Statistics in Medicine: Not everything you read is true

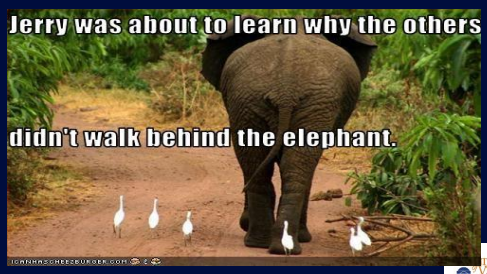
Jeff Sloan, Ph.D.
Mayo Clinic

AAPM, Denver, July 31, 2017

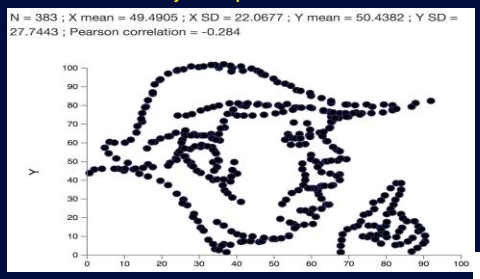
Flaws in statistical analysis

- 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

Some Practical Hard Lessons Learned



Statistical analysis can produce some odd results



Department of Radiation Oncology

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?

Department of Radiation Oncology

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 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

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?

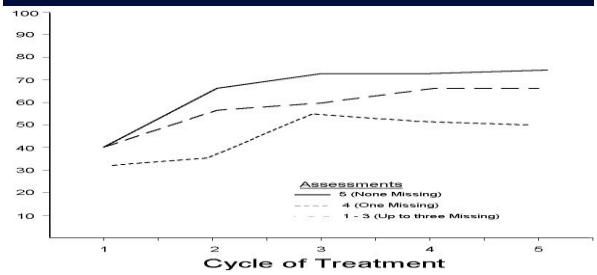
How do you handle the problem of missing data?

We all hate it when things go missing

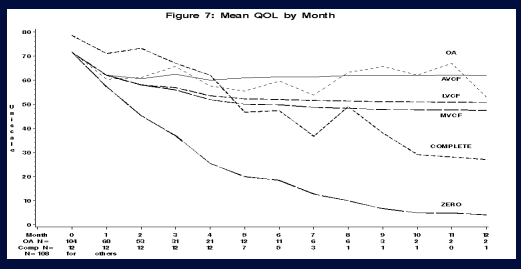
Recent Impetus for this work

- Reporting and dealing with missing quality of life data in RCTs: has the picture changed in the last decade? S. Fielding • A. Ogbuagu • S. Sivasubramaniam • G. MacLennan • C. R. Ramsay, QLR Dec. 2016.
- A random selection (50 %) of all RCTS published during 2013–2014 in BMJ, JAMA, Lancet and NEJM:
  - in 35% the amount of missing primary QoL data was unclear
  - 36% used imputation.
  - Only 23 % discussed the missing data mechanism.
  - Nearly half used complete case analysis.
- There is a large gap between statistical methods research relating to missing data and the use of the methods in applications.

Non-random Missing-ness: The worst performers leave

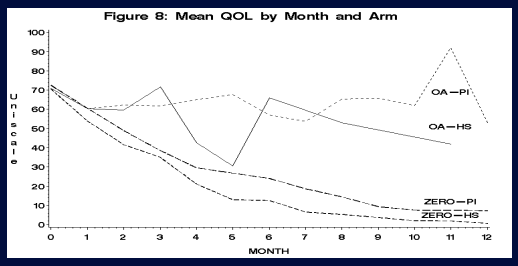


Impact of hydrazine sulfate on colorectal cancer patient QOL



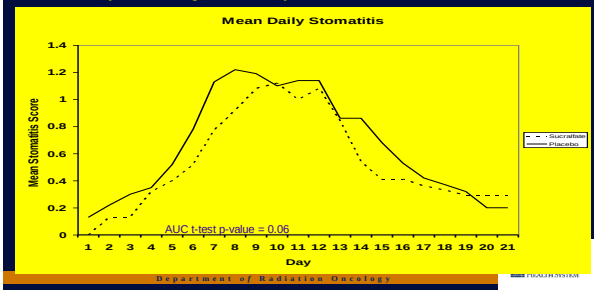
Impact of different imputation methods for missing data

Effect of imputation method on treatment comparison



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

A study examining the efficacy of sucralfate to alleviate stomatitis



Department of Radiation Oncology

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

Department of Radiation Oncology

Missing Data Macro Demonstration

Randomized study of Epoetin Alpha vs. Placebo for Anemia in Advanced Cancer Patients

Applied to the LASA Fatigue scale (higher scores are better)

