Designing and Analyzing RCTs

David L. Streiner, Ph.D.

Emeritus Professor, Department of Psychiatry & Behavioural Neurosciences, McMaster University

Emeritus Professor, Department of Clinical Epidemiology & Biostatistics, McMaster University

Professor, Department of Psychiatry, University of Toronto

Master Woodworker, Halton County Radial Railway Museum
Conflict of Interest

- I wish!

- Nobody pays psychologists, so I have none.
Don’t worry about taking notes.

If you want a copy of the talk, send me an e-mail:

streiner@mcmaster.ca
What I’ll Cover

1. Designing the Study
   a. Types of trials (with a side-trip about threats to validity)
   b. Choosing the comparison group
   c. Choosing an outcome
   d. When conditions vary in length

2. Analyzing the Study
What I Won’t Cover

1. Analyzing baseline differences
2. Dealing with missing data
3. Non-normal data
4. Dealing with multiplicity
5. Co-intervention and contamination
6. Protocol deviations and violations
7. Data Safety Monitoring Boards
DESIGNING THE STUDY
TYPES OF TRIALS
Classifying Trials

(At least) two ways of classifying RCTs:

- Efficacy vs. effectiveness
  - Efficacy: *Can* the intervention work?
  - Effectiveness: *Does* the intervention work?

- Superiority vs. non-inferiority vs. equivalence
  - Superiority: Is A better than B?
  - Non-inferiority: Is A not too much worse than B?
  - Equivalence: Is A about the same as B?
BUT FIRST, A SIDE-TRIP INTO THREATS TO VALIDITY
A Bit of Jargon

- **Internal validity**
  - How well the study was done
  - How free from bias
  - How free from alternative hypotheses
  - The fidelity of the intervention
  - The adequacy of the statistical analyses

- **External validity**
  - The degree to which we can generalize the results to people and settings outside the study
Internal Validity

- RCTs eliminate many threats to internal validity:
  - History
  - Maturation
  - Regression to the mean
  - Reactive effect of testing
  - Instrumentation
Internal Validity

But not all:

- **Differential drop-out** – participants drop out of treatment and control groups for different reasons
  - Treatment group – because of adverse events
  - Control group – because of failure to improve
Internal Validity

But not all:

- Differential drop-out

- Contamination – those in control group receive intervention
  - Family doc prescribes medication
  - In social interventions, people in different conditions talk to each other
Internal Validity

But not all:

- Differential drop-out
- Contamination
- Compensatory rivalry – those in comparison group try harder (“John Henry” effect)
Internal Validity

But not all:

- Differential drop-out
- Contamination
- Compensatory rivalry
- Resentful demoralization – those in comparison group angry they didn’t get intervention
External Validity

• The degree to which findings can be generalized across people, settings, and time

• Two types:
  • Population validity
  • Ecological validity
Population Validity

- How representative is the sample of the population?
- Jeopardized by many inclusion/exclusion criteria:
  - Must meet all diagnostic criteria
  - Cannot have comorbidities
  - Cannot have history of treatment failures
  - Etc.
- How many real patients are like this?
Population Validity

• **Selection bias** – those who volunteer for an RCT different from those who don’t
  • Tend to be healthier / non-smokers
  • More likely to be Protestant or Jewish
  • More likely to be married
  • More highly educated
  • In clinical settings:
    • Berkson’s bias
    • Referral bias
Ecological Validity

• Can the treatment be applied in the real world?
• Jeopardized by tight control over delivery of therapy
  • Supervision
  • Manualized treatment
  • Reminder calls
  • Serum drug level monitoring
  • Etc.
• Can these be done in real life?
Internal and External Validity

Like two ends of a scale:
As one increases, the other decreases
Efficacy and Effectiveness Trials

• Efficacy trials also referred to as “explanatory trials”
  • “Explanatory” is a poor choice of terms, as trials rarely look “how” or “why,” but rather at “can”

• Effectiveness trials also called “pragmatic” or “management”
  • Both make sense
Efficacy and Effectiveness Trials

