Minimum statistical knowledge required for clinical trials on cancer

~Key points for interpreting the results of phase III randomized controlled trials~

Part 2 of 2

Ryunosuke Machida

JCOG(*) Data Center

The 23rd JCOG Clinical Trial Seminar

10/10/2020

JCOG Japan Clinical Oncology Group

*** Japan Clinical Oncology Group** (https://jcog.jp/en/)

ATLAS ASIAN CLINICAL TRIALS NETWORK FOR CANCERS PROJECT

ICRweb: https://www.icrweb.jp/icr_index.php?lang=en

Outline ~What to know for interpreting a randomized trial~

- Interpretation of results of a **survival curve**
  - Annual survival rate, median survival time
- Why is **randomization** necessary?
  - **Confounding and randomization**
- Result verification method
  - Concept of **hypothesis testing** and meaning of **p-value**
  - Comprehension of **α error, β error, and statistical power**
- Views on magnitude of treatment effects
  - Meaning of **hazard ratio**
- What is an analysis set?
  - **Intention-to-treat analysis** (ITT analysis)

Did the CRT group win?

We now know that a comparison is possible because of randomization. Indeed, the CRT group's survival curve was higher than the RT group's survival curve, but is it safe to say that the CRT group wins if its survival curve was higher after randomization?

Lung Cancer Medical Group
JCOG0301

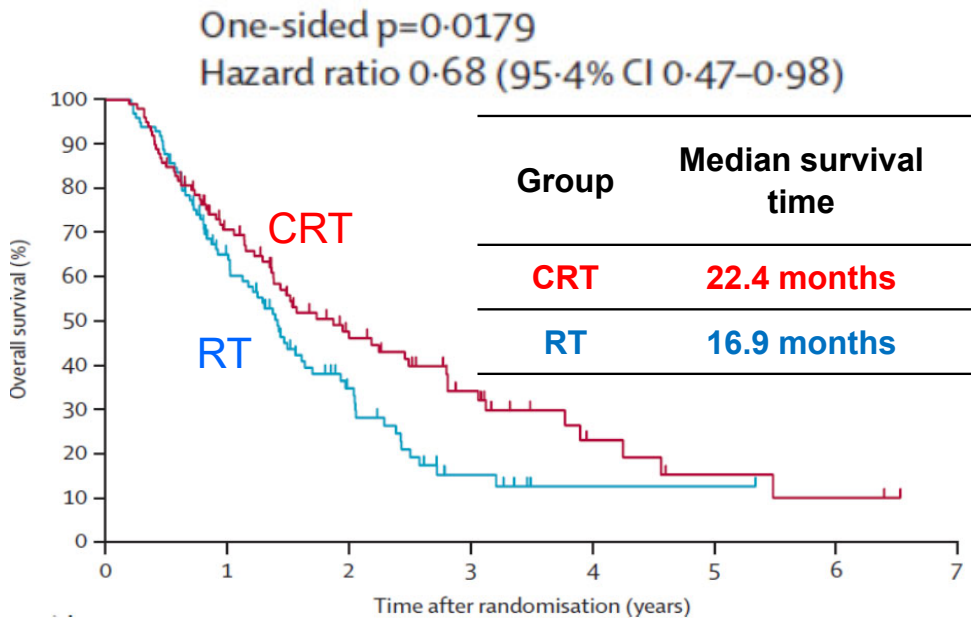
Unresectable stage III non-small cell lung cancer in patients aged ≥71 years

↓

Randomization

100 cases → **Radiotherapy (RT)** **(standard treatment)**

100 cases → **Radiotherapy + chemotherapy (CRT)** **(new treatment)**

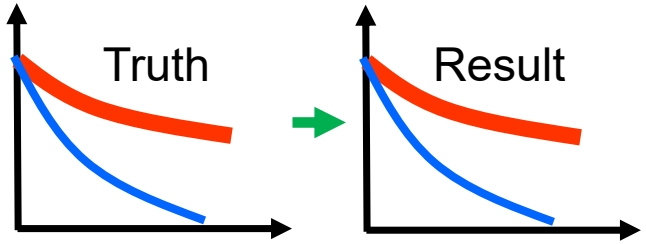


Group	Median survival time	Two-year survival rate
CRT	22.4 months	46.3%
RT	16.9 months	35.1%

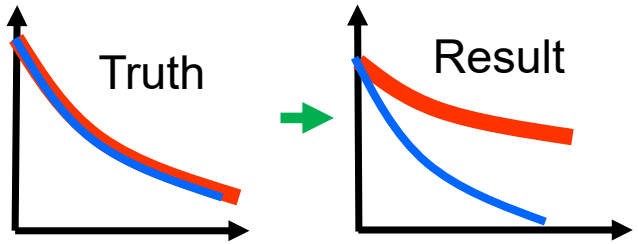
Atagi *et al.* (2012) Lancet Oncology 13(7): 671-8.

Interpretation when there is a gap in survival curves

- **There are two possibilities. Which one is correct?**
  - There really is a "difference between **RT** and **CRT,**" so an actual difference is observed
    - Correct conclusion is obtained



  - Even when there is actually "no difference between **RT** and **CRT**," a difference in noted by chance
    - Wrong conclusion is reached



I want to confirm which is correct from the obtained results!

