Multivariate Analysis 1

National Cancer Center, Japan
Biostatistics Division
Shogo Nomura

Outline

- Review of the Fourth Lecture
- What Is a (Statistical) Model?
- Multivariate Analysis for Confounding Adjustment
- Notes on the Use of Regression Models
- Uses of Multivariate Analysis

Review of the Fourth Lecture

- What is “confounding”?

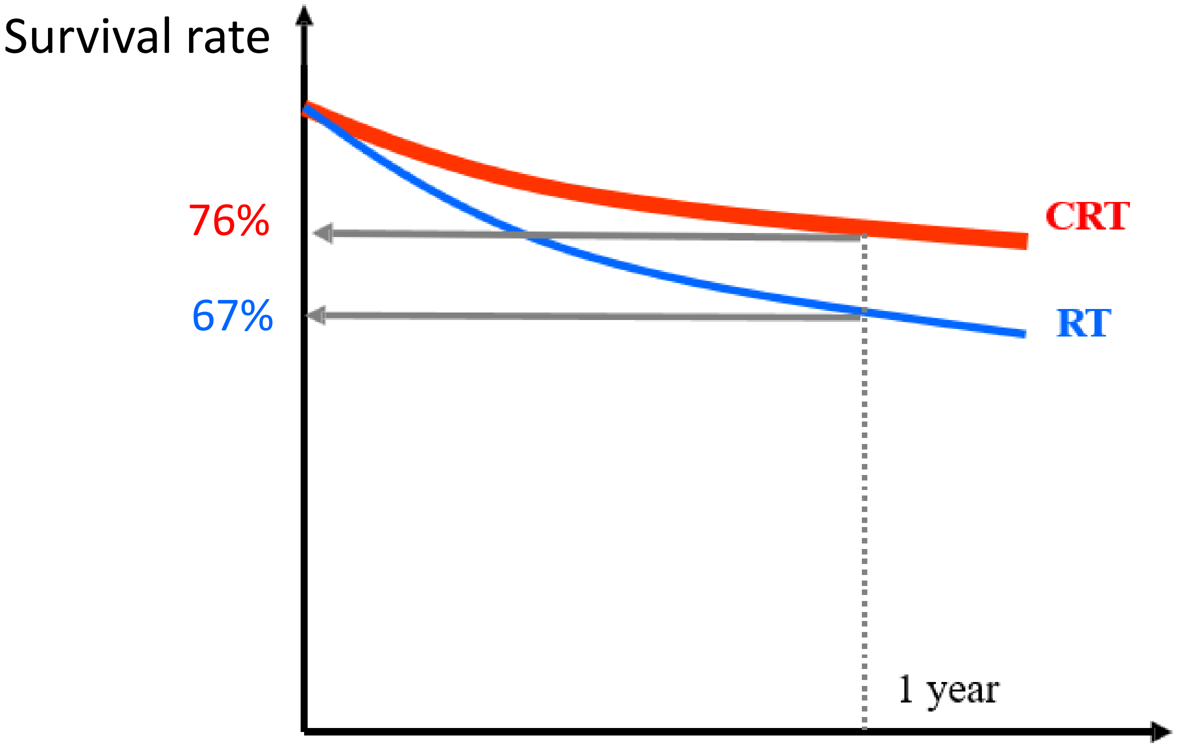
- What is randomization?

- Methods for eliminating confounding
  - Design phase innovation: randomization
  - Analysis phase innovation

Hypothetical Example

CRT: Chemoradiotherapy
RT: Radiation Therapy

- The CRT group (250 cases) had a better prognosis than the RT group (60 cases).
- Is CRT recommended for this subject?

