Biostatistics 226, Winter 2018

Course Overview

Jan 9 – Design of RCTs: How to avoid bias and achieve valid results

Jan 16 – Analyses: Baseline, Efficacy, and Safety

Jan 23 – Monitoring trials for efficacy, futility, and patient safety

Jan 30 – Cluster Randomized Trials

Feb 6 – Precision medicine I: SMART (Adaptive) Trials

Feb 13 – Precision medicine II: N-of-1 Trials

Textbook References

Lawrence Friedman, Curt Furburg, David DeMets. *Fundamental of Clinical Trials, 5th edition*, 2015.
e-copy is available online to UCSF community

Vittinghoff, Glidden, Shiboski, McCulloch, *Regression Methods in Biostatistics*. 2012.
e-copy: http://www.springer.com/us/book/9781461413523

Conducting Valid Trials
The Critical (and Misunderstood) Roles
of Randomization and Complete Follow-Up

Society for Clinical Trials, May 15, 2016, Montréal, Québec

Presenters:
Thomas D. Cook, Department of Biostatistics and Medical Informatics, University of Wisconsin – Madison
Kevin A. Buhr, Department of Biostatistics and Medical Informatics, University of Wisconsin – Madison
T. Charles Casper, Departments of Pediatrics and Family and Preventive Medicine, University of Utah School of Medicine

What is a *Valid Trial?*

Definition …

> A *valid trial* is a trial that establishes, up to a predetermined level of *statistical certainty*, whether an experimental treatment is superior (or non-inferior) to a control treatment.

Importantly, the *sole* source of uncertainty allowed by this definition is *statistical,* by which we mean *non-systematic, random variation*.

Therefore, the goals of an investigator conducting a valid trial are to:

- limit the impact of random variation, in particular by ensuring that the trial has sufficient sample size
- *eliminate* (to the greatest extent possible) all other sources of uncertainty.

If *any* systematic sources of uncertainty remain, the validity of the trial is diminished, and whether the trial is salvageable will depend on a *subjective* assessment of the potential impact of the systematic error.

Alternative statement of the goal of the trial:

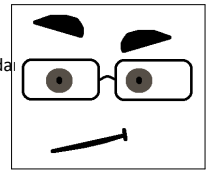
- to establish a treatment's effectiveness to the satisfaction of a *skeptic* who is not prepared to believe the treatment is effective without convincing evidence.

This goal explicitly puts the burden on the investigator/sponsor of the trial to conduct a trial that will be convincing, not just to themselves, but to the broader scientific community.

Study investigators & sponsors have an ethical obligation to

- patients in the trial
- future patients
- the scientific community

to conduct a trial according to the highest possible scientific standa



The skeptic

Conventions and Context

Clinical Trials

For the purpose of this course, we will consider trials in which we have

- two treatments (study "arms"): control (A) and experimental (B)
- a single primary outcome of interest
- each subject assigned a single treatment
- a statistical test of superiority (except as otherwise noted):

  Is treatment B *superior* to treatment A?

RCT Inference in Five Easy Steps

1. Enroll ***a group*** of subjects from the target population
2. ***Randomly assign*** individuals to either the experimental or control arm
3. Ascertain outcomes for each enrolled subject
4. *Statistically* compare the two treatments

… and the final step:

5. If, to the desired level of statistical certainty, *on average* the outcomes in the experimental treatment arm are *superior* to those in the control arm, then conclude that the experimental treatment is superior to the control treatment.

*Step 5 requires a **profound** leap of logic! … because …*

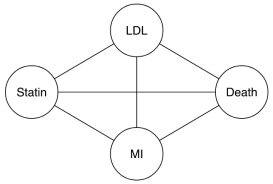
- Statistical comparisons are simply computations involving observed data from which we draw conclusions about the distributions of the underlying data.
- However, claims regarding the effectiveness of treatment are *inherently* about **causal relationships in the physical universe.**
  → Statistical analysis and causation are independent, unrelated concepts.

It is by the *magic* of randomization that we are able to make this leap.

Statistics versus Causation

Consider this diagram.

For a given dataset, statistics can assess whether there are associations among these four features.