Confirmation method: hypothesis testing

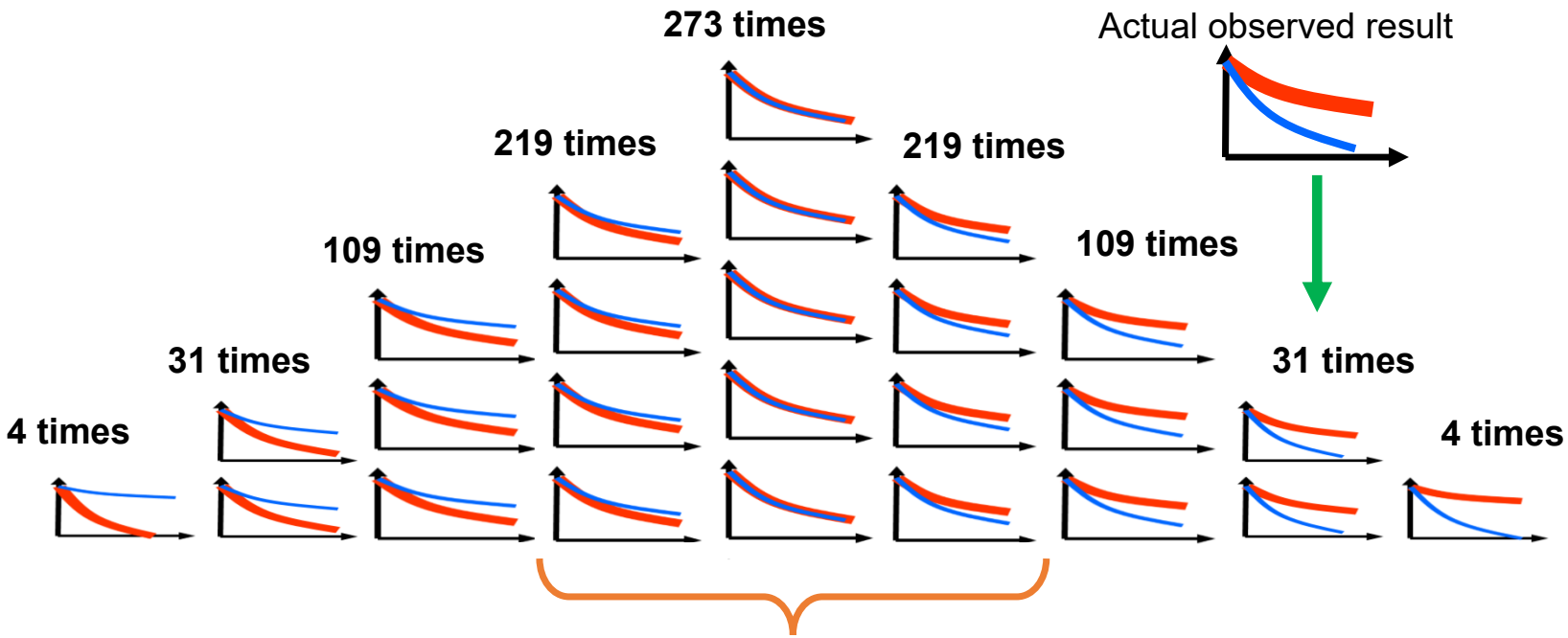
- We want to prove that "there is a difference between **RT** and **CRT**"

1. Set the hypothesis that "there is no difference between **RT** and **CRT**"
   - This hypothesis is called the **null hypothesis**
2. Under the hypothesis that "there is no difference between **RT** and **CRT,**" examine the distribution of results obtained when the trial is repeated multiple times
3. Under the hypothesis that "there is no difference between **RT** and **CRT,**" examine the probability that the difference is larger than the observed difference between **RT** and **CRT**
4. If this probability is small, then the hypothesis that "there is no difference between **RT** and **CRT**" (null hypothesis) is judged to be wrong in the first place
5. The hypothesis that "there is a difference between **RT** and **CRT**" is judged to be correct

Distribution of results under [no difference] between the survival curves of RT and CRT

If “[no difference] between the survival curves of RT and CRT” was true…

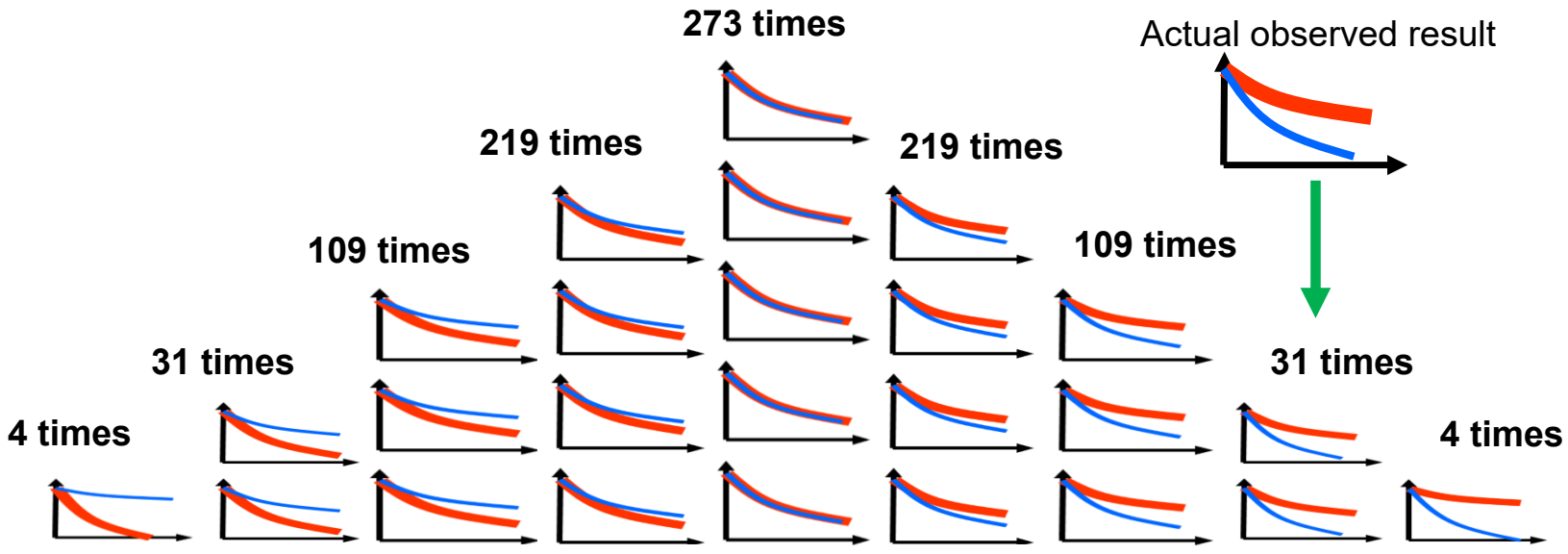
If selecting 200 patients from those aged ≥71 years who had unresectable stage III non-small cell lung cancer from all over Japan and conducting 1,000 trials…



