Statistical significance, statistical power, and clinical significance: How to understand your results in context

Michael B. Altman, Ph.D.

Department of Radiation Oncology
Physics Division

Washington University in St. Louis
School of Medicine

Hypothesis Testing

- H_0 = Null Hypothesis
- H_1 = Alternative Hypothesis

Possible Hypothesis Test Outcomes		
Decision	Accept H_o	Reject H_o (i.e. Accept H_1)
H_o *is true*	Correct Decision (No error)	Type I Error
	Probability = $1 - \alpha$	Probability = α
H_o *is false* (i.e. H_1 is true)	Type II Error	Correct Decision (No error)
	Probability = β	Probability = $1 - \beta$

https://keydifferences.com/difference-between-type-i-and-type-ii-errors.html

Hypothesis Testing

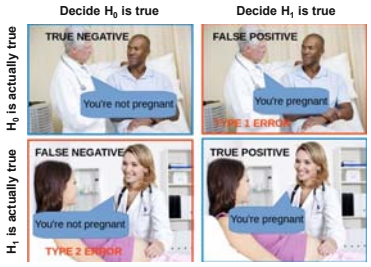
- H_0 = Null Hypothesis
- H_1 = Alternative Hypothesis

Possible Hypothesis Test Outcomes		
Decision	Accept H_o	Reject H_o (i.e. Accept H_1)
H_o *is true*	Correct Decision (No error)	Type I Error **FALSE POSITIVE**
	Probability = $1 - \alpha$	Probability = α
H_o *is false* (i.e. H_1 is true)	Type II Error **FALSE NEGATIVE**	Correct Decision (No error)
	Probability = β	Probability = $1 - \beta$ **POWER**

https://keydifferences.com/difference-between-type-i-and-type-ii-errors.html

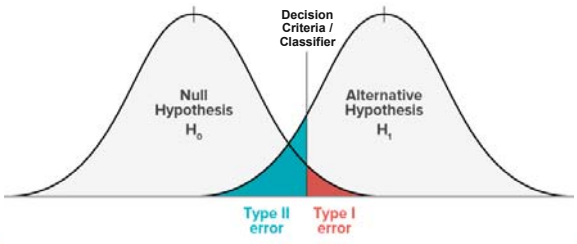
Type I v Type II Errors

- H_0 = You are not pregnant
- H_1 = You are pregnant



https://keydifferences.com/difference-between-type-i-and-type-ii-errors.html

Type I v Type II Errors



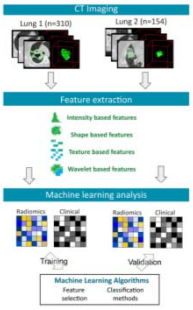
https://grasshopper.com/blog/the-errors-of-ab-testing-your-conclusions-can-make-things-worse/

Type I Errors and Statistical Significance

- α = Type I error probability = "false positive rate"
- Probability of getting one or more false positives
- Typically, we determine statistical significance by setting a certain threshold of acceptable α, the p-value (i.e. α = p)
- Common p (hence α) values: 0.05, 0.01, etc.
- Multiple simultaneous statistical comparisons: A given assigned p-value only applies to each individual test → does not apply to all tests together even if the p-values are the same for every test

Radiomics

- Multiple comparisons is an integral part of radiomics studies.
- Radiomics: "High throughput extraction of quantitative imaging features from medical images with the intent of mining predictive or prognostic associations between outcomes and medical images" (Paraphrased from Kumar et al., MRI, 2012).
- Statistical comparisons of multiple features (large p), using univariate statistical tests (ex: t-tests) from some number of images, n (where n<<p in some cases).



Parmar *et al., Sci. Rep.*, 2015

Family Wise Error Rate (FWER)

- For m independent comparisons, the multiple comparison Type-I error (false positive) rate is given by the Family Wise Error-Rate:

$$\bar{\alpha} = 1 - (1 - \alpha_{per\ comparison})^{m}$$

- So as m increases, the FWER (total false positive rate) increases.
- Example: If you do a study with 15 independent features and evaluate each one at p = 0.05:

$$\bar{\alpha} = 1 - (1 - 0.05)^{15} = 0.537$$

- On average, 53.7% of null (truly non-significant) features will be called significant instead of 5%!

