UC Irvine Data Science Initiative Short Course

Experimental Design

The Key to Reliable & Reproducible Science

Navneet R. Hakhu, MS

2nd Year PhD Student

Department of Statistics
University of California, Irvine

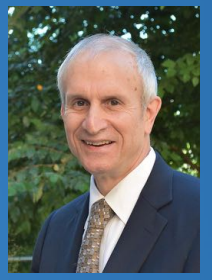
August 28, 2020

Sponsors

- UCI Data Science Initiative (DSI)
- UCI Center for Statistical Consulting (CSC)
- UCI Institute for Clinical & Translational Science (ICTS)

Acknowledgements

- Materials presented based on courses by:
  - Tom Fleming, PhD
    - Professor, Biostatistics & Statistics
      University of Washington
  - Scott Emerson, MD, PhD
    - Emeritus Professor, Biostatistics
      University of Washington
  - Dan Gillen, PhD
    - Professor & Chair, Statistics
      UC Irvine



Course Title according to Scott Emerson:

The Use of Statistics to Answer Scientific Questions

Science and Statistics

- Statistics is about science
  - (Science in the broadest sense of the word)

- Science is about proving things to people
  - (The validity of any proof rests solely on the willingness of the audience to believe it)

Course Objectives

1. To introduce and motivate experimental design*

2. To explore the utility and appropriateness of different types of experimental designs, including clinical trials, to address a given scientific question of interest

* Design in the broadest sense of the word

Course Format

- Session 1 (10:00 AM – Noon Pacific)
  - Lecture 1, Q&A
  - Short Break (~10 min)
  - Lecture 2, Q&A
- [1-hour break: Noon – 1:00 PM Pacific]
- Session 2 (1:00 PM – 3:00 PM Pacific)
  - Lecture 3, Q&A
  - Short Break (~10 min)
  - Lecture 4, Q&A
- [1-hour break: 3:00 PM – 4:00 PM Pacific]
- Session 3 (4:00 PM – 5:00 PM Pacific)
  - Lecture 5, Q&A, Wrap-up

Experimental Design (ExpDsn)

- ...or, Design of Experiments

- Experiments can be in a variety of forms, applied to many different disciplines
  - Agriculture
  - Medicine
  - Clinical Science
  - Basic Science
  - Epidemiology
  - Computer Science
  - Engineering
  - Finance
  - etc.

But what is an experiment?

- Why would we need an experiment?
- Do the experiments differ depending on the scientific context?
- Can we classify experiments? The designs of experiments? Where to start?

Clinical Trials

- Experimentation in human volunteers
- Investigates a new treatment, preventive agent, or diagnostic method
- Safety:
  - Are there adverse effects that clearly outweigh any potential benefit?
    - *Benefit-to-risk*
- Efficacy:
  - Can it alter the disease process in a beneficial way?
- Effectiveness:
  - Would its adoption as a standard affect morbidity / mortality in the population?

The Enemy

- “Let’s start at the very beginning,
  a very good place to start…”

  – Maria von Trapp (nee Kutschera)
  (as quoted by Rodgers and Hammerstein)

Course Outline – Session 1

- Lecture 1:
  - Motivation
  - Prespecifying the Primary Analysis (of the Primary Endpoint to address the Primary Question of Scientific Interest)
    - What if we don't?

- Lecture 2:
  - Study Design Considerations
    - Sampling Scheme and Target Population
      - To what population are we trying to generalize?
  - Choice of Outcome (Response) measure
    - What about surrogates or biomarkers?

Course Outline – Session 2

- Lecture 3:
  - Role of Variables and Relationships [DAGs]
    - Predictor of Interest, Confounder, Precision, Effect Modifier
  - Randomization
  - (Common) Study Designs
    - Single vs. Multi-factor studies

- Lecture 4:
  - Missing Data
    - Retention (as close to 100%)
      - Stopping Study Treatment vs. Withdrawal of Consent
  - Analysis populations
    - Per randomization (ITT)
      - Adherence
        - Real-world achievable (not usually 100%)
    - Different analysis populations

Course Outline – Session 3

- Lecture 5:
  - Non-inferiority designs
    - (compared to typical superiority designs from earlier)

UC Irvine Data Science Initiative Short Course

Experimental Design

The Key to Reliable & Reproducible Science

Lecture 1:

Pre-specifying the Primary Analysis

Navneet R. Hakhu, M.S.

2nd Year PhD Student, Statistics

Department of Statistics
University of California, Irvine

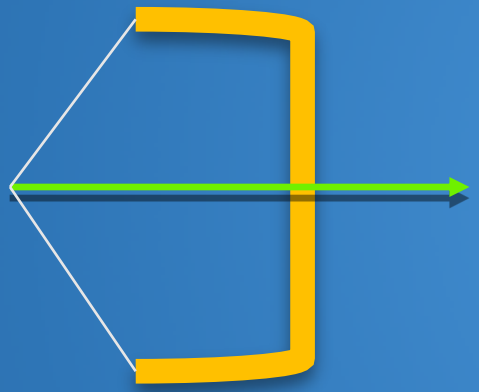
August 28, 2020

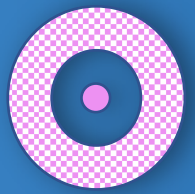
A Common Theme

- Reliability and Reproducibility
  - We want to reduce bias and reduce variability
    - Why?

The Target

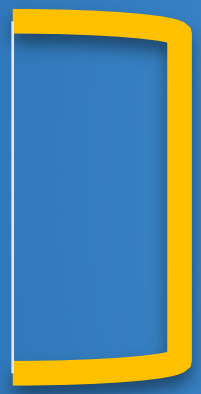
- (The solid circle in the middle)
  - What do we want?

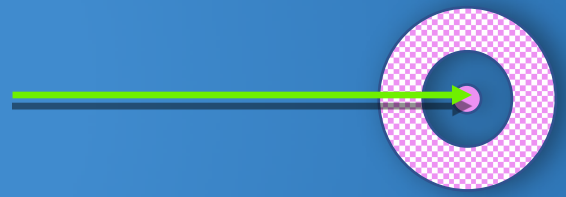




The Target

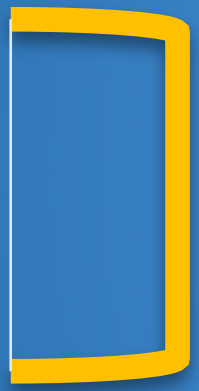
- Goal: Hit the target in one shot



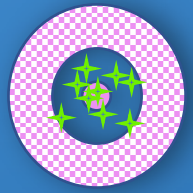


The Target

- Archer 1

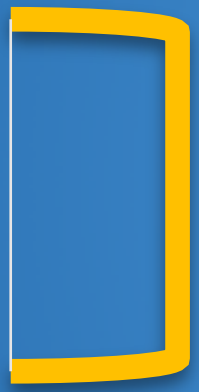


- Where Archer 1's arrows hit

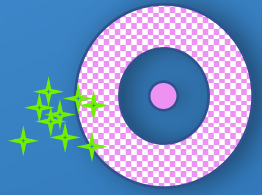


The Target

- Archer 2

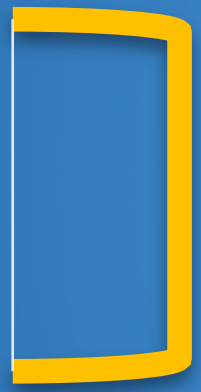


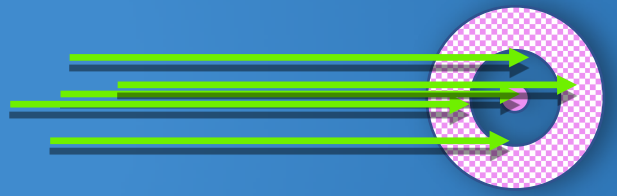
  - Where Archer 2's arrows hit



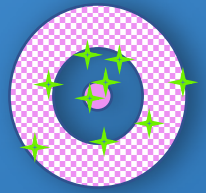
The Target

- Archer 3



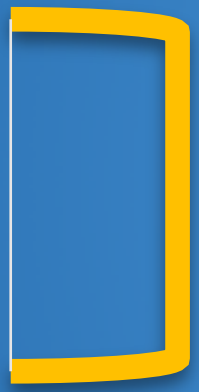


- Where Archer 3's arrows hit

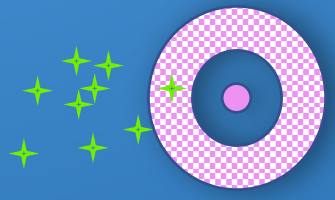


The Target

- Archer 4

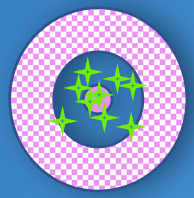


  - Where Archer 4's arrows hit

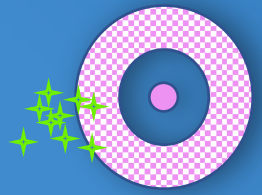


Preference to which archer to hit target?

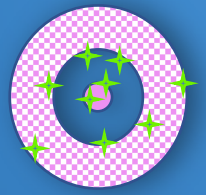
Archer 1



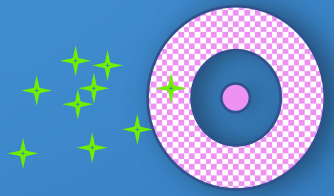
Archer 2



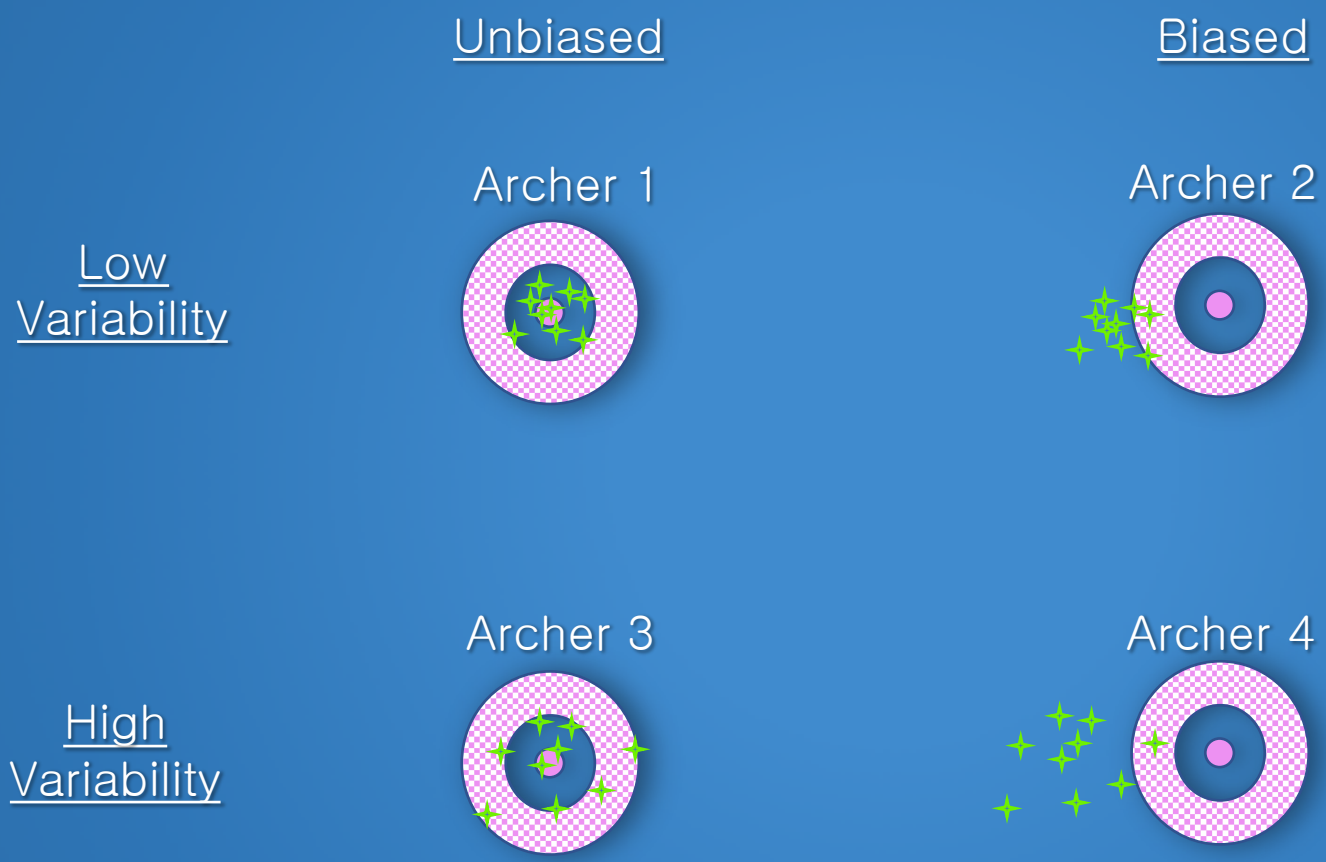
Archer 3



Archer 4

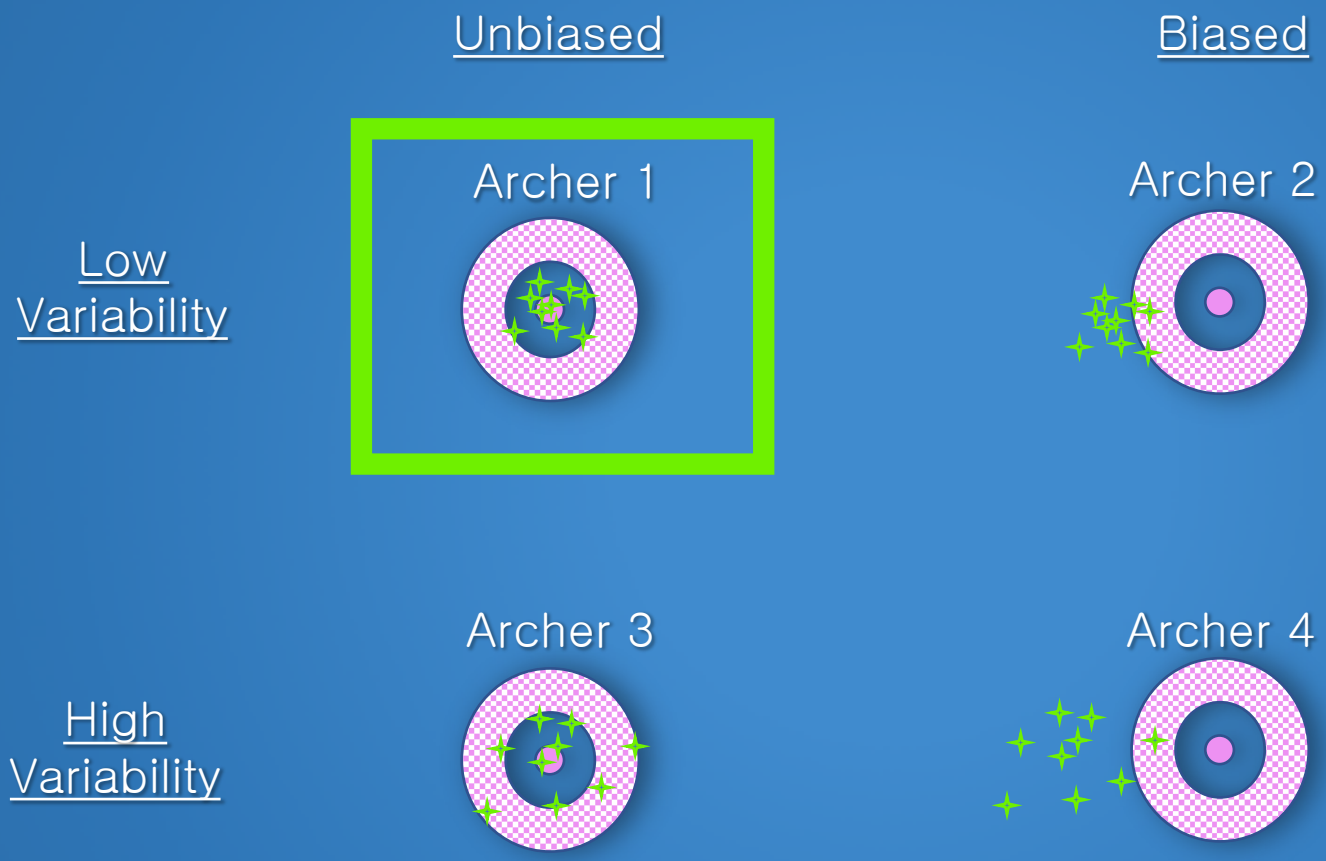


Preference to which archer to hit target?