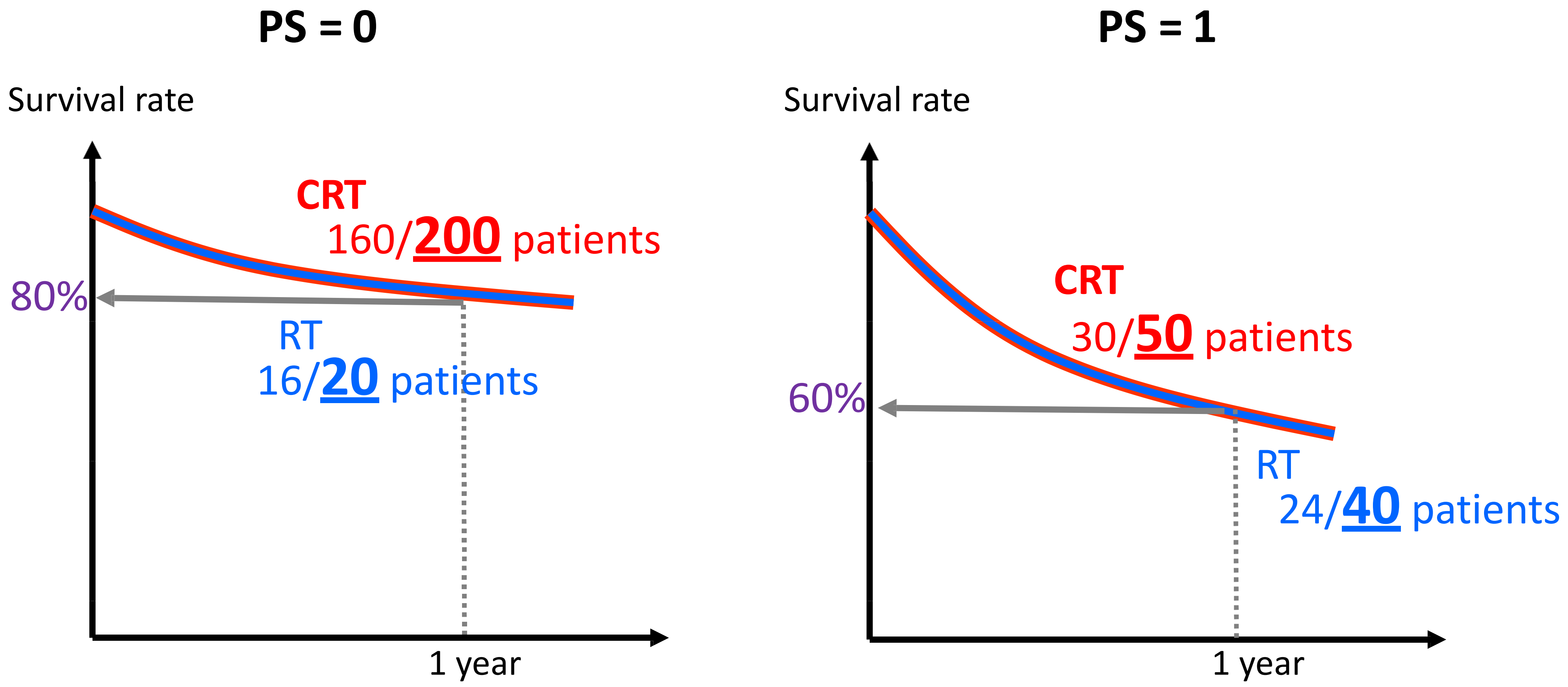


190/250 patients

40/60 patients

Note: Hypothetical example

Prognosis by PS



The prognosis for **CRT** **and** **RT** is the same regardless of PS

Note: Hypothetical example

Why Did CRT Outperform?

Treatment	PS = 0		PS = 1	Total
CRT	**200 people (80%)**	>>	50 people	250 people
RT	20 people (33.3%)	<<	**40 people**	60 people

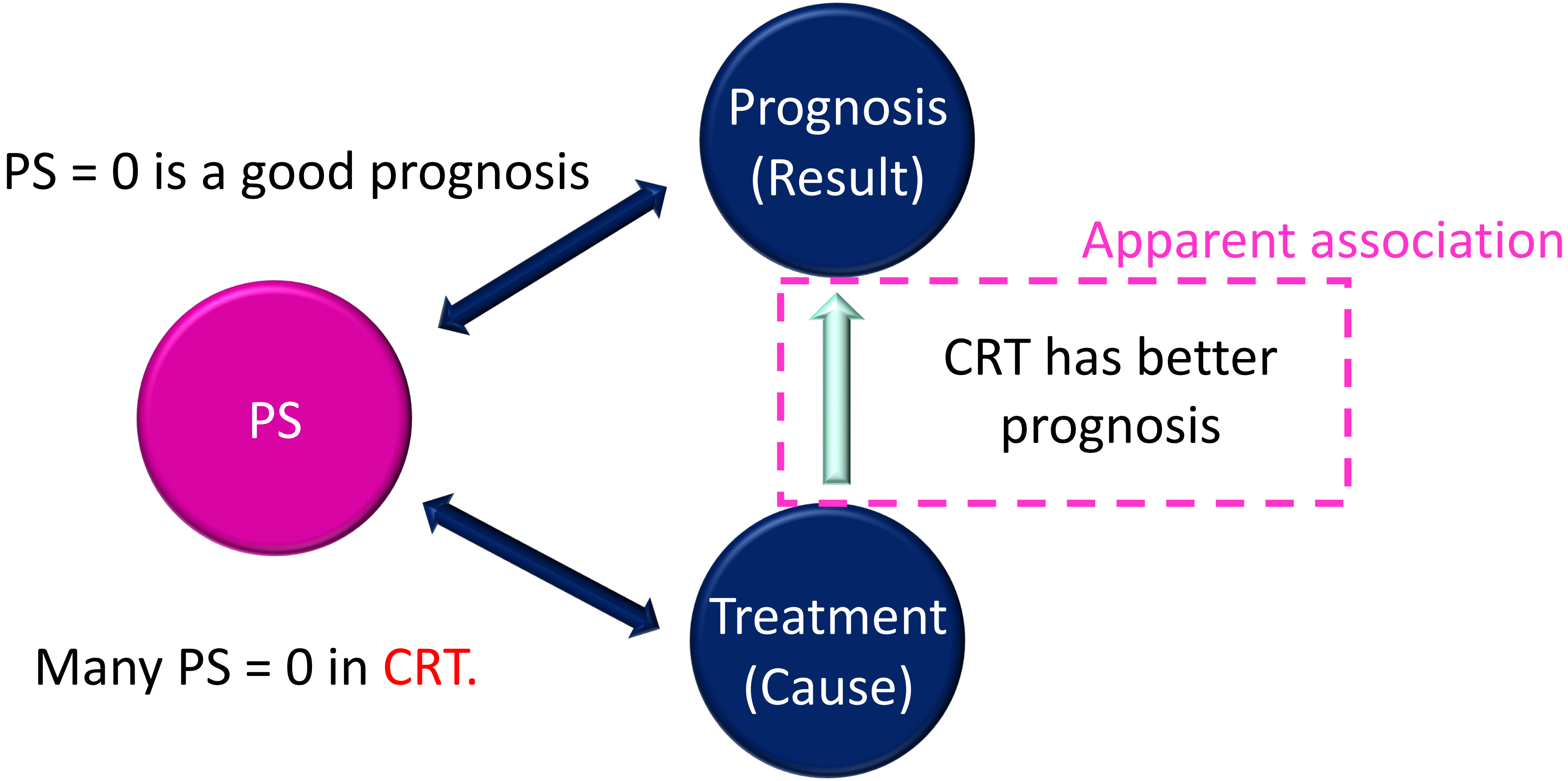
- CRT has many cases of PS = 0
- (In general) If PS = 0, the prognosis is good.
- Unless the conditions of factors affecting prognosis other than the treatment method are the same, there is no "comparison"!!