Percent of Missing Values by Time

Percent Missing

Time	Epoetin Alfa	Placebo	Total Pct	p-value
Overall	24.7	24.7	24.7	0.9864
0	1.8	3.1	2.4	0.4581
1	15.7	16.0	15.8	0.9429
2	26.5	27.0	26.7	0.9204
3	34.3	36.8	35.6	0.6395
4	45.2	40.5	42.9	0.3901

Missing Data Patterns
X=Not Missing -=Missing

	Placebo (N=163)	Epoetin Alfa (N=166)	Total (N=329)	p value
Missing Data Pattern for Fatigue				0.8780
-----	1 (1%)	1 (1%)	2 (1%)	
-XX--	2 (1%)	1 (1%)	3 (1%)	
-XXXX	2 (1%)	1 (1%)	3 (1%)	
X----	12 (7%)	14 (8%)	26 (8%)	
X---X	1 (1%)	0 (0%)	1 (0%)	
X--X-	1 (1%)	1 (1%)	2 (1%)	
X--XX	0 (0%)	2 (1%)	2 (1%)	
X-X--	4 (2%)	3 (2%)	7 (2%)	
X-X-X	3 (2%)	0 (0%)	3 (1%)	
X-XX-	1 (1%)	3 (2%)	4 (1%)	
X-XXX	3 (2%)	2 (1%)	5 (2%)	
XX---	14 (9%)	17 (10%)	31 (9%)	
XX--X	4 (2%)	1 (1%)	5 (2%)	
XX-X-	5 (3%)	4 (2%)	9 (3%)	
XX-XX	6 (4%)	4 (2%)	10 (3%)	
XXX--	13 (8%)	14 (8%)	27 (8%)	
XXX-X	6 (4%)	6 (4%)	12 (4%)	
XXXX-	13 (8%)	17 (10%)	30 (9%)	
XXXXX	72 (44%)	75 (45%)	147 (45%)	

Monotone = missing only values at the end
Intermittent=Has last value but missing other values
Mixed=missing last value and other values

	Placebo (N=163)	Epoetin Alfa (N=166)	Total (N=329)	p value
Missing Data Type				0.5701
Complete Case	72 (44%)	75 (45%)	147 (45%)	
Intermittent	25 (15%)	16 (10%)	41 (12%)	
Mixed	13 (8%)	12 (7%)	25 (8%)	
Monotone Dropout	52 (32%)	62 (37%)	114 (35%)	
No Data	1 (1%)	1 (1%)	2 (1%)	

Missing data patterns were not significantly different between arms

Little's (JASA 1998) Test for MCAR

MCAR means the probability of an observation being missing does not depend on any observed or unobserved measurements

Chi-Square = 82.909
df = 49
p=0.002

If p<0.05 missing values are significantly different than MCAR
If p>=0.05 the hypothesis of MCAR can not be rejected

Department of Radiation Oncology

Odds Ratios From Logistic Models To Find Variables Associated With Missing Data

Variables Associated With Missing Values

Time	AGE
0	0.93 (0.86,1.00)
1	1.00 (0.98,1.03)
2	1.01 (0.99,1.03)
3	1.00 (0.98,1.02)
4	1.00 (0.98,1.02)

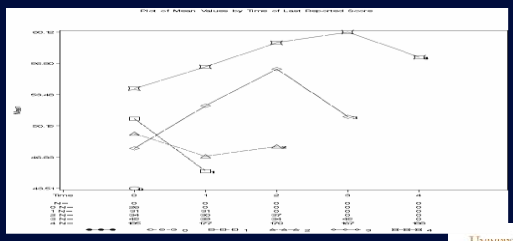
SEX	f vs m
	0.32 (0.06,1.64)
	0.64 (0.34,1.20)
	0.99 (0.60,1.63)
	1.10 (0.69,1.75)
	0.97 (0.62,1.53)

newarm	Placebo vs
	0.61 (0.14,2.62)
	0.99 (0.55,1.79)
	0.98 (0.60,1.59)
	0.90 (0.57,1.41)
	1.21 (0.78,1.88)

Age, sex, and arm were not associated with missing values

Department of Radiation Oncology

Mean Fatigue Scores
Grouped By When The Patient Last Reported Values



Subjects dropping out early have worse scores than subjects with complete data

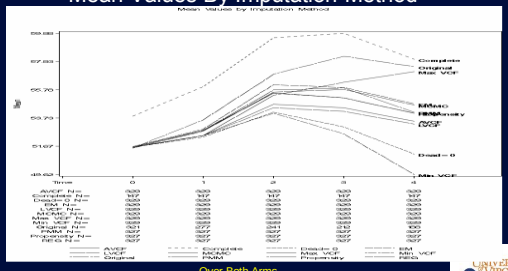
Department of Radiation Oncology

Imputation Methods

Original Data (No Imputation)	Uses All Available Data
Complete	Uses Only Subjects With No Missing Data
AVCF	Average Value Carried Forward
LVCF	Last Value Carried Forward
Max VCF	Maximum Value Carried Forward
Min VCF	Minimum Value Carried Forward
Dead=0	Imputes Zero After a Subject Dies
EM	EM Algorithm Estimates Based on Known Covariates
Regression	Regression Estimates Based on Known Covariates
MCMC	Bayesian Markov Chain Monte Carlo
PMM	Predictive Mean Matching
Propensity	Propensity Scores

