Methodological and Statistical Issues in Research Proposals

Rich Jones, Tom Travison
Fah Vasunilashorn, Dae Kim, John Devlin

CEDARTREE 4th Annual Delirium Boot Camp
November 8, 2016
The Inn at Longwood Medical

In Five Parts

Part 1. Common problems (0:15)
Part 2. Tom (0:15)
Part 3. A checklist for a sample size justification (0:15)
Part 4. Focus topic: Propensity scores (0:15)
Part 5. Pilot proposals (1:15)
Part 1

Common methodological and statistical issues in research proposals

Specific Aims

• Clear and addressable and represent clear and potentially falsifiable research questions
• Hypotheses: are specific, include a contrast, and are testable given the design
Significance / Premise

• **High quality supporting evidence** supports the scientific premise (adequately powered existing, preliminary studies, pilot data are appropriately used). And/or
• Limitations of supporting evidence are acknowledged and addressed with respect to the scientific premise

Approach / Rigor

• **Data collection** descriptions are complete and clear
• What data points are being measured, by whom, at what occasion, for what purpose?
  – Ensure that potentially confounding variables are collected and specified
  – Clinical trials to be consistent with CONSORT must pre-specify adjustment variables and pre-specify subgroup analysis
• **Data quality** and preprocessing are appropriately described
• **Sample size** is explicit and clear (and justified, see separate sample size and power checklist)
Data Analysis

• Complexity is appropriate as complex as warranted, not overly so
• Well suited to answer the questions or test hypotheses
• Missing data is addressed in
  – design (avoiding drop out) and
  – analysis
• Sensitivity analyses are considered to assess impact of important assumptions

Relevant biological variables

• If sex differences are not specifically hypothesized, then at least include a plan to separately report effects by gender
Sample size/power

- Each aim has a power/sample/minimum detectable effect size documented
- Match between model for power/sample size and planned analysis
- Estimates on which power/sample size are based are
  - Appropriate
  - Derive from adequately powered preliminary studies or otherwise well justified
- Clarity and transparency in power/sample size presentation

Part 2

Tom
By Example:
Principles of Visual Data Display

Example: Randomized Clinical Trial

• Intervention: resistance exercise training to increase appendicular lean body mass (ALBM) among frail older adults
  – 3 dose groups (1 Hr, 2 Hrs, 4 Hrs per week training)
  – Attention control: Literature concerning benefits of physical activity, phone contacts
  – Duration: six months
  – Sample Size: N = 200 randomized (50 per group)
    • Assuming 10% cumulative attrition and missingness (45 participants evaluable per group at trial end), design obtains 80% power to detect standardized differences of at least 0.6 between any two groups
  – Primary Endpoint: Change in ALBM at 6 months post-randomization
Hypothesis

• Resistance training will be associated with greater mean increases in ALBM than attention control, and more frequent exercise will be associated with greater increases than less frequent (i.e. dose-response).

Results

• 192 individuals (96%) evaluable at 6 months (great!)
• Adherence to intervention (60% of participant contacts or greater): 83% (pretty great!)

• Some evidence that mean gains in ALBM behave in a dose-responsive fashion as expected
  – Control: 0.56 kg increase
  – 1 Hr: 0.52 kg increase
  – 2 Hr: 0.99 kg increase
  – 4 Hr: 1.48 kg increase

• How to display these data (192 values) for inspection?
• How to display the result in the average?
A regrettably common approach

Figure 1. ALBM by group

Extraneous ink / “information”

Figure 1. ALBM by group

Legend: Nonsensical, unnecessary, needlessly divides attention

Confusing use of frequency-type plotting for continuous mean

Three dimensions when 2 (1?) are needed

Variation in color – needlessly reproduces X axis, confuses eye
Missing information

Unexplained abbreviations
Units not given
No quantification of uncertainty
Failure to note these are means
No display of actual measurements

Figure 1. ALBM by group

A superior treatment ... as far as it goes

• If all one aims to do is show the means per group (not that this is recommended...), the following sophisticated display is superior:

  – Attention Control: 0.56 kg mean increase in ALBM
  – 1 Hr Training per Week: 0.52 kg mean increase in ALBM
  – 2 Hr Training per Week: 0.99 kg mean increase in ALBM
  – 4 Hr Training per Week: 1.48 kg mean increase in ALBM

