Northwestern Medicine®
Feinberg School of Medicine

Interim analysis in randomized controlled trials: utility and use cases

Jody D. Ciolino, PhD
Associate Professor
Northwestern University Data Analysis and Coordinating Center (NUDACC)
Biostatistics Collaboration Center (BCC)
jody.ciolino@northwestern.edu

Northwestern Medicine®
Feinberg School of Medicine

Disclaimer:

The views presented today are my own. They do not represent those of the statistical community nor those of Northwestern University at large.

I have no relevant conflicts of interest.

Coauthors:
Dr. Alex Kaizer at University of Colorado
Dr. Lauren Bonner at Northwestern University

Recent review…

Journal of Clinical and Translational Science

www.cambridge.org/cts

Clinical Research Review Article

Cite this article: Ciolino JD, Kaizer AM, and Bonner LB. Guidance on interim analysis methods in clinical trials. *Journal of Clinical and Translational Science* **7**: e124, 1–9. doi: 10.1017/cts.2023.552

Received: 25 January 2023
Revised: 8 May 2023
Accepted: 9 May 2023

Keywords:
Interim analysis; clinical trials; randomized controlled trial; guidance; efficacy; futility

Corresponding author: J. D. Ciolino;
Email: jody.ciolino@northwestern.edu

Guidance on interim analysis methods in clinical trials

Jody D. Ciolino[1], Alexander M. Kaizer[2] and Lauren Balmert Bonner[1]

[1]Department of Preventive Medicine (Biostatistics), Northwestern University Feinberg School of Medicine, Chicago, Illinois, USA and [2]Department of Biostatistics & Informatics, Colorado School of Public Health, Aurora, Colorado, USA

Abstract

Interim analyses in clinical trials can take on a multitude of forms. They are often used to guide Data and Safety Monitoring Board (DSMB) recommendations to study teams regarding recruitment targets for large, later-phase clinical trials. As collaborative biostatisticians working and teaching in multiple fields of research and across a broad array of trial phases, we note the large heterogeneity and confusion surrounding interim analyses in clinical trials. Thus, in this paper, we aim to provide a general overview and guidance on interim analyses for a nonstatistical audience. We explain each of the following types of interim analyses: efficacy, futility, safety, and sample size re-estimation, and we provide the reader with reasoning, examples, and implications for each. We emphasize that while the types of interim analyses employed may differ depending on the nature of the study, we would always recommend prespecification of the interim analytic plan to the extent possible with risk mitigation and trial integrity remaining a priority. Finally, we posit that interim analyses should be used as tools to help the DSMB make informed decisions in the context of the overarching study. They should generally not be deemed binding, and they should not be reviewed in isolation.

Goals for today

Inspired by "Guidance on interim analysis methods in clinical trials" by Ciolino, Kaizer, and Bonner (Journal of Clinical and Translational Science; 2023)

- Clarify what we often mean by "interim analysis" and "interim monitoring"
- Provide an overview of common types of interim analyses
- Present example studies that involved interim analysis methods
- Illustrate potential successes and challenges in implementing these methods

Interim Monitoring vs. Analysis

These terms are often confused + there are a lot of variations of meaning even within each overarching topic

- For purposes of our discussion…
- Interim **monitoring**:
  - **Processes** and **systems** important for study conduct and reporting
  - **Data integrity** and general cleanliness = primary focus
  - There is almost always an **ethical/safety component** when we talk about human subjects' research

Interim Monitoring vs. Analysis

These terms are often confused + there are a lot of variations of meaning even within each overarching topic

- For purposes of our discussion…

- Interim **analysis**:
  - Referring to **statistical <u>tools</u>** used to **guide** study design modifications (most often revolving around recruitment targets) – "go/no-go", stopping bounds, adding arms, removing arms, sample size re-estimation, etc.
  - **<u>Need not involve a formal hypothesis test (often does, not always!)</u>**

Reasoning for interim monitoring + analyses

Interim monitoring

Ensuring trial conduct according to protocol, **efficiently** and ethically

- Consent verification
- Data quality checks
- Process measures – screening rates, dropout rates, adherence, etc.
- **Safety** and major event reporting – adverse events, deviations, etc.

Findings on interim monitoring help drive study conduct decisions.

Interim analysis

Ensuring the study is set up for success (it addresses the research question + is equipped to address the research question) with an adequate risk-benefit profile

- Making **efficient** use of participant time, data, and general resources
- Continual assessment of **safety** signals

Findings on interim analysis help drive study design considerations.

Diving deeper into **interim analysis**

- The term “interim analysis” in clinical trials has multiple meanings.
- In general, interim analyses help **guide decisions on overall clinical trial** modifications, specifically those pertaining to the study sample size or recruitment targets.
- We often encounter a lot of confusion and misunderstanding when it comes to interim analysis…

Some common misconceptions

"But I don't want to spend any of my alpha early on, especially for a small study."

"Won't an interim analysis raise red flags and be met with more scrutiny when we go to publish?"

"We can't do an interim analysis if this is the first study to explore this intervention in this population."

"I don't like using these statistical rules to dictate trial decisions because we won't be able to keep going even if we think there is reason to finish the study if we cross some threshold."

"We crossed our threshold, and it is significant…we have to stop."

"Don't we have to do an interim analysis?"

Shedding some light on the general topic

We "bucket" interim analysis methods into 4 main areas

1. Efficacy
2. Futility
3. Safety
4. Sample Size Re-Estimation

…the next set of slides give high-level take home points for each type…

Table 1. Summary of Interim Analysis Types

	Explanation	Justification for Use
Efficacy		
Futility		
Safety		
Sample size re-estimation		

Table 1. Summary of Interim Analysis Types

	Explanation	Justification for Use
Efficacy	• Early termination of a trial that is **showing promising results** • **Control of type I error** through group sequential methods or alpha spending functions	• Usually for longer, larger studies and later phases of research • **Ethical imperative for a promising treatment to reach the entire target clinical population**
Futility		
Safety		
Sample size re-estimation		

Table 1. Summary of Interim Analysis Types

	Explanation	Justification for Use
Efficacy		
Futility	• Early termination of a trial that **is not likely to achieve the intended objective** (e.g., little chance of finding a "significant" treatment effect at the end of the study) • Employed through group sequential methods, error spending functions, conditional power, or predictive power	• **Reduces costs, resources, and patient burden** for a trial with a low probability of "success" • Usually for mid-late phase studies • Helpful in the context of recruitment and retention challenges
Safety		
Sample size re-estimation		

Table 1. Summary of Interim Analysis Types

	Explanation	Justification for Use
Efficacy		
Futility		
Safety	• **Early termination** (or pausing) of a trial **for safety concerns** • Should be coupled with efficacy analyses to evaluate the benefit-to-risk ratio	• Incorporated across all phases of research • **Particularly important for vulnerable populations and high-risk interventions** with more "serious" outcomes (e.g., death)
Sample size re-estimation		

Table 1. Summary of Interim Analysis Types

	Explanation	Justification for Use
Efficacy		
Futility		
Safety		
Sample size re-estimation	• Reassessment of the sample size required to **ensure adequate power** using updated information from interim trial data • Can be blinded or unblinded • May not necessarily spend alpha	• Allows for **interim look at assumptions** (standard deviations, event rates, correlations, etc.) • May be particularly useful for mid-late phase studies

Table 1. Summary of Interim Analysis Types

	Explanation	Justification for Use
Efficacy	• Early termination of a trial that is **showing promising results** • **Control of type I error** through group sequential methods or alpha spending functions	• Usually for longer, larger studies and later phases of research • **Ethical imperative for a promising treatment to reach the entire target clinical population**
Futility	• Early termination of a trial that **is not likely to achieve the intended objective** (e.g., little chance of finding a "significant" treatment effect at the end of the study) • Employed through group sequential methods, error spending functions, conditional power, or predictive power	• **Reduces costs, resources, and patient burden** for a trial with a low probability of "success" • Usually for mid-late phase studies • Helpful in the context of recruitment and retention challenges
Safety	• **Early termination** (or pausing) of a trial **for safety concerns** • Should be coupled with efficacy analyses to evaluate the benefit-to-risk ratio	• Incorporated across all phases of research • **Particularly important for vulnerable populations and high-risk interventions** with more "serious" outcomes (e.g., death)
Sample size re-estimation	• Reassessment of the sample size required to **ensure adequate power** using updated information from interim trial data • Can be blinded or unblinded • May not necessarily spend alpha	• Allows for **interim look at assumptions** (standard deviations, event rates, correlations, etc.) • May be particularly useful for mid-late phase studies

Northwestern Medicine®
Feinberg School of Medicine

Efficacy

Interim analysis for efficacy

- Involves a statistical hypothesis test evaluating one arm over another.
- If there is a “large enough” signal early on in the study, it may be ethically imperative and most efficient to stop the study early.
- Threshold for sufficient evidence is subject to debate and topic of statistical research and literature.
- We cannot simply use a “statistically significant” finding to guide this decision…

Interim analysis for efficacy

- Recall: **type I error** = probability of finding a significant result [usually $p<0.05$] when in fact we should not, as there is no effect.

- The more times we "look" at the data (i.e., conduct statistical tests), the more likely we are to find a significant result (i.e., make a type I error).

Interim analysis for efficacy

Type I error illustration for hypothetical null effect trial

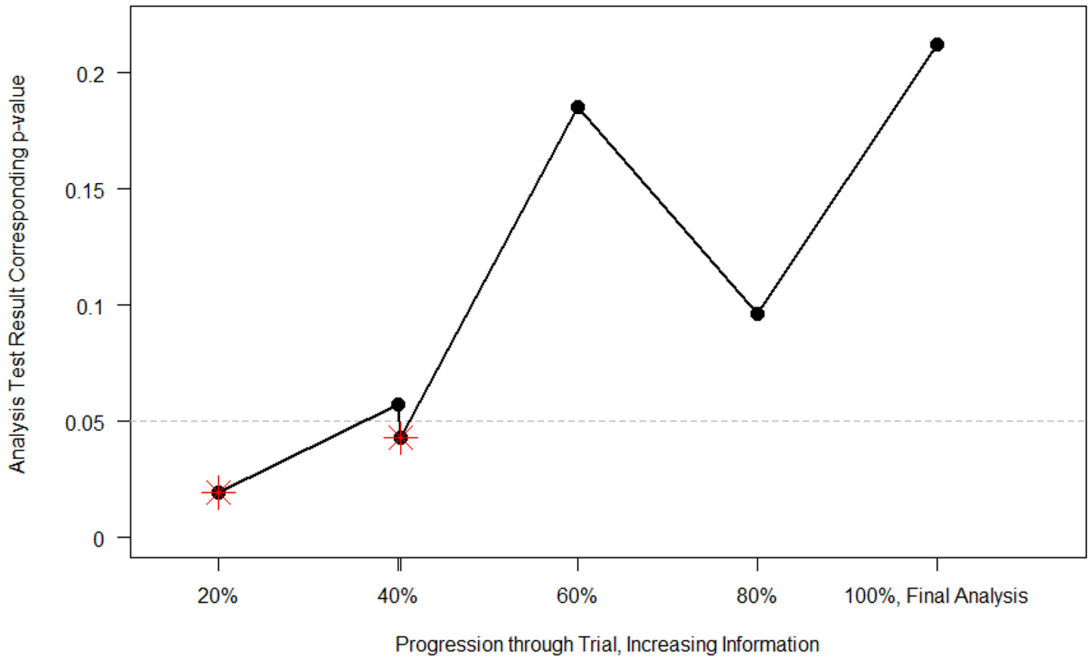
- Consider a hypothetical, two-arm clinical trial
- Binary outcome = "success" of intervention
- Plan to conduct a statistical test at the 5% level of significance

- We can simulate no underlying difference between the two study arms
  - Assume p(success) = 0.30 across arms
  - N=355 planned per arm
  - Simulate under null (H0)
  - IF we conduct a statistical test and find $p<0.05$, we are making a type I error

- There are countless ways the test statistic could "behave" over the course the study, but consider one hypothetical scenario…

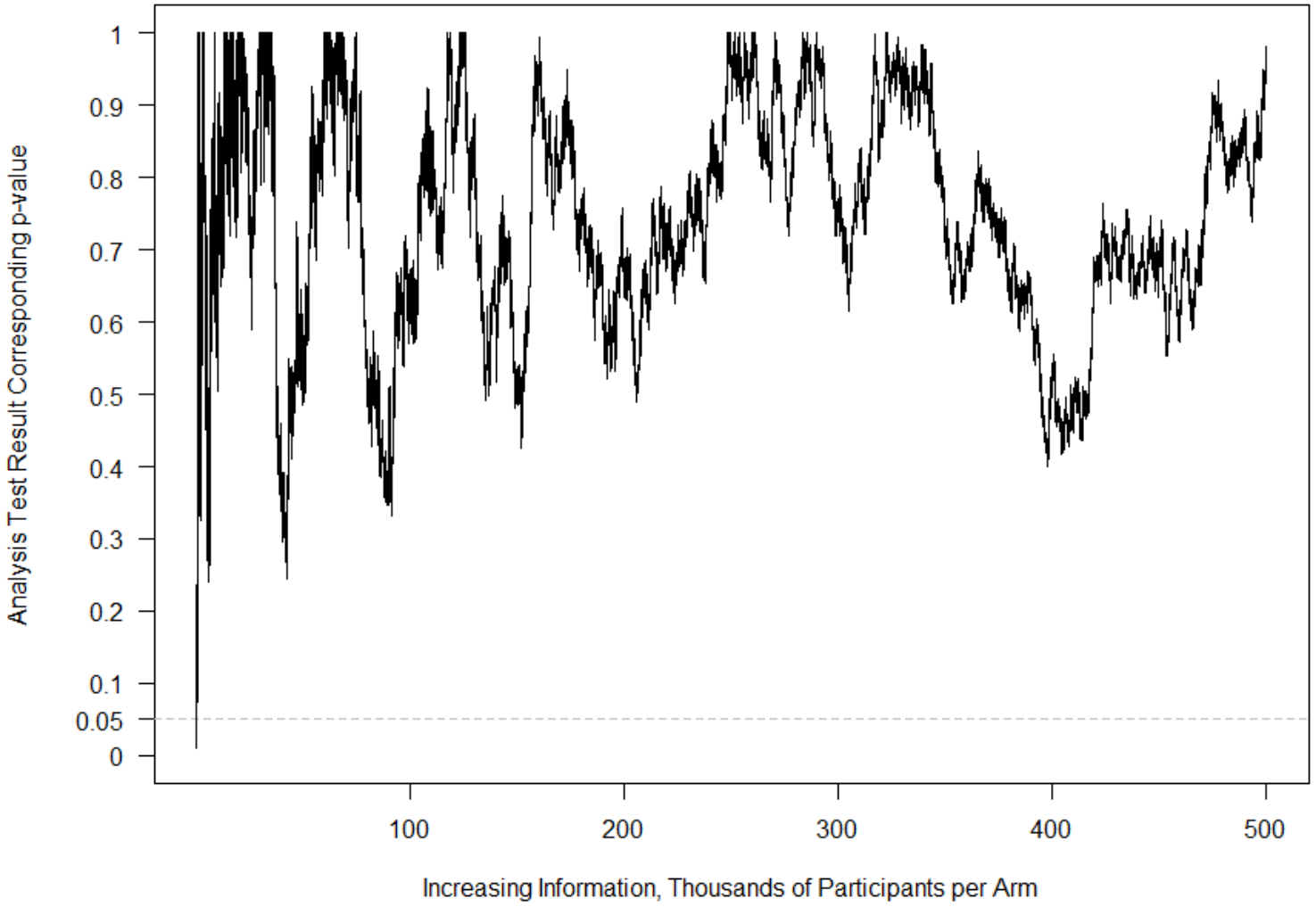
Interim analysis for efficacy

Type I error illustration for hypothetical null effect trial



% of Participants Enrolled with Outcome Data Observed	20%	40%	41%**	60%	80%	100%
N per Arm	71	142	143	213	284	355
p-value	**0.019**	**0.057**	**0.042**	0.185	0.096	0.212
Arm 1 Successes	11	29	29	49	73	93
Arm 2 Successes	24	44	45	62	92	109
Difference in Successes (N)	13	15	16	13	19	16
Difference in Proportions	0.183	0.105	0.112	0.061	0.067	0.045
****between 40-41%, included for illustrative purposes**						