- Unbiased: on average, archer hits the target
- Variability (or variance): the spread

Preference to which archer to hit target?



- Which archer is next best?

The Second Best Archer

- Do we know where the target is in truth?
  - If yes, then Archer 2 is second best
    - Biased, but we can fix
      - Adjust the aim

- If we don't know where the target is in truth
  - Then Archer 3 is second best
    - At least on average Archer 3 hits the target (unbiased)
      - Why does this matter?

In Statistics

- ...we do not know the true value of the quantity we wish to estimate from a sample (i.e. our data)
  - We only get one shot at hitting the target

- Unbiased (estimate) means that on average we will estimate the true value of the quantity of interest
  - We use statistics when we do not know what the true value is for the population of interest (target population)
  - I rather obtain an unbiased estimate with high variability than a biased estimate with low variability (i.e., Archer 3 over Archer 2)
    - Caveat: ...unless I know how to correct for the bias of Archer 2 (when Archer 1 is not available)
      - Not always possible, however

What does this have to do with ExpDsn?

- The bias–variance tradeoff
  - Ideally we want:
    - No bias (unbiased)
    - Minimum variability
  - In practice, we may not be able to identify Archer 1
    - We may have to settle for a different archer (e.g., the second best, or maybe third best, or …)

- We need an approach in which we consider how we can *a priori* minimize bias and variance
  - What are the sources of bias?
  - What are the sources of variability?
  - Won't statistics solve our concerns?
    - "It depends" – Spoken by almost every statistician

Remembering the Enemy

- So let's start at the end
  - Where do we want to be at the end of a study?
  - What do we want to conclude?
    - "Positive" result
      - e.g., (two-sided) P-value < 0.05?

Maternity Wards (Fleming, 2010)

- In a hospital nursery ~50 years ago, Fleming noticed
  - 20 babies of one sex and 2 of the other sex
    - 20 vs. 2 seems very different from 10 vs. 10 (50% - 50%)
      - Two-sided P-value was 0.0001
        - Statistical significance (!)
        - Wait…why did it take Fleming 40 years to publish this result?

Fleming TR "Clinical Trials: Discerning Hype from Substance." *Annals of Internal Medicine* 2010; 153:400-406

What is the P-value?

- Definition: The probability of observing a result at least as extreme as that which was observed (based on the data) *assuming in truth there is no difference*
  - (Statistical jargon: *assuming the null hypothesis is true*)

- In Fleming's situation, the P-value of 0.0001 was:
  - The probability of an imbalance that extreme (20 vs. 2) occurring by chance was 1 in 10,000
    - Better than the typical P-value < 0.05 (1 in 20) threshold often used in scientific disciplines
      - So 1 in 10,000 is really unlikely, no?
      - Seems like he could have published this result in less than 40 years…

The Problem

- Data-driven hypothesis
  - Fleming did not *a priori* hypothesize that there might be a potential difference in the sex distribution of babies born (i.e., not a 50-50 split) before getting to the nursery
    - He saw something "unusual" and then tested that data-driven hypothesis
      - ***IMPORTANT POINT***: "a P-value is only interpretable when you understand the sampling context from which it is derived"
    - Instead, that observed result could have served as a form of *hypothesis generation*
      - Problem still exists if a new set of data was collected based on the data-generated hypothesis (11 vs. 11) and both samples combined in a "meta-analysis"
        - 31 vs. 13 (two-sided P-value = 0.0096)

Think Before You Look (↑ type 1 error)

- Proof of concept simulation study
  - Truth: no difference (here the distributions the same)

# of outcomes "looked" at	Type 1 error (False positive rate)					
	corr=0	corr=0.1	corr=0.3	corr=0.5	corr=0.7	corr=0.9
1						
(outcome 1)	0.059	0.056	0.047	0.053	0.050	0.049
(outcome 2)	0.053	0.049	0.051	0.049	0.047	0.050
(outcome 3)	0.049	0.051	0.050	0.047	0.052	0.048
2						
(outcomes 1,2)	0.107	0.101	0.093	0.092	0.083	0.070
(outcomes 2,3)	0.099	0.098	0.096	0.087	0.083	0.068
(outcomes 1,3)	0.105	0.104	0.093	0.092	0.084	0.069
3						
(all 3 outcomes)	0.151	0.147	0.135	0.125	0.108	0.082

Single Binary Predictor of Interest (n=500 per group);
Three Continuous Outcomes, conditional on each predictor group (level),
assuming MVN with mean 0, variance 1,
and specified exchangeable correlation
Two-sample t-test used to analyze difference in means between 2 groups
(10,000 sims per setting)

Multiple Analyses: During 4-Year Trial

- The (log rank) P-value was less than 0.05:
  - At the final test (at 4 years) in 5 of 100 studies
  - At either 2- or 4-year test in 10 of 100 studies
  - At least 1 of 4 yearly tests in 17 of 100 studies
  - At least 1 of 8 semi-annual tests in 21 of 100 studies
  - At least 1 of 16 three-month tests in 26 of 100 studies

Everyone wants "Positive" Results

- (Not only in clinical trials)

- Industry Sponsors
  - Company profits, stocks, promotion
- Government Sponsors
  - Claims of success in advancing health care
  - Leverage for increase in federal funding
- Journal Editors
  - "Positive" results sell (increase in readership)
- Academic Investigators
  - Increase publications, notoriety, salary, earlier promotion
- Caregivers
  - More options for patients/loved ones

Conflicts of Interest

- Each entity brings their own conflicts of interest
  - Each of these are not bad in their own right
  - However, if these conflicts inhibit furthering of science (the truth), then there is a problem

Bias for "Positive" Results in Trials

- What is the definition of a successful clinical trial?

- Very common response:
  - "A clinical trial that achieves a *positive* result."

Bias for "Positive" Results in Trials

- What is the definition of a successful clinical trial?

- Very common response:
  - "A clinical trial that achieves a *positive* result."

- The proper scientific response:
  - "A clinical trial that *reliably answers* the questions the trial was designed to address."
    - (Objective/Aim: "To determine whether…")

The Public Health Objective

- Objective: high prevalence of truly beneficial therapies among all therapies approved for clinical practice
  - i.e., clinical trials must have high Positive Predictive Value
    - PPV, also called Predictive Value of the Positive (PV+)
- PPV
  - Diagnostic testing: prevalence of diseased individuals among those with a positive diagnostic test
  - Clinical trials: prevalence of truly beneficial therapies among those which are identified by a positive clinical trial
  - PPV is calculated using Bayes rule:

$$PPV = \frac{sensitivity \times prevalence}{(sensitivity \times prevalence) + (1-specificity) \times (1-prevalence)}$$

Clinical Trials as Diagnostic Tests

- A statistical hypothesis test can be viewed as a test for beneficial treatments
  - α–level: Probability of observing a positive (statistically significant) test in absence of a true treatment effect
    - Level of significance is 1 – specificity (false positive error rate)
    - Choosing $\alpha = 0.05$ gives 95% specificity
  - Statistical power (β): Probability of observing a positive (stat. significant) test when there is a true treatment effect
    - Power is sensitivity (true negative error rate)
    - Common choices are 80% or 90% (for me it depends)
  - Prevalence (π_0): the percentage of effective treatments among all tested treatments
  - Positive Predictive Value (PPV): Probability that a stat. significant trial indicates a truly effective treatment

$$\text{PPV} = \frac{\beta\pi_0}{\beta\pi_0 + \alpha(1-\pi_0)}$$

Ex: Colorectal Cancer

- Consider all experimental therapies for colon adjuvant setting
  - Suppose only 4% (= 40/1000 = π_0) are truly positive (prevalence) and 96%(= 960/1000) are truly negative
  - Suppose false negative error rate is 10%
    - (i.e., Statistical power = β = 0.90 = 36/40)
  - Suppose false positive error rate is 5%
    - (i.e., α = 0.025 = 24/960)

		Truth		
		Positive	Negative	
Result of Experiment	Positive	36	24	60
	Negative	4	936	940
		40	960	1000

  - Then, the probability that a positive trial is a true positive (PPV) is 36/60 = 0.60

PPV ↑ via Good Experimental Practice

$$\text{PPV} = \frac{\beta\pi_0}{\beta\pi_0 + \alpha(1-\pi_0)}$$

- Increase π_0:
  - Careful planning of preliminary studies
  - Avoid "novel" and "innovative" ideas
  - Careful specification of hypothesis-riven research (avoid "science by hunch")
- Increase β:
  - Good practice (no missing data, low variation in outcome assessment, good adherence, etc.)
- Reduce α:
  - Pre-specify outcomes
  - Pre-specify all analyses
  - Avoid multiple comparisons
  - Avoid surrogate outcomes
  - Avoid subgroups

Sanchez BJ (2014) Evaluation of Strategies for the Phase II to Phase III Progression in Treatment Discovery (MS Thesis) http://rctdesign.org/TechReports/BSanchezThesis20150220.pdf

Legal Requirements for Good Science

- 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 reducing 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 Amendment Act (2007)
  - Registration of RCTs, Pediatrics, Risk Evaluation and Mitigation Strategies (REMS)

The Public Health Objective

- A wide range of situations/therapies are studied in clinical trials
- Globally, clinical trials need to assure:
  - Scientific credibility
  - Ethical experiments
  - Efficient experiments
    - Minimize time
    - Minimal number of extra subjects
    - Minimize cost
  - A high prevalence of truly beneficial therapies among all therapies used in routine care

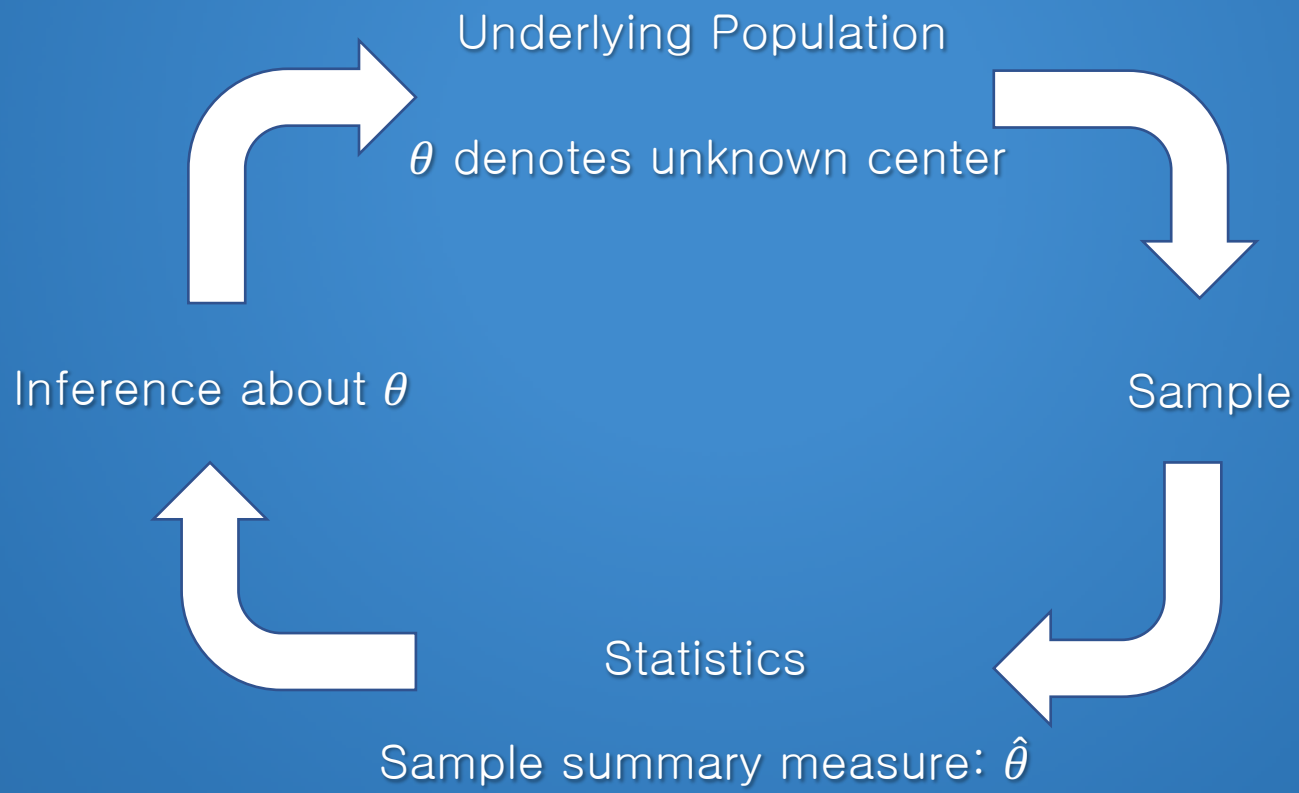
Statistical Foundations

- Key elements
  - Empirical objective
  - Four required elements for inference
  - Properties of estimators
  - Interpretation of interval estimates

- Why (briefly) review foundations?
  - We are discussing the scientific setting, and:
    - As a scientific experiment, the results of a clinical trial are used to rule out (or rule in) hypotheses about treatment effects
    - The standards for rejecting (or accepting hypotheses) are based on statistical criteria

The Empirical Objective

- Use observed trial result ($\hat{\theta}$) to make inference about underlying population θ



Four Main Inferential Elements

1. Point estimate: $\hat{\theta}$ is the "best" estimate of θ

2. Interval estimate: Values of θ that are consistent with the trial results

3. Expression of uncertainty (P-value): To what degree is a particular hypothesis (the "null" hypothesis) consistent with the observed trial results?