Department of Radiation Oncology

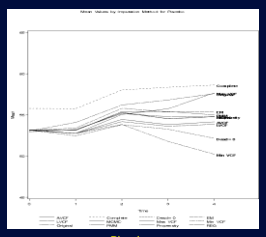
Mean Values By Imputation Method

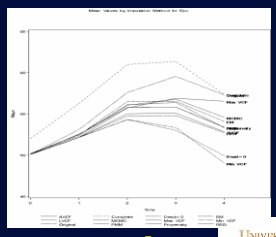


Over Both Arms

Department of Radiation Oncology

Mean Values By Imputation Method and Arm





Placebo Epo

Department of Radiation Oncology

Mean Changes From First to Last Values

	Placebo	Epoetin Alfa	Total	p value
Original	1.6	4.9	3.2	0.1782
LVCF	0.7	2.6	1.7	0.2068
AVCF	1.0	2.8	1.9	0.1503
Min VCF	-2.9	-1.0	-2.0	0.2712
Max VCF	4.5	6.4	5.5	0.1097
Complete	2.8	5.4	4.1	0.2222
Dead=0	-1.0	-0.2	-0.6	0.3668
EM	2.3	3.9	3.1	0.3961
MCMC	1.8	4.4	3.1	0.1312
Regression	1.6	3.3	2.5	0.3155
Propensity	1.6	3.3	2.5	0.3155
Predictive Mean Matching	1.8	3.3	2.5	0.3189

Changes from first to last cycle were not significantly different between arms for any of the imputation methods

Department of Radiation Oncology

Mean AUC Values by Imputation Method

	Placebo	Epoetin Alfa	Total	p value
Original	170.5	166.9	168.7	0.7466
LVCF	214.1	213.1	213.6	0.8766
AVCF	214.7	213.7	214.2	0.8766
Min VCF	210.1	209.3	209.7	0.8912
Max VCF	218.9	218.0	218.5	0.8415
Complete	229.3	233.3	231.3	0.8509
Dead=0	212.2	209.4	210.8	0.7193
EM	218.8	217.6	218.2	0.7267
MCMC	217.0	217.5	217.3	0.9496
Regression	216.9	215.9	216.4	0.8096
Propensity	216.9	215.9	216.4	0.8096
Predictive Mean Matching	218.0	216.7	217.4	0.7521

AUC values were not significantly different between arms for any of the imputation methods

Department of Radiation Oncology

What do you think?...

Missing Dta Macro Availability

– Need more datasets to test robustness

– Happy to collaborate (jsloan@mayo.edu)

UNIVERSITY OF VIRGINIA HEALTH SYSTEM

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

EORTC QLQ-LC13

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
  - 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. The real issue underlying the controversy over QOL is the relative novelty and lack of experience that presently exists with QOL. With time and familiarity this too shall pass.

(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

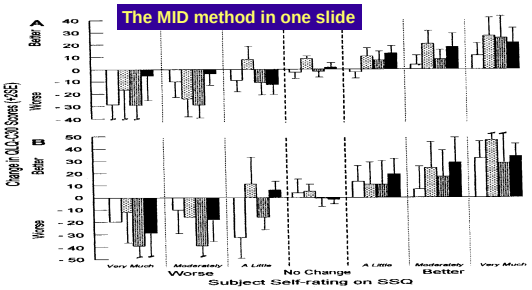


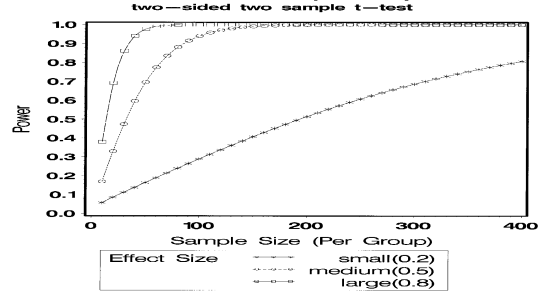
Fig 1. Relationship between SSQ ratings of change and QLQ-C30 scores from T1 to T2 for patients receiving chemotherapy for either breast cancer (A) or SCLC (B). Columns represent mean scores + 2 SE. □, physical functioning; ▨, emotional functioning; ▤, social functioning; ■, global QL.