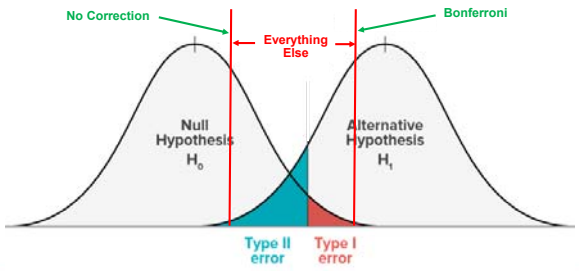
Bonferroni Correction

- Bonferroni correction: The most common correction to the FWER

$$\alpha_{per\ comparison} = \alpha / m$$

- Earlier example (m=15): Each α = 0.05/15 = 0.0033
- Very conservative approach
- Best when you really only expect very few true positive (significant) features
- Greatly increases the probability of Type II (false negative) errors → can diminish power of study

Type I / Type II Error Tradeoff



https://grasshopper.com/blog/the-errors-of-ab-testing-your-conclusions-can-make-things-worse/

Alternative Methods

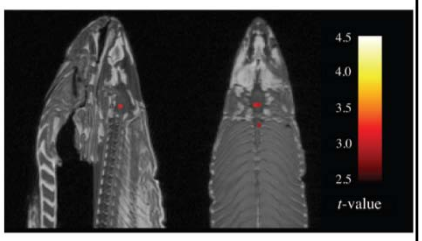
- Pick a primary comparison and make all others secondary
  - Likely impractical, inapplicable, or both for radiomics studies
- Alternative FWER corrections
  - Sidak Correction
  - Hochberg's Step-Up Procedure
  - O'Brien's Global Test
  - Many others
- False discovery rate (FDR)-based methods
  - FDR = # False Positives / # Total Significant Features
  - Ex: q – value (Storey and Tibshirani, PNAS, 2003)
  - Controlling FDR controls the percentage of features called significant that truly are not
  - 5% FDR rage means that among all features called significant, 5% of these are actually not.

Are Multiple Comparison Corrections Necessary?

- Example study: Bennett *et al.* Neuroimage, 2009.
- EPI images of single subjects brain while being subjected to visual stimuli.
- Images analyzed for regions of brain with significant BOLD signal changes during stimuli photos relative to rest.
- Several active voxels observed in a cluster located within the brain.
  - Statistics uncorrected for multiple comparisons showed activated cluster size was significant at p = 0.0001 level.
  - Both Hochberg correction and setting FDR at 0.05 showed no significant voxels in dataset.

Multiple Comparison Corrections Are Necessary

Subject Data	
Height	45.7 cm
Weight	1.73 kg
Sex	????
Species	Salmo salar
Common Name	Atlantic Salmon
Status	DEAD
Suggested Wine Pairing	Depends on the preparation, but a full bodied white like a Chardonnay or a rosé or light-tannin red

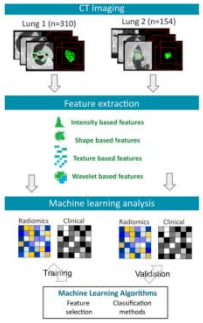


Power

- Power = 1-β → As power ↑, probability of making at Type II error (i.e. b) ↓
- A typical desired power is 0.8
- This means 80% chance that a feature called significant is actually significant (true positive rate)
- Increasing power typically comes at the expense of increasing Type I error rate.
- Power decreasing to 0.5 (or worse) means a test's ability to discriminate between false and true values is at or worse than random chance.

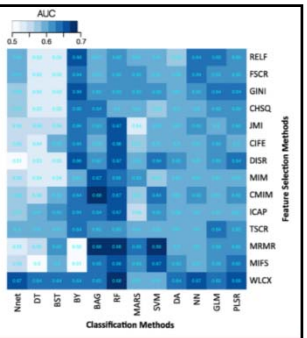
Factors Affecting Power

- In general, power impacted by the p-value used (decision threshold), sample size, magnitude of the effect
- For radiomics studies, additionally: feature selection method, classification method, # of selected features (Parmar *et al. Sci. Rep.* 2015).
- Parmar *et al.* study: compared 14 feature selection and 12 classification methods.
  - Five different feature numbers used per combination from 5-50 features (at 5 feature intervals)



Factors Affecting Power

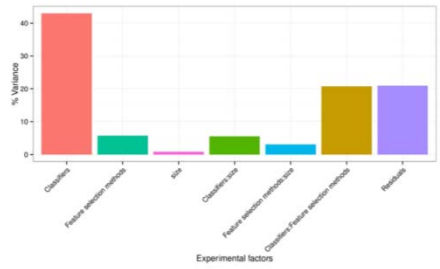
- 440 radiomics features extracted from CT images of 464 NSCLC patients.
- Attempted to find features that could predict overall survival
- Average ROC AUC used as metric for predictive performance (i.e. power)
- Large variance in predictive power across combinations of classification methods and feature selection methods.