Illustration of test statistic behavior with increasingly large sample sizes



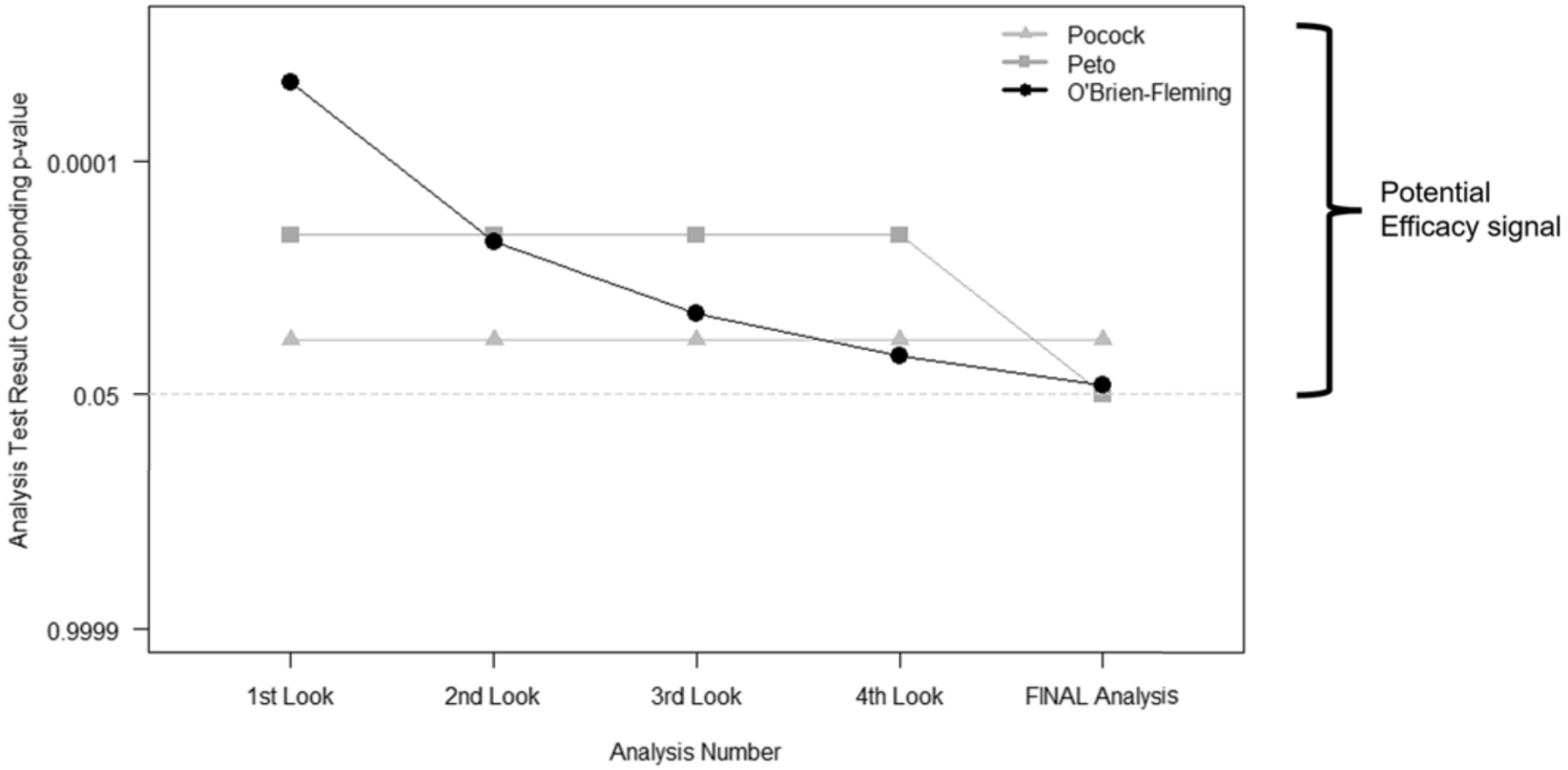
Interim analysis for efficacy

- Recall: type I error = probability of finding a significant result [usually $p<0.05$] when in fact we should not as there is no effect.
- The more times we "look" at the data (i.e., conduct statistical tests), the more likely we are to find a significant result (type I error).
- Utilize **methodology to control type I error** with repeated looks at the data.
- These methods help us decide whether the results at an interim analysis are **"significant enough" to warrant early stopping.**

Interim analysis for efficacy

Common methods for controlling type I error

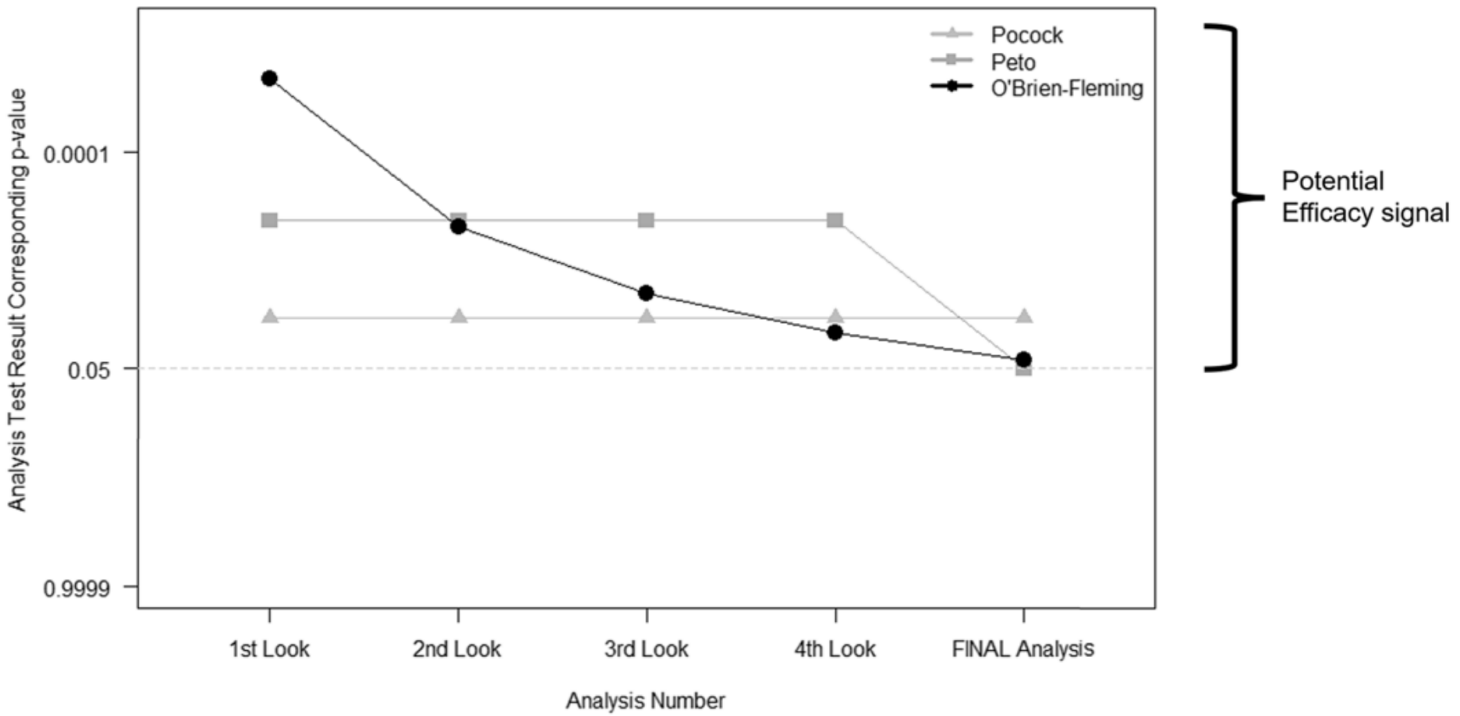
- Another two-arm hypothetical trial
- 4 interim looks + 1 final analysis
- Typical "group sequential" stopping bounds look like this…



Interim analysis for efficacy

Common methods for controlling type I error

- **Pocock** bounds have a constant threshold (p<0.016) for all 5 analyses
- **Peto** bounds have a stringent (p<0.001) threshold for the first 4, then uses p<0.05 at the end
- **O'Brien-Fleming** (perhaps most common) – stringent thresholds early on, and increasingly less stringent as time goes on; final analysis uses p<0.04



Interim analysis for efficacy

- Note: “alpha spending functions” incorporate these ideas but allow for more flexibility.
- Can accommodate different ways of “spending” type I error (differing weights), and the timing of analyses needed not be evenly spaced.
- Sometimes these terms “group sequential” and “alpha spending” are used interchangeably since they are so closely linked.

Alpha Spending Functions Background

- Relies on **information fraction** (roughly how far we are through the trial, statistically), for example…
  - Current ***n*** relative to overall ***N***
  - Current **number of events** relative to **overall expected events**

- An **alpha spending function** is a monotone function relating the information fraction t^* to an alpha level $\alpha(t^*)$
  - t^* goes from 0 to 1
  - $\alpha(t^*)$ goes from 0 to α (the overall Type I error rate you hope to preserve for the study)
  - Alpha spending function can take on different forms, as long as it meets the above criteria

Visuals of Alpha Spending

Cook TD, DeMets DL. Introduction to statistical methods for clinical trials. CRC Press; 2007 Nov 19.

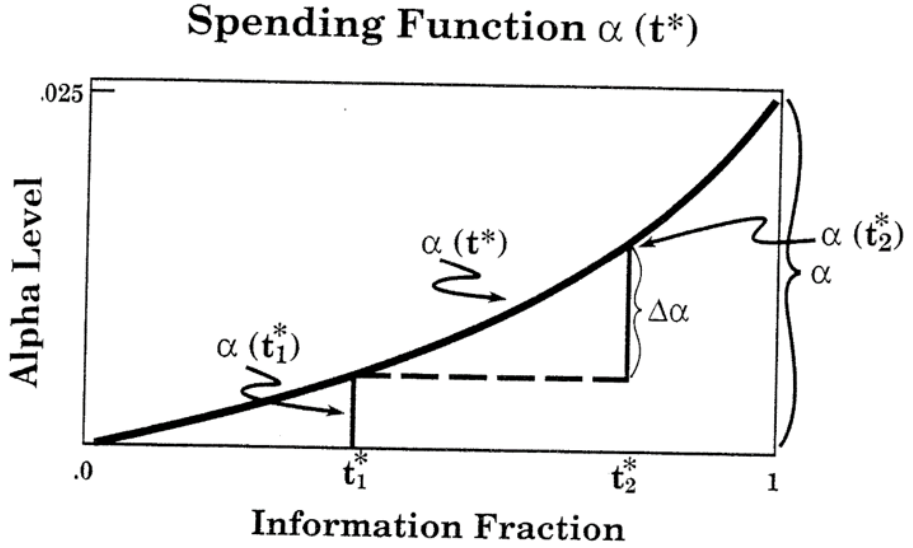


Figure 1. Alpha spending function indicating additional type I error rate, Δα, allocat between interim analyses t_1^* and t_2^*.

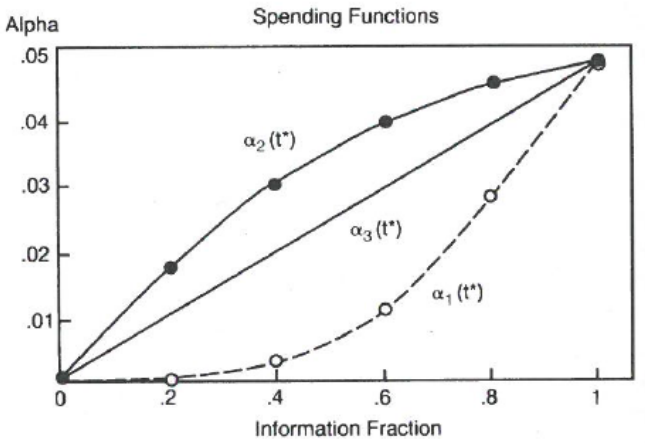


Fig. 16.9 Alpha-spending functions for $K=5$, two-sided $\alpha=0.05$ at information fractions $t^*=0.2$, 0.4, 0.6, 0.8, and 1.0 where $\alpha_1(t^*)$ ~ O'Brien–Fleming; $\alpha_2(t^*)$ ~ Pocock; $\alpha_3(t^*)$ ~ uniform alpha spending functions [211]

Visuals of Alpha Spending + Corresponding Equations

Cook TD, DeMets DL. Introduction to statistical methods for clinical trials. CRC Press; 2007 Nov 19.

