Most clinical research evaluates an association between exposure/intervention and outcome.
Does the investigator decide who gets the intervention/exposure?

Yes  
Experimental/Interventional Study  
Usually Randomized  

No  
Observational Study  
Comparison Between Groups  
Yes  
Analytical Study  

No  
Descriptive Study  

Cohort Study  
Usually Prospective  
Case-Control Study  
Retrospective  
Cross-sectional Study  

Cohort Design

Exposure  
Observer  
Time  
? Disease  

Case Control Design

? Exposure  
Disease  
Time  
Observer
Hormone Replacement Therapy

- Reduces symptoms of menopause, but should it be used to reduce/prevent cardiovascular disease?
- Large number of observational studies (case-control and cohort) concluded that hormone replacement therapy reduced the risk of cardiovascular disease in post-menopausal women.

- Women’s Health Initiative - randomized controlled trial – relative risk of coronary heart disease 1.29 (1.02-1.63) with use of hormones
- What’s the likely explanation?

Women's Health Initiative Group. JAMA 2002

Postmenopausal Women

Take hormone replacements
- More affluent
- More Caucasian
- Better health care access
- More healthy behaviors
- Less cardiac disease

Do not take hormone replacements
- Less affluent
- Less Caucasian
- Poorer health care access
- Less healthy behaviors
- More cardiac disease
Postmenopausal Women Randomly Assigned

- Equivalent risk of cardiac disease
  - Hormone replacements
    - More cardiac disease
  - Placebo
    - Less cardiac disease

Randomized Trials

- Most rigorous design for assessing the effects of interventions
- Groups are comparable at study entry - should differ only by intervention under investigation, even if there are unrecognized important variables
- But...
Even RCTs Can Get the “Wrong” Answer

• Part of the problem is our slavish devotion to the “p value”.

• And the common practice of saying, based on a p value, that there is or isn’t a difference.

• What does the p value really mean?

p Value

The likelihood that a difference this large could have been observed between the study groups, just by chance, when there really is no difference in the underlying populations (conditions).
What?

We never measure outcomes in the entire group we’re interested in. We study a “sample” from each group.

If you just pick a few people at random from one group and a few people at random from another group and you measure something (blood sugar, BMI, whatever), the mean in one sample will probably be a little different than the mean in the other sample, just because of chance variation in sampling.
The World My Sample

Blood Sugar

BMI

First Grade Fourth Grade

First Grade Fourth Grade

p > 0.05

p > 0.05

p < 0.05

p < 0.05

The World

My Sample
“Statistically significant” difference can be due to...

- True effects of the exposure/intervention/treatment
- Sampling variation or chance (5% of the time when p value of 0.05 is used)
- Inherent differences between the groups apart from the exposure or treatment of interest
- Differences between handling or evaluation of the groups

“No significant difference” can be due to...

- There truly is no difference between the two populations (exposures/treatments).
- Inherent differences in the study groups or in their handling or evaluation.
- Because of a small sample size, there appears to be “no difference”, based on the statistical test.
“No significant difference” because sample size was too small

- How does that work?
- Ability to identify a “significant” difference depends on
  - the size of the difference between the two groups
  - the prevalence of the outcome (if it’s a yes/no outcome like death)
  - the variability in the outcome (if it’s a continuous outcome like BP)
  - and the number of patients studied.
What is study power?

- Calculated before the study is started
- Power = likelihood that you will be able to identify a statistically significant difference if there really is a difference between the populations
- BUT
  - It depends on how big a difference you’re willing to miss or fail to identify
  - Depends on the SD of your measurement if your outcome is continuous, like BP
  - Depends on the prevalence of the outcome if it’s yes/no or categorical, like death

Hypothermia Study

- Hypothermia vs Control for infants with perinatal asphyxia
- Calculated sample size of 208 to have 80% power to detect a reduction in death or mod-severe disability from 50% to 30% (RR=0.60).
- What does that mean?