[No difference] results are the most frequently observed findings

P-value calculation

- **P**robability of obtaining a larger difference than the actual observed result is 35/1000 = **3.5%**
  - This probability is called the **p-value**
- If it was true that the actual observed result had [no difference], this would be a **rare result** (?) that occurs about 35 put of 1000 times



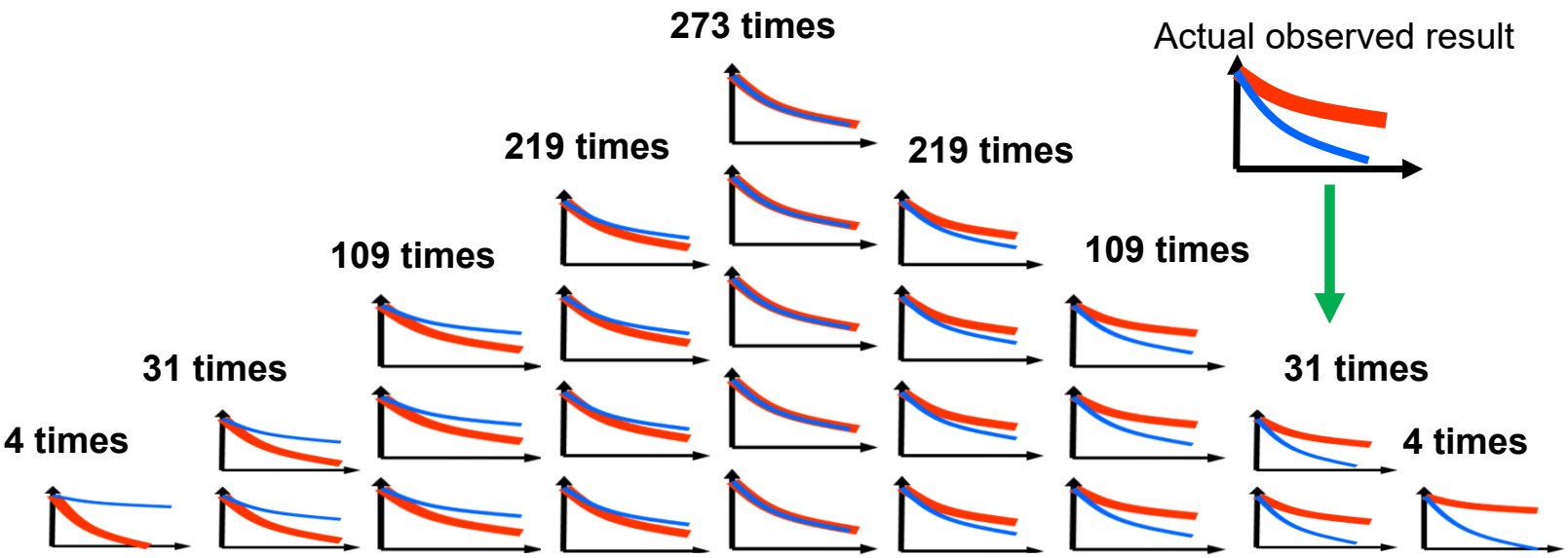
Is 3.5% a rare result?

- If 3.5% **is considered a rare result**
  - It is judged that the hypothesis of [no difference] is wrong in the first place, and **the conclusion is that there is a difference** between RT and CRT = [there is a significant difference]
- If 3.5% **is not considered a rare result**
  - It cannot be said that the hypothesis of [no difference] is wrong; therefore, **the conclusion is that there is no difference** between RT and CRT = [no significant difference]

- Judging whether a result is rare after analyzing it will be an afterthought, so the criteria for whether the result is rare is decided in advance
  - This criterion is called the **significance level (α level)**
  - If the P-value falls below the significance level, the conclusion is that [there is a significant difference]

