Retrospective Translational Research Projects

Kathryn Winter, M.S.

Radiation Therapy Oncology Group (RTOG)

Director of Statistics

Outline for the talk

- Planning issues for a retrospective project
- Analyzing and interpreting results of retrospective analyses
- Determining cutpoints for continuous markers
- From retrospective to prospective

How should you plan a retrospective translational research (TR) project?

(a) Find out how many samples you can get and figure that'll work

(b) Randomly (*that's statistical, right?*) choose a sample size

(c) **Work with a statistician!!!!!!!**

How should you plan a retrospective TR project?

(c) **Work with a statistician!!!!!!!**

Planning a Retrospective TR Project

- Basic hypothesis
  - If your hypothesis is “Will I get an abstract accepted to a meeting being held in a fun spot?” – rethink your hypothesis………
  - High levels of marker x are associated with poorer overall survival
  - Marker x is associated with overall survival
- Need an estimate of effect size
  - Hazard Ratio (HR)

Hazard Rate for Survival

$$\text{Hazard Rate} \quad = \quad \text{death rate per time unit}$$

$$= \quad \frac{\text{\# deaths}}{\text{sum of follow-up times}}$$

Hazard Ratio (HR)

Hazard Ratio = 1 ⇒ no difference

= 2 ⇒ death rate twice as high for abnormal group

Planning a Retrospective TR Project

- **Basic hypothesis**
- **Estimate of effect size: Hazard Ratio (HR)**
- **Determine power to detect an <u>association</u> given data you have**
  - Number of events (death, local failure, etc) are fixed
  - Based on number of events, not sample size
  - 200 patients with 10 deaths vs. 200 patients with 150 deaths
    - Give different levels of power

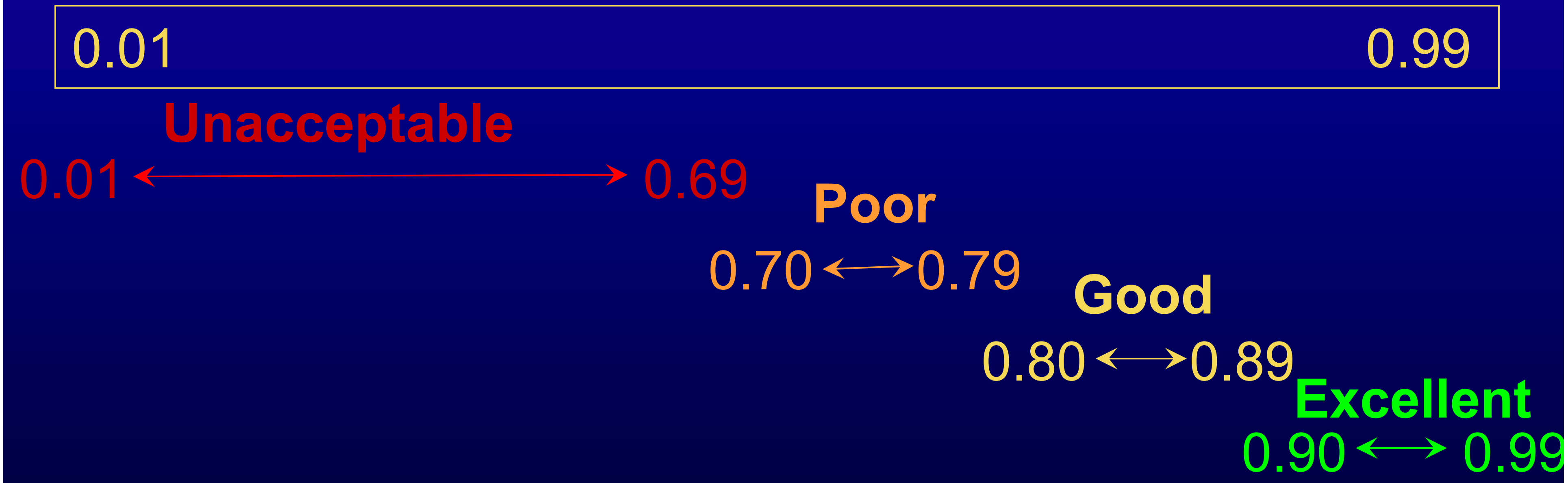
Statistical Power

Power = 1.0 – β (type II error)

Probability of detecting the hypothesized difference Δ or greater, if it exists.

Statistical Power

Acceptability Scale



0.01 0.99

Unacceptable

0.01 ↔ 0.69

Poor

0.70 ↔ 0.79

Good

0.80 ↔ 0.89

Excellent

0.90 ↔ 0.99

Schoenfeld's Equation

$$\#\ \text{events} = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{(\ln HR)^2\, \omega\, (1-\omega)}$$

HR = hazard ratio (measure of difference)

ω = prevalence rate for patients with the abnormal tumor marker

$z_{1-\alpha/2}$ = the normal deviate for the significance level (α=0.05 / two-sided)

