Hypothesis Testing 1

National Cancer Center, Japan

Biostatistics Division

Aya Kuchiba

Hypothesis Testing

- Hypothesis Testing 1: Fundamentals
  - Theory, interpretation, and limitations
- Hypothesis Testing 2: Specific methods
  - Introduction of *t*-test, X^2 test, etc.

When Is A Test Used?

- Comparison of mean quality of life (QOL) scores

	Group A	Group B	Group C	*P* value
QOL Score	19.4	12.6	16.7	0.0015

Dabakuyo et al. (2009) 20, 1352-1361 Ann Oncol

- Family history

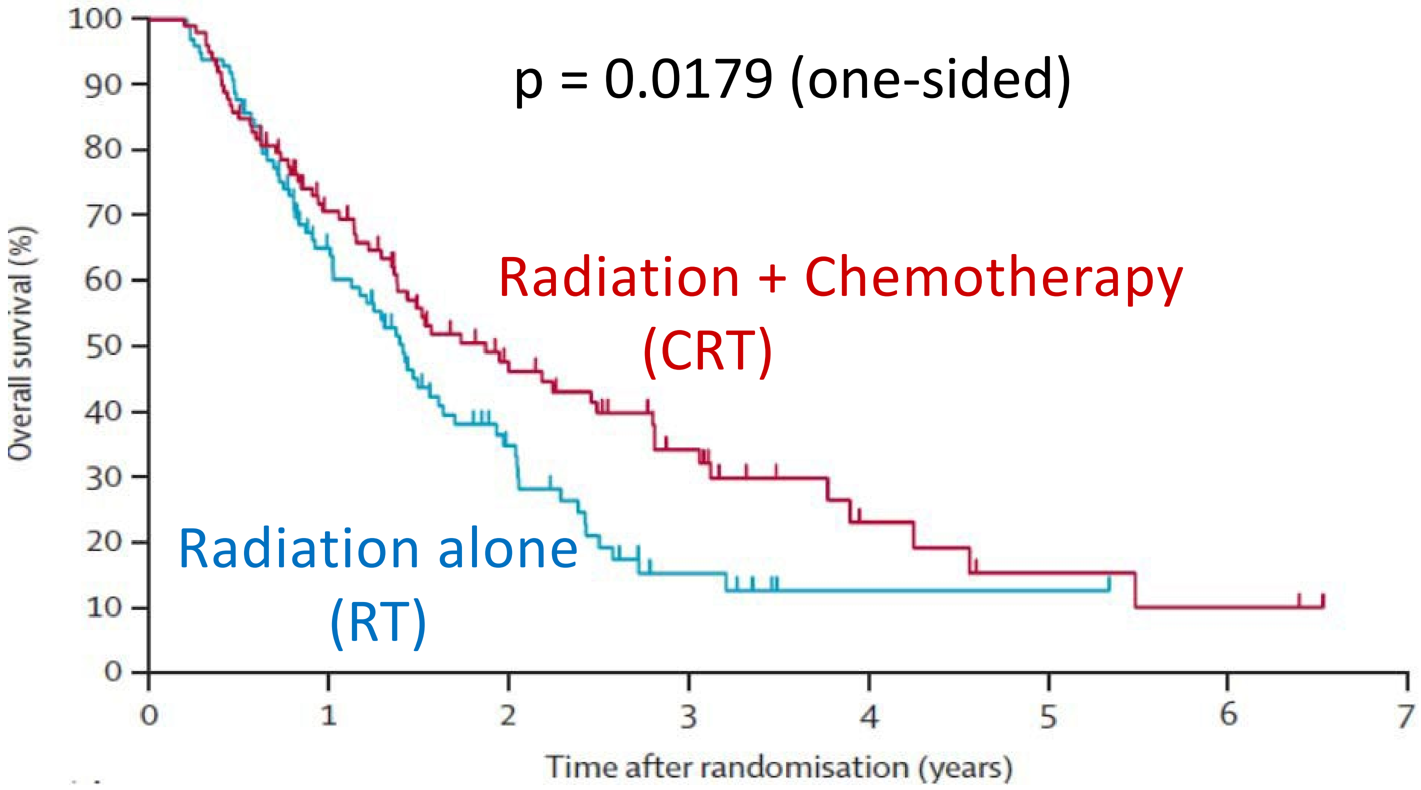
Family history	Case Group	Control group	*P* value
Yes	120 people	156 people	
No	268 people	232 people	0.201

Ma et al. (2009) 61(4), 447-456 Nutr Cancer

- Comparison of survival curves and so on

Today's Aim:

To be able to interpret results of hypothesis testing



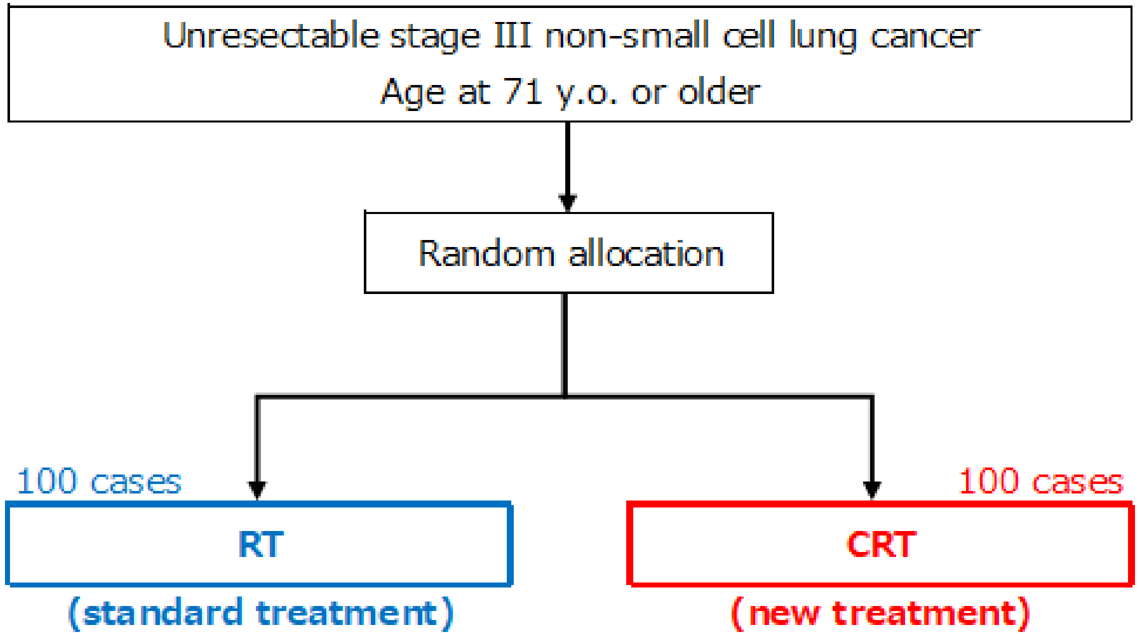
Atagi *et al.* (2012) Lancet Oncol 13(7): 671-8.