4. Decision: Based on the above measures, what decision should be reached about the use of a new therapy?

Desirable Properties

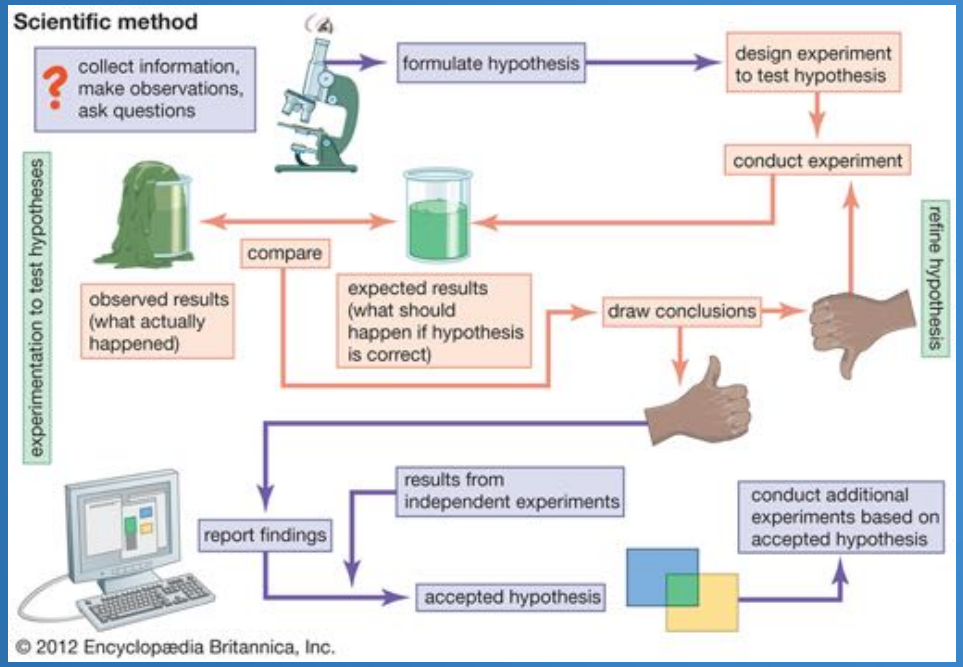
- Point estimate
  - Unbiased and consistent: the long-run average of $\hat{\theta}$ is very close to θ
  - Small variance
    - (Uniform Minimum Variance Unbiased Estimator)
- Interval estimate (non-rejection region)
  - Correct coverage probability (e.g., 95% of all 95% confidence intervals include θ)
  - As narrow as possible while maintaining the correct coverage probability
- P-value
  - Correct size
- Decision
  - Decision criteria maintain the appropriate type 1 statistical error rate

Remembering the Enemy

- So let's start at the end
  - Where do we want to be at the end of a study?
  - What do we want to conclude?
    - "Positive" result
      - e.g., (two-sided) P-value < 0.05?
  - How do we get there?
    - The Scientific Method?

The Scientific Method

- Britannica



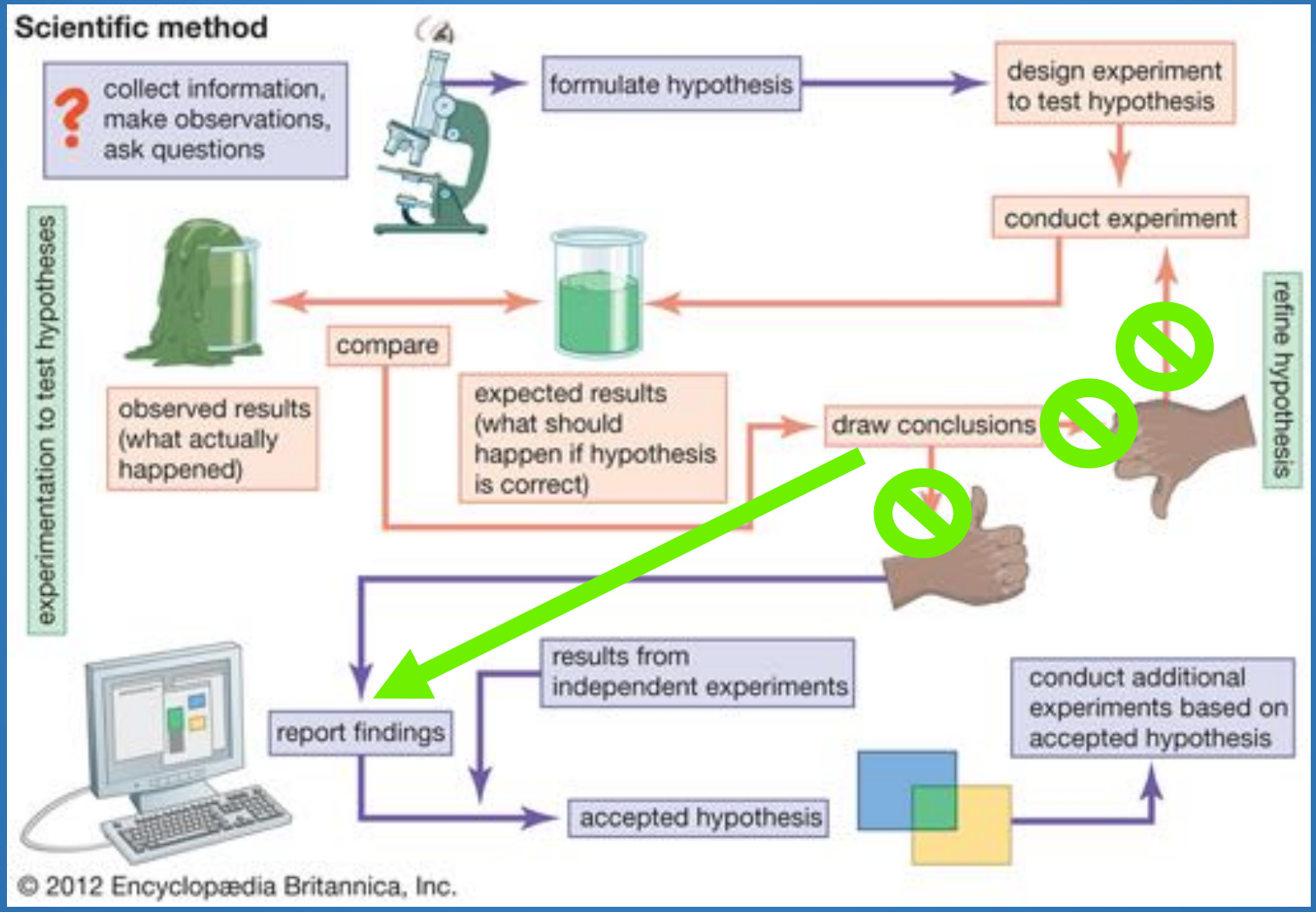
https://www.britannica.com/science/scientific-method#/media/1/528929/159476

- Wikipedia
  1) Define a question
  2) Gather information and resources (observe)
  3) Form an explanatory hypothesis
  4) Test the hypothesis by performing an experiment and collecting data in a reproducible manner
  5) Analyze the data
     Interpret the data and draw conclusions that serve as a starting point for new hypothesis
  6) Publish results
  7) Retest (frequently done by other scientists)

  The iterative cycle inherent in this step-by-step method goes from point 3 to 6 back to 3 again

https://en.wikipedia.org/wiki/Scientific_method#Elements_of_the_scientific_method

Modifications



The Scientific Method

- The scientific method is an iterative process of posing and evaluating hypotheses using carefully designed experiments
- A clinical trial is an experiment and should be built on carefully-framed hypotheses:
  - What is the treatment?
  - What is θ (the measure of treatment effect)?
  - What are important differences?
  - What differences support recommending use of a new treatment?
- The trial must be designed to be informative relative to the hypotheses (the scientist game)
- Upon completion, the range of viable hypotheses that remain is determined by the experimental results

The Scientist Game

- Try the scientist game:
  http://www.emersonstatistics.com/ScientistGame
- Careful consideration of what you want to know upon trial completion is essential
- The scientist game is illustrative of the scientific importance of all aspects of the design including:
  - Specification of the treatment
  - Selection and definition of the outcome(s)
  - Choice of control group
  - Definition of design hypotheses
  - Statistical standard for evidence
  - Choice of sample size

Summary

- A general goal in any study is to reduce bias and reduce variability
- Pre-specifying the primary analysis to prevent inflation of the type 1 error
  - Avoid spurious results
    - "If you torture your data long enough, they will confess" (Fleming, 2010)
  - P-values are only interpretable when you understand the sampling context from which they were derived
- Exploratory analyses are for hypothesis generation
  - (vs. Confirmatory trials that can enhance reliability)
- The scientific method is our friend
  - (the "correct" version)

UC Irvine Data Science Initiative Short Course

Experimental Design

The Key to Reliable & Reproducible Science

Lecture 2:

Study Considerations, Outcome Measures & Surrogates/Biomarkers

Navneet R. Hakhu, M.S.

2nd Year PhD Student, Statistics

Department of Statistics
University of California, Irvine

August 28, 2020

Scientific Setting

- The goal of medical science is to produce the evidence that can be used to
  - Gain approval of new treatments and diagnostic tests
  - Provide evidence to be used in applying those treatments and tests

Goals of Medical Research

- Identify methods to diagnose disease
- Identify risk factors for disease
- Identify treatments for disease
- Identify methods for disease prognosis
- Identify strategies for prevention of disease
- Basic science

Legal Requirements for Good Science

- 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 reducing 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 Amendment Act (2007)
  - Registration of RCTs, Pediatrics, Risk Evaluation and Mitigation Strategies (REMS)

Typical Chronology

- Observational epidemiology of disease, risk
- Preclinical experiments
  - Laboratory, animal studies of mechanisms, toxicology
- Clinical trials
  - Safety for further investigations / dose
  - Safety for therapy
  - Measures of efficacy
  - Confirmation of efficacy / effectiveness
- Synthesis and quantification of evidence
- Adoption of new treatment indication

Types of Studies – 1

- Anecdotal observations
  - Case report
  - Case series
  - Hypothesis generation

Types of Studies – 2

- Designed observational study: Case–control
  - Sample diseased and nondiseased
  - Examine rates of exposures
  - Efficient for rare diseases
  - Can look at multiple risk factors
  - Limitation: Cannot infer cause and effect
    - Correlations with other factors
    - Protopathic associations

Types of Studies – 3

- Designed observational study: Cohort study
  - Sample exposed and nonexposed
  - Examine rates of disease
  - Efficient for common diseases
  - Can look at multiple diseases
  - Can identify "retrospective cohort"
  - Limitation: Cannot infer cause and effect
    - Correlations with other factors
    - Protopathic associations

Types of Studies – 4

- Designed interventional study: Clinical trial
  - Assign subjects to treatments
  - Examine outcomes
  - Can look at multiple diseases
  - Can infer cause and effect

Clinical Trials

- Experimentation in human volunteers
- Investigates a new treatment, preventive agent, or diagnostic method
- Safety:
  - Are there adverse effects that clearly outweigh any potential benefit?
    - *Benefit-to-risk*
- Efficacy:
  - Can it alter the disease process in a beneficial way?
- Effectiveness:
  - Would its adoption as a standard affect morbidity / mortality in the population?

Clinical Trials

- Experimentation in human volunteers
- Investigation of a new treatment or preventive agent
  - Safety: Do adverse effects outweigh any benefit?
  - Efficacy: Can treatment beneficially alter disease?
  - Effectiveness: Would adoption of the treatment help population's health?
- Investigation of existing treatments
  - Relative benefits: Is one treatment clearly superior?
  - Harm: Should a therapy currently in use be removed?
- Some questions cannot be answered by a clinical trial
  - e.g., establishing harm of a new substance

Competing Goals of a Clinical Trial

- Scientific
  - Questions regarding mechanistic pathways
- Ethical
  - Minimize harm (due to treatment or disease) done to patients
- Clinical
  - Improve the overall health of patients
- Statistical
  - Quantifying scientific questions in a precise manner

Minimum Scientific Standards

- Trial must address a meaningful question
  - Discriminate between viable hypotheses (Science)
- Trial results must be credible to the scientific community
  - Valid materials, methods (Science, Statistics)
  - Valid measurement of experimental outcome (Science, Clinical, Statistics)
  - Valid quantification of uncertainty in experimental procedure (Statistics)

Individual Ethics

- Conducted in human volunteers, the clinical trial must be ethical for participants on the trial
  - Minimize harm and maximize benefit for participants in clinical trial
  - Avoid giving trial participants a harmful treatment
  - Do not unnecessarily give trial participants a less effective treatment

Group Ethics

- The clinical trial must ethically address the needs of the greater population of potential recipients of the treatment
  - Approve new beneficial treatments as rapidly as possible
  - Avoid approving ineffective or (even worse) harmful treatments
  - Do not unnecessarily delay the new treatment discovery process

Ethical Issues

- Mechanisms for ensuring ethical treatment of study subjects
  - Before starting the study
    - Institutional review board (IRB)
      - also known as Independent Ethics Committee (IEC)
  - During conduct of the study
    - Data monitoring committee (DMC)
      - also known as Data Safety Monitoring Board (DSMB)
  - After studies completed
    - Regulatory agencies (e.g., FDA, EMA, PMDA, etc.)

Safety

- Safety to conduct trials
  - No serious adverse effect appears in first few subjects treated
- Safety in practice
  - Manageable adverse reaction profile
  - Severity: mild or moderate
  - Serious: rarely leads to hospitalization, disability, death, congenital abnormalities
  - Risk factors: patients at risk for adverse reactions can be identified before lasting clinical harm occurs
    - Errors of commission: treatment causes lasting disability
    - Errors of omission: lack of tolerability causes delay in better therapy
- Risk / benefit tradeoffs
  - Lack of efficacy in presence of any adverse reactions may be a safety problem

Efficacy: A Moving Target

- Definition of efficacy can vary widely according to choice of endpoint and magnitude of importance
  - Basic science
    - Does treatment have any effect on the pathway?
  - Clinical science
    - Does treatment have a sufficiently large effect on a clinically relevant endpoint in some subpopulation of the target population?

Effectiveness: A Moving Target

- A treatment is "effective" if its introduction improves health in the population
  - Considers the net effect of safety and efficacy in the population as a whole
  - Takes into account such issues as
    - Noncompliance (nonadherence)
    - Off-label use

Efficacy vs. Effectiveness

- A treatment can be both efficacious and ineffective depending on such factors as
  - Target population
    - Restricted eligibility due to toxicity, compliance
  - Intervention
    - Training, quality control, compliance
  - Comparison treatment
    - No treatment, active treatment, ancillary treatments
  - Measurement of outcome(s)
    - Clinical disease vs. subclinical markers
  - Summary measure of outcome distribution
    - Effects on mean, median, outliers

Disease

- Efficacy and effectiveness study populations may differ with respect to
  - Certainty of diagnosed disease
  - Subgroups with more (less) severe disease

Target Population

- Efficacy and effectiveness study populations may differ with respect to
  - Properly diagnosed disease
  - Subgroups with more (less) disease
  - Tolerance of treatment
  - Willingness to comply with treatment
  - Ancillary treatments
  - Different risk factors

Ex: Desensitization in Allergy

- Efficacy trial might consider
  - Patients with proven allergy who have shown "response" in open label study (perhaps due to genetic profile?)
  - Exclusion criteria for safety in trial
    - Cannot tolerate oral food challenge
    - Patients likely to be noncompliant
  - Exclusion criteria to ensure adequate data

- Effectiveness populations might include
  - All patients with reported allergy

Intervention

- Efficacy and effectiveness populations may differ with respect to
  - Dose
  - Administration
  - Duration
  - Training
  - Quality control

Ex: Insulin Dependent Diabetes

- Efficacy trial might consider
  - Glucose monitoring according to protocol
  - Lengthy training
  - Close monitoring and retraining when necessary

- Effectiveness trial should strive for realistic setting
  - What would instructions and training, monitoring be if treatment were efficacious?
  - What if treatment fails (use another)?