$z_{1-\beta}$ = the normal deviate for the statistical power

Statistical Power

	HR = 1.5			HR = 2.0			HR = 2.5			HR = 3.0		
	# Events			# Events			# Events			# Events		
	25	50	100	25	50	100	25	50	100	25	50	100
ω=	Statistical Power =											
0.1	0.08	0.13	0.22	0.17	0.31	0.54	0.27	0.49	0.78	0.37	0.64	0.90
0.2	0.12	0.20	0.36	0.28	0.50	0.79	0.44	0.73	0.95	0.59	0.87	0.99
0.3	0.15	0.25	0.45	0.35	0.61	0.88	0.55	0.84	0.98	0.71	0.94	0.99
0.4	0.16	0.28	0.51	0.39	0.67	0.92	0.61	0.88	0.99	0.76	0.96	0.99
0.5	0.17	0.29	0.52	0.41	0.68	0.93	0.62	0.89	0.99	0.78	0.97	0.99

Statistical Power

	HR = 1.5			HR = 2.0			HR = 2.5			HR = 3.0		
	# Events			# Events			# Events			# Events		
	25	50	100	25	50	100	25	50	100	25	50	100
ω=	Statistical Power =											
0.1	0.08	0.13	0.22	0.17	0.31	0.54	0.27	0.49	0.78	0.37	0.64	0.90
0.2	0.12	0.20	0.36	0.28	0.50	0.79	0.44	0.73	0.95	0.59	0.87	0.99
0.3	0.15	0.25	0.45	0.35	0.61	0.88	0.55	0.84	0.98	0.71	0.94	0.99
0.4	0.16	0.28	0.51	0.39	0.67	0.92	0.61	0.88	0.99	0.76	0.96	0.99
0.5	0.17	0.29	0.52	0.41	0.68	0.93	0.62	0.89	0.99	0.78	0.97	0.99

Statistical Power

	HR = 1.5			HR = 2.0			HR = 2.5			HR = 3.0		
	# Events			# Events			# Events			# Events		
	25	50	100	25	50	100	25	50	100	25	50	100
ω=	Statistical Power =											
0.1	0.08	0.13	0.22	0.17	0.31	0.54	0.27	0.49	0.78	0.37	0.64	0.90
0.2	0.12	0.20	0.36	0.28	0.50	0.79	0.44	0.73	0.95	0.59	0.87	0.99
0.3	0.15	0.25	0.45	0.35	0.61	0.88	0.55	0.84	0.98	0.71	0.94	0.99
0.4	0.16	0.28	0.51	0.39	0.67	0.92	0.61	0.88	0.99	0.76	0.96	0.99
0.5	0.17	0.29	0.52	0.41	0.68	0.93	0.62	0.89	0.99	0.78	0.97	0.99

Statistical Power

	HR = 1.5			HR = 2.0			HR = 2.5			HR = 3.0		
	# Events			# Events			# Events			# Events		
	25	50	100	25	50	100	25	50	100	25	50	100
ω=	Statistical Power =											
0.1	0.08	0.13	0.22	0.17	0.31	0.54	0.27	0.49	0.78	0.37	0.64	0.90
0.2	0.12	0.20	0.36	0.28	0.50	0.79	0.44	0.73	0.95	0.59	0.87	0.99
0.3	0.15	0.25	0.45	0.35	0.61	0.88	0.55	0.84	0.98	0.71	0.94	0.99
0.4	0.16	0.28	0.51	0.39	0.67	0.92	0.61	0.88	0.99	0.76	0.96	0.99
0.5	0.17	0.29	0.52	0.41	0.68	0.93	0.62	0.89	0.99	0.78	0.97	0.99

Statistical Power

	HR = 1.5			HR = 2.0			HR = 2.5			HR = 3.0		
	# Events			# Events			# Events			# Events		
	25	50	100	25	50	100	25	50	100	25	50	100
ω=	Statistical Power =											
0.1	0.08	0.13	0.22	0.17	0.31	0.54	0.27	0.49	0.78	0.37	0.64	0.90
0.2	0.12	0.20	0.36	0.28	0.50	0.79	0.44	0.73	0.95	0.59	0.87	0.99
0.3	0.15	0.25	0.45	0.35	0.61	0.88	0.55	0.84	0.98	0.71	0.94	0.99
0.4	0.16	0.28	0.51	0.39	0.67	0.92	0.61	0.88	0.99	0.76	0.96	0.99
0.5	0.17	0.29	0.52	0.41	0.68	0.93	0.62	0.89	0.99	0.78	0.97	0.99