What is Confounding?

- A phenomenon in which an apparent association occurs due to a third factor related to the cause and effect.
  - Factors that cause confounding are called confounders

- Confounder for hypothetical example = PS

What is Confounding?



PS = 0 is a good prognosis

Prognosis (Result)

Apparent association

CRT has better prognosis

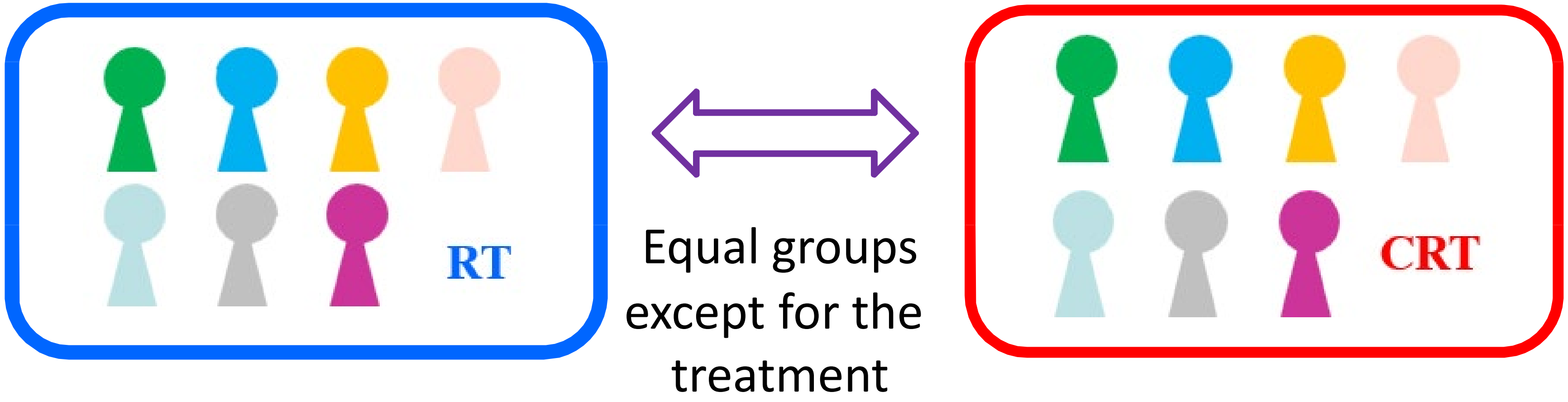
PS

Treatment (Cause)

Many PS = 0 in CRT.

