Prominent Design Issues in Clinical Research

Al M Best, Ph.D.
Virginia Commonwealth University

Task Force on Design and Analysis in Oral Health Research
Satellite Symposium, AADR

Boston, MA: March 9, 2015

Outline

- Core Concepts
- Prominent Design Issues in Clinical Research
  - Randomization
  - Masking/blinding
  - Analysis: Intention-to-treat or Per Protocol?
  - Biomarkers and surrogate measures
  - Superiority, equivalence, or non-inferiority

Randomization: What it is

- Randomization: assignment of treatments to patients (equivalently, patients to treatments) based on a chance mechanism of known characteristics
- Can take many different forms, all acceptable
  - The simplest is a coin-flip for each patient
  - A computed “pseudo-random-number” process is probably psychologically better for the patient
- Randomization can be purposely unbalanced
  - Example: 2/3 patients on experimental therapy

Randomization: What it is

- Randomization is often not entirely random, but constrained in some way to promote better balance
- Stratified randomization: Different subgroups of patients may be randomized separately with protection against major imbalances across groups
- Stratified randomization without constraints is the same as simple randomization. It must be constrained by blocking to do any good
  - Randomize to even group size within blocks
  - The block size must not be known or predictability ensues!
Randomization: Why?

- What it accomplishes
  - Virtually eliminates opportunities for intentional or inadvertent skewing of patient allocation to favor a treatment
  - Eliminates other selection biases of all sorts affecting treatment comparisons, period!
  - Eliminates reverse causation, period!

Randomization: Why?

- Tends to distribute patients across treatments more equitably than humans especially in large studies - even when we systematically try to make groups similar
- This applies to characteristics we measure, and also to any behind-the-scenes characteristics, e.g., lurking variables, that could produce confounding in ways we are unaware
- So protects against confounding:
  - but not entirely
  - more so in larger, less so in smaller studies
- Can provide an assumption-free basis for statistical analysis

Randomize, But

- What it can’t do

Randomization can’t assure that the actual treatment groups in a particular trial are comparable at baseline.
  - The laws of chance that produce the tendency towards equitable allocation ALSO
    - guarantee inequities will occur in some studies, especially if many characteristics are considered
    - guarantee important inequities will occur more frequently in small than in large studies
- Randomization’s benefits vanish if it is performed before patients enter the trial, or when patients are not committed to its results

Randomize, But

- What it can’t do

Randomization gives no help whatsoever with regard to
  - measurement bias
  - placebo effect
  - Hawthorne effect
  - regression to the mean
- So other tools are needed to cope with these, and possibly to deal with imbalances/inequities between randomized groups
Hi. You’ve been randomly selected to participate in a sex survey upstairs in fifteen minutes

Blinding, AKA: Masking
- Masking and Blinding refer to concealment of the randomized intervention received by a patient. Who may be blind:
  - Case/patients/participants
  - Interventionists, those treating participants
  - Those measuring outcomes: Clinicians and technicians who do not treat case/patients, but are involved in evaluating their outcomes
  - Investigators involved in decision-making about policies during the trial, and about statistical analyses to interpret the resulting data

Blinding: How to describe
- Thus, in principle trials can be single, double, triple, or quadruple-masked/blind
- Single masking/blinding always refers to concealment from case/patients
- Beyond that, there is no standard terminology.
  - Double-masked/blind is a carry-over term from a time when treating and evaluating clinicians were not differentiated, and when investigators were never blinded
  - The current usage is ambiguous

Double Blind
- “Blinding is intended to prevent bias on the part of study personnel.
- The most common application is double-blinding, in which participants, caregivers, and outcome assessors are blinded to intervention assignment.”

**Blinding: How to describe**

- Be explicit
  - Communicate which of the trial participants were masked, and **how** treatment was concealed
  - Describe what the blinding accomplished
- If not blind, why not?
  - Clearly, some types of masking can be impractical, inconvenient, expensive, ineffective, and/or even unethical depending upon the specific treatment and setting
  - However, exactly when such constraints apply is highly controversial

---

**Sham Surgery**

- Background Many patients report symptomatic relief after undergoing arthroscopy of the knee for osteoarthritis, but it is unclear how the procedure achieves this result. We conducted a randomized, placebo-controlled trial to evaluate the efficacy of arthroscopy for osteoarthritis of the knee.
- Methods A total of 180 patients with osteoarthritis of the knee were randomly assigned to receive arthroscopic débridement, arthroscopic lavage, or placebo surgery. Patients in the placebo group received skin incisions and underwent a simulated débridement without insertion of the arthroscope. Patients and assessors of outcome were blinded to the treatment group assignment. Outcomes were assessed at multiple points over a 24-month period with the use of five self-reported scores — three on scales for pain and two on scales for function — and one objective test of walking and stair climbing. A total of 165 patients completed the trial.
- Results At no point did either of the intervention groups report less pain or better function than the placebo group. For example, mean (±SD) scores on the Knee-Specific Pain Scale (range, 0 to 100, with higher scores indicating more severe pain) were similar in the placebo, lavage, and débridement groups: 48.9 ± 21.9, 54.8 ± 19.8, and 51.7 ± 22.4, respectively, at one year (P=0.14 for the comparison between placebo and lavage; P=0.51 for the comparison between placebo and débridement) and 51.6 ± 23.7, 53.7 ± 23.7, and 51.4 ± 23.2, respectively, at two years (P=0.64 and P=0.96, respectively). Furthermore, the 95 percent confidence intervals for the differences between the placebo group and the intervention groups exclude any clinically meaningful difference.
- Conclusions In this controlled trial involving patients with osteoarthritis of the knee, the outcomes after arthroscopic lavage or arthroscopic débridement were no better than those after a placebo procedure.


