Conducting Valid Trials

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

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

A note to participants: Our workshop will include a number of interactive demontrations and diagrams that haven't been reproduced here. Still, we hope that this transcript will be helpful, and we look forward to showing you the "missing data" on Sunday at SCT!

From Hardy (*Divergent Series*, 1949, as quoted by Ellenberg, 2014)

> [I]t does not occur to a modern mathematician that a collection of mathematical symbols should have a "meaning" until one has been assigned to it by definition.
>
> It was not a triviality even to the greatest mathematicians of the 18th century. They had not the habit of definition: it was not natural to them to say, in so many words, "by X we mean Y." … this habit of mind led them into unnecessary perplexities and controversies which were often really verbal.

All too often, decisions regarding design, conduct and analysis of clinical trials are made without defining what we need to achieve and how we know if we have achieved it.

The purpose of this workshop is to present a formal basis for making these decisions.

The Purpose of a Clinical Trial

Typically, the stated objective of a randomized clinical trial is to

> estimate the effect of treatment A
> relative to treatment B

This is **not the best objective.**

Meaning of "Treatment Effect"

What we mean by "treatment effect" is unclear.

- Are the subjects in the trial representative of the target population?
- Is the treatment administered in the same way as it will be in the target population?
- Is the outcome measured in the same way as it will be in the target population?

- Trials are designed around hypothesis tests with specified power. No mention of estimation in the sample size calculation.
- Finally, it's difficult to conceive of a universe in which the effect of treatment can be fully captured by a single number.

Note that we will, by necessity, frequently refer to the "treatment effect" or "effect size," but it must be kept in mind that (in our opinion) this is not a well defined concept.

Convincing a Skeptic

A better objective:

> to prove the treatment's effectiveness to a skeptic who is not prepared to believe the treatment is effective without convincing evidence

Why is this a better objective?

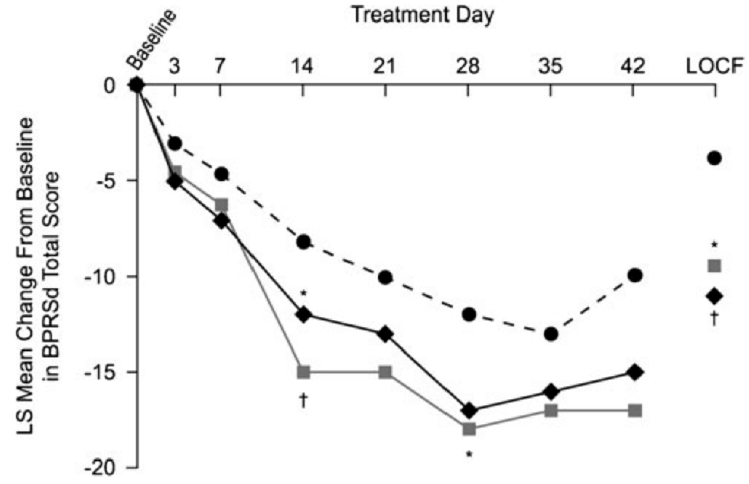
- It changes the question from "what is the treatment effect" to "is the new treatment better" which, we think, is a more well defined concept.
- It explicitly puts the burden on the investigator/sponsor of the trial to conduct a trial that will be convincing, not just to themselves, but to the broader scientific community.

- Too often, study sponsors are compelled to design, conduct and analyze trials in a way that they believe will maximize the chance that they will "win."
    - "The FDA approved my protocol, so it doesn't matter whether it satisfies the statistical *purist*."
- On the other hand, study 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 standards.

A Real-World Example

From Ogasa *et al.* (*Psychopharmacology,* 2012),



Primary outcome: change from baseline in BPRSd score at 6 weeks.

Treatment 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%)

- Primary outcome observed on less than half of randomized subjects.
- Primary analysis uses last-observation-carried-forward (LOCF) for subjects not completing study.
- The plotted trajectories are based on least squares means from an ANCOVA model, a hard-to-understand “black box”.
- Because we don’t observe the 6-week outcome on all randomized subjects, our “skeptic” has no way to know whether the active treatments are really better than placebo or whether the differences are due to lack of adherence and follow-up.

Hope-Based Analysis

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 ANCOVA model is valid,

and we *hope* that they’re correct.

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

Convincing Analysis

So, what would our 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.

Anything short of this requires our skeptic to trust in some unverifiable assumption made by the investigator/analyst.