Measurement of Outcome

- Efficacy and effectiveness populations may differ with respect to
  - Clinical measurement
  - Timing of measurement

Ex: Hypercholesterolemia

- Efficacy trial might consider
  - Lowering serum cholesterol
  - Means
- Effectiveness trial should strive for relevant outcomes
  - Proportion exceeding acceptable thresholds
    - Normal cholesterol levels
  - Time of survival

Primary Outcomes

- Criteria
  - The goal of a RCT is to find effective treatment indications
  - The primary outcome is a crucial element of the indication

Scientific Basis

- A clinical trial is planned to detect the effect of a treatment on some outcome

- Statement of the outcome is a fundamental part of the scientific hypothesis

Ethical Basis

- Generally, subjects participating in a clinical trial are hoping that they will benefit in some way from the trial

- Clinical endpoints are therefore of more interest than purely biological endpoints

Primary Endpoint: Clinical

- Consider (in order)
  - The most relevant clinical endpoint
    - Survival, quality of life
  - The endpoint the treatment is most likely to affect
  - The endpoint that can be assessed most accurately and precisely
    - (in an appropriate/feasible timeframe)

Clinically Relevant Outcome

- Consistently and readily measurable
- Sensitive
- Well defined and reliable
- Clinically meaningful
  - "a direct measure of how a patient feels, functions or survives" (Robert Temple, FDA)
    - Feels
      - E.g., chest pain, breathlessness, fatigue, dizziness
    - Functions
      - Ability to conduct normal activities
        - Ability to walk, to engage in recreational activities, for self care, risk of syncope (fainting)
        - Time in hospital or missing school (overall, or cause specific)
    - Survives

Potential Clinically Meaningful Benefit

- Patient Reported Outcomes (PROs)
  - "Any report of the status of a patient's health condition that comes directly from the patient, without interpretation of the patient's response by a clinician or anyone else."
    - FDA Guidance for Industry. Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims. (December, 2009)
  - Note: These are subjective and specific to each subject
    - Clinically meaningful endpoints, but need to confirm:
      - Reliability, sensitivity, validity (content, construct, etc.), Clinical Relevance, Interpretability
      - Integrity, including need for:
        - Blinded assessment and control of missing data
      - Mobilize disease specific interest groups before sponsors plan clinical trials

Additional Endpoints

- Other outcomes are then related to a “secondary” status
  - Supportive and confirmatory
  - Safety

- Some outcomes are considered “exploratory”
  - Subgroup effects
  - Effect modification (interaction)

Primary Endpoint: Clinical

- Consider (in order of importance)
  - The phase of study: What is current burden of proof?
  - The most relevant clinical endpoint
    - Survival, quality of life
    - Proven surrogates for the above
      - But how can we be sure?
  - The endpoint the treatment is most likely to affect
    - Therapies directed toward improving survival
    - Therapies directed toward decreasing AEs
  - The endpoint that can be assessed most accurately and precisely
    - Avoid unnecessarily high invasive measurements
    - Avoid poorly reproducible endpoints

Multiple Endpoints

- Sometimes we must consider multiple endpoints
- We then control experiment-wise error
- Possible methods
  - Composite endpoint
    - AND: Individual success must satisfy all
    - OR: Individual success must only satisfy one
    - AVERAGE: Sum of individual scores
    - EARLIEST: e.g., event-free survival
  - Co-primary endpoints
    - Must show improvement in treatment group on all endpoints
      - No guarantee that the same subjects are experiencing the improvement
      - (Gate-keeping: but how to decide?)

Goal of a Clinical Trial

- Establish whether an experimental treatment will prevent a particular clinical outcome

Problems with Clinical Outcomes

- Relevant clinical outcomes are often relatively rare events that occur after a significant delay
  - Believe that earlier interventions have greater chance of benefit
- Difficulty in measuring clinical outcome
  - Quality of life needs to be assessed over a sufficiently long period of time

Impacts on Clinical Trial Design

- Large sample size required to assess treatment effect on rare events
- Long period of follow-up needed to assess endpoints
- Can we do something else?

Surrogate Outcomes

- Statistical and logistical constraints often lead to the desire for surrogate outcomes
  - But are they reliable?

Motivation for Surrogate Endpoints

- Hypothesized role of surrogate endpoints
  - Find a biological endpoint which
    - Can be measured in a shorter timeframe,
    - Can be measured precisely, and
    - Is predictive of the clinical outcome
  - Use of such an endpoint as the primary measure of treatment effect will result in more efficient trials

Identifying Potential Surrogates

- Typically use observational data to find risk factors for clinical outcome
- Treatments attempt to intervene on those risk factors
- Surrogate endpoint for the treatment effect is then a change in the risk factor

Examples

- AIDS
  - HIV leads to suppression of CD4 cells
  - Decreased CD4 levels correlates with development of AIDS
  - Treatment effects measured by following CD4 counts
  - True clinical outcome is prevention of morbidity and mortality

Examples

- Coronary heart disease
  - Poor prognosis in patients with arrhythmias following heart attack
  - Therapies directed toward preventing arrhythmias
  - Treatment effects measured by prevention of arrhythmias
  - True clinical outcome is prevention of mortality

Problem

- Establishing biologic activity does not always translate into effects on the clinical outcome

- May be treating the symptom, not the disease
  - Examples
    - Concorde: Zidovudine (ZDV) improves CD4, not survival
    - CAST: encainide, flecainide prevents arrhythmias, worsens survival

- May be missing effect through other pathways
  - Example
    - Int'l CDG group: Gamma-INF no affect on biomarkers, decreases serious infections

Example: Concorde Trial

- (Lancet, April 3, 1993)
  - Asymptomatic HIV positive patients
  - Randomize to
    - Immediate ZDV (n = 877)
    - Placebo then progression to ZDV (n = 872)
  - Mean follow-up: 3 years

Concorde Trial: Surrogate Results

- CD4 changes
- 3 mos relative to baseline
  - Immediate ZDV: +20 cells
  - Placebo: −10 cells
- Difference between treatment arms
  - 3 mos: 30 cells ($P < 0.0001$)
  - 6 mos: 35 cells ($P < 0.0001$)
  - 9 mos: 32 cells ($P < 0.0001$)

Concorde Trial: Clinical Results

- Clinical Results:

	ZDV (n = 877)	Placebo (n = 872)
AIDS / Death Death	175 95	171 76
3 year survival	92%	93%

- Conclusions:
  - "Results cast doubt on the value of using changes over time in CD4 count as a predictive measure for effects of antiviral therapy on disease progression and survival."

Example: CAST

- Cardiac Arrhythmia Suppression Trial
  - Arrhythmia a risk factor for sudden death following a myocardial infarction (MI)
  - Anti-arrhythmic drugs (encainide and flecainide) successfully decrease incidence of arrhythmias
  - CAST
    - Placebo controlled trial using mortality as outcome
    - Encainide and flecainide TRIPLE the death rate

Example: CGD

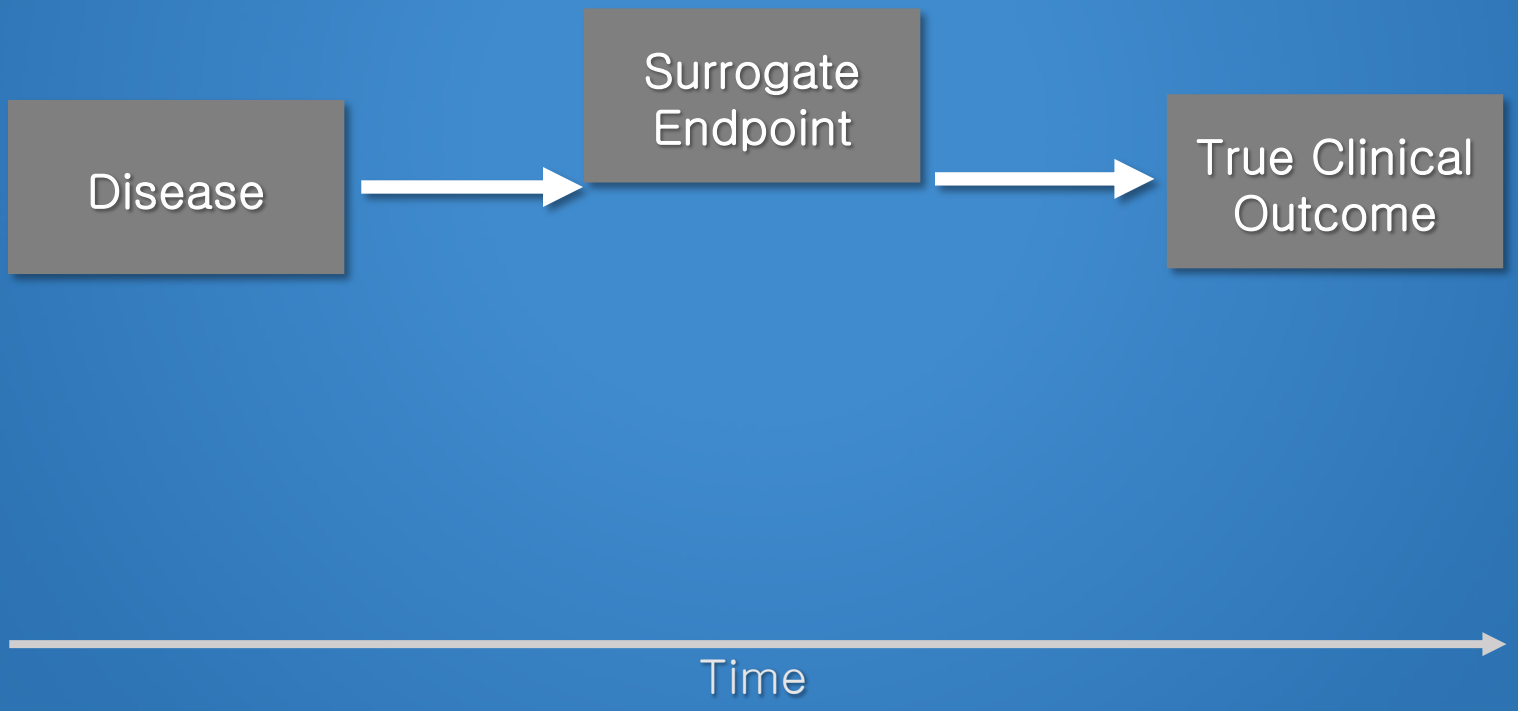
- Chronic Granulomatous Disease (CGD)
  - CGD leads to recurrent serious infections
  - Gamma interferon increases bacterial killing and superoxide production?
  - International CGD Study Group of Gamma-INF
    - 70% reduction in recurrent serious infections
    - Essentially no effect on biological markers

Surrogate Outcomes

- Possible mechanisms
  - Understanding the pitfalls of surrogate outcomes requires thinking about the mechanisms of treatments
    a) ideal
    b) inefficient
    c) misleading
    d) dangerous