---

**Blinding the participant**

- Reduces potential for confounding by placebo and/or Hawthorne effects, by promoting equal benefits from placebo effects, and equal Hawthorne effects, across treatment groups
- Reduces potential for confounding by drop-outs or patient-stimulated treatment switches based on subjective failure or side effects (which often occur at similar rates in actively treated and placebo patients)
- These benefits are hard to gain in any other way
  - Totally objective outcome criteria help, but aren’t sufficient since objective outcomes also respond to placebos

---

**Blinding interventionists**

- Directly, reduces the potential for confounding by discretionary treatment differences that might be influenced by therapy
- Indirectly, further reduces potential for confounding by drop-outs or treatment switches, since clinicians can stimulate either based on subjective failure or side effects
- Indirectly, further reduces potential for confounding by placebo and/or Hawthorne effects, since contacts with clinicians impacts patient perceptions, e.g., about prognosis, and can inadvertently unmask the patient
Outcome measures blinded to group membership

- Directly and totally protects against “diagnostic suspicion bias,” a measurement bias when evaluation is skewed by treatment-influenced expectations
- Protects against indirect consequences of such bias from communication of skewed intermediate evaluations back to the study investigators, treating clinicians, and patients during conduct of the trial while they can influence
  - further treatment decisions, and
  - whether the trial continues

Blinding investigators

- Masking Study Principal and Other Investigators
  - Protects operational decisions from biasing the study due to over interpretation of events that happen to occur (randomly) in conjunction with the experimental treatment
  - Protects against biases resulting from premature release of intermediate results, which may actually be random and subsequently reversed
  - ...
Intention-to-treat analysis

- Analyzes all randomized patients as if they received the full randomized treatment
  - NO MATTER WHAT!!!
  - Where randomized, there analyzed
  - NO MATTER WHAT!!!
- Examples
  - switched to the other treatment due to side effects after 6 months (2-year study)
  - switched to the other treatment due to side effects after 1 day (2-year study)
  - received the other treatment
  - ineligible patient, inadvertently admitted

Intention-to-treat analysis

- Why ITT?
  - Intention-to-treat analysis prevents bias caused by loss of participants, which may disrupt the baseline equivalence established by random assignment and may reflect non-adherence to the protocol

ITT vs. Per Protocol

- ITT is distinguished from “per-protocol” or “as-treated” and other approaches, based on what the patient actually received, and for how long
- ITT is currently the “gold-standard” primary analysis for pivotal trials
- ITT is widely-misunderstood and frequently misapplied
  - “Intention-to-treat analysis was performed on the analysis population, defined as . . .”
  - “Reasons for patient exclusion from the ITT analysis were . . .”
ITT analysis accomplishes...

- Recall: randomization protects against finding spurious treatment differences due to selection bias and/or confounding.
- ITT analysis assures that selection bias or confounding in favor of one treatment over the other can’t be reintroduced by selective implementation of the randomized allocation.
- This is true whatever the reason for selective implementation, whether:
  - patient behavior,
  - investigator preconception,
  - miscommunication,
  - or other reason,
  - intended or not,
  - scientifically/medically sensible, or not.

Limitations of ITT

- ITT can’t distinguish between:
  - scientifically/medically appropriate departures of individual patients from the randomized treatment, and
  - departures generating bias towards a treatment difference.
- That’s because we generally can’t either, and they may be the same!
- Virtually all reasons for departures from randomized allocation operating in groups of patients tend to bias a Per Protocol or As Treated Analysis towards finding or exaggerating a treatment difference.

Limitations of ITT

- ITT can’t protect against spurious differences without cost.
- When departures from the randomized treatment occur, ITT introduces a conservative bias:
  - Treatments that really do have different effects tend to look more similar than they really are.
- In other words, ITT favors regulatory protection of the public against treatments that don’t work. ITT sacrifices statistical power.
- Pharma, particularly, dislikes this. Depending on your perspective, this is either appropriate protection or restraint of trade.