Today's Lecture

- Hypothesis Testing
  - Null hypothesis
  - Alternative hypothesis
  - *P*-value
- Errors Associated with Hypothesis Testing
  - α error
  - β error
  - Statistical power
- Sample Size Design
- Limitations of Hypothesis Testing

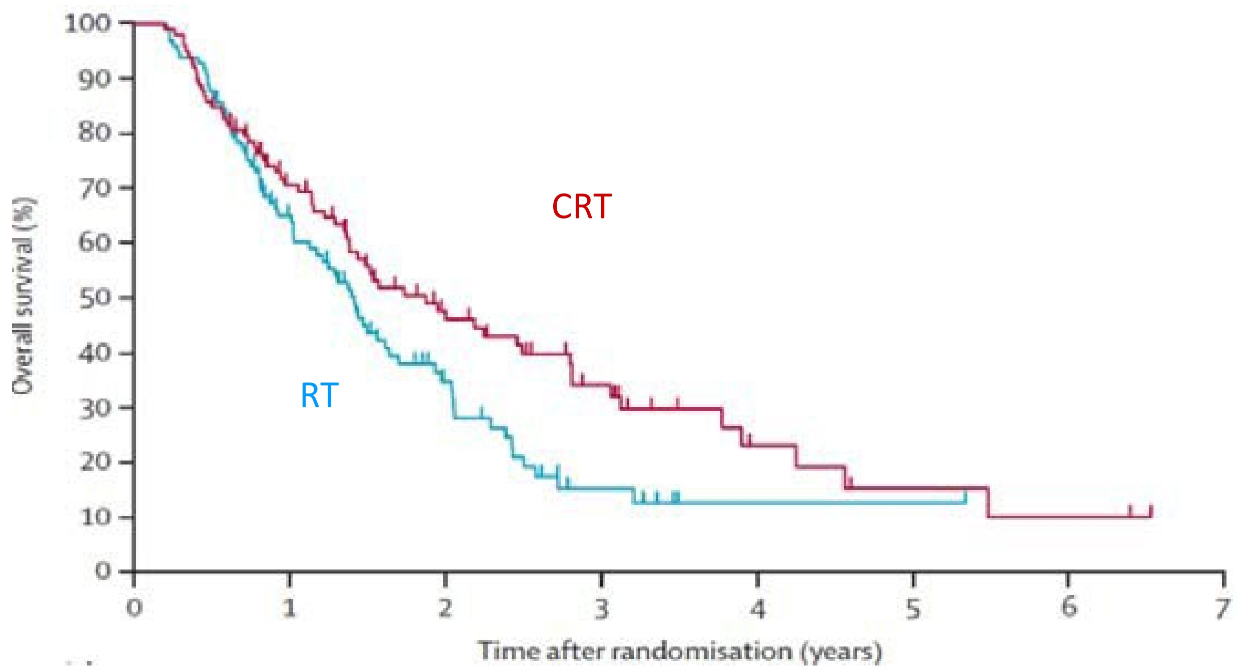
JCOG0301 Study

- Participants: 71 years and up with unresectable stage III non-small cell lung cancer
- Research question: Can radiation + chemotherapy (CRT) reduce mortality than radiation alone (RT)?



Evaluation of Research Questions

- Research Question: “Can RCT reduce mortality more than RT?”
- Comparison of survival curves between the two groups
- Evaluate whether “there is a difference between the survival curves.”



Two-Group Comparison: Hypothesis Testing

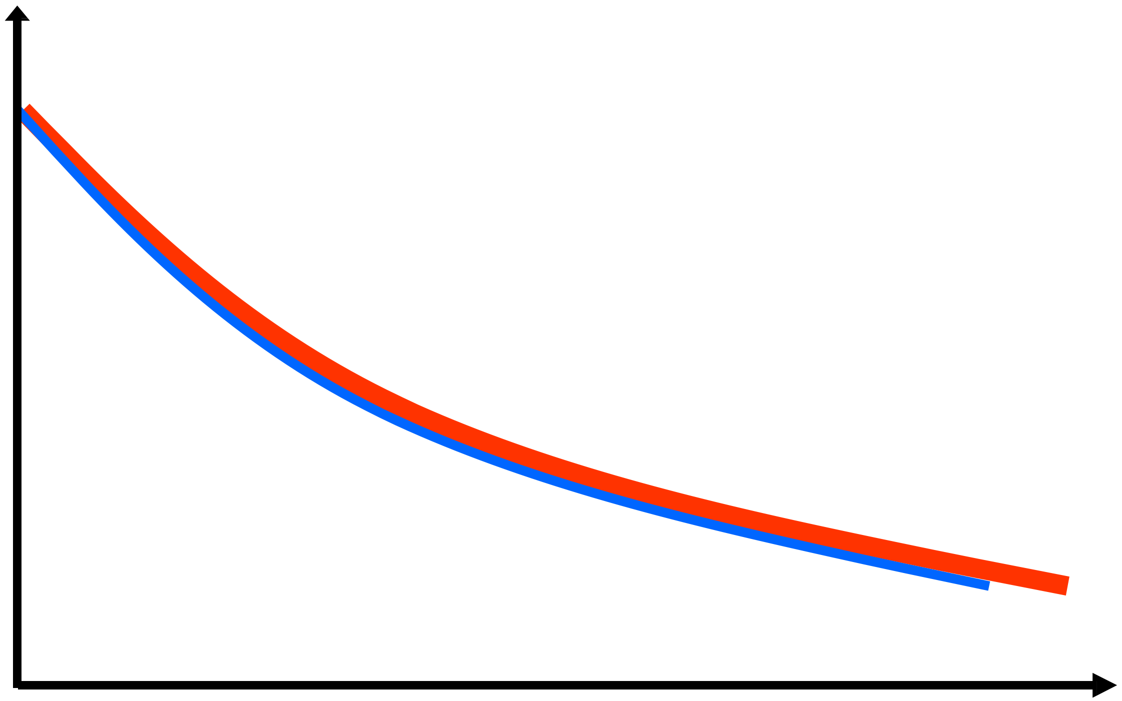
- I want to examine whether there is a “difference between RT and CRT.”
- This method makes the opposite assumption that “there is no difference between RT and CRT,” and draws conclusions by evaluating the extent to which the data contradict the assumption.
  - Drawing conclusions by *reductio ad absurdum*