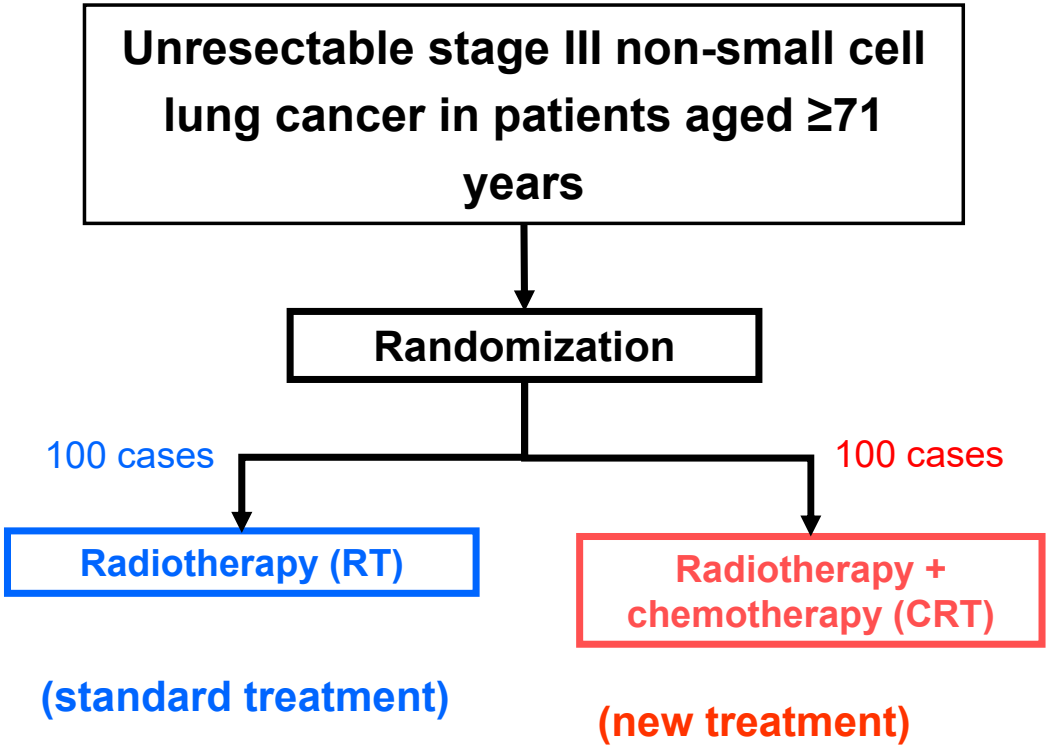
Test result

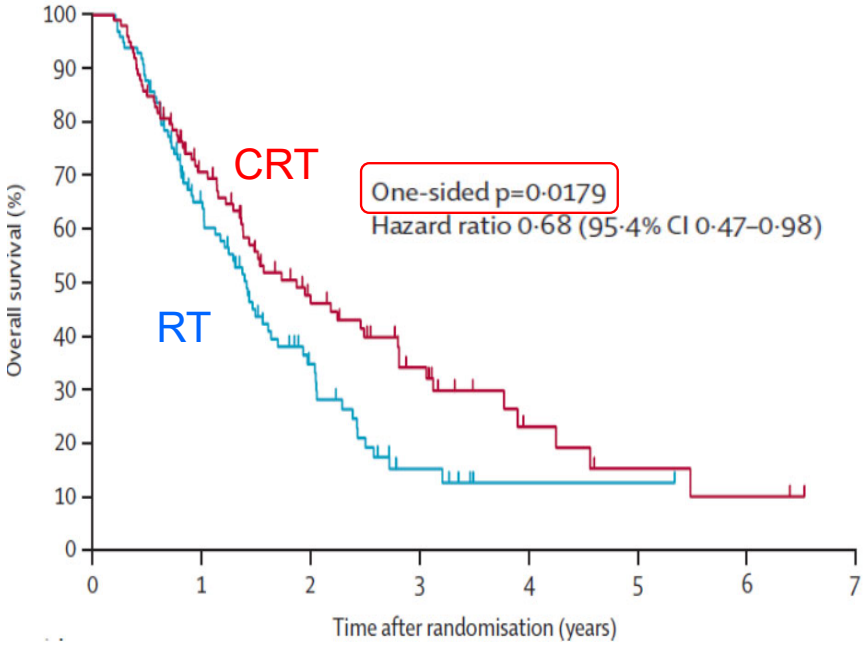
- P-value=3.5%
  - If [no difference] were true, the actual observed result would occur about 35 out of 1,000 times
- If the significance level was set to **5%**, a significant difference would be considered
- If the significance level was set to **2.5%**, no significant difference would be considered



JCOG0301 case

- **p=0.0179:** Result occurs only 1–2 out of 100 if there is no difference between the groups
  - Satisfies the pre-determined criterion value of p-value ≤5% (significance level 5%)
  - Judged that CRT is better than RT





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

Outline ~What to know for interpreting a randomized trial~

- Interpretation of results of a **survival curve**
  - Annual survival rate, median survival time
- Why is **randomization** necessary?
  - **Confounding and randomization**
- Result verification method
  - Concept of **hypothesis testing** and meaning of **p-value**
  - Comprehension of **α error, β error, and statistical power**
- Views on magnitude of treatment effects
  - Meaning of **hazard ratio**
- What is an analysis set?
  - **Intention-to-treat analysis** (ITT analysis)

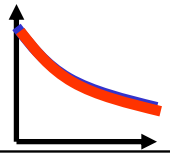
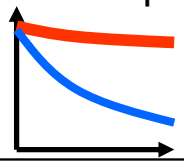
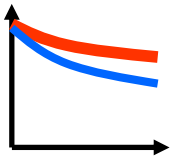
Test result is not always correct

- The actually obtained result has P-value=3.5%
  - This happens only rarely, so the hypothesis of [no difference] was judged to be wrong
  - Conversely, cases wherein [no difference] is true would occur rarely

- If the judgment is that [there is a difference] when the truth is that [there is no difference], the wrong judgment would be made
  - This error is called an **α error**
  - The probability of judging as [difference present] when there is [no difference] is below the significance level, so the probability of an α error is below significance level

Concluding that there is no significant difference despite [difference present]

- This error is called **"β error"**
  - Error of eliminating an actual effective treatment, contrary to truth
- **statistical power** (probability is **1-β**)
  - Probability of correctly judging "difference present" as "difference present"

		Truth	
		Null hypothesis (no difference) 	Alternative hypothesis (difference present) 
Test result 	No significant difference	Correct	Error (**β error**)
	Significant difference	Error (**α error**)	Correct (**statistical power 1-β**)

Outline ~What to know for interpreting a randomized trial~

- Interpretation of results of a **survival curve**
  - Annual survival rate, median survival time
- Why is **randomization** necessary?
  - **Confounding and randomization**
- Result verification method
  - Concept of **hypothesis testing** and meaning of **p-value**
  - Comprehension of **α error, β error, and statistical power**
- Views on magnitude of treatment effects
  - Meaning of **hazard ratio**
- What is an analysis set?
  - **Intention-to-treat analysis** (ITT analysis)

How good of a treatment is CRT?

