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

Variable	P-value
HF frequency	0.0001
HF severity	0.04
HF Score	0.007
Uniscale QOL	0.0002
Hot flash affects QOL	0.01
Toxicity (11 vars)	all >0.25

Bonferroni-type correction

- 16 variables tested, divide experiment-wise Type I error rate of 5% by 16 ➔ 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

Variable	P-value
HF frequency	0.0001
HF severity	0.04
HF Score	0.007
Uniscale QOL	0.0002
Hot flash affects QOL	0.01
Toxicity (11 vars)	all >0.25

Hochberg's Step-up Procedure

Variable	P-value	α
HF frequency	0.0001	0.0031
Uniscale QOL	0.0002	0.0033
HF Score	0.007	0.0036
Hot flash affects QOL	0.01	0.0038
HF severity	0.04	0.0042
Toxicity (11 vars)	all >0.25	

Hochberg's Step-up Procedure

Variable	P-value	α
HF frequency	0.0001	0.0031
Uniscale QOL	0.0002	0.0033
HF Score	0.007	0.0036
Hot flash affects QOL	0.01	0.0038
HF severity	0.04	0.0042
Toxicity (11 vars)	all >0.25	

O'Brien Global Test for Multiple Outcomes

- Example: Venlafaxine for Hot Flashes *(Sloan et at, 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

Endpoints Included	p-value
Hot Flash Frequency	
Hot Flash Average Severity	0.0071
Hot Flash Score	0.0050
Uniscale QOL	0.7528
Hot Flash Affects QOL	
Toxicity	

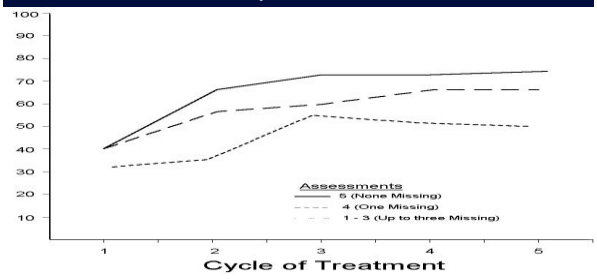
Summary

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

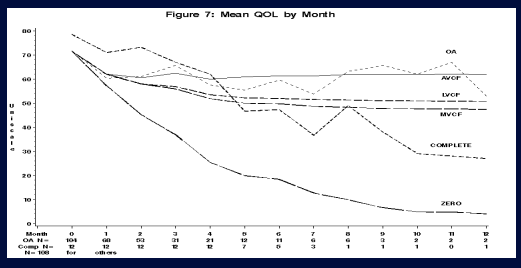
- Question: have you ever ignored a p-value <0.05? Even in the presence of multiple testing?

How do you handle the problem of missing data?

Non-random Missing-ness: The worst performers leave

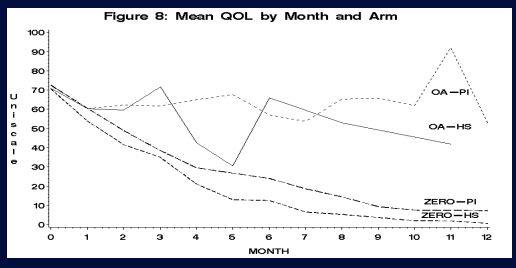


Impact of hydrazine sulfate on colorectal cancer patient QOL



Impact of different imputation methods for missing data

Effect of imputation method on treatment comparison

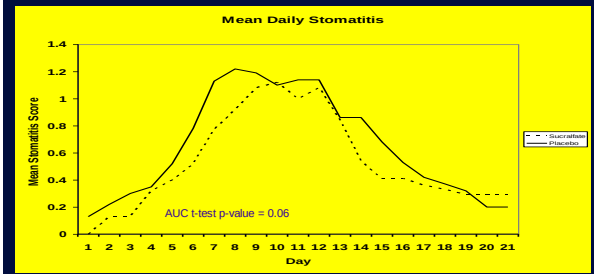


Sloan et al, JCO 16:3662-3673, 1998.

AUC Comparison Between HS and Placebo by Imputation Method

Method	P-value
Complete	0.03
AVCF	0.79
LVCF	0.79
MVCF	0.79
ZVCF	0.29
OA	0.99

A study examining the efficacy of sucralfate to alleviate stomatitis



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

A trend of trends

(barely) not statistically significant (p=0.052) a barely detectable statistically significant difference (p=0.073) a borderline significant trend (p=0.09) a certain trend toward significance (p=0.08) a clear tendency to significance (p=0.052) a clear trend (p<0.09) a clear, strong trend (p=0.09) a considerable trend toward significance (p=0.069) a decreasing trend (p=0.09) a definite trend (p=0.08) a distinct trend toward significance (p=0.07) a favorable trend (p=0.09) a favourable statistical trend (p=0.09) a little significant (p<0.1)

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

A trend of trends

"a 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 your 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?

Item	n=537	n=346	Effect Size
Coughing	46.2	44.3	small
Dyspnea	17.2	16.2	small
Pain	26.9	25.5	small

- 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
  - Norman: Exp Rev Pharmaco Outcomes Res, 2004
- 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.

(Sloan, J Chronic Obs. Pul. Dis. 2: 57-62, 2005.)