Two extremes of a continuum
Most studies fall somewhere in between
Efficacy and Effectiveness Trials

Where we are on the continuum dictates:

- Who is selected for the study
- How the intervention is delivered
- How drop-outs and deaths are handled
- How we deal with people who get the “wrong” treatment
- How we analyze the results
- How we interpret the results
Who Is Enrolled

- Every study has inclusion/exclusion criteria
- Effects:
  - Clinically, we know whom we’re dealing with
  - Statistically, makes the within-group variance smaller, increasing our ability to detect between-group differences

- *But*, the more criteria, the less the groups resemble actual clinic patients
Who Is Enrolled

1. Efficacy studies
   - Many inclusion and exclusion criteria
   - “Pure” groups
     - Patients meet all DSM or ICD criteria
     - No comorbidities
     - No other concurrent treatments
     - No previous treatment failures
   - Internal validity – High
   - External validity – Low
Who Is Enrolled

2. Effectiveness studies:
   - Few inclusion and exclusion criteria
   - Patients more similar to those seen clinically
     - May not meet all DSM or ICD criteria
     - May have comorbidities
     - May be taking other treatments
     - May have failed previous treatments
   - External validity – High
   - Internal validity – Low
What Is Done

1. Efficacy studies:
   - Best treatment, ideal circumstances
   - Standardized (e.g., manualized) treatment
   - Experienced therapists
   - Drop centres that do not meet standards
   - Reminders to take drugs, complete diaries
   - Monitoring of drug levels
What Is Done

2. Effectiveness studies:
   - Treatment as it’s usually delivered
   - More flexibility in delivering treatment
   - Mix of experienced and novice therapists
   - Range of treatment centres
   - No special reminders
   - Monitoring of drug levels only if SOP
Who Gets Counted

1. **Efficacy trials:**
   - Only those who get treated in their intended condition
   - Don’t count drop-outs
   - Don’t count those who died before starting treatment
   - *But*, this may affect internal validity if there is differential drop-out
Who Gets Counted

2. Effectiveness trials:
   - Use intention to treat (ITT) – people counted even if they received “wrong” treatment
   - Drop-outs included
   - Dying before treatment starts is part of reality, so they’re counted, too
   - *But*, this may affect internal validity if there is differential drop-out
What Gets Counted

1. Efficacy trials:
   - Outcomes related to the intervention

2. Effectiveness trials:
   - All outcomes
Analysis

1. Efficacy trials
   • Since only completers are counted, no special techniques needed to deal with missing data

2. Effectiveness trials:
   • Because of ITT, must find a way to handle drop-outs and missing data
     • Usual way is Last Observation Carried Forward (LOCF)
     • We’ll return to this
Which to Do

• Both are often necessary
  • If effectiveness trial is positive, efficacy likely not needed
• But, if negative, is it because
  • Treatment ineffective
  • Treatment effective, but delivery is less than optimal
    • Therapists need more training
    • Patients too heterogeneous
    • Problems with compliance
    • Etc.
THE COMPARISON GROUP
The Comparison Group

• Easy in theory, especially in an RCT
  • Treatment and control groups the same except for the intervention

• More difficult in practice
  • Intervention group gets more than therapy
    • Attention
    • Non-specific effects of treatment
  • Wait-list control does not get nothing
    • Pre- and post testing
What’s The Question?

• **Question 1**
  - Treatment vs. No Treatment?
  - Treatment A vs. Treatment B?
  - Treatment A vs. Treatment A + B?

• **Question 2**
  - Aside from therapy, what else is treatment group receiving?
    - More contact time?
    - Adjustment of meds? Blood levels?
    - Between-session telephone contact?
What’s The Question?

• **Question 3**
  - What determines content of treatment?
    - Same for all patients?
    - Modules tailored to patients’ needs?
      - Advantage: May be more clinically relevant
      - Disadvantage: Hard to describe what happened

• **Question 4**
  - How is termination determined?
    - Fixed number of sessions?
    - Until criterion reached?
      - Same advantage and disadvantage
What’s The Question?