# Events needed for HR=1.5 with at least 80% Power

ω	# events
0.10	531
0.20	299
0.30	228
0.40	199
0.50	191

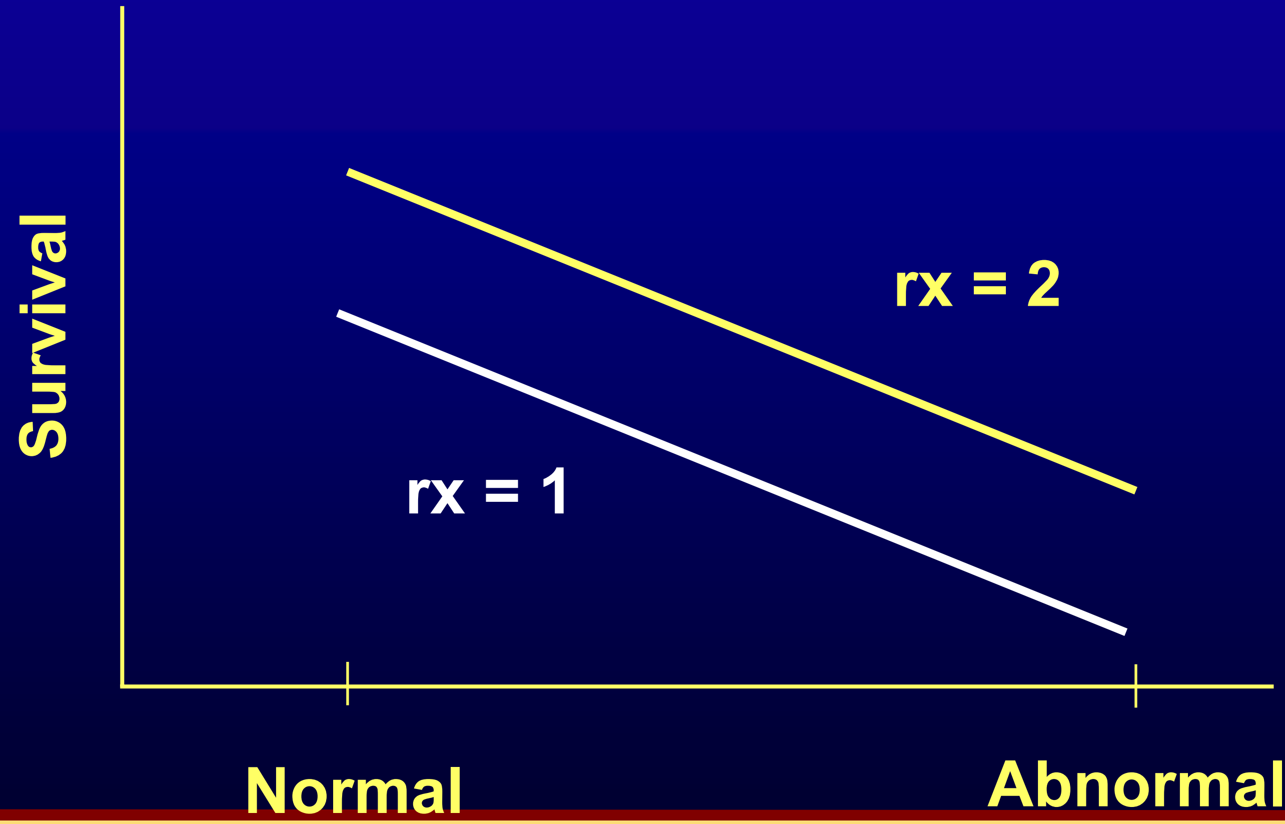
Statistical Power Considerations

- **If power is too low for realistic HR**
  - Don’t waste the specimens on an underpowered study
    - Specimens are a valuable, finite, resource
    - Need to make the best use of them
  - Consider other studies that would be applicable to combine

