A Case Study in What Can Go Wrong: Perspectives on FDA's Accelerated Approval of Aducanumab in Alzheimer's Disease

Scott S. Emerson, M.D., Ph.D.

Professor Emeritus of Biostatistics

University of Washington

January 6, 2022

Abstract

On June 7, 2021, the US Food & Drug Administration (FDA) Center for Drug Evaluation and Research (CDER) announced the accelerated approval of aducanumab, a monoclonal antibody targeting amyloid, in the treatment of Alzheimer's Disease. This decision was made in the face of two confirmatory clinical trials that had been terminated for futility, as well as a negative opinion from the CDER Office of Biostatistics and an overwhelmingly negative recommendation (0 yes, 10 no, 1 abstention) from the Peripheral and Central Nervous System (PCNS) Advisory Committee that had been asked to review the evidence in support of the indication. The societal impact of the decision is far-reaching: Medicare insurance rates have been markedly increased for 2022 in anticipation of the high cost of a drug that many consider to be unproven. In this talk I will present a statistical perspective on the impact that clinical trial design issues had on the available evidence, the controversial data analyses that were presented to the committee, and the arguments put forth by the Office of Clinical Pharmacology in defense of the accelerated approval using a surrogate endpoint that had not been validated. In particular, I will touch on the important role that screening pilot studies play in drug discovery, the appropriate choice and implementation of sequential sampling in clinical trial designs where time varying treatment effects are of concern, the dangers of conditioning on post-randomization variables, and the proper validation of surrogate endpoints that might be used in drug approval.

Executive Summary

“He was against it.”

-Calvin Coolidge

Case Study: Aducanumab

Overview of Regulatory History

"Treatment Discovery"

Where am I going?

Alzheimer's Disease is a highly prevalent, serious disease having great impact on patient quality of life **and** societal resources

There is much controversy about the interpretation of the results of the pivotal RCT investigating aducanumab, as well as the accelerated approval that was ultimately granted by the FDA

Alzheimer's Disease Unmet Need

- Prevalence in US: 1.6% overall, 70% of dementia
  - 19% among 75-84 yo

- Economic impact (per AD Assn)
  - ~20% of medical care dollars
  - $355 billion / year
  - 11 million unpaid caregivers (valued at $230 billion / year)

- Available treatments
  - Cholinesterase inhibitors
  - Memantine (approved 2003 for moderate – severe AD)

- Failed trials: 244 compounds 2002-2014
  - 26 targeting amyloid, including 6 monoclonal Ab

Aducanumab Timeline

- 2005 Biogen intensified interest in Alzheimer's Disease
- 2011 IND and phase 1 trials of aducanumab
- 2012 – 2019 Phase 1b (Study 103)
- 2014 – 2015 EoP2 and development of SPA for phase 3
- 2015 Aug Start of phase 3 (Studies 301 and 302)
- 2017 Mar Amendment 4 high dose in Apo E4 carriers
- 2018 Dec Data cutoff for futility analysis
- 2019 Mar Futility declared, trials stopped
- 2019 Jun Final analysis discussed with FDA
- 2019 – 2020 Collaboration with FDA
- 2020 Nov PCNS votes 0 for approval, 10 against, 1 abstain

Fallout From PCNS AdCom

- Biogen stock price

Fallout From PCNS AdCom

NOVEMBER 8, 2020

Biogen, Inc. (BIIB) Neutral

Cast from That '70s Show Takes Over Aducanumab AdCom. BIIB and FDA Spanked with Slide Rule. Approval Still Possible.

WHAT YOU SHOULD KNOW:

Firstly, the **FDA is not required to follow the guidance of an AdCom**, and the agency seems fairly vested in the aducanumab regulatory process and seems to have given the submission the benefit of the doubt in several instances (likely triggering the gag-reflex in some panel members). Second, bad panels happen all the time and Friday reminded us of the Exubera panel that was nearly derailed on theoretical lung-function issues until a passionate diabetologist suggested: you are trivializing this disease, patients may lose their feet. We think bad-panel-syndrome may be an ongoing problem as the strong conflict of interest rules at the FDA make it hard to assemble seasoned expert panels.

There is little question the **FDA has dug itself a real hole if it still wants aducanumab**. The key questions from the panel (Q2 and Q8) asked if the 302 trial showed "strong" evidence to support approval. Q2 got one "yes" vote and Q8 got none. It's in the hands of the agency now, but the panel discussion also made it clear that **no one in the room was really excited** about the treatment effects seen in the 302 trial. As a result, we continue to see strong commercial headwinds for the drug based on the expected complex and expensive PET imaging required for diagnosis and the uncertainty about

Fallout From PCNS AdCom

1600 20th Street, NW • Washington, D.C. 20009 • 202/588-1000 • www.citizen.org

December 9, 2020

The Honorable Christi A. Grimm
Principal Deputy Inspector General
Office of Inspector General
U.S. Department of Health and Human Services
330 Independence Avenue SW
Washington, DC 20201

RE: Request for an Office of Inspector General investigation of the Food and Drug Administration's inappropriate close collaboration with Biogen before and after the submission of the biologics license application for aducanumab for treatment of Alzheimer's disease

Aducanumab Timeline: 2021-2022

- Jun 7 : PDUFA date - FDA granted accelerated approval
  - » Based on unvalidated surrogate of “reduction in amyloid”
  - » No restriction to patients with confirmed plaques, MCI, mild AD; Misleading presentation of clinical results
  - » Confirmatory trial results not required until 9 years
  - » Biogen announces $56,000/year cost of drug (~$100 K total)
  - » Several prominent members of PCNSAC resign
- Jun 29 : Two Congressional investigations announced
- Jul – Aug : HHS OIG investigation requested and announced
- Jul – now : Multiple insurers deny coverage, some hospitals will not use
- Nov 12 : CMS announces 14.5% increase in Medicare rates
- Nov 15 : Biogen Chief Scientific Officer retires, other layoffs
- Dec 16 : EMA recommends no approval
- Dec 20 : Biogen announces 50% reduction in price
- Jan 11 : Preliminary CMS decision re Medicare coverage
- Mar 18 : Publication of primary phase 3 results
- Apr 7 : Final CMS decision denying Medicare coverage except in RCT

Fallout After Accelerated Approval

- Biogen stock price

My View: Many Unforced Errors

- Biogen (and the FDA Office of Neuroscience) made many poor decisions
  - Rushing into phase 3 before safety data available
  - Poorly specified hierarchy for multiple endpoints
  - Poorly chosen (and implemented) futility rule
  - Failure to consider full potential impact of protocol amendments
  - ?Failure to explore results before accepting futility decision
  - Incorrect handling of discordant results from RCTs
  - Dubious analyses based on post-randomization conditioning
  - Accelerated approval based on unproven surrogate
  - Accelerated approval much broader than tested in RCT
  - Inadequate description in FDA package insert
  - Primary publication misleading in claims about secondary endpoints

This Module

- The ultimate impact of any errors made in the development and testing of aducanumab has large potential impact

- If the drug is truly ineffective
  - It should not have been approved

- If the drug is truly effective
  - Design, conduct, and analyses of the RCT did not provide the necessary compelling evidence
  - Biogen greatly delayed an appropriate approval of the drug

- Using the aducanumab setting, I will highlight
  - What was done appropriately vs inappropriately
  - What could have been done instead

Drug Discovery

Phases of Investigation

"Treatment Discovery"

Where am I going?

US regulation of drugs requires evidence that the drug is effective and safe

Such evidence typically accumulates through a several "phases" of clinical trials

Failure to consider the statistical properties of such phased investigations can lead to failures of late stage RCT

Science and Statistics

- Statistics is about science
  - Estimating and quantifying precision of answers to questions deemed important to scientists
  - (Science in the broadest sense of the word)

- Science is about proving things to people
  - Discriminating between the most important competing hypotheses at the time
  - (The validity of any proof rests solely on the willingness of the audience to believe it)

Overall Goal

- "Drug discovery"
  - More generally
    - a therapy / preventive strategy or diagnostic / prognostic procedure
    - for some disease
    - in some population of patients

- A series of experiments to establish
  - Safety of investigations / dose
  - Safety of therapy
  - Measures of efficacy
    - Treatment, population, and outcomes
  - Confirmation of efficacy
  - Confirmation of effectiveness

U. S. Regulation of Drugs / Biologics

- Wiley Act (1906)
  - Labeling
- Food, Drug, and Cosmetics Act of 1938
  - Safety
- Kefauver – Harris Amendment (1962)
  - Efficacy / effectiveness
    - " [If] there is a lack of substantial evidence that the drug will have the effect ... shall issue an order refusing to approve the application. "
    - "...The term 'substantial evidence' means evidence consisting of **adequate and well-controlled investigations, including clinical investigations**, by experts qualified by scientific training"
- FDA Amendments Act (2007)
  - Registration of RCTs, Pediatrics, Risk Evaluation and Mitigation Strategies (REMS)

U.S. Regulation of Medical Devices

- Medical Devices Regulation Act of 1976
  - Class I: General controls for lowest risk
  - Class II: Special controls for medium risk - 510(k)
  - Class III: Pre marketing approval (PMA) for highest risk
    - "…**valid scientific evidence** for the purpose of determining the safety or effectiveness of a particular device … adequate to support a determination that there is reasonable assurance that the device is safe and effective for its conditions of use…"
    - "Valid scientific evidence is evidence from well-controlled investigations, partially controlled studies, studies and objective trials without matched controls, well-documented case histories conducted by qualified experts, and reports of significant human experience with a marketed device, **from which it can fairly and responsibly be concluded by qualified experts that there is reasonable assurance of the safety and effectiveness**…"
- Safe Medical Devices Act of 1990
  - Tightened requirements for Class 3 devices

Treatment “Indication”

- Disease
  - Therapy: Putative cause vs signs / symptoms
    - May involve method of diagnosis, response to therapies
  - Prevention / Diagnosis: Risk classification

- Population
  - Therapy: Restrict by risk of AEs or actual prior experience
  - Prevention / Diagnosis: Restrict by contraindications

- Treatment or treatment strategy
  - Formulation, administration, dose, frequency, duration, ancillary therapies

- Outcome
  - Clinical vs surrogate; timeframe; method of measurement

Ideal Results

- Goals of “drug discovery” are similar to those of diagnostic testing in clinical medicine
- We want a “drug discovery” process in which there is
  - A low probability of adopting ineffective drugs
  - A high probability of adopting truly effective drugs
  - A high probability that adopted drugs are truly effective

Application to Drug Discovery

- We consider a population of candidate drugs using RCT to "diagnose" truly beneficial drugs

- Use both frequentist and Bayesian criteria
  - Sponsor:
    - High probability of adopting a beneficial drug (freq power)
  - Regulatory:
    - Low probability of adopting ineffective drug (freq type 1 error)
    - High probability that adopted drugs work (Bayes post prob)
  - Public Health:
    - Maximize the number of good drugs adopted
    - Minimize the number of ineffective drugs adopted

Frequentist and Bayesian

- Bayes rule: PPV depends on type I error, power, and prevalence
  - Maximize new information by maximizing Bayes factor

$$PPV = \frac{power \times prevalence}{power \times prevalence \quad + \quad type\ I\ err \times (1 - prevalence)}$$

$$\frac{PPV}{1 - PPV} = \frac{power}{type\ I\ err} \times \frac{prevalence}{1 - prevalence}$$

$$posterior\ \ odds = Bayes\ Factor \times prior\ \ odds$$

  - **KEY POINT:** Inflation of type 1 error has major impact on the probability that an approved drug truly works
    - Need to consider relative increase in type 1 error, not difference
    - Type 1 error of 0.06 is a 20% relative increase over 0.05

Statistics and Art

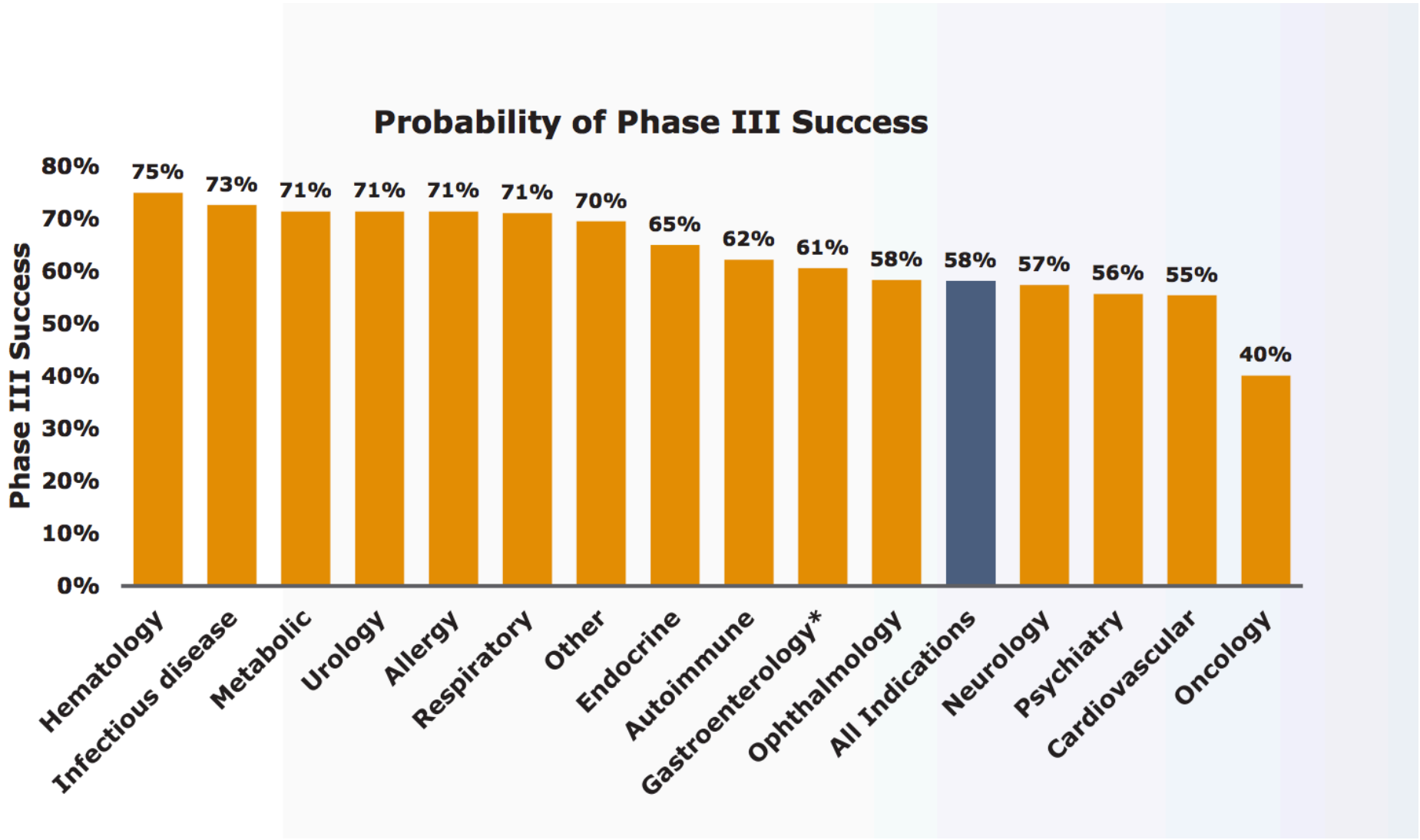
- "In statistics, as in art, never fall in love with your model."
  - G.E.P. Box: "All models are false, some models are useful"

- "In statistics, as in art, value depends heavily on provenance."
  - Unless we understand where the data is coming from, the resulting statistical analyses are of little value

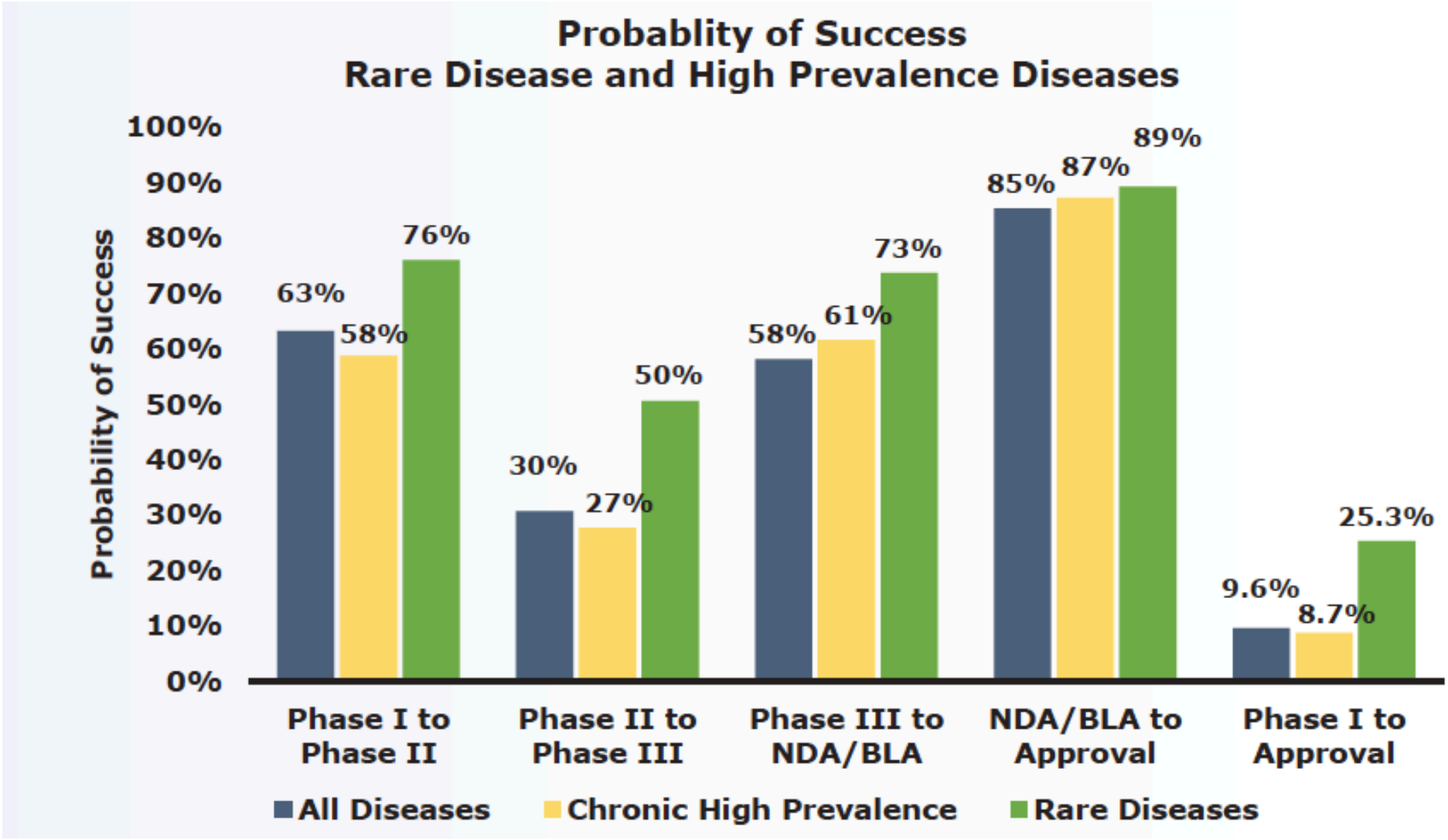
Inflation of the Type I Error

- We must be concerned about data dredging ("data mining")
  - Selective reporting of studies
  - Revising outcomes to reflect the most promising results
    - Both clinical outcome or statistical summarization important
  - Revising eligibility criteria based on subgroup analyses
  - Changing from surrogate efficacy to effectiveness endpoints
    - "Treating the symptom not the disease"

