Understanding Design & Statistics in Clinical Trials
Why bother?
Experience-Based Medicine?

Benjamin Rush

“Father of American Psychiatry”
Experience-Based Medicine?

• Supported blood-letting as treatment for various diseases

• George Washington probably died from excess blood-letting for a throat infection

• One of Rush's students was a major player in that episode!

Yes, we can!

• Design of the study, approach to statistical testing, interpreting the results

• **Not** details of the statistical tests done

• 6 Questions: 3 about how the study was done and 3 about the findings

• Slides at **www.simpleandpractical.com/clinicaltrials**
Caution!

- “Critical appraisal” of the medical literature
Question no. 1:
What is the STUDY DESIGN used by the clinical trial?
Clinical Trials

1. Uncontrolled vs. Controlled
   A. Placebo control
   B. Active control

2. Randomized vs. Non-randomized

3. Open-label vs. Single blind vs. Double blind
1. Uncontrolled vs. Controlled

Q. Why do we need a control group?
James Lind

- Father of (controlled) clinical trials
- First to introduce control groups (in 1747)
- Showed that citrus fruits in the diet could prevent scurvy
• All scurvy patients given same general diet

• Different groups of patients supplemented with different items - oranges/ lemons OR cider OR vinegar OR seawater OR nutmeg OR others

• In just six days, patients taking citrus fruits were fit for duty
Uncontrolled Studies

• Of 131 published uncontrolled studies in psychiatry in a defined period, 93% reported the treatment to be efficacious*

• Only 27% were followed up by published, positive RCT

* Mago et al. (unpublished data)
Effectiveness of Adjunctive Antidepressant Treatment for Bipolar Depression

Gary S. Sachs, M.D., Andrew A. Nierenberg, M.D., Joseph R. Calabrese, M.D., Lauren B. Marangell, M.D., Stephen R. Wisniewski, Ph.D., Laszlo Gyulai, M.D., Edward S. Friedman, M.D., Charles L. Bowden, M.D., Mark D. Fossey, M.D., Michael J. Ostacher, M.D., M.P.H., Terence A. Ketter, M.D., Jayendra Patel, M.D., Peter Hauser, M.D., Daniel Rapport, M.D., James M. Martinez, M.D., Michael H. Allen, M.D., David J. Miklowitz, Ph.D., Michael W. Otto, Ph.D., Ellen B. Dennehy, Ph.D., and Michael E. Thase, M.D.

- From STEP-BD: Bipolar depression, 26-week randomized, double-blind study
- Mood stabilizer plus antidepressant (bupropion or paroxetine)
• Mood stabilizer plus antidepressant (n=179) – 10.1% switched to hypomania/mania

• Mood stabilizer plus placebo (n=187) – 10.7% switched

• Without placebo control, could not interpret the data!
2. Randomized vs. Non-Randomized

Q1: Why do we need randomization?

• Prevents one treatment group from having a better prognosis

Q2. Well, then why can’t we just make the groups similar ourselves rather than leaving it to chance?

• Tends to equally distribute both known and unknown factors
3. Open-Label vs. Blinded

Q. Why do we need blinding?

• Can be:

  Open-label vs. Single blind vs. Double blind

• Ideally, double-blind
Rule #1:

Only consider randomized, double-blind, controlled trials (& syntheses based on them)

(with rare exceptions)
Rule #1:

Only consider randomized, double-blind, controlled trials (& syntheses based on them)

(with rare exceptions)
Exclude from consideration:

- Case reports
- Naturalistic studies
- Uncontrolled, open label trials
- Will be able to exclude over 95% of treatment-related publications!
Question no. 2:

What is the POPULATION being studied?
Inclusion/ Exclusion Criteria

• Increased efficiency

• Increased safety

• Efficacy vs. Effectiveness studies
Implications

• Internal validity vs. Generalizability
Generalizability

• Patients in clinical trials are usually a selected sample of real world patients

• “Were the patients in the study similar to the patients I’m applying the results to?”