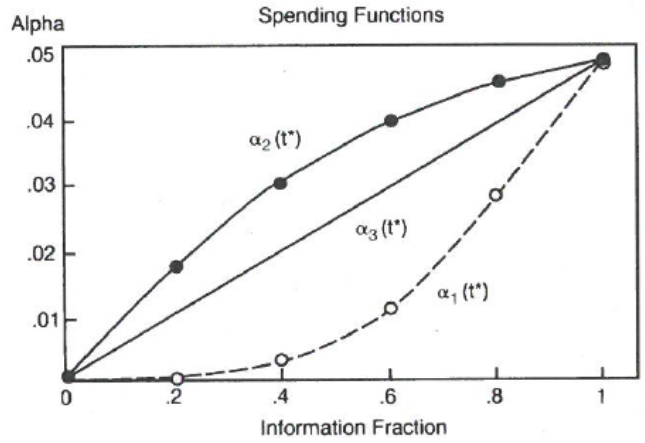


Fig. 16.9 Alpha-spending functions for $K=5$, two-sided $\alpha=0.05$ at information fractions $t^*=0.2$, 0.4, 0.6, 0.8, and 1.0 where $\alpha_1(t^*)$ ~ O'Brien–Fleming; $\alpha_2(t^*)$ ~ Pocock; $\alpha_3(t^*)$ ~ uniform alpha spending functions [211]

Many different spending functions can be specified. The O'Brien–Fleming $\alpha_1(t^*)$ and Pocock $\alpha_2(t^*)$ type spending functions are specified as follows:

$$\alpha_1(t^*) = 2 - 2\Phi\left(Z_{\alpha/2}/\sqrt{t^*}\right) \quad \sim \text{O'Brien-Fleming}$$

$$\alpha_2(t^*) = \alpha \ln(1+(e-1)t^*) \quad \sim \text{Pocock}$$

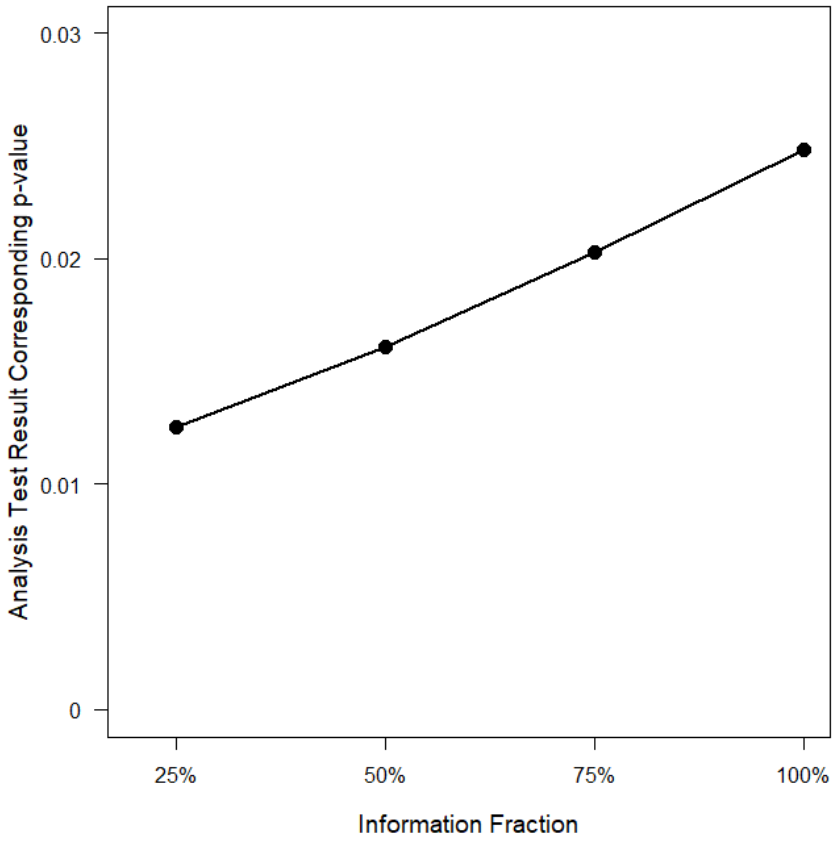
$$\alpha_3(t^*) = \alpha\, t^{*\theta} \quad \text{for } \theta > 0$$

The spending function $\alpha_3(t^*)$ spends alpha uniformly during the trial for $\theta=1$, at a rate somewhat between $\alpha_1(t^*)$ and $\alpha_2(t^*)$. Other spending functions have also been defined [165, 166].

Thrombectomy for Stroke in the Public Health Care System of Brazil

Martins SO, Mont'Alverne F, Rebello LC, et al. Thrombectomy for stroke in the public health care system of Brazil. *New England Journal of Medicine*. 2020;382(24):2316-2326

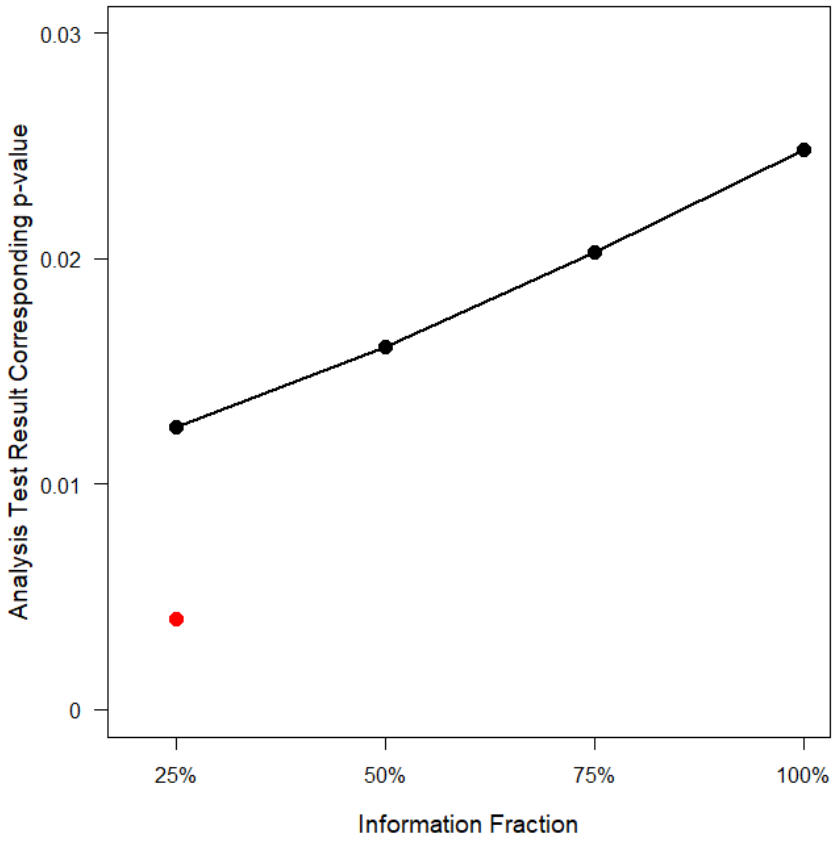
- Randomized stroke patients to standard-of-care (SOC) or SOC + mechanical thrombectomy.
- Target N = 690 participants.
- Proposed interim analysis using information from 90-day follow-up.
- Primary outcome = modified Rankin scale (measure of disability) at 90 days.



Thrombectomy for Stroke in the Public Health Care System of Brazil

Martins SO, Mont'Alverne F, Rebello LC, et al. Thrombectomy for stroke in the public health care system of Brazil. *New England Journal of Medicine*. 2020;382(24):2316-2326

- Results at first interim analysis (N=174):
  - OR = 2.24 (1.30, 3.88) p-value = 0.004, in favor of thrombectomy
- Data and Safety Monitoring Board (DSMB) recommended early stopping.
- At time of early termination, N=221 had been randomized and were included in final analyses
  - OR = 2.28 (1.41, 3.69) p-value = 0.001



Interim Analysis for Efficacy: Implications

- Controlling type I error in any trial is important.
- Things get more complicated if we plan to incorporate interim analyses for efficacy.
- Group sequential methods and alpha spending functions allow researchers a tool to maintain control over type I error.

- In the example trial, investigators were able to address their research question with fewer participants than planned → more efficient use of participant time and study resources(!)
  - This is the heart of the reasoning behind these analyses.
  - Caveat – intervention being studied should be in later stages (e.g., phase III clinical trial) of development; must consider the big picture.

Northwestern Medicine®
Feinberg School of Medicine

Futility

Futility Analyses

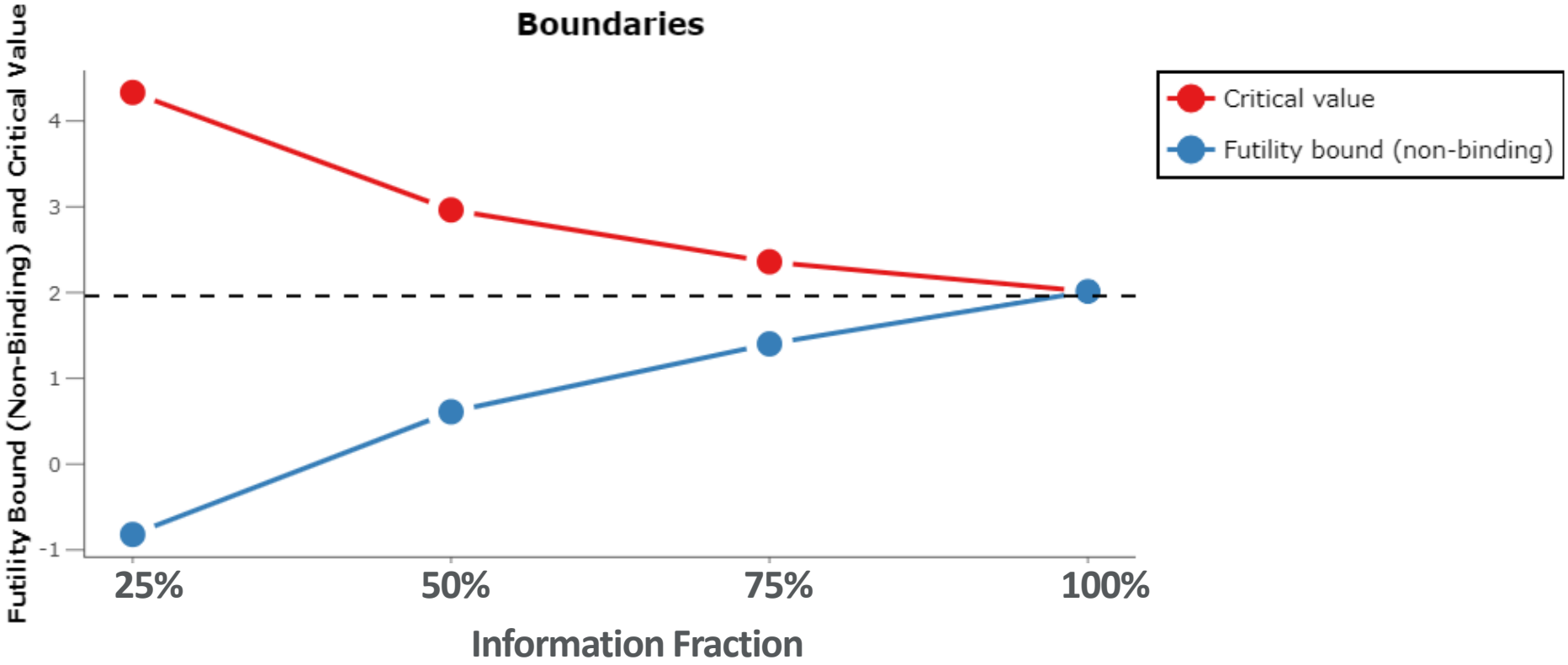
- Stopping a trial for futility suggests that observing a statistically significant result at the end of the study is unlikely.
- This can **increase efficiencies** with respect to cost, resources, and participant burden.
- There are similar methods to the group sequential methods for futility (or error spending functions).

Figure 1 A futile action can never achieve its goals.

Kite S, Wilkinson S. Beyond futility: to what extent is the concept of futility useful in clinical decision-making about CPR? Lancet Oncol. 2002 Oct;3(10):638-42. doi: 10.1016/s1470-2045(02)00878-1. PMID: 12372726.

Beta spending function

One-sided test with efficacy and futility stopping bounds



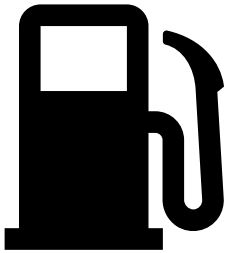
The Stroke and Hyperglycemia Insulin Network Effort (SHINE)

Johnston KC, Bruno A, Pauls Q, et al. Intensive vs Standard Treatment of Hyperglycemia and Functional Outcome in Patients With Acute Ischemic Stroke: The SHINE Randomized Clinical Trial. JAMA. 2019;322(4):326-335. doi:10.1001/jama.2019.9346

- Randomized trial to evaluate **efficacy of intensive glucose control** during ischemic stroke (N=1400.)
- **Non-binding futility thresholds** using error spending function approach.
- **4th interim analysis** (N=1151), trial was **stopped for futility.**
- Final results: no significant difference in proportion with 90-day favorable outcome between groups **(20.5% vs. 21.6%).**

Interim Analysis Sample Size	P-value Futility Threshold
500	0.949
700	0.896
900	0.652
1100	0.293

Conditional Power



- Probably a bit more common are **conditional power or predictive power** approaches…
- In general: power = probability(reject null | some assumption with respect to true underlying effect)

		True	
		Null is True (No Difference)	Null is not True (Difference)
Test	Reject Null	Type I Error	**Power**
	Fail to Reject Null	Confidence	Type II Error

- **Conditional power** = probability(reject null | **data observed up to this point in the trial** and assumption with respect to effect)

Conditional Power

- Based on…
  - Information fraction (how far we are through the trial)
  - Current test statistic – from interim analysis
  - **Assumption of effect at end of trial**
    - Current trend
    - H1 (from beginning of trial)
    - H0 or null effect
  - If this calculated conditional power is low (usually below 20%, 15%, 10%), then we might consider stopping a study

- *Example using Conditional Power in a non-conventional sense later through QUARTET USA trial.*

Conditional Power – How it's calculated

- Let t = trial information fraction (as before)
  - →roughly the fraction of total N that you have at given point
  - Range = 0 to 1

- Z(t) = value of test statistic at information fraction t
- B(t) = "B-value" coined by Lan et al. = $(\sqrt{t})Z(t)$
  - Follows a Brownian Motion process
  - Stochastic literature
  - Random process, with predefined properties → allows us to make inferences based on Brownian Motion properties