Variance

- Variability in the outcome of the study – i.e. "stability" of the results across multiple samples
- Relates to standard deviation for continuous variables
- Parmar et al:
  - N/2 subsampling of training cases repeated 100 times using a boostrap approach
  - Repeated approach for each feature selection method; for classifiers used Wilcoxon feature selection method will all classifiers.
  - Relative standard deviation (RSD) of AUC used as variance benchmark.
  - Multifactor ANOVA used to quantify variation due to each factor / interaction

Impact on Variance by Experimental Factor



Trends Toward Significance

- Verbiage in the literature: (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 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) … among many others.
- There is no 'trend to significance' in any direction, and nowhere for the trend to be 'towards'.
- Think of it as pregnancy, you either are or your are not.

Clinical Significance vs. Statistical Significance

- Statistical significance is not clinical significance
- Clinical significance: Differences representing something that is relevant to a level of significantly impacting patient care.
- Example drawn from JCO 2001 (anonymous)
  - Health status questionnaire (HSQ) before / after scores on 1300 patients
  - All p-values <0.0001
  - Conclusion: all domains of QOL were statistically significantly different across treatment groups
  - Problem: 1300 patients provides 80% power to detect a change of 1 unit on 0-100 point scale

Clinical Significance vs. Statistical Significance

- EORT QLA-LC13: Clinical trial assessing change in QOL for lung cancer patients before and during TX using QOL questionaires and targeted modules

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 Approaches

- Assessing clinical significance can be difficult and subjective
- Two Primary Method Classes:
  1. Anchor-based methods requirements
     - Independent interpretable measure (the anchor) which has appreciable correlation between anchor and target
  2. Distribution-based methods
     - Rely on expression of magnitude of effect in terms of measure of variability of results (effect size or ES)
     - Many version attempt to divide ES into different levels of magnitude: small, moderate, large

ES Level Calculations

Table 1
ES Indexes and Their Values for Small, Medium, and Large Effects

Test	ES index	Effect size		
		Small	Medium	Large
1. m_A vs. m_B for independent means	$d = \frac{m_A - m_B}{\sigma}$	.20	.50	.80
2. Significance of product–moment r	r	.10	.30	.50
3. r_A vs. r_B for independent rs	$q = z_A - z_B$ where z = Fisher's z	.10	.30	.50
4. $P = .5$ and the sign test	$g = P - .50$	.05	.15	.25
5. P_A vs. P_B for independent proportions	$h = \phi_A - \phi_B$ where ϕ = arcsine transformation	.20	.50	.80
6. Chi-square for goodness of fit and contingency	$w = \sqrt{\sum_{i=1}^{k} \frac{(P_{1i} - P_{0i})^2}{P_{0i}}}$	.10	.30	.50
7. One-way analysis of variance	$f = \frac{\sigma_m}{\sigma}$	.10	.25	.40
8. Multiple and multiple partial correlation	$f^2 = \frac{R^2}{1 - R^2}$	.02	.15	.35

Note. ES = population effect size.

- J Cohen, 1992

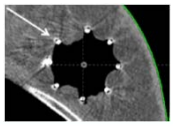
Four Guidelines from QOL studies

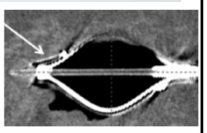
- Sloan, Cella, Hays, JCE 2005:

1. The method used to obtain an estimate of clinical significance should be scientifically supportable.
2. ½ SD is a reasonable and scientifically supportable conservative estimate of ES that is likely to be clinically meaningful and can be used in the absence of other information.
3. ESs < ½ SD may be meaningful if they are supported by data regarding the specific characteristics of a particular assessment or application, may also be meaningful.
4. 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 these should not preemptively restrict the application of information gained.