Prognostic vs. Predictive

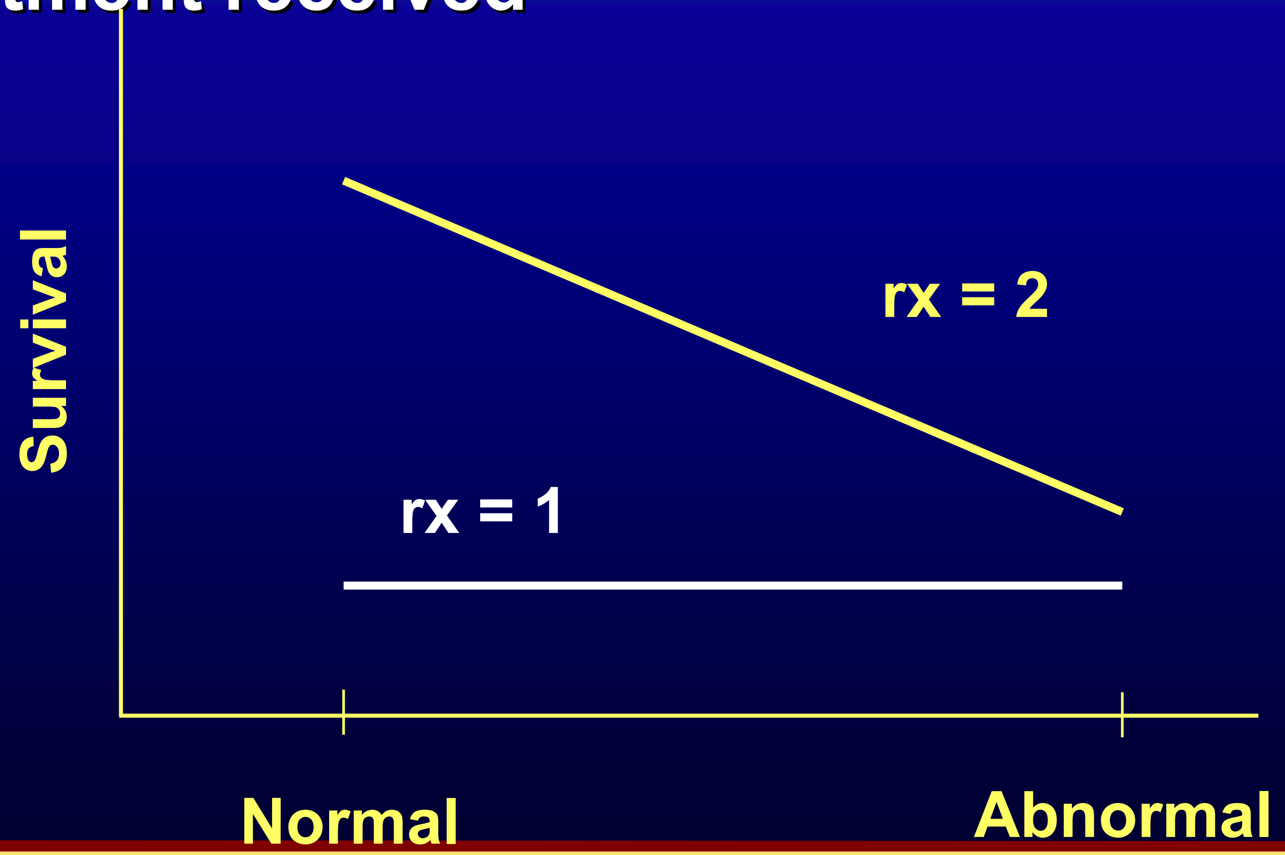
Prognostic vs. Predictive

- **Prognostic marker:** level of the marker is associated with different efficacy regardless of treatment received



Prognostic vs. Predictive

- **Predictive marker:** level of the marker is associated with different efficacy based on treatment received



Interactions

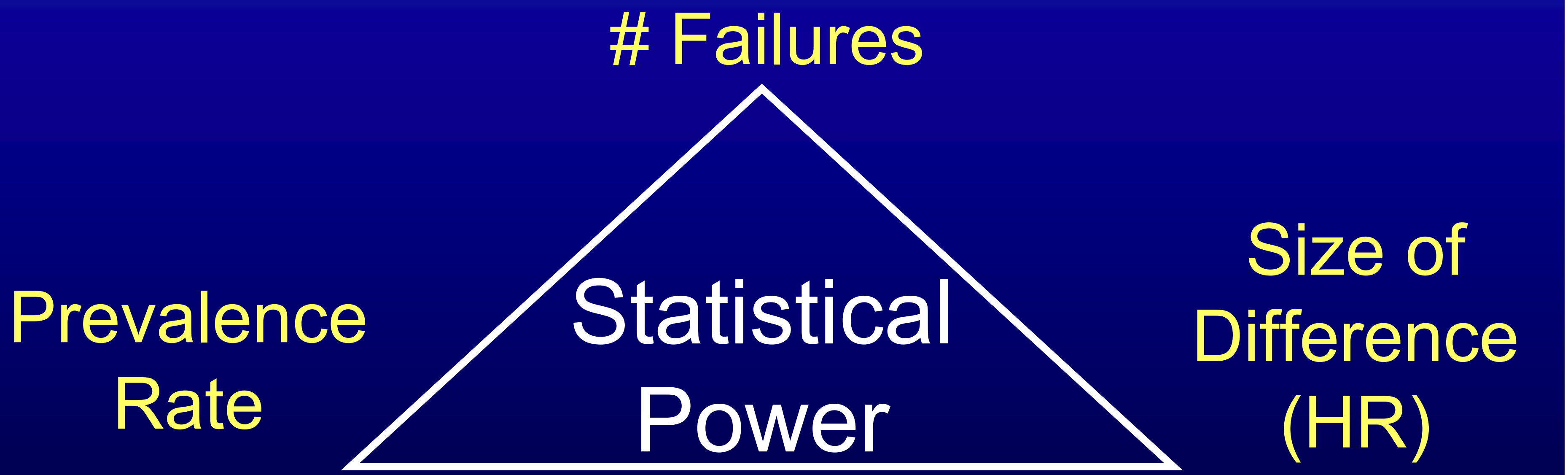
Is the tumor marker associated with response or lack of response to a particular therapy?

- Really testing for an *interaction* between marker status and treatment.

Sample Size Considerations

- Test of interaction can require 4 times more failures than test for treatment main effect. (Peterson and George)
- Marker status is not randomized and imbalance must be taken into account.

Summary



# Failures

Prevalence Rate

Statistical Power

Size of Difference (HR)

Analyzing and Interpreting Results

How should you analyze and interpret results of a retrospective TR project?

(a) Get a hold of any statistical computer package and do it yourself.

(b) Get your resident/fellow/grad student to do it.

(c) **Work with a statistician!!!!!!**

How should you analyze and interpret results of a retrospective TR project?

(c) Work with a statistician!!!!!!!

Interpreting Results

A p-value is a probability of obtaining
a result as extreme or more extreme
than the one observed, if due to chance alone.

Statistical Reality!

Any difference HOWEVER SMALL
can be shown to be statistically significant
with enough patients.

Statistical Significance

All a p-value tells is how likely chance alone can account for the observed result. It tells nothing about the magnitude of the observed difference or about the number of patients.

Interpreting Results

- Statistically Significant vs. Clinically Important
- Is a statistically non-significant result NOT clinically important?

Interpreting Results

- Possible reasons for a non-significant result
  - The difference really doesn’t exist
  - Study is underpowered for the difference of interest
  - Study is underpowered for a clinically meaningful difference

Interpreting Results

Noordzij et al reported a

non-significant

cause-specific survival result for expression of neuroendocrine cells in prostate cancer patients

To Calculate Statistical Power

# Observed Cancer Deaths (*<u>not total # of patients</u>*)	=	14
Prevalence rate of patients with neuroendocrine cells (observed)	=	0.47
Significance Level (α) (set by statistician)	=	0.05
Hazard Ratio - measure of difference (estimated by statistician)	=	2.0

What is the statistical power?