Conditional Power – How it's calculated

- $\Theta = E\{Z(1)\} = E\{B(1)\}$ = Expected value of test statistic at end of trial
- Θ also known as *drift parameter*
- Some properties:
  - $E\{B(t)\} = \theta t$
  - $E\{Z(t)\} = \theta\sqrt{t}$
  - For $s \leq t$, $Cov\{B(s),B(t)\} = s$
  - For $s \leq t$, $Cov\{Z(s),Z(t)\} = \sqrt{(s/t)}$

- Note: we can plot B(t) vs. t to determine whether observed trend is better/worse than expected → slope of this line = θ

Conditional Power – How it's calculated

- CP = probability(statistically significant result at end of trial | current trend and assumption about drift parameter)
- $CP = p\{B(1) > Z_{1-\alpha/2} \mid B(t) = b\}$
- $B(1) = Z(1) = B(t) + \{B(1) - B(t)\}$
- $E\{B(1) - B(t)\} = \theta - \theta t = \theta(1-t)$
- $Var\{B(1) - B(t)\} = V\{B(1)\} + V\{B(t)\} - 2Cov\{B(1),B(t)\}$

  $= 1 + t - 2t$

  $= 1-t$
- Let $B(t) = b \rightarrow B(1) = b + \{B(1)-B(t)\}$
- Therefore, $\mathbf{B(1) \mid B(t)=b \sim N(b + \theta(1-t), 1-t)}$

Conditional Power – How it's calculated

- $B(1)|B(t)=b \sim N(b + \theta(1-t), 1-t)$

- $CP = p\{B(1) > Z_{1-\alpha/2} \mid B(t) = b\}$

  $= 1-p\{B(1) \leq Z_{1-\alpha/2} \mid B(t) = b\}$

  $= \mathbf{1-} \boldsymbol{\phi} \left\{ \frac{\boldsymbol{Z_{1-\alpha/2} - [b + \theta(1-t)]}}{\boldsymbol{\sqrt{1-t}}} \right\}$

- Notes for the three typical calculations:
  - Under H0: $\theta = 0$
  - Under H1: $\theta = Z_{1-\alpha/2} + Z_{1-\beta}$
  - Under current trend, use $\hat{\theta} = b/t$

Predictive Power

- Bayesian alternative to conditional power
- Likelihood of demonstrating treatment efficacy at the end of the study
- Estimated by…
  - Updating prior assumptions with observed data
  - Average conditional power over this distribution
- Avoids having to assume a specific treatment effect (as in the conditional power approach)

Futility Analysis: Implications

- Incorporating a futility assessment can **increase efficiency of the trial**, allowing trials that are unlikely to meet their objectives to stop early ultimately reducing costs, preserving resources, and limiting patient burden.

- Particularly in large clinical trials or vulnerable patient populations, an interim futility assessment may be essential to **prevent patients from being unnecessarily randomized to ineffective treatments.**

- Despite common misconceptions, an interim analysis incorporating a futility assessment **alone does not inflate the type I error.**

- It can, however, have **implications on type II error**, reducing the overall power of the study by stopping early.

Stopping a study early

For efficacy or futility

- Potential to pose challenges in interpretations – smaller than expected sample size decreases precision around treatment effect estimates.

- Subgroup analyses and analyses examining heterogeneity of intervention effects will inevitably suffer (they are typically underpowered already).

Northwestern Medicine®
Feinberg School of Medicine

Safety

jody.ciolino@northwestern.edu

Interim analysis for safety

- Note: all clinical studies (regardless of phase) should incorporate safety monitoring.

- It is impossible to foresee all potential safety issues ahead of time, and **safety analyses are seldom adequately powered.**

- Unanticipated safety issues should merit consideration for early stopping; for example,
  - **Serious** and **unanticipated** events or event rates
  - Events that are **related and unexpected.**

Interim analysis for safety

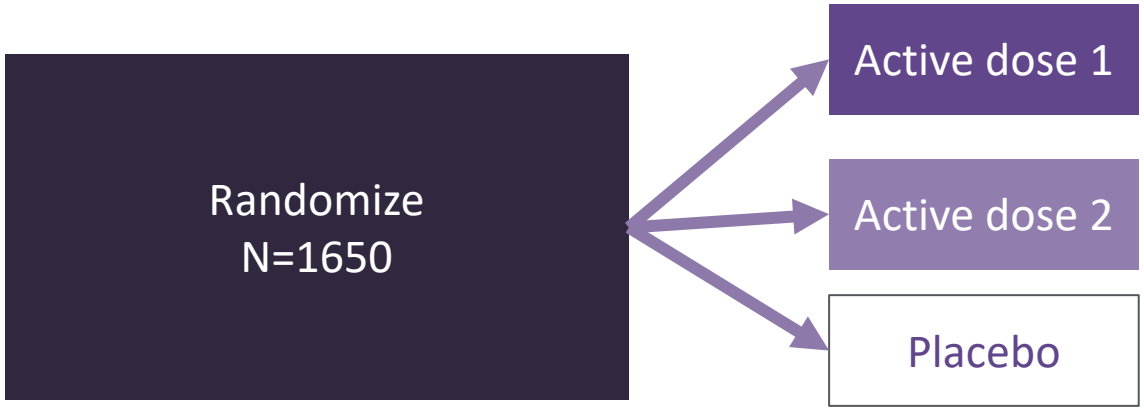
- Discussion on formal interim safety analyses should consider the **benefit-to-risk assessment**: i.e., it **should be paired with an interim efficacy analysis.**

- **Context is key** in this assessment…. *For example, a cardiovascular secondary prevention trial to prevent subsequent myocardial infarction may have the expectation of some myocardial infarction events, whereas it may be extremely concerning to observe the same events in a behavioral intervention in a generally healthy population.*

The EARLY Trial – will skip for presentation

Sperling R, Henley D, Aisen PS, et al. Findings of efficacy, safety, and biomarker outcomes of atabecestat in preclinical Alzheimer disease: a truncated randomized phase 2b/3 clinical trial. JAMA neurology. 2021;78(3):293-301

- Randomized phase 2b/3 trial assessing effects of atabecestat in preclinical Alzheimer's disease
  - 3 arms (2 doses of active intervention + placebo)
  - Double-blind
  - Primary outcome: change from baseline in cognitive composite score
- **Plan:**
  - N=1650 across 143 sites
  - Interim analysis for futility
  - NO formal interim efficacy analyses



The EARLY Trial - will skip for presentation

Safety and tolerability monitored throughout study

- Treatment-emergent adverse events
- Vital signs
- Serial MRIs
- Clinical laboratory results

0%

33%

N=557

100%

N=1650

- Safety concerns
- Serious elevation in liver enzymes
- Unexpected
- Related to dosage

Benefit

Risk

The study terminated early due to unacceptable safety profile

Interim analysis for safety: Implications

will skip for presentation

- Monitoring safety outcomes and adverse events is crucial for **maintaining study integrity and protecting study participants.**

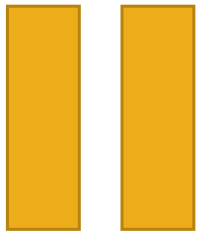
- Decision to stop a trial early for safety concerns should generally be made in the context of the **benefit-to-risk ratio**
  - No "one size fits all" approach to assessing benefit-to-risk ratio

Remember: **Context is key** in this assessment…. *For example, a cardiovascular secondary prevention trial to prevent subsequent myocardial infarction may have the expectation of some myocardial infarction events, whereas it may be extremely concerning to observe the same events in a behavioral intervention in a generally healthy population.*

Interim analysis for safety: Implications

- Safety monitoring may raise concerns about **multiple "looks"** at the data
  - When coupled with efficacy analysis, incorporate conservative alpha-spending approach

- Safety concerns may warrant a **temporary pause** to further assess causality – consider whether to pause vs. completely terminate the study.



VS.

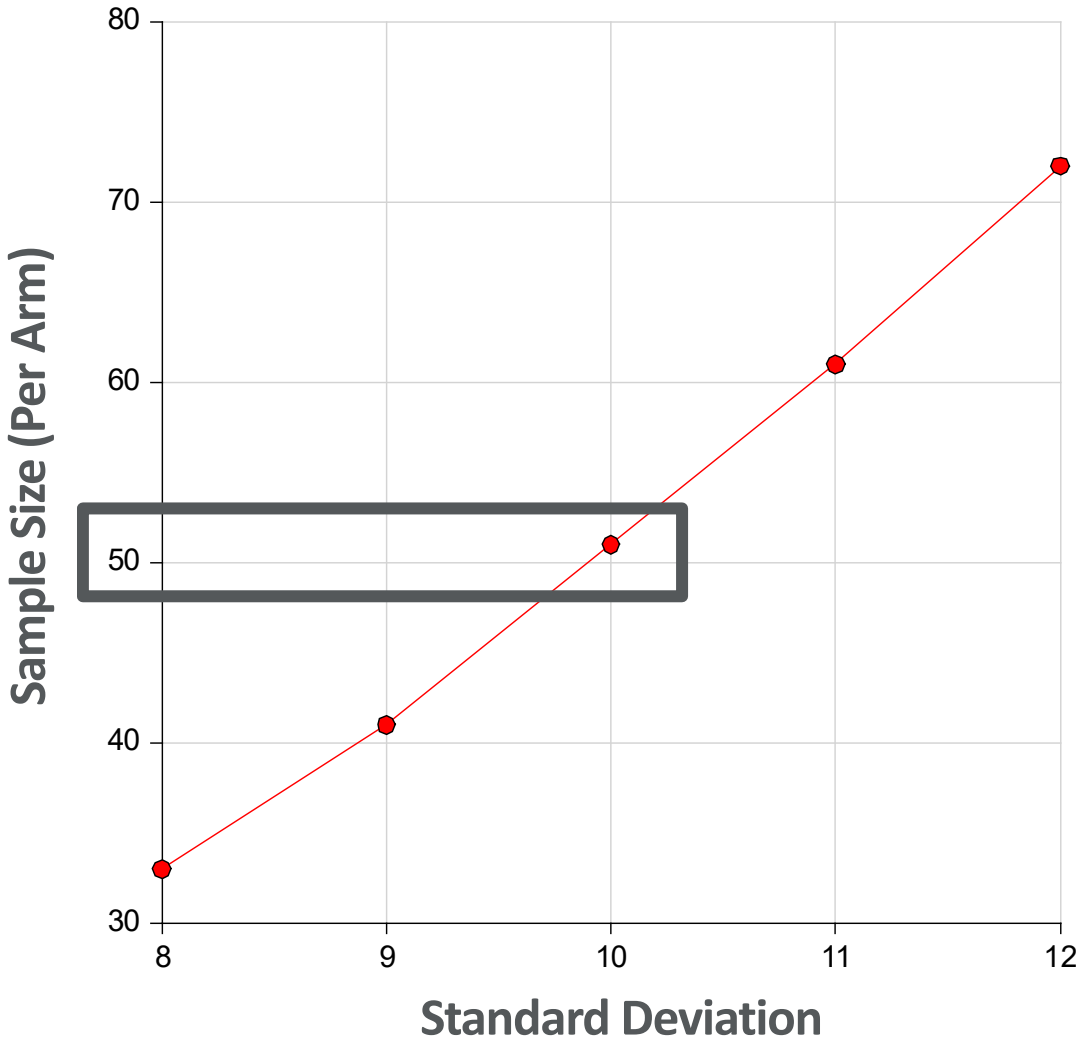
Northwestern Medicine®
Feinberg School of Medicine

Sample Size Re-estimation

jody.ciolino@northwestern.edu

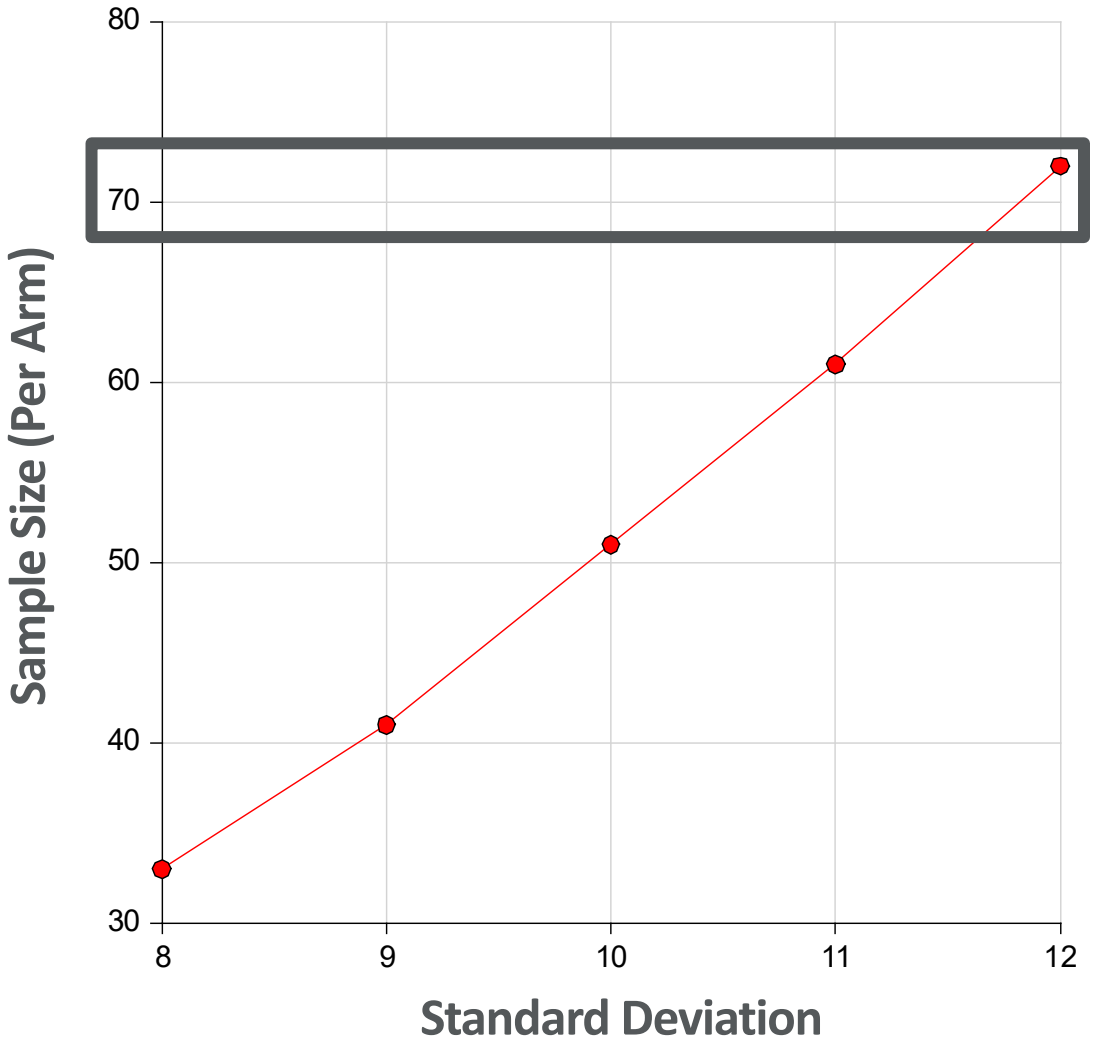
Sample size re-estimation

- Designed to modify the planned sample size based on the accumulating data.
- Accounts for **uncertainty when conducting power calculations** during the initial planning of the study.