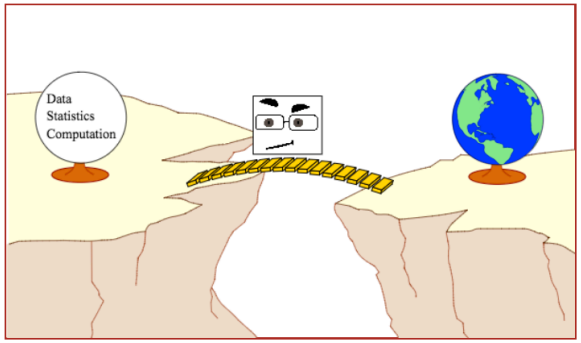


Add arrows to represent *causal pathways* by which treatment with a statin might provide benefit.

- Statistics quantifies strengths of associations. Statistics can't tell us *anything* about the *direction* of the causal pathways, or even that they're causal at all!
- Non-statistical evidence informs us of credibility of causality (eg, logic tells us that causes can only work forward in time; biology tells us …).

The Causal Bridge

Our skeptic is acting as the troll preventing anyone from crossing the chasm between statistical result and scientific interpretation *unless* their *study design and conduct*, combined with proper statistical analysis, allow for meaningful inference about how the world works.



Statistics versus Science

Statistical analyses fall into two broad categories …

- **Descriptive statistics**, where there is *no* attempt to draw conclusions beyond the sample at hand.
  - The mean value of variable *X* is 53.
  - 54% of our sample has *Y* = 2.
- **Statistical inference**, where the goal is to gain generalizable knowledge about the distribution of *X*s and *Y*s in the universe from which a sample has been drawn.
  - The mean value of *X* in the population is 53 with 95% CI (42,64).
  - The mean of *X* is larger when *Y* = 2 than when *Y* = 1.

Scientific inference also can fall into two primary categories …

- **Correlational** or **predictive** inference, where the goal is to identify characteristics that vary together in the population.
  - Older people have higher risk of mortality than younger people.
  - People with high HDL ("good" cholesterol) are at lower risk for cardiovascular disease than those with low HDL.
- **Causal inference**, where the goal is to infer that X causes Y.
  - Statins reduce risk of cardiovascular disease relative to placebo.

Correspondingly,

A ***statistical*** *hypothesis* is a claim regarding a statistical parameter.

For example, given two groups A and B, if β is the difference between the mean response in group A and the mean response in group B, some statistical hypotheses are:

- $\beta=0$
- $\beta\neq0$
- $\beta>0$
- $10<\beta<20$

A ***scientific*** *hypotheses* is a claim regarding the true state of nature.

For example, given two treatments, A and B and given a participant population, some scientific hypotheses are:

- Treatments A and B are equally effective
- Treatment A is less effective than treatment B
- Treatment A is more effective than treatment B by 10 units

A *valid trial aligns* the statistical and scientific hypotheses.

We reject:

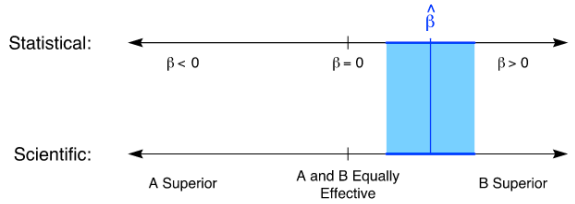
- H_0: $\beta=0$

if and only if we reject

- H_0: treatments A and B are equally effective

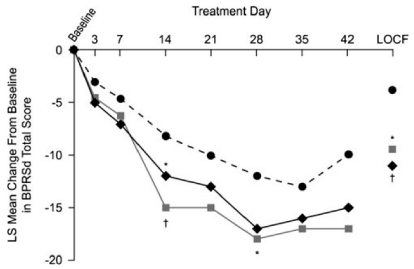
The scientific goal is to use data and statistics to assess the scientific hypotheses of interest. This requires a link between *statistical* and *scientific* hypotheses.

- Suppose we construct a confidence interval for β, the "treatment effect."
- We would like for there to be a corresponding "confidence interval" for scientific hypotheses as well.



Randomization coupled with *complete follow-up* and an *intention-to-treat analysis* is the only objective way in which statistical and scientific hypotheses can be aligned.