Two-Group Comparison: Hypothesis Testing

1. Setting the null hypothesis
   - "There is no difference between RT and CRT."
2. Distribution under the null hypothesis
   - Distribution of results obtained when the same test is repeated many times under the hypothesis that "there is no difference between RT and CRT."
3. Calculation of the p-value
   - Determine the probability that the difference becomes greater than or equal to the observed difference between RT and CRT.
4. Decision based on the p-value
   - If the P-value is small, the hypothesis "there is no difference between RT and CRT" is wrong and judge that "there is a difference between RT and CRT."
   - The hypothesis, "there is a difference between RT and CRT," is called the "alternative hypothesis."

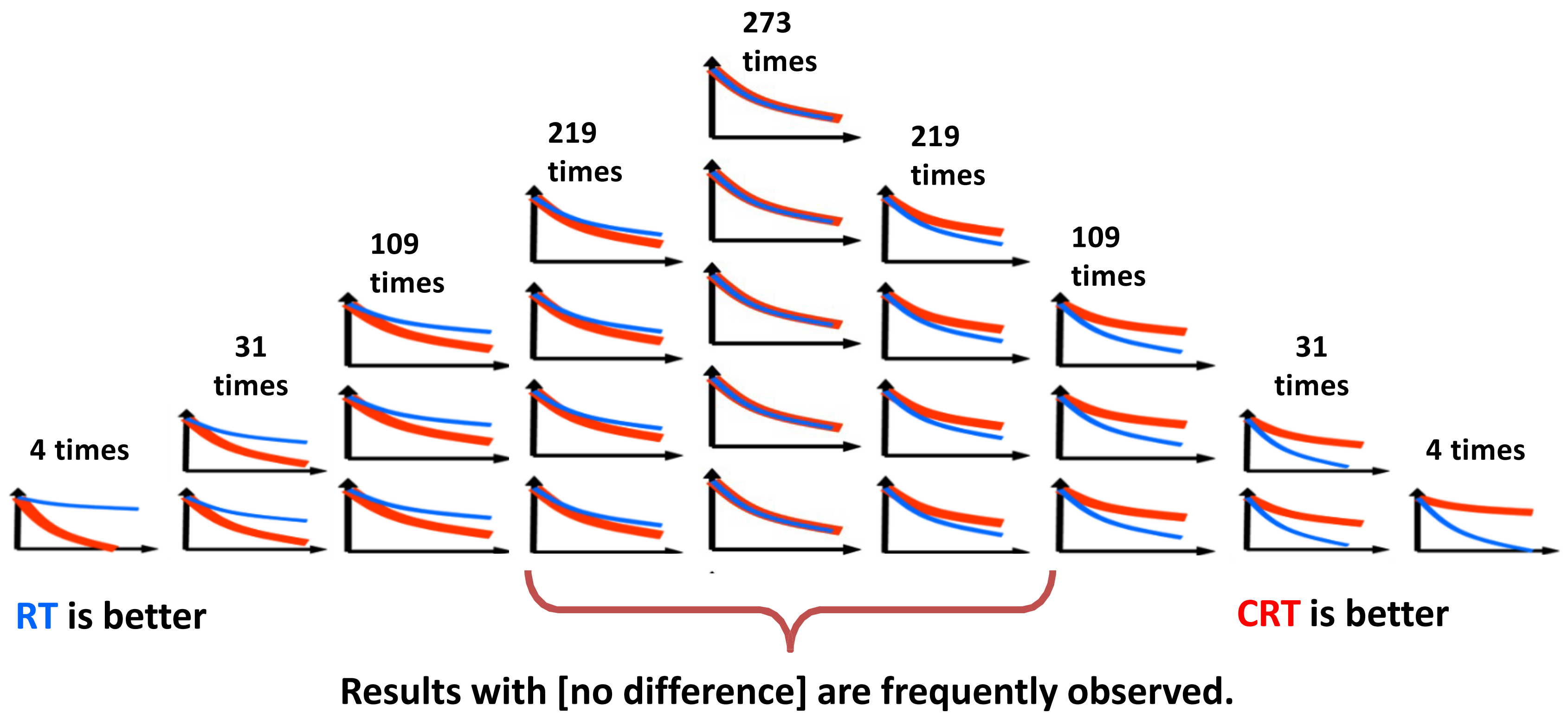
1. Setting the Null Hypothesis

- “There is no difference between RT and CRT.”