• Based on the biology of the illness, whom can I generalize the findings of the study to?
The Efficacy and Safety of Aripiprazole as Adjunctive Therapy in Major Depressive Disorder: A Multicenter, Randomized, Double-Blind, Placebo-Controlled Study

Robert M. Berman, M.D.; Ronald N. Marcus, M.D.; René Swanink, M.S.; Robert D. McQuade, Ph.D.; William H. Carson, M.D.; Patricia K. Corey-Lisle, Ph.D., R.N.; and Arif Khan, M.D.

• 362 patients with inadequate response to antidepressants
• Randomized to augmentation with aripiprazole or placebo
• Aripiprazole superior to placebo
Whom can we generalize to?

- Failed 1 to 3 antidepressants by history, and 1 more during the study

- Traditionally, if no response → switch
  if partial response → augment

- Included both non-responders & partial responders

- Subgroup analysis (consider tentative) – no difference between non-responders & partial responders
Acute bipolar mania: Mood Stabilizer vs. Mood Stabilizer plus Second-Generation Antipsychotic?

- 8 studies, 1124 patients
- Difference in mean scores on YMRS was 4.4
- Significantly more participants on co-therapy responded. Relative “Risk” 1.53
Acute bipolar mania:
Mood Stabilizer vs. Mood Stabilizer plus Second-Generation Antipsychotic?

BUT

Subjects could already have been on mood stabilizer before entering the study

So, it's possible (but don't really know) that we should add a SGA only in patients who have not done well on mood stabilizer alone
Rule #2: POPULATION?

Look at inclusion/exclusion criteria.

Ask yourself: “Whom can the findings be generalized to?”
Question no. 3:

Should be only ONE EXPLANATION for the treatment group doing better than the control group:

Is everything kept (or made) the same except the treatment being studied?
1. Baseline Characteristics

• With randomization, baseline characteristics (known and unknown) tend to be equally distributed BUT...

• Are sometimes different

• Look at Table 1 of most papers
Table 1. Baseline Demographic and Clinical Characteristics of Adult Patients With Major Depressive Disorder in the Safety Population Who Were Randomly Assigned to Either Vilazodone or Placebo

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Vilazodone, N = 205</th>
<th>Placebo, N = 204</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female gender, n (%)</td>
<td>127 (62.0)</td>
<td>130 (63.7)</td>
</tr>
<tr>
<td>Race, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White&lt;sup&gt;a&lt;/sup&gt;</td>
<td>181 (88.3)</td>
<td>157 (77.0)</td>
</tr>
<tr>
<td>Black/African American&lt;sup&gt;a&lt;/sup&gt;</td>
<td>20 (9.7)</td>
<td>36 (17.6)</td>
</tr>
<tr>
<td>Other</td>
<td>4 (2.0)</td>
<td>11 (5.4)</td>
</tr>
<tr>
<td>Age, mean (SD), y</td>
<td>40.0 (12.1)</td>
<td>39.8 (12.7)</td>
</tr>
<tr>
<td>Range, y</td>
<td>18–63</td>
<td>18–65</td>
</tr>
<tr>
<td>Age at onset of depression, mean (SD), y</td>
<td>33.4 (13.4)</td>
<td>31.9 (13.8)</td>
</tr>
<tr>
<td>First episode of depression, n (%)</td>
<td>72 (35.1)</td>
<td>76 (37.3)</td>
</tr>
<tr>
<td>Duration of current episode, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1–6 mo</td>
<td>107 (52.2)</td>
<td>110 (53.9)</td>
</tr>
<tr>
<td>&gt; 6–12 mo</td>
<td>68 (33.2)</td>
<td>78 (38.2)</td>
</tr>
<tr>
<td>&gt; 12 mo&lt;sup&gt;a&lt;/sup&gt;</td>
<td>30 (14.6)</td>
<td>16 (7.9)</td>
</tr>
<tr>
<td>Severity of current episode, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Moderate</td>
<td>147 (71.7)</td>
<td>147 (72.1)</td>
</tr>
<tr>
<td>Severe</td>
<td>58 (28.3)</td>
<td>57 (27.9)</td>
</tr>
<tr>
<td>Melancholia, n (%)</td>
<td>99 (48.3)</td>
<td>89 (43.6)</td>
</tr>
<tr>
<td>Recurrent depression, n (%)</td>
<td>133 (64.9)</td>
<td>128 (62.7)</td>
</tr>
<tr>
<td>GAD or SAD within past 6 mo, n (%)</td>
<td>13 (6.3)</td>
<td>15 (7.4)</td>
</tr>
</tbody>
</table>