Example 1: 3-arm trial of change from baseline in BPRSd score at 6 weeks.
Ogasa *et al.* (*Psychopharmacology,* 2012)



Group	Number Randomized	Number (%) completing
Lurasidone, 40 mg/day	50	16 (32.0%)
Lurasidone, 120 mg/day	49	20 (40.8%)
Placebo	50	15 (30.0%)

Example 1 – Hope-Based Analysis

In this example, the extent of missingness compromises the result of the trial.

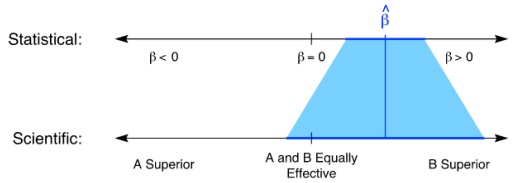
This analysis might be considered to be based on hope.

We make certain assumptions:

- LOCF reflects the 6-week outcome
- Observations are "missing at random," so the *statistical hypotheses* tested using the ANCOVA model align with the *scientific hypotheses* of interest.

. . . and we *hope* that they're correct.

Because a skeptic has no way to verify that these assumptions are correct, she might have good reason to conclude that **this is a completely failed trial.**



Example 1 – Convincing Analysis

So, what evidence does a skeptic want to see?

- A simple direct comparison of responses at 6 weeks
- for all randomized subjects
- analyzed according to their assigned treatment groups
- regardless of their adherence to their assigned treatments.

Otherwise, skeptic must trust unverifiable assumption(s) made by the investigator and/or analyst.

Our scientific conclusions require two steps:

1. Statistical inference: "Are the two groups different?"
2. Scientific inference: "Is the difference due to an effect of treatment?"

STATISTICAL INFERENCE

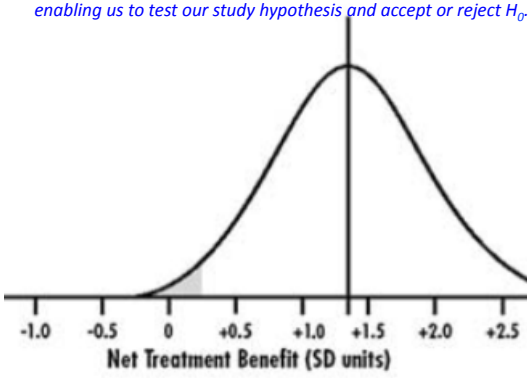
Study populations are inherently *heterogeneous* with respect to prognosis. Thus, arbitrarily chosen individuals have different baseline characteristics, different prognoses, and different outcomes. → *random variation*

- To characterize the outcomes across the population, we might take a sample and calculate its mean and variability.
- Because of heterogeneity, the magnitude *and* variability of the sample mean depend on the kinds of subjects we sample.

Impacts of eligibility criteria compared with heterogeneous samples:

1. Draw sample from a relatively homogeneous "healthy" subpopulation
   - Outcomes (say, FEV_1%) in the two groups are larger on average and less dispersed than a sample from the full population.
   - If we sampled from a subpopulation with a smaller average group difference, the observed difference is smaller.
2. Draw sample from a relatively homogeneous "sicker" subpopulation
   - Outcomes (FEV_1%) in the two groups are smaller on average, but still less dispersed
   - If we sampled from a subpopulation with a larger average group difference, the observed difference is larger.

Randomization guarantees that, under H_0, the distribution of outcomes is the same in both study arms. This establishes the critical value corresponding with the type-1 error rate, enabling us to test our study hypothesis and accept or reject H_0.



Hypothesis Testing

We want to *compare* mean outcomes of two arms, A and B.

- For a superiority trial, there are two *hypotheses* of interest:
  - **Null hypothesis, H_0:** The arms are *not* different (*e.g.,* the *true* average response is the same in both arms)
  - **Alternative hypothesis, H_A:** The arms are *different* (*e.g.,* the *true* average in one arm is larger than the average in the other)
- Before collecting any data, $\mathbf{H_0}$ is assumed to be true and the *statistical* goal is to show that $\mathbf{H_0}$ is false.

Typically, we set up the test to *reject* $\mathbf{H_0}$ if the difference between the group means is large compared to their *standard errors.*

- Statistics can only tell us *if* groups are or are not different.
- The *reason* they are different is a *causal* question and can't be identified statistically.