How to Eliminate Confounding during the Design Phase

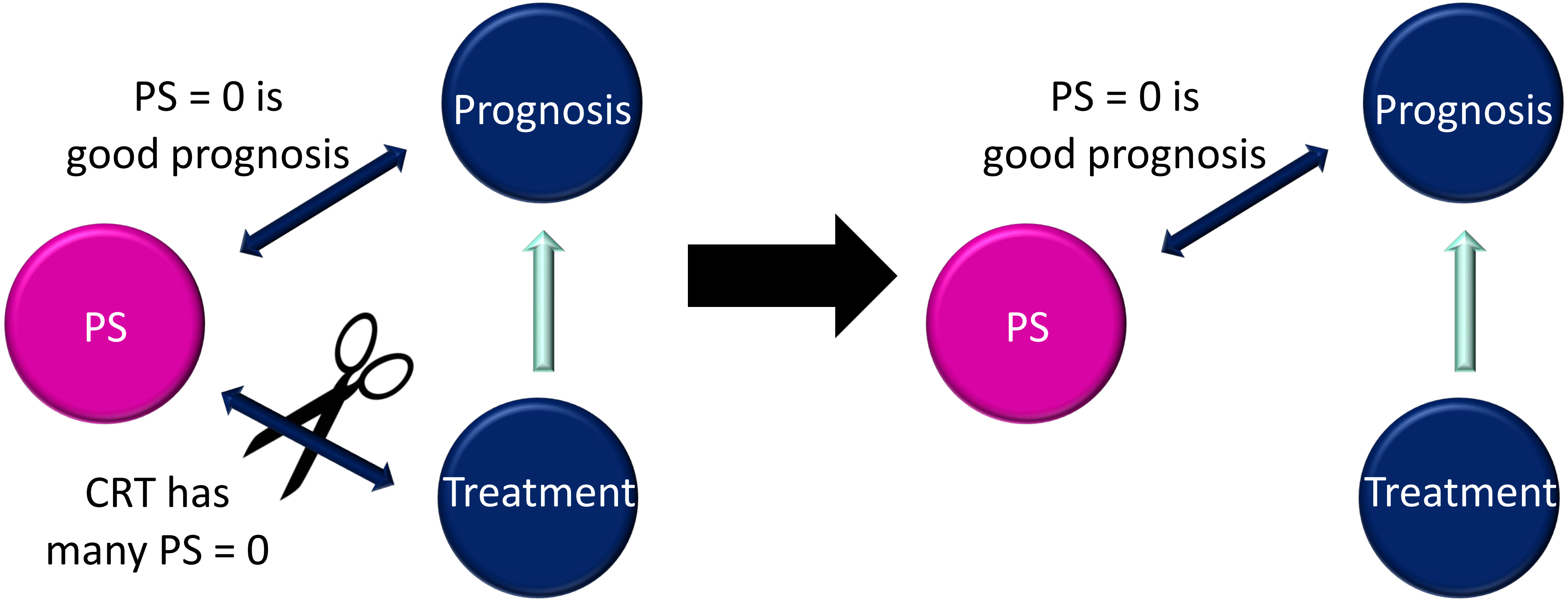
- Randomization
  - Assign patients to each treatment group based on probability, <u>independent of the physician's or patient's will.</u>
  - Equal groups, except for treatment
    →Differences in treatment if effectiveness has differences