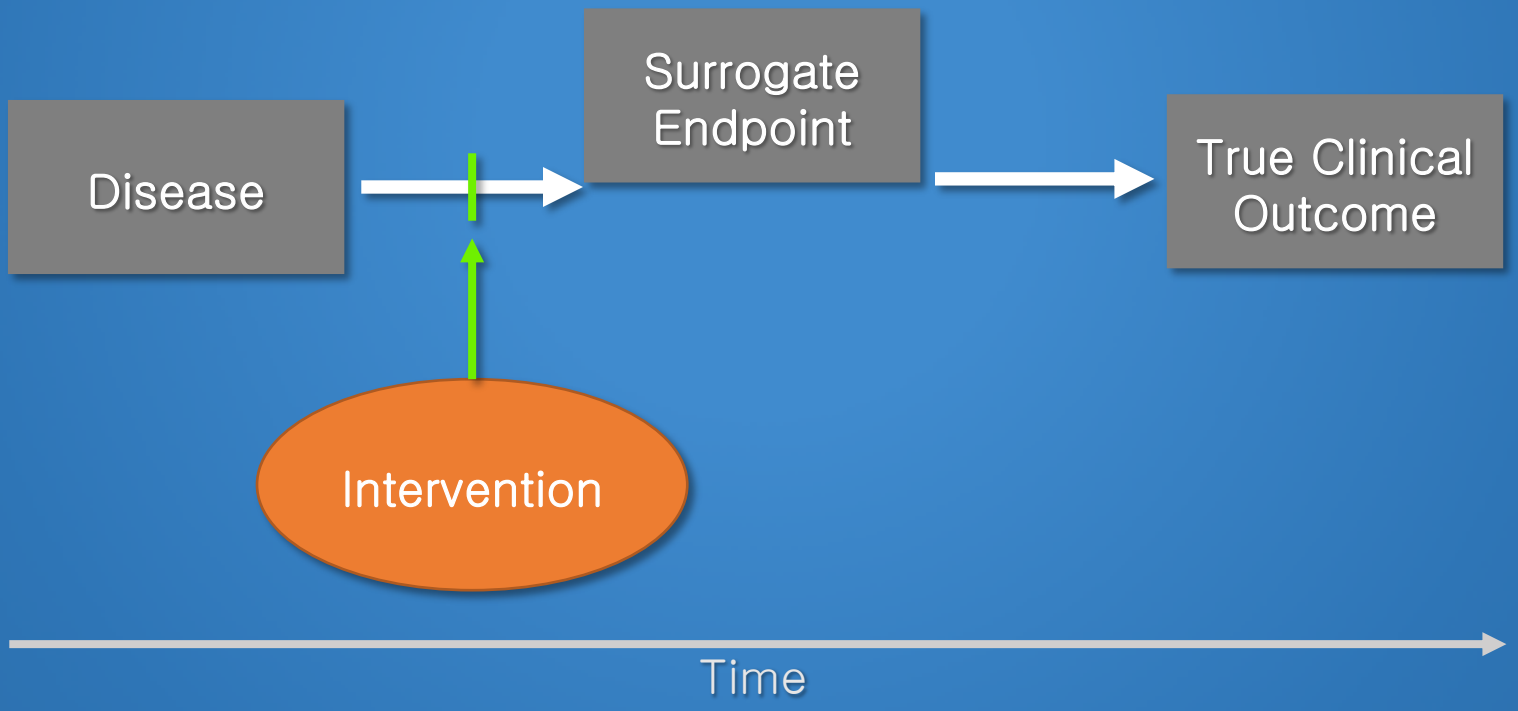
Scenario 1: The Ideal

- Disease progresses to Clinical Outcome only through the Surrogate Endpoint



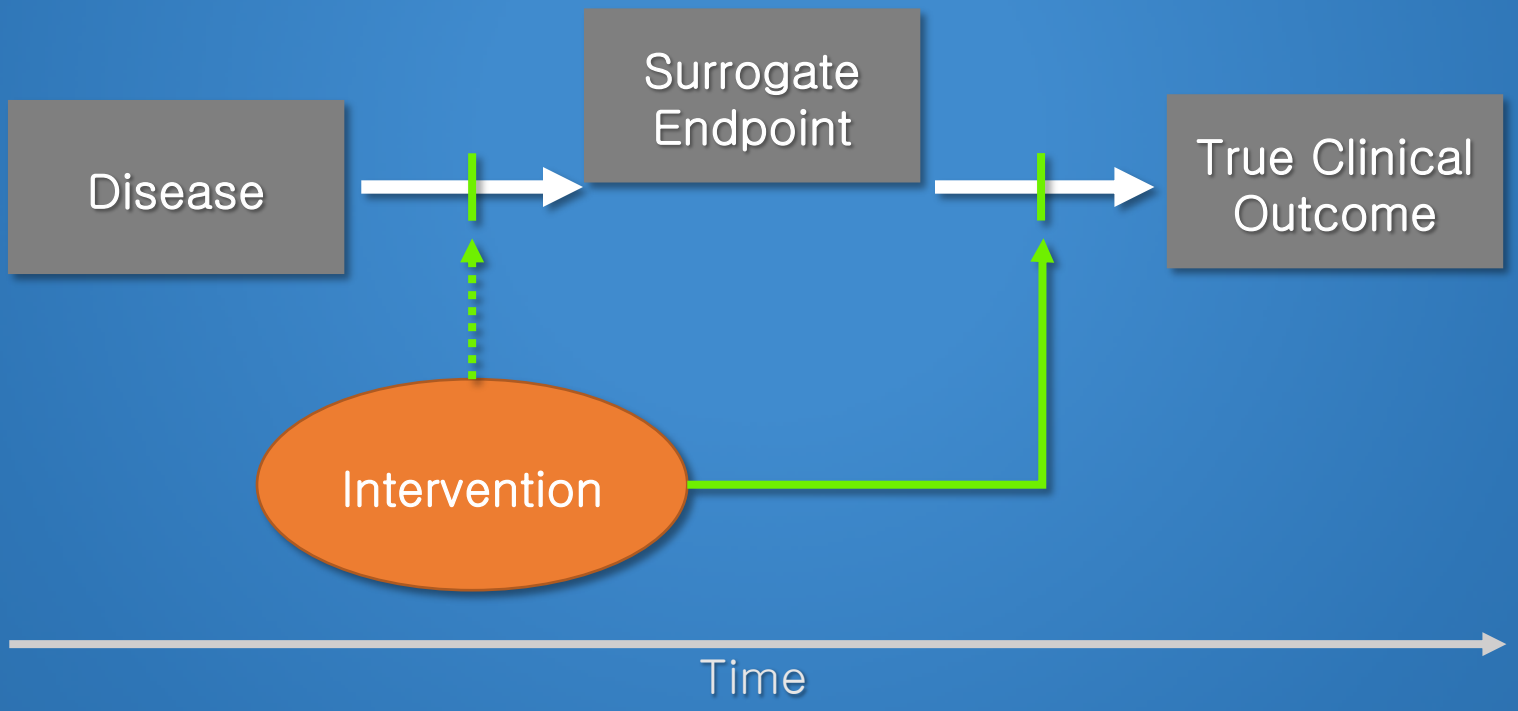
Scenario 1a: Ideal Surrogate Use

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



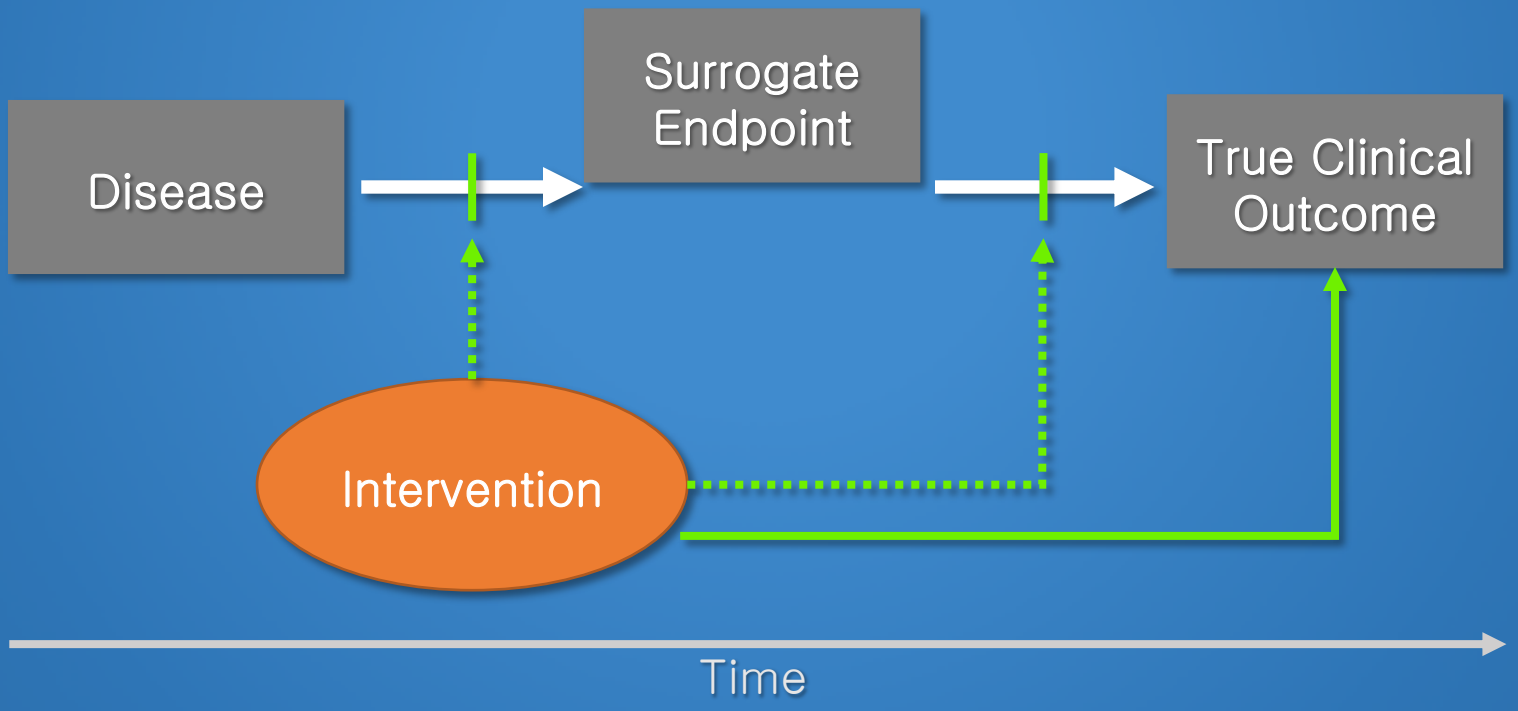
Scenario 1b: Inefficient Surrogate

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



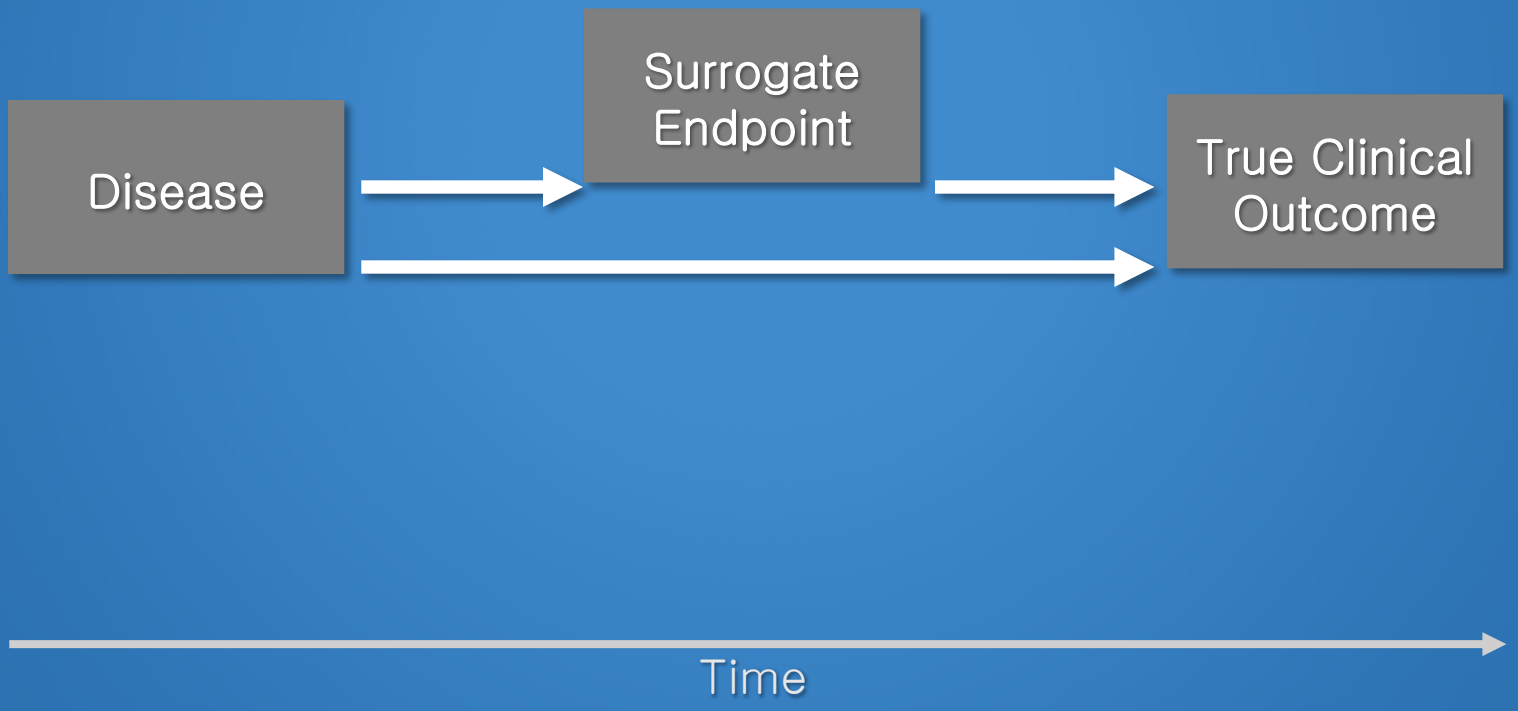
Scenario 1d: Dangerous Surrogate

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



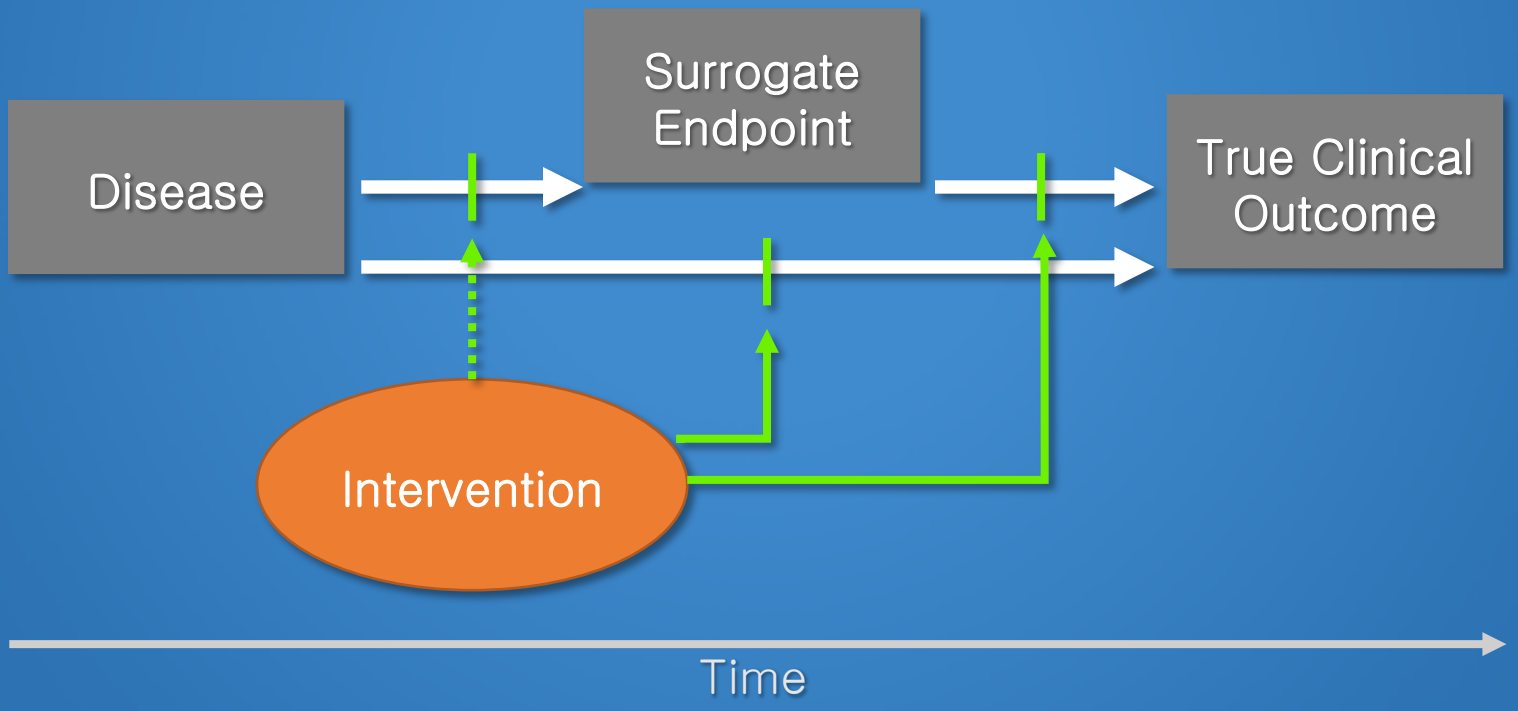
Scenario 2: Alternate Pathways

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