Note: hypothetical example

Lung Cancer Medical Group
JCOG0301

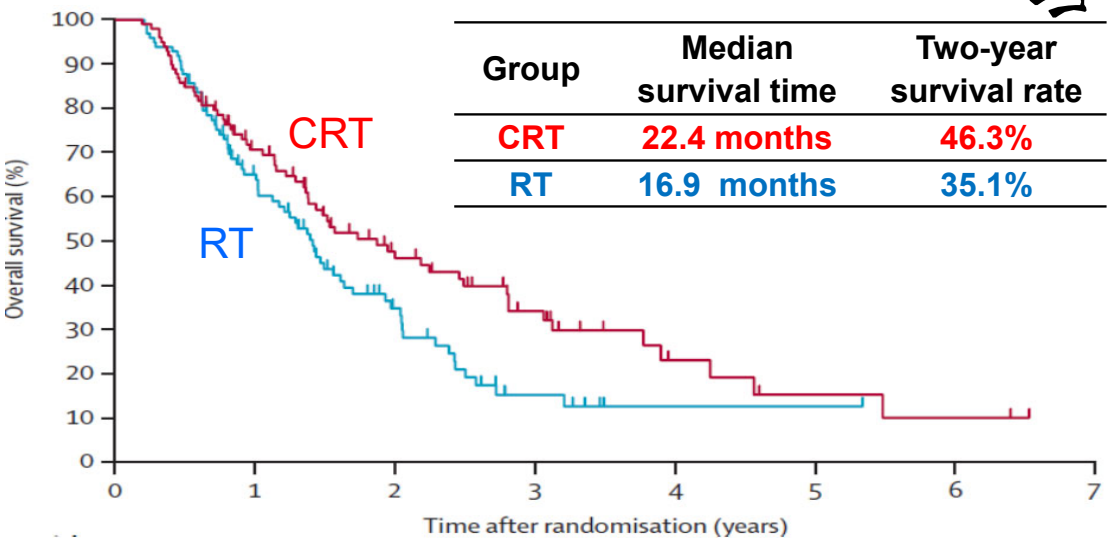
Unresectable stage III non-small cell lung cancer in patients aged ≥71 years

Randomization

100 cases → **Radiotherapy (RT)** **(standard treatment)**

100 cases → **Radiotherapy + chemotherapy (CRT)** **(new treatment)**

We know that the CRT group was better than the RT group, but how good is the treatment method? Does a smaller P-value signify better treatment?



Group	Median survival time	Two-year survival rate
CRT	22.4 months	46.3%
RT	16.9 months	35.1%

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

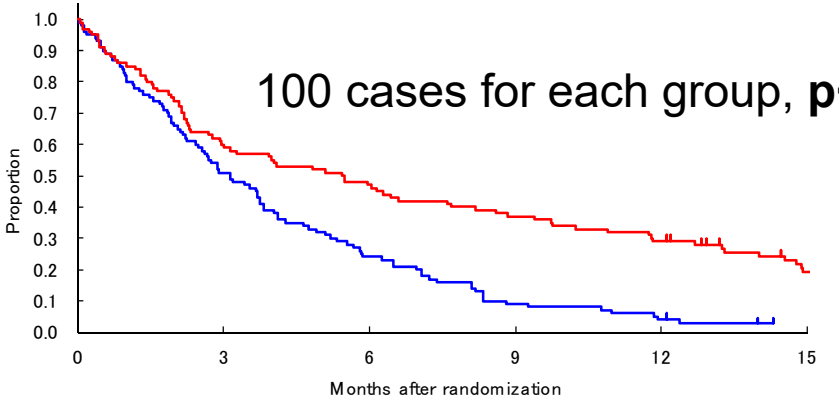
Statistically significant difference ≠ clinically significant difference