Equal groups except for the treatment

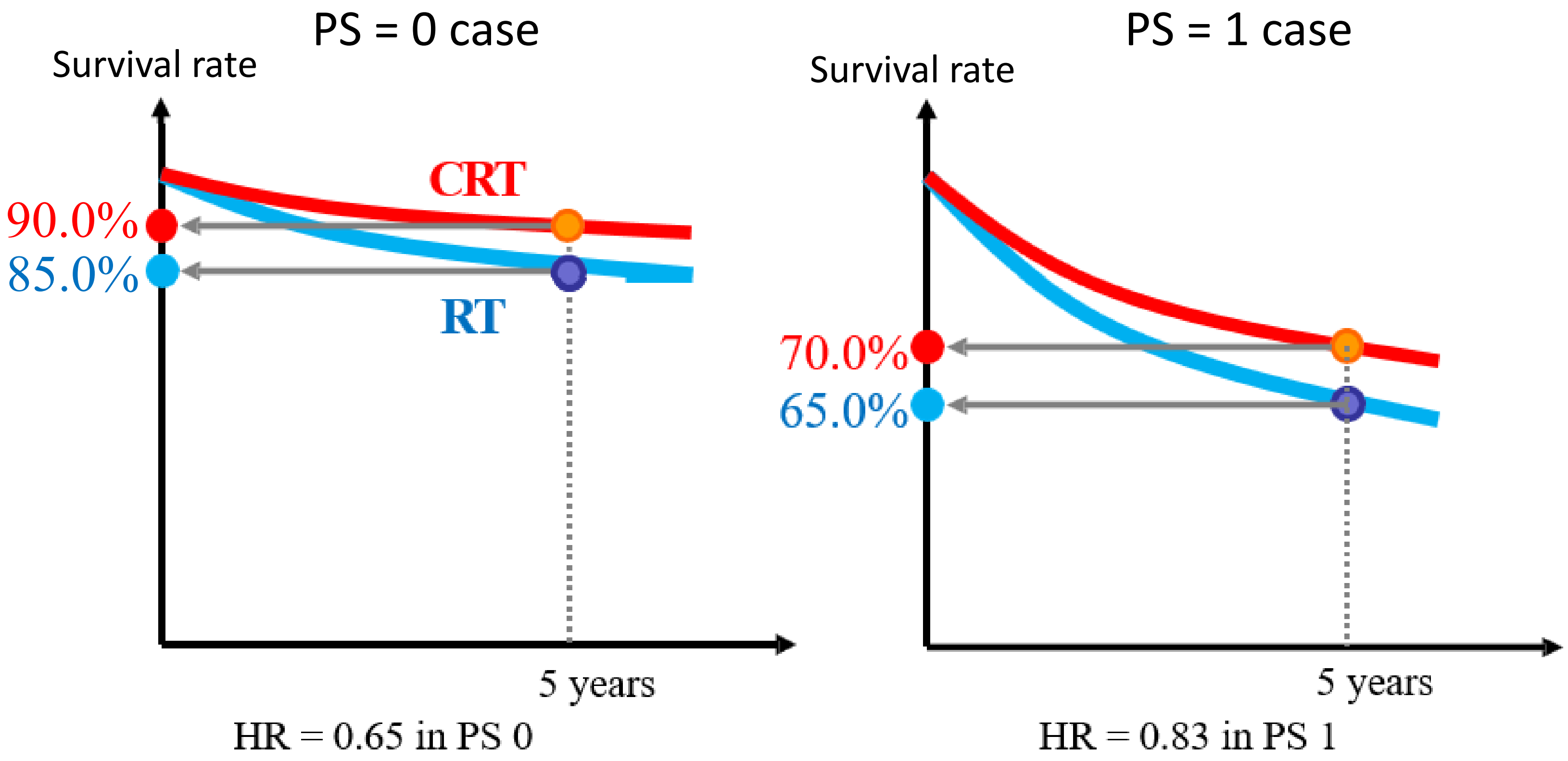
Significance of Randomization

- The association between treatment and PS can be eliminated
  - Confounding by PS is eliminated so that the relationship between treatment and prognosis can be assessed
  - Note: The relationship between PS and prognosis remains



Stratified Analysis

- Analysis of merging treatment effects under the assumption that treatment effects are common for PS 0/1



Integrated HR = 0.78

Advantages and Disadvantages of Stratified Analysis

- <u>Advantages</u>
  - The effect of treatment on the entire population can be determined
  - Fewer assumptions (compared with analysis using models)
- <u>Disadvantages</u>
  - When there are too many subgroups,

    sample size for each subgroup becomes too small
    - If there are 5 confounders, at least $2^5 = 32$ subgroups
    - Subgroup analysis can only be performed after categorization when confounders are continuous variables.
  - The magnitude of the effect of the confounders themselves is unclear
    - Unable to assess the prognostic impact of PS 1 on PS 0

How to Deal with Confounding

Design phase

- Randomization
- Matching

Analysis phase

- Subgroup analysis
- Stratified analysis
- **Multivariate analysis using models**
  - **Logistic regression, Cox regression, etc.**

Outline

- Review of the Fourth Lecture
- What Is a (Statistical) Model?
- Multivariate Analysis for Confounding Adjustment
- Notes on the Use of Regression Models
- Uses of Multivariate Analysis

Hypothetical Example

- Can the rent of a rental property be expressed using an equation?
  - Possibly related factors
    - Number of floors and size
    - Suppose rent can be expressed as a weighted sum of these

Rent = Market rent in the area + 1.5 × Number of floors + 2.5 × Square footage

- This equation does not hold for all properties.
  - It is because variabilities exist.

What is a Statistical Model?

- Mathematical model that accounts for variability (error)

Rent = Market rent in the area + 1.5 × Number of floors + 2.5 × Square footage + Error

  - The outcomes of interest are called ”response and outcome variables."
  - The variables that explain the outcomes are called "explanatory variable”, “causal variable” and “covariate."

- Most statistical models have a "linear“ structure
  - Linear equation: Expressions that can be multiplied, added, or subtracted

Response variable = $\boldsymbol{\beta_0}$ + $\boldsymbol{\beta_1}$ × Explanatory variable1 + $\boldsymbol{\beta_2}$ × Explanatory variable2 + ... + Error