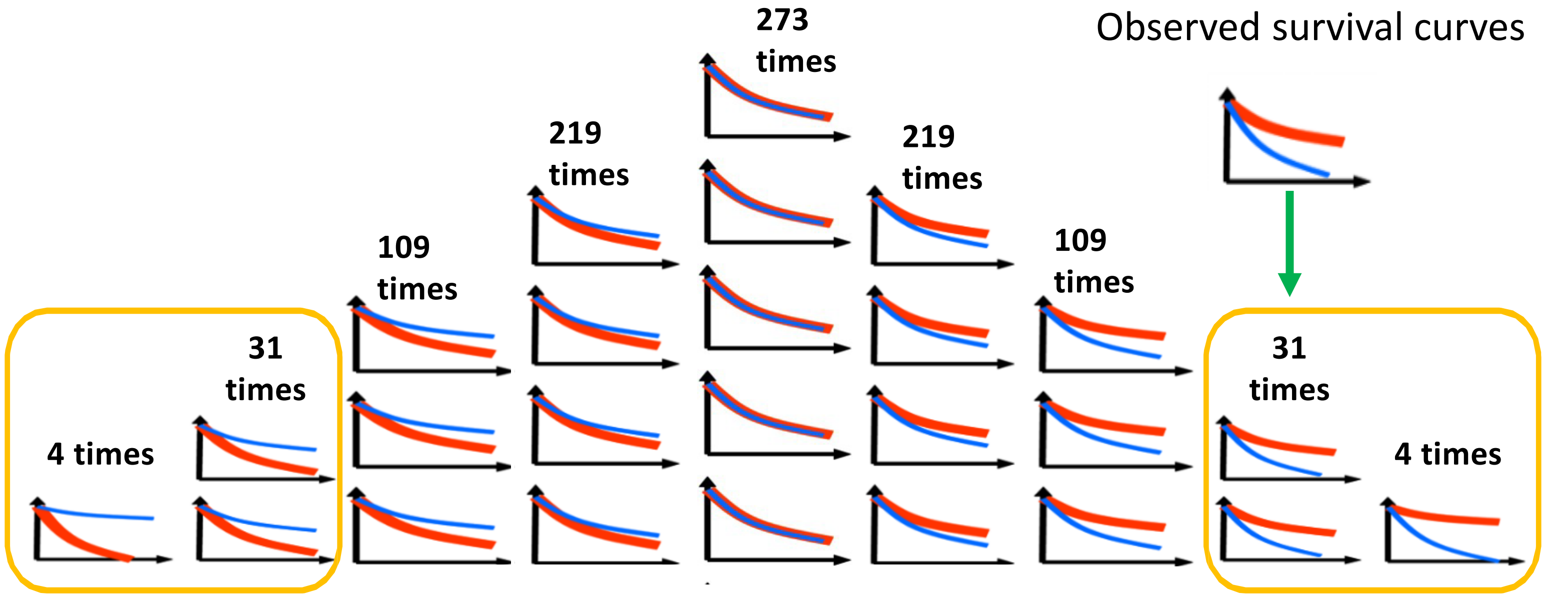
2. Distribution under the Null Hypothesis

- If a test is repeated 1,000 times with 200 participants…



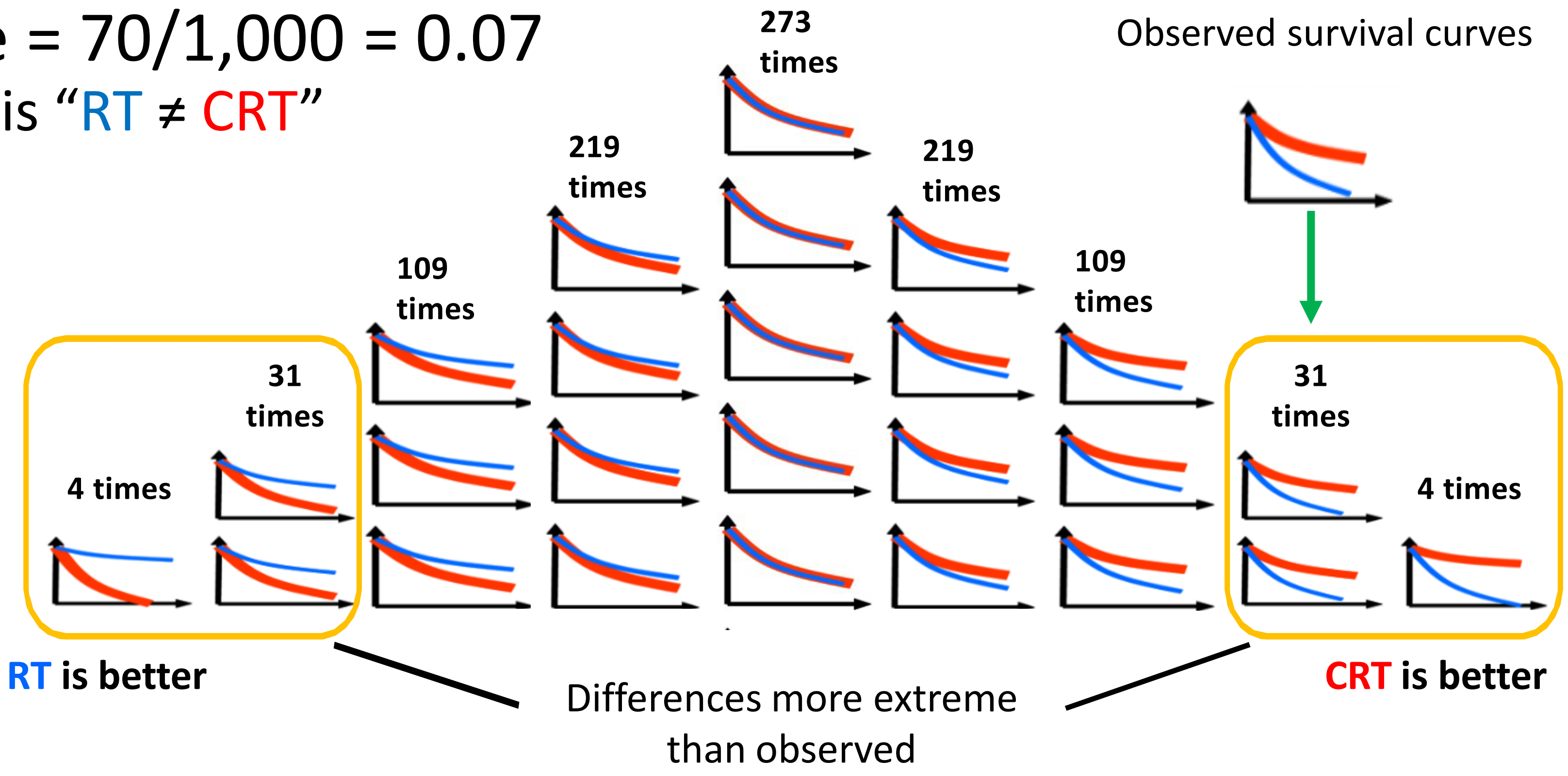
3. Calculating the p-value

- Comparison of distributions and observed data under the null hypothesis
- Probability of the difference being greater than the observed difference between RT and CRT
- p-value = 35/1,000 = 0.035
  If [no difference] is true, the result is likely to occur about 35 times out of 1,000 times.