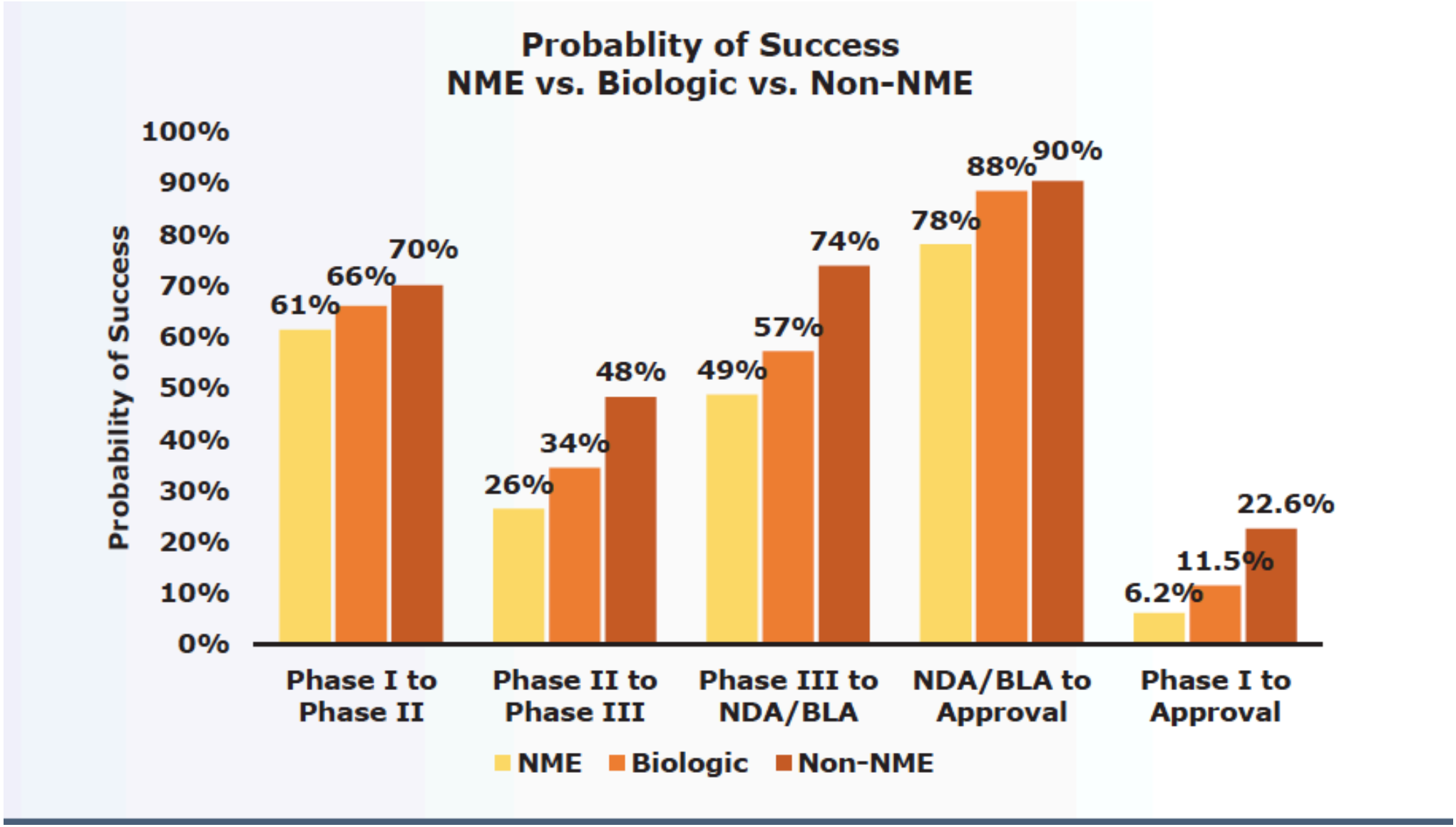
- In order to avoid inflation of type I error, we require confirmatory studies using prespecified indication and statistical analysis
  - Protocols
  - Statistical analysis plans (SAP)
  - Registration of RCT on ClinicalTrials.gov



From Declan Doogan, MD



From Declan Doogan, MD



From Declan Doogan, MD

Preliminary Studies in Screening

- Two general approaches to studying new treatments

- Scenario 1:
  - Study every treatment in a large definitive experiment
    - Only do Phase III studies
      - Level of significance 0.025, high power
    - (Ignore, for now, the safety / ethics of this)

- Scenario 2:
  - Perform small screening trials, with confirmatory trials of promising treatments passing early tests
    - Phase II studies
      - Level of significance, power (sample size) to be determined
    - Confirmatory
      - Level of significance 0.025, high power

Scenario 1: Only Phase III

- Only large trials using 1,000,000 subjects
  - 10% of drugs being investigated truly work
  - Level of significance .025, .025, or 0.05

  - Sample size / power
    - 979 subjects, α=0.025, 97.5% power ➔ 1,021 RCT
    - 500 subjects, α=0.025, 80.0% power ➔ 2,000 RCT
    - 394 subjects, α=0.050, 80.0% power ➔ 2,538 RCT

  - Results
    - N= 979: 99 effective / 23 ineffective (PV+ = .81)
    - N= 500: 160 effective / 45 ineffective (PV+ = .78)
    - N= 394: 202 effective / 114 ineffective (PV+ = .64)

Scenario 2a: Screening Phase II

- Use 700,000 subjects in Phase II studies
  - 10% of drugs being investigated truly work
  - Level of significance .025
  - Sample size / power
    - 100 subjects provide 24% power ➔ 7,000 RCT
  - Results
    - N= 100: 168 effective / 158 ineffective (PV+ = .52)

- Use 300,000 subjects in confirmatory Phase III studies
  - 52% of drugs being investigated truly work
  - Level of significance .025
  - Sample size / power
    - 921 subjects provide 96.7% power ➔ 326 RCT
  - Results
    - N= 921: 162 effective / 4 ineffective (PV+ = .98)

Scenario 2b: Screening Phase II

- Use 700,000 subjects in Phase II studies
  - 10% of drugs being investigated truly work
  - Level of significance .10
  - Sample size / power
    - 342 subjects provide 85% power ➔ 2,047 RCT
  - Results
    - N= 342: 173 effective / 184 ineffective (PV+ = .49)

- Use 300,000 subjects in confirmatory Phase III studies
  - 49% of drugs being investigated truly work
  - Level of significance .025
  - Sample size / power
    - 839 subjects provide 95% power ➔ 357 RCT
  - Results
    - N= 839: 165 effective / 5 ineffective (PV+ = .97)

Summary: “Drug Discovery”

	Scenario 1	Scenario 2a	Scenario 2b
Phase 2			
Number RCT	**2,000 (10% eff)**	**7,000 (10% eff)**	**2,047 (10% eff)**
N per RCT	0	100	342
Type 1 err; Pwr		0.025; 24%	0.100; 85%
“Positive” RCT		168 eff; 158 not	173 eff; 184 not
Confirmtaory Phase 3			
Number RCT	2,000 (10% eff)	326 (52% eff)	357 (49% eff)
N per RCT	500	921	839
Type 1 err, Pwr	0.025; 80%	0.025; 97%	0.025; 95%
# Effctve Adopt	**160**	**162**	**165**
# Ineff Adopt	**45**	**4**	**5**
Pred Val Pos	**78%**	**98%**	**97%**
N per Adopt	**500**	**1,021**	**1,181**

Furthermore

- Additional advantages of screening trials
  - Gathering more detailed preliminary safety data before embarking on expensive, large scale Phase 3 trials
  - Gathering preliminary efficacy data that allows fine tuning
  - Fine tune eligibility criteria
    - Include only susceptible patient populations
    - Exclude patients at high risk for AEs
  - Optimal treatment strategies
    - Fine tune formulation, dose, administration, frequency, duration
    - Develop dose modification strategies
    - Prophylactic treatments, rescue treatments for AEs
  - Optimal clinical endpoints
- Major disadvantage
  - “White space” (time delay) between phase 2 and phase 3
  - (Truly an issue for sponsors, rather than public health)

Inflation of the Type I Error

- Recall that in order to avoid inflation of type I error, we require confirmatory studies using prespecified
  - Patient population
  - Treatment
  - Primary clinical outcome
  - Statistical analysis

- Hence, we must be concerned about data dredging ("data mining") of the phase 2 data, because it may lead to differences between phase 2 and phase 3 due to
  - Random high bias in "positive" phase 2 results
  - Revising outcomes to reflect the most promising results
  - Revising eligibility criteria based on subgroup analyses
  - Changing from surrogate efficacy to effectiveness endpoints
    - "Treating the symptom not the disease"

The Problem of Small Studies

- Using 700,000 patients
  - Small sample size ➔ Big bias of "positive" studies

			Null: Δ =0			Alt: Δ= .125		
N per RCT	**RCTs**	**Crit Value**	**Prob Sig**	**N Sig RCT**	**Expected Estimate**	**Prob Sig**	**N Sig RCT**	**Expected Estimate**
7000	**100**	**0.0234**	0.025	2	0.028	1.000	100	0.125
3500	**200**	**0.0331**	0.025	5	0.039	1.000	200	0.125
700	**1000**	**0.0741**	0.025	25	0.089	0.912	912	0.132
350	**2000**	**0.1048**	0.025	50	0.125	0.649	1,298	0.156
70	**10000**	**0.2343**	0.025	250	0.280	0.180	1,801	0.299
35	**20000**	**0.3313**	0.025	500	0.390	0.114	2,271	0.407

Phase 3 Sample Size from Phase 2

- Phase 2: type 1 error 0.10 with N= 342 provides 85% power
  - Estimated treatment effects: mn (sd; min – max)
    - Ineffective drugs 0.095 (0.022; 0.069 - 0.313)
    - Effective drugs 0.140 (0.043; 0.069 - 0.396)

- Phase 3 using Phase 2 results
  - Estimated N for 95% power:
    - Ineffective drugs 1665 (610; 134 – 2745)
    - Effective drugs 893 (571; 84 – 2745)
  - Type 1 error 0.025, avg power 86%

- Screen 1,759 drugs with 1,000,000 patients
  - End of phase 2: 150 effective, 159 ineffective (PVP= 0.49)
  - End of phase 3: 129 effective, 4 ineffective (PVP= 0.97)

Examples without Error Control

- Consideration of multiple summary measures
  - Mean, geometric mean, Wilcoxon, median, two proportions
  - Type 1 error 0.10 ➔ 0.23
  - Power 0.85 ➔ 0.92

- Consideration of subgroups
  - Overall sample and equal subgroups defined by three variables
  - Type 1 error 0.10 ➔ 0.33
  - Power 0.85 ➔ 0.95

- Consideration of change of endpoint between phase 2 and 3
  - Phase 2: potential surrogate and Phase 3: clinical outcome
  - Type 1 error 0.10 ➔ 0.19 (10%) ➔ 0.27 (20% misleading)
  - Power 0.85 ➔ 0.85

Frequentist and Bayesian

- Bayes rule: PPV depends on type I error, power, and prevalence
  - Maximize new information by maximizing Bayes factor

$$PPV = \frac{power \times prevalence}{power \times prevalence \quad + \quad type\ I\ err \times (1 - prevalence)}$$

$$\frac{PPV}{1 - PPV} = \frac{power}{type\ I\ err} \times \frac{prevalence}{1 - prevalence}$$

$$posterior\ \ odds = Bayes\ Factor \times prior\ \ odds$$

  - **KEY POINT:** Inflation of type 1 error has major impact on the probability that an approved drug truly works
    - Need to consider relative increase in type 1 error, not difference
    - Type 1 error of 0.06 is a 20% relative increase over 0.05

Compare: Fix Resources, 10% Prevalence

	RCT	Eff (TP)	Not(FP)	n
No Screening trials				
• Homogeneous effect	2,000	160(160)	45	500
Nonadaptive				
• Homogeneous effect	2,040	165(165)	5	1,181
• Homogeneous,10% bad surrog	1,812	147(147)	8	1,181
• Homogeneous,20% bad surrog	1,627	132(132)	12	1,181
• Inhomogeneous effect	2,123	99(0)	5	1,181
Adaptive subgroups: inflate error				
• Homogeneous effect	1,488	134(43)	11	1,181
• Inhomogeneous effect	1,493	122(88)	11	1,181
Adaptive subgroups: control error				
• Homogeneous effect	2,040	153(56)	4	1,277
• Inhomogeneous effect	2,067	135(103)	4	1,277

Confirmatory Trials

- Screening Phase II trials followed by confirmatory RCT provide great protection
  - Ensure overwhelming majority of adopted drugs are truly effective

- Control of type I and II errors are of great importance at phase 2
  - But note that type 1 error of 0.025 not necessarily indicated

- Adaptive designs can help provide that control
  - But need to re-power the study to get greatest benefit
  - The added benefit over nonadaptive designs is not huge, but
    - Higher power and predictive value of the positive
    - More beneficial drugs identified with more safety data

- Adaptation cannot protect against false surrogates

Seamless Phase 2 / 3

- In cases that no changes will be made between Phase 2 and Phase 3, can try to use same trial
  - Need to ensure that same level of evidence is provided as would be in two independent trials
    - Pivotal 0.005 vs 0.000625 in two independent trials?
    - One RCT setting vs two RCT settings (random effects)

- Such would eliminate “white space”
  - Nothing presented here specific to separate Phase 2 / Phase 3
  - But note that white space is truly an issue for those whose focus is on a particular agent
  - During “white space” other agents in the pipeline can be investigated
  - Eliminating “white space” limits scientific, regulatory, and ethical review of phase 2 results

Major Conclusions

- There is no substitute for planning a study in advance
  - At Phase 2, adaptive designs may be useful to better control parameters leading to Phase 3
    - Most importantly, learn to take "NO" for an answer
  - At Phase 3, there seems little to be gained from adaptive trials
    - We need to be able to do inference, and poorly designed adaptive trials can lead to some very perplexing estimation methods

- **"Opportunity is missed by most people because it is dressed in overalls and looks like work."** -- Thomas Edison

- In clinical science, it is the steady, incremental steps that are likely to have the greatest impact.

Case Study: Aducanumab

Early Phase RCT

Where am I going?

Early investigation of a drug might include phase 1 dose finding studies based on concurrent or staggered cohorts

Phase 2 studies gather preliminary evidence for an indication closer to the ultimate goal

- Safety
- Efficacy (often surrogate)

Results of these early phase studies are meant to inform design of a confirmatory study

Alzheimer's Disease Unmet Need

- Prevalence in US: 1.6% overall
  - 19% among 75-84 yo
- Economic impact (per AD Assn)
  - ~20% of medical care dollars
  - $355 billion / year
  - 11 million unpaid caregivers (valued at $230 billion / year)
- Available treatments
  - Cholinesterase inhibitors
  - Memantine (approved 2003 for moderate – severe AD)
- Failed trials: 244 compounds 2002-2014
  - 26 targeting amyloid, including 6 monoclonal Ab

Aducanumab Efficacy Hypothesis

- Previous failure of anti-amyloid treatments may be overcome by
  - Exploring monoclonal antibodies from human B-cells collected from elderly subjects with no or only minimal cognitive impairment
  - Monoclonal antibody selective for Aβ aggregates, but not monomers
  - Shifting focus to earlier disease
  - Outcome measures more sensitive to early declines In cognition

Aducanumab Safety Hypothesis

- Amyloid related imaging abnormalities (ARIA-E and ARIA-H) remain a risk especially when targeting aggregates and amyloid deposition
  - (hence a suggestion that ARIA might go hand in hand with efficacy)

- ARIA
  - occurs early in treatment,
  - typically asymptomatic,
  - more common in ApoE4 carriers at higher doses

Phase 1b: Study 103 Design

- Key features of patient eligibility
  - Earlier stage of disease, including predementia
  - Brain amyloid pathology confirmed by PET imaging
  - Stratified by APOE4 positivity

- Endpoints
  - Primary: safety and tolerability
  - Secondary: Brain amyloid by PET; Pharmacokinetics; Immunogenicity
  - Exploratory: Clinical cognition: CDR-SB, MMSE, others?

- Staggered cohorts differing by increasing dose and eligibility criteria across 12 protocol amendments
  - 12 month study
  - 4:1 randomization

Phase 1b: Study 103 Results

- Publication used pooled placebo cohorts
  - (FDA later noted issue with handling of missing data)

- Major findings informing design of phase 3 studies
  - Focus on 10mg dose (with uptitration phase)
  - Focus on earlier disease
  - ARIA worse in ApoE4
    - 18 month study to allow for uptitration
    - Lower dose in ApoE4 carriers
  - CDR-SB sensitive enough to detect early changes

Case Study: Aducanumab

Confirmatory Phase 3: Indication

Where am I going?

