Flaws in statistical analysis

• How much time do we have?
• There are lies, damn lies, and statistics (B. Disraeli)
• If you use statistics to lie, you are the liar not the statistic

Most common flaws

• inappropriate or incomplete analysis, including violations of model assumptions and analysis errors,
• improperly addressing missing data, and
• power/sample size concerns.
  • Fernandes-Taylor, BMC, 2011

How do you deal with multiple endpoints?
Example Study (Loprinzi, JCO, 2002)

- A study for the efficacy of venlafaxine for hot flashes involved two treatment groups (Venlafaxine and placebo respectively) and the following endpoints:
  - Hot flash frequency per day
  - Hot flash average severity per day
    - none, mild, moderate, severe, very severe
    - scored 0, 1, 2, 3, 4
  - Hot flash score (severity times frequency)
  - Uniscale QOL
  - Hot flash affect on QOL
  - Toxicity incidence on 11 variables

Challenge

- What is the optimal way to deal with the multiplicity of endpoints available for analysis in this study?
  a) Pick a primary and make all else secondary
  b) Use a Bonferroni-type correction
  c) Use Hochberg’s step-up procedure
  d) Use an O’Brien global test

Results: Venlafaxine versus placebo

<table>
<thead>
<tr>
<th>Variable</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>HF frequency</td>
<td>0.0001</td>
</tr>
<tr>
<td>HF severity</td>
<td>0.04</td>
</tr>
<tr>
<td>HF Score</td>
<td>0.007</td>
</tr>
<tr>
<td>Uniscale QOL</td>
<td>0.0002</td>
</tr>
<tr>
<td>Hot flash affects QOL</td>
<td>0.01</td>
</tr>
<tr>
<td>Toxicity (11 vars)</td>
<td>all &gt;0.25</td>
</tr>
</tbody>
</table>
Bonferroni-type correction

- 16 variables tested, divide experiment-wise Type I error rate of 5% by 16 \( \Rightarrow 0.003125 \), use as comparison-wise significance level
- 2 of 16 p-values meet this criteria
- Four of 5 QOL-related p-values <0.01
- No toxicity p-values <0.05

Results: Bonferroni Approach

<table>
<thead>
<tr>
<th>Variable</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>HF frequency</td>
<td>0.0001</td>
</tr>
<tr>
<td>HF severity</td>
<td>0.04</td>
</tr>
<tr>
<td>HF Score</td>
<td>0.007</td>
</tr>
<tr>
<td>Uniscale QOL</td>
<td>0.0002</td>
</tr>
<tr>
<td>Hot flash affects QOL</td>
<td>0.01</td>
</tr>
<tr>
<td>Toxicity (11 vars)</td>
<td>all &gt;0.25</td>
</tr>
</tbody>
</table>

Hochberg’s Step-up Procedure

<table>
<thead>
<tr>
<th>Variable</th>
<th>P-value</th>
<th>( \alpha )</th>
</tr>
</thead>
<tbody>
<tr>
<td>HF frequency</td>
<td>0.0001</td>
<td>0.0031</td>
</tr>
<tr>
<td>Uniscale QOL</td>
<td>0.0002</td>
<td>0.0033</td>
</tr>
<tr>
<td>HF Score</td>
<td>0.007</td>
<td>0.0036</td>
</tr>
<tr>
<td>Hot flash affects QOL</td>
<td>0.01</td>
<td>0.0038</td>
</tr>
<tr>
<td>HF severity</td>
<td>0.04</td>
<td>0.0042</td>
</tr>
<tr>
<td>Toxicity (11 vars)</td>
<td>all &gt;0.25</td>
<td></td>
</tr>
</tbody>
</table>
Hochberg's Step-up Procedure

<table>
<thead>
<tr>
<th>Variable</th>
<th>P-value</th>
<th>α</th>
</tr>
</thead>
<tbody>
<tr>
<td>HF frequency</td>
<td>0.0001</td>
<td>0.0031</td>
</tr>
<tr>
<td>Uniscale QOL</td>
<td>0.0002</td>
<td>0.0033</td>
</tr>
<tr>
<td>HF Score</td>
<td>0.007</td>
<td>0.0036</td>
</tr>
<tr>
<td>Hot flash affects QOL</td>
<td>0.01</td>
<td>0.0038</td>
</tr>
<tr>
<td>HF severity</td>
<td>0.04</td>
<td>0.0042</td>
</tr>
<tr>
<td>Toxicity (11 vars)</td>
<td>all &gt;0.25</td>
<td></td>
</tr>
</tbody>
</table>