- If the real truth (we can’t know that, especially before doing the study) is that 50% of control infants will have death or disability and hypothermia will reduce this to 30%, we have an 80% chance of finding a “statistically significant” difference if we enroll 208 subjects.
Hypothermia Study

<table>
<thead>
<tr>
<th></th>
<th>Hypothermia (n=102)</th>
<th>Control (n=103)</th>
<th>RR (95%C)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Death or Disability</td>
<td>45 (44%)</td>
<td>64 (62%)</td>
<td>0.72 (0.54, 0.95)</td>
<td>0.01</td>
</tr>
</tbody>
</table>

- If outcome had been reduced from 50% to 40%, we would have concluded there was no “significant” difference. RR 0.80 (0.59, 1.09), p=0.16.

- If you were the parent, you might think that was an important benefit.

- It’s even worse if the outcome is less common and worse still if there are more than 2 treatment groups.

- Proposed study of lactoferrin and lactobacillus for prevention of NEC

<table>
<thead>
<tr>
<th></th>
<th>Lactoferrin</th>
<th>No Lactoferrin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lactobacillus</td>
<td>13%</td>
<td>18%</td>
</tr>
<tr>
<td>No Lactobacillus</td>
<td>18%</td>
<td>23%</td>
</tr>
<tr>
<td></td>
<td>15%</td>
<td>20%</td>
</tr>
</tbody>
</table>

- For 80% power, 1828 subjects will be needed.
Why is study power important?

- At best, it’s an educated guess.
- People designing a study should calculate it because they should have some clue about how big or small a difference they will be able to identify.
- There may be other reasons, eg feasibility pilot, for doing a small study, but it needs to be justified.
- If it’s not big enough to answer the questions, what’s the justification for putting patients through the hassle or risk?

What’s the deal with “non-inferiority” studies?

- In traditional statistical analyses, when investigator is trying to show a difference, the deck is more or less stacked against concluding that there is a difference.
- We require that the p value (likelihood that the observed difference was due to chance) is less than 5%.
- At most, we expect as power of 80-90% to identify a difference if there is one.
- We typically accept a greater likelihood of falsely concluding that there is no difference.
What’s the deal with “non-inferiority” studies?

- Non-inferiority study is one type of “equivalence” study (investigator wants to conclude that the treatments are the same).
- The statistical analysis is flipped the other way so that it’s not so easy to falsely conclude that there’s “no difference”.
- Investigator must specify the (usually fairly small) difference that they will accept as “equivalent”.
- Then the sample size required will usually be very large.

Let’s say you want to compare a new antihypertensive drug to placebo, and you want to have 80% power to detect a difference of ≥ 5 mm Hg in BP (SD = 15)?
- You need a sample size of 142/group

Let’s say you want to compare a new drug to an old drug and have 80% power to conclude that the BP is the same within 2 mm Hg (SD = 15)?
- You need a sample size of 964/group
Even RCTs Can Get the “Wrong” Answer

- Benefits may be overestimated because of
  - unmasked allocation
  - failure to mask caregivers or assessors
  - use of surrogate outcomes
  - analysis does not include all enrolled subjects ("intention to treat")

Unmasked Allocation

- The person(s) making/influencing the decision about whether the subject is enrolled are aware of which treatment the next subject will get.

- Not the same as unmasked treatment or unmasked outcome assessment.
Postmenopausal Women Randomly Assigned

Equivalent risk of cardiac disease

Postmenopausal Women “Randomly” Assigned

Next patient gets hormones

Symptoms: Enroll

No symptoms: Don’t enroll

Sicker, higher risk

Next patient gets placebo

Symptoms: Don’t enroll

No symptoms: Enroll

Less sick, lower risk

Unmasked Caregivers or Assessors

- RCT of steroids to facilitate weaning and extubation in premature infants:
- I think steroids will be beneficial so I wean the ventilator and attempt extubation more aggressively when I know the infants have received steroids → steroids appear to be beneficial only because I treated those infants differently.
- I think steroids will cause glucose intolerance so I give fewer calories to babies when I know they’re on steroids → steroids appear to have a harmful effect on growth only because I treated those infants differently.
Use of Surrogate Outcomes

- Laboratory test or physical finding that is believed to predict an outcome that is clinically important.
- Previous randomized trial showed that anti-arrhythmic drugs reduced EKG abnormalities in post-MI patients.
- Arrhythmias can be fatal.
- In a larger RCT, encainide or flecainide, as compared to placebo, increased post-MI mortality.