> Over the next two hours, we will demonstrate why this kind of skepticism is justified for all trials, not just “completely failed” ones.

A Common Objection

The objection:

- How can the 6-week response for a subject who stops treatment at 2 weeks be important to our understanding of the effect of treatment?

Our response:

- First, clinical trials and statistical analysis provide inference about *populations* and not *individuals*. Non-adherent subjects are still important for understanding the effect of treatment within the *population*.[†]
- Second, even if a sponsor or investigator remains unconvinced, by not making every effort to collect this information, they deprive our skeptic (and the broader community) of their opportunity to make use of it.

[†] More on this later!

A Hypothetical Example

Suppose we conduct a hypothetical trial in COPD. Subjects with COPD will be assigned to

- placebo (treatment “A”), or
- experimental treatment (treatment “B”)

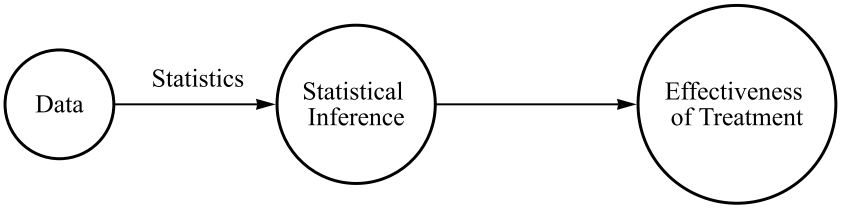
For simplicity, the outcome is percent predicted FEV_1 measured 6 months after baseline.

We collect information at baseline:

- age
- smoking status (current, former, never)
- height

- weight
- (baseline) $FEV_1\%$
- disease duration

The Purpose of Our Trial



Our scientific conclusions require two steps:

- Statistical inference: “Are the two groups different?”
- Scientific inference: “Is the difference due to an effect of treatment?”

Let’s talk about the role of statistics in this process.

The Role of Statistics

Heterogeneity

Here’s the hypothetical population we want to study in our trial.
Unfortunately, populations don’t look like this:
Study populations are inherently *heterogeneous*. This figure illustrates how prognoses vary over a hypothetical population.

Prognosis and Outcomes

Let’s imagine our heterogeneous population distributed along the *x*-axis of this graph, ordered by severity of the underlying condition (i.e., health status or prognosis).

The *y*-axis represents the expected “placebo” response. I.e., mean response if no treatment is administered.

Note that the *x*-axis represents subject status before treatment and is (at least in part) unobservable. The *y*-axis represents the measured outcome after treatment.

Because populations are heterogeneous, an arbitrarly chosen individual will have different baseline characteristics, different prognosis, and different outcome.

To try to characterize the outcomes across the population, we might take a sample and calculate its mean. This will vary from sample to sample.

The variability of the mean is dependent on the sample size, of course.

However, because the population is heterogeneous, both the magnitude *and* variability of the sample mean depend on the kinds of subjects we sample.

Comparing Groups

But we're not interested in describing placebo response. We want to *compare* treatments. Let's consider two treatment groups (not necessarily randomly assigned).

We can imagine two possible outcome curves, **one** representing placebo (treatment "A"), and **the other** representing the experimental treatment (treatment "B").

As before, we can take samples from each group and calculate the sample means.

On what basis should we compare them to see if they are different?

Hypothesis Testing

There are two *hypotheses* of interest:

- The *null* hypothesis says that the groups are *not* different. E.g., the *true* average response is the same in both groups.
- The *alternative* hypothesis says that the groups are *different.* E.g., the average in one group is larger than the average in the other.

Before collecting any data, the null hypothesis is assumed to be true and the *statistical* goal is to show that the *null* hypothesis is false. **This is the state of mind of our skeptic!**

A common approach is to *reject* the null hypothesis if the difference between the group means is large compared to their *standard errors.*

We note that were we to sample different kinds of subjects the observed responses differ in several ways, which can affect our hypothesis test.

What if we sample from a more homogeneous, "healthier" subpopulation? The responses in the two groups are larger on average and less dispersed.

Also, because we've sampled from a subpopulation that has a smaller average treatment difference, the observed difference is smaller.

Instead, what if we sample from a more homogeneous, "sicker" subpopulation. The responses in the two groups are now smaller on average but, as above, are less dispersed.

Also, because we've sampled from a subpopulation that has a larger average treatment difference, the observed difference is larger.