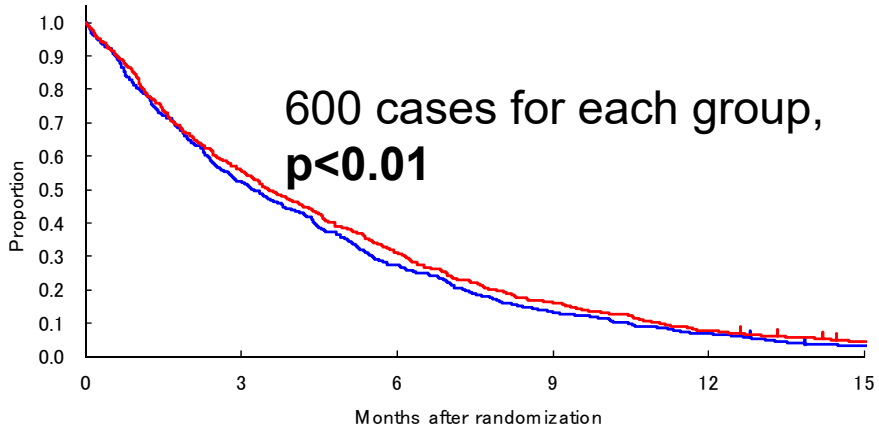
National Cancer Center Japan

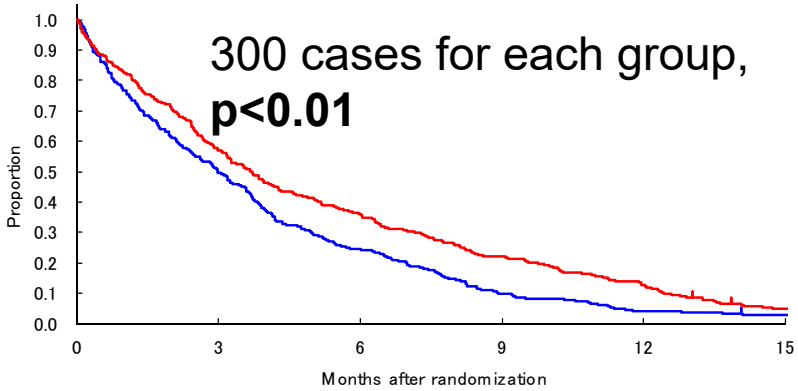
Note: hypothetical example

Same p<0.01 value but different clinical implications

p-value is not a measure of magnitude of treatment effect





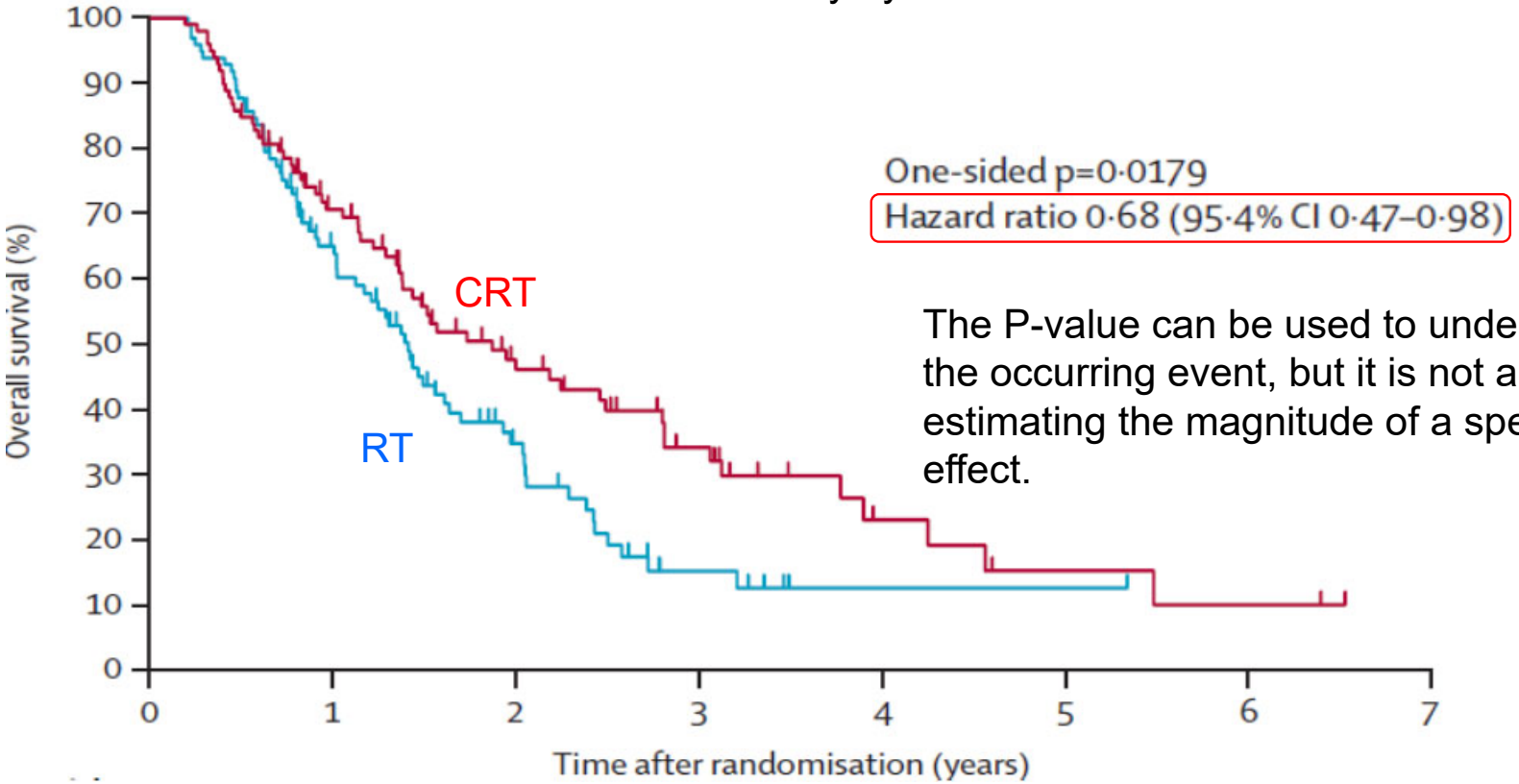


Index that shows magnitude of treatment effect

- Index that focuses on one time point on a curve
  - Difference in annual survival rate
    - Two-year survival rate with CRT was 46.3% vs with RT was 35.1%
  - Difference in median survival time (MST)
    - CRT: 22.4 months vs RT: 16.9 months
- Index that combines an entire curve into a single effect
  - **Hazard ratio** (**HR**: Hazard Ratio)
    - Ratio of **hazard rate (instantaneous mortality rate)** between groups

Interpretation in JCOG0301

- **Hazard ratio (HR)** of the CRT group relative to the RT group is **0.68**
  - CRT increases the risk of mortality by 0.68 times
  - CRT reduces the risk of mortality by 32%



The P-value can be used to understand the rarity of the occurring event, but it is not an index for estimating the magnitude of a specific treatment effect.