Failure to randomize can introduce confounding:

- If draw arm A (placebo) subjects from healthy subpopulation and arm B (treatment) subjects from unhealthy subpopulation, baseline health status will be associated with both arm and outcome. → "Simpson's Paradox"

Is it important to Balance baseline covariates ?

Beliefs in the importance of baseline balance have led some to propose "balancing strategies" as supplements or alternatives to randomization.

Arguments in favor of balancing strategies seem to …

- take it for granted that covariate balance, at least for "important" covariates, is desirable or even necessary
- make unsubstantiated claims that covariate imbalance biases or invalidates statistical analysis
- make the specific claim that a trial demonstrating a treatment difference in the presence of a baseline covariate imbalance is not credible: the covariate imbalance, rather than the treatment itself, might have caused the observed treatment difference

The question of interest is:

- In comparison to balancing strategies, does randomization *cause* biased or invalid statistical analysis?
- Specifically, does randomization **cause** detection of spurious treatment differences that would not have been observed **if only** we had employed a balancing strategy?

→ This is a causal question, and we can evaluate it in our causal framework.

Minimization – a dynamic method of allocation to study arms that aims to achieve balance in baseline covariates without using randomization

Table 5.7 shows the numbers of patients on each of the two treatments, A and B, according to each of the four patient factors. Thus, each one of the 80 patients appears four times in this table, once for each factor. Suppose the next patient is ambulatory, age < 50, has disease-free interval ⩾ 2 years and visceral metastasis (as indicated by the arrows in table 5.7). Then for each treatment one adds together the number of patients in the corresponding four rows of the table.

Thus, for A this sum = 30 + 18 + 9 + 19 = 76
while for B this sum = 31 + 17 + 8 + 21 = 77

Table 5.7. Treatment assignments by the four patient factors for 80 patients in an advanced breast cancer trial

Factor	Level	No. on each treatment A	No. on each treatment B	Next patient
Performance status	Ambulatory	30	31	←
	Non-ambulatory	10	9	
Age	< 50	18	17	←
	⩾ 50	22	23	
Disease-free interval	< 2 years	31	32	
	⩾ 2 years	9	8	←
Dominant metastatic lesion	Visceral	19	21	←
	Osseous	8	7	
	Soft tissue	13	12	

Example from Pocock SJ. *Clinical Trials – A Practical Approach*, 1973, Wiley.

Compare Minimization with Randomization with respect to error rates

1. Study 100,000 simulated trials with H_0 true to determine spurious false positive (FP) rates.
   - Overall, 8.5% (expected 5%)
   - Randomization 4.7%; Minimization 3.8% → Difference = 0.8%
2. Study 100,000 simulated trials with H_A true and 90% power to determine spurious false negative (FN) rates.
   - Overall, 0.79%
   - Randomization 0.35%; Minimization 0.44% → Difference = -0.09%

***Inefficiently* Improving Efficiency**

Normally, we use statistical techniques (eg, adjust for covariates) to improve efficiency:

- We want to increase power (*i.e.,* decrease type 2 error)
- while maintaining the same level of type 1 error (*e.g.,* nominal α=0.05)

However, minimization uses baseline covariates inefficiently:

- Power does increase
- But type-1 error rate decreases → The statistical test becomes too conservative.

Randomization and Baseline Covariates

We've established that randomization ensures the validity of the trial, in spite of any random covariate imbalances.

Some Unspoken Assumptions

We've made two key assumptions so far:

- Full exposure to interventions → All subjects fully adhere to their treatment and realize the full causal effect on their outcomes.
- No Missing Data → Outcomes are available on all subjects.

Missingness

Causal comparisons:

- The data are missing at random.
- The data are not missing at random and missingness isn't influenced by treatment.

Cannot claim causal association: If neither assumption holds, as we'd expect in most real-world trials, then causal bridge is cut. (*e.g.,* Scientifically nonsignificant difference may test as statistically significant. → Cannot claim difference is causal.)

SCIENTIFIC INFERENCE

Two possible outcomes for Paul following intervention/not on Jan 1, 2018:

- Paul receives heart transplant and dies 5 days later.
- Paul does not receive heart transplant and lives at least 5 days more.

→ *Did the transplant cause the death? Yes, for Paul.*

Two possible outcomes for Sue following intervention/not on Jan 1, 2018:

- Sue receives heart transplant and lives at least 5 days more.
- Sue does not receive heart transplant and lives at least 5 days more.