First step in design of confirmatory RCT is pre-specification of

- Disease
- Patient population
- Treatment
- Clinical outcome

Treatment "Indication"

- Disease
  - Therapy: Putative cause vs signs / symptoms
    - May involve method of diagnosis, response to therapies
  - Prevention / Diagnosis: Risk classification

- Population
  - Therapy: Restrict by risk of AEs or actual prior experience
  - Prevention / Diagnosis: Restrict by contraindications

- Treatment or treatment strategy
  - Formulation, administration, dose, frequency, duration, ancillary therapies

- Outcome
  - Clinical vs surrogate; timeframe; method of measurement

Disease

- Met clinical criteria for MCI due to AD or mild AD dementia, with amyloid pathology confirmed by visual assessment of amyloid positron emission tomography

- This patient population is consistent with stage 3 and 4 patients as described in the FDA 2018 Guidance for Industry Early Alzheimer's Disease: Developing Drugs for Treatment

Population

- Exclusion due to concomitant disease
  - Confounding pathology on brain MRI
  - Other disease that might be associated with dementia

- (Not really clear whether the above are merely RCT issues, or actually part of the indication
  - Brain pathology could conceivably be a contraindication, though my best guess is that it is not
  - The other disease is probably not a contraindication, so long as the patients have amyloid on PET)

Treatment

- Consider low dose and high dose

- Definition of “low” and “high” vary by ApoE4 status
  - ApoE4 carriers
    - Low is 3 mg/kg uptitrated over 8 weeks
    - High is 6 mg/kg uptitrated over 24 weeks
  - ApoE4 noncarriers
    - Low is 6 mg/kg uptitrated over 24 weeks
    - High is 10 mg/kg uptitrated over 24 weeks

- Treatments injections administered every 4 weeks

- Dosing suspended with observed ARIA

Clinical Outcome

- Improvement or slower progression in cognitive impairment

- Multiple instruments possible
  - Clinical Dementia Rating – Sum of Boxes (CDR-SB)
  - Mini-mental status exam (MMSE)
  - Alzheimer's Disease Assessment Scale- Cognitive Subscale (ADAS-Cog13)
  - Alzheimer's Disease Cooperative Study Activities of Daily Living – Mild Cognitive Impairment (ADCS-ADL-MCI)

- Note: AD is associated with shortened survival
  - I would consider that slight degradations in survival might be acceptable with major beneficial effects on cognition
  - Hence, I would be reluctant to use overall survival as primary

Case Study: Aducanumab

Confirmatory Phase 3: Scientific Study Design

Where am I going?

Review the specification of

- Treatment arms (doses of investigational drug, control)
- Treatment assignment (randomization, blinding)
- Patient eligibility
- Assessment of outcomes
  - Efficacy: primary, secondary, exploratory
  - Safety: Special monitoring, adverse events

Treatment Arms and Eligibility

- Two identically designed RCTs (301 and 302)

- Double blind randomization in 1:1:1 ratio
  - High dose, low dose, placebo
  - Stratified by ApoE4 carrier status and site

- Patient eligibility excludes subjects based on typical criteria that detract from RCT procedures
  - Liver, kidney, metabolic, infectious disease
  - Inability to comply with study procedures
  - etc.

Assessment of Outcomes

- Primary comparison: High dose vs placebo

- Efficacy
  - Cognitive tests 6, 12, 18 months
    - CDR-SB primary
    - Summarized by **mean change** from baseline **at 78 weeks**
  - PET imaging 6 and 18 months (in subset)

- Safety
  - Brain MRI for ARIA weeks 14, 22, 30, 42, 54, 78 or prn
    - Incidence by severity
  - AEs via phone calls q4w or spontaneous reports continuously

Case Study: Aducanumab

Confirmatory Phase 3: Statistical Study Design

Where am I going?

Review the specification of

- Summarization of treatment effect
- Statistical analysis models
- Planning for missing data
- Sequential sampling
- Hierarchy of multiple endpoints

Statistical Analysis Model

- Primary endpoint: Mean change in CDR-SB over 78 weeks
  - Intent to treat (ITT) population (modified: only dosed subjects)

- Mixed model repeated measures (MMRM) using available data
  - Adjusted for
    - treatment, categorical visit, treatment-by-visit interaction,
    - baseline score, baseline score-by-visit interaction,
    - baseline MMSE score
    - AD symptomatic medication use at baseline,
    - region, and
    - ApoE4 status (carrier and noncarrier).

Planning for Missing Data

- In the presence of missing data, such a model presumes a missing at random model
  - Subjects with missing data “imputed” to behave like similar patients who remain under study
  - (a typical starting place for analyses)

- Statistical analysis plan (SAP) stipulated some sensitivity analyses to other patterns of missing data
  - (I would place greatest emphasis on the tipping point analysis)

SAP: Sensitivity to Missing Data

Table 2. Analysis for Primary and Secondary Endpoints

Endpoint	Analysis	Analysis Population	SAP Section
CDR-SB	Primary: Analysis of change from baseline at Week 78 (MMRM)	ITT	4.3.2.1
	Sensitivity: Pattern mixture model (ANCOVA)	ITT	4.3.2.2.1
	Sensitivity: Copy increment from reference method (ANCOVA)	ITT	4.3.2.2.2
	Sensitivity: Imputation by natural disease progression (ANCOVA)	ITT	4.3.2.2.3
	Sensitivity: Tipping point analysis (ANCOVA)	ITT	4.3.2.2.4
	Supplementary: Censoring after intercurrent events (MMRM)*	ITT	4.3.2.3.1
	Supplementary: Per-protocol analysis (MMRM)	Per-protocol	4.3.2.3.2
	Supplementary: Responder analysis (Logistic regression)	ITT	4.3.2.3.3
	Supplementary: Slope analysis (MMRM)	ITT	4.3.2.3.4
	Supplementary: Divergence effect analysis (MMRM)	ITT	4.3.2.3.5

Sample Size Determination

- Calculated based on a 90% power to detect a mean difference of 0.5 in change from baseline in CDR-SB score at week 78, based on a two-sided .05 test
  - N = 450 / arm in each study estimated initially
  - Modified to N = 535 / arm in pre-specified blinded sample size reestimation

- Ideally, such an analysis should have considered
  - Attenuation of effect due to treatment discontinuation
    - ARIA called for suspension of treatment
    - Presumably patients who do not take the drug will have lesser response in the data they contribute to ITT analysis
  - Imprecision of MAR model due to loss to follow-up

Futility Interim Analysis

- Analysis of data after half of subjects have 78 week data available

- Goal is to terminate the study if results suggestive of a treatment effect that is not clinically important

- Many of the issues that arose during scientific and regulatory review can be traced to
  - A poor choice of futility rule
  - An apparent lack of understanding of how it might behave during the eventual conduct of the study

Evaluation of Designs

- Process of choosing a trial design
  - Define candidate design
  - Evaluate operating characteristics
  - Modify design
  - Iterate

Evaluation of Designs: Fixed Sample

- Operating characteristics for fixed sample studies
  - Level of Significance (often pre-specified)
  - Sample size requirements
  - Power Curve
  - Decision Boundary
  - Frequentist inference on the Boundary
  - Bayesian posterior probabilities

Evaluation of Designs: Sequential

- Additional operating characteristics for group sequential studies
  - Probability distribution for sample size
  - Stopping probabilities
  - Boundaries at each analysis
  - Frequentist inference at each analysis
  - Bayesian inference at each analysis
  - Futility measures at each analysis

Evaluation of Designs

- Futility measures
  - Consider the probability that a different decision would result if trial continued
  - Can be based on
    - particular (design) hypotheses rejected by the boundary,
    - current best estimate, or
    - Bayesian predictive probabilities
  - Perhaps best measure of futility is whether the stopping rule has changed the power curve substantially

Group Sequential Approach

- Perform analyses when sample sizes $N_1 \ldots N_J$
  - Can be randomly determined if independent of effect
- At each analysis choose stopping boundaries
  - $a_j < b_j < c_j < d_j$
  - Often chosen according to some boundary shape function
    - O'Brien-Fleming, Pocock, Triangular, …
- Compute test statistic $T_j = T(X_1 \ldots X_{Nj})$
  - Stop if $T_j < a_j$ (extremely low)
  - Stop if $b_j < T_j < c_j$ (approximate equivalence)
  - Stop if $T_j > d_j$ (extremely high)
  - Otherwise continue

Stopping Boundary Scales

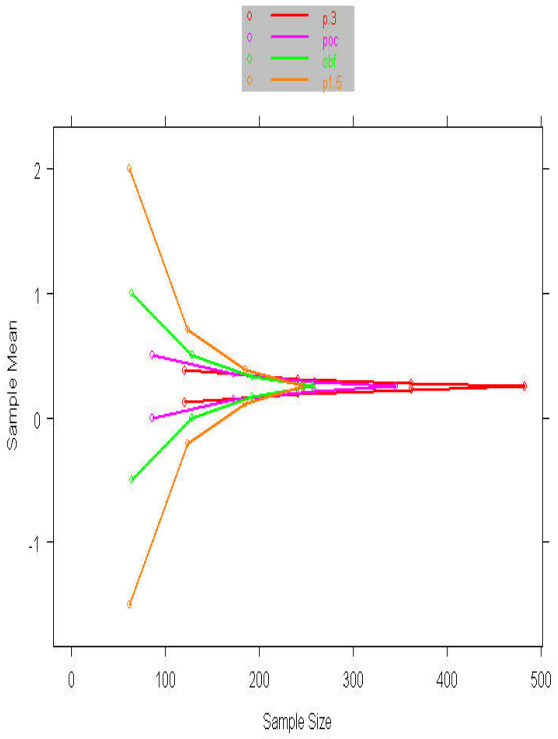
- Boundary scales (1:1 transformations among these)
  - Z statistic
  - P value
    - Fixed sample (so wrong)
    - Computed under sequential sampling rule (so correct)
  - Error spending function
  - Estimates
    - MLE (biased due to stopping rule)
    - Adjusted for stopping rule
  - Conditional power
    - Computed under design alternative
    - Computed under current MLE
  - Predictive power
    - Computed under flat prior (possibly improper)

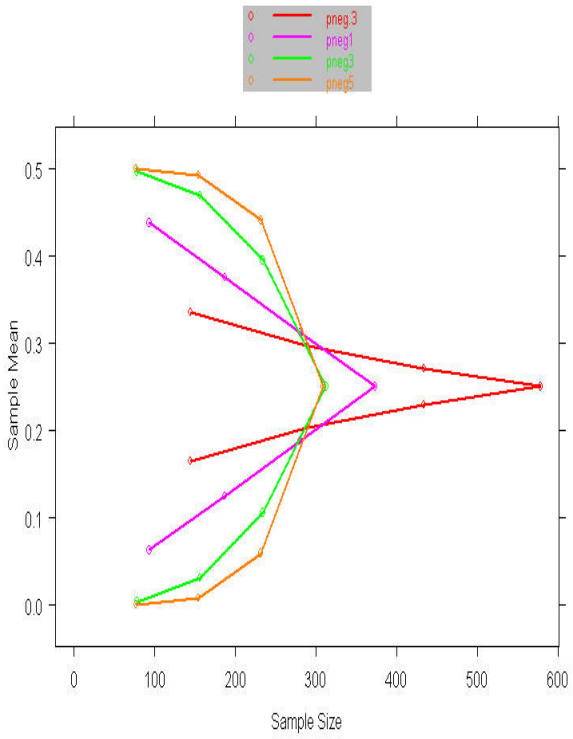
Exploring Group Sequential Designs

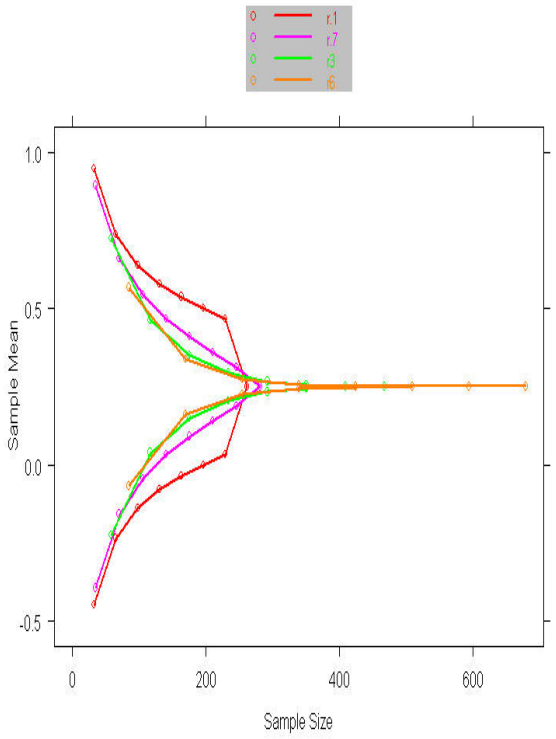
- Examining operating characteristics
  - Stopping boundaries
    - Z scale
    - Conditional power under hypothesized effects
    - Conditional power under current MLE
    - Predictive power under flat prior
  - Estimates and inference
    - MLE (Bias adjusted estimates suppressed for space)
    - 95% CI properly adjusted for stopping rule
    - P value properly adjusted for stopping rule
  - Power at specified alternatives
  - Sample size distribution (as function of true effect)
    - Maximal sample size
    - Average sample size

Spectrum of Boundary Shapes

- All of the rules depicted have the same type I error and power to detect the design alternative



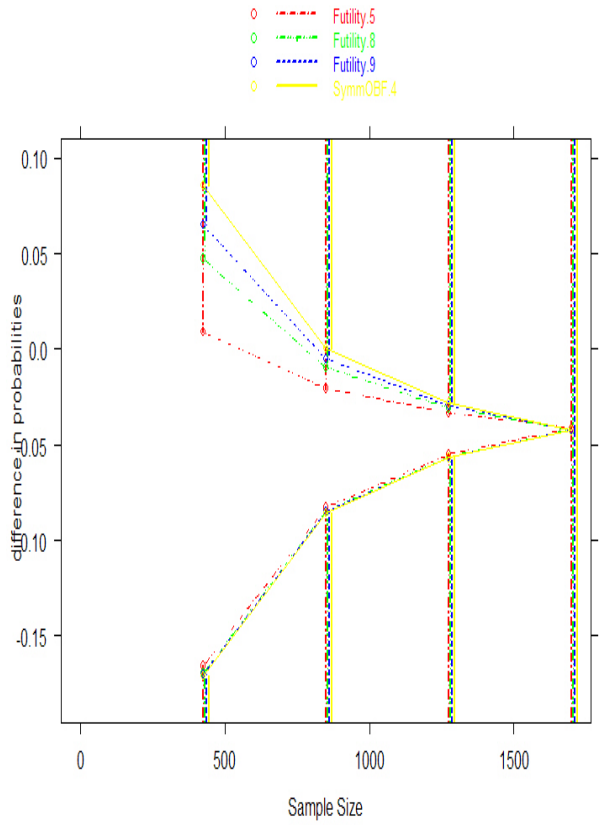




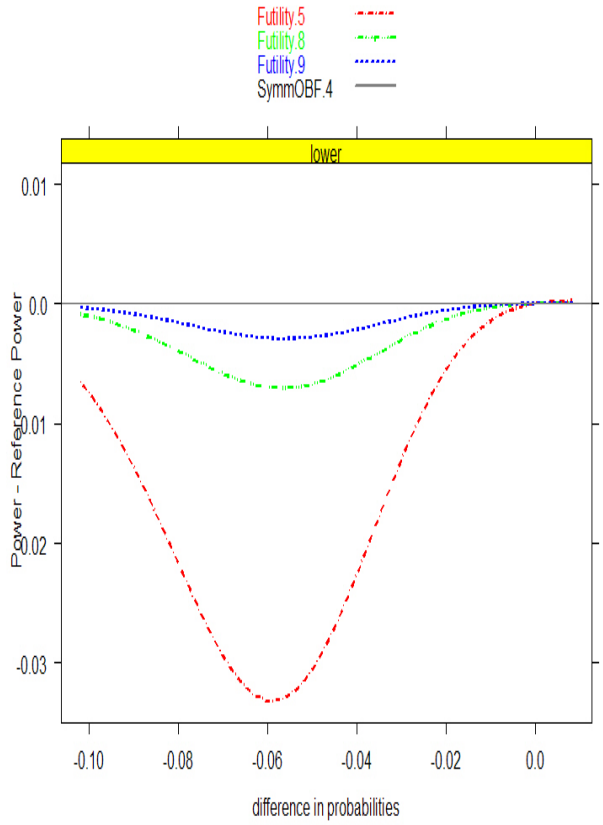
Efficiency / Unconditional Power

- Tradeoffs between early stopping and loss of power

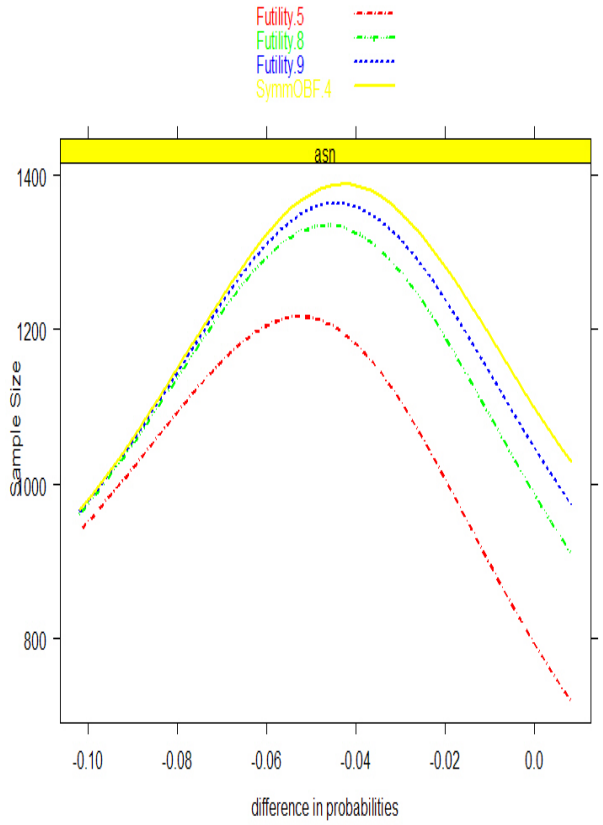
Boundaries



Loss of Power



Avg Sample Size



Illustration

- Can illustrate some basic ideas using a GSD I often recommend

- Efficacy: O'Brien-Fleming
  - Only stop when data suggests a major treatment effect
  - (coincidentally: rules out null with extreme confidence)

- Futility: Intermediate to Pocock and O'Brien-Fleming
  - Stop study when results have convincingly ruled out clinically important differences, AND
  - Use of such a rule does not materially affect study power / precision

O'Brien-Fleming & Futility: J = 4

- Introduce four evenly spaced analyses
  - Type 1 error of 0.025
- Stopping boundary table
  - (cf: Z statistic threshold of 1.96 in fixed sample test)

Info Frac	Futility				Efficacy			
	Z	CP_{alt}	CP_{est}	PP_{flat}	Z	CP_{null}	CP_{est}	PP_{flat}
0.25	-1.108	0.719	0.000	0.008	3.976	0.500	0.999	0.999
0.50	0.321	0.648	0.015	0.063	2.811	0.500	0.997	0.977
0.75	1.258	0.592	0.142	0.177	2.295	0.500	0.907	0.874
1.00	1.988	--	--	--	1.988	--	--	--

O'Brien-Fleming & Futility: J = 4, N = 1

- Introduce four evenly spaced analyses
  - Maintain sample size $N_J = 1$
- Estimates and inference table