Hazard Ratio	Statistical Power
2.0	0.25

The probability of detecting that patients with neuroendocrine cells are dying from prostate cancer twice as fast as patients without them if the true hazard ratio is 2.0 is only 25/100.

Thus, 75 times out of 100, this difference would not be detected.

RTOG 8610
Prostate Cancer

S	**Clinical Stage**	R	
T	**B_2**	A	**1) Radiation Therapy**
R	**C**	N	**+**
A	**Differentiation**	D	**Zoladex and Flutamide**
T	**Well**	O	
I	**Moderate**	M	**2) Radiation Therapy Alone**
F	**Poor**	I	
Y		Z	
		E	

RTOG 8610

Eligibility:

- bulky, locally advanced adenocarcinoma of the prostate
- stage T2 and T3
- no prior hormonal therapy
- no metastatsis

Hazard Ratio (HR)

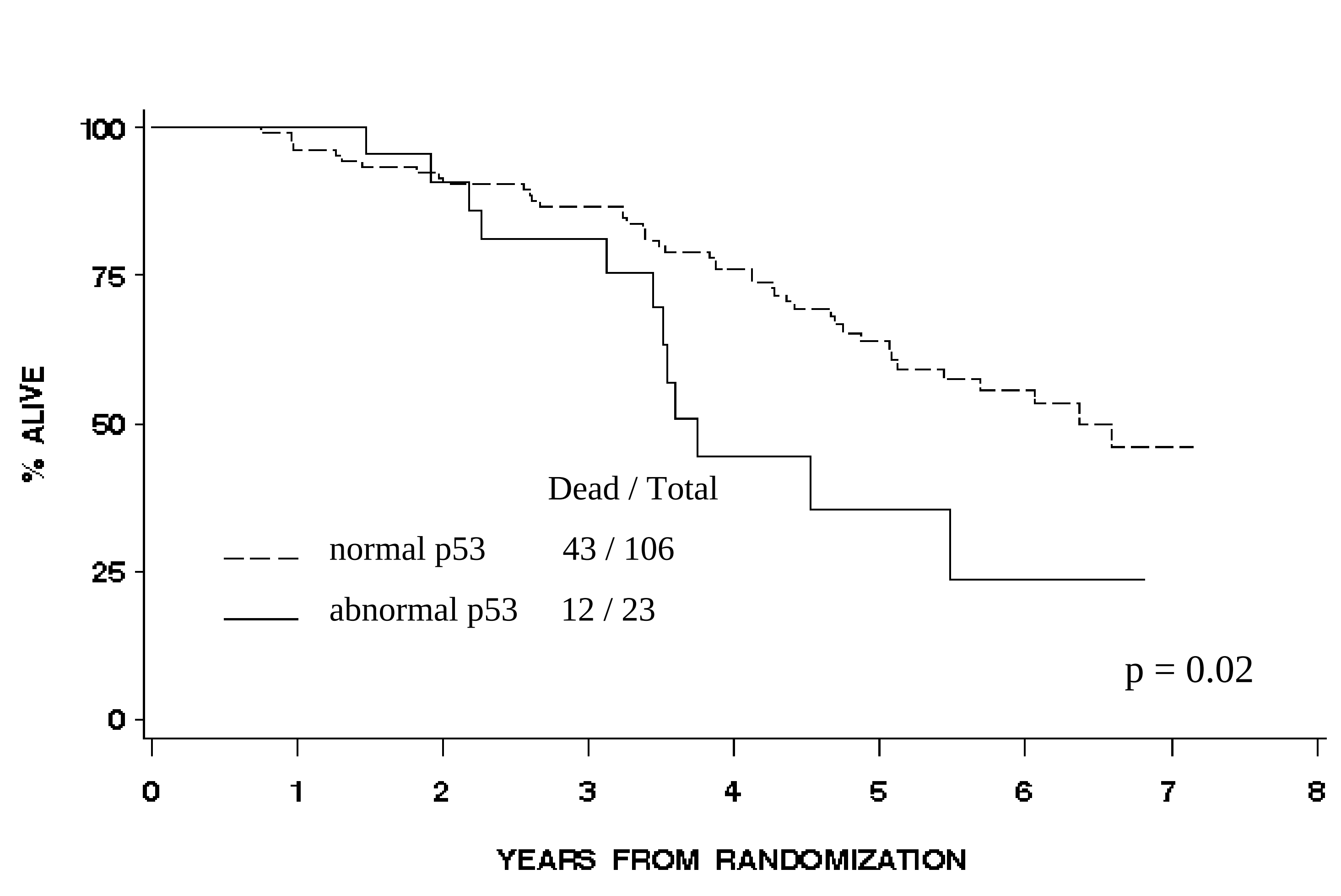
Grignon et al

Overall Survival

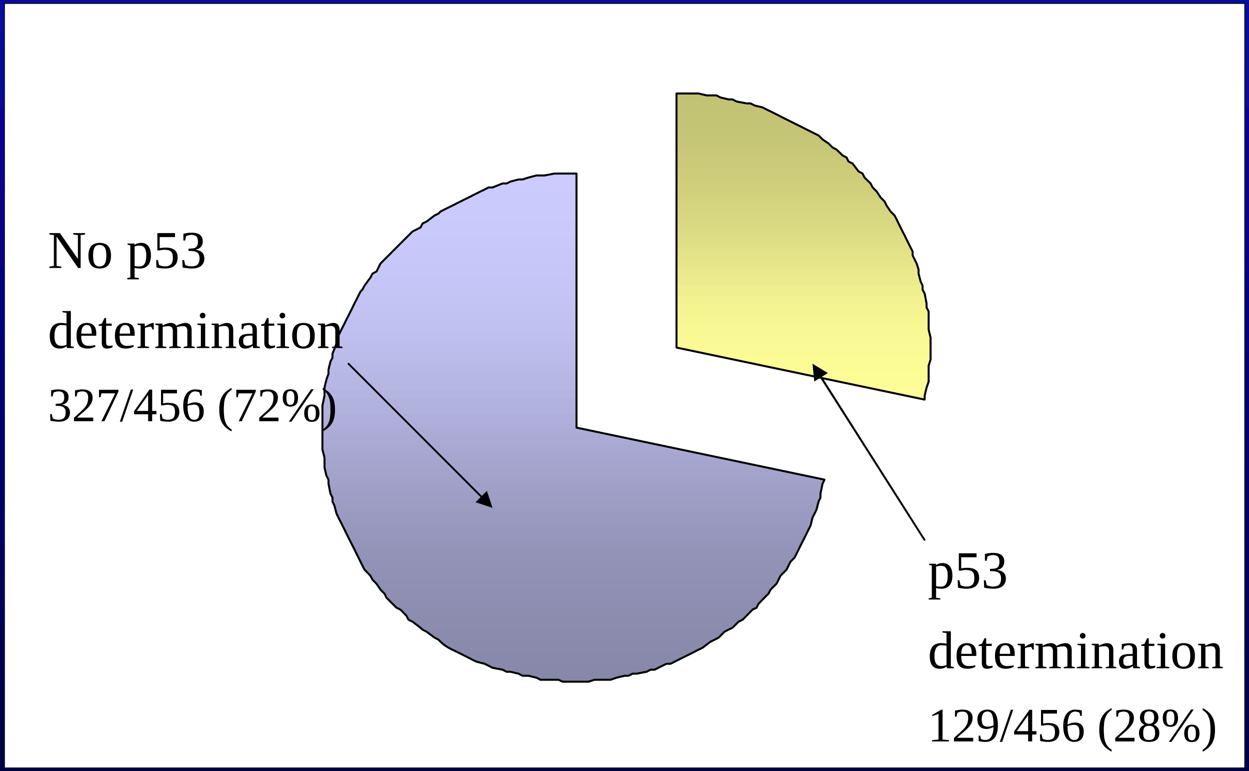
$$\frac{\text{hazard rate with abnormal p53 expression}}{\text{hazard rate with normal p53 expression}} = 2.3$$

RTOG 8610 – Overall Survival

Normal p53 vs. Abnormal p53 (Grignon et al)