3. Note on Calculating the p-value: One-Sided and Two-Sided p-values

- One-sided p-value = 35/1,000 = 0.035
  Alternative hypothesis "RT < CRT"
- Two-sided p-value = 70/1,000 = 0.07
  Alternative hypothesis "RT ≠ CRT"



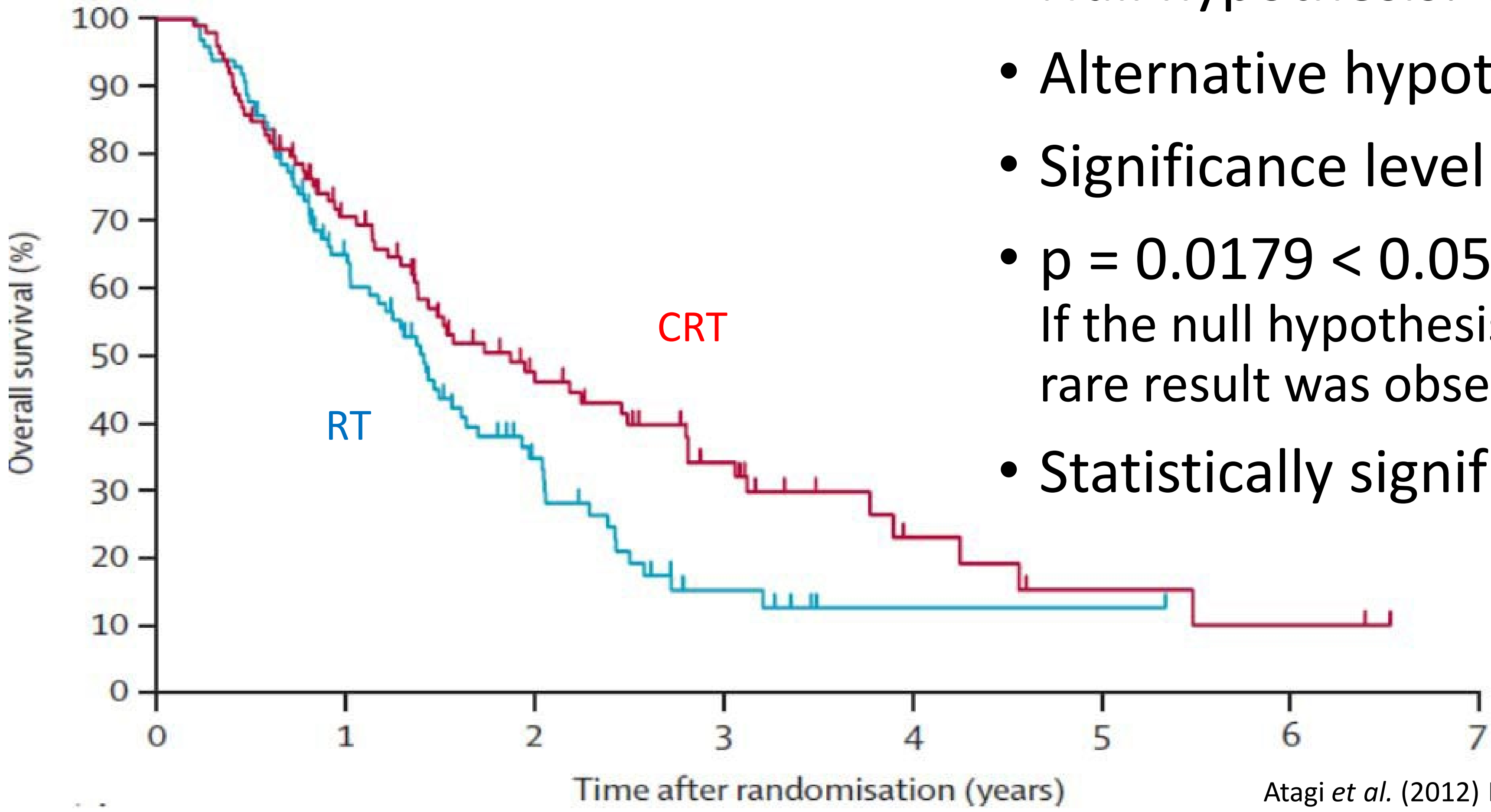
3. Calculating the p-value: Interpretation

- The p-value is the probability of observing a result more extreme than the data under the null hypothesis “no difference between RT and CRT”
- Large p-value
  …The obvious happened.
  …The null hypothesis was probably correct.
- Small p-value
  …Something rare happened.
  …Perhaps the assumption of the null hypothesis was wrong.

4. Decisions Based on the p-value

- p-value = 35/1,000 = 0.035
  - If [no difference] is true, a rare result was observed.
  - Perhaps the assumption that [there is no difference] is wrong.
- If the result is considered sufficiently rare, reject the null hypothesis [no difference] and adopt the alternative hypothesis [there is a difference]
- Significance level: Criteria for whether the results are sufficiently rare or not.
  - Decide and specify in advance to avoid retrofitting.

Results of JCOG0301



- Null hypothesis: CRT = RT
- Alternative hypothesis: CRT > RT
- Significance level: 5%
- $p = 0.0179 < 0.05$
  If the null hypothesis is correct, then a rare result was observed.
- Statistically significant difference

Atagi *et al.* (2012) Lancet Oncol 13(7): 671-8.