Note that we haven't made any claims about the *reason* the groups are or are not different.

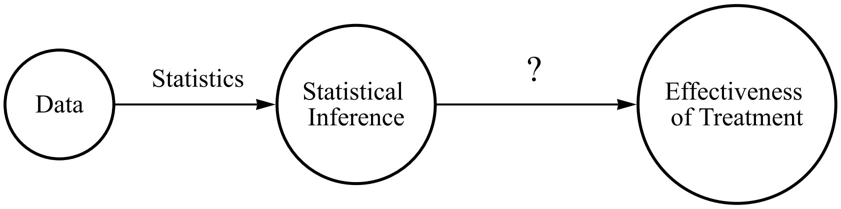
- Statistics can only tell us *if* groups are or are not different.
- The *reason* they are different is a *causal* question and can't be identified statistically.

In particular, suppose our samples are chosen in a treatment-dependent way (healthy placebo subjects and sick treatment subjects).

This is *confounding* — a baseline characteristic that is associated with both treatment and outcome.

The result of the (extreme) confounding in this example is a reversal of the effect of treatment, commonly known as “Simpson’s paradox”.

Again, the Purpose of Our Trial



But we *do* want to show that the reason for a statistical difference is treatment. We want to show that treatment *causes* a difference in outcomes. So, what do we do?

Causation

Causation

From Hernán and Robins (to be published in 2015):

Zeus is a patient waiting for a heart transplant. On January 1, he receives a new heart.

Five days later, he dies.

Imagine that we can somehow know, perhaps by divine revelation, that had Zeus not received a heart transplant on January 1, he would have been alive five days later.

Equipped with this information most would agree that ***the transplant caused Zeus’s death.*** *The heart transplant intervention had a causal effect on Zeus’s five-day survival.*

Another patient, Hera, also received a heart transplant on January 1. Five days later she was alive.

Imagine we can somehow know that, had Hera not received the heart on January 1, she would still have been alive five days later.

Hence ***the transplant did not have a causal effect*** *on Hera’s five-day survival. …*

In our context, we say that “X causes Y” if Y would have been different if X had been different. Causation is, at it’s heart, about the question:

> What would have happened if …

What would have happened if …

Here’s a “typical” human female in our trial.

We imagine that everyone has labels attached to them.

In this case we know that (at baseline) this is a 45 year old female with a BMI of 29 and FEV_1% of 82.

The labels marked “A” and “B” are special.

These represent the *outcomes* that we would observe if this subject were to receive either treatment A or treatment B.

Suppose that we assign this subject to treatment A, conduct the study, and observe the outcome, in this case “79”.

The quantity behind “Door B” remains hidden.

On the other hand, this subject could have been assigned treatment B.

In this case we would have observed a different outcome, “93” and the quantity behind “Door A” remains hidden.

If we were omniscient, we could open *both* doors, we would know that the “treatment effect” for B relative to A is $93 - 79 = 14$.

(Or maybe, $14/82 \times 100\% = 17\%$, if we define the “treatment effect” to be percent change from baseline.)

Maybe a secondary outcome is vital status, the person is assigned treatment “A” and they survive.

We don’t know if they would have survived if assigned B.

The subject may also have survived if they had been assigned treatment “B”.

In this case, we wouldn’t know if they would have survived had they been assigned “A”.

Of course, if we were omniscient, we could open *both* doors, we would know that there is *no* “treatment effect” on survival for B relative to A.

There is no causal relationship between treatment and survival in this subject.

In our heterogeneous population, if we were omniscient, we would see both outcomes of every subject and could identify the causal effect of treatment for every individual.

In Reality, We Aren’t Omniscient!

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 for the moment that we *aren't* omniscient, but our population is special.

Every individual has a **perfect twin.**

Perfect twins are more than identical. They are *perfectly* identical. Not only are all their baseline characteristics, observed and **unobserved**, exactly the same …

Even their hypothetical outcomes ("what would have happened if …") match precisely.

With such a population of twins, we could assign one twin to group A (on the left) and the other twin to group B (on the right).

Note that the observed outcome difference between twins ($93 - 79 = 14$) is the causal effect (in either twin).

And we could do this for everyone in our trial, creating a sample made up of pairs of perfect twins.

By definition, the average causal effect is the average difference in outcome between twins.

Mathematically, this is equal to the difference between average outcomes of the two groups (because the difference of averages is the average of differences).