<sup>a</sup>p < .05, Fisher exact test.
2. Intention-to Treat analyses

- 100 patients randomized to study drug, 30 drop out before completing the study

- Why is this a problem? What to do with these 30 pts in data analysis?
2. Intention-to Treat analyses

- Or “Modified Intention-to treat” analyses

- Vs. “Completer analysis”
Abbreviation: ITT = intention to treat.
Intention-to-Treat Analysis: How?

1. **LOCF**
   
   Last Observation Carried Forward

2. **Mixed Models**
Last Observation Carried Forward (LOCF)
Rule #3: Both groups should be or be made equal except for the intervention:

A. Baseline characteristics similar, unimportant, or controlled for,

AND

B. Intention-to-Treat analyses
Question no. 4:

STATISTICAL “SIGNIFICANCE”
Question no. 4:

STATISTICAL “SIGNIFICANCE”

Could the difference found be just due to chance?
Research question: Is Drug A superior to citalopram for reducing severity of depression in MDD over 8 weeks?

Recruit 200 patients with MDD

Randomize to Drug A or citalopram

58% of pts on Drug A remitted vs. 48% of pts on citalopram
A lot of research is based on **Hypothesis testing**

What is a **Null Hypothesis**?

What difference in score would convince you?
Increasingly improbable

• Started with assumption that study drug and control were equal?

• 60% improved on study drug vs. 40%

• 80% improved on study drug vs. 40%

• 90% improved on study drug vs. 40%
Hypothesis testing is like Legal trials
Legal Trials

• Assume innocent, burden of proof on prosecution

• Hard to conclusively “prove” innocence. Can only show that it is unlikely, given the evidence, that the person is innocent

• Must find guilt “beyond a reasonable doubt.” (There can always be doubt, but it is very unlikely)
Clinical Trials

• Assume innocent = “Null hypothesis”

• Null hypothesis = start with assumption that there is no difference between the 2 groups
Clinical Trials

• Guilt beyond a reasonable doubt = = Statistically very unlikely that this difference between the 2 groups is just by chance

• I.e., given the results we got – Null Hypothesis is very unlikely

• Same as “Statistically significant difference”
What does P value mean?

A. The extent to which the study deserves to bepeed upon (lower pee value is good)!

B. How likely the treatment is to help the patients

C. How likely it is that this result could have been obtained by chance

D. How significant this finding is for subsequent clinical application
What does P value mean?

A. The extent to which the study deserves to be peed upon (lower peek value is good)!

B. How likely the treatment is to help the patients

C. How likely it is that this result could have been obtained by chance

D. How significant this finding is for subsequent clinical application
P value

P value means if in reality there was no difference, the probability that the current results could have been obtained just by chance
P value

• If more patients on study drug improved vs. control group, the LOWER the p-value...

• LESS likely that the observed difference is not real and simply due to chance
P value

<table>
<thead>
<tr>
<th>p</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.20</td>
<td>20%</td>
</tr>
<tr>
<td>0.05</td>
<td>5%</td>
</tr>
<tr>
<td>0.04</td>
<td>4%</td>
</tr>
<tr>
<td>0.01</td>
<td>1%</td>
</tr>
<tr>
<td>0.001</td>
<td>0.1%</td>
</tr>
</tbody>
</table>
P value

<table>
<thead>
<tr>
<th>p</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>.20</td>
<td>20%</td>
</tr>
<tr>
<td>.05</td>
<td>5%</td>
</tr>
<tr>
<td>.04</td>
<td>4%</td>
</tr>
<tr>
<td>.01</td>
<td>1%</td>
</tr>
<tr>
<td>.001</td>
<td>0.1%</td>
</tr>
</tbody>
</table>

"Statistically significant"
Statistical “Significance”

• Instead of “statistically significant,” think of it as “statistically UNLIKELY”