• No right or wrong answers
• Depends on what you’re trying to show

• But, *these questions must be considered*
CHOOSING AN OUTCOME
The Outcome

• Ideally, should be *directly* related to the intervention and be the variable of interest
  • Decrease in suicide attempts
  • Decrease in hospitalizations
  • Increase in quality of life

• Seems obvious, but …
Beware the Surrogate Outcome!

- Examples of surrogate outcomes:
  - Scores on a scale
  - Blood levels
  - Number of diagnostic criteria

- The purpose of treatment is not to decrease these values
So Why Do We Use Them?

- “Real” outcome
  - Too rare
  - Too temporally distant
  - Can’t be assessed directly
- Assume they’re highly correlated with outcome of interest, but they aren’t always
  - Homocystine levels in AD
  - Plaque levels in AD
Surrogate Outcomes

• Use them only if:
  • They are highly correlated with what’s important
  • More importantly, a change in the surrogate is correlated with a change in the outcome
How Many Outcomes?

- Move toward simpler trials with single outcome
  - May make sense in drug trials
  - Doesn’t make as much sense in psychosocial interventions
    - Interventions are complex
    - Outcomes are varied
- Measure what’s important, but don’t go overboard
Which Primary Outcome?

• Easy if there’s a gold standard (e.g., BDI, Ham-D)
• More of a problem when outcomes vary for each individual
• No easy answer; may have to use global assessments
  • Goal Attainment Scaling
  • Global Assessment of Functioning
WHEN CONDITIONS VARY IN LENGTH
Why Conditions Vary in Length

• By design
  • Compare 6 months DBT vs. 12 months
  • Compare time-limited vs. unlimited therapy

• By outcome
  • Treatment until criterion reached

• Note: We are not looking at time to completion as the outcome
Why It’s a Problem

Group 1

- Treatment
- Follow-up

Group 2

- Treatment
- Follow-up

Time 1
Group 1 – Treatment over
Group 2 – Only half way through treatment
Why It’s a Problem

Group 1
- Treatment
- Follow-up

Group 2
- Treatment
- Follow-up

Time 2
- Group 1 – 6 months of follow-up
- Group 2 – Treatment just ending
Why It’s a Problem

Group 1: Treatment for 12 months followed by follow-up

Group 2: Treatment for 6 months followed by follow-up

Time 3:
- Group 1: 12 months of follow-up
- Group 2: 6 months of follow-up
Why It’s a Problem

• No matter when groups are assessed, they differ with regard to:
  • Whether or not therapy is on-going
  • How long the follow-up time is

• Are differences due to:
  • More treatment?
  • More non-specific effects?
  • Consolidation of treatment effects?
  • Length of follow-up?
The Solution

• There ain’t none!

• Easier if looking for non-inferiority rather than superiority

• The longer the follow-up period, the less it matters
ANALYZING THE STUDY
TYPES OF ANALYSES
Types of Analysis

- Four main types of analyses:
  - Intention-to-Treat (ITT)
  - Modified Intention-to-Treat (mITT)
  - Per-Protocol (PP)
  - As-Treated

- (There are others, but they depend on sophisticated statistical methods)
What Are The Issues?

• Especially in longitudinal studies, there will always be drop-outs
• Introduces two problems:
  • Reduced power
  • Potential for bias
Reduced Power

- Relatively easy to deal with:
  - Recruit more participants to begin with
  - Replace drop-outs

- More people doesn’t solve problem with bias
  - May even exacerbate it
Potential for Bias

- People do not drop out studies for trivial reasons
  - In treatment group:
    - Patient achieved desired goal
    - Adverse effects
  - In comparison group:
    - Disappointment
    - Lack of improvement
- Major problem if differential drop-out
Differential Drop-Out

• Assume:
  • 50 people/group
  • Scores at baseline ~ $N(15,5)$
  • Treatment is useless

• At each session:
  • 5 with highest scores drop out of treatment group because of adverse events
  • 5 with lowest scores drop out of placebo group because of lack of effect
The Solution