In a population of twins, this makes a comparison of group means just as good as being omniscient.

In Reality, There Are No Twins!

Well, okay, there *are* twins (and one of us is married to one), but there are no identical subjects ("perfect twins").

Nonetheless, we can approximate a population of perfect twins by applying randomization to a twin-less population.

This only works at the population level.

> Randomized clinical trials study *populations*, not *individuals*. Thinking too much about individuals can lead one astray.

Randomization

Usual Rationale for Randomization

The typical rationale for randomization (from FFD)

> Randomization tends to
>
> - produce study groups comparable with respect to known and unknown risk factors,
> - removes investigator bias in the allocation of participants, and
> - guarantees that statistical tests will have valid significance levels.

This is largely a true statement (although it's not clear what "comparable groups" means).

However, it makes no *explicit* reference to *causation*.

Our skeptic would argue that the three features above are *side effects* of randomization and are neither necessary nor sufficient to ensure that we can make valid causal claims.

Real Purpose of Randomization

The act of randomization can be thought of as randomly choosing to open either Door A or Door B.

At a population level, this is the same as opening *both* doors …

… or (again, only at a population level) this is the same as studying a population of perfect twins.

This works because randomization takes *random subsamples* of the "A" and "B" twins, and random subsamples have the same distribution as the originals.

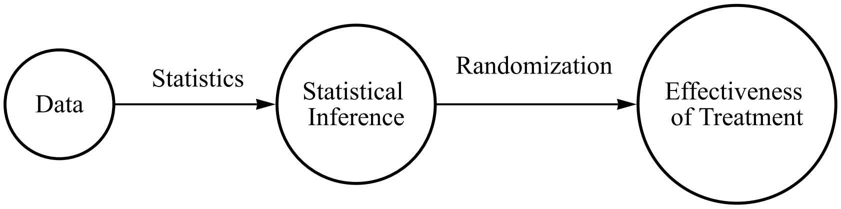
By the way, using this way of thinking about randomization, it is impossible for the other labels, age, sex, bmi, etc, to have any effect on the difference in outcomes between treatment groups.

That is, we randomly open either Door A or Door B, and the other labels just "come along for the ride."

Randomization tends to provide approximate balance among baseline characteristics, but this is neither *necessary* nor *sufficient* for the validity of our inference.

In particular, chance imbalances do *not* invalidate our conclusions.

Again, the Purpose of Our Trial



Now our trial can fulfil its purpose. **Statistics** allows us to establish that there is a difference between two groups. **Randomization** allows us to attribute this difference to the causal effect of treatment.

Some Unspoken Assumptions

We’ve made two key assumptions so far:

1. Outcomes are available on all subjects.
2. All subjects fully adhere to their treatment and realize the full causal effect on their outcomes.

Missing Data

The Effect of Missing Data

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

What impact will missing data have on our hypothetical trial?

The assumption is frequently made that observations are *missing at random* (we’re using this term informally and won’t define it).

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 the assumption is made that the missingness is not influenced by treatment. (Again, we won’t define exactly what we mean here.)

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.

Non-random, treatment-dependent missingness can easily subvert the benefits of randomization.

Straightforward analysis of such data no longer has a causal interpretation.

Missing Twins

As before, let’s imagine we sample from a population of perfect twins.

Now suppose there are some *pairs of twins* with a missing outcome.

If we include only complete pairs, we will still have a valid causal analysis.

Having missing pairs in a population of twins is analogous to “missingness is not influenced by treatment” in a randomized, twin-less population.

This is because, even if missingness depends on the attributes of an individual, if twins always go missing in pairs, then treatment (the only difference between twins) is not influencing missingness.

Partly Missing Pairs

But why would data on twins only go missing in pairs? We have to assume that, in some pairs, **one twin** has complete outcome data, but the other does not.

Again, if we include only complete *pairs* in our analysis, we will still have a valid causal analysis

However, this requires that we are able to identify “twins”, specifically the twin with **complete data** that correponds to the one with missing data, so the former can be removed from the analysis.

If we really were studying a population of perfect twins, this would be easy.

But we are using randomization to approximate a population of twins. If an individual in group A has missing data, there is no perfect twin in group B to remove!

Unmatched Twins

If we proceed to compare everyone in groups A and B anyway, we include in our analysis the differences between twins in the complete pairs (which *is* a causal effect), plus some extra twins in group A that don’t have a match, and some extra twins in group B that don’t have a match.