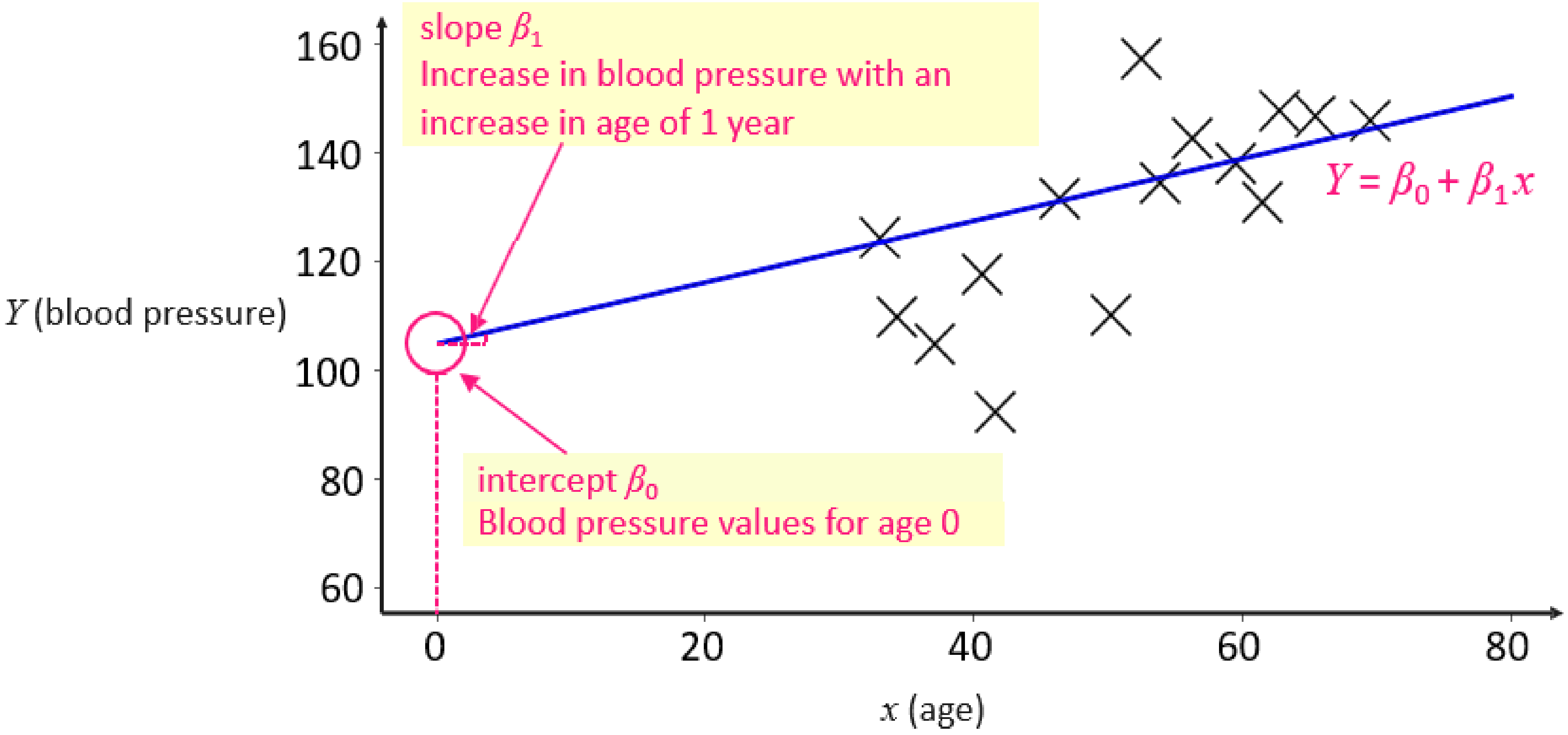
    - Bolded parts are called "parameters."
    - Model assumes "additive effects of explanatory variables"

Simple Regression Model

- $Y = \beta_0 + \beta_1 x + \text{Error}$
  - Y: response variable, x: explanatory variable (**only one**)
  - β_0: intercept, β_1: slope
- Model the relationship between blood pressure*(Y)* and age *(x)*
  - Model: blood pressure (Y) $= \beta + \ \beta_1 \times \text{Age}\ (x) + \ \text{Error}$
    - The blood pressure value shall be the systolic blood pressure.
    - A linear model of the relationship between blood pressure and age