• To repeat, 4 major approaches:
  • Intention-to-Treat (ITT)
  • Modified Intention-to-Treat (mITT)
  • Per-Protocol (PP)
  • As-Treated

• Differ in terms of who is/is not counted in each group
Intention-to-Treat (ITT) Analysis

• “Once randomized, always analyzed”

• Participant included in group to which originally randomized

• Missing data imputed (stay tuned)
Advantages of ITT

• Preserves sample size

• Preserves randomization

• Mandated by FDA (if not primary analysis, must be a secondary one)
Disadvantages of ITT

• May underestimate effect of intervention
  • Untreated people counted in treatment group
  • Wrongly treated people counted in control
  • A particular problem with BPD studies because of non-compliance with treatment

• Therefore more likely for Type II error or under-estimate of effect of treatment
Disadvantages of ITT

• Doesn’t answer question we’re asking
  
  • We ask, “Does treatment work for people who get it?”
  
  • It answers, “Does it work for people prescribed it?”
Disadvantages of ITT

- Should *never* be used in equivalence or non-inferiority trials
  - These are shown if groups don’t differ
  - ITT biases outcome to finding no difference (Type II error)
- Should *not* be used when looking for adverse events
  - Underestimates those in treatment group
  - Overestimates those in control group
Modified ITT Analysis

- ITT with some exceptions
  - Those later found to not meet inclusion/exclusion criteria (e.g., new info; lied)
  - Those who never started treatment
  - Never completed baseline assessment
  - Etc.

- Appropriate if modifications specified *a priori*

- No consistency from study to study
Per-Protocol Analysis

- Also called ‘completer analysis’

- Patients counted in group only if treated according to protocol
  - If in treatment group, only if received treatment
  - If in control group, only if did not receive treatment
Per-Protocol Analysis

• Advantages
  • Does not dilute treatment effects
  • Best way to analyze adverse effects data

• Disadvantages
  • Reduced sample size
  • Disrupts randomization scheme
  • May introduce selection bias
As-Treated Analysis

- Patients counted according to what they received
  - If treatment patients not treated, put in placebo group
  - If control groups treated, put in treatment group
As-Treated Analysis

- **Advantage**
  - Fewer dropped participants than in PP

- **Disadvantages**
  - Increased probability of bias due to cross-overs
  - Undoes randomization
Which to Use

• Depends on aim of analysis
  • For *efficacy* (can it work), primary should be PP or As Treated
  • For *effectiveness* (does it work), primary should be ITT or mITT
  • For non-inferiority or equivalence, PP or As Treated
  • For adverse events, PP or As Treated
• Best to use 2 or 3 (one must be ITT if FDA approval needed)
DEALING WITH MISSING DATA
Why It’s a Problem

• Data set consists of:
  • $N = 10$
  • 5 variables (A through E)
  • One data point missing for each variable
  • Each variable missing only 10% of data
<table>
<thead>
<tr>
<th>Subject</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✗</td>
</tr>
<tr>
<td>2</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>3</td>
<td>✔</td>
<td>✔</td>
<td>X</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>4</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>5</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>6</td>
<td>✔</td>
<td>✔</td>
<td>X</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>7</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>X</td>
<td>✔</td>
</tr>
<tr>
<td>8</td>
<td>X</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>9</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>10</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>
Missing Data

• If we do any analysis using >1 variable
  
  • Most programs use *list-wise* deletion
    
    • If any variable has missing data, entire case deleted
    
    • Can use *pair-wise* deletion, but you never should
  
  • Assume you do a regression using all variables
<table>
<thead>
<tr>
<th>Subject</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>2</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>3</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>4</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>5</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>6</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>7</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>8</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>9</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>10</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>
Missing Data

• You have lost 50% of your data

• In one simulation:
  • If 2% of data are missing at random, 18% of subjects lost using list-wise deletion
  • If 10% missing at random, 59% lost
Types of Missing Data (1)

• Missing Completely at Random (MCAR)
  • No pattern to missingness; completely random
  • Rarely happens
  • Called ‘ignorable’
Types of Missing Data (2)