Sample size re-estimation

- Designed to modify the planned sample size based on the accumulating data.
- Accounts for **uncertainty when conducting power calculations** during the initial planning of the study.

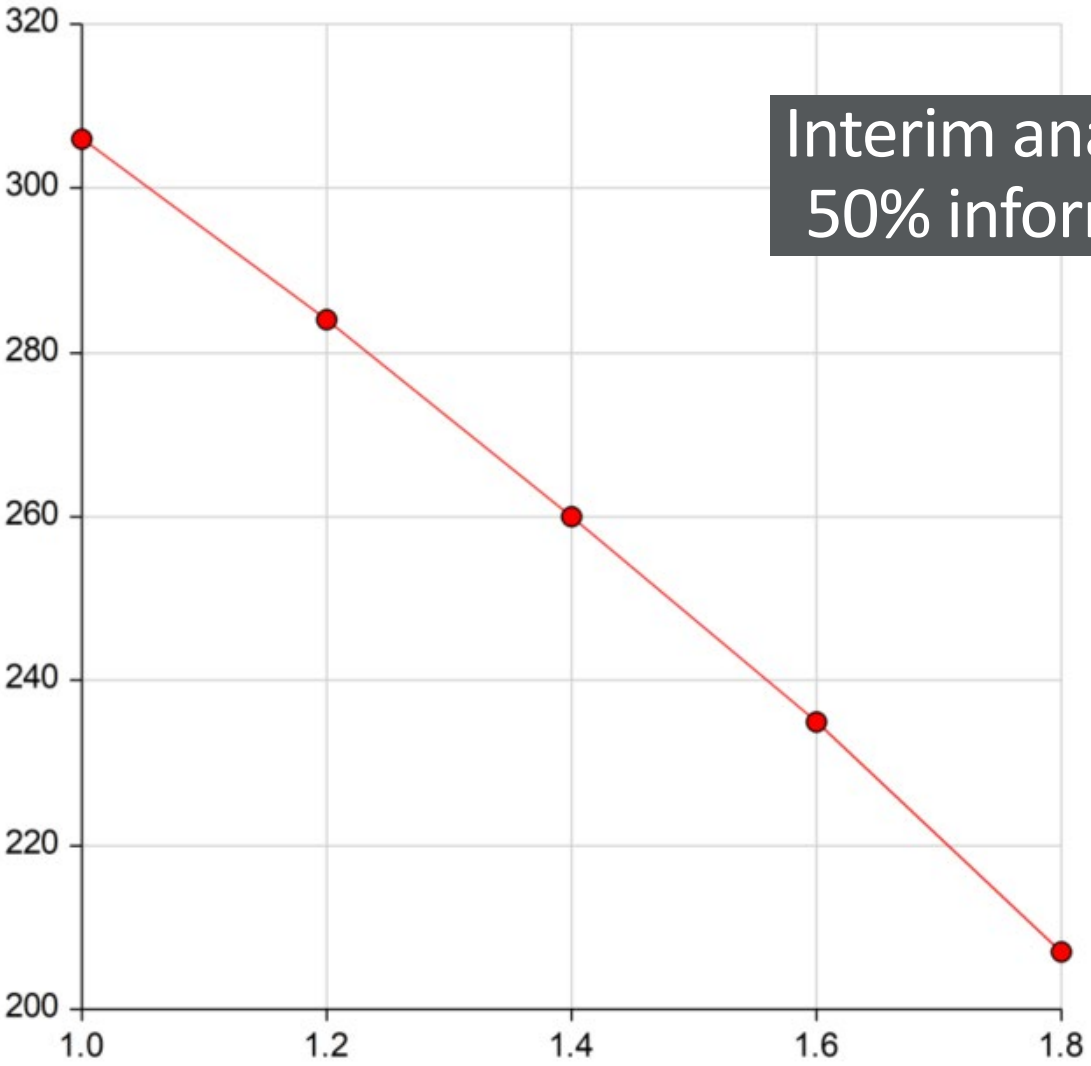


Sample size re-estimation

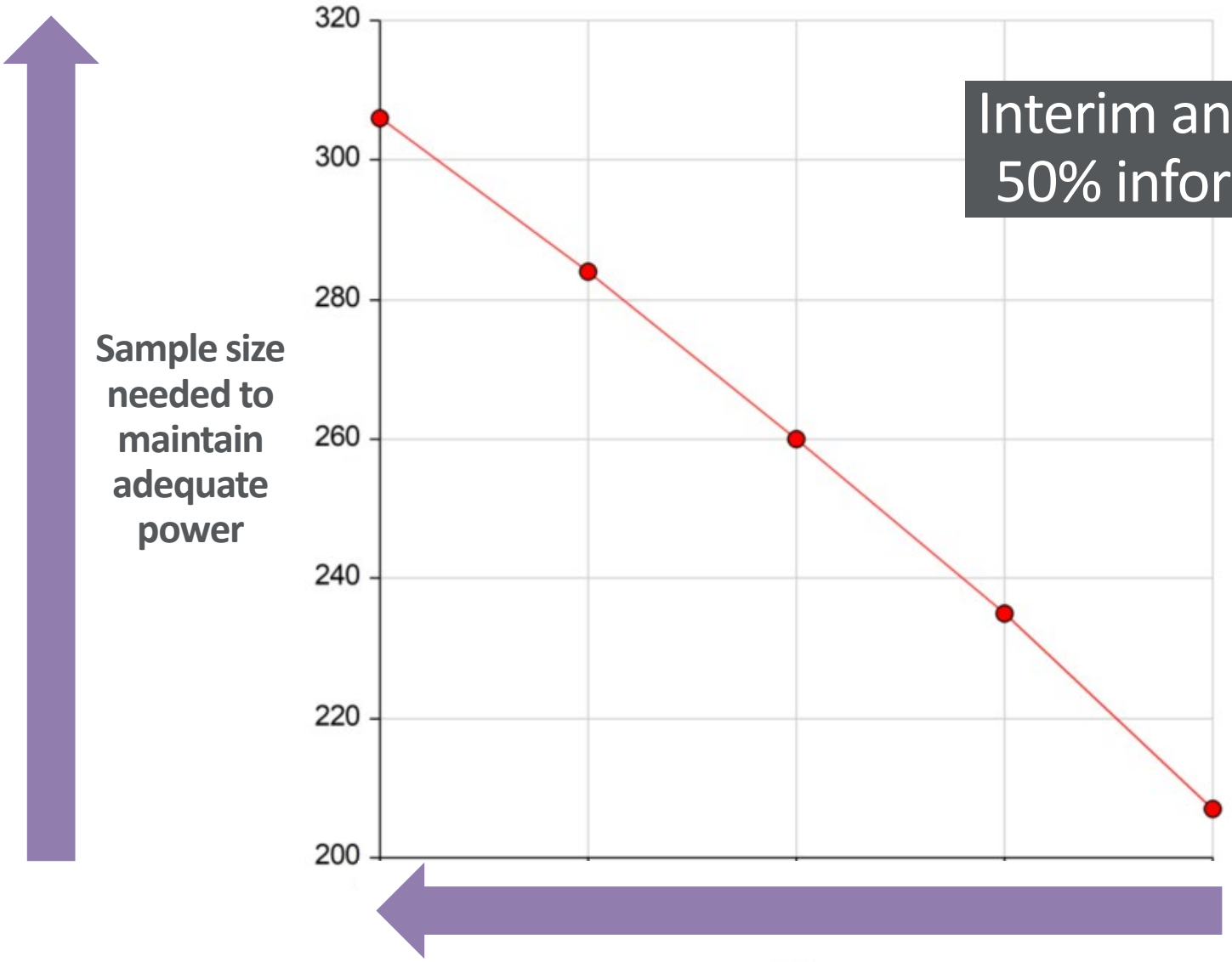
- Approaches facilitate a revised sample size calculation using information for the **ongoing assessment of event rates, the estimation of nuisance parameters** (e.g., the variance of a continuous outcome), or the effect size expected.

- Re-estimating a sample size at an interim stage can increase the likelihood of a successful trial, but **may result in a substantial increase** in the needed sample size if the initial sample size **assumptions** were very **different** from what is observed.

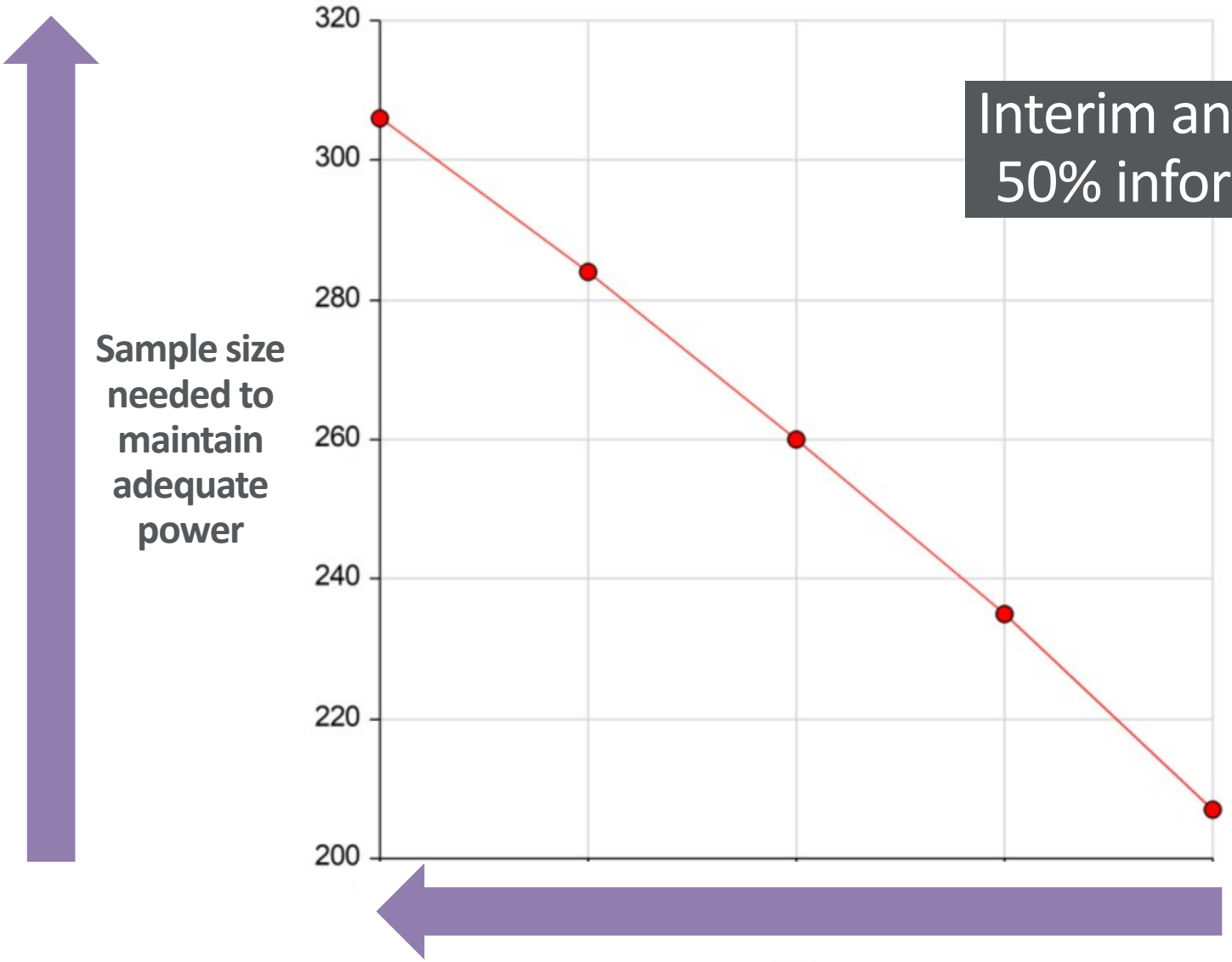
Sample size re-estimation (unblinded; incorporating effect)



Sample size re-estimation



Sample size re-estimation



Sample size re-estimation

- Two categories: blinded or unblinded
- **Blinded:** used to revise estimation of nuisance parameters (e.g., variance) – often uses pooled estimates
- **Unblinded**: based on comparative interim results; ideal when uncertainty about estimates of effect size and nuisance parameters – allows for capturing an effect that may still be clinically meaningful, but differs from original assumption.