Two-Group Comparison: Hypothesis Testing

- Setting the null hypothesis
  “There is no difference between RT and CRT.”
- Distribution under the null hypothesis
  Distribution of results obtained when the same test is repeated many times under the hypothesis that "there is no difference between RT and CRT."
- Calculation of the p-value
  Determine the probability that the difference becomes greater than or equal to the observed difference between RT and CRT
- Decision based on the p-value
  If the P-value is small, the hypothesis “there is no difference between RT and CRT” is wrong and judge that “there is a difference between RT and CRT.”
  The hypothesis, “there is a difference between RT and CRT,” is called the “alternative hypothesis.”

Today's Lecture

- Hypothesis Testing
  - Null hypothesis
  - Alternative hypothesis
  - *P*-value
- Errors Associated with Hypothesis Testing
  - α error
  - β error
  - Statistical power
- Sample Size Design
- Limitations of Hypothesis Testing

α Error

- Error in judging that [there is a difference] when [there is really no difference].
- False positive
- Also known as a Type I error

β Error

- Incorrectly concluding [no difference] when [there is a difference].
- False negative
- Also known as a Type II error
- Power = 1 – β
  Probability of determining [there is a difference] when [there is a difference]

Summary of Decisions Based on Hypothesis Testing

		True	
		Null hypothesis is true (No difference)	Null hypothesis is false (There is a difference)
Decision	Accept the null hypothesis	Correct	Wrong β error
	Reject the null hypothesis	Wrong α error	Correct

- Test results are not always correct.

Relationship Between α and β Errors

- Want to reduce α error

▼

- Always accept the null hypothesis

▼

- α error = 0%
- β error = 100%

- Want to reduce β error

▼

- Always reject the null hypothesis

▼

- α error = 100%
- β error = 0%

The two types of errors cannot be minimized simultaneously.

Today’s Lecture

- Hypothesis Testing
  - Null hypothesis
  - Alternative hypothesis
  - *P*-value
- Errors Associated with Hypothesis Testing
  - α error
  - β error
  - Statistical power
- Sample Size Design
- Limitations of Hypothesis Testing

Sample Size Design

- How many participants are needed to make a reliable judgement on a research question?
  - No answer to the research question can be obtained if the sample size is too small (small power)
  - Costly and unethical if too large

Three Pieces of Information Needed for Sample Size Design

- Size of treatment effect: Δ (delta)
  - Difference in minimum clinically meaningful treatment effect required for a new treatment to replace the standard of care
- Significance level: α
  - Allowable α error
- Power: $1 – \beta$
  - Probability of judging a truly effective treatment to be correctly effective

Sample Size Design for JCOG0301

- Magnitude of treatment effect:
  - Median survival time in RT group, 10 months
  - Median survival time in CRT group, 15 months
  - Treatment effect size Δ = 5 months
- Significance level = 5%
- Power = 80%
- Sample Size = 200

Sample Size Design for Single Group and Response Rate (JCOG1002)

Katayama *et al.* (2012) Jpn J Clin Oncol 42(6): 556-9.

- Effect size:
  - Threshold response rate = 65%
  - Expected response rate = 80%
- Significance level = 10%
- Power = 80%
- Sample size = 50

Sample Size Design Tool
SWOG Statistical Center: Statools

http://www.swogstat.org/stat/public/

SWOG
Leading cancer research. Together.

One Sample Binomial

Select Calculation and Test Type

(●) Sample Size	(●) 1 Sided
( ) Power	( ) 2 Sided

Select Hypothesis Test Parameters

Null Proportion	Alternative Proportion	Alpha
0.65	0.80	.1

Calculate Power/Sample Size

Power	Sample Size	Approx Lower Count Critical Value	Aprox Upper Count Critical Value
0.80	40	-1	30

Setting of α, 1 – β

- Significance level α: Usually 5 – 10%
  - (From the perspective of protecting a participant) minimize the error judging the ineffective treatment as good if at all possible
- Power 1 – β: Usually 80 – 90%
  - (Primarily from the perspective of the trial conductor) so that effective treatment can be said to have the right effect
- From the perspective of participant protection
  - Reducing α should be a priority
  - Decide the number of participants required to achieve a power of 80% or more with a small α