Two perspectives on ITT

- Treatment Efficacy = its effect under optimal conditions of use.
- Treatment Effectiveness = its effect when applied in general clinical practice, with the usual degree of clinical imperfection and patient noncompliance.
- From this viewpoint,
  - As Treated/Per Protocol analyses attempt to compare efficacy, but tend to exaggerate it.
  - ITT analyses are compare something close to effectiveness.
  - Neither is quite on the mark.

**Biomarkers and surrogate outcomes**

- Biomarker: An objectively measurable characteristic correlated with processes distinguishing disease from health
- Biomarkers may be
  - anatomic,
  - physiologic, or
  - molecular
- and derived from
  - physical examination,
  - tissue samples, or
  - images

**Biomarkers, surrogacy: Examples**

- LDL, blood pressures for cardiovascular disease
- PSA for prostate disease
- white spots for caries
- GCF-AST for gingival inflammation
- gadolinium-enhancing MRI lesions for MS
- CD4 lymphocytes for AIDS
- hemoglobin A1c for diabetic glucose control
- glomerular filtration rate (GFR) for renal disease
- EKG changes, creatine kinase-MB fraction, and troponin for myocardial infarction
- arrhythmia frequency for myocardial infarction
Biomarkers, surrogacy: Example


Example of a potential mechanism of infectious agents in atherosclerosis.

LDL: Low-density lipoprotein. HDL: High-density lipoprotein

Biomarkers and surrogacy: Warning

- Correlation with outcome is insufficient for a biomarker to be useful as a diagnostic or surrogate clinical trial outcome!
- Diagnostic utility depends on
  - biomarker distributions in the sick and healthy
  - population disease frequency
  - overlap of biomarker information with available information from cheaper available sources including hx and px
- Validity as a surrogate outcome depends on
  - causal pathway(s) between biomarker and relevant clinical outcomes
  - position and prominence of biomarker in causal pathways between studied interventions and these outcomes
  - the latter usually being unknown in advance of study

Biomarkers and surrogacy: Warning

- There is a tremendous thirst for biomarkers and surrogates
- Everyone wants to find them
- Venture capitalists want to fund them
- Clinicians and hospitals want to offer them
- Industry wants them
  - as new diagnostics and screening tools, and
  - to cut therapeutic development costs
- But they are much harder and more expensive to develop and validate than is realized by most basic scientists and clinicians
- Most proposed biomarkers and surrogates fail
  - Some in medicine have been problematic (example: diabetes products to lower HbA1c may increase risk of cardiovascular events)

Drug development is difficult

CDER 2005 Report to the Nation

Drug Recalls

0 300 600


Prescription Over-the-counter

V I R G I N I A   C O M M O N W E A L T H   U N I V E R S I T Y
Superiority, Equivalence, & Non-Inferiority

- Treatment effect: The difference in outcomes between intervention groups
  - Continuous outcomes: difference in means
  - Binary outcomes: odds ratio, relative risk, risk diff.
- Effective: non-zero effect, as compared to control
  - Valid scientific evidence from well controlled studies for the intended use
  - Clinically meaningful endpoint
- Superiority test: a positive effect, as compared to a reference

Superiority Testing

- Goal: Reject the “no difference” hypothesis
  - $H_0: \text{Mean}_{\text{test}} - \text{Mean}_{\text{ref}} = 0$
  - $H_a: \text{Mean}_{\text{test}} - \text{Mean}_{\text{ref}} \neq 0$
- Or, define $\delta = \text{Mean}_{\text{test}} - \text{Mean}_{\text{ref}}$
  - $H_0: \delta = 0$
  - $H_a: \delta \neq 0$
- Decide using either:
  - Statistical test
  - Confidence intervals

Recall: $t$-test = $\delta/SE$

- Decide:
  - Large P-value $\rightarrow t \approx 0 \rightarrow \delta \approx 0$
  - P-value < .05 $\rightarrow |t| > 2 \rightarrow \delta \neq 0$

Recall: 95%CI $\approx \delta \pm 2 \text{SE}$

- Decide:
  - 95%CI includes 0 $\rightarrow \delta \approx 0$
  - 95%CI excludes 0 $\rightarrow \delta \neq 0$

Efficacy (AKA Effectiveness)
Equivalence

- AKA: “just as good as” claims require clinical judgment of:
- Equivalence margins (−delta, +delta): the largest difference between test and reference that would be clinically acceptable
- TOST: tested using Two One-Sided statistical Tests or
- 95% CI within the equivalence margins

Equivalence, Non-inferior

Examples:

  - Conclusion: The Tooth Morphology CD taught the anatomy of the adult dentition as well as traditional lecture, as measured by exams
  - Conclusion: The 2-implant treatment provided equivalent or more favorable treatment outcomes for most measured parameters relative to the more complex and costly 2- and 4-implant bar attachments
Thanks

... for your participation

Acknowledgement: I’m grateful to Peter B Imrey, PhD for sharing his previous version of this talk.

"... significant linear correlation between chocolate consumption per capita and the number of Nobel laureates per 10 million persons ...