O'Brien Global Test for Multiple Outcomes

- Example: Venlafaxine for Hot Flashes (Sloan et al., JCO, 19(23):4280-4290, 2001)
- Hot flash frequency per day
  - Hot flash average severity per day
  - none, mild, moderate, severe, very severe
  - scored 0, 1, 2, 3, 4
- Hot flash score (severity times frequency)
- Uniscale QOL
- Hot flash affect on QOL
- Toxicity incidence on 11 variables

O'Brien p-values

<table>
<thead>
<tr>
<th>Endpoints Included</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hot Flash Frequency</td>
<td>0.0071</td>
</tr>
<tr>
<td>Hot Flash Average Severity</td>
<td>0.0050</td>
</tr>
<tr>
<td>Uniscale QOL</td>
<td>0.7528</td>
</tr>
<tr>
<td>Hot Flash Affects QOL</td>
<td></td>
</tr>
<tr>
<td>Toxicity</td>
<td></td>
</tr>
</tbody>
</table>
Summary

- Pick one: hf frequency ➔ significant
- Bonferroni ➔ significant
- Hochberg ➔ significant
- O’Brien ➔ significant

- Question: have you ever ignored a p-value <0.05? Even in the presence of multiple testing?

How do you handle the problem of missing data?

Non-random Missing-ness: The worst performers leave
Impact of hydrazine sulfate on colorectal cancer patient QOL.

AUC Comparison
Between HS and Placebo
by Imputation Method

<table>
<thead>
<tr>
<th>Method</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete</td>
<td>0.03</td>
</tr>
<tr>
<td>AVCF</td>
<td>0.79</td>
</tr>
<tr>
<td>LVCF</td>
<td>0.79</td>
</tr>
<tr>
<td>MVCF</td>
<td>0.79</td>
</tr>
<tr>
<td>ZVCF</td>
<td>0.29</td>
</tr>
<tr>
<td>OA</td>
<td>0.99</td>
</tr>
</tbody>
</table>
A study examining the efficacy of sucralfate to alleviate stomatitis

Mean Daily Stomatitis

Mean Stomatitis Score

Sucralfate
Placebo

Intent to treat analysis results

- AUC analysis, sucralfate vs placebo \( p \)-value=0.06 in favor of sucralfate
- twice as many patients went off study early on sucralfate arm
- all but 3 patients on sucralfate arm were off due to gagging
- add these folks back in as failures: \( p \)-value=0.06 in favor of placebo

How do you determine clinical significance

...and this is where we put the non-significant results.
A trend of trends

(barely) not statistically significant (p=0.052); a barely detectable statistically significant difference (p=0.072); a borderline significant trend (p=0.08); a certain trend toward significance (p=0.11); a clear tendency to significance (p=0.052); a clear trend (p=0.08); a clear, strong trend (p=0.08); a considerable trend toward significance (p=0.08); a decreasing trend (p=0.08); a definite trend (p=0.08); a distinct trend toward significance (p=0.07); a favourable trend (p=0.08); a favourable statistical trend (p=0.08); a little significant (p=1.0)

https://mchankins.wordpress.com/2013/04/21/still-not-significant-2/

A trend of trends

"trend towards significance" expresses non-significance as some sort of motion towards significance, which it isn't; there is no 'trend', in any direction, and nowhere for the trend to be 'towards'.

Think of it AS PREGNANCY, you either are or you are not.

Or "Do or do not, there is no try"  Yoda

What is a clinically meaningful effect?
What Clinical significance is NOT

- Statistical significance

- Example drawn from JCO 2001 (anonymous)
  - HSQ before / after scores on 1300 patients
  - all p-values <0.0001
  - conclusion: all domains of QOL were significantly different across treatment groups
  - problem: 1300 patients provides 80% power to detect a change of 1 unit on 0-100 point scale

Are these differences clinically meaningful?

<table>
<thead>
<tr>
<th>Item</th>
<th>n=537</th>
<th>n=346</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coughing</td>
<td>46.2</td>
<td>44.3</td>
<td>small</td>
</tr>
<tr>
<td>Dyspnea</td>
<td>17.2</td>
<td>16.2</td>
<td>small</td>
</tr>
<tr>
<td>Pain</td>
<td>26.9</td>
<td>25.5</td>
<td>small</td>
</tr>
</tbody>
</table>

- all p-values were statistically significant

Clinical Significance:

- Key Literature

- Developed ½ standard deviation method as accepted criterion (10 points on 0-100 scale)
  - Sloan: Cancer Integrative Medicine, 2003
  - Dueck: 2007 J Biopharm Stats (under review)
  - Sloan: J Chronic Obs Pul Dis, 2005
- Fostered development of state of the science consensus and standards
  - Guyatt, MCP 2002 – over 75 citations
  - Wyrwich, QOLR, 2005
  - Over 20 publications since 2001
Bottom Line

• Assessing the clinical significance of QOL can be as simple as a 10-point change on a 100-point scale, if that is consistent with the goals of the scientific enquiry.