Intention-to-Treat

- Analyzing all subjects according to their group assignment, regardless of whether they received or completed the assigned treatment.

Postmenopausal Women Randomly Assigned

Hormones Equivalent risk of cardiac disease

High risk patients have more side effects, stop study drug

Survivors who continue study drug are relatively low risk

More deaths

Placebo Equivalent risk of cardiac disease

High risk patients continue taking placebo

Fewer deaths

Survivors who continue study drug are relatively high risk

Even RCTs Can Get the “Wrong” Answer

- Benefits or hazards may be underestimated because of
  - poorly implemented intervention
  - inadequate follow-up
Even RCTs Can Get the “Wrong” Answer for the Real Target Population

- Benefits may be overestimated for the target population (the ones where the treatment will be applied in clinical practice)
  - highly selected patients (highly sensitive to treatment, highly compliant, low risk of complications)
    - overestimates beneficial effects and underestimates adverse effects of use in a more diverse population
  - high risk of primary outcome

RSV Immunoglobulin for Prevention of RSV Hospitalization

- Eligibility criteria for hospitalized infants
  - < 35 weeks gestation at birth
  - No reporting of eligible non-enrolled subjects
  - Average gestational age was ≤ 32 weeks

Risk Difference = 
RD = 53/500 (10.6%) – 48/1002 (4.8%) = 4.8%

Is it worth it?

NNT: You need to treat 21 infants (5 injections each at $1500/injection) to prevent 1 RSV hospitalization

Palivizumab was FDA approved and recommended for infants < 35 weeks gestation

Risk of RSV hospitalization is ~ 2% for 33-34 week infants

If you decrease it by half (2%-1%), RD = 1%

You need to treat 100 infants (5 injections each at $1500/injection) to prevent 1 RSV hospitalization
Hypothermia study had FU for all but 3 subjects in the control group (98.6% follow-up).

Loss to follow-up isn’t a big problem if it’s random (non-differential loss).

If you lost 10% of the patients from each cell below, the RR would be the same; your p value would be higher but it would still be significant.

<table>
<thead>
<tr>
<th></th>
<th>Hypothermia (n=102)</th>
<th>Control (n=103)</th>
<th>RR (95%CI)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Death or Disability</td>
<td>45</td>
<td>64</td>
<td>0.72 (0.54, 0.95)</td>
<td>0.01</td>
</tr>
<tr>
<td>No death/disability</td>
<td>57</td>
<td>39</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

But what if it’s not random?
Why is high follow-up important?

- What if we had 0% loss of cooled infants and 0% loss of good outcomes and 15% loss of control infants with bad outcomes?

<table>
<thead>
<tr>
<th></th>
<th>Hypothermia (n=102)</th>
<th>Control (n=103)</th>
<th>RR (95%CI)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Death or Disability</td>
<td>45</td>
<td>64</td>
<td>0.72 (0.54, 0.95)</td>
<td>0.01</td>
</tr>
<tr>
<td>No death/disability</td>
<td>57</td>
<td>39</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Hypothermia (n=102)</th>
<th>Control (n=93)</th>
<th>RR (95%CI)</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Death or Disability</td>
<td>45</td>
<td>54</td>
<td>0.76 (0.58, 1.00)</td>
<td>0.052</td>
</tr>
<tr>
<td>No death/disability</td>
<td>57</td>
<td>39</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Key Design Features of a Treatment Study

- Randomized allocation
- Masked allocation
- Masked intervention (or other measures to ensure similar treatment of intervention and controls)
- Masked outcome assessment (or other measures to standardize assessment)
- Minimal loss to follow-up
- Intention-to-treat analysis
- Reporting of enrolled vs non-enrolled subjects
Clear as mud