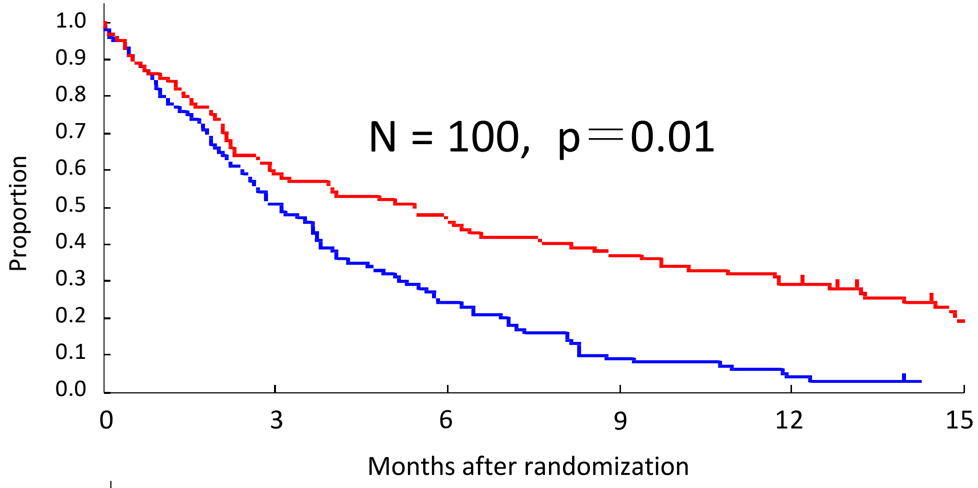
Today's Lecture

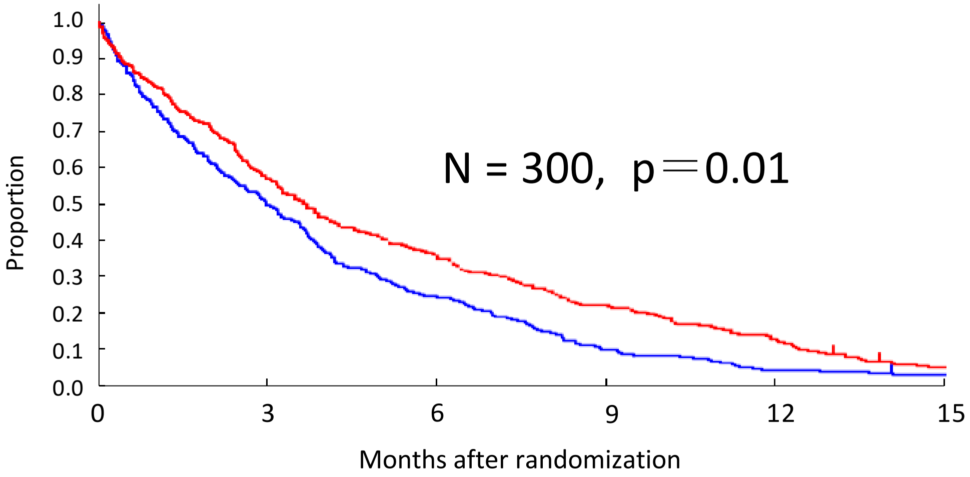
- Hypothesis Testing
  - Null hypothesis
  - Alternative hypothesis
  - *P*-value
- Errors Associated with Hypothesis Testing
  - α error
  - β error
  - Statistical power
- Sample Size Design
- Limitations of Hypothesis Testing

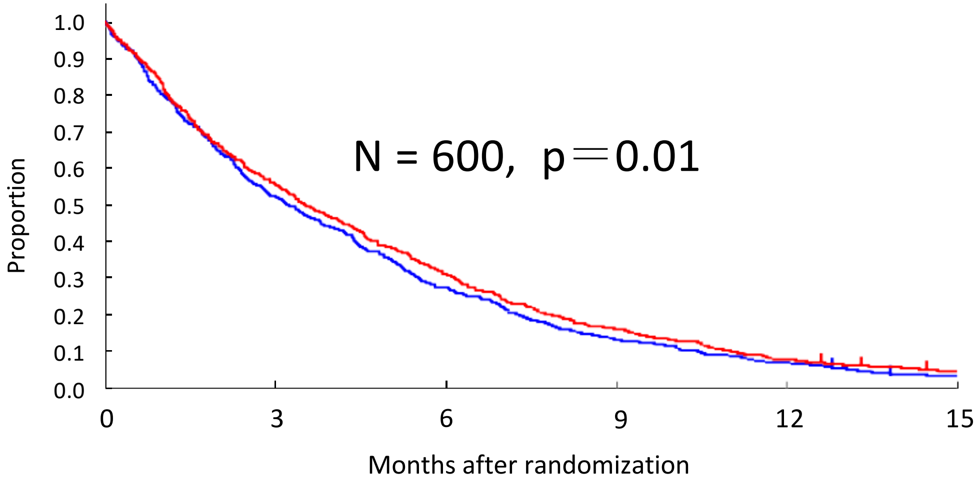
Limitations: p-value Does Not Depict the Size of the Effect

- Studies with smaller p values do not imply greater treatment effects
- The size of the treatment effect needs to be “presumed”