We have no reason to believe that the unmatched group A twins are comparable to the unmatched group B twins. We might *hope* that their differences will “cancel out” to give a causal effect, but this requires untestable assumptions.

Missingness in Our Trial

If the data are missing at random, there is no effect on the causal analysis.

If the data are not missing at random but missingness isn’t influenced by treatment, the analysis changes, but we are still making a causal comparison.

If neither assumption holds, as we’d expect in most real-world trials, then all bets are off.

Even under the null hypothesis (where there is no causal effect of treatment by definition), missingness can create a statistically significant difference.

Note that the statistics is working fine: there *is* a difference between these groups.

But the randomization has been subverted. The difference can’t be a causal effect of treatment.

Handling Missing Data

What can we do if we have missing data?

- Complete-case analysis
- Imputation
- Principal Stratification
- Sensitivity analysis

Complete-case Analysis

In complete-case analysis, we just analyze all subjects with complete data.

We've already seen that this is **not a valid causal analysis.**

Imputation

Imputation means "filling in" the missing values in some principled way (e.g., LOCF). Of course, this can only done be making untestable assumptions.

We have to make assumptions and **hope they're correct.**

Principal Stratification

Principal stratification means that, if we had perfect twins, we would include only complete pairs.

Since we don't have perfect twins, we use statistical models to simulate what would have happened if we did.

Again, we make assumptions and **hope they're correct.**

Sensitivity Analysis

A sensitivity analysis or "stress test" means:

- 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

Limited empirical evidence suggests that unless the result is *extremely strong,* if there is more than about 10-15% missingness, it won't withstand this stress test.

Incomplete Adherence

Non-Adherent Subjects

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.

Treatment-Dependent Adherence

Suppose that we have a subject whose FEV_1% would be

- 79 if they received treatment A
- 93 if they received treatment B

The causal effect of *receipt* of B relative to A is 93 − 79 = 14.

Now, suppose that the subject adheres if assigned treatment B but does not adhere if assigned treatment A.

We'll say this subject is a “B-adherer”, but not an “A-adherer”.

If A is placebo so receipt doesn’t matter, we might *still* observe their FEV_1% to be

- 79 if assigned treatment A
- 93 if assigned treatment B

and the causal effect of *assignment* of B versus A would be 93 − 79 = 14, the same as the causal effect of *receipt*.

On the other hand, if they are an “A-adherer”, but not a “B-adherer,” they might receive partial benefit from B and we might see

- 79 if assigned treatment A
- ~~93~~ 83 if assigned treatment B

The 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.

Note that there is no way for us to see the outcome (namely, 93) that *would have* occurred if the subject had *received* the full dose of B.

Therefore, **there is no way for us to calculate the causal effect of receipt, however much we might want to.**

Many investigators view non-adherence as subverting the purpose of the trial:

- 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 the treatment
- Non-adherence prevents us from correctly “estimating the treatment effect”

This objection is certainly valid.

Full-Adherence Effect

What's the real problem?

- By "estimate the treatment effect" we assume that our investigator means 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**!

The "correct" experiment is the one in which we strictly enforce full adherence.

At least in human trials, this isn't possible.

If we want to recover the "full-adherence treatment effect" from a trial with incomplete adherence …

… we somehow need to recover the full-adherence outcomes in subjects with partial or no adherence.

Unless we're omniscient, **we *can't* know what the full-adherence outcomes for these people *would have* been.**

Without somehow correctly guessing them, it is **impossible** to estimate the "full-adherence treatment effect."

One common proposed solution to this problem is to exclude non-adherent subjects and include in the analysis only those subjects who are fully adherent (or nearly so).

This is the so-called "per-protocol" analysis set.

What's the problem with the "per-protocol" analysis?

By excluding non-adherers, we have (deliberately!) induced missing data! In terms of our twin model, we have excluded some non-adhering twins, leaving some unmatched twins.

We have already seen that when data are missing like this, we *can't* know if the differences between groups are causally related to treatment, or induced by the missing data.

The "per-protocol" analysis *can't* possibly recover the full-adherence treatment effect!

Non-Adherence in Our Trial

The Right Question

If we have non-adherence, are we without hope?

What's the "right" answer to the non-adherence problem?

The answer is to **change the question!**

- We know that the best basis for assessing the causal effect of treatment 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.

- Unless we have full adherence, our causal conclusions **can't be** conclusions about the effect of *full adherence* to 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.