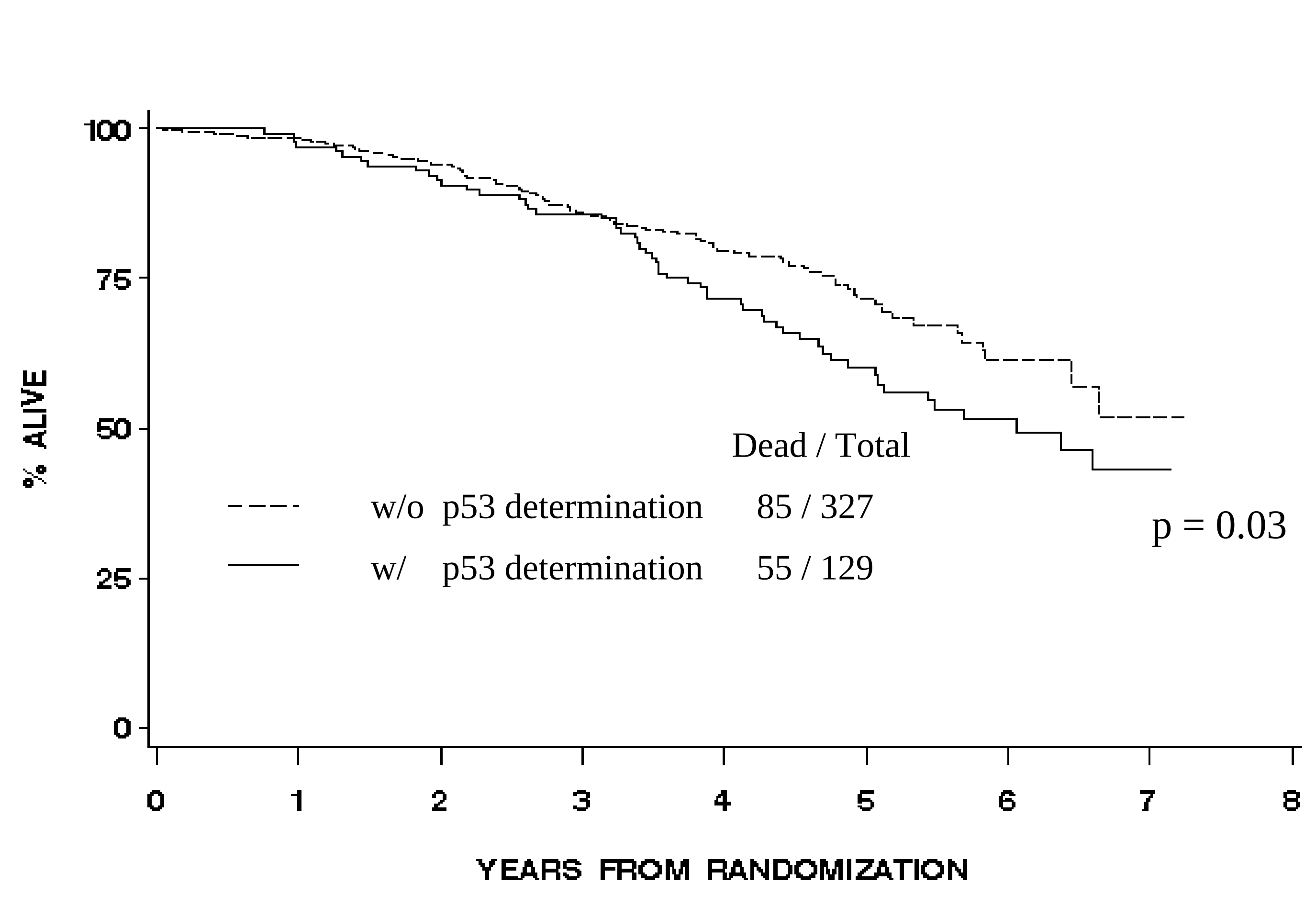


RTOG 8610
p53 Expression



RTOG 8610 – Overall Survival

Patients w/ & w/out p53 (Grignon et al)



RTOG 8610

Pretreatment Characteristics

Combined Gleason	With p53 Value	Without p53 Value
2-5	17 (13%)	51 (16%)
6-7	69 (53%)	184 (58%)
8-10	43 (35%)	85 (26%)
T-Stage		
T2	34 (26%)	103 (32%)
T3	95 (74%)	224 (68%)

RTOG 8610

Randomized Treatment

Randomized Treatment	With p53 Value	Without p53 Value
RT	72 (56%)	158 (48%)
RT+Hormones	57 (44%)	169 (52%)

Missing Data

- Common practice: to delete cases with missing data

  - loss of statistical power at best

  - severe bias at worse

Missing Data
Conflicting Results

RTOG 8610 Survival

Marker	Patient Population	#	p-value
Ploidy	With ploidy data	149	p = .03
(diploid vs. non-diploid)			
p53	With p53 data	129	p =.02
(normal vs. abnormal)			
Ploidy	With both ploidy and p53 data	113	p = .22
(diploid vs. non-diploid)			

Explanation

Patient Group	# Pts	# Deaths	p-value	Hazard Ratio
Ploidy (diploid vs. non-diploid)	149	102	0.03	1.54
Ploidy and p53 (diploid vs. non-diploid)	113	78	0.22	1.32

RTOG 8610

Pre-treatment Tumor Markers

- p53
- DNA contents (ploidy)
- Microvessel density (MVD)
- Neuroendocrine
- PSA density/extent
- PAP density/extent

Statistician's Nightmare:
Missing Data!!!

Tumor Marker	# Patients w/ Marker
A	129
B	147
C	149
D	155
E	139
F	153
Total # Patients on RTOG 8610	456

Number of patients with all 6 markers: 70 (15%)

Missing Data

- One solution: Imputation

  - Statistical Method "Multiple Imputation"

Assessing Possible Biases

- **Difference between patients with normal and abnormal levels of tumor marker respect to:**
  - Baseline demographics and tumor characteristics
  - Treatment received

Cox Proportional Hazards Model

$$\ln(\mathrm{HR}) = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

X = patient value, e.g.

$0 = T_2$

$1 = T_3$

β_i = parameter for “risk ratio” to be estimated

Cox Proportional Hazards Model

Cox Model 1 = known prognostic factors

Cox Model 2 = known prognostic factors
+ tumor marker under test

Cox Proportional Hazards Model

Model 1 = 0.59(Gleason) + 0.40(T-stage) + 0.22(RX)

Model 2 = 0.58(Gleason) + 0.49(T-stage) + 0.26(RX) + 0.85(p53)