• The actual sample mean values are given
• Units are provided
• Values are associated naturally with the participant groups (no legend)
• No extra colors, dimensions, distractions
But ... we should aim to do more

• For displaying data, show the data

• For estimation / inference concerning means, show uncertainty

• Provide more information in general

Candidate solution: data display

More appropriate use of boxplot / scatter for continuous measures (numerous alternatives)

Two dimensions - plenty

No duplication of information – vertical axis handles differentiation by group without color or shape

Direct labeling of groups (no legend) with horizontal text, the way humans read

Powerful combination of tabular and graphical information

Basic good practice: sample sizes, units provided; proper labeling, informative caption. Figure is self-explanatory.

Figure 1. Change in ALBM by group. Boxplots and participant measurements (dots) displayed.
Candidate solution: estimation of means

Figure 1. ALBM by group

Candidate solution: estimation of means

Figure 1. Change in ALBM by group. Means and 95% confidence intervals displayed
Rules for improvement

• Strive for decreased ink per information, and be sure ‘information’ is real
• Utilize tools appropriate to measurement types
• Inspect raw data, and where appropriate provide this to readers
• Good practice: give group sizes, units, proper scaling
• Annotation is powerful: provide tabular information as appropriate, kill legends if possible
• Figures must stand on their own, at minimum with assistance of captioning.
Part 3

A checklist for preparing a complete sample size justification

Bookmarks

Checklist
https://goo.gl/lGjfYY

Explanatory text
https://sites.google.com/site/ifarwf/home/samplesizeandpower
SAMPLE SIZE AND POWER ANALYSIS CHECKLIST

Click here to read an elaboration and explanation of the points below.

1. **Effect size specification.** The *minimum detectable* effect size is specified, OR the *hypothesized effect size is specified*.
   - Each hypothesis/aim has its own effect size specification

2. **Effect size justification.** There is a *scientific* justification for the effect size to be detected. The effect to be detected must be convincingly:
   - Feasible to observe
   - Based on previously published or pilot data from adequately powered studies, an essential part of scientific rigor
   - Clinical relevance/practical importance of effect to be detected is justified
   - Each hypothesis/aim has its own effect size justification

3. **Sample size justification.** The number of units to be studied (people, animals) is specified.

4. **Sample size justification.** The proposal contains a *scientific* justification that the number of units to be studied is both:
   - Feasible, the predicted and recruitment plan will support target enrollment
   - Feasible, the research team has the capacity to achieve goals

5. **Statistical power.** Statistical power is a function of the *hypothesized effect size* and the sample size.

6. **Sample size justification.** Each hypothesis/aim has its own statement of statistical power regarding the number of units to be studied.

7. **Special design considerations.** The following design considerations are described in the experimental design and sample size justification:
   - Missing data from other sources
   - Selection bias or other non-independence of observations
   - Measurement error or other non-independence of observations
   - Multiple endpoints in terms of all outcomes and/or pre-planned post-hoc outcomes; whether or not to adjust significance levels
   - Randomization or any other strategies for the effect on P-values
   - Comparison or other constraints and adjustments described that will affect statistical results

8. **Strategies to deal with challenges.** Strategies are described for monitoring and dealing with challenges in achieving target sample sizes.

9. **Statistical power.** Statistical power is a function of the *hypothesized effect size* and the sample size.

10. **Sample size justification.** Each hypothesis/aim has its own statement of statistical power regarding the number of units to be studied.

11. **Special design considerations.** The following design considerations are described in the experimental design and sample size justification:
    - Missing data from other sources
    - Selection bias or other non-independence of observations
    - Measurement error or other non-independence of observations
    - Multiple endpoints in terms of all outcomes and/or pre-planned post-hoc outcomes; whether or not to adjust significance levels
    - Randomization or any other strategies for the effect on P-values
    - Comparison or other constraints and adjustments described that will affect statistical results

12. **Strategies to deal with challenges.** Strategies are described for monitoring and dealing with challenges in achieving target sample sizes.

13. **Statistical power.** Statistical power is a function of the *hypothesized effect size* and the sample size.

14. **Sample size justification.** Each hypothesis/aim has its own statement of statistical power regarding the number of units to be studied.