Samp Size	Futility			Efficacy		
	MLE	95% CI	P	MLE	95% CI	P
0.25	-2.216	(-4.71, 1.74)	0.846	7.951	(4.00, 10.5)	0.000
0.50	0.454	(-1.60, 3.31)	0.263	3.976	(1.14, 6.04)	0.003
0.75	1.452	(-0.36, 3.85)	0.053	2.650	(0.30, 4.48)	0.013
1.00	1.988	(0.00, 4.06)	0.025	1.988	(0.00, 4.06)	0.025

O'Brien-Fleming & Futility: J = 4, N = 1

- Introduce four evenly spaced analyses
  - Maintain sample size $N_J = 1$
- Power and sample size table

True Effect	Power	Avg N
0.00	0.025	0.580
1.96	0.478	0.783
2.80	0.776	0.761
3.24	0.882	0.723
3.92	0.966	0.650

O'Brien-Fleming & Futility: J = 4, Power

- Introduce four evenly spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Estimates and inference table

Samp Size	Futility			Efficacy		
	MLE	95% CI	P	MLE	95% CI	P
0.27	-2.141	(-4.55, 1.68)	0.846	7.682	(3.86, 10.1)	0.000
0.54	0.439	(-1.55, 3.20)	0.263	3.841	(1.10, 5.84)	0.003
0.80	1.403	(-0.34, 3.72)	0.053	2.561	(0.29, 4.33)	0.013
1.07	1.920	(0.00, 3.92)	0.025	1.920	(0.00, 3.92)	0.025

O'Brien-Fleming & Futility: J = 4, Power

- Introduce four evenly spaced analyses
  - Maintain power 0.975 at alternative 3.92
- Power and sample size table

True Effect	Power	Avg N
0.00	0.025	0.622
1.96	0.504	0.840
2.80	0.803	0.808
3.24	0.902	0.762
3.92	0.975	0.680

Take Home Messages

- Effect of adding more analyses
  - Greater loss of power if maximal sample size not increased
  - Greater increase in maximal sample size if power maintained
  - But, improvement in average efficiency

- Can also use this example for guidance in how to judge thresholds for conditional and predictive power
  - The same threshold should not be used at all analyses
  - It is not, however, clear what threshold should be used
    - I look at tradeoffs between average efficiency and power
  - We can look at optimal (on average) designs for more guidance

Criteria for Futility

- Best: Consider inference, efficiency, statistical power
  - Convincingly ruled out minimal clinically important difference (MCID), AND
  - Use of futility rule provides appropriate tradeoff of average efficiency and power within available resources

- Often problematic: Stochastic curtailment **(used in 301, 302)**
  - Estimated probability of statistical significance based on some hypothesis (prior) of treatment effect in future data
  - Poorly understood thresholds in presence of imprecise estimates
  - Issue in 301 and 302 with estimation of future treatment effect

- How to treat futility when two identically designed clinical trials
  - Usually only terminate when both trials meet criteria

Implementation of Futility Rules

- Must explicitly consider time-varying treatment effects
  - Calendar time: patterns of patient accrual
  - Study time: Delayed efficacy, tachyphylaxis
  - (If treatment works, Amendment 4 introduced time varying aspect to RCT, but Biogen did not modify futility rule)

- Must explicitly consider partial follow-up
  - In absence of varying patient characteristics over calendar time, incomplete data due to administrative censoring is MAR
  - Best to include partial data
  - But at least consider “over-run” before accepting DMC decision

- Must explicitly consider eventual inference
  - Sequential sampling changes sampling distribution
  - How will over-running data be included
  - Binding vs nonbinding futility rules

Delayed Measurement of Outcome

- Longitudinal studies
  - Measurement might be 6 months – 2 years after randomization
  - Interim analyses on variable lengths of follow-up
    - Use of partial data can improve efficiency (Kittelson, et al.)
- Time to event studies
  - Statistical information proportional to number of events
  - Calendar time requirements depend on number accrued and length of follow-up
- In either case: Interim analyses may occur after accrual completed
    - Group ethics of identifying beneficial treatments faster
    - Savings in calendar time costs, rather than per patient costs

Phase 3 Studies 301, 302

- Formal futility interim analysis based on conditional power
  - Computed under current estimate of treatment effect as computed from the combined data from both studies
  - Criteria poorly specified
    - Complete case analysis vs use partial data?
    - Threshold – which was it?
      - "less than 20% probability of both studies reaching statistically significant differences at the final analysis"
      - "not futile unless both studies had less than 20% probability of reaching statistically significant differences in doses"

- Note that a futility analysis I would use would have required conditional power < 0.015 in both studies at 50% information
  - I would have explicitly included partial data in analysis
  - (I would have examined "overrun" before acting)

O'Brien-Fleming & Futility: J = 4

- Introduce four evenly spaced analyses
  - Type 1 error of 0.025
- Stopping boundary table
  - (cf: Z statistic threshold of 1.96 in fixed sample test)

Info Frac	Futility				Efficacy			
	Z	CP_{alt}	CP_{est}	PP_{flat}	Z	CP_{null}	CP_{est}	PP_{flat}
0.25	-1.108	0.719	0.000	0.008	3.976	0.500	0.999	0.999
0.50	0.321	0.648	0.015	0.063	2.811	0.500	0.997	0.977
0.75	1.258	0.592	0.142	0.177	2.295	0.500	0.907	0.874
1.00	1.988	--	--	--	1.988	--	--	--

Multiple Endpoints

- Usually require statistical significance in pre-specified statistical analysis of primary clinical endpoint in two confirmatory RCTs

- Supporting analyses on
  - Additional doses
  - Alternative summary measures
  - Alternative measures of same (latent) clinical endpoint
  - Important subgroups

- Supplementary analyses on other possible benefits

- Type 1 error control through testing hierarchy
  - Require statistical significance at each level to proceed to next

Testing Hierarchy in 301, 302

- According to FDA statistical reviewer at time of AdCom, but confirmed in SAP that was eventually posted
- High dose vs placebo on CDR-SB
- Low dose vs placebo on CDR-SB
- High dose vs placebo on MMSE
- High dose vs placebo on ADAS- Cog13
- High dose vs placebo on ADAS-ADL-MCI
- Etc.

SAP re Hierarchy

Considerations for multiple comparison adjustments

A sequential (closed) testing procedure will be used to control the overall Type I error rate due to multiple comparisons for the primary endpoint. The order of treatment comparisons is as follows: aducanumab high-dose versus placebo and aducanumab low-dose versus placebo. All comparisons after the initial comparison with $p > 0.05$ will not be considered statistically significant.

Secondary endpoints have been rank prioritized, in the order shown in Section 1.2. In order to control for a Type I error for the secondary endpoints, a sequential closed testing procedure will be used and will include both the order of the secondary endpoints and treatment comparisons. Specifically, for each of the secondary endpoints, a sequential (closed) testing procedure, as for the primary endpoint, will be used to control the overall Type I error rate due to multiple treatment comparisons. If statistical significance is not achieved for 1 or 2 treatment comparisons, all endpoint(s) of a lower rank will not be considered statistically significant for that 1or 2 treatment comparisons, respectively.

There will be no multiple comparison adjustments for the sensitivity and supplementary analyses for the primary and secondary efficacy endpoints, the tertiary efficacy endpoints, the subgroup analyses or the additional analyses.

Case Study: Aducanumab

Conduct of Study: Protocol Amendments

Where am I going?

During the course of the phase 3 studies, additional analyses of safety data suggested higher dose levels could be managed in the key subgroup

However, in the protocol amendment, insufficient attention was paid to how changes in the dosing might affect clinical relevance of the pre-specified analyses

Protocol Amendments

- During the course of the study, additional data became available from Study 103

- Protocol Amendment 3 eased the conditions under which ARIA might lead to permanent discontinuation of the drug

- Protocol Amendment 4 modified the “high” dose to be taken by ApoE4 carriers
  - Increased to 10 mg/kg as in the noncarriers
  - Affected patients previously accrued, as well as new randomizations

- In *post hoc* analyses, Sponsor noted that Study 301 had more subjects whose treatment had completed prior to Amendment 4

Impact of Protocol Amendments

- Under the null hypothesis of no treatment effect:
  - Increasing the dose and/or duration of higher dose would have no impact

- Under the Sponsor's hypothesis of a beneficial treatment effect with a positive dose response
  - We would anticipate that the protocol amendment introduced a time varying treatment effect based on calendar time
  - Analysis of the interim data at 50% trial completion would not be expected to be representative of the future

- The Sponsor should have modified the futility rule
  - My recommendation would have been to discard the futility analysis entirely

Case Study: Aducanumab

Conduct of Study: Interim Futility Analysis

Where am I going?

At the pre-specified interim assessment of futility, the DMC reported that the trial results suggested futility according to the pre-specified nonbinding boundary

It is not clear the degree to which the Sponsor examined the available data prior to accepting the DMC recommendation

Interim Analysis

- In December 2018, trigger for futility analysis was reached
  - 50% of subjects had completed 78 weeks post randomization
  - (About 15 – 20% of subjects would have had some 26 or 52 week data available)

- Analysis of interim data showed results that met futility criteria
  - Careful inspection of the two trials by the DMC would have shown some suggestion that 301 had less promising results than 302

- Sponsor accepted DMC recommendation re futility in Mar 2019

Interim Analysis

- Not clear that Sponsor examined the results in any greater detail

- In March 2019 about 15-20% additional statistical information would have been available, but not included in interim analysis
  - Futility boundary was nonbinding, so could have been ignored

- **Intent to cheat analysis**: Could sponsor have looked at the data and strategized to go to FDA with current data?
  - Study 301 was destined to be negative
  - Study 302 currently looked positive, so did not want to risk regression to the mean causing it to be less dramatic

Case Study: Aducanumab

Regulatory Issues: Analyses of Phase 3 Data

Where am I going?

Early stopping for futility does not preclude a careful analysis of the results, and such an analysis could suggest efficacy.

Results for the two phase 3 RCTs were discordant.

The Sponsor and FDA collaborated on presentations of

- inappropriate analyses of the “best of two” RCT
- Inappropriate conditioning on post-randomization variables

No member of the PCNS Advisory Committee voted in favor of approval of the drug

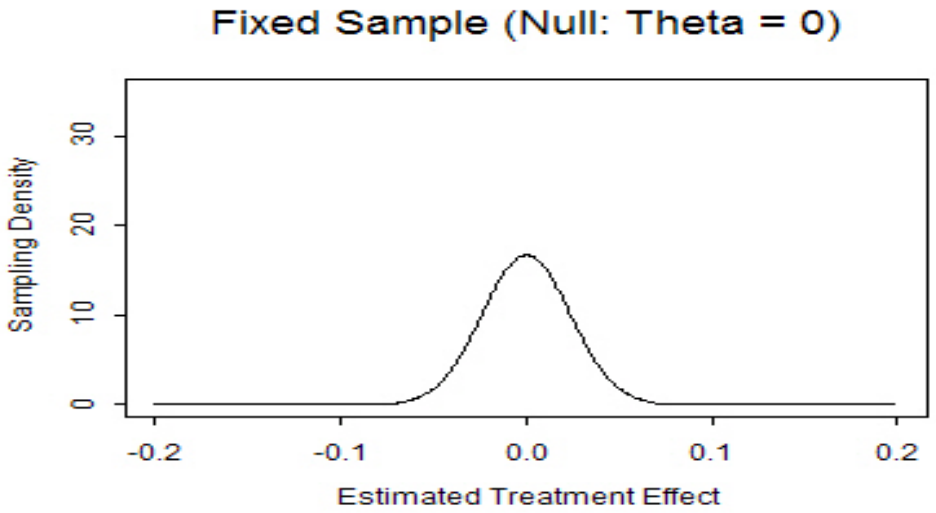
Impact of Stopping for Futility

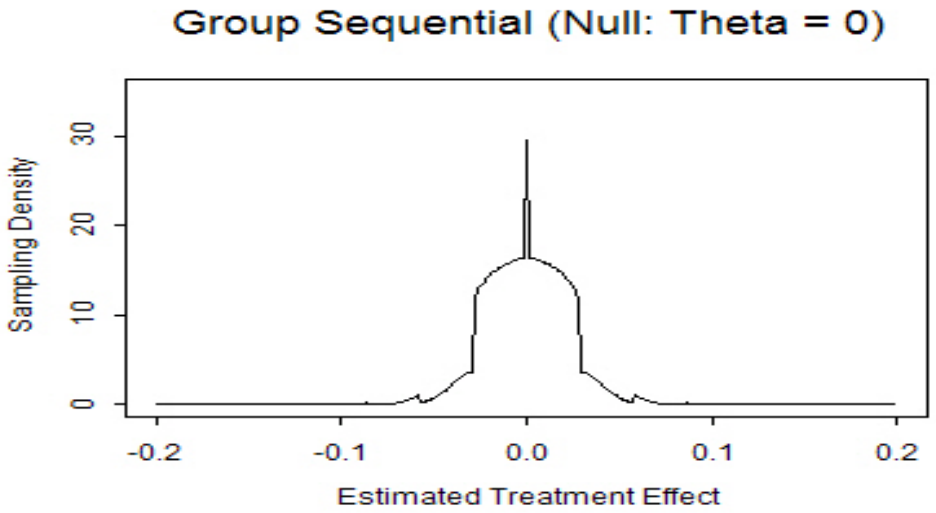
- Scientifically:
  - Changes timeframe that efficacy / safety observed

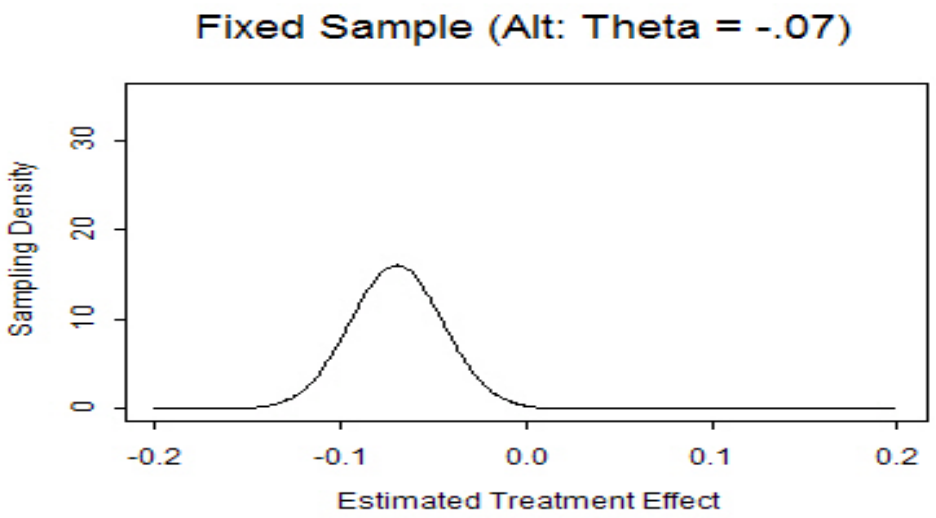
- Statistically:
  - Reduces precision of estimates relative to prior plan
    - After incorporating “overrun” had only about 75% of maximal statistical information
  - Changes sampling distribution
    - But in this case, no impact analyses pre-specified in SAP
  - Does NOT change how to interpret point estimates, confidence intervals, p values
    - *Providing* the method to maintain relevance of clinical endpoint is prespecified

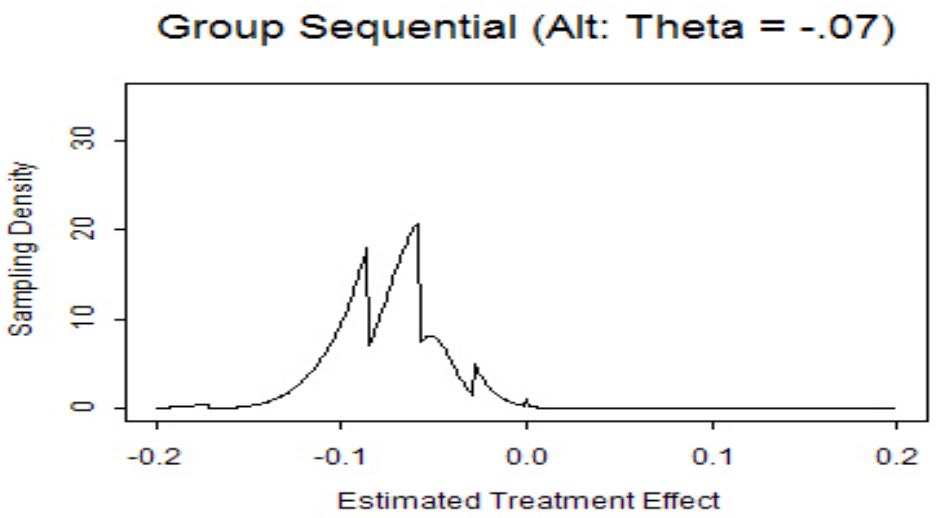
Sampling Distribution of MLE

- Depends on exact sampling rule and schedule of analyses









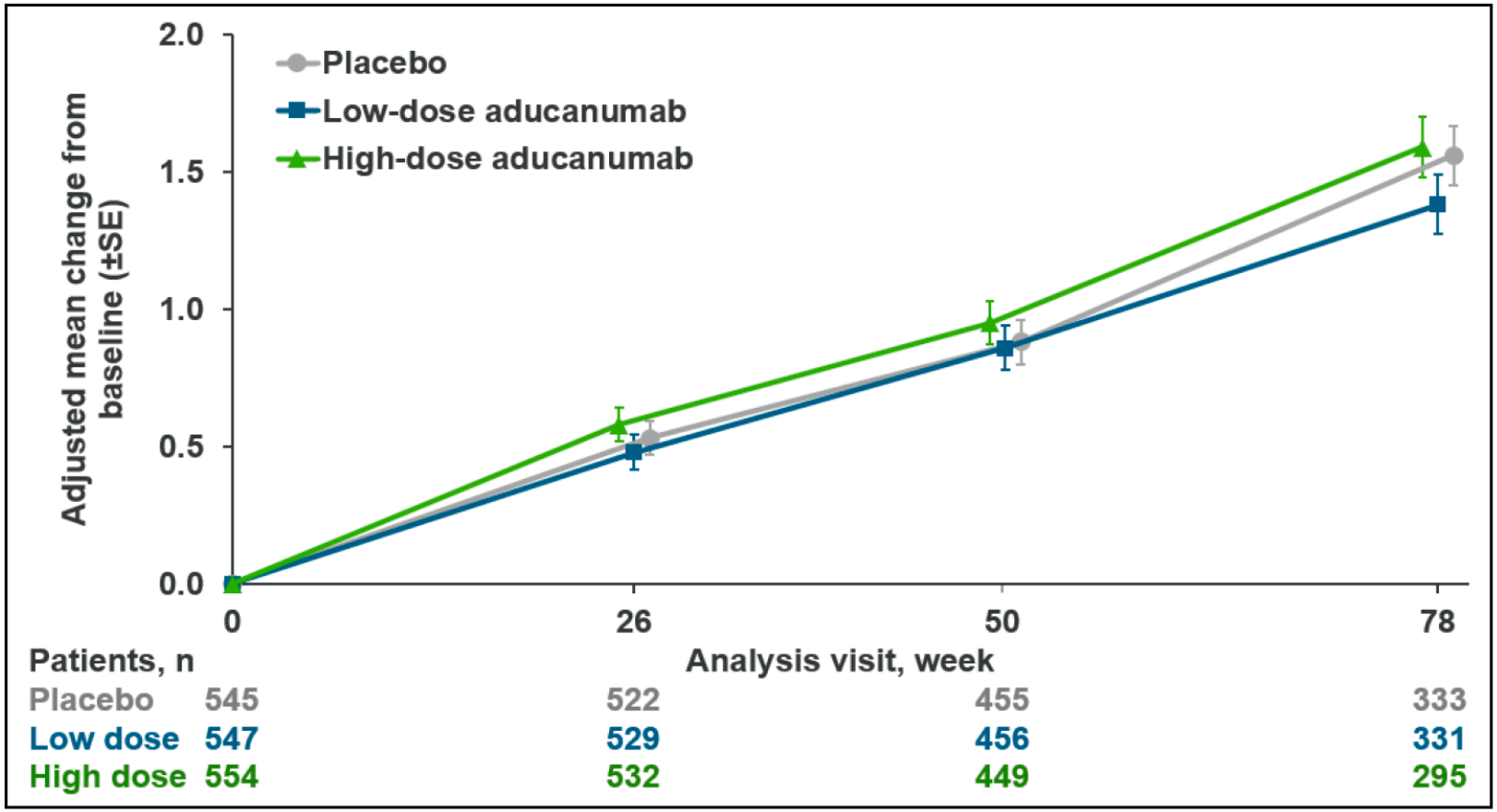
Study Efficacy Results

- Scientifically rigorous, statistically valid analysis of mean progression of CDR-SB
  - 301: high dose 2% more (p=0.83)
  - 302: high dose 22% less (p = 0.01), low dose 15% less (p = 0.09)

- Descriptive and exploratory (not all in AdCom documents)
  - Impact of missing data
  - Clinical importance: responder analysis, more severe endpoints
  - Dose response (ITT, per protocol)
  - Alternative measures of cognition
  - Mechanism of action
    - Intended: Correlation with changes in amyloid
    - Unintended: Change physician behavior (unblind, medications)
  - Consistency across subgroups

301: Primary Endpoint Over Time

Figure 13: Change From Baseline on the CDR-SB Over Time, Study 301



Abbreviations: SE = standard error.