Example: Sample Size Re-estimation Plan

The Liver Cirrhosis Network (LCN) "RESCU" Trial

- **R**osuvastatin **E**fficacy and **S**afety for **C**irrhosis in the **U**nited States: A Double-Blind Randomized Placebo-Controlled Phase 2 Study

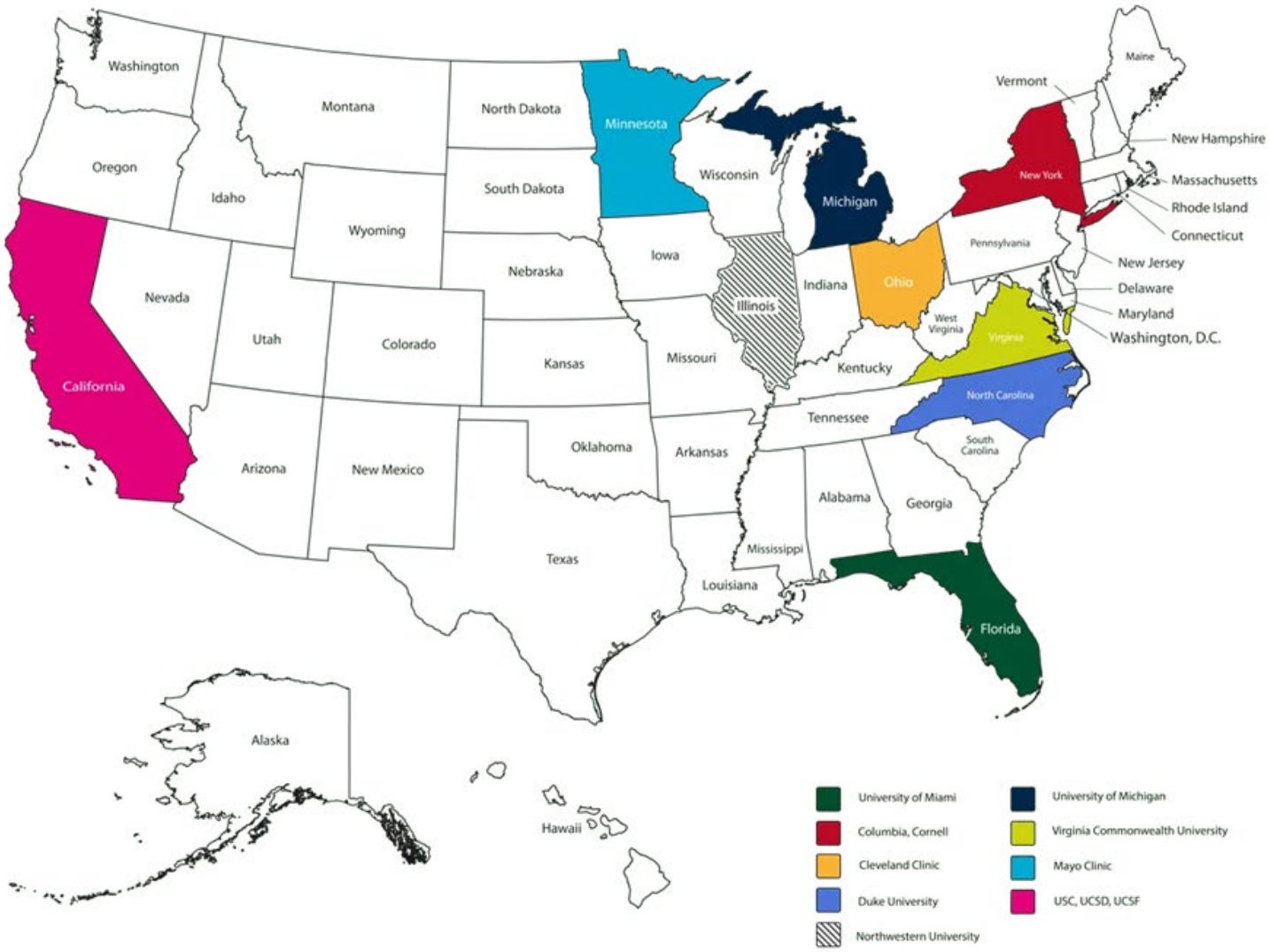
- clinicaltrials.gov ID: NCT05832229

- **Primary Objective**: To evaluate the **safety and efficacy** of **rosuvastatin** 10-20mg compared to **placebo in patients with compensated cirrhosis** in modifying disease progression as measured by **liver stiffness**.

- **Primary Outcome**: Mean liver stiffness (kPa) as measured by vibration controlled transient elastography (VCTE) at week 96.

Background – about the LCN

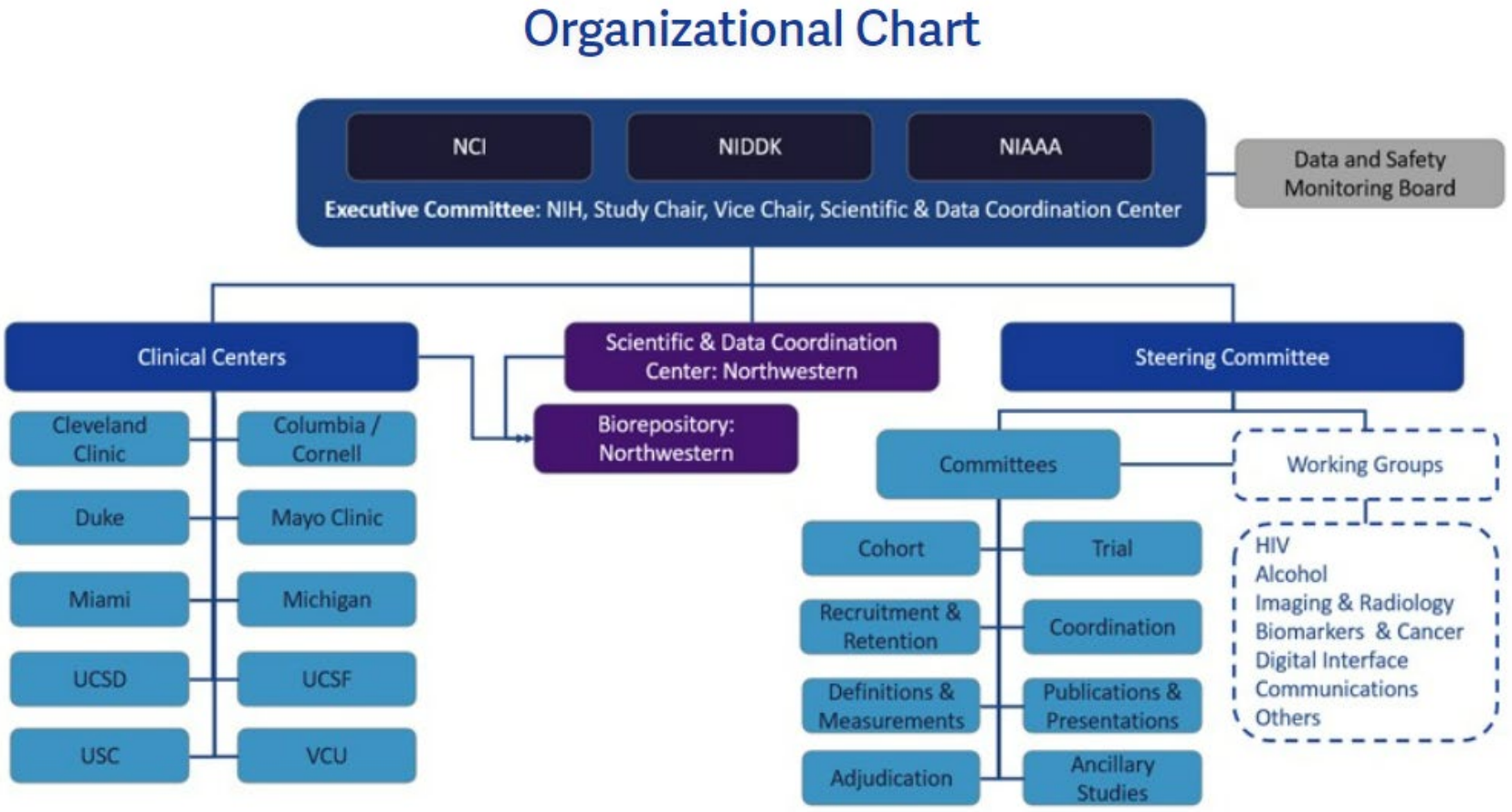
- The Liver Cirrhosis Network (LCN) was **formed in Fall of 2021.**
- Goals: of designing and implementing **two studies** across **10 clinical centers**:
  - a prospective, longitudinal cohort study of patients with cirrhosis to study disease progression, and
  - a **double-blind, randomized, placebo-controlled trial** to evaluate **statin** therapy in patients with cirrhosis for both safety and efficacy.



10 clinical centers + a Scientific and Data Coordinating Center (SDCC; Northwestern) + Biorepository (Northwestern)

Structure: 10 U01 awardees + 1 U24 awardee (SDCC; NU)

3 institutes from NIH involved



The LCN RESCU Study

1 of 2 major studies as part of the LCN; nearly ready to launch

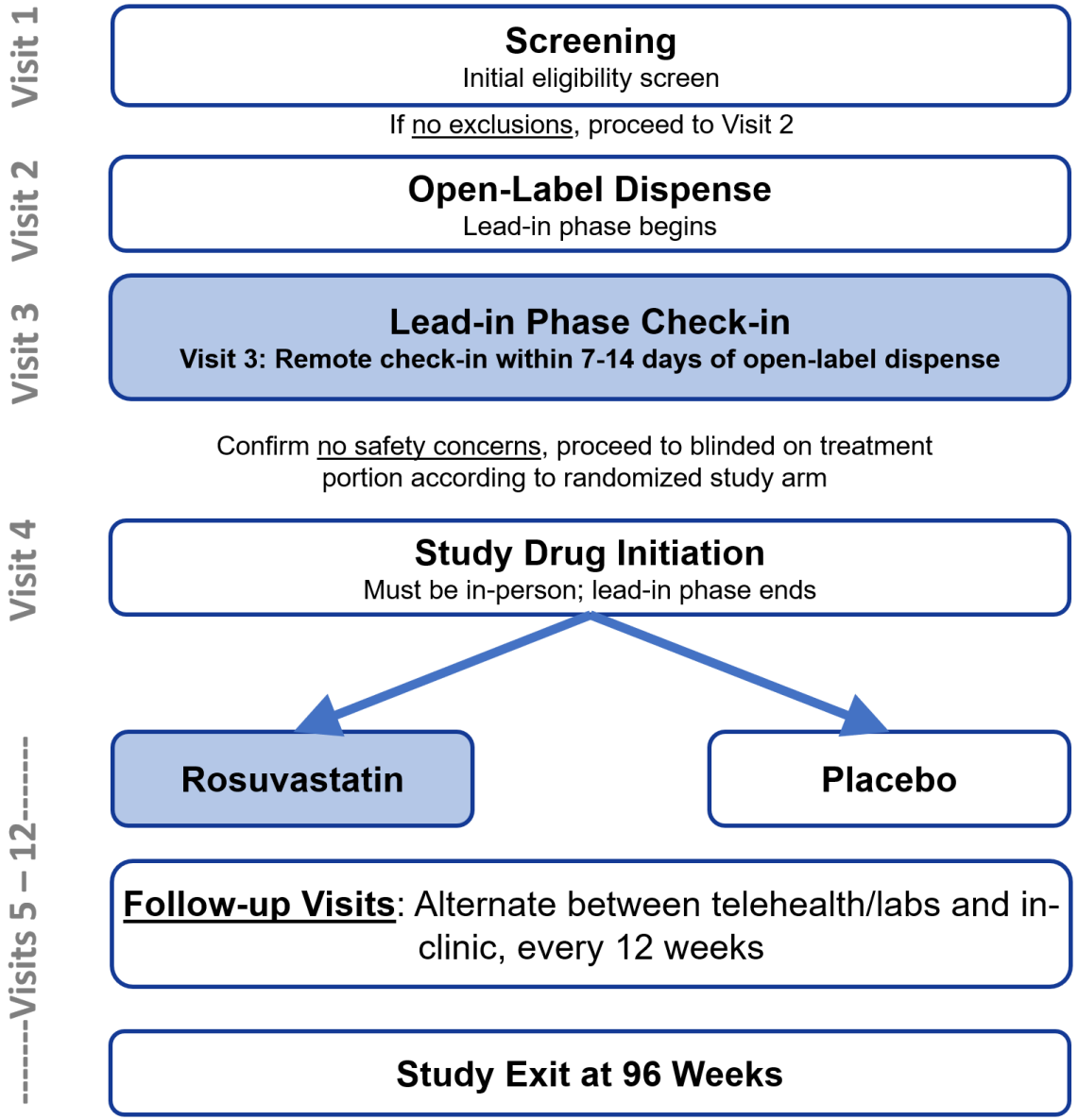
- The first several years of the LCN: developing infrastructure, policies, procedures, agreeing on common protocol(s), launching the first of the 2 studies.

- A LOT of debate, time energy, resources are going into the LCN RESCU study
  - Original plan was an adaptive phase IIb/III study
  - Original considerations to expand / extend follow-up to increase power and ability to detect meaningful differences across study arms in key clinical endpoint(s): **decompensation**

  - **Problem:** median time-to-decompensation is estimated at about 10-12 years among patients with cirrhosis → sample size estimates are infeasible with the anticipated yearly event rate on this outcome.
  - **Solution**: use a different outcome as a surrogate of sorts for decompensation: VCTE.

The RESCU Trial

clinicaltrials.gov ID: NCT05832229



RESCU – Sample Size Considerations

Assuming:

- *11 kPa standard deviation*
- 2-sided 5% type I error rate
- We would like to be powered to detect a mean 5 kPa difference across arms
- 80% power

→ We would need an analytic sample size of 162 study participants (81 participants per arm

→ Assuming 20% loss to follow-up prior to providing data at the specified primary time point of interest, we plan to enroll a total of **204 study participants (102 per study arm)** into the study drug initiation phase of the study, after lead-in.

→ To account for dropout during the lead-in period, we plan to **consent (into lead-in) 256** study participants.

Sample Size Re-Estimation for RESCU

SO. MANY. ASSUMPTIONS.

- The SDCC suggested a blinded sample size re-estimation based on the nuisance parameter.
- The Data and Safety Monitoring Board (DSMB) agreed, and they suggested blinded sample size re-estimation after about 1/3 of participants have been followed for 1 year.
- Current interim analysis plan (under review at the moment)….

Text taken from the Statistical Analysis Plan (SAP)

- After one-third of target number of participants for analyses (N = 162 ÷ 3 = 54 participants) have been followed through Week 48, we plan for a blinded interim sample size re-estimation.

- The primary justification: initial uncertainty around the parameter assumptions in sample size calculations.

- Assumed standard deviation of 11kPa, and the sample size calculations do not account for the multiple observations per study participant (not enough published data on behavior of the liver stiffness measure over time and its serial correlations).

- Sample size re-estimation will allow for improved estimates on variance of stiffness within the population and further allow for initial estimates of correlation structure of this variable over time within a participant.

Text taken from the SAP

Specifically, the estimate of the variance will be based on the pooled sample computed without unblinding as described in Gould and Shih [9].

$$\hat{\sigma}^2 \approx \frac{n-1}{n-2}\left(s^2 - \Delta^2/4\right)$$

Where $\hat{\sigma}^2$ is the unknown within-group estimate of variance, Δ = hypothesized mean difference under the alternative and s^2 is the pooled sample variance at the Week 48 time point. This updated estimate for σ^2 will be used in an updated sample size calculation under the same original assumptions: 80% power, two-sided 5% level of significance, equal allocation, effect size of 5kPa, in several different sample size formula:

1. Independent two-sample t-test, as in the original sample size calculation.
2. Incorporating an estimate of correlation between baseline mean stiffness (across Visits 1 and 2) and Week 48 stiffness (i.e., an ANCOVA approach). In this instance we would use the accumulated data to estimate this correlation.

Recognizing that the primary time point of interest is the Week 96 stiffness rather than the Week 48 interest, the estimates for variance and correlation used in the interim analysis planned will be subject to bias and may exhibit more or less variability compared to the true parameter values at Week 96. For this reason, we plan to provide a range of sample size re-estimations under varying assumptions to inform plans to update the study enrollment targets.