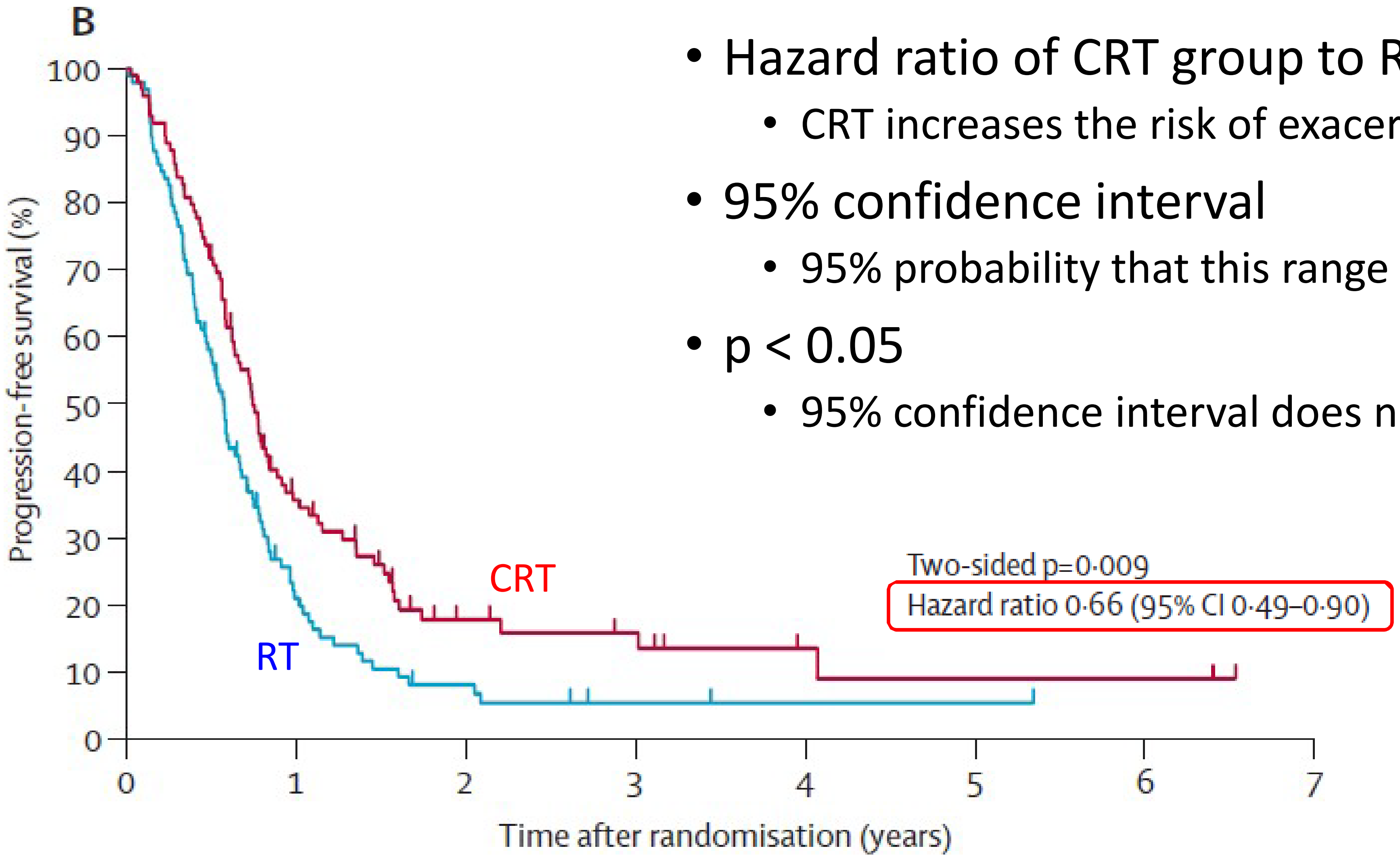
Hypothetical examples







Estimation of Effects



- Hazard ratio of CRT group to RT group is 0.66
  - CRT increases the risk of exacerbation by 0.66 times.
- 95% confidence interval
  - 95% probability that this range contains a true value
- $p < 0.05$
  - 95% confidence interval does not include 1 (no effect)

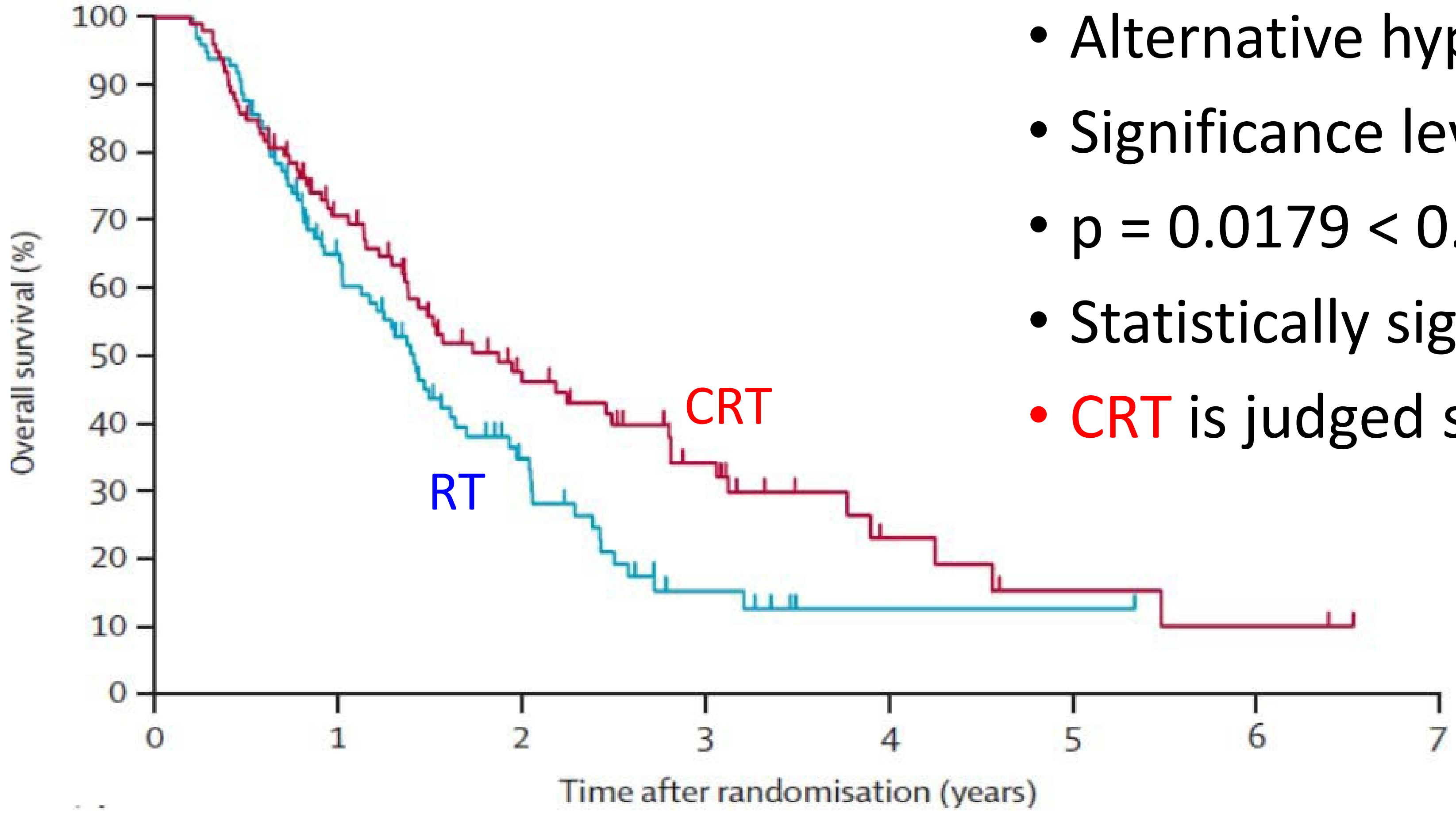
Multiplicity of Tests

- Repeated tests increase the probability that at least one of the null hypotheses will be falsely rejected.
  - Across the studies, many treatments are erroneously assumed to be effective.

- I want to compare three treatment methods, A, B, and C
- A vs. B, A vs. C, B vs. C (5% significance level) three times.
- Probability of finding a group difference somewhere by chance > 5%

- Multiplicity adjustment: Controlling α error for the entire test by reducing the α error allowed for individual tests
  - For example, Bonferroni's method: Significance level for each test = 0.05/(number of testing)
  - Even if we obtain a P value of 0.03 for a group comparison, 0.03 > 0.05/3 = 0.017, as we test 3 times
  - No significant difference

Today's Summary: Interpret Test Results Correctly



- Null hypothesis: CRT = RT
- Alternative hypothesis: CRT > RT
- Significance level 5%
- $p = 0.0179 < 0.05$
- Statistically significant difference
- CRT is judged superior to RT

Atagi *et al.* (2012) Lancet Oncol 13(7): 671-8.