15. **Special design considerations.** The following design considerations are described in the experimental design and sample size justification:
    - Missing data from other sources
    - Selection bias or other non-independence of observations
    - Measurement error or other non-independence of observations
    - Multiple endpoints in terms of all outcomes and/or pre-planned post-hoc outcomes; whether or not to adjust significance levels
    - Randomization or any other strategies for the effect on P-values
    - Comparison or other constraints and adjustments described that will affect statistical results

16. **Strategies to deal with challenges.** Strategies are described for monitoring and dealing with challenges in achieving target sample sizes.

17. **Statistical power.** Statistical power is a function of the *hypothesized effect size* and the sample size.

18. **Sample size justification.** Each hypothesis/aim has its own statement of statistical power regarding the number of units to be studied.
3. Sample size specification. The number of units to be studied (people, animals) is specified.

4. Sample size justification. The proposal contains sufficient justification that the number of units to be studied is both
   - Feasible: the population and recruitment plan will support target enrollment
   - Achievable: the research team has the capacity to achieve goals

5. Statistical power. Statistical power is a function of the hypothesized effect size and the sample size
   - Each hypothesis/aim needs its own statement of statistical power

6. Special design considerations. The following design considerations are reflected in the power analysis and sample size justification
   - Attrition or loss to follow-up
   - Missing data from other causes
   - Clustering or other non-independence of observations
   - Repeated measurements of outcome variables
   - Multiple comparisons in terms of multiple outcomes and/or pre-planned subgroup analyses, and whether or not to adjust significance levels
   - Preliminary analysis, or early stopping rules and their effect on P-values
   - Covariates or other confounders and adjustment accounted for
Part 4

Focus Topic: Propensity scores
Use and Interpretation of Propensity Scores in Aging Research: A Guide for Clinical Researchers

Dae Hyun Kim, MD, MPH, ScD,* † Carl F. Pieper, DrPH,‡ Ali Ahmed, MD, MPH,§ ¶ and Cathleen S. Colo’-Emeric, MD, MHS**††

- Confounding in observational studies
- What is propensity score?
- How to estimate propensity score
- How to use propensity score to estimate treatment effect
- Limitations of propensity score analysis

Example: antipsychotic safety in cardiac surgery patients with postoperative delirium

- Consider a database-based study to compare mortality associated with atypical vs. typical antipsychotics.

<table>
<thead>
<tr>
<th></th>
<th>Oral Atypical (N=2,580)</th>
<th>Oral Typical (N=1,126)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age, years</td>
<td>69</td>
<td>71</td>
</tr>
<tr>
<td>Charlson Comorbidity Index</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Ventilation days before drug initiation</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Blood transfusion, %</td>
<td>27</td>
<td>19</td>
</tr>
<tr>
<td>Teaching hospital, %</td>
<td>58</td>
<td>65</td>
</tr>
<tr>
<td>Total length of hospital stay, days</td>
<td>17</td>
<td>14</td>
</tr>
<tr>
<td>In-hospital mortality, %</td>
<td>7</td>
<td>4</td>
</tr>
</tbody>
</table>
Baseline imbalance in risk factors of outcome results in biased treatment effect: confounding

Confounder → Treatment → Outcome

How to adjust for confounding
- Restriction
- Matching
- Standardization
- Stratification
- Regression
- Weighting
- Other (e.g., G-estimation)

Achieve balance in risk factors between the treatment groups

Regression models are commonly used to adjust for multiple confounders

- Logistic regression model for outcome
  \[
  \text{Logit}(\Pr(Y=1)) = \beta_0 + \beta_1 X + \beta_2 C_1 + \beta_3 C_2 + \ldots + \beta_{11} C_{10}
  \]
  - \( Y = 1 \) if in-hospital death or 0 if survive
  - \( X = 1 \) if atypical drugs or 0 if typical drugs
  - \( C_1 = \text{age}; C_2 = \text{sex}, \ldots, C_{10} = \text{comorbidity score} \)

- Models the relationship of treatment and confounders with the outcome
  - This relationship should be correctly specified.
  - May not adjust for many confounders (8-10 outcomes per covariate)
We can develop a summary score that contains information on all confounders

\[
\text{Logit}(\Pr(Y=1)) = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot C_1 + \beta_3 \cdot C_2 + \ldots + \beta_{11} \cdot C_{10} + \beta_S \cdot C_S
\]