Presenting global solutions is always interesting

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)*

The MID method in one slide

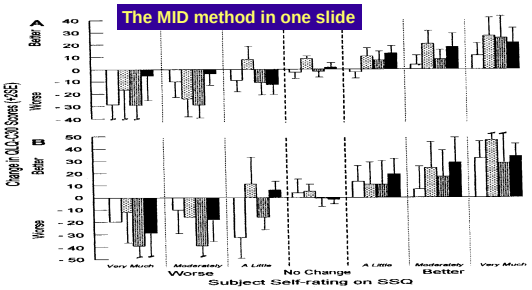
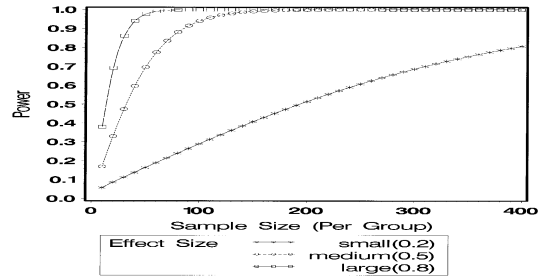


Fig 1. Relationship between SSQ ratings of change and QLQ-C30 scores from T1 to T2 for patients receiving chemotherapy for either breast cancer (A) or SCLC (B). Columns represent mean scores + 2 SE. □, physical functioning; ▨, emotional functioning; ⊟, social functioning; ■, global QL.

The Empirical Rule Effect Size (ERES) Approach *(Sloan et al, Cancer Integrative Medicine 1(1):41-47, 2003)*

- QOL tool range = *6 standard Deviations*
- SD Estimate = *100 percent / 6*
  = 16.7% of theoretical range
- Two-sample t-test effect sizes *(J Cohen, 1988)*:
  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

(Norman, Sloan, Wyrwich. Pharmaco. and Outcomes Research 4(5):515 – 519, 2004)

- 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 ½ SD is a conservative estimate of an effect size that is likely to be clinically meaningful. An effect size greater than ½ SD is not likely to be one that can be ignored. In the absence of other information, the ½ 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?

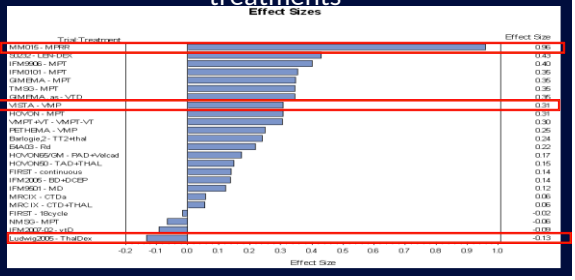
- The question arises as to whether this sort of calibration can be made for non-QOL endpoints such as survival and tumor response using the same ½ standard deviation approach.
- Major et al, 2014, ASCO, "Effect sizes for phase II and Phase III clinical trials using the ½ SD rule.
- So we can now produce a calibrated effect size for any endpoint

Calibrated Effect Size Example

San Miguel et al. N Engl J Med 2008; 359:906-17

- 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

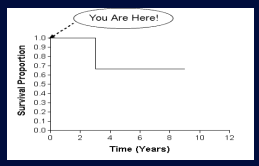
Cancer Type	Patient Population	Current Baseline Median OS	Improvement Over Current OS That Would be Clinically Meaningful	½ SD (column a)	Effect Size (column b)
Pancreatic Cancer	FOLFRONOX Eligible Patients	10 – 11 months	4-5 months	7.21-7.93 months	0.35-0.25
Pancreatic Cancer	Gemcitabine Eligible Patients	6 - 8 months	3-4 months	4.33-5.77 months	0.46-0.26
Lung Cancer	Non-squamous cell carcinoma	13 months	3.25-4 months	9.38 months	0.17-0.21
Lung Cancer	Squamous cell carcinoma	10 months	2.5-3 months	7.21 months	0.17-0.21
Breast Cancer	Metastatic triple negative, previously untreated for metastatic disease	18 months	4.5-6 months	12.98 months	0.17-0.23
Colon Cancer	disease progression on all prior therapies (or not a candidate for standard 2nd or 3rd line options)	4-6 months	3-5 months	2.89-4.33 months	0.87-0.35

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



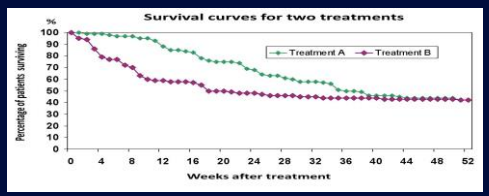
Censoring on survival curves

Survival analysis assumes censoring is random.

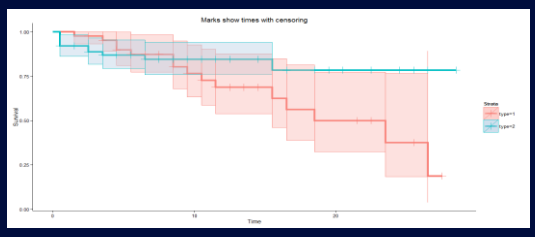
Censoring times vary across individuals and are not under the control of the investigator.

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 apects of the curves
- With n>100, all converge
- Take your pick

Hopefully these ideas will enable you to make advances in science