• Important: p value < .0001 is NOT more clinically “significant” than p < .05

• What is the difference between p < .0001 and p < .05
Efficacy of Duloxetine and Selective Serotonin Reuptake Inhibitors

Comparisons as Assessed by Remission Rates in Patients With Major Depressive Disorder

Michael E. Thase, MD,*†‡ Yili Lu Pritchett, PhD§ Melissa J. Ossanna, PhD,∥ Ralph W. Swindle, PhD∥∥ Jimmy Xu, PhD∥∥ and Michael J. Detke, MD, PhD∥∥∥

• Meta-analysis of 6 studies comparing duloxetine to an SSRI and placebo

J Clin Psychopharmacol 2007
• HAMD-17 remission rates 40.3% on duloxetine, 38.3% on SSRI (N.S.)

• No difference on other outcome measures either

• *%$#&^@ it !!!!

• What to do?
• Subgroup of pts with more severe depression remission rates 35.9% on duloxetine, 28.6% on SSRI

\[(p = .046)\]

• Conclusion “...therapy with SNRI resulted in a significantly higher remission rate among patients with moderate-to-severe depression”
Problem of Multiple Testing

• If 3 outcome measures: HAM-D scores, MADRS scores, CGI scores

• Use < 5% as cut off for chance of finding a difference between two treatments just by chance

• What if we tested all 3 ways of looking at it?
Two Options

1. See if they defined a “Primary Outcome Measure” in advance

   Consider all their other statistical testing as tentative (“hypothesis generating”)

2. Set the bar higher with the p value. If \( p < .05 \) is the bar & 5 tests are done, use \( p < .01 \) as the cutoff – “Bonferroni correction”
Rule #4:

Believe $p < .05$ only for the prespecified primary outcome measure

Use a different $p$ value for others OR consider them preliminary
Question no. 5:

“NEGATIVE STUDIES”:

Statistically significant difference NOT found

What to make of it?
Three possibilities:

1. Treatments really did work equally well

2. Neither treatment had a significant effect

3. One treatment worked better but the study lacked the “power” to detect this difference
STAR*D Study

Level 1

Citalopram

Level 2

Switch to:
bupropion (sustained release), or
venlafaxine (extended release), or
sertraline, or
cognitive therapy

or

Augment with:
bupropion (sustained release), or
buspirone, or
cognitive therapy

Level 2a

(only for those receiving cognitive therapy in level 2)

Level 3

Switch to:
mirtazapine or
nortriptyline

or

Augment with:
lithium or
T₃ thyroid hormone

Level 4

Switch to:
tranylcypromine or
mirtazapine + venlafaxine (extended release)
• If one treatment works better than the other one – no problem

• But, what if each treatment works equally well?

• Maybe NEITHER worked in this particular sample?

• How is that possible, you say?
“Assay Sensitivity”
STAR*D Study

Level 1

Citalopram

Level 2

Switch to:
bupropion (sustained release), or
venlafaxine (extended release), or
sertraline, or
cognitive therapy

or

Augment with:
bupropion (sustained release), or
buspirone, or
cognitive therapy

Level 2a

(only for those receiving cognitive therapy in level 2)

Switch to:
bupropion (sustained release) or
venlafaxine (extended release)

Level 3

Switch to:
mirtazapine or
nortriptyline

or

Augment with:
lithium or
T₃ thyroid hormone

Level 4

Switch to:
tranylcypromine or
mirtazapine + venlafaxine (extended release)
• “Non-inferiority” studies MUST have a placebo control
Placebo-controlled study?

• When is absence of an active control a problem?