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

Outline ~What to know for interpreting a randomized trial~

- Interpretation of results of a **survival curve**
  - Annual survival rate, median survival time
- Why is **randomization** necessary?
  - **Confounding and randomization**
- Result verification method
  - Concept of **hypothesis testing** and meaning of **p-value**
  - Comprehension of **α error, β error, and statistical power**
- Views on magnitude of treatment effects
  - Meaning of **hazard ratio**
- What is an analysis set?
  - **Intention-to-treat analysis** (ITT analysis)

How to handle patients who were not properly treated?

National Cancer Center Japan

Lung Cancer Medical Group
JCOG0301

Unresectable stage III non-small cell lung cancer in patients aged ≥71 years

Randomization

100 cases

Radiotherapy (RT)

RT cases: 98
RT completion: 93
RT non-completion: 5

Untreated cases: 2

(standard treatment)

100 cases

Radiotherapy + chemotherapy (CRT)

CRT cases: 96
CRT completion: 88
CRT non-completion: 8

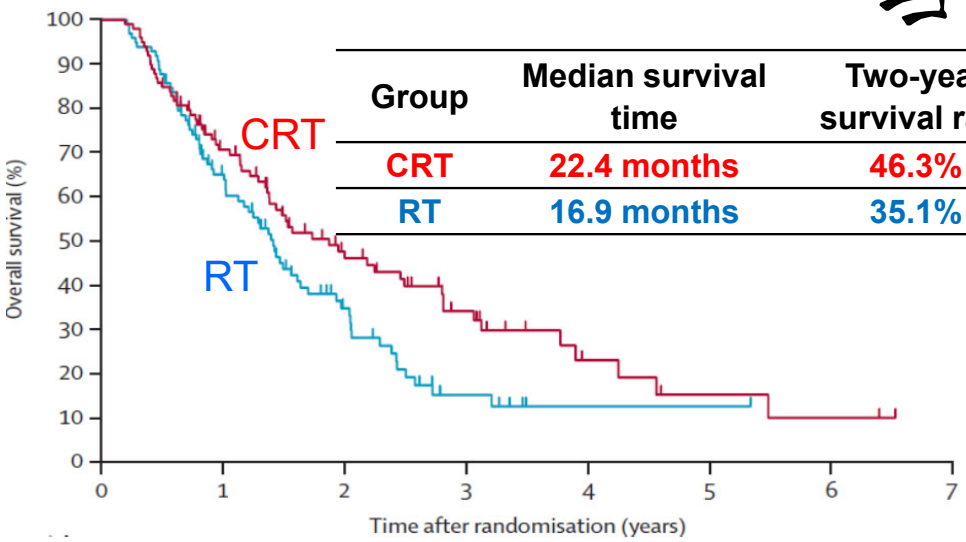
Untreated cases: 4

(new treatment)

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

But there are patients in both the RT and CRT groups who haven't been able to receive proper treatment!
Are these patients included in the group comparison analysis of survival curves?

One-sided p=0·0179
Hazard ratio 0·68 (95·4% CI 0·47–0·98)



Group	Median survival time	Two-year survival rate
CRT	22.4 months	46.3%
RT	16.9 months	35.1%

Question 2: Which analysis method would you choose?

- In the case like that on the right, which analysis method should you choose?
- Please select one of the following.