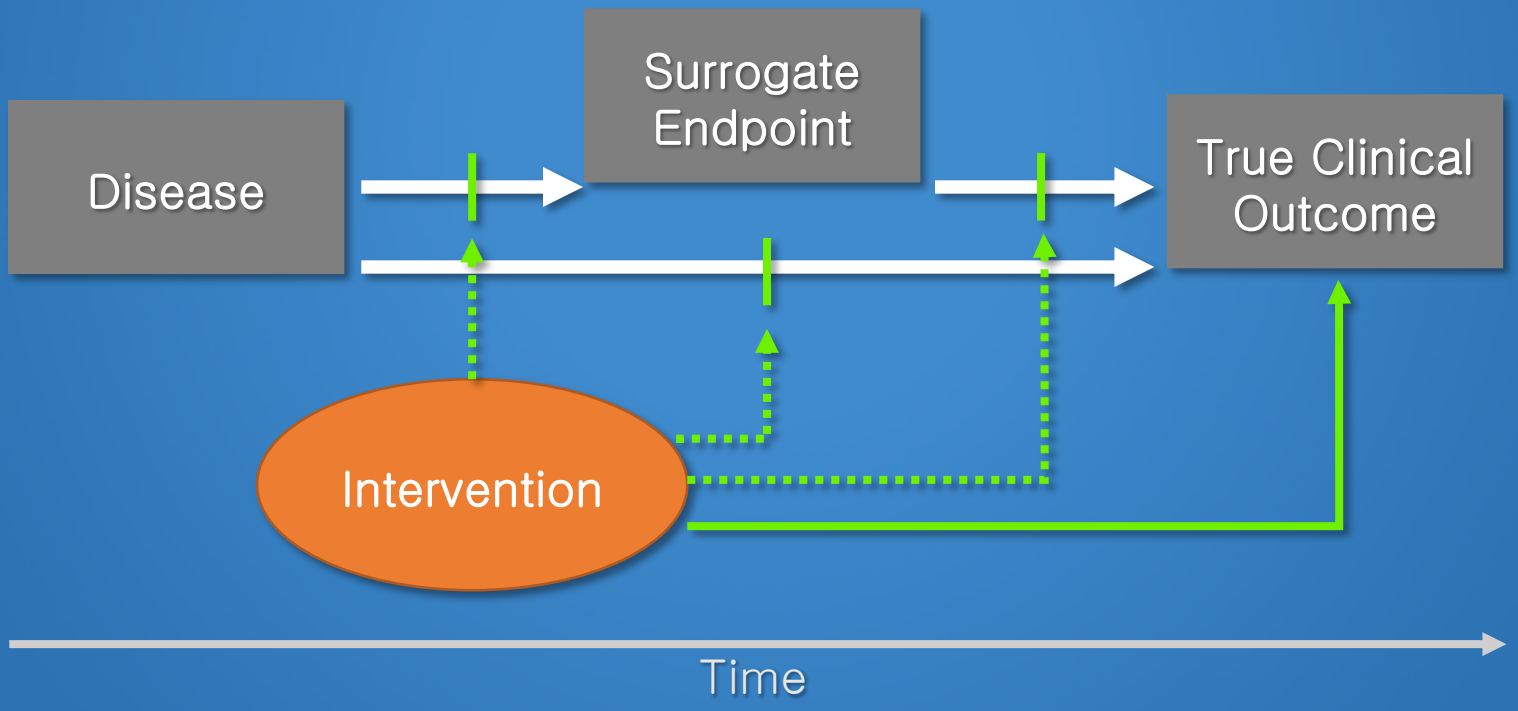
Scenario 2b: Inefficient Surrogate

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



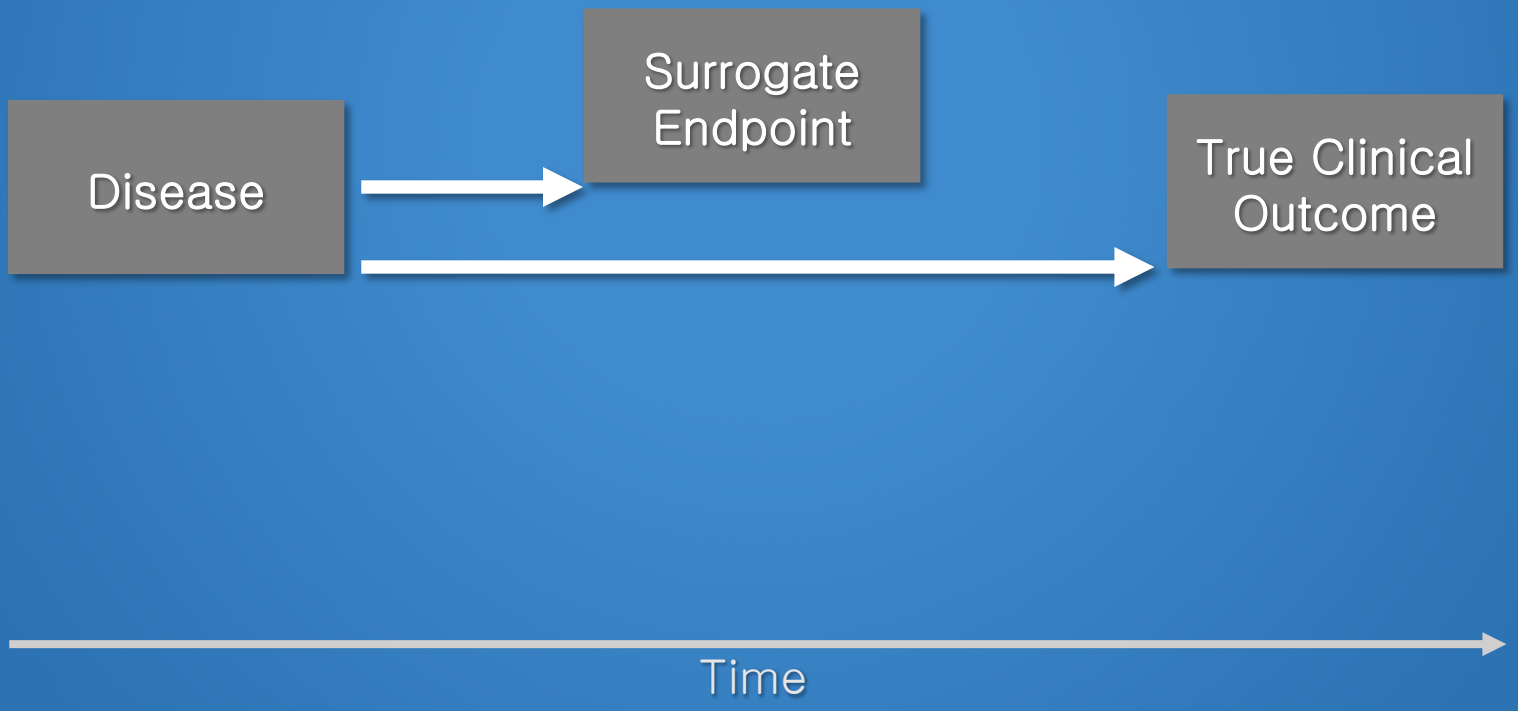
Scenario 2d: Dangerous Surrogate

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



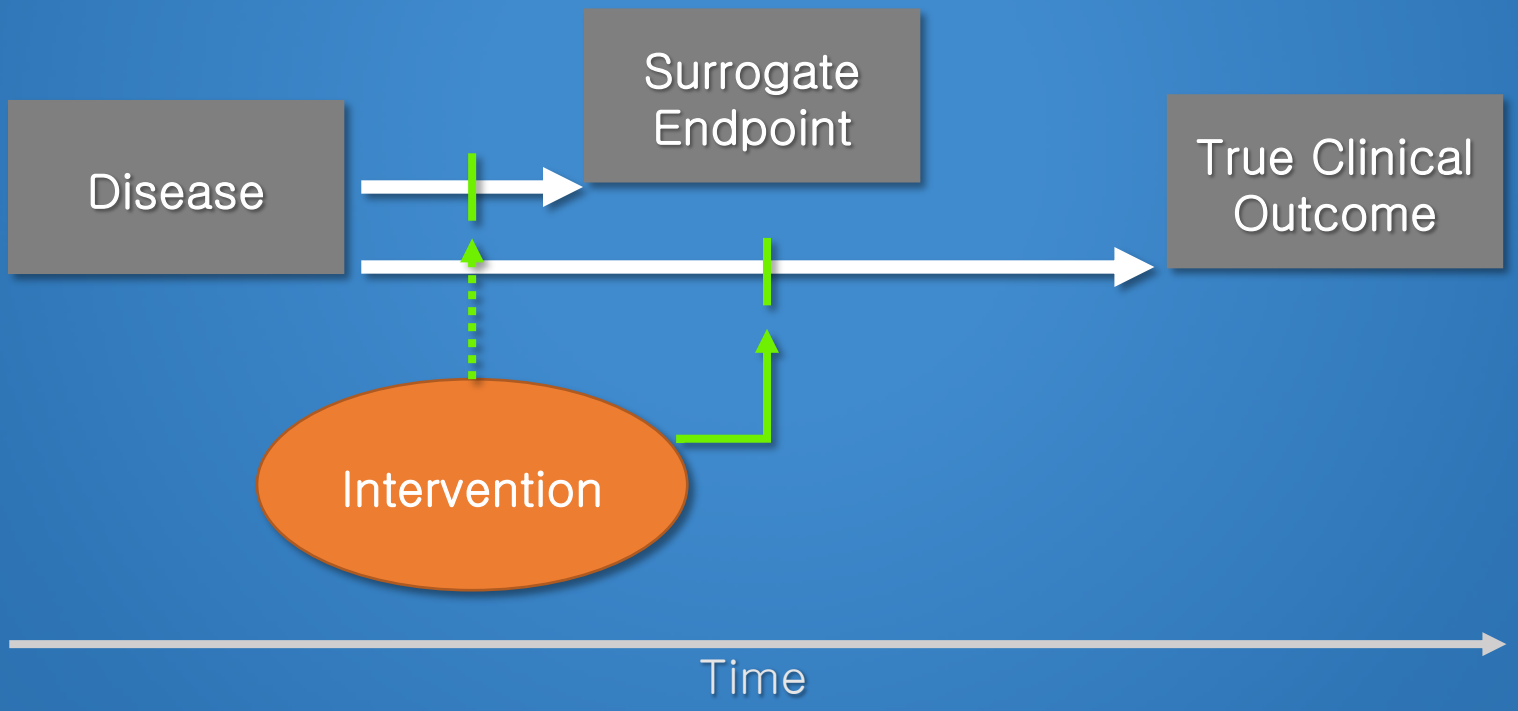
Scenario 3: Marker

- Disease causes Surrogate Endpoint and Clinical Outcome via different mechanisms



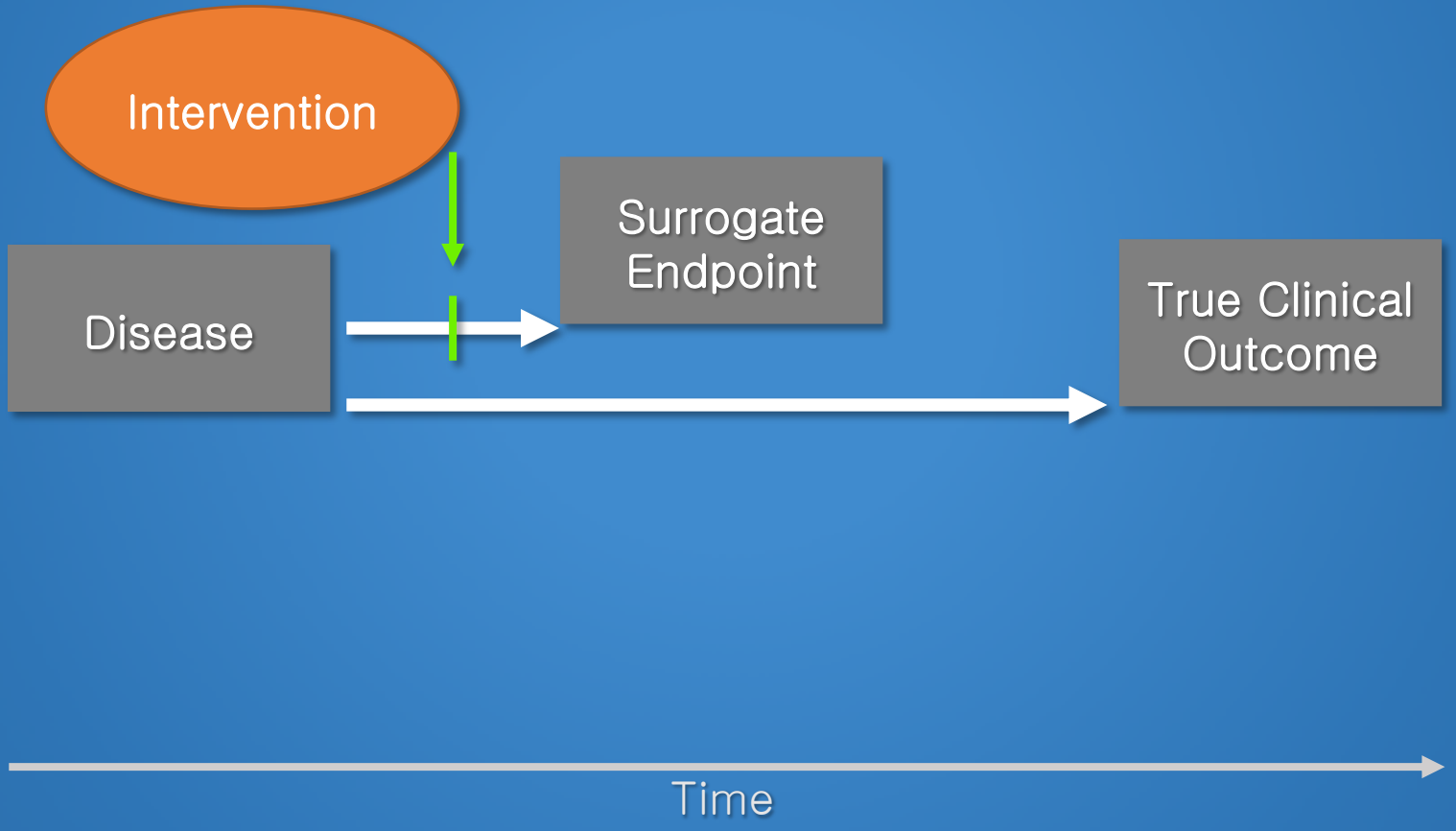
Scenario 3b: Inefficient Marker

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



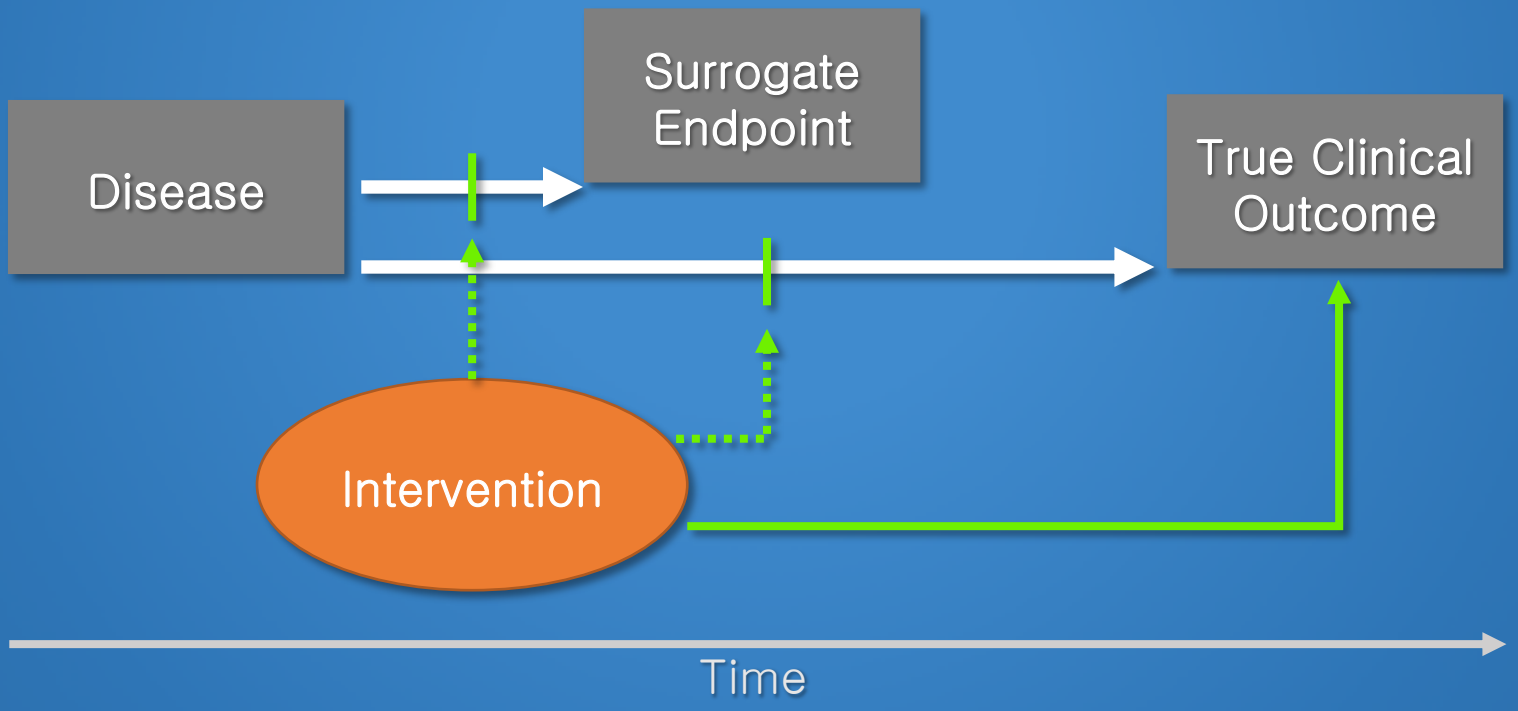
Scenario 3c: Misleading Surrogate

- 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)