Note: Hypothetical example

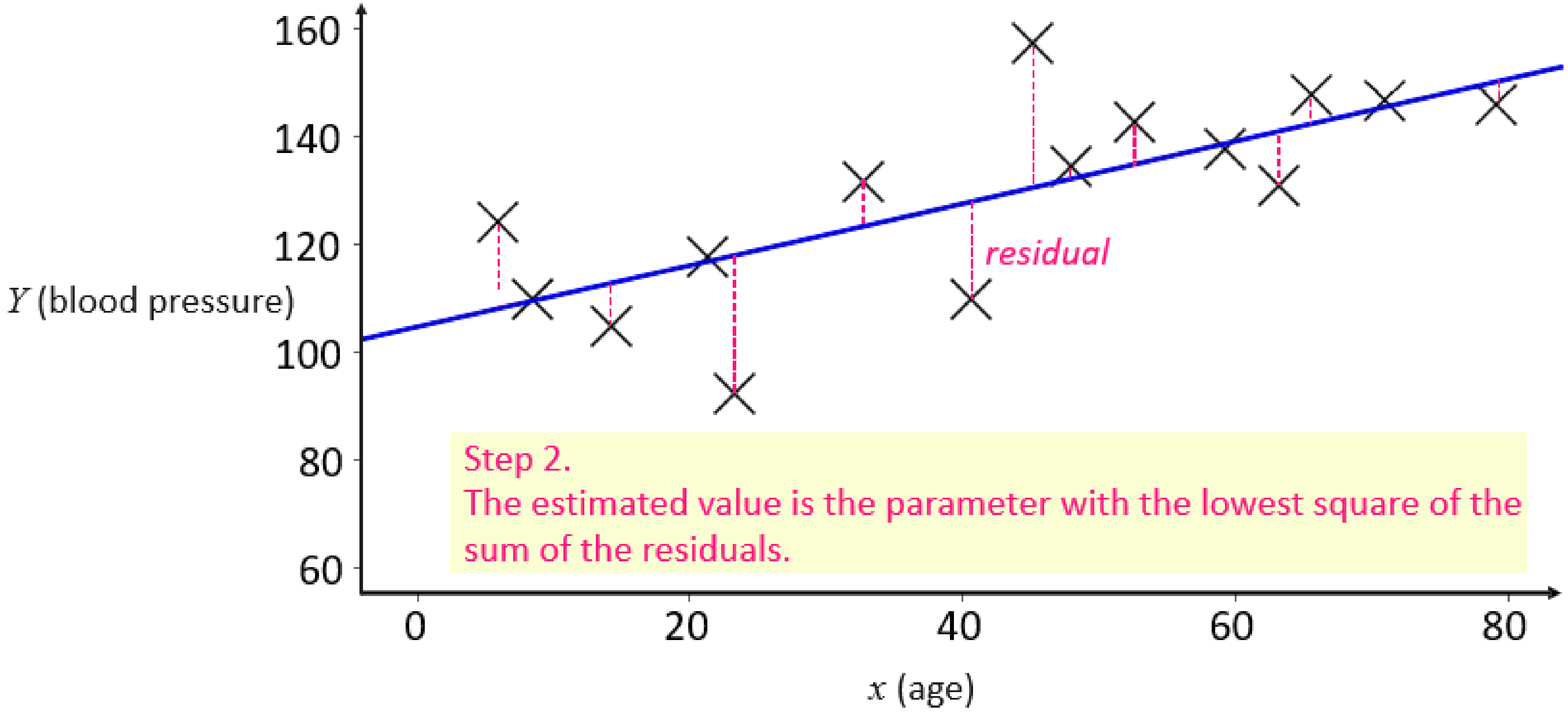
Fitting a Straight Line to the Scatter Plot (n = 15)



Note: Hypothetical example

How To Determine the Parameters?

Step 1.
Compute the sum of the two squares of the residuals (= predicted value − measured value) by entering certain values for the parameters.



Step 2.
The estimated value is the parameter with the lowest square of the sum of the residuals.

Logistic Regression Model

- Statistical models for binary variable outcomes
  - For example, consider that you are interested in response rates

- Logistic regression model (with one explanatory variable)

  p: response rate

$$\underset{\text{log odds}}{\boxed{\log\left(\frac{p}{1-p}\right)}} = \underset{\text{linear equation}}{\overset{\text{intercept}}{\beta_0} + \overset{\text{slope}}{\beta_1} x}$$

What Is an Odds Ratio?

Odds: = (number of events occurring / number of events not occurring) ratio

Therapeutic group	Responded	No response	Total
Standard	20	80	100
Test	40	60	100
Total	60	140	200

Odds for standard group $= \frac{20}{100} \quad \frac{80}{100} = \frac{20}{80} = \frac{1}{4}$

Odds for study group $= \frac{40}{100} \quad \frac{60}{100} = \frac{40}{60} = \frac{2}{3}$

$$\text{Odds ratio} = \frac{2}{3} \Big/ \frac{1}{4} = \frac{8}{3} \approx 2.67$$

The odds for the experimental group were 2.67 times the odds for the standard group.

Relationship between Logistic Regression Model and Odds

$$\log\left(\frac{p}{1 - p}\right) = \beta_0 + \beta_1 x \qquad x = \begin{cases} 0 & \text{Standard group} \\ 1 & \text{Experimental group} \end{cases}$$

$$\log\left(\frac{A}{B}\right) = \log A - \log B$$

$$\log\left(\text{experimental group odds}\right) = \beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1$$

$$\log\left(\text{standard group odds}\right) = \beta_0 + \beta_1 \times 0 = \beta_0$$