Sample size re-estimation: Implications

- Goal to **prevent underpowered** studies

- Be careful as this **may have impact on type I error** without using appropriate methods to control.
- What if re-estimated sample size is **not feasible**?
- What if re-estimated sample size is based on an observed effect that is **no longer clinically meaningful**?
- Care should be taken in **how results from interim re-estimation are reported** for ongoing studies -> may be possible to back-calculate the effect size!

Northwestern Medicine®
Feinberg School of Medicine

Tying it all together

jody.ciolino@northwestern.edu

Table 1. Summary of Interim Analysis Types

	Explanation	Justification for Use
Efficacy	• Early termination of a trial that is **showing promising results** • **Control of type I error** through group sequential methods or alpha spending functions	• Usually for longer, larger studies and later phases of research • **Ethical imperative for a promising treatment to reach the entire target clinical population**
Futility	• Early termination of a trial that **is not likely to achieve the intended objective** (e.g., little chance of finding a "significant" treatment effect at the end of the study) • Employed through group sequential methods, error spending functions, conditional power, or predictive power	• **Reduces costs, resources, and patient burden** for a trial with a low probability of "success" • Usually for mid-late phase studies • Helpful in the context of recruitment and retention challenges
Safety	• **Early termination** (or pausing) of a trial **for safety concerns** • Should be coupled with efficacy analyses to evaluate the benefit-to-risk ratio	• Incorporated across all phases of research • **Particularly important for vulnerable populations and high-risk interventions** with more "serious" outcomes (e.g., death)
Sample size re-estimation	• Reassessment of the sample size required to **ensure adequate power** using updated information from interim trial data • Can be blinded or unblinded • May not necessarily spend alpha	• Allows for **interim look at assumptions** (standard deviations, event rates, correlations, etc.) • May be particularly useful for mid-late phase studies

Considerations for Interim Analyses

- **Pre-specify** as much as possible
- Describe anticipated timing, proposed methodology, pre-specified rules to **guide** decisions
- Timing of analysis can be flexible
- Often specified when some proportion of participants is enrolled and meet a particular study milestone (e.g., 50% of participants completed 6-week follow-up)
- Balance between **maximal information** (later interim analysis) vs. ensuring **adequate time to make any modifications** and reducing potential risk to participants as much as possible

Considerations for interim analysis

- Think about potential logistical implications…if an interim sample size re-estimation is proposed, are there adequate resources to support an increase in sample size if indicated?
- Evaluation of interim analysis results **should not be interpreted in isolation**, but rather in the context of other internal study factors and external contemporaneous issues.
- Any interim analysis results and statistical tools are intended to serve as **guidelines.**
- **Transparency** in disseminating trial results when interim analyses were conducted is also critical.

Tying it all together

- **Interim monitoring** of data quality and integrity ≠ **interim analysis** to guide study design modifications.
- Both **require coordination and pre-specification** of protocol and procedural elements to the extent possible.
- Monitoring does not have implications for type I error issues; interim analyses can, but it is not always the case.
- Remember that there is no "one size fits all" + recommendations should be made with a big picture view.

Northwestern Medicine®
Feinberg School of Medicine

If time allows…

jody.ciolino@northwestern.edu

QUARTET USA

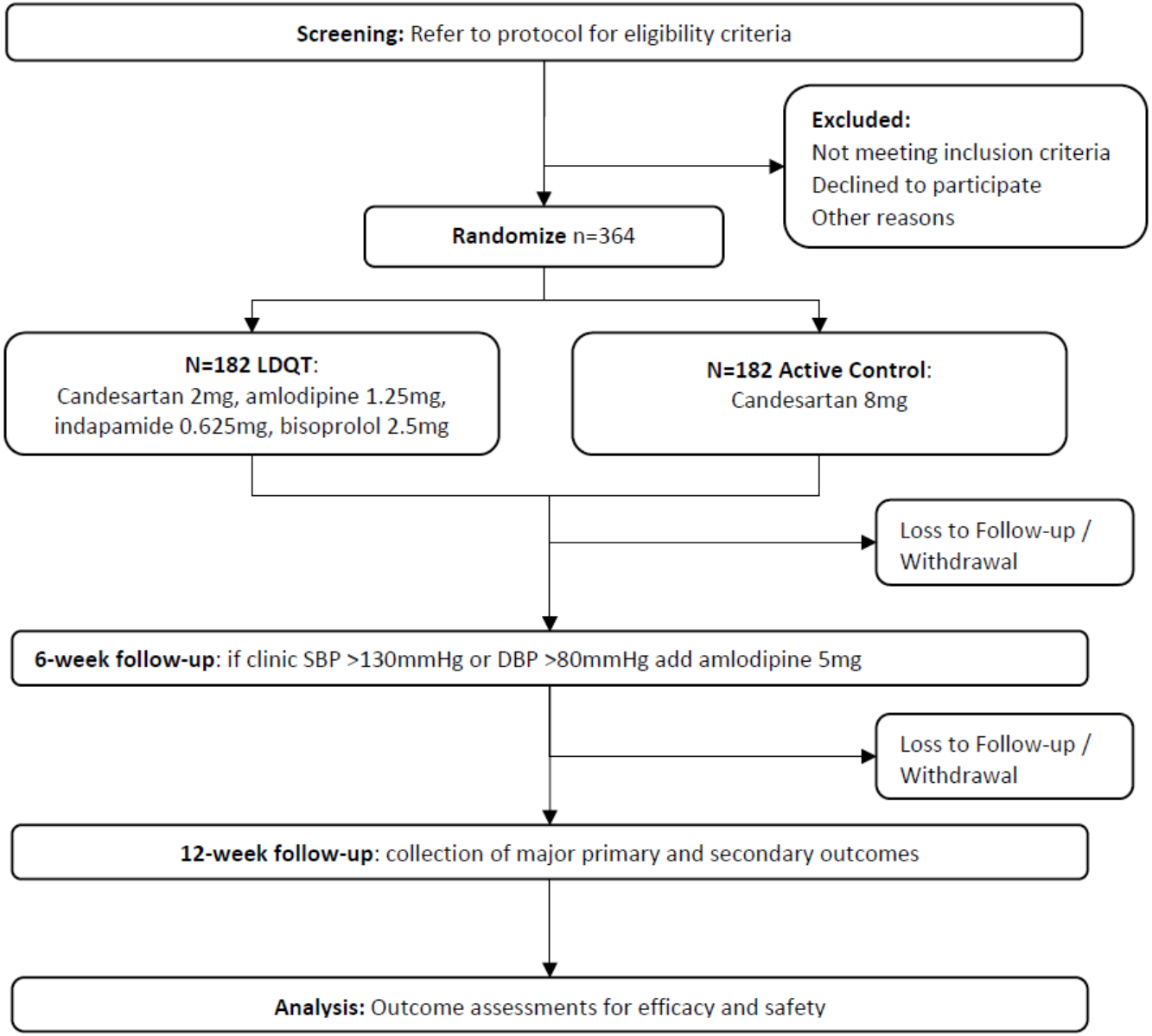
QUARTET USA (inspired by QUARTET – original study in Australia)

- Efficacy and Safety of a Quadruple Ultra-low-dose Treatment for Hypertension (QUARTET USA)

- Clinicaltrials.gov: NCT03640312

- Goals: to investigate, in a double-blind randomized controlled trial, whether initiating treatment with ultra-low-dose quadruple-combination therapy ("LDQT") will lower office blood pressure more effectively, and with fewer side effects, compared to initiating standard dose monotherapy in patients with hypertension.

Figure 2. Study Schematic



Study Participants

Patients within Access Community Health Network

ACCESS COMMUNITY HEALTH NETWORK
TRANSFORMING COMMUNITY HEALTH CARE

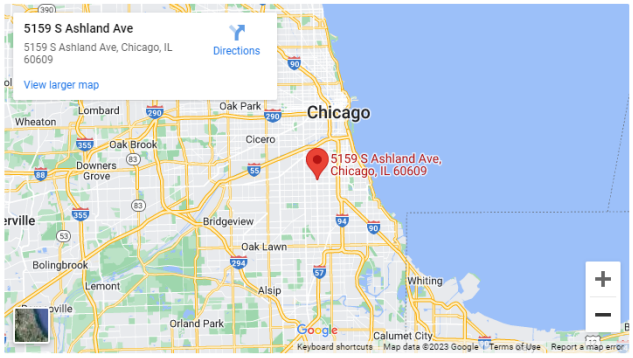
- https://www.achn.net/

ACCESS Ashland Family Health Center

5159 S. Ashland Ave.
Chicago, IL 60609

Phone 773.434.9216
Fax 773.434.2670

Hours
Monday: 8:00 a.m. - 5:00 p.m.
Tuesday: 8:00 a.m. - 5:00 p.m.
Wednesday: 8:00 a.m. - 5:00 p.m.
Thursday: 8:00 a.m. - 5:00 p.m.
Friday: 8:00 a.m. - 5:00 p.m.



Doctors and other Medical Providers

Psychiatry
- Motaz Alshami, M.D.
- Jorge Luis Castillo-Gonzalez, M.D.

Family Medicine
- Victoria Viveen, P.A.-C.
- Katie McKeough, P.A.-C.

Obstetrics and Gynecology

Pediatrics
- Brett Ballard, M.D.
- Daneen Woodard, M.D.

Internal Medicine
- Eleanor Teoh, D.O.
- Daneen Woodard, M.D.

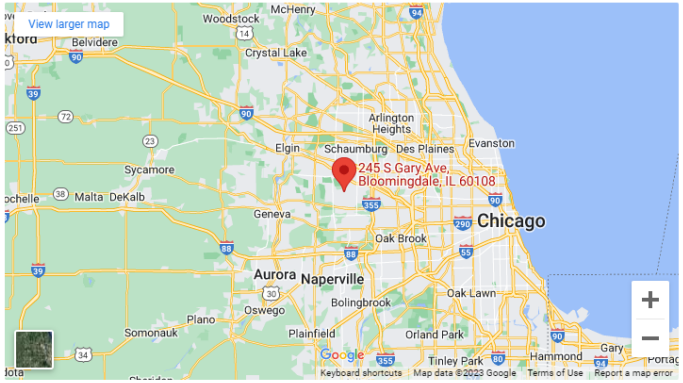
Cardiology

ACCESS Martin T. Russo Family Health Center

245 South Gary Avenue,
Lower Level
Bloomingdale, IL 60108

Phone 630.893.5230
Fax 630.893.5837

Hours
Monday: 8:00 a.m. - 8:00 p.m.
Tuesday: 8:00 a.m. - 7:00 p.m.
Wednesday: 8:00 a.m. - 8:00 p.m.
Thursday: 8:00 a.m. - 8:00 p.m.
Friday: 8:00 a.m. - 5:00



Doctors and other Medical Providers

Family Medicine
- Charity Alikpala, D.O.
- Preyanshu Parekh, D.O.
- Steven Miguel Chapa, P.A.-C.
- Marion R. Tan, F.N.P.-B.C.
- Mary Winokur, P.A.-C.
- Nicole Locascio, P.A.-C.
- Joy De Leoz, M.D.

Obstetrics and Gynecology
- Jessica Ocampo, M.D.
- Dale Liaugminas, M.D.
- James Kim, M.D.
- Josephine Rios, C.N.M.

Internal Medicine
- Jairo Mejia, M.D.

Outcomes

Primary Outcome

Automated office systolic blood pressure (SBP) at 12 weeks, and analyses will compare this change across arms for primary outcome analyses, adjusting for baseline.

Secondary Outcomes

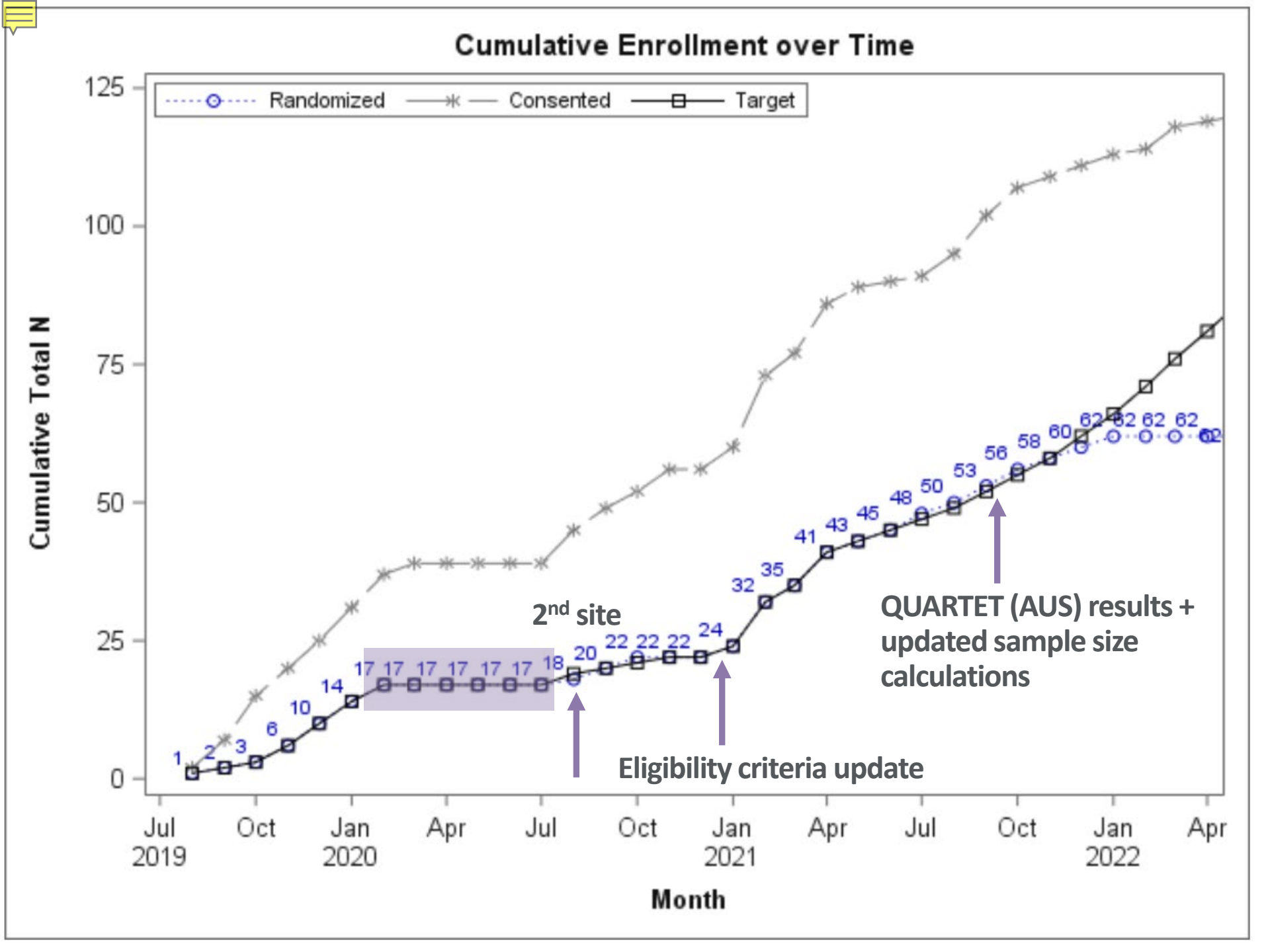
a. Automated office diastolic blood pressure (DBP) at six and 12 weeks.
b. Proportion of patients with hypertension control (SBP < 130 mmHg and DBP <80 mmHg) at six and 12 weeks.
c. Proportion of patients requiring step-up treatment.
d. Proportion of patients with adverse event-free hypertension control (SBP < 130 mmHg and DBP < 80 mmHg).
e. Medication Adherence (pill counts, participant-report).
f. Health-related Quality of Life: PROMIS Global Health.