① Comparison between completed treatment cases

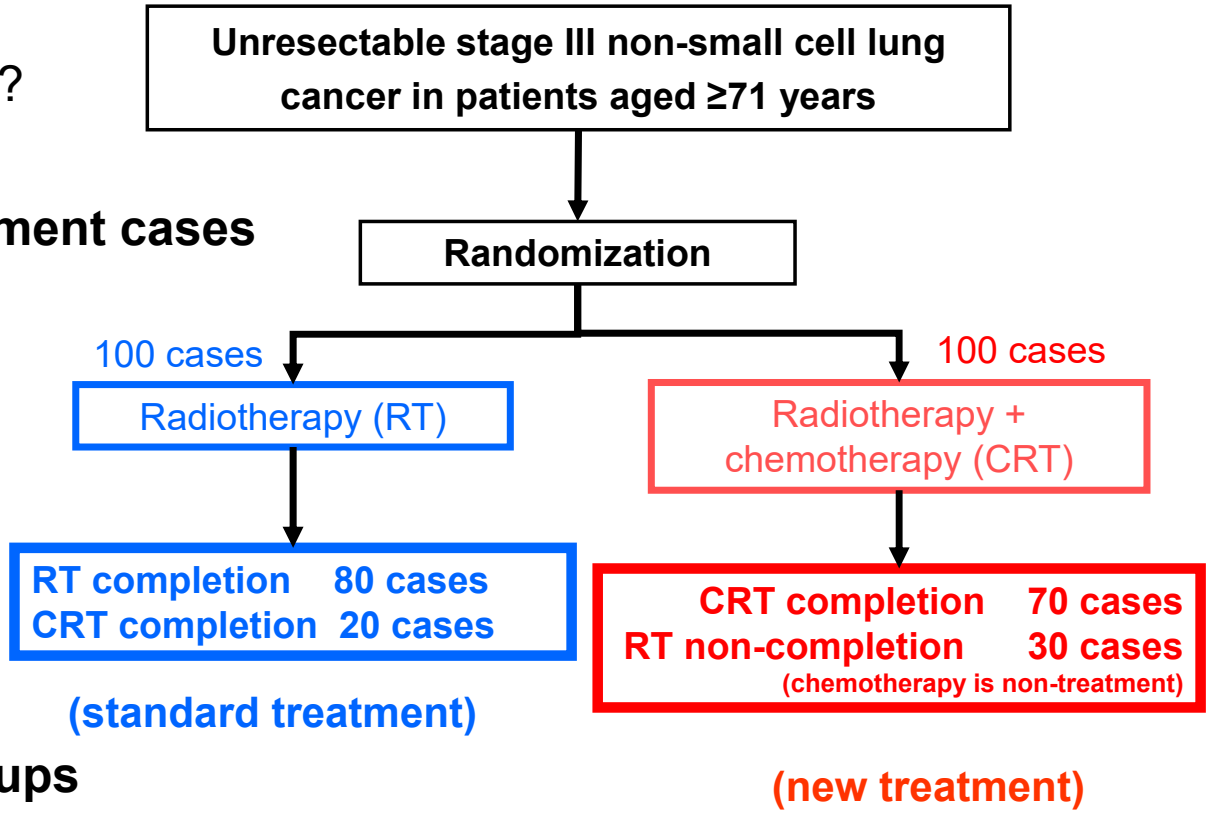
RT 80 cases vs. CRT 70 cases

② Comparison between treatments actually conducted

RT (80+30) cases vs. CRT (70+20) cases

③ Comparison between randomized groups

RT 100 cases vs. CRT 100 cases



When considering which result is predicted

Increased α error

Increased probability of mistakenly saying there is a difference when there is no difference

Decreased statistical power

① **Comparison between completed treatment cases:** RT 80 cases vs. CRT 70 cases

- RT cases are those excluding healthy individuals who could undergo CRT
- CRT cases are those excluding unhealthy individuals who could only undergo RT

② **Comparison between actual treatments conducted:** RT (80+30) cases vs. CRT (70+20) cases

- RT cases are those including unhealthy individuals who could only undergo RT
- CRT cases are those including healthy individuals who could undergo CRT

⇒ In the comparison of (1) and (2), the background factors are not aligned between groups, and the randomization loses its significance

When there is no difference, it can correctly be said that there is no difference

③ **Comparison between randomized groups:** RT 100 cases vs. CRT 100 cases

- If CRT is truly effective, then if there are patients who were allocated to the CRT group but received RT, the treatment effect would be diluted

Intention-to-treat analysis (ITT analysis)

- Analysis conducted with treatment groups **as allocated** by randomization (method of analyzing with subjects in (3))
  - If conducting **ITT analysis**, the probability of α error does not increase
  - This is “conservative” analysis, such that it is less likely to be significant

- If there is a significant difference even with ITT analysis, it can confidently be judged that there is a difference
  - Can be said that “treatment effect of at least XXX is present!”
  - **ITT analysis is standard method for primary analysis in randomized controlled trials**

JCOG0301 case

- Main analysis is the **ITT analysis**, which includes treatment of non-completion cases
  - CRT can be judged to outperform RT

Unresectable stage III non-small cell lung cancer in patients aged ≥71 years

Randomization

100 cases

Radiotherapy (RT)

RT cases: 98
RT completion: 93
RT non-completion: 5

Untreated cases: 2

(standard treatment)

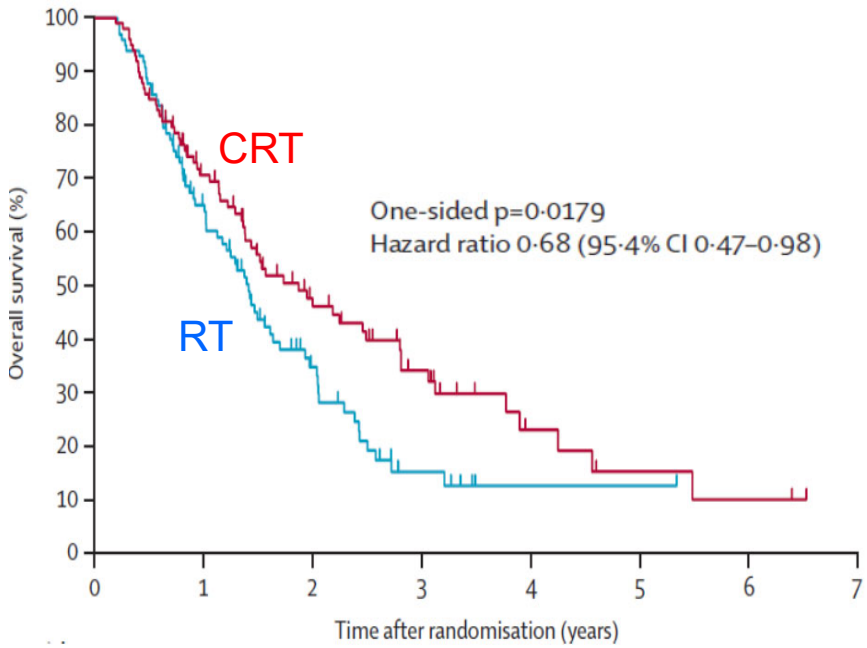
100 cases

Radiotherapy + chemotherapy (CRT)

CRT cases: 96
CRT completion: 88
CRT non-completion: 8

Untreated cases: 4

(new treatment)



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

Summary

- **Survival curves** are used to plot survival rate against time. Results can be judged visually
- **Randomization** eliminates confounding and allows for proper comparison of treatments
- Results are judged as having a difference if the **p-value** obtained from **hypothesis testing** is below the **significance level (α)**
- Magnitude of treatment effect is judged not with p-value but with **hazard ratio** or survival rate
- Primary analysis of clinical trials will involve comparison of treatment groups as allocated (**ITT analysis**)