→ *Did the transplant have a causal effect? No, for Sue.*

Causation

In our context, we say that "X causes Y" if Y would have been different if X had been different.

Beth is a 45 year old female with a BMI of 29 and FEV_1% of 82 at baseline.

- If she's assigned to arm A,
  - her FEV_1% outcome at 6 weeks will be 79; her survival outcome will be "alive."
- If she's assigned to arm B,
  - her FEV_1% outcome at 6 weeks will be 93; her survival outcome will be "alive."
- If we could observe both arms in Beth, her treatment effect would be
  - B-A absolute difference = (93-82) – (79-82) = 14
  - B-A percent change from baseline = 14/82 x 100% = 17%

→ *The intervention has a causal effect on Beth's 6-week change in FEV_1% but not on her survival status.*

If we were omniscient, we would see both outcomes of every member of the heterogeneous population and could identify the causal effect of treatment for every individual.

The difficulty is that we only get to see what *actually* happened. We can never know

what would have happened if …

at least not for individual subjects.

Imagining Twins

Suppose our population is special: Every individual has a **perfectly identical twin …** baseline characteristics, observed and **unobserved**, exactly the same … even their hypothetical outcomes ("what would have happened if …") match precisely.

- We could assign Beth to A and her twin to B. The observed outcome difference between twins (93 – 79 = 14) is the causal effect (in either twin).
- We could do the same for all trial participants. The average causal effect is the average difference in outcome between twins; this is equal to the mean outcome on arm B minus the mean outcome on A.

Real Purpose of Randomization

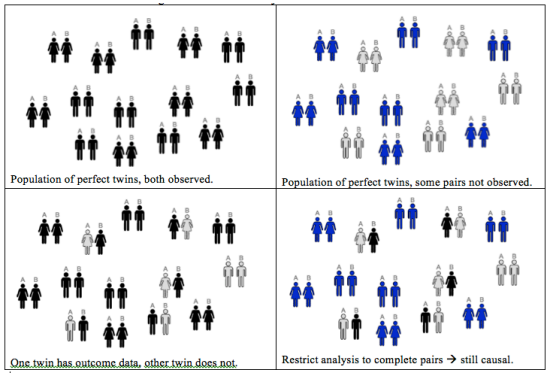
Since these don't exist, we approximate a population of perfect twins by applying **RANDOMIZATION** to a twin-less population.

- Randomization provides approximate balance in baseline characteristics:
  - this is neither *necessary* nor *sufficient* for the validity of our inference.
  - chance imbalances do *not* invalidate our conclusions; thus testing for balance is not needed or relevant.
- This only works at the population level: RCTs study *populations*, not *individuals*. Thinking too much about individuals can lead one astray.

Imaginary Perfect Twins

Each individual has a "perfect twin:" All characteristics match except the intervention and the associated outcome.

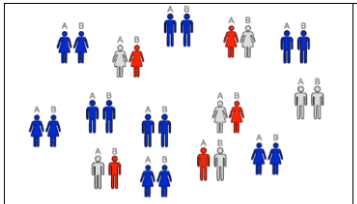
- Each pair generates a covariate-adjusted estimate of the treatment effect.
- Mean over pairs yields trial estimate.



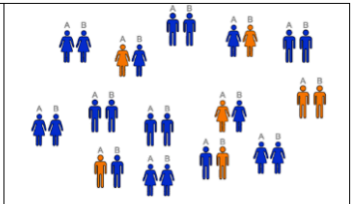
The Real World

We use randomization to approximate "perfect pairs."

- In complete data, group-wise results do this well, because *difference between means* equals *mean of paired differences*.
- In contrast to ideal world, in real world we have no way to identify paired individuals. If one "twin" has missing outcome data, we cannot exclude the whole pair in order to avoid confounding of missingness with causation.



The **complete-case analysis** combines a causal effect, based on paired observations (blue), and an arbitrary effect, based on unpaired observations (red). There is no way to identify which are "paired" and which are not, so we cannot test whether the arbitrary effect is biased and cannot adjust for it.



Imputation: We hope our guessed values (orange) are correct. However,

- differences can be biased in either direction,
- sample size can be falsely increased, and
- variability can be underestimated.