Surrogate Outcomes

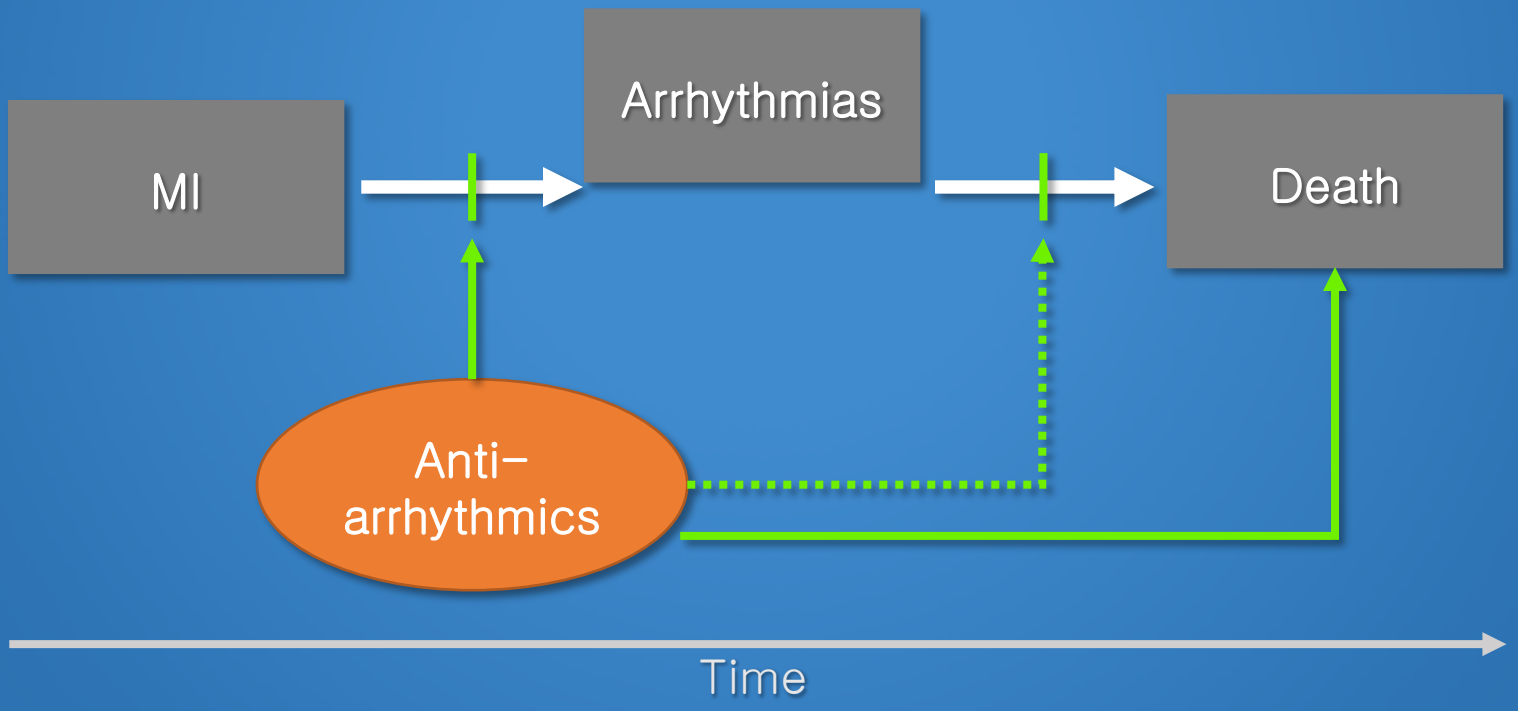
- Statistical and logistical constraints often lead to the desire for surrogate outcomes
  - But these have led us astray in the past
    - Illustration of the Problem (to follow)

Example: CAST

- Cardiac Arrhythmia Suppression Trial
  - Arrhythmia a risk factor for sudden death following a myocardial infarction
  - Anti-arrhythmic drugs (encainide and flecainide) successfully decrease incidence of arrhythmias
  - CAST
    - Placebo controlled trial using mortality as outcome
    - Encainide and flecainide TRIPLE the death rate

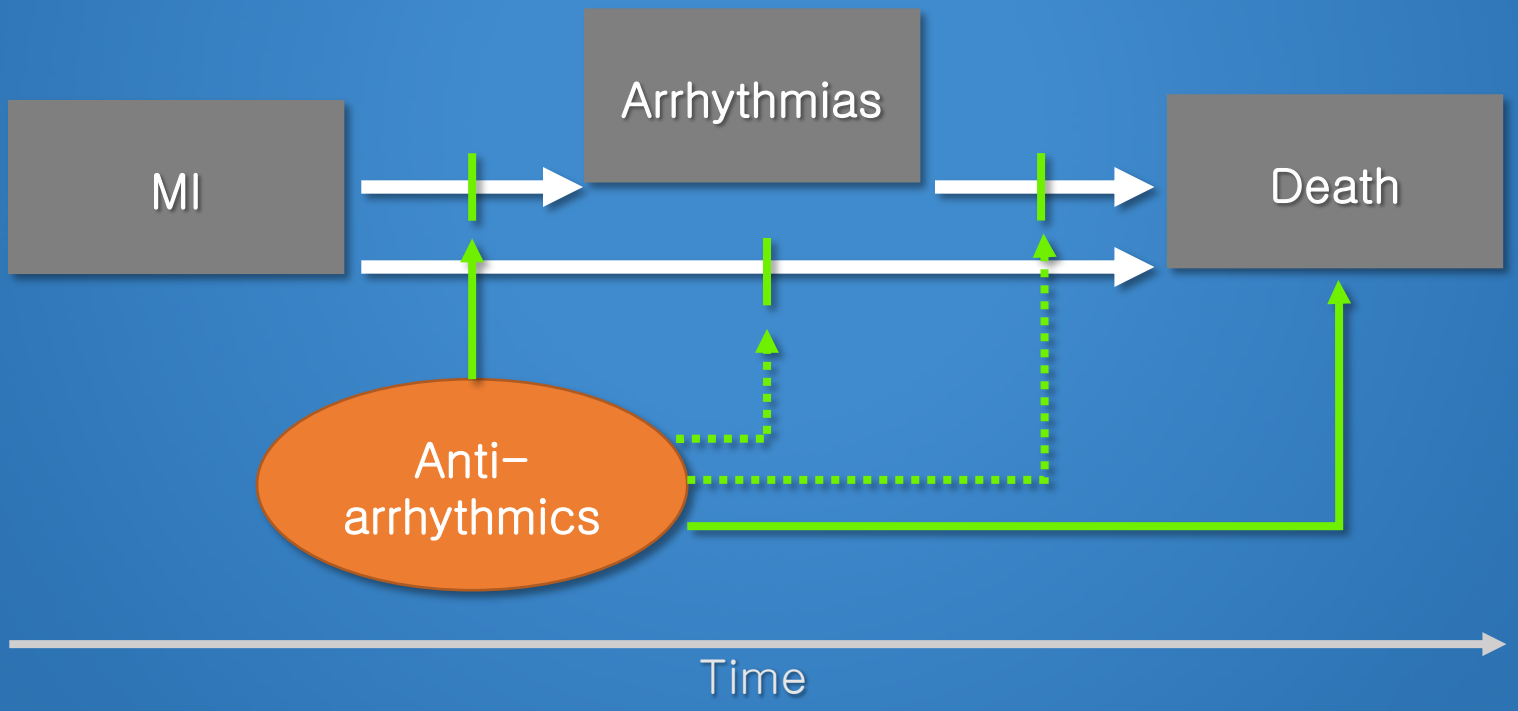
Scenario 1d: Dangerous Surrogate

- CAST



Scenario 2d: Dangerous Surrogate

- CAST

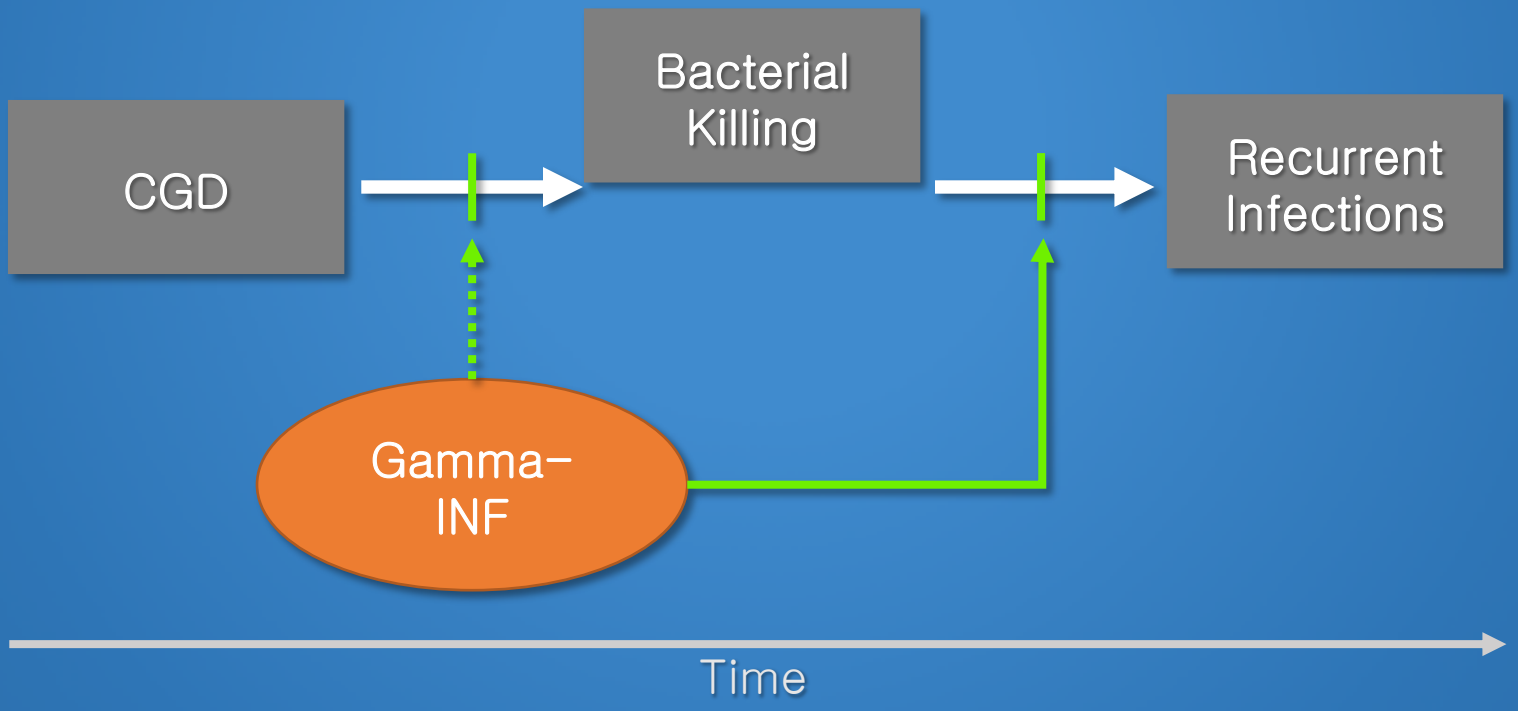


Example: CGD

- Chronic Granulomatous Disease (CGD)
  - CGD leads to recurrent serious infections
  - Gamma interferon increases bacterial killing and superoxide production?
  - International CGD Study Group of Gamma-INF
    - 70% reduction in recurrent serious infections
    - Essentially no effect on biological markers

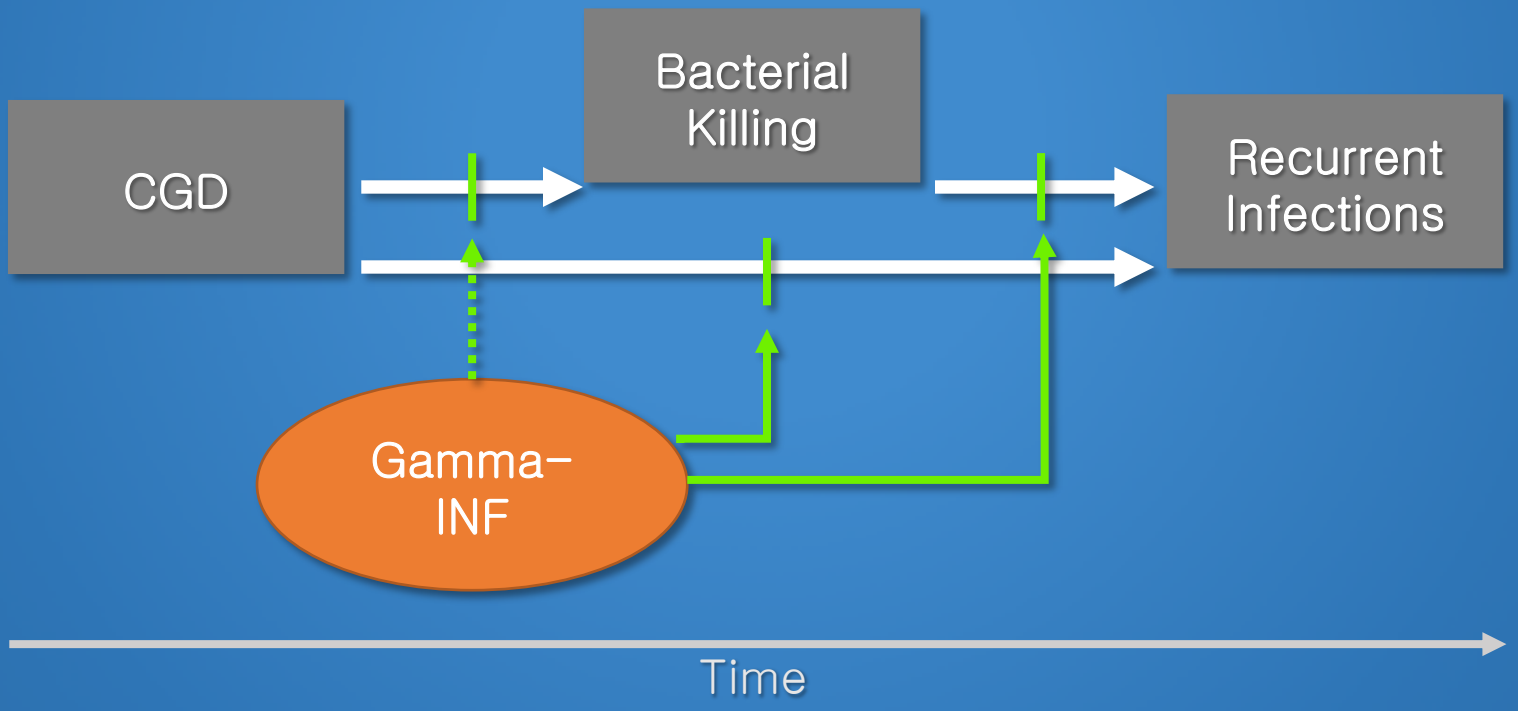
Scenario 1b: Inefficient Surrogate

- CGD



Scenario 2b: Inefficient Surrogate

- CGD



Surrogate Outcomes

- Validation
  - Many proposed fixes for surrogate outcomes revolve around “validation” of particular surrogate outcomes
    - This is generally very difficult to do
  - Question
    - Is there a way to validate a surrogate endpoint by establishing which causal pathway holds?