- **Disease risk score**: predicted risk of outcome given all confounders
  \[
  \text{Logit}(\Pr(Y=1)) = \alpha_0 + \alpha_1 \cdot C_1 + \alpha_2 \cdot C_2 + \ldots + \alpha_{10} \cdot C_{10}
  \]

- **Exposure propensity score**: predicted risk of exposure given all confounders
  \[
  \text{Logit}(\Pr(X=1)) = \gamma_0 + \gamma_1 \cdot C_1 + \gamma_2 \cdot C_2 + \ldots + \gamma_{10} \cdot C_{10}
  \]

Propensity score analysis has 2 steps

1. **Estimation of propensity score**
   - Use logistic model for treatment as a function of confounders
   - Evaluation of propensity score model (i.e., diagnostics)

2. **Estimation of treatment effect using propensity score**
   - Matching on propensity score
   - Weighting using propensity score-based weights
   - Stratification by propensity score quantiles (e.g., quintile)
   - Covariate adjustment in outcome regression model
     (discouraged)
What variables should be included in the propensity model?

- Model 1: X1
- Model 2: X1, X2
- Model 3: X1, X2, X3

Propensity score model should include confounders

- PS model is not the best prediction model of treatment status.
- Including X2 (instrumental variable) reduces precision and may increase bias by unmeasured confounders.
- X3 (intermediate variable) in the causal pathway may obscure part of treatment effect.
How can propensity score model be evaluated?

- Assess balance of confounders between the groups.
- Use a metric that is specific to sample and not affected by sample size (standardized mean difference < 0.1)
- Significance testing, measures of model fit, or C statistics do not inform whether PS model is correctly specified.

### Before PS matching

<table>
<thead>
<tr>
<th>Age, years</th>
<th>Typical</th>
<th>Atypical</th>
<th>SMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>N=1,126</td>
<td>70.9</td>
<td>68.9</td>
<td>-0.18</td>
</tr>
<tr>
<td>N=2,580</td>
<td>70.8</td>
<td>69.9</td>
<td>0.09</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Charlson index</th>
<th>Typical</th>
<th>Atypical</th>
<th>SMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>N=1,126</td>
<td>2.6</td>
<td>3.9</td>
<td>0.19</td>
</tr>
<tr>
<td>N=2,580</td>
<td>2.7</td>
<td>2.7</td>
<td>0.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Transfusion, %</th>
<th>Typical</th>
<th>Atypical</th>
<th>SMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>N=1,126</td>
<td>19.4</td>
<td>26.9</td>
<td>0.18</td>
</tr>
<tr>
<td>N=2,580</td>
<td>23.0</td>
<td>21.9</td>
<td>-0.03</td>
</tr>
</tbody>
</table>

### After PS matching

<table>
<thead>
<tr>
<th>Age, years</th>
<th>Typical</th>
<th>Atypical</th>
<th>SMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>N=832</td>
<td>70.8</td>
<td>69.9</td>
<td>0.09</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Charlson index</th>
<th>Typical</th>
<th>Atypical</th>
<th>SMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>N=832</td>
<td>2.7</td>
<td>2.7</td>
<td>-0.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Transfusion, %</th>
<th>Typical</th>
<th>Atypical</th>
<th>SMD</th>
</tr>
</thead>
<tbody>
<tr>
<td>N=832</td>
<td>23.0</td>
<td>21.9</td>
<td>-0.03</td>
</tr>
</tbody>
</table>

Propensity score distribution may provide insights into confounding and inference

- PS model separates treated vs. untreated patients very well (i.e., small overlap).
  - Consistent clinical practice (guideline)
  - Strong confounding
  - Inclusion of instrumental variable

- PS model does not separate treated vs. untreated patients (i.e., large overlap).
  - Clinical uncertainty (no guideline)
  - Weak confounding
  - Omission of important confounders

Treatment effect is estimated for study population within the range of common support (PS overlap)
Propensity score distribution of patients treated with atypical vs. typical drugs

- No clinical guidelines for choice of antipsychotic drugs
- Confounding is likely less in active comparator design
- Missing important confounders: e.g., delirium severity

Matching achieves balance by selecting treated and untreated patients with similar features

Study Population (N = 10)  PS = Pr(A=1)

Untreated Patients (A=0)  Treated Patients (A=1)