The Effect of Missing Data

Unfortunately, virtually all trials have some degree of missing data.

The assumption is frequently made that observations are *missing at random*.

- *If* this assumption is correct, the only consequence of missing data is that we have a smaller sample size and, therefore, lower power and precision. Otherwise, the analysis of the available data will give the same results as if there were no missing data.

Sometimes, it's acknowledged that observations are *not* missing at random, but it's assumed that the missingness is not influenced by treatment.

- *If* this assumption is correct, it can change the results of the analysis, but at least the analysis still has a causal interpretation.

The problem is that, by their very nature, missing data are unobserved and it is **impossible to know** whether they are actually missing at random.

- With non-random, treatment-dependent missingness a straightforward statistical analysis remains sound; however, its relationship to the scientific hypotheses may yield FP or FN conclusions.
- We cannot conclude there is a causal association between intervention and outcome.

Effect of Missingness on Type 1 Error Rates

1) Assume for simplicity that the rates of missingness are the same in the two groups, and let's consider two probabilities:

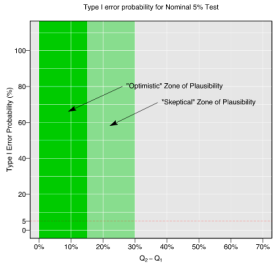
   Q_A=*Pr*(dead | missing, group A)

   Q_B=*Pr*(dead | missing, group B)

   If these two probabilities are equal, on average there is no net effect of missingness on the overall type 1 error rate.

2) Now consider the effect of a difference, $Q_B - Q_A$, on type 1 error rate. Imagine there are two "optimistic" (smaller differences) and "skeptical" zones (larger differences).

- The actual type 1 error rate increases as $Q_B - Q_A$ increases and the type 1 error rates may be seriously inflated, even within the "optimistic" zone.
- If missingness rates exceed 15%, nearly all reasonably sized trials can be seriously compromised.



Handling Missing Data

Common approaches when data are missing:

- Complete-case analysis → Not a valid causal analysis
- Imputation → Hope-based: Assumptions may be incorrect
  - LOCF: No causal basis
  - Single imputation: Worse than complete case
  - Multiple imputation
- Principal Stratification → No causal basis
- Joint modeling of outcome and missing indicator → No causal basis
- Inverse probability methods → No causal basis

Sensitivity Analysis ("Stress Test")

- We choose a primary analysis that is based on the best assumptions (maybe complete case, maybe imputation)
- We conduct a series of analyses that deviate from the *best* assumptions in a controlled way.
- We consider our results robust, *if* the conclusions are unchanged over a sufficiently wide range of plausible assumptions

Importance of Complete Follow-Up

- Missingness subverts randomization. The resulting naïve complete-case comparison is no longer causal.
- This affects all types of data (*e.g.,* safety, efficacy, adverse events, laboratory measures).
- The magnitude and direction of the error (the difference of the obtained effect from the real causal effect) cannot be known.

The Panel on Handling Missing Data in Clinical Trials …

> "… encourages increased use of [modern statistical analysis tools]. However, all of these methods ultimately rely on untestable assumptions concerning the factors leading to the missing values and how they relate to the study outcomes …
>
> … There is no 'foolproof' way to analyze data subject to substantial amounts of missing data." → "Substantial" amounts of missing data might be 10 or 15%.

All this applies to the importance of complete follow-up **after withdrawal from treatment** (i.e., despite lack of adherence).

Intervention Exposure: Incomplete Adherence

Unfortunately, non-adherence to assigned treatment occurs in virtually every clinical trial.

> Subjects are *non-adherent* if they are *assigned* a treatment but do not *receive* the intended dose.

Subjects may be non-adherent for lots of reasons:

- They took one dose, didn't like it or experienced a side effect, and stopped their study medication.
- They took their assigned dose for part of the intended follow-up period, but their disease worsened and they stopped their study medication.
- Study procedures became too onerous and they stopped their study medication.

→ Most of these reasons will probably result in adherence that depends on the assigned treatment.

Intervention Exposure: Treatment-Dependent Adherence

Recall Beth. Her FEV_1% would be 79 is she received arm A (placebo), and 93 if arm B (active). Her FEV_1% causal effect is 93 – 79 = 14.