Based on the experiment we have done, we can draw valid conclusions about the causal effect of *assignment* to treatment 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.

which is precisely what our skeptic wanted to see.

Intention to Treat (ITT)

Intention to Treat

A definition:

> "[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)

Criticisms of ITT

Some researchers are outright *hostile* towards ITT (these are actual quotes):

> "Intention-to-treat analysis (ITT) more often than not gives erroneous results. These erroneous results are then reported as gospel, when in reality they are simply erroneous."
>
> "What if you believe in a more accurate way of presenting the data?"
>
> "When unbiased, intelligent people consider ITT, they cannot understand how it can be used by scientists trying to make sense out of their data, but, unfortunately, it is in almost every experiment."

> “Why, you may ask, could seemingly intelligent people do something so stupid as use ITT to evaluate data?”

More Criticisms

An objection that is less extreme but frequently heard:

> “But this patient didn’t adhere to treatment, so they can’t possibly provide information about the treatment effect. They are just adding noise and shouldn’t be in the analysis.”

Usual Rationale for ITT

These criticisms and objections may be partly a response to the typical rationale for ITT that is sometimes offered by its proponents:

- ITT assesses the benefit of a treatment **strategy**, rather than the biological effect of a particular treatment (the investigator **intended** to treat the subject)
- ITT estimates the overall clinical effectiveness most relevant to “real life” use of the therapy

While this may be true (although it is unclear how close the conditions in a trial mimic those of the real world), 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.

To reiterate, 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.

Example: Coronary Drug Project (CDP)

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

CDP Adherence

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

Some subjects did not take all of their assigned drug.

Some might argue that, since adherence was about the same in both groups, we can use per-protocol or as-treated analysis.

However, even if rates of non-adherence are equal, we **cannot assume** (hope) that non-adherence is independent of treatment.

CDP Analysis

ITT analysis showed no difference between groups.

Here is a summary of survival in the clofibrate group by adherence to treatment.

There is a large, highly statistically significant difference in mortality between people taking at least 80% of their assigned medication and those taking less than 80%.

This might suggest that if you take your medication, you will receive a mortality benefit, relative to those who only take part of their medication.

Here we see a difference *at least as large* between those taking at least 80% **of their placebo** and those taking less than 80%.

Either:

- placebo is highly effective at improving survival, or
- subjects adhering to placebo have better prognoses than those who fail to adhere.

Of course, it's most likely that subjects discontinue their placebo *because* they're doing poorly. That is, subjects who adhere are *different* than subjects who don't.

It is likely that the difference within the clofibrate group is driven by a similar phenomenon and is unrelated to any benefit of treatment.

We can see how this works in our hypothetical COPD trial.

An Unfortunately Common Practice

1. Plan to do "ITT" analysis
2. Will include all randomized subjects **with complete outcome data** in the analysis, as they were randomized
3. Subjects do not adhere to treatment

4. **Stop data collection** when subjects fail to adhere
5. Can’t include them in the analysis, because the outcome is missing
6. **A per-protocol analysis is sold as ITT**

Summary

Randomization

The purpose of randomization is **NOT**:

- to “estimate the treatment effect”
- to balance the baseline covariates

Neither of these is required for for valid causal inference.

> The purpose of a randomized trial is to establish a causal relationship between an intervention and outcome.

In order to determine whether a treatment has a causal effect on an outcome, we must either:

1. Be omniscient
2. Have a population of perfect twins
3. **Randomize**

Randomization is the only one of these that we can do in the real world.

Missing Data

Missing data compromise the causal interpretation of the analysis of a trial.

Analytical methods for dealing with missing data (including complete-case analysis) do not adequately restore the pure causal relationship.

We cannot emphasize strongly enough that

> **WITHDRAWAL FROM TREATMENT ≠ WITHDRAWAL FROM STUDY**

There must be a process in place for withdrawing from treatment without withdrawing from data collection.

Withdrawal from data collection should be **extremely** rare.

Follow and collect data on all subjects, regardless of adherence to treatment.

Non-adherence and ITT

Randomization and intention-to-treat analysis are tied together, and **both are required** for valid causal inference.

The only question that can be answered in a valid way is the question of treatment *assignment*.

No analysis can give a valid answer to the question

> “If you take your medicine, does it work?”

unless you force subjects to adhere to their assigned treatment or make unverifiable assumptions.