Note: Results were based on an MMRM model, with change from baseline in CDR-SB as dependent variable and with fixed effects of treatment group, categorical visit, treatment-by-visit interaction, baseline CDR-SB, baseline CDR-SB by visit interaction, baseline MMSE, Alzheimer's disease symptomatic medication use at baseline, region, and laboratory ApoE ε4 status.

Data source: 221AD301/CSR/T-CDR-MMRM- PC

301: Primary, Secondary Endpoints

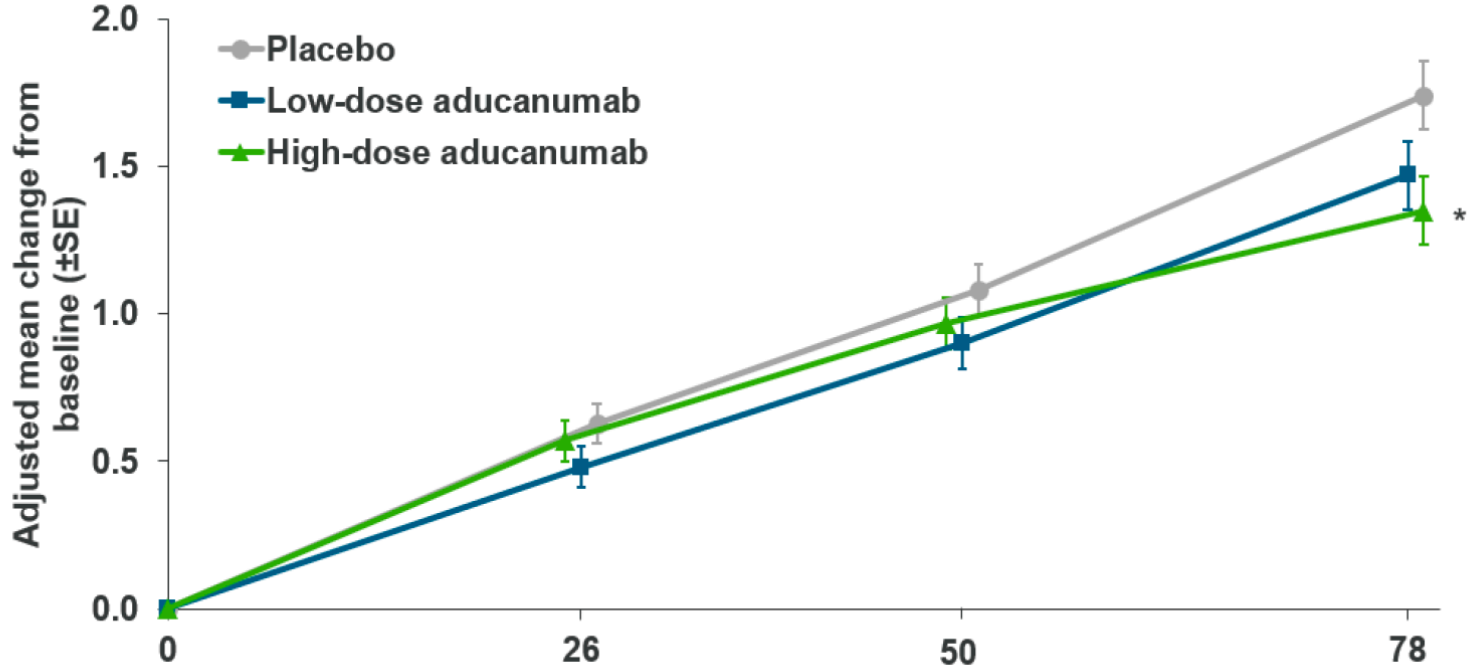
Table 10: Primary and Secondary Endpoints at Week 78 in Study 301: ITT Population

		Diff vs PBO [a] (%) p-value	
	PBO decline (N=545)	**Low dose (N=547)**	**High dose (N=555)**
CDR-SB	n=333 1.56	n=331 -0.18 (-12%) 0.2250	n=295 0.03 (2%) 0.8330
MMSE	n=332 -3.5	n=334 0.2 (-6%) 0.4795	n=297 -0.1 (3%) 0.8106
ADAS-Cog13	n=331 5.140	n=332 -0.583 (-11%) 0.2536	n=294 -0.588 (-11%) 0.2578
ADCS-ADL-MCI	n=331 -3.8	n=330 0.7 (-18%) 0.1225	n=298 0.7 (-18%) 0.1506

Abbreviations: ADAS-Cog13 = Alzheimer's Disease Assessment Scale - Cognitive Subscale (13 items); ADCS-ADL-MCI = Alzheimer's Disease Cooperative Study - Activities of Daily Living Inventory (Mild Cognitive Impairment version); CDR-SB = Clinical Dementia Rating – Sum of Boxes; MMRM = mixed model for repeated measures; MMSE = Mini-Mental State Examination; PBO = placebo

302: Primary Endpoint Over Time

Figure 5: Change From Baseline on the CDR-SB Over Time in Study 302



Patients, n				
Placebo	548	532	430	288
Low dose	543	513	420	290
High dose	547	513	431	299

* $p < 0.05$

Note: Results were based on a mixed model for repeated measures, with change from baseline in CDR-SB as dependent variable and with fixed effects of treatment group, categorical visit, treatment-by-visit interaction, baseline CDR-SB, baseline CDR-SB by visit interaction, baseline MMSE, Alzheimer's disease symptomatic medication use at baseline, region, and laboratory ApoE ε4 status.

Data source: 221AD302/CSR/T-CDR-MMRM- PC

302: Primary, Secondary Endpoints

Table 7: Primary and Secondary Endpoints at Week 78 in Study 302: ITT Population

		Difference vs PBO [a] (%) p-value	
	PBO decline (N=548)	**Low dose (N=543)**	**High dose (N=547)**
CDR-SB	n=288 1.74	n=290 -0.26 (-15%) 0.0901	n=299 -0.39 (-22%) 0.0120
MMSE	n=288 -3.3	n=293 -0.1 (3%) 0.7578	n=299 0.6 (-18%) 0.0493
ADAS-Cog13	n=287 5.162	n=289 -0.701 (-14%) 0.1962	n=293 -1.400 (-27%) 0.0097
ADCS-ADL-MCI	n=283 -4.3	n=286 0.7 (-16%) 0.1515	n=295 1.7 (-40%) 0.0006

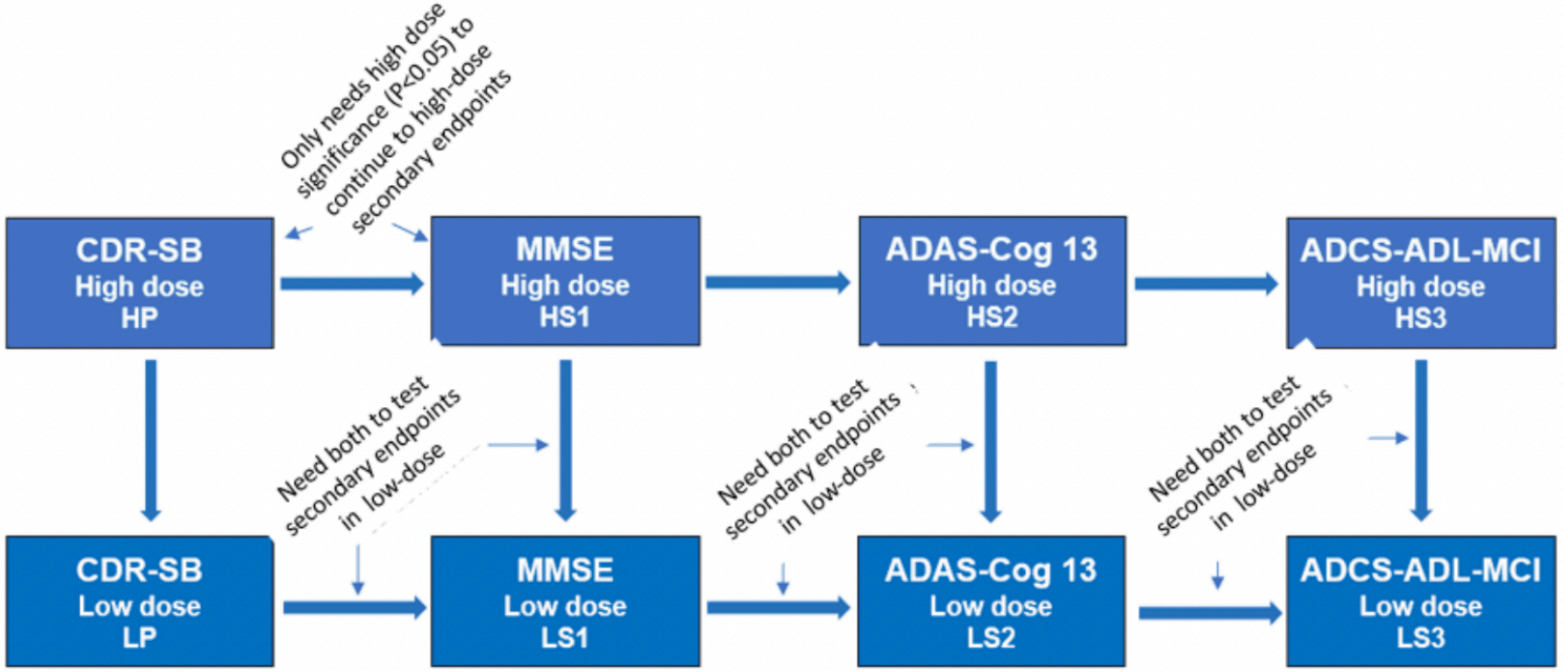
Abbreviations: N = numbers of all randomized and dosed participants that were included in the analysis; n: numbers of randomized and dosed subjects with endpoint assessment at Week 78.
PBO = placebo.

Post hoc Hierarchy

- Budd Haberlein letter to Editor, MedPage Today:

The clinical principle underpinning the testing strategy was that high-dose (10 mg/kg) was the target dose. Therefore, failure of low-dose on any endpoint should not preclude testing of the high-dose.

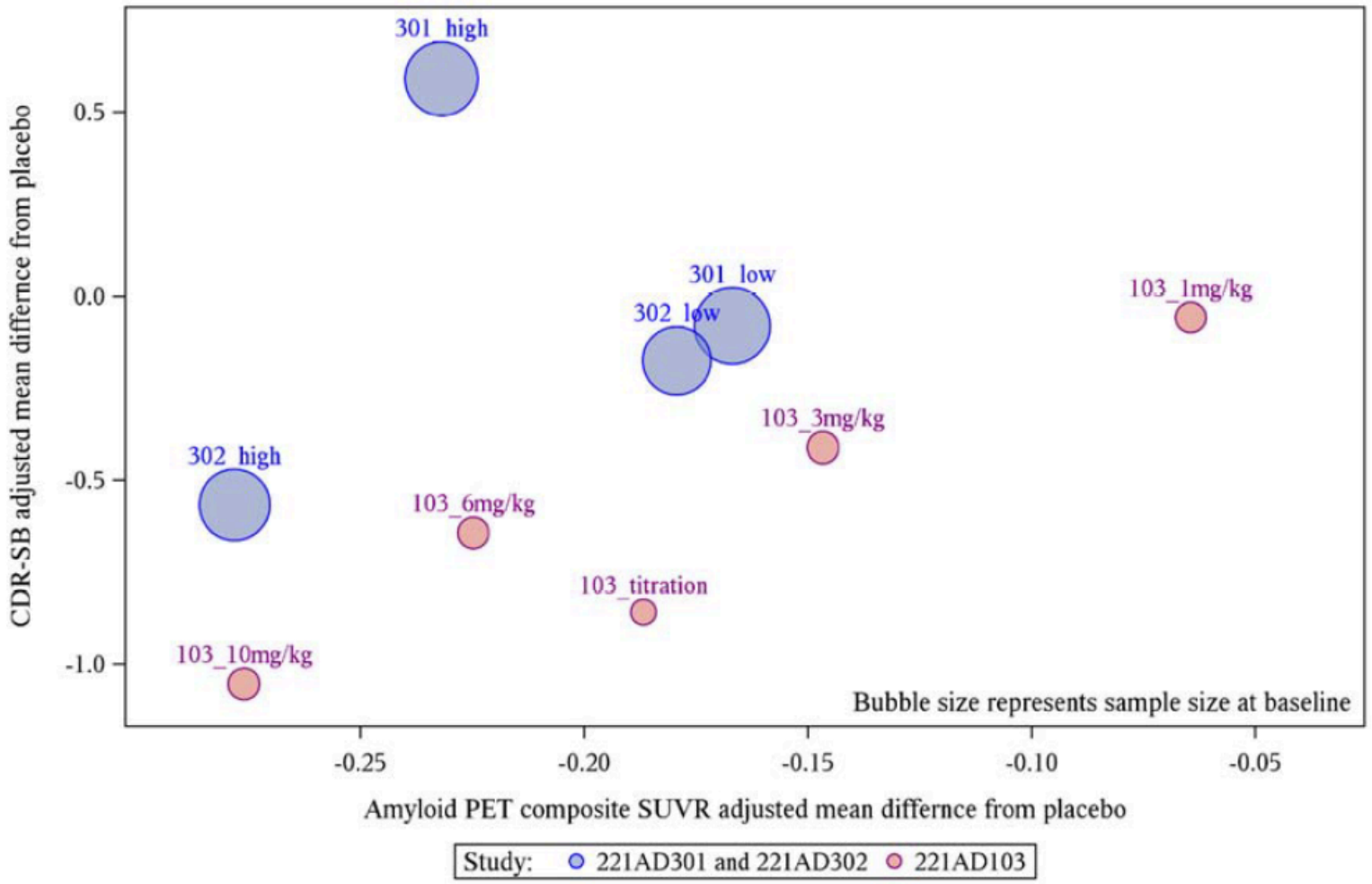
Figure 1



As currently written, the article assumes an endpoint-by-endpoint review, first reviewing the high-dose in CDR-SB (for example) and then the low-dose in CDR-SB, and then followed by secondary endpoints in the high dose. This is simply not the case based on the prespecified SAP.