Be Careful

- It is not sufficient to establish that the surrogate endpoint predicts the clinical outcome in each treatment group separately
- Treatment can affect the distribution of the surrogate endpoint while increasing mortality in every level

Hypothetical Example

Surrogate	Treatment		Control	
	n	% die	n	% die
Low	30	50%	10	30%
Medium	40	60%	30	40%
High	30	70%	60	50%
Total	100	60%	100	45%

Example: CARET

- Beta-carotene supplementation for prevention of cancer in smokers
- Treatment group had excess cancer incidence and death
- Within each group, subjects having higher beta-carotene levels in their diet had better survival
  - Similar to hypothetical example scenario

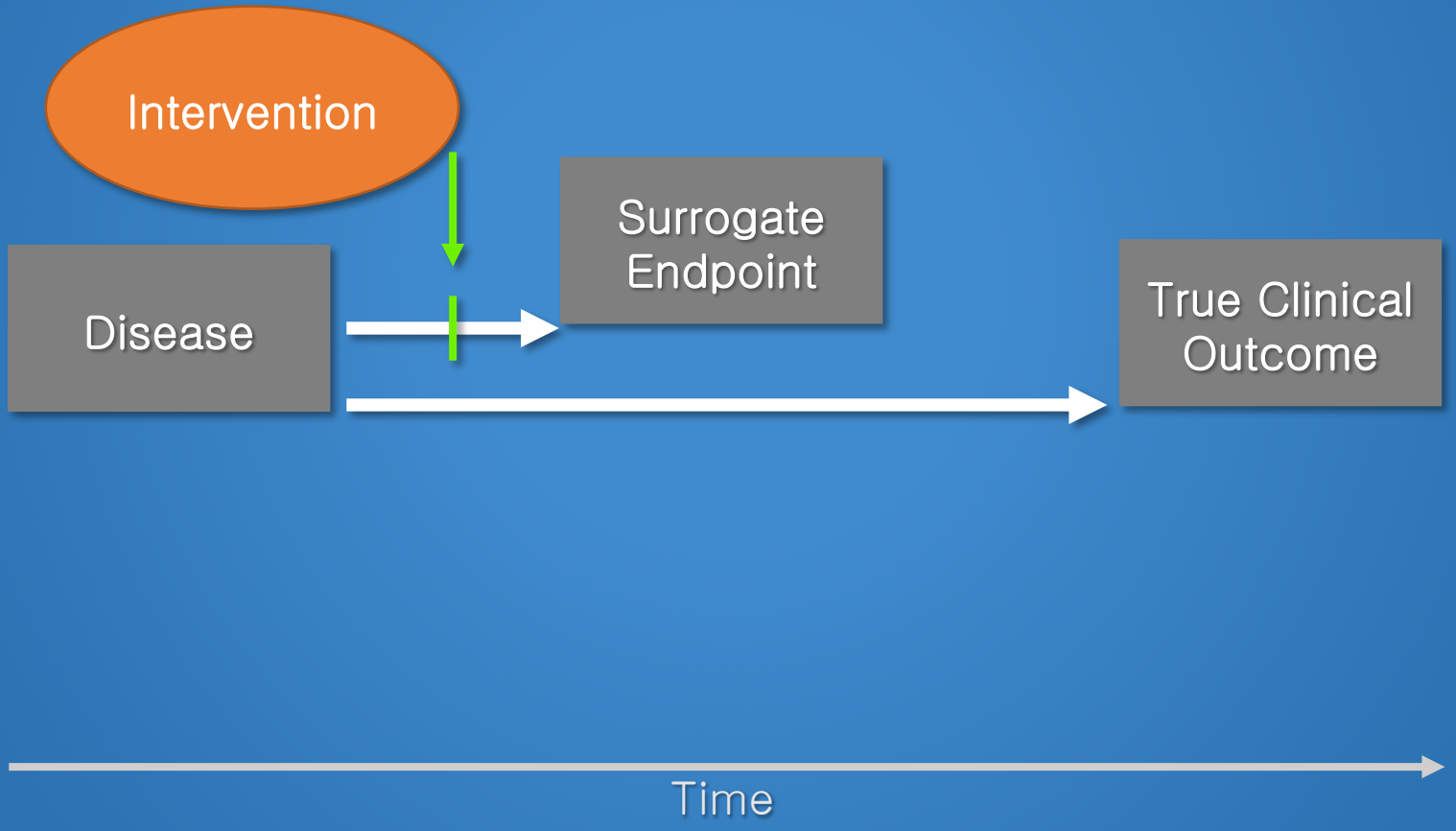
Prentice's Criteria

1. A surrogate endpoint must be correlated with the clinical outcome

2. A surrogate endpoint must fully capture the net effect of treatment on the clinical outcome
   - After adjustment for the surrogate endpoint, there must be no treatment effect on the clinical outcome
     - (Treatment effect mediated through surrogate)
     - (Would need to test pathways not through surrogate)

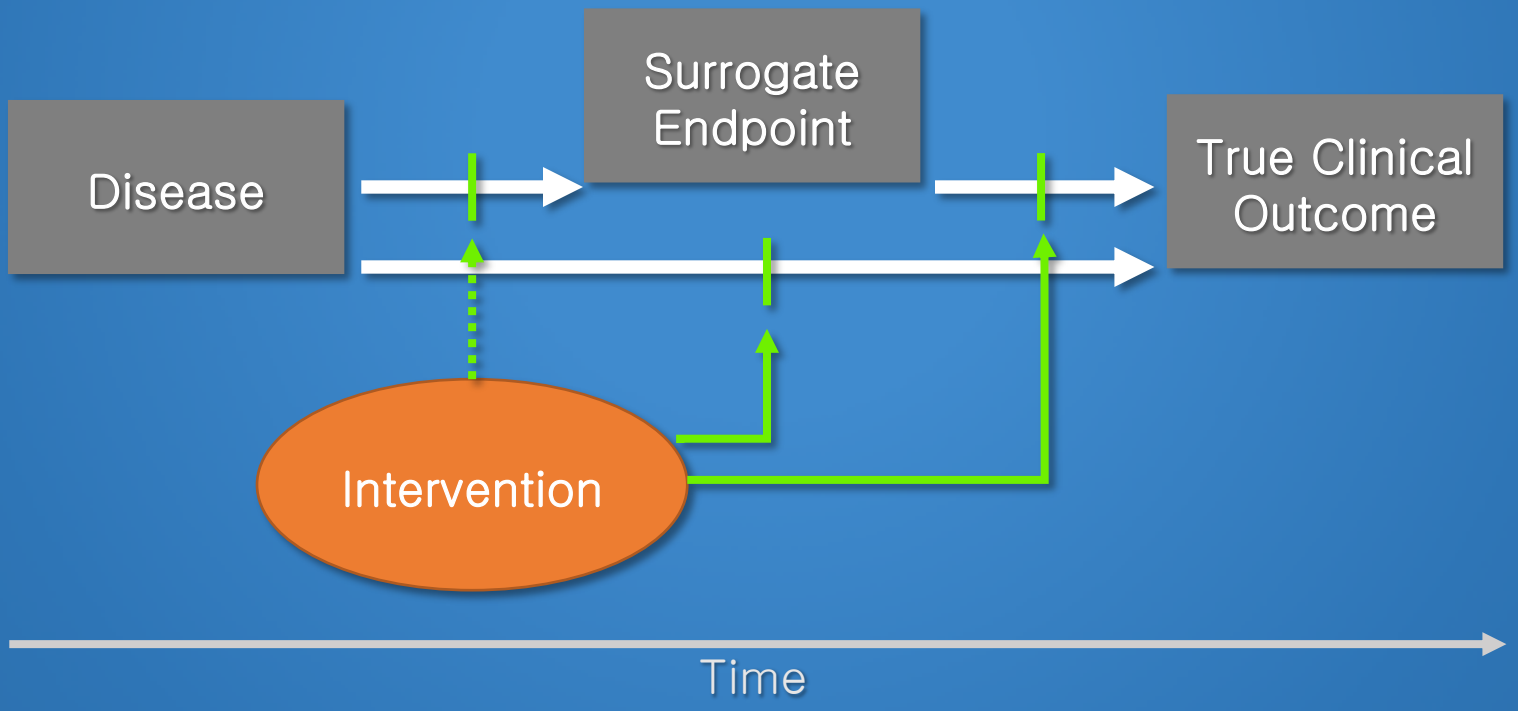
Does Not Satisfy Criterion #2

- Treatment has no effect on Clinical Outcome



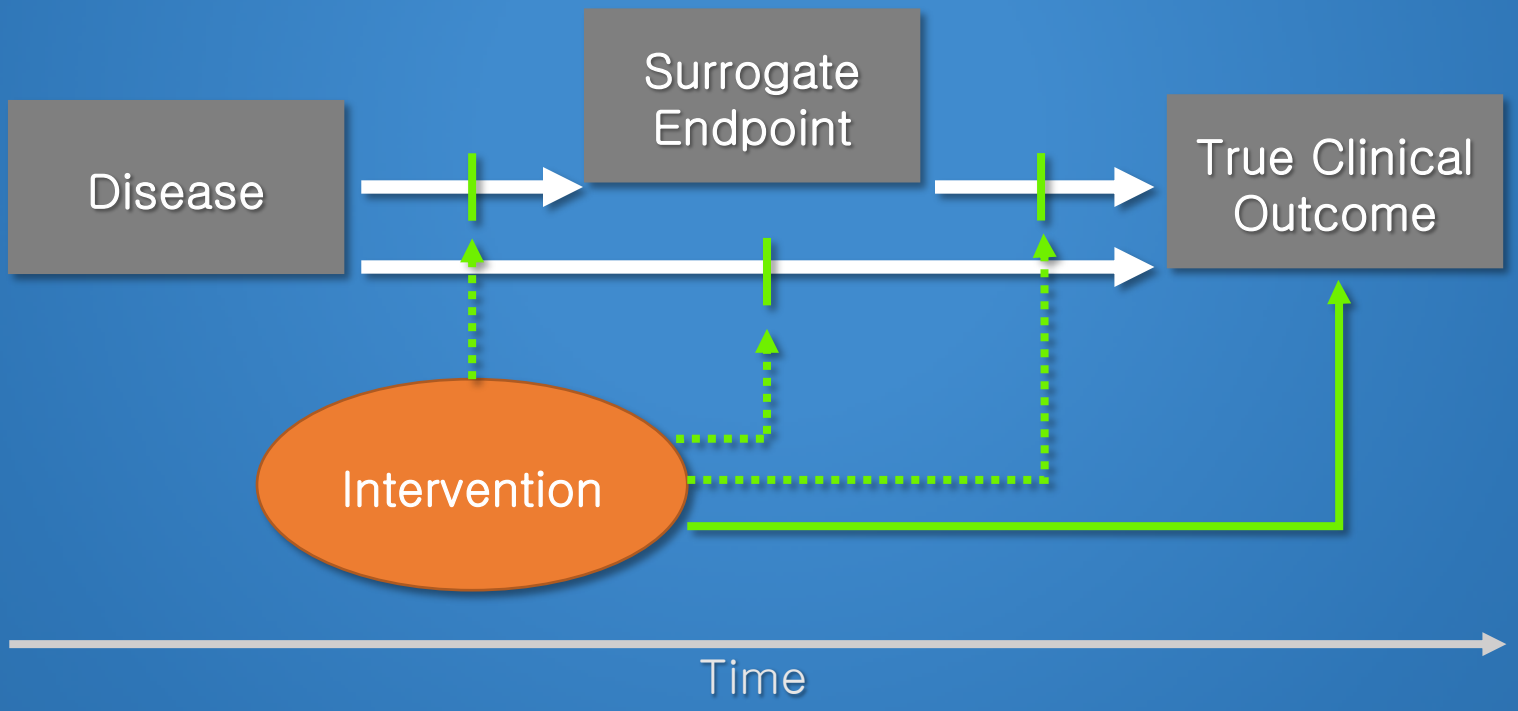
Does Not Satisfy Criterion #2

- Adjusting for Surrogate Endpoint will not capture all of Treatment effect



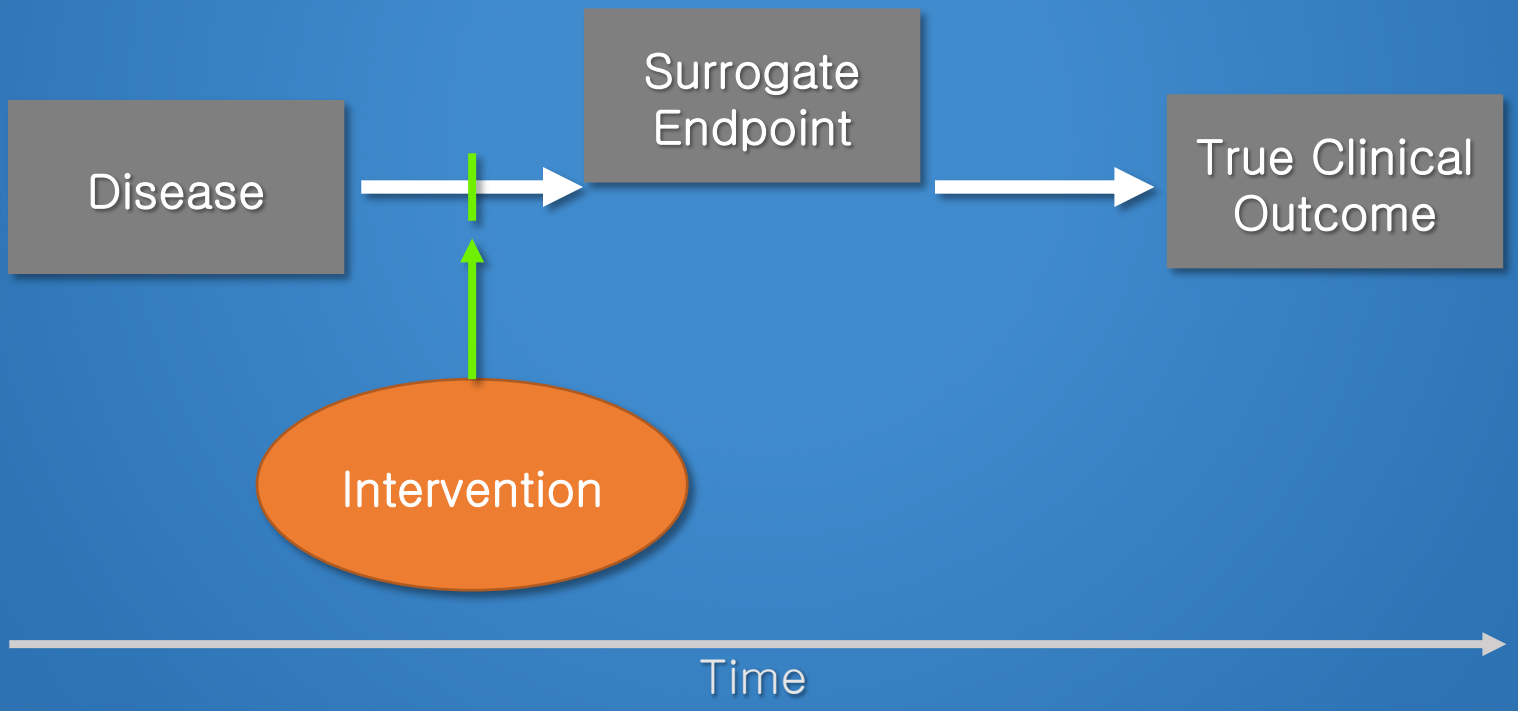
Does Not Satisfy Criterion #2

- Adjusting for Surrogate Endpoint will not capture all of Treatment effect



Satisfies Criterion #2

- Adjusting for Surrogate Endpoint will remove effect of Treatment on Clinical Endpoint



However…

- The validity of a surrogate endpoint is dependent upon the
  - Disease
  - Clinical outcome
  - Treatment

- Thus it is not possible to validate a surrogate endpoint for every combination of treatment and disease without doing a trial looking at the clinical outcome

Bottom Line…

- Surrogate endpoints have a place in screening trials where the major interest is identifying treatments which have little chance of working

- But for confirmatory trials meant to establish beneficial clinical effects of treatments, use of surrogate endpoints can lead to the introduction of harmful treatments

Summary

- Be careful using a (non-validated) surrogate as a replacement endpoint (outcome)
  - "A correlate does not a surrogate make."
  - Setting-specific:
    - Multiple causal pathways of disease process
    - Magnitude and duration of effects
    - Intended and unintended effects of interventions
  - Impact on public health:
    - Need reliable and timely evaluation (informed choice)

Fleming TR, DeMets DL. "Surrogate endpoints in clinical trials: Are we being misled?" *Annals of Internal Med* 1996; 125:605-613.

IOM, 2010. "Evaluation of Biomarkers & Surrogate Endpoints in Chronic Disease." Washington DC. National Academies Press

Fleming TR, Powers JH. "Biomarkers and Surrogate Endpoints in Clinical Trials." *Statistics in Medicine* 2012; 31: 2973-2984