• If drug is better than placebo, no problem

• Active control was unnecessary
If drug = placebo

• Maybe population was not response to the treatment (assay not sensitive)
  
e.g., treatment resistant patients or “situational” depression

OR

population was too responsive to placebo (Tip: look to see if response rate in control group was high)
Masand et al. (2001)

<table>
<thead>
<tr>
<th>Reduction in ASEX score</th>
<th>Placebo</th>
<th>Bupropion</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt; 50%</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>25-49%</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>TOTAL</td>
<td>2/15</td>
<td>2/15</td>
</tr>
</tbody>
</table>
If drug = placebo

• “Failed study” vs. “Negative study”
Statistical Power

= the ability of the study to detect a difference (with statistical confidence) when a real difference is present

Depends on number of subjects, size of the difference, and degree of variability
Comparative efficacy study of haloperidol, olanzapine and risperidone in delirium

Sandeep Grover*, Vineet Kumar, Subho Chakrabarti

Department of Psychiatry, Postgraduate Institute of Medical Education & Research, Chandigarh, India

- Significant improvement on Delirium Rating Scale & MMSE in all 3 groups
- No significant difference between the 3 groups
- “Risperidone and olanzapine are as efficacious as haloperidol in the treatment of delirium”
• Problems?

• Obviously, no placebo group. All 3 groups may have improved due to other reasons

• Also, 64 patients randomized to 3 groups – probably underpowered
Rule #5: If a study does not find a statistically significant difference,

1. Appropriate control group (active or placebo), i.e., assay sensitive?

2. Enough power (patients) to detect a meaningful difference?

3. Unexpectedly high response rate in control group?

(Look at numerical difference)
Question no. 6:

Statistically significant difference found

Should we implement the treatment right away?
Two possible problems:

1. The finding may still have occurred only by chance

2. The difference found may be real, but not large enough to be CLINICALLY significant
Found statistically significant difference

• p < .05 is not the ONLY criterion

• Lower p value ➞ more unlikely that by chance alone (i.e, more likely the difference is real)

• Other issues in drawing inference from a study
Other issues in inference

1. Similar results with different approaches?

2. Biological plausibility?

3. Dose response relationship?

4. Replication?
Are SNRIs more efficacious than SSRIs for MDD??

• Papakostas et al. (2007)

• 93 trials (n = 17,036)

• SNRIs response rates statistically significantly more than with SSRIs (p = .003)

• Should we preferentially use SNRIs for most patients?
Statistical vs. Clinical Significance

• In articles, when they say “…was significantly greater than with placebo”

they mean statistically significantly greater

• You have to ask – are results clinically significant enough?
Size Does Matter!

Look for the **size of the effect**

- “Risk” difference
  (should usually be 10% or more)

- Odds ratio

- Relative “Risk”

- Hazard Ratio

- Number Needed to Treat
Number Needed to Treat (NNT)

• If 10% more chance of responding with one tx and tx 10 pts, have 100% chance of getting 1 more responder

• If 20% more chance, need to treat only 5 patients to get 100% chance of getting 1 more responder

• NNT = 100% / % difference between groups
Number Needed to Treat (NNT)

- NNT of 5 to 10 is usual for effective treatments
- But – depends on other factors like seriousness of illness, adverse effects, cost of treatment, etc
Number Needed to Treat (NNT)

- SNRIs vs. SSRIs:
  ~64% versus 59%

What is the NNT?

\[
\text{NNT} = \frac{100}{5} = 20
\]
The Efficacy and Safety of Aripiprazole as Adjunctive Therapy in Major Depressive Disorder: A Multicenter, Randomized, Double-Blind, Placebo-Controlled Study

Robert M. Berman, M.D.; Ronald N. Marcus, M.D.; René Swanink, M.S.; Robert D. McQuade, Ph.D.; William H. Carson, M.D.; Patricia K. Corey-Lisle, Ph.D., R.N.; and Arif Khan, M.D.

• 362 patients with inadequate response to antidepressants

• Randomized to augmentation with aripiprazole or placebo
• “Mean change in MADRS score was significantly greater with adjunctive aripiprazole (-8.8) than adjunctive placebo (-5.8; p < .001)”

• For clinical significance, go to response rates

• Response rates at end of study: ~24% on placebo, 34% on aripiprazole

• “Risk” difference = 34 – 24 = 10
Rule #6: If a study does find a statistically significant difference:

A. Look for other evidence of a valid inference

B. Look at the size of the effect. Is it clinically significant? (Percentage of responders in each group is one way of doing that)
Conduct of the Study

1. Study design?

2. Population?

3. Only one explanation?
Results of the Study

1. True statistical significance?

2. Truly no difference?

3. Clinical significance?