Correlation CDR-SB and Amyloid

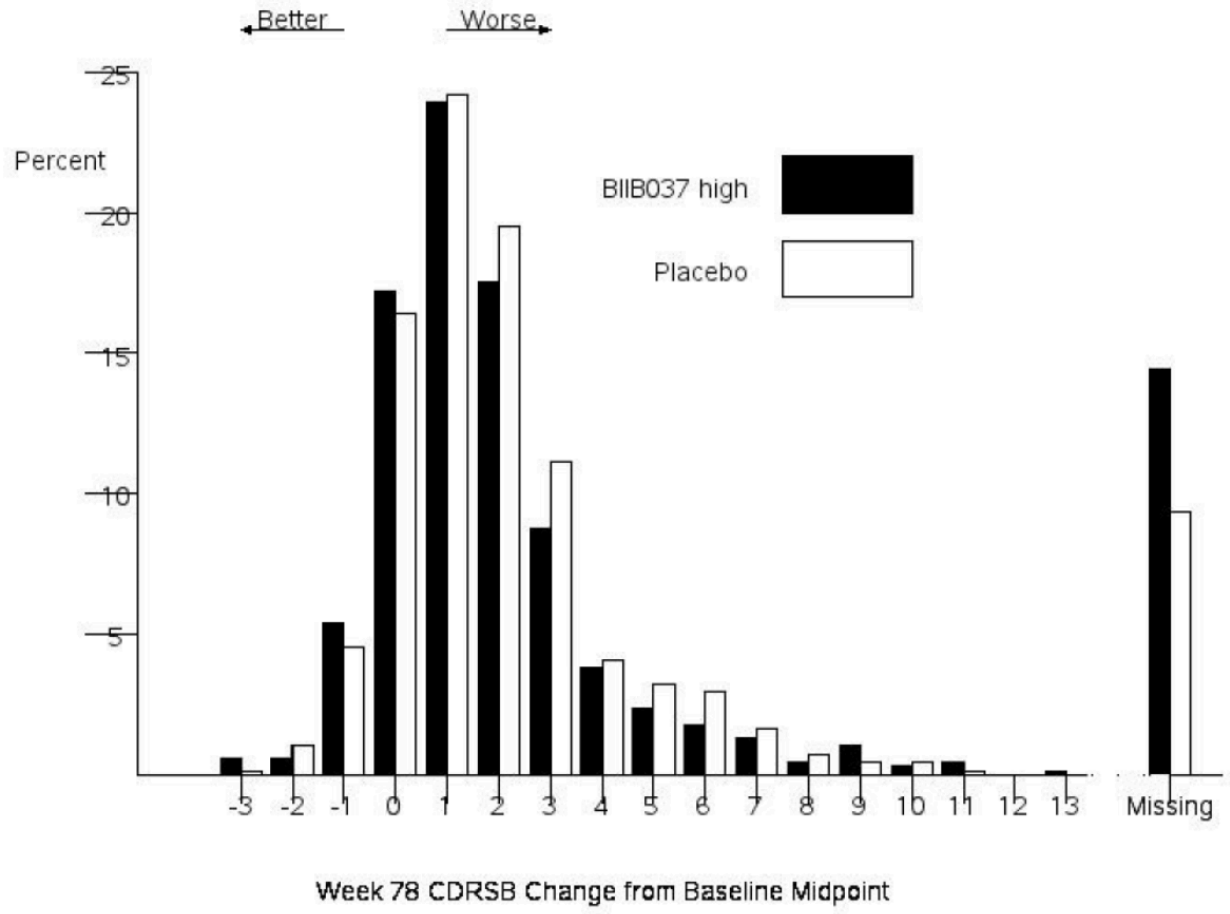
Figure 12: Group-Level Correlation Between Adjusted Mean Difference from Placebo in Aβ PET Composite SUVR and CDR-SB



Source: Figure 61 in ISE

Pseudo-Responder Analysis

Figure 2. CDRSB changes at Week 78 in those with opportunity to complete (studies 301 and 302 pooled)



Study Safety Results

- Adverse events of special interest (AESI)
  - ARIA higher in both studies in both dose groups
  - Only slight dose response relationship
  - Two-thirds asymptomatic, one-third radiographically severe

- Other adverse events
  - Low signal of more falls (~ 3%)
  - No other major concerns

Differential Diagnosis of Discordance

- Under null hypothesis: Consistency with type 1 error
  - 26 prior negative studies for agents targeting amyloid
  - Lack of dose response in 301: High dose worse than low dose

- Presuming effective treatment
  - 301 from a different population
  - Unlucky randomization of “rapid progressors”
  - Fewer subjects exposed to highest dose post Amendment 4
    - Due to date of randomization
    - Due to study treatment discontinuation

Collaboration with FDA

- Alternative pathway using a single pivotal study
  - Guidance not specific on criteria
  - Historically, $p < 0.01$ on primary endpoint
  - Must have strong supportive evidence
    - Secondary endpoints
    - Other trials

- FDA supported exploration of treating 302 as primary, 103 supportive

Statistical Issues

- Treating 302 as independent trial
  - Sampling distribution is different for a sole trial than for the best of two trials
  - Many issues with inflation of type 1 error
    - At a minimum, the type 1 error for 302 is now 0.02

- Post hoc nature of focusing on opportunity for higher dose
  - Lack of dose response within 301
  - Still does not explain lack of dose response in 301

- Deletion of subjects based on “rapid progressors” totally without rigor

- Study 103 was not analyzed by randomization, hence cannot be “adequate and well controlled”

Why Confirmation: Real-life Examples

- Effects of arrhythmias post MI on survival
  - Observational studies: high risk for death
  - CAST: Specific anti-arrhythmics have higher mortality

- Effects of beta-carotene on lung CA and survival
  - Observational studies: high dietary beta carotene has lower cancer incidence and longer survival
  - CARET: beta carotene supplementation in smokers leads to higher lung CA incidence and lower survival

- Effects of hormone therapy on cardiac events
  - Observational studies: HT has lower cardiac morbidity and mortality
  - WHI: HT in post menopausal women leads to higher cardiac mortality

My Suggestion for Further Study

- For scientific rigor, need a confirmatory study
- Biogen urged trial participants to testify that they had seemed to benefit
- Equipoise can be addressed with a randomized withdrawal design
  - All participants get drug initially
  - Tolerators, compliers, responders are then randomized to withdrawal to compare progression
  - Benefit of longer term safety among subjects most likely to take drug
  - Benefit of examining tachyphylaxis

Case Study: Aducanumab

Regulatory Issues: Accelerated Approval

Where am I going?

Leadership of the FDA ultimately relied on Office of Clinical Pharmacology opinions, rather than the assessment of the Office of Biostatistics

Accelerated approval was granted on the belief that amyloid reduction was a surrogate endpoint likely to predict improvements in cognitive function

These decisions run counter to the literature on the use of surrogate endpoints in general, and the evidence on amyloid reduction in particular

Accelerated Approval

- During Advisory Committee Meeting Director of Neuroscience:

  "This is Dr. Dunn. I can speak to the second one. We're not using the amyloid as a surrogate for efficacy."

- Meetings of FDA CDER personnel reviewing situation
  - Office of Biostatistics unwavering against approval or accelerated approval
  - Clinical reviewer supported approval
  - Office of Clinical Pharmacology supported full approval

- Ultimate accelerated approval apparently relied on OCP
  - Dunn: "*There is substantial evidence that aducanumab reduces amyloid beta plaques, and this reduction is reasonably likely to result in clinical benefit for patients.*"

Dunn's Decision Memorandum

- "The accelerated approval pathway is intended to provide a path to approval for drugs in certain situations where there is some **uncertainty at the time of approval regarding the drug's ultimate clinical benefit.**

- "Accelerated approval is based on an **outcome that is reasonably likely to predict clinical benefit**, rather than on the clinical benefit itself.

- "These outcomes predictive of benefit are **generally surrogate markers** of disease of some sort, but may also be an intermediate clinical endpoint that can be measured earlier than the outcome of ultimate clinical importance.

- "**Substantial evidence of effectiveness is required on such an endpoint to support accelerated approval**, just as it is required for an endpoint supporting standard approval.

- "Accelerated approval (AA) is **intended for serious conditions** where the drug provides a meaningful advantage over available therapies

(emphasis mine)

Comments re Accelerated Approval

- Institute for Clinical and Economic Review AA 1992-2016f
  - 76.5% converted to full approval, 10.3% were withdrawn, and 13.1% on market a median of 9.5 years without confirmatory RCT

- 2018 paper by FDA (Beaver, et al.) reports on 93 AA
  - 51 later satisfied requirements, 37 pending, 5 withdrawn

- 2019 Paper by Gyawali, et al. considers same 93
  - Of 58 drugs receiving full approval: 19 drugs demonstrated benefit in overall survival, 19 used same surrogate as AA, 20 used different surrogate
  - 5 confirmatory trials delayed, 10 pending, 9 ongoing
  - 3 confirmatory trials “failed”, but 1 drug indication still approved

- 2021 paper by Cherla, et al. considers same 93
  - 12 not recommended in Europe, 30 not reviewed
  - Of the 51 available in UK, 86% have additional restrictions

General Comments re Surrogacy

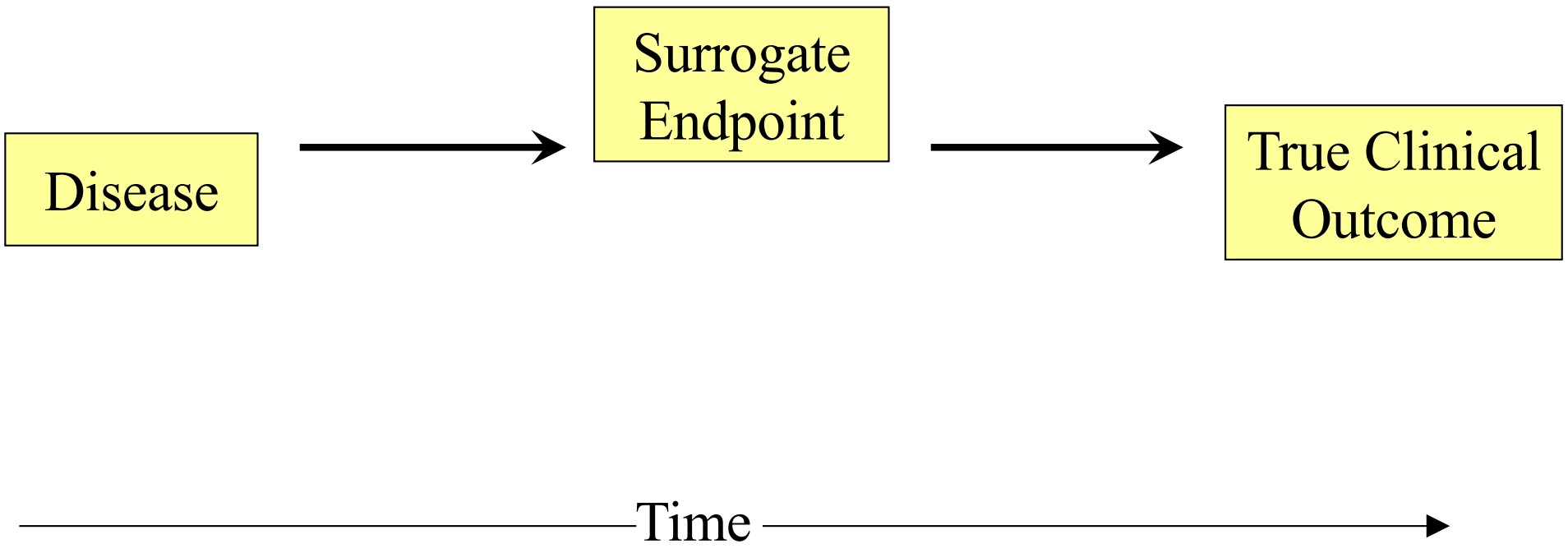
- What is goal?
  - Population benefit of drugs vs Individual benefit
  - All treatments or only treatments within some class of drugs

- Is potential surrogate indicative of a patient receiving benefit?
  - A mediator analysis will demonstrate reduced effect of drug after adjusting for (time-varying?) surrogate.

- Is it just a biomarker of drug effect?
  - Group level correlation analyses across a population of drugs
  - Are we estimating a common relationship or is the drug a random effect?
    - In cancer, we can demonstrate that bone marrow suppression is correlated with clinical response across drugs, though the patients with worst suppression may get no benefit

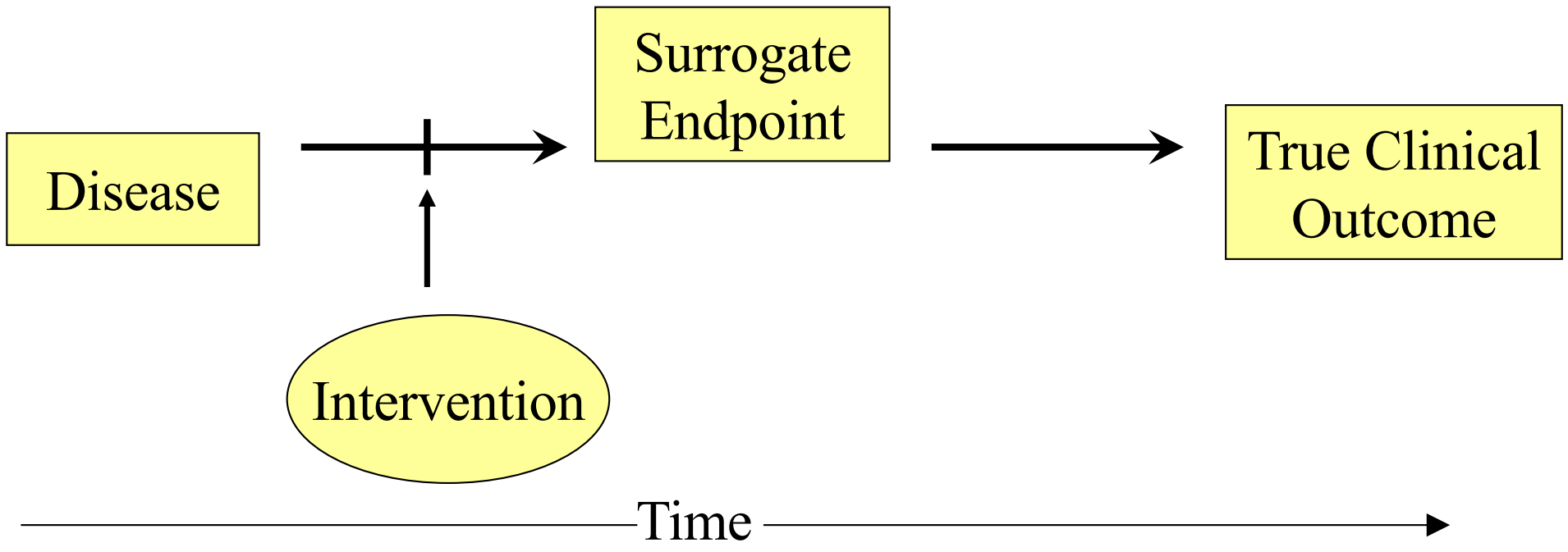
Scenario 1: The Ideal

- Disease progresses to Clinical Outcome only through the Surrogate Endpoint



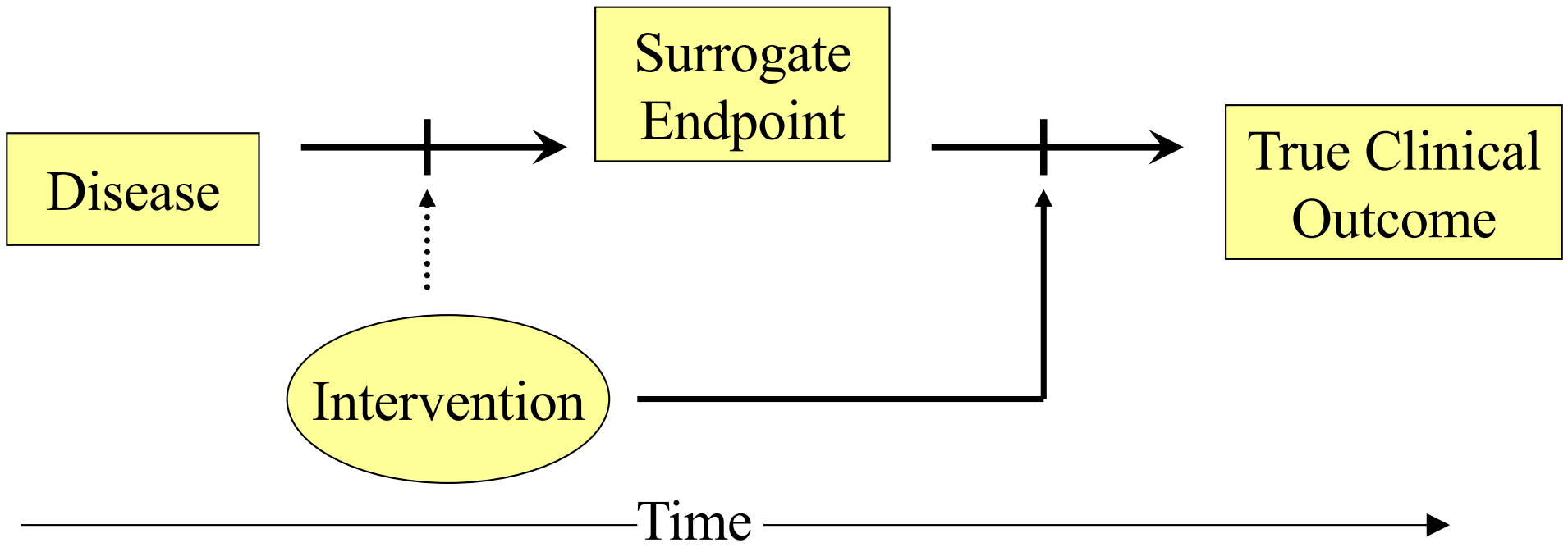
Scenario 1a: Ideal Surrogate Use

- The intervention's effect on the Surrogate Endpoint accurately reflects its effect on the Clinical Outcome



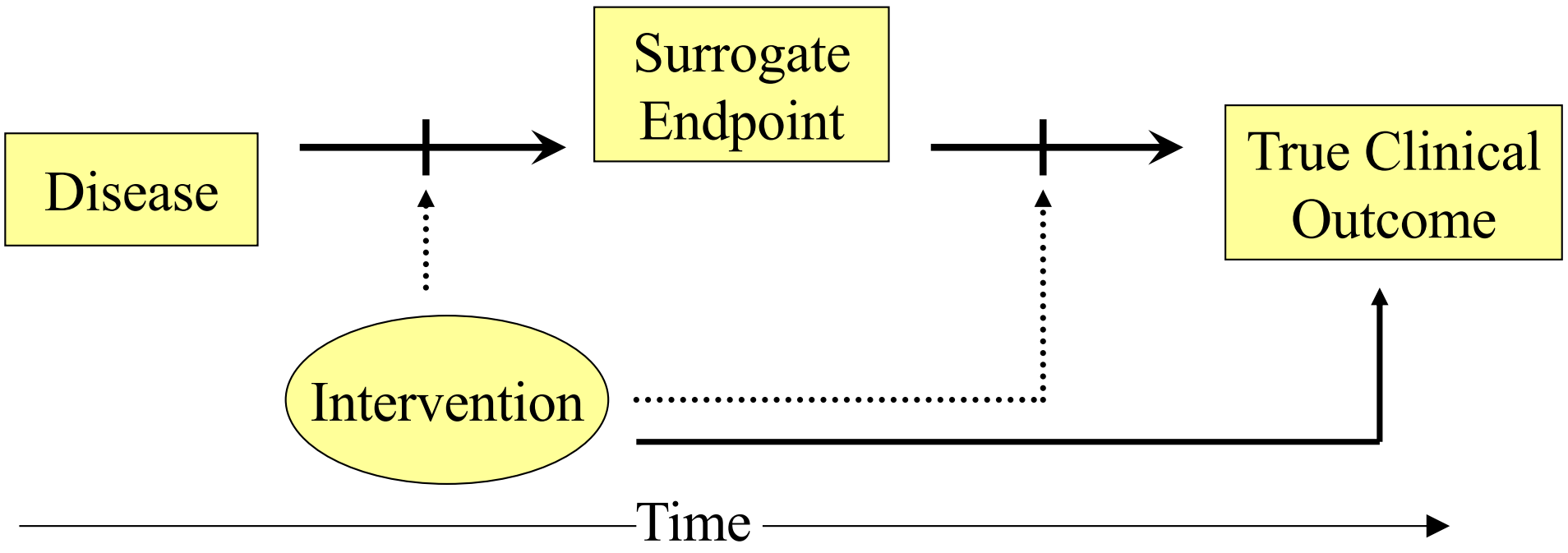
Scenario 1b: Inefficient Surrogate

- The intervention's effect on the Surrogate Endpoint understates its effect on the Clinical Outcome



