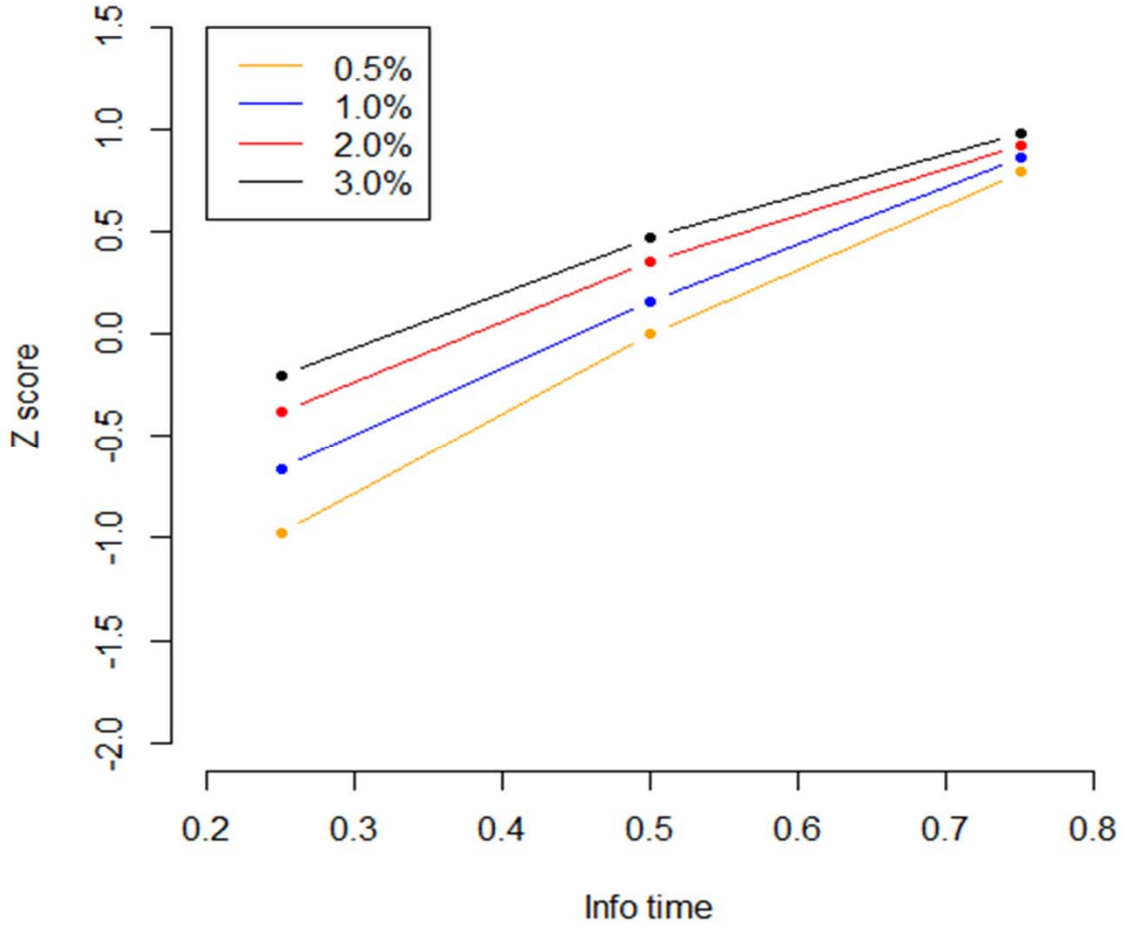


1% Power Loss Boundaries

			Futility boundary on Z-scale		
Boundary type	**Common value**	**ASN**	**1st look**	**2nd look**	**3rd look**
Equal CP	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
Optimal	-	0.585	-0.660	0.160	0.860

What do “Good” Boundaries Look Like?

- Optimal boundaries for various amounts of power loss:



What do “Good” Boundaries Look Like?

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

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

Message

- My experience: trial teams *encouraged* by the knowledge that their study proceeded beyond a futility analysis, 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*

Back to (Flawed) Consultation Examples

- *"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

Consultation Example

- *"⅔ into the trial, continue the study only if the conditional chance of success, computed under the assumption that the observed effect is the true effect, is at least 70%"*
- As stated, this must correspond to an observed effect greater than the value that would be significant at the end of the trial

Conclusion

- 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
- Predictive probability seems to have some benefits in terms of easy description of a scheme which might have desirable properties
- Sensible futility criteria often correspond to quite poor observed outcomes, and it is important that trial personnel understand this