$p = 0.025$

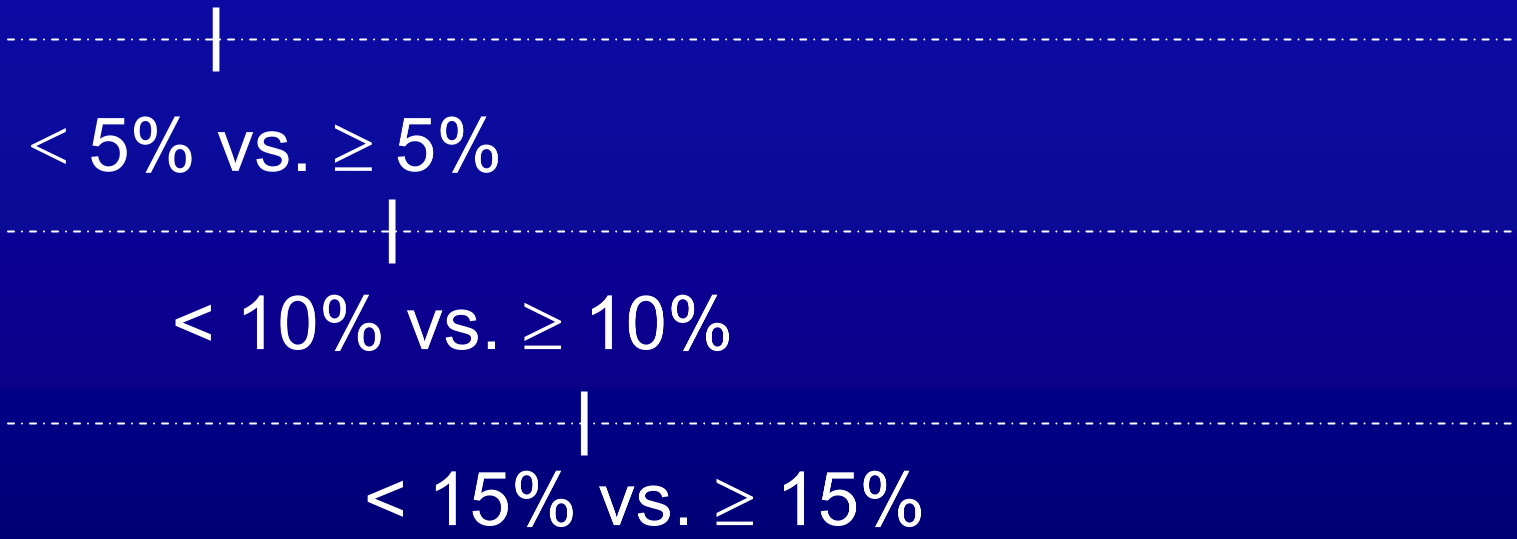
Considerations of the Cox Model

- Estimates of the hazard ratio
- Statistically more powerful than multiple subset analyses
- However, *for every factor in the model, there should be ~ 10 failures (death, local failure etc.)*

Determining Cutpoints for Continuous Markers

Fishing: Keep this in the water?

Evaluating Cutpoints



< 5% vs. ≥ 5%

< 10% vs. ≥ 10%

< 15% vs. ≥ 15%

1. 19 different thresholds

2. Report lowest p-value with log rank test

3. Probability of finding one p-value < 0.05 = 0.53 (multiple testing)

Approaches to the Cutpoint Problem

- p-value adjustment
- Literature based cutpoint
- Separate validation sets of data

Multiple Testing

Bonferroni Method

- To preserve an overall significance level of 0.05 with 19 tests
- p-value ≤ 0.0026 (=0.05/19)

PICKING CUTPOINT(S)

Literature Based

e.g. Grignon et al, p53 cutpoint

- Positive survival study in prostate cancer
- Same cutoff point used in other organ systems
- High degree of correlation with presence of a mutation

Separate Validation

- **Confirm the observation with another dataset**
- **Randomly split dataset in half**
  - Training dataset
  - Validation dataset

From Retrospective to Prospective

- Phase III trial w/ 4 years of accrual and 3 years follow-up and projected 280 deaths
- Design/activate in 2009, efficacy results available 2016
- What markers do you prospectively project in 2009 to evaluate in 2016?
- Will these markers still be relevant in 2016?
- Translational research landscape changes quickly

Possible Solution

- Include a table in the protocol showing statistical power for various HRs and prevalence rates based on the number of events in the trial.

Statistical Power

	HR = 1.5			HR = 2.0			HR = 2.5			HR = 3.0		
	# Events			# Events			# Events			# Events		
	280	210	140	280	210	140	280	210	140	280	210	140
ω=	Statistical Power =											
0.1	0.53	0.42	0.30	0.93	0.85	0.69	0.99	0.97	0.90	0.99	0.99	0.97
0.2	0.77	0.65	0.48	0.99	0.98	0.90	0.99	0.99	0.99	0.99	0.99	0.99
0.3	0.87	0.76	0.59	0.99	0.99	0.96	0.99	0.99	0.99	0.99	0.99	0.99
0.4	0.91	0.82	0.65	0.99	0.99	0.98	0.99	0.99	0.99	0.99	0.99	0.99
0.5	0.92	0.83	0.66	0.99	0.99	0.98	0.99	0.99	0.99	0.99	0.99	0.99

Possible Solution

- **Include text such as:**

  > **“As the trial gets closer to the time of efficacy analysis, relevant markers based on the current state of the science for x cancer will be chosen to be evaluated prospectively in this trial.”**

- **When those markers are chosen, officially amend the protocol**
  - Define markers with scientific justification
  - Power info and analysis plan

Summary

- Sufficiently powered projects to make the best use of the valuable, finite specimen resources
- Power driven by the number of events (not the number of patients), and the effect size (HR), prevalence of marker
- *Not statistically significant* is not synonymous with *clinically meaningless*.
- **Work with a statistician!!!!!!!**

“Statistics are no substitute for judgment”

- Henry Clay

Acknowledgements

Patients that participate in RTOG and all clinical trials

Thomas F. Pajak, PhD
(RTOG H&N Senior Statistician)