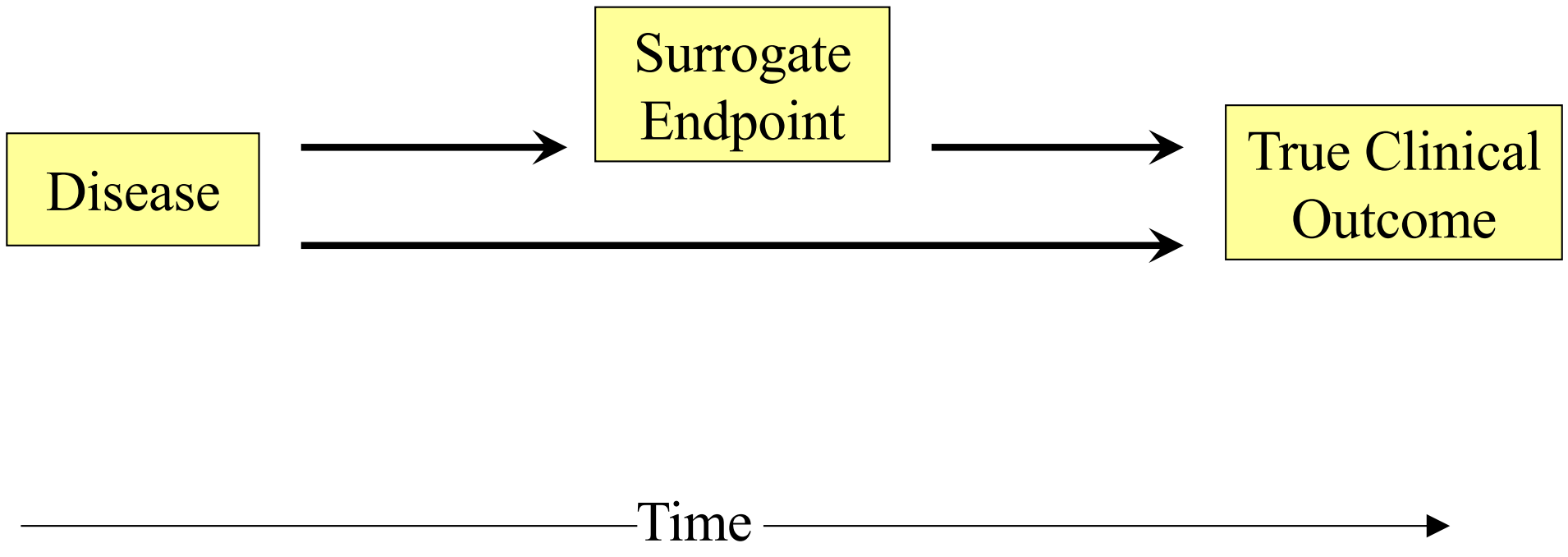
Scenario 1d: Dangerous Surrogate

- Effect on the Surrogate Endpoint may overstate its effect on the Clinical Outcome (which may actually be harmful)



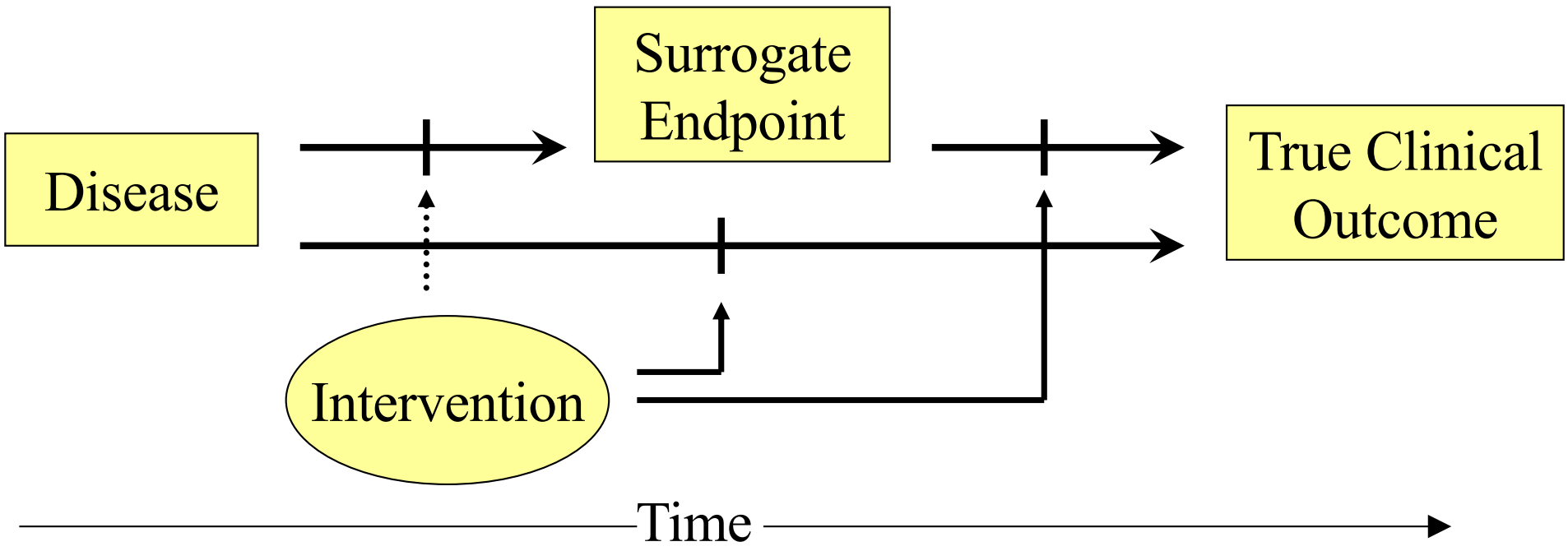
Scenario 2: Alternate Pathways

- Disease progresses directly to Clinical Outcome as well as through Surrogate Endpoint



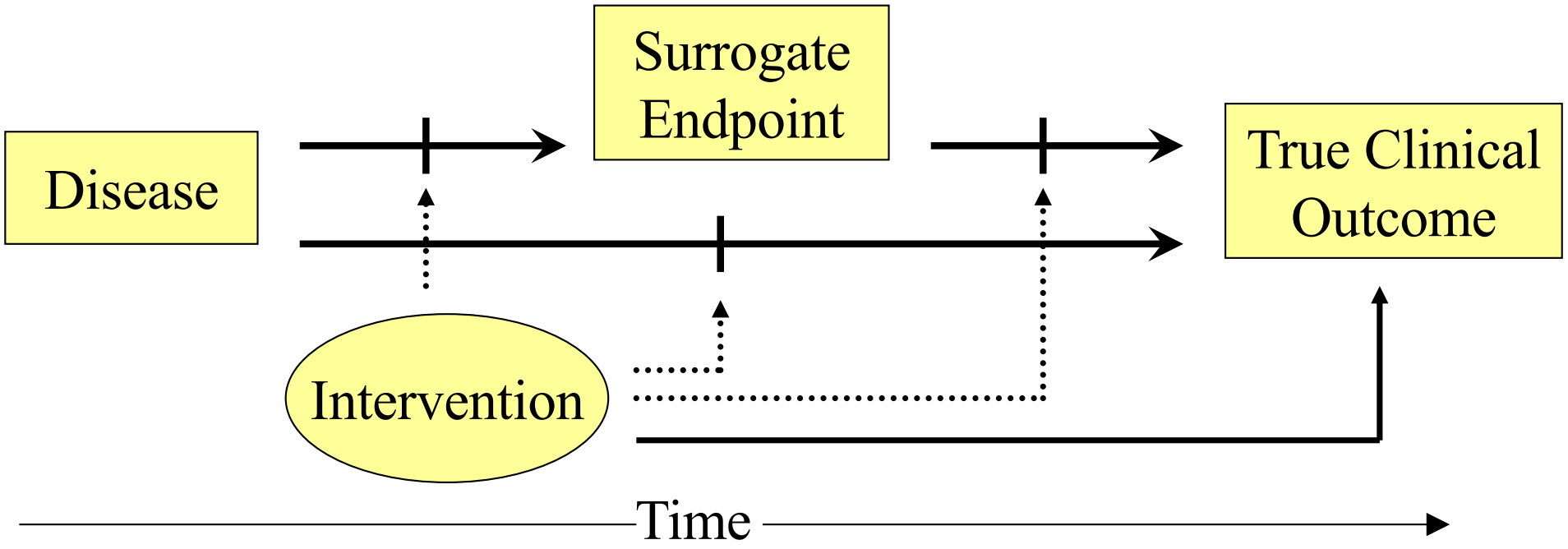
Scenario 2b: Inefficient Surrogate

- Treatments' effect on Clinical Outcome is greater than is reflected by Surrogate Endpoint
  - E.g., Gamma interferon in CGD



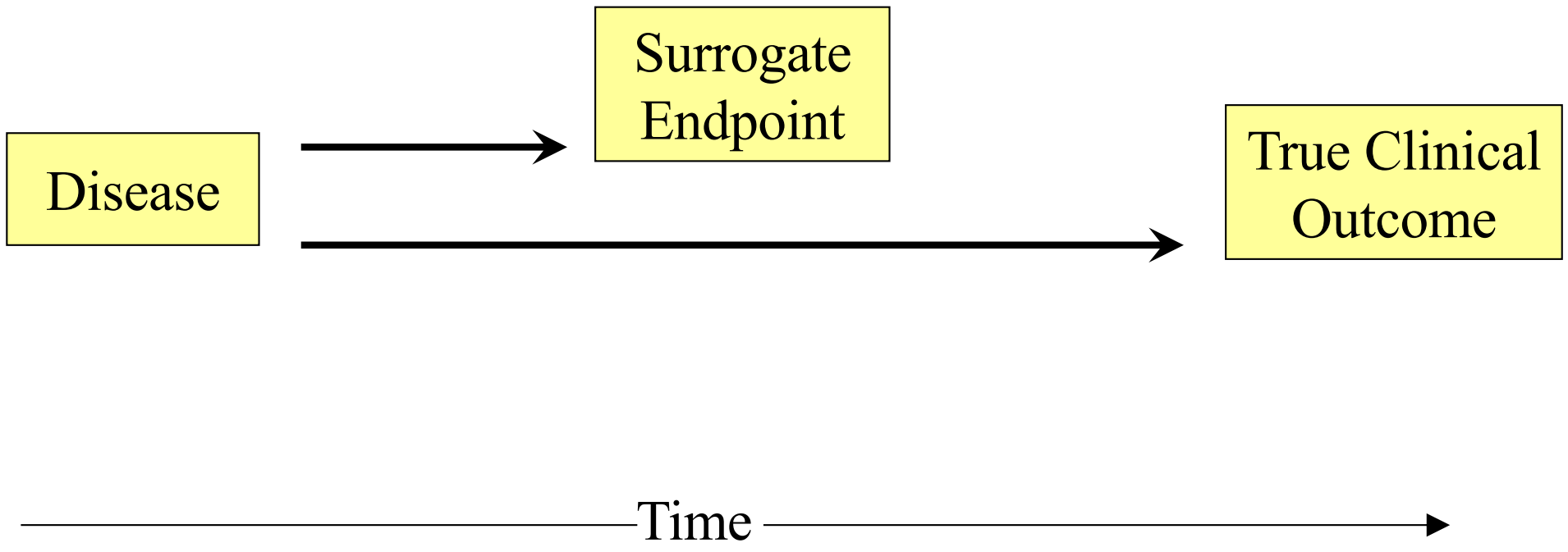
Scenario 2d: Dangerous Surrogate

- The effect on the Surrogate Endpoint may overstate its effect on the Clinical Outcome (which may actually be harmful)



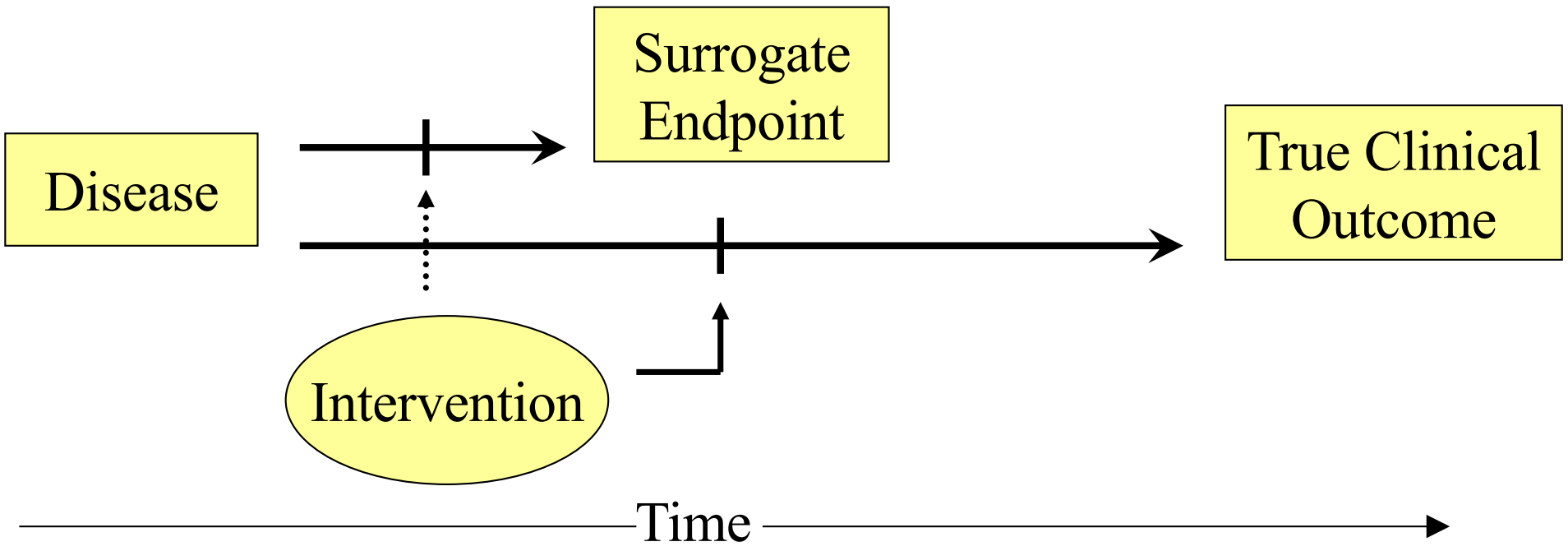
Scenario 3: Marker

- Disease causes Surrogate Endpoint and Clinical Outcome via different mechanisms



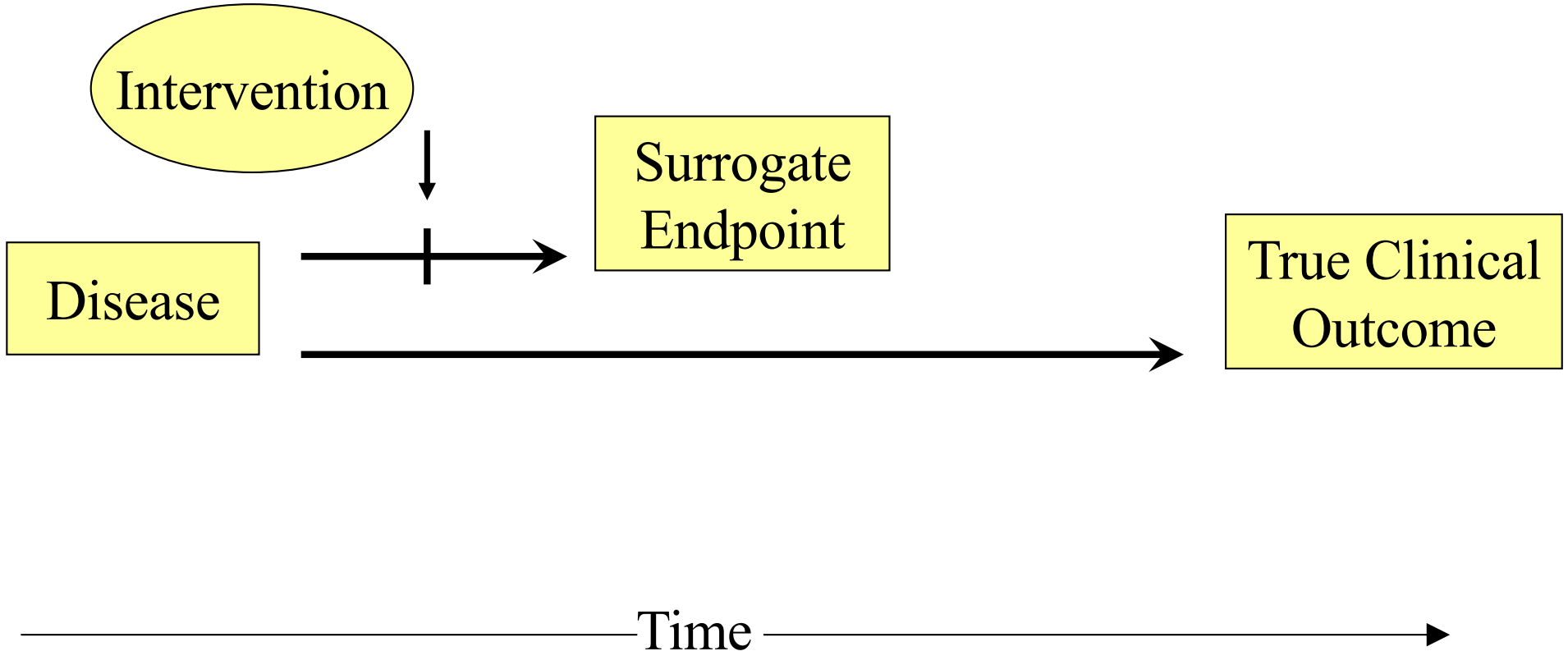
Scenario 3b: Inefficient Marker

- Treatments' effect on Clinical Outcome is greater than is reflected by Surrogate Endpoint



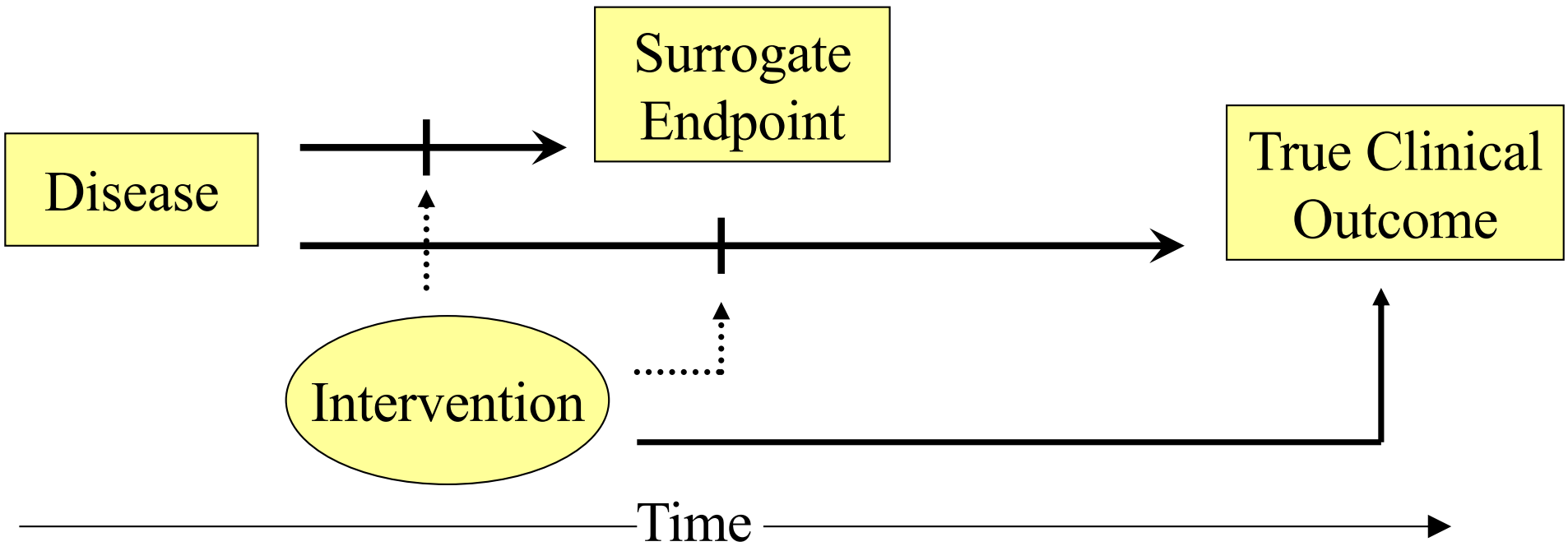
Scenario 3c: Misleading Surrogate

- Effect on Surrogate Endpoint does not reflect lack of effect on Clinical Outcome



Scenario 3d: Dangerous Surrogate

- Effect on the Surrogate Endpoint may overstate its effect on the Clinical Outcome (which may actually be harmful)
  - E.g., anti-arrhythmics?, laromustine in AML?