0.20  0.50  0.66  0.66

Matched Population (N = 6)

Untreated Patients (A=0)  Treated Patients (A=1)

0.20  0.30  0.66  0.66

Matching
Matching allows intuitive and transparent analysis to estimate treatment effect

- Matching results in 2 groups with similar characteristics
  - Matching algorithm: optimal or greedy, ratio (1:1 or 1:n)
  - Use of caliper (maximum distance in PS allowed within a pair)
- Allows transparent assessment of confounder balance
- Estimates average treatment effect (ATE) for a smaller group (typically, treated patients)
- Limits generalizability and reduces power

<table>
<thead>
<tr>
<th></th>
<th>Atypical (%)</th>
<th>Typical (%)</th>
<th>RR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crude analysis</td>
<td>183 / 2580 (7.1)</td>
<td>49 / 1126 (4.4)</td>
<td>1.7 (1.2, 2.3)</td>
</tr>
<tr>
<td>1:1 nearest neighbor matching</td>
<td>45 / 832 (5.4)</td>
<td>44 / 832 (5.3)</td>
<td>1.0 (0.7, 1.5)</td>
</tr>
</tbody>
</table>

Weighting achieves balance by giving different weights to treated and untreated patients
Weighting vs. matching

- Both methods achieve balance better (less bias) than stratification or covariate adjustment.
- Weighting is less efficient than matching (wider 95% CI)
- May be subject to few participants with extreme weight
- Can be extended to account for time-dependent confounding (IPTW) and censoring (IPCW)

<table>
<thead>
<tr>
<th></th>
<th>Atypical (%)</th>
<th>Typical (%)</th>
<th>RR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crude analysis</td>
<td>183 / 2580 (7.1)</td>
<td>49 / 1126 (4.4)</td>
<td>1.7 (1.2, 2.3)</td>
</tr>
<tr>
<td>1:1 nearest neighbor matching</td>
<td>45 / 832 (5.4)</td>
<td>44 / 832 (5.3)</td>
<td>1.0 (0.7, 1.5)</td>
</tr>
<tr>
<td>Weighting (ATE in total population)</td>
<td>237 / 3665 (6.5)</td>
<td>187 / 3630 (5.2)</td>
<td>1.3 (0.9, 1.8)</td>
</tr>
<tr>
<td>Weighting (ATE in patients treated with atypical)</td>
<td>183 / 2580 (7.1)</td>
<td>138 / 2505 (5.5)</td>
<td>1.3 (0.9, 1.9)</td>
</tr>
<tr>
<td>Weighting (ATE in patients treated with typical)</td>
<td>54 / 1085 (5.0)</td>
<td>49 / 1126 (4.4)</td>
<td>1.1 (0.8, 1.6)</td>
</tr>
</tbody>
</table>

Stratification estimates treatment effect using weighted average of stratum-specific effects

- Assess confounder balance within each PS quantile
- Weighted average of stratum-specific treatment effects
  - Non-uniform stratum-specific treatment effects: residual confounding or treatment effect heterogeneity?

<table>
<thead>
<tr>
<th></th>
<th>Atypical (%)</th>
<th>Typical (%)</th>
<th>RR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Crude analysis</td>
<td>183 / 2580 (7.1)</td>
<td>49 / 1126 (4.4)</td>
<td>1.7 (1.2, 2.3)</td>
</tr>
<tr>
<td>PS quintile 1</td>
<td>8 / 236 (3.4)</td>
<td>16 / 505 (3.2)</td>
<td>1.1 (0.5, 2.5)</td>
</tr>
<tr>
<td>PS quintile 2</td>
<td>25 / 464 (5.4)</td>
<td>10 / 277 (3.6)</td>
<td>1.5 (0.7, 3.2)</td>
</tr>
<tr>
<td>PS quintile 3</td>
<td>35 / 569 (6.2)</td>
<td>13 / 172 (7.6)</td>
<td>0.8 (0.4, 1.6)</td>
</tr>
<tr>
<td>PS quintile 4</td>
<td>55 / 629 (8.7)</td>
<td>7 / 112 (6.3)</td>
<td>1.4 (0.6, 3.2)</td>
</tr>
<tr>
<td>PS quintile 5</td>
<td>60 / 682 (8.8)</td>
<td>3 / 59 (5.1)</td>
<td>1.8 (0.5, 5.9)</td>
</tr>
<tr>
<td>PS stratification (weighted average)</td>
<td>183 / 2580 (7.1)</td>
<td>138 / 2505 (5.5)</td>
<td>1.3 (0.9, 1.9)</td>
</tr>
</tbody>
</table>
What propensity score analysis cannot do

• Propensity score analysis cannot adjust for confounders that are unmeasured or measured with error.