If Beth adheres to B but not to A: Since A is placebo, receipt doesn't matter; we might *still* observe her causal effect of *assignment* of B versus A would be 93 − 79 = 14 → the same as the causal effect of *receipt*.

If Beth adheres to A but not to B: Beth might receive partial benefit from B … {79 if assigned treatment A; ~~93~~ 83 if assigned treatment B} for a causal effect of *assignment* to B relative to *assignment* to A is 83 − 79 = 4.

> The causal effect of *assignment* to B versus A is, in general, different than the causal effect of *receipt* of B versus A, but it is still a **real causal effect,** and it is the object of an ITT analysis.

We would prefer to see the causal effect of *receipt* of B versus A but cannot.

Intervention Exposure: Treatment-Dependent Adherence

Critics say non-adherence subverts the purpose of the trial. Their objection is certainly valid.

- The goal of the trial is to "estimate the treatment effect" on the outcome in question.
- Subjects who do not receive the treatment as intended do not benefit from it.
- Therefore non-adherence prevents us from correctly "estimating the treatment effect."

Full-Adherence Effect

What's the real problem?

- By "estimate the treatment effect" we assume they mean the "full-adherence treatment effect."
- If the goal is to estimate the "full-adherence treatment effect," then, unless we have achieved full adherence, we've done the **wrong experiment**!

Solutions:

A. The "correct" experiment strictly enforces full adherence. → In humans this isn't possible but a well run trial will come close to achieving this. Adherence information is critical to trial interpretation.

B. We *want* to recover the "full-adherence treatment effect" from a trial with incomplete adherence. However, we *cannot correctly guess* full-adherence values from partial-adherence values.

C. Conduct a Per-protocol analysis: Only analyze subjects who are fully adherent (or nearly so). This subgroup analysis …
   - … induces missing data (see Slide 26). We *cannot determine* if differences between groups are causally related to treatment or are induced by missing data.
   - … does not recover the full-adherence treatment effect.

D. Change the question!

***Intervention Exposure:* The Right Question**

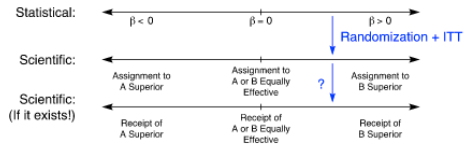
What's the "right" answer to the non-adherence problem?
The answer is to **change the question!**

- Unless we have full adherence, our causal conclusions **can't be** conclusions about the effect of *full adherence* to (receipt of) treatment.
- We *can*, however, reach valid conclusions about the effect of **assignment** to treatment.
- After all, in randomized trials in humans, we can randomly *assign* people to treatments, but we can't force them to *receive* treatments.

The best basis for assessing the causal effect of treatment, and that which will satisfy skeptics, is inference based on randomization.
- We've done a randomized experiment.
- If we have complete data, we can draw valid causal conclusions from our trial if we perform
  - A simple direct comparison of responses
  - for all randomized subjects (no missing data!)
  - analyzed according to their assigned treatment groups
  - regardless of their adherence to their assigned treatments.

***Intervention Exposure:* Assignment versus Receipt**

When we do not have full adherence, we actually have (at least) three sets of hypotheses:
1. statistical hypotheses
2. scientific hypotheses regarding the effect of **assignment to treatment**
   - Randomization, complete data, and ITT analyses ensure alignment of hypotheses 1 & 2.
3. scientific hypotheses regarding the effect of **receipt of treatment**.
   - Because we cannot observe full-adherence responses for some subjects, there is no *objective* way to ensure alignment of hypotheses 2 & 3; they can be "off" in either direction.



***Intervention Exposure:* Example 2** – Coronary Drug Project (CDP)

- 8341 patients enrolled between 1966 and 1969
- Males, <3 months since MI
- Exposures:
  - 5 active treatment groups (one was clofibrate) + placebo
  - 2:2:2:2:2:5 randomization (→ 1/3 randomized to placebo)
- Primary Outcome: all-cause mortality
- 8.5 year total study length

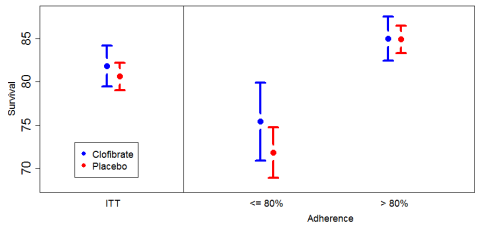
	Clofibrate	Placebo
	N (%)	N (%)
<80% Adherent	357 (32.4%)	882 (31.6%)
≥80% Adherent	708 (67.6%)	1813 (68.4%)
	1103	2789

Adherence rates were similar by arm, but we cannot assume (or check) non-adherence is independent of treatment.