General Comments re Surrogacy

- Best approach: Meta-analysis across multiple classes of drugs
  - Slopes, not correlation

- But still:
  - Need to include all available studies
  - Need to be clear on measure of clinical endpoint
  - Need to be clear on measure of surrogate
  - Need to try to assess random effects of drugs / classes

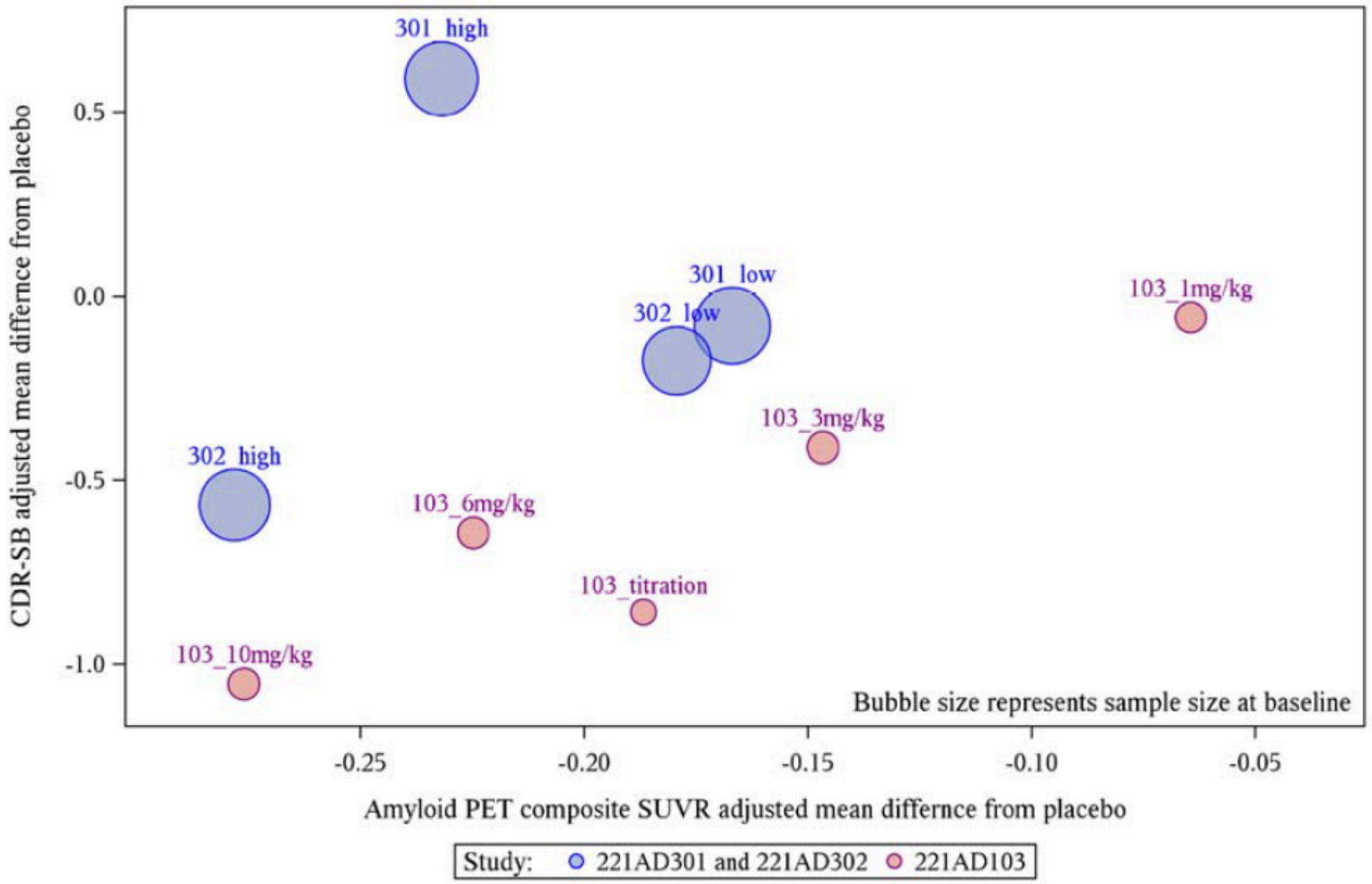
OCP Analysis of Surrogacy

- Based on correlation of group effects
  - Noted that correlation of individual data will be attenuated low, but erroneously attributed that to confounding rather than precision
  - (Also tried to simulate results to dispute a comment I made in AdCom meeting but did not model selection of data)

- Considered
  - Aducanumab studies
  - Aducanumab studies except 301
  - Aducanumab studies excluding “fast progressors”
  - Aducanumab studies plus published studies of monoclonal antibodies
    - Interestingly: Almost 3 years out from RCT termination, the efficacy results of 301 and 302 have not been published in a peer reviewed journal

Correlation CDR-SB and Amyloid

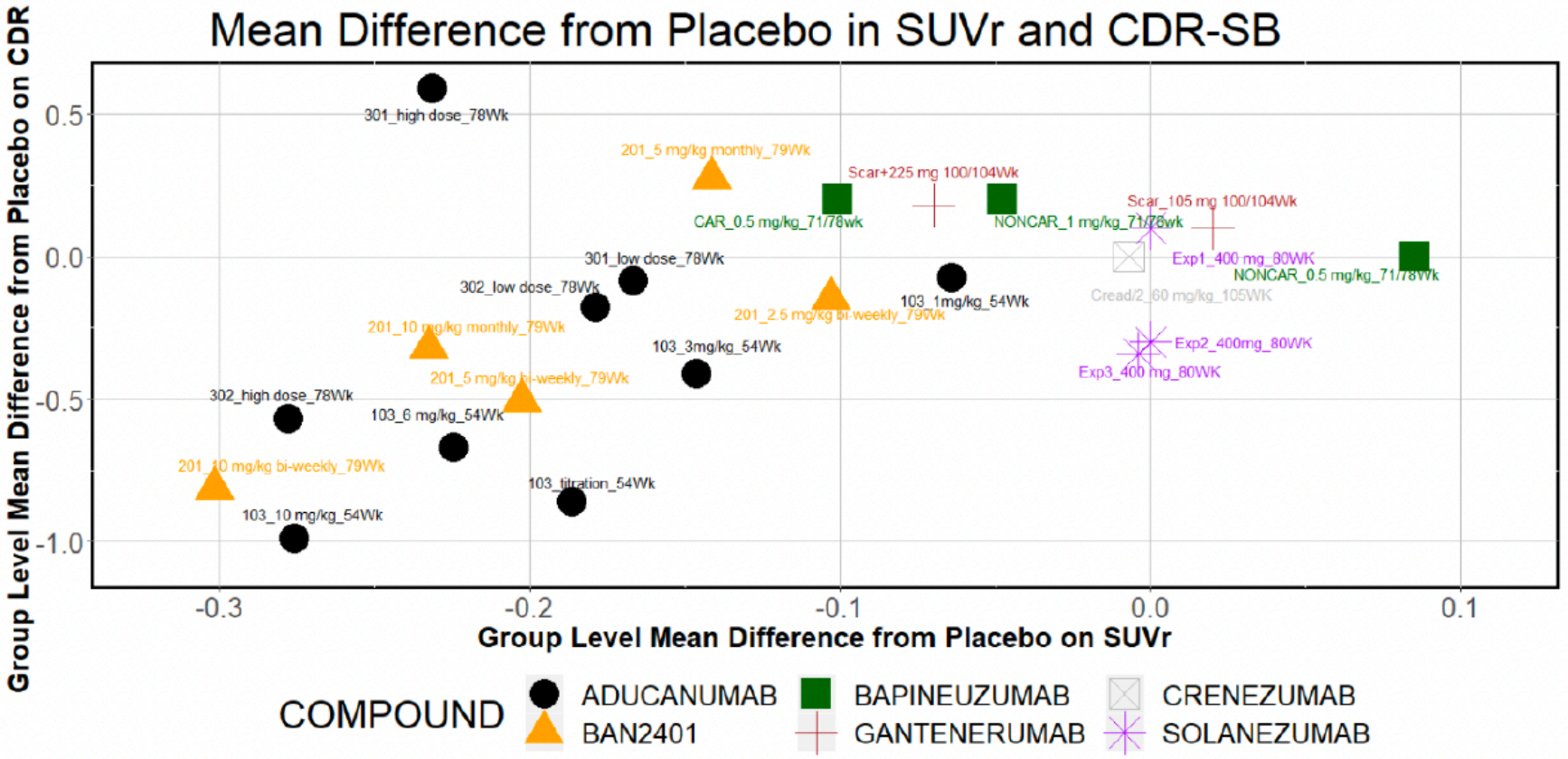
Figure 12: Group-Level Correlation Between Adjusted Mean Difference from Placebo in Aβ PET Composite SUVR and CDR-SB



Source: Figure 61 in ISE

Correlation CDR-SB and Amyloid

Figure 5 Group Level Comparison of Reduction in Aβ Plaque Burden (Described by SUVR (A.) and Centiloid (B.)) and Preserving of Clinical Function across All Available Programs of Anti-Aβ Antibodies under Development. Multiple dose levels for each drug candidate are presented.



Published Meta-analysis 26 Feb 2021

Effect of reductions in amyloid levels on cognitive change in randomized trials: instrumental variable meta-analysis

Sarah F Ackley,[1] Scott C Zimmerman,[1] Willa D Brenowitz,[1,2] Eric J Tchetgen Tchetgen,[3] Audra L Gold,[1] Jennifer J Manly,[4] Elizabeth Rose Mayeda,[5] Teresa J Filshtein,[6] Melinda C Power,[7] Fanny M Elahi,[8] Adam M Brickman,[4] M Maria Glymour[1]

[1]Department of Epidemiology and Biostatistics, University of California, San Francisco, 550 16th Street, San Francisco, CA, USA

[2]Department of Psychiatry, University of California, San Francisco, CA, USA

[3]Department of Statistics, University of Pennsylvania, PA, USA

[4]Taub Institute for Research on Alzheimer's Disease and the Aging Brain, G H Sergievsky Center, Department of Neurology, Columbia University, New York, NY, USA

[5]Department of Epidemiology, University of California, Los Angeles, CA, USA

[6]23&Me, Sunnyvale, CA, USA

[7]Department of Epidemiology, George Washington University, Milken Institute School of Public Health, Washington DC, USA

[8]UCSF Weill Institute for Neurosciences, University of California, San Francisco, CA

ABSTRACT

OBJECTIVE

To evaluate trials of drugs that target amyloid to determine whether reductions in amyloid levels are likely to improve cognition.

DESIGN

Instrumental variable meta-analysis.

SETTING

14 randomized controlled trials of drugs for the prevention or treatment of Alzheimer's disease that targeted an amyloid mechanism, identified from ClinicalTrials.gov.

POPULATION

Adults enrolled in randomized controlled trials of amyloid targeting drugs. Inclusion criteria for trials vary, but typically include adults aged 50 years or older with a diagnosis of mild cognitive impairment or Alzheimer's disease, and amyloid positivity at baseline.

MAIN OUTCOME MEASURES

Analyses included trials for which information could be obtained on both change in brain amyloid levels estimate supporting the benefit of reducing amyloid levels.

CONCLUSIONS

Pooled evidence from available trials reporting both reduction in amyloid levels and change in cognition suggests that amyloid reduction strategies do not substantially improve cognition.

Introduction

Amyloid plaques and oligomers are hypothesized to cause a cascade of pathological events resulting in cognitive decline in Alzheimer's disease.[1-3] Motivated by the amyloid cascade hypothesis, a primary aim of many new treatments for the prevention or management of Alzheimer's disease has been to reduce amyloid β levels in the brain.[4] Although the presence of amyloid plaques and oligomers in the brain is highly correlated with the progression of Alzheimer's disease,[5 6] the mechanisms by which amyloid might mediate neuronal pathology are currently not well understood.[7] To date, no anti-amyloid treatments have progressed sufficiently to receive approval from the Food and Drug Administration (the regulatory agency

Published Meta-analysis 26 Feb 2021

- SUVR vs MMSE

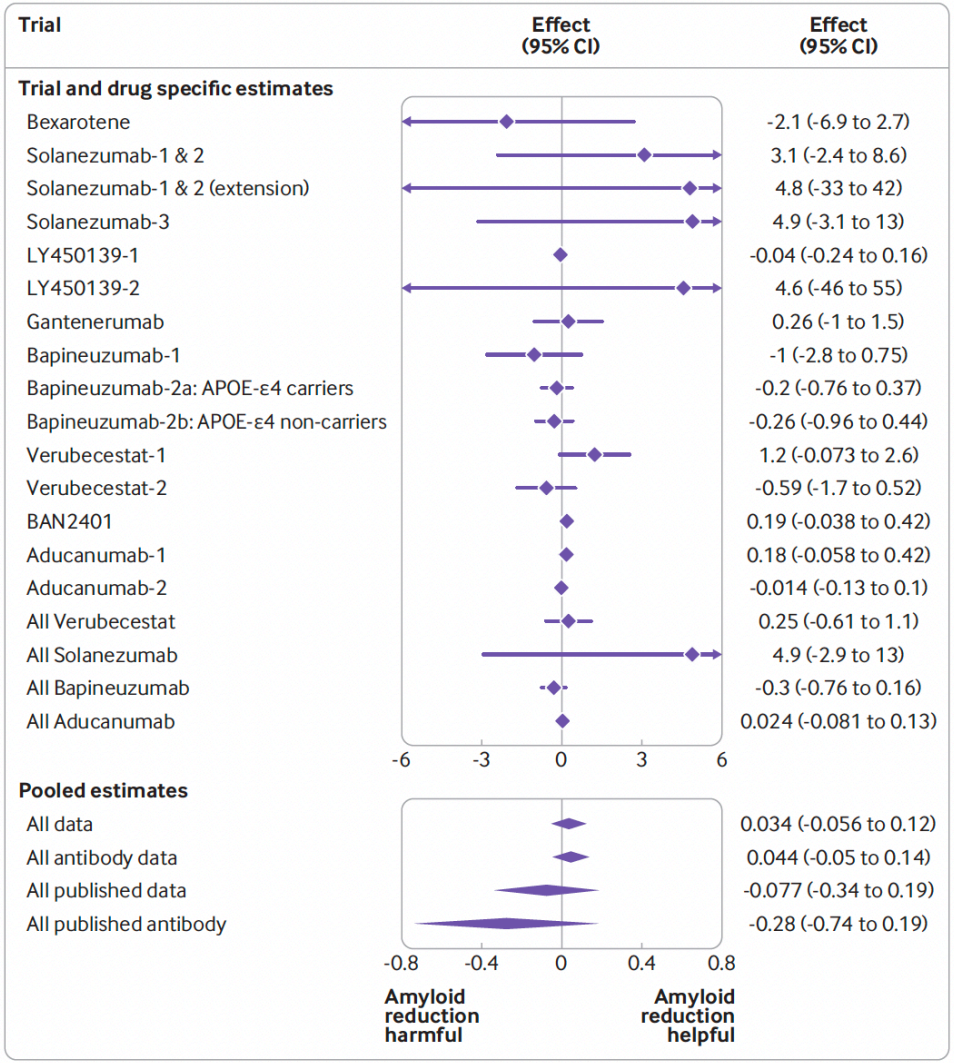


Fig 2 | Forest plot of estimated effects (95% confidence intervals) of a 0.1 decrease in standardized uptake value ratio on mini-mental state examination score for each trial and drug (top panel) and pooled across all drugs and by drug type (bottom panel). Trials of BAN2401 (lecanemab) and aducanumab are unpublished and were excluded from the "all published antibody" category. Centre and width of diamonds represent pooled estimates and 95% confidence intervals, respectively. The numbered key shows multiple trials of the same drug (see appendix table S2 for clinical trial numbers)

Final Comments

- "Statistics means never having to say your certain"
  - I sincerely hope that progress is made in Alzheimers' Disease
  - But as yet, the evidence does not seem to be there

- An unfortunate byproduct of cutting corners in the design, conduct, and analysis of the RCTs
  - First choice is to spend more time understanding clinical trial operating characteristics when faced with the (inevitable) unanticipated issues that arise in an RCT
  - But 3 years of valuable time has been spent trying to have substandard data accepted as evidence
    - We could certainly be 2 years into the conduct of a better RCT had they decided to obtain proper confirmation

Bottom Line

"You better think (think)
about what you're
trying to do…"

-Aretha Franklin, "Think"