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*

Department of Radiation Oncology

Figure 1 Power Example for Varying n



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?

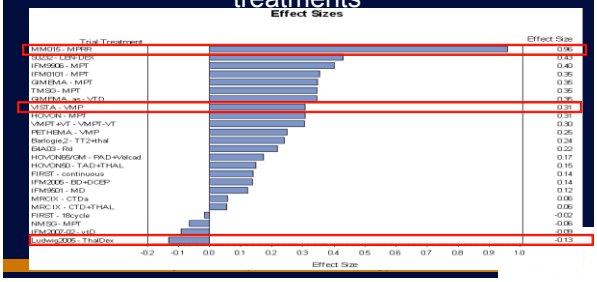
- The question arises as to whether this sort of calibration can be made for non-QOL endpoints such as survival and tumor response using the same ½ standard deviation approach.
- Major et al, 2014, ASCO, "Effect sizes for phase II and Phase III clinical trials using the ½ SD rule.

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	FOLFIRONOX 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?

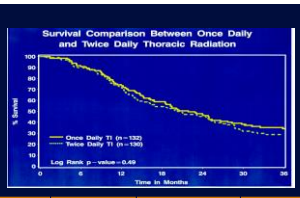
- This method makes for ready comparison across different oncology trials
- Clinicians can use calibrated effect size in the design of future clinical trials
- Provides a mathematically based effect-size that can be gauged by clinical opinion
- It provides a mechanism for comparing the effect sizes of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

Combining PRO/QOL data with survival: a peak into the near future

NCCTG Trial TDTRT vs ODTRT

Twice-Daily Thoracic Radiotherapy (TDTRT) for the treatment of lung cancer versus Once-Daily Thoracic Radiotherapy (ODTRT)

Median Overall Survival:
TDTRT 20 months
ODTRT 22 months
log-rank p=0.49



Toxicities	TDTRT	ODTRT(control)	Difference
Esophagitis	12.3%	5.3%	7%
Lethargy	7.7%	3%	4.7%
Overall Toxicity (nonhematologic)	53.9%	39.4%	14.5%

Bonner JA et al. JCO 1999;17:2681-2691

Non-significant survival comparison and significant unfavorable toxicity comparison
(89-20-52 Lung Cancer TDTRT vs ODTRT)
(equal weighting)

Endpoint	TDTRT	ODTRT	Difference	Effect Size	Quality Adjusted Effect Size
Median Overall Survival	20	22	-2	-0.06	
Toxicity					
Overall Toxicity (nonhematologic)	0.54	0.39	0.15	0.30	-0.18

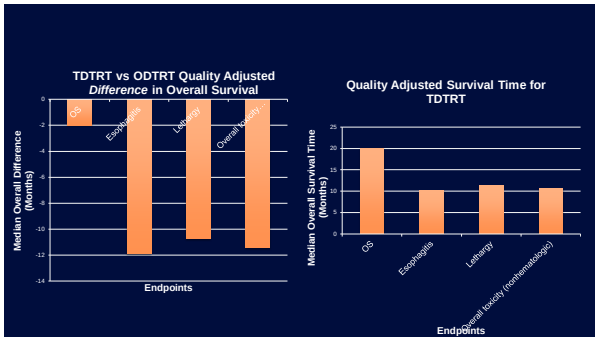
$$Quality-adjusted\ Effect\ Size = \frac{w_1 ES_A - w_2 ES_B}{w_1 + w_2}$$

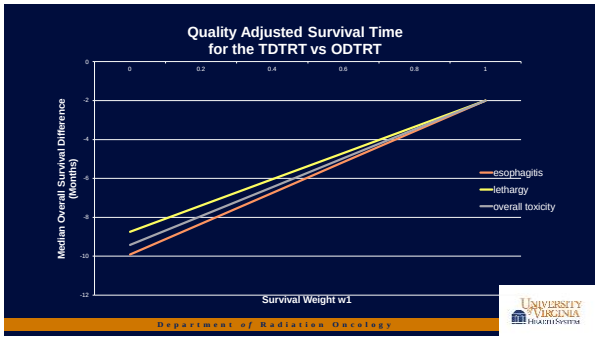
$$Quality-adjusted\ survival\ difference = Effect\ Size\ x\ OS\ Standard\ Deviation$$

$$\Delta\ Median\ OS = 0.18 * \left(\frac{22}{\ln 2}\right) = -5.7\ months.$$

Quality-adjusted survival estimate considering overall toxicity

- The quality-adjusted effect size is 0.18 in favor of the control (ODTRT) arm.
- The quality adjusted survival difference is –5.7 months.
- The median quality-adjusted OS for the TDRT arm is 16.3 months compared to 22 months in the ODTRT arm





…because we are not all the same

Summary

- Simple, sound, statistical practices produce credible results