***Intervention Exposure:* Example 2**

Regardless of arm, subjects who adhered to assigned treatment had better outcomes than those who failed to adhere.

- Commonly, subjects who adhere are *different* than subjects who don't.



Intervention Exposure: **Intention to Treat**

> "[According to the Intention to Treat principle,] all subjects meeting admission criteria and subsequently randomized should be counted in their originally assigned treatment groups without regard to deviations from assigned treatment." *L. Fisher et al (1990)*

The real reason we use ITT is this:

> ITT is the only analytical approach that relies solely on randomization to infer the effectiveness of treatment.

We do *not* use ITT because the ITT estimate is more realistic or more clinically relevant or more useful to calculate.

> We use ITT because **it is the only assumption-free, causal analysis we can do** with the experiments we can perform. It is the only analysis that can convince our skeptic.

Intervention Exposure: **Deviations from ITT**

Strict ITT Principle

A catchy phrase commonly used to refer to the ITT principal is "analyze as you randomize". This means:

1. A subject is later found to be ineligible → analyze him/her anyway
2. A subject doesn't receive treatment → analyze anyway
3. A subject receives wrong treatment → analyze in assigned (not received) group

Let's discuss each of these cases and whether we can ever deviate from the rules …

Case #1: Ineligible

When consider removing from the analysis, adhere to the spirit of ITT by reviewing every subject for a particular set of eligibility criteria in a systematic way:

- Pre-specify (a) which criteria and (b) the review process
- Conduct review in a blinded way (which is often harder than it sounds)

Intervention Exposure:

Case #2: Received no treatment

A modified ITT analysis excludes those who never started treatment.
The decision to remove a case should be made blinded to the treatment assignment.

Case #3: Randomization errors

(Refers to application of the list (e.g., transcription errors), not the list itself.)

- Different treatment arm than correct assignment:
  - Analyze as assigned, making note of the error for potential reporting & sensitivity analysis.
  - If the wrong item from the list was taken, retain that as originally specified.
- A batch of wrong treatments. Regardless of arm …
  - Keep all, or
  - Remove all who could have been in the bad batch, if blinding (esp of results) is secure.
- Stratified design, with wrong stratum recorded, led to "wrong" treatment
  - Randomization is still valid.
  - Analysis should use corrected stratum and assigned arm

Summary

The purpose of randomization is:

* to establish a causal relationship between an intervention and outcome	YES
* to "estimate the treatment effect"	NO
* to balance the baseline covariates	NO

Randomization is the only available method we have to determine whether a treatment has a causal effect on an outcome.

Threats to Randomized Controlled Trials

1. **Intervention exposure: Non-adherence and lack of ITT**. Randomization and intention-to-treat are tied together, and **both are required** for a valid causal inference.

   The only question that can be answered in a valid way, when participants are free to adhere or not, is the question of treatment *assignment*.

   Adherence information is critical to trial interpretation, but not to whether causal inference is possible.

2. **Outcomes: Missing data.** Can arise from incomplete follow-up, or be induced by excluding participants who are not adherent ("per protocol").

   Exclusions compromise the causal interpretation of the analysis of a trial. Analytic methods for dealing with missing data (including complete-case analysis) do not adequately restore the pure causal relationship.

   If the study design calls for outcome assessments beyond the intervention phase, these must continue even if exposure ends early.

To address the study goal:

- **Design the trial** using randomized assignments.
- **Execute the trial** with high fidelity to the study protocol, ensuring high levels of adherence and complete outcome ascertainment.
  - Example 3: VOICE Trial … http://www.nejm.org/doi/full/10.1056/NEJMoa1402269
  - Example 4: Aspire Trial … http://www.nejm.org/doi/full/10.1056/NEJMoa1506110
- **Analyze the trial** applying the intention-to-treat principle, using straightforward methods of comparing study arms.