• Reduce measurement error in confounder assessment

• Alternative approaches for unmeasured confounding
  – Compare two active treatments instead of treated vs. untreated
  – Sensitivity analysis under various confounding assumptions
  – Find another dataset with information on unmeasured confounders in similar population (e.g., PS calibration)
  – Instrumental variable analysis

Take-home points

• The aim of propensity score is to balance confounders between treatment groups.

• Matching and weighting achieve better balance (less bias) than stratification or covariate adjustment.
  – Target population for inference may be different across methods.

• Propensity score does not adjust for confounders that are unmeasured or measured with error.
  – Conduct sensitivity analysis.
Part 5

Proposals

2:00 – 2:15
Rich Jones
Shanna Burke
Group differences in measurement properties of diagnostic or screening tools

Prediction noninvariance is not indicative of measurement bias
Measurement Invariance Versus Selection Invariance: Is Fair Selection Possible?

Denny Borsboom
University of Amsterdam

Jan-Willem Romeijn
Groningen University

Jelte M. Wicherts
University of Amsterdam

This article shows that measurement invariance (defined in terms of an invariant measurement model in different groups) is generally inconsistent with selection invariance (defined in terms of equal sensitivity and specificity across groups). In particular, when a unidimensional measurement instrument is used and group differences are present in the location but not in the variance of the latent distribution, sensitivity and positive predictive value will be higher in the group at the higher end of the latent dimension, whereas specificity and negative predictive value will be higher in the group at the lower end of the latent dimension. When latent variances are unequal, the differences in these quantities depend on the size of group differences in variances relative to the size of group differences in means. The effect originates in a special case of Simpson's paradox, which arises because the observed score distribution is collapsed into an accept-reject dichotomy. Simulations show the effect can be substantial in realistic situations. It is suggested that the effect may be partly responsible for overprediction in minority groups typically found in empirical studies on differential academic performance. A methodological solution to the problem is suggested, and social policy implications are discussed.


Sensitivity, Specificity

<table>
<thead>
<tr>
<th>Disease</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>0</td>
<td>FN</td>
<td>TN</td>
</tr>
</tbody>
</table>

$SN = \frac{TP}{TP + FN}$

$SP = \frac{TN}{TN + FP}$

Screening test
Screening test

Assumed proportion with disease given screening test

Actual proportion with disease as determined by gold standard

Disease

Low Test Score

Distribution of People

TP FP
FN TN

SN = TP / TP + FN
SP = TN / TN + FP

1.0 Proportion with disease

0.5

Test Score
Suggestions

• Clarify question
• Re-specify population, sample
• Identify instruments
• Consider
  – novel methods approach: use weighting
  – or, latent class analysis for diagnostic agreement
Dr. Racine: neuroimaging markers, delirium, and long-term cognitive decline

• N=146 (up to 60 months of follow-up)
• Aim 3: linear mixed effects model for repeated measures
  – **Outcome**: global cognitive performance (continuous)
  – **Main effects**: cortical atrophy (low, medium, high), delirium (yes/no), time, age, sex, education (continuous)
  – **Interaction**: cortical atrophy*delirium, cortical atrophy*time, delirium*time, cortical atrophy*delirium*time
  – **Random effects**
• The study is 80% powered to detect standardized effect size 0.63 at type 1 error rate 5%, which is a large effect.
Small studies are less likely to detect a true non-null effect

- The probability that your results with p<.05 reflect a true non-null effect depends on 2 factors:
  - Pre-study odds that the effect is truly non-null
  - Statistical power of your study

Suppose 1 in 5 tested hypotheses are truly non-null in the neuroscience field (e.g., pre-study odds = 1/4 = .25).

If you find p<.05, the chance that your findings are true is:
- If statistical power .10: 33%
- If statistical power .30: 60%
- If statistical power .80: 80%

Even if true effect is detected in small studies, the effect is likely exaggerated

- Small studies can only detect large effects.
- If the true effect is modest, the estimate of the true effect that happened to be large will only be detected.