• Missing at Random (MAR)
  • Missingness can be explained by some other variable
    • E.g., depressed people may be less likely to report quality of life
  • But, if non-reporting is not related to QOL *within* depressed group, data are MAR

• Also called ‘ignorable’
Types of Missing Data (3)

- Not Missing at Random (NMAR)
  - Reason for missingness related to the variable itself
    - More depressed people don’t report degree of depression
    - More physically impaired children can’t do tests of physical ability
    - More mentally impaired children can’t do IQ tests
    - People with high and low incomes won’t report them
  - Called ‘non-ignorable’ for a reason!
Ways of Imputing Data

- Replacement with the mean
- Using multiple regression
- Using multiple regression + error
- Last observation carried forward
- Multiple imputation
- Maximum likelihood
- HLM (Growth curve analysis)
### Replacement With the Mean

<table>
<thead>
<tr>
<th>Subject</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>53</td>
</tr>
<tr>
<td>2</td>
<td>27</td>
</tr>
<tr>
<td>3</td>
<td>38</td>
</tr>
<tr>
<td>4</td>
<td>Missing</td>
</tr>
<tr>
<td>5</td>
<td>46</td>
</tr>
</tbody>
</table>

Mean = 41, so replace missing value with 41
Replacement With the Mean

• Advantages
  • Easy to do and explain
  • Doesn’t bias estimate of mean

• Disadvantages
  • Reduces estimate of SD and therefore SE
  • Increases probability of finding significance
Replacement With the Mean

- Can improve estimate if missing values correlated with another variable
  - E.g., If weight is missing, we know it’s correlated with sex
  - Use separate means for males and females
Multiple Regression

- We used one variable (sex) to improve estimate of weight
- Can go further; use many variables to predict weight (sex, age, height, etc.)
- Multiple regression
Multiple Regression

• Advantages
  • Easy to do and explain
  • Doesn’t bias estimate of mean
  • Reduces SD less than replacement with mean

• Disadvantages
  • Still reduces estimate of SD
  • People with same combination of predictors get identical estimates
Multiple Regression + Error

Variations on a theme:

- Deliberately add some error to each estimate
- Chain the estimates
  - Use variables B, C, and D to predict missing values in A
  - Use variables A, C, and D to predict missing values in B
  - etc.
- Reiterate with more complete data
Last Observation Carried Forward (LOCF)

- Used when people drop out of longitudinal studies
- Used in drug trials
- Has imprimatur of FDA
  - Sure sign it has flaws
LOCF – What We Assume

Time

Score

Treatment group improves

Control group doesn’t improve as much
LOCF – What We Assume

Person drops out at Time 3; value carried forward
Underestimates effect in treatment group – this is good
LOCF – What We Don’t Assume

Assumes no measurement error – this is bad
LOCF – What We Don’t Assume

Underestimates effect in control group – this is also bad
LOCF – What We Don’t Assume

- So far, one good assumption:
  - Drop-outs in treatment group don’t improve

- Two bad assumptions:
  - Constant value carried forward
  - Drop-outs in control group don’t improve

- Worst (fatal) flaw is in studies of interventions to slow decline:
  - “Memory drugs” in Alzheimer’s disease
  - Erythropoietin in cancer treatment
LOCF – What We Don’t Assume

- Treatment group declines
- Control group declines faster
LOCF – What We Don’t Assume

The earlier the patient drops out, the better the drug looks.
LOCF – What We Don’t Assume

• From a drug company’s perspective, the best study is one where:
  • Many people drop out
  • They drop out early

• How can this be good?
• In studies trying to show improvement, use with extreme caution.

• In studies trying to slow decline, never use at all.
Advanced Techniques

- Multiple imputation
- Maximum likelihood
- HLM (Growth curve analysis)

- Still assume data MCAR or MAR
- Can take other variables into account
- Much less biased estimates of variance
Advanced Techniques

• Implemented in high-end programs:
  • Stata, SAS, Mplus, etc.
  • SPSS is limited

• Don’t try these at home; need a friendly statistician (they do exist)

• The way to go.
QUESTIONS?