Radiomics-(ish) Example Study

- Altman et al. *Brachytherapy*, 2018
- Strut adjusted volume implant (SAVI) used for accelerated partial breast irradiation (APBI)
- Implant device → "Day 0" CT → Create TX plan
- CT 48-72 hours later → Replan if necessary → Treat
- Question: Can we use imaging-based features to predict which patients will need replanning from the Day 0 CT?





Methods and Metrics

- 62 consecutive patients, 36% requiring replans

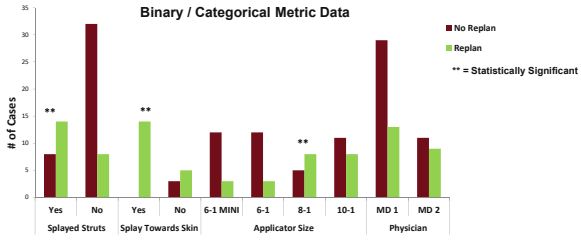
Table 1
Summary of metrics used in the retrospective analysis

Metric	Definition	Variable type
Plan adapted	Yes or no: was treatment plan replanned using Verify CT?	Binary
Applicator size	Use of which size SAVI applicator: 6-1 mini, 6-1, 8-1, or 10-1	Categorical
Attending physician	Which one of two attending physicians implanted the device and evaluated the need to replan the DOI CT plan	Binary
Splayed struts	Yes or no: were any of the struts in the applicator splayed?	Binary
Splay toward skin	Yes or no: was the direction of the splay in the applicator directed toward the skin	Binary
Absolute volume of air	Absolute volume of air in the PTV_EVAL	Continuous
Percent volume of air	Percent volume of air in the PTV_EVAL	Continuous
Minimum distance to skin or ribs	Minimum distance from any part of the applicator volume to the skin or ribs	Continuous
(SAVI + 1)/(SAVI + 1 − SKIN)	Ratio of volume of 1-cm thick shell around the applicator volume to that same shell cropped from a margin of 0.5 cm from the outer surface of the skin	Continuous

DOI = day-of-implant; SAVI = strut-adjusted volume implant.
Definitions of Verify CT, PTV_EVAL, and splay discussed in the text.

- Evaluation: Fisher's exact test, Wilcoxon rank-sum, 2 tailed p-test for categorical, binary, and continuous variables, respectively.

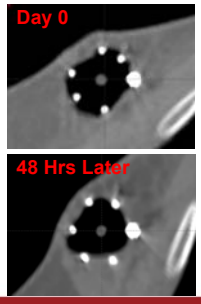
Results



- No continuous variables showed statistical significance

Statistical vs. Clinical Significance

- Splay (Towards Skin)
  - Flexible struts can be pushed around by patient's anatomy
  - Anatomy (and device) can shift/change over the 24 hours after implant → altered dose distribution vs. Day 0 plan
  - Statistical significance seems to correlate to a clinical result → clinically significant result → change in clinical workflow
- Use of 8-1 SAVI
  - Use of only 1 (out of 4) middle-sized device was statistically significant
  - Clinically intuitive reason why: ??? → Not deemed clinically significant



Summary

- For "big data" type studies as in radiomics with multiple hypotheses tested simultaneously, corrections must be made to statistical significance thresholds to limit Type I errors.
- Too conservative corrections can result in a rise in Type II errors (i.e. limited power).
- Other factors such as feature selection methods, classification methods, and number of features used can also impact power.
- Clinical significance statistical significance.
- Clincal significance should be assessed case-wise using accepted methods and/or guidelines in concert with input from clinicians
- Don't go it alone – statisticians are your friends!

Bibliography

1. Altman *et al. Brachytherapy*. 17, 2008; 40-49
2. Bennett CM *et al. Neuroimage,* 47, 2009; S125
3. Cohen J. *Psycholog Bulletin,* 112(1), 1992; 155-159
4. Kumar *et al.* MRI 30, 2012; 1234-1248
5. Parmar *et al.* Sci Rep, 5(13087), 2015
6. Sloan JA, Cella D, Hays RD. J Clin Epid 58(12), 2005; 1217-1219.
7. Storey JD and Tibshirani R., *Proc Natl Acad Sci USA*, 100 (16), 2003; 9440-9445
8. Yip SSF and Aerts HJWL. Phys Med Bio. 61, 2016. R150-166

Michael B. Altman
Assistant Professor of Radiation Oncology
Campus Box XXXX
Street address
St. Louis, MO 63110
(314) 123-4567

m.altman@wustl.edu