Sample Size Calculations – Original

From SAP version 1.0

A total of **365 participants will be randomized** (1:1 allocation). We anticipate an analytic sample size of **292 based on 365 participants at randomization and a 20% dropout rate by the 12-week follow-up time point.** We base sample size and power calculations conservatively on an independent two-sample t-test. The analysis methods, ANCOVA, will increase precision on intervention effect when controlling for relevant baseline covariates, thereby providing additional power of detecting intervention effect.

For primary outcome analyses, an independent two-sample t-test **provides 80% power to detect a 5 mmHg difference in SBP between the intervention and comparator arms assuming a two-sided 5% level of significance and a 15 mmHg standard deviation in outcome.** This estimate is based on a 2017 Cochrane systematic review update evaluating the effects of fixed-dose combination therapy and systematic review on quarter dose combination therapy, and a pilot trial of quarter-dose combination therapy [3]. We assume baseline SBP has a moderate correlation with follow-up SBP (r≈0.50-0.6); under this assumption, sample size calculations based on ANCOVA has the potential to allow for over 90% power under the same assumptions for remaining parameters.


Cumulative Enrollment over Time
Randomized
Consented
Target
Cumulative Total N
125
100
75
50
25
0
Jul 2019
Oct
Jan 2020
Apr
Jul
Oct
Jan 2021
Apr
Jul
Oct
Jan 2022
Apr
Month
1 2 3 6 10 14 17 17 17 17 17 17 18 20 22 22 22 24 32 35 41 43 45 48 50 53 56 58 60 62 62 62 62 62
2nd site
Eligibility criteria update
QUARTET (AUS) results + updated sample size calculations

Updates and challenges

- (understandably) Pressures from funder re: enrollment numbers: continued difficulty in justification for funds for study that is underperforming on recruitment
  - Suggested interim analysis
  - Worked with DSMB to develop an interim analysis plan
  - Updated sample size requirements

- **All of these things carried logistical, operational challenges + required protocol amendments (and all the things that go along with those)**

Interim Analysis

Enrollment was slower than we would have hoped

- We were in continued discussion with the funder and the DSMB
  - Suggested interim analysis
  - Worked with DSMB to develop an interim analysis plan
  - Updated sample size requirements
- We conducted an unplanned **interim conditional power analysis**

- At the same time, we **updated the analytic plan** to a Linear Mixed Model (LMM), using the **six-week time point in analyses as well**
  - Originally, it was going to be a simpler ANCOVA model – looking at Week 12, controlling for baseline
  - QUARTET (AUS) results also came out at this time – their analytic strategy used an LMM
  - Thus, to make most use of our data AND to align with QUARTET (AUS), we performed an interim analysis using this new analytic strategy

Interim Analysis

In collaboration with the DSMB chair and DSMB statistician

Solution for Fixed Effects: SBP = SBP_baseline + Arm					
Effect	trt	Estimate	Standard Error	DF	t Value
Intercept		63.9718	23.5595	40	2.72
Baseline SBP		0.4771	0.1692	32	2.82
Arm	1	-7.2374	3.4698	32	-2.09
Arm	2	--Reference--			

In the calculation for conditional power (CP), we take this test statistic value and evaluate what power may be if:

1. From here to the end of the study, there is **no trend** (Null assumption).
2. From here to the end of the study, this **current trend** continues (Current trend assumption).
3. From here to the end of the study, the originally hypothesized trend (80% power, 5mmHg difference, etc.) were to hold (**alternative trend** assumption).

We allow the assumed **information fraction** to vary according to different projected recruitment numbers.

Interim Analysis

Recall – our initial sample size called for analytic sample size = 292 (planned to recruit N=364 total!)

Total N	Scenario	CP Estimate
71	Null	0.29
71	**Alternative**	**0.89**
71	**Current Trend**	**0.88**
77	Null	0.27
77	**Alternative**	**0.90**
77	**Current Trend**	**0.90**
84	Null	0.25
84	**Alternative**	**0.90**
84	**Current Trend**	**0.92**
92	Null	0.23
92	Alternative	0.91
92	Current Trend	0.94

Interim Analysis

Our NEW sample size justification

The initial sample size calculations called for a total of 365 participants to be randomized (1:1 allocation). We anticipated an analytic sample size of 292 based on 365 participants at randomization and a 20% dropout rate by the 12-week follow-up time point. We originally based sample size and power calculations conservatively on an independent two-sample t-test.

"Based on results of interim analyses…we updated our recruitment target to 87 participants (1:1 allocation). The **analytic sample size of 77 is anticipated based on 87 participants at randomization and a conservatively estimated 12% dropout rate by the 12-week follow-up time point based on 8% dropout rate observed through September 2021."**

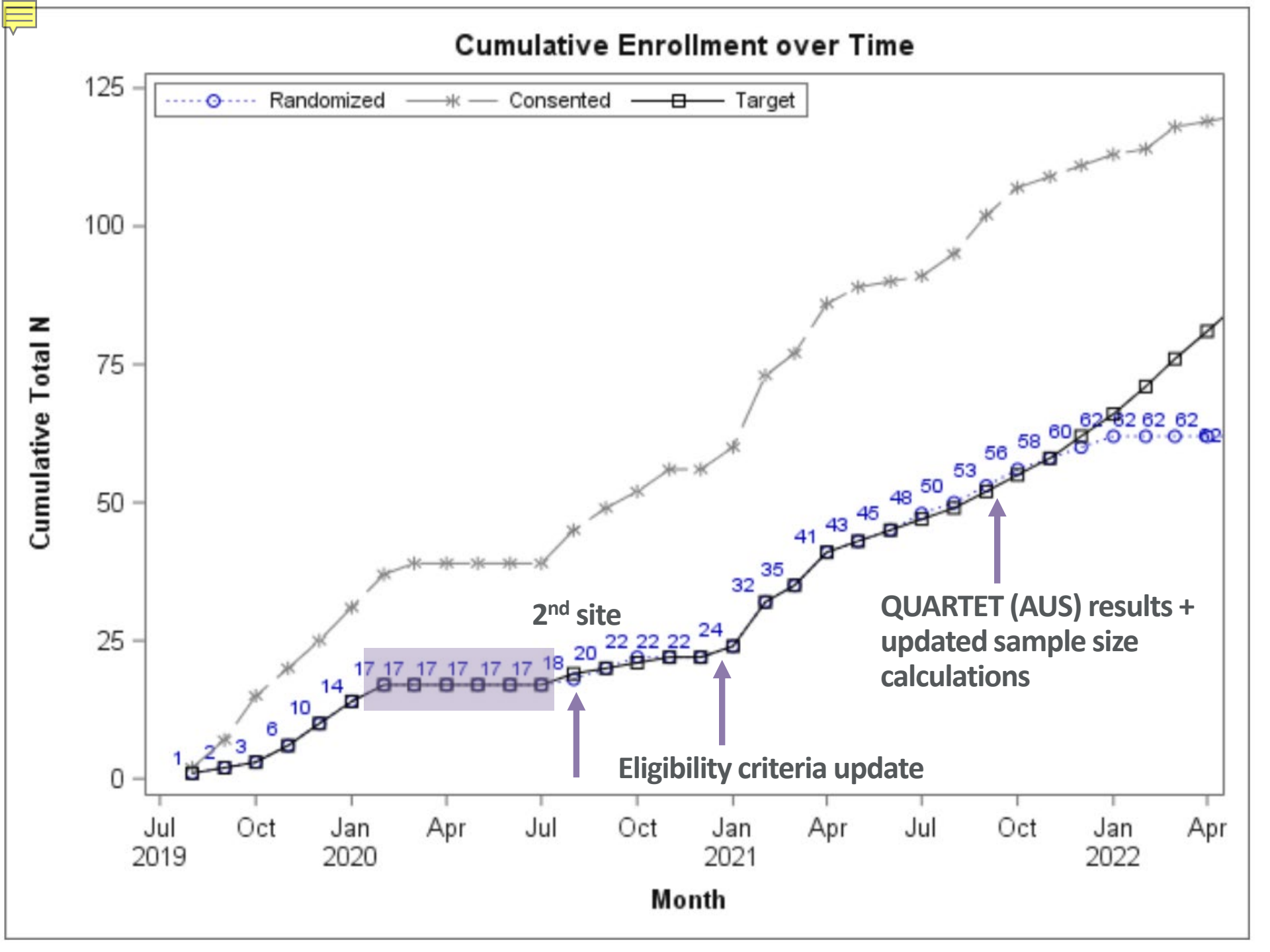
Interim Analysis

Our NEW sample size justification

"…we **conducted an interim conditional power analysis**, taking into consideration information from both the QUARTET USA trial data as of August 2021 and further the QUARTET (Australia) results. These interim analyses, **incorporating information to date**, suggested that a **sample size of at least 77**, and a **12% dropout rate**, would **provide over 90% conditional power** based on a sample of 87 randomized participants."

Note: understandably, the **huge** difference between original (N=364) and the final (N=87) recruitment targets required A LOT of **very careful, detailed but vague** justification

Baldridge, A. S., Huffman, M. D., Lazar, D., Abbas, H., Flowers, F. M., Quintana, A., ... & Ciolino, J. D. (2022). Efficacy and safety of a quadruple ultra-low-dose treatment for hypertension (QUARTET USA): Rationale and design for a randomized controlled trial. *American heart journal*, *254*, 183-193.


Cumulative Enrollment over Time
Randomized
Consented
Target
Cumulative Total N
125
100
75
50
25
0
Jul 2019
Oct
Jan 2020
Apr
Jul
Oct
Jan 2021
Apr
Jul
Oct
Jan 2022
Apr
Month
1 2 3 6 10 14 17 17 17 17 17 17 18 20 22 22 22 24 32 35 41 43 45 48 50 53 56 58 60 62 62 62 62 62
2nd site
Eligibility criteria update
QUARTET (AUS) results + updated sample size calculations

The decision on stopping the study

...came in collaboration with the funder and DSMB...

- Recruitment continued to be slow
- We ultimately ran out of time as we had agreed to halt the study by **May 2022**, regardless of recruitment numbers
- Now, we are in the process of analyzing the writing up study results
  - Primary results manuscript
  - clinicaltrials.gov updates pending
  - Qualitative analysis pending
  - We will be pooling our data with the Australia study data as well (they had over 600 study participants)

Thank you for your attention today + please feel free to reach out with comments/questions!

jody.ciolino@northwestern.edu

Acknowledgments – QUARTET USA

- **QUARTET USA** patients, study team members, and **DSMB** members (Paul Muntner [chair], Emily Anderson, Perla Herrera, Ken Jamerson, Chris Lindsell)

- **Northwestern**: Mark Huffman (MPI), Jody Ciolino (MPI), Abi Baldridge, Namratha Kandula, Sadiya Khan, Don Lloyd-Jones, Steve Persell, Jay Paparello

- **Access Community Health Network:** Dani Lazar, Jairo Mejia, Hiba Abbas, Fallon Flowers, Adriana Quintana, Patricia Helbin, Edgar Pizarro

- **University of Sydney:** Clara Chow

- **The George Institute for Global Health/UNSW:** Anthony Rodgers, Emily Atkins, Bruce Neal, Anushka Patel, MA Salam

- **Sponsors:** NHLBI (R33HL139852), NUCATS (U01TR003528), Northwestern

Funding Acknowledgements

The Liver Cirrhosis Network

- This research is supported by NIDDK, NIAAA and NCI via the following grants:
  - U01DK130177, Duke University
  - U01DK130197, University of Southern California
  - U01DK130185, University of Miami
  - U01DK130180, Cleveland Clinic
  - U01DK130221, Cornell / Columbia
  - U01DK130134, Virginia Commonwealth University
  - U01DK130168, University of California, San Francisco
  - U01DK130190, University of California, San Diego
  - U01DK130113, University of Michigan
  - U01DK130181, Mayo Clinic
  - U24DK130164, Northwestern University
- **The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health**.

Potentially Helpful References

- Friedman, LM, Furberg, CD, DeMets, DL, Reboussin, DM, Granger, CB. Fundamentals of Clinical Trials. Switzerland: Springer; 2015.
- Proschan, MA, Lan, KG, Wittes, JT. Statistical Monitoring of Clinical Trials: A Unified Approach. New York, NY: Springer Science & Business Media; 2006.
- Ellenberg, SS, Fleming, TR, DeMets, DL. Data Monitoring Committees in Clinical Trials: A Practical Perspective. Hoboken, NJ: John Wiley & Sons; 2019.
- Fleming, TR, DeMets, DL, Roe, MT, et al. Data monitoring committees: Promoting best practices to address emerging challenges. Clinical Trials. 2017;14(2):115–123.
- Jennison, C, Turnbull, BW. Group Sequential Methods With Applications to Clinical Trials. Boca Raton, FL: Chapman & Hall/CRC Press Taylor & Francis Group; 1999.
- Ciolino, JD, Renee'H, M, Zhao, W, Jauch, EC, Hill, MD, Palesch, YY. Continuous covariate imbalance and conditional power for clinical trial interim analyses. Contemp Clin Trials. 2014;38(1):9–18.
- Ciolino JD, Kaizer AM, Bonner LB. Guidance on interim analysis methods in clinical trials. Journal of Clinical and Translational Science. 2023 Jan;7(1):e124.