Two general methods for clinical significance

• Anchor-based methods requirements
  – independent interpretable measure (the anchor) which has appreciable correlation between anchor and target

• Distribution-based methods
  – rely on expression of magnitude of effect in terms of measure of variability of results (effect size)

- QOL tool range = 6 standard Deviations
- SD Estimate = 100 percent / 6
  = 16.7% of theoretical range
- Two-sample t-test effect sizes
  small, moderate, large effect (0.2, 0.5, 0.8 SD shift)
- S,M,L effects = 3%, 8%, 12% of range

Figure 1 Power Example for Varying n
n = observations per sample
two—sided two sample t—test
All Methods Give Similar Answers

- Cohen - 1/2 SD is moderate effect
- MCID - 1/2 point on 7-point Likert
  - 7-1 = 6 point range ==> SD of 1 unit
  - so 1/2 point = 1/2 SD
- Cella - 10 point on FACT-G
  - 10/1.12 = 8.9% / 16.7% = 1/2 SD
- Feinstein - correlation approach
  - Cohen was arbitrary, should be 0.6 SD

There are more similarities than differences

- Statistical, Philosophical, Empirical, Clinical, Historical, Practical
  significant differences are all in the same ballpark
- All are animals of a slightly different shape and size but none are
  clinically distinct from one another
- The different approaches produce differences that are within the
  measurement error of the scales used

Four Guidelines
(Sloan, Cella, Hays, JCE 2005)

- The method used to obtain an estimate of clinical significance should be
  scientifically supportable.
- The 1/2 SD is a conservative estimate of an effect size that is likely to be
  clinically meaningful. An effect size greater than 1/2 SD is not likely to be
  one that can be ignored. In the absence of other information, the 1/2 SD is
  a reasonable and scientifically supportable estimate of a meaningful
  effect.
Four Guidelines
(Sloan, Cella, Hays, JCE 2005)

• Effect sizes below ½ SD, supported by data regarding the specific characteristics of a particular QOL assessment or application, may also be meaningful. The minimally important difference may be below ½ SD in such cases.

• If feasible, multiple approaches to estimating a tool's clinically meaningful effect size in multiple patient groups are helpful in assessing the variability of the estimates. However, the lack of multiple approaches with multiple groups should not preemptively restrict application of information gained to date.

Summary

• Defining clinical significance is today where pain was 25 years ago, tumor response was 50 years ago and blood pressure was 100 years ago

• Define clinical significance a priori, and use the definition in the analytical process

• Consensus is building as the answers from different approaches are similar and relatively robust

A ½ standard deviation for other endpoints?

• The question arises as to whether this sort of calibration can be made for non-QOL endpoints such as survival and tumor response using the same ½ standard deviation approach.

• Major et al, 2014, ASCO, “Effect sizes for phase II and Phase III clinical trials using the ½ SD rule.

• So we can now produce a calibrated effect size for any endpoint
Calibrated Effect Size Example

- VISTA: median PFS of melphalan and prednisone with bortezomib in previously untreated patients with multiple myeloma who were ineligible for high-dose therapy was 24 months compared to 16.6 months without bortezomib (p<0.001)
- ES=(24-16.6)/(16.6/ln2)=0.31
- Small/Medium Effect Size

Effect Sizes for 23 multiple myeloma treatments

Summary of recommended targets for meaningful clinical trial goals
So What?

- This method makes for ready comparison across different oncology trials.
- Clinicians can use calibrated effect size in the design of future clinical trials.
- Provides a mathematically based effect-size that can be gauged by clinical opinion.
- It provides a mechanism for comparing the effect sizes of QOL outcomes, survival outcomes and toxicity outcomes on one scale.

Interpreting survival curves

A few points about survival curves
Censoring on survival curves

Survival analysis assumes censoring is random. Censoring times vary across individuals and are not under the control of the investigator. Random censoring also includes designs in which observation ends at the same time for all individuals, but begins at different times.

Censoring is important

Confidence intervals are helpful
Which model to use?
- Kaplan-Meier, Logrank, nonparametric, gamma, Wilcoxon alternatives
- Give different emphasis to different aspects of the curves
- With $n>100$, all converge
- Take your pick

Hopefully these ideas will enable you to make advances in science