"Winner’s Curse"

Suppose the true effect is OR 1.2. Due to random error and sampling variation, your study may find an OR of 1.0, 1.2, or 1.6.

Since OR 1.0 or 1.2 does not reach p<.05, you will only claim discovery of non-null effect when random error creates OR 1.6.
Some recommendations

• Perform an a priori power calculation based on the effect size from the existing literature, and design your study
• If your study is underpowered, acknowledge this and disclose methods and findings transparently
• Clarify your analysis as confirmatory or exploratory
• Pre-register your study protocol
• Make raw study data available for meta-analysis
• Work collaboratively to increase power and replicate findings

Primary Objective

• To investigate the effect of pharmacological conversion of hyperactive delirium into hypoactive delirium on hospital mortality of acutely ill older adults.
• (Null hypothesis: hospital mortality of acutely ill older adults is not associated with pharmacological conversion of hyperactive delirium into hypoactive delirium)
Approach

• Prospective cohort study; N = 65 ‘per group’

• Primary endpoint: time to death in hospital

• Multiple measures of delirium and delirium subtypes

• Analysis of associations between exposures and delirium subtypes and transitions

• Biomarker profiles for subtypes (hyperactive, hypoactive, ‘mixed’)

Strengths

• Significance and novelty seem clear

• Design seems appropriate overall, though diversity within cohorts may cause difficulty
Points for clarification / discussion

• As described, analytic approach is sound
  —i.e. choice of methods seems appropriate

• Major source of confusion: lack of definition of comparison groups
  —Defined by delirium subtypes, or rather by exposures, or both?

• Project appears oriented toward transition, but design and analytic plan
do not make clear how this should be measured and attacked

• If groups are ill-defined, unclear if biomarker analysis can succeed

• Sample size is not reassuring given above complexities

2:45 – 3:00

Fah Vasunilashorn

Sophia Wang
Quantitative Challenges - Wang

• Each Aim linked to a hypothesis

Example: Aim 1

“Estimate changes in cognitive, functional and behavioral systems...in patients receiving Critical Care Recovery Center (CCRC)...”

Hypothesis: Relative to patients ‘not in the CCRC’ (define the group), patients in the CCRC will have a significantly less steep decline in the Healthy Aging Brain Care Monitor (HABC-M).
Quantitative Challenges - Wang

- Each Aim linked to a hypothesis
- Matching vs. multivariable adjustment
- Interaction effect sizes

Qualitative Challenges - Wang

- Sampling strategies
- Analyzing data – thematic coding
- Reliability
- Validity
The Role of Pilot Studies in Clinical Research

John W. Devlin, PharmD
Northeastern University
Tufts Medical Center
Role of Pilot Studies

• Also known as ‘feasibility’ or ‘proof of concept’ studies
• Examine the feasibility of an approach that is intended to be used in a larger scale study
  – Will enhance the probability of success in larger, subsequent RCTs.
• Should not be a hypothesis-testing study
  – Safety, efficacy and efficiency are generally not evaluated
  – Does not have a role in providing a ‘signal’ of efficacy
  – Power analysis should not be included
    • Sample size should be based on “pragmatic” considerations
  – Should not be used to guide the sample size of future RCTs

Chmura Kraemer H et al. Arch Gen Psychiatry 2006; 63:484-9

Structure of Pilot Investigations

• Feasibility:
  – Recruitment
  – Randomization
  – Retention
  – Intervention
    • Implementation
    • Education
    • Adherence
    • Satisfaction
  – Assessment procedures
    • Efficacy
    • Safety
• A control group should still be incorporated as there may be distinct feasibility issues when a blinded, “placebo” intervention is incorporated in future RCT

Chmura Kraemer H et al. Arch Gen Psychiatry 2006; 63:484-9
Immortal time bias produces results in favor of the treatment group

• Determination of treatment status involves a wait period during which follow-up time is accrued.
  – This wait period is immortal time (i.e., the study outcome cannot occur by design).

Levesque et al. BMJ 2010; 340: b5087
rich_jones@brown.edu
tgt@hsl.harvard.edu
svasunil@bidmc.harvard.edu
dkim2@bidmc.harvard.edu
j.devlin@northeastern.edu

questions