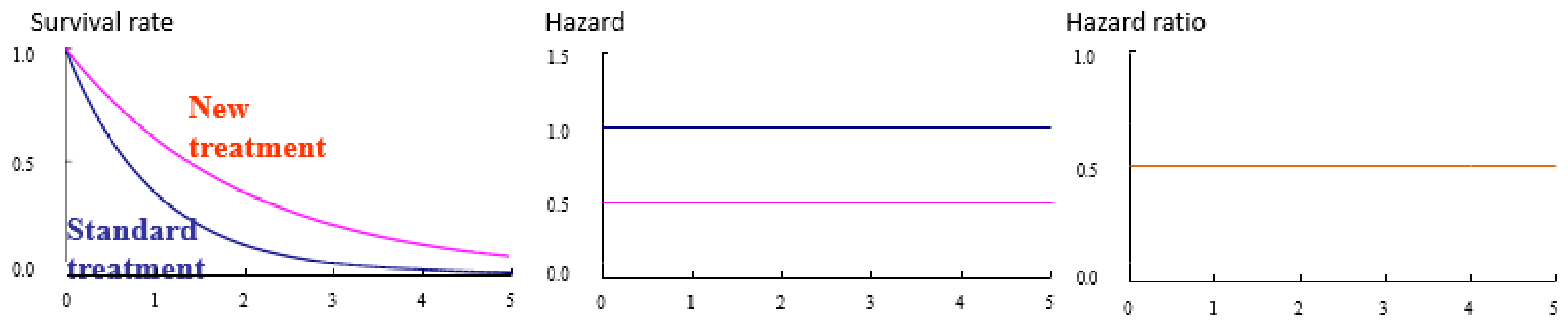
$$\text{Log odds ratio} = \log \frac{\text{Experimental group odds}}{\text{Standard group odds}} = \text{log (experimental group odds) - log (standard group odds)} = \beta_1$$

$$\text{Odds ratio} = \exp(\beta_1)$$

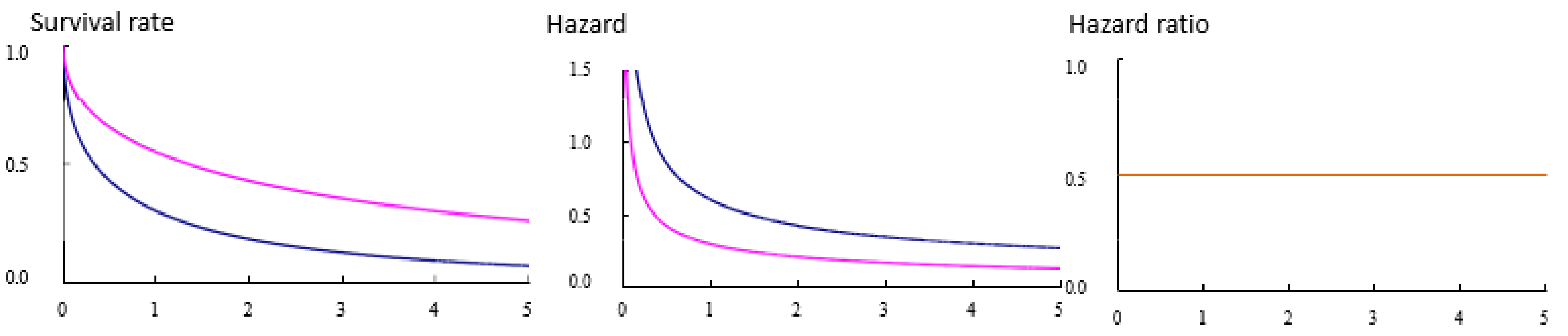
***exp (β) means e$^{\beta}$**

Review: Proportional Hazards

Exponential distribution: The hazard is constant throughout the time point. HR is constant if both groups have an exponential distribution.



Weibull distribution: The hazard varies across time points. HR is constant for Weibull distributions of the same shape for both groups



Cox's Proportional Hazards Model

- Statistical model for survival time outcomes
  - For example, consider that you are interested in the overall survival

- Cox’s proportional hazards model (one explanatory variable case)

$_{h0}(t)$: *baseline hazard* (expressed as a function of time)

$$\boxed{h(t)} = h_0(t) \text{ x } \exp(\beta x)$$

Hazard function with death as an event

$$\log \left[\frac{h(t\ )}{h_0(t)}\right] = x$$ The structure is the same as the logistic regression model

The Relationship between Cox Proportional Hazards Model and Hazard Ratio

$$h(t) = h\ (t) \times \exp(x)$$

$$x = \begin{cases} 0 & \text{Standard group} \\ 1 & \text{Experimental group} \end{cases}$$

$$\text{Hazard of the standard group} = h_0(t) \times \exp(\beta \times 0) = h_0(t)$$

$$\text{Hazard of the experimental group} = h_0(t) \times \exp(\beta \times 1) = h_0(t) \times \exp(\beta)$$

$$\text{Hazard ratio} = \frac{\text{Hazard of the experimental group}}{\text{Hazard of the standard group}} = \frac{\cancel{h_0(t)} \times \exp(\beta)}{\cancel{h_0(t)}} = \exp(\beta)$$

Statistical Model – Summary

- "Linear equation" to account for variability

  Response variable $= \beta_0 + \beta_1 \times$ Explanatory variable 1 $+ \beta_2 \times$ Explanatory variable 2 $+ \ldots +$ Error

  - For univariate models, we assume a linear relationship (intercept: β_0, slope: β_1)

- Logistic Regression Model
  - Statistical models for binary outcomes
  - Odds ratio can be estimated from the estimated parameters.

- Cox Proportional Hazards Model
  - Statistical model for survival time outcomes
    - Model assuming proportional hazard property (hazard ratio is constant regardless of the time point)
  - Hazard ratio can be estimated from the estimated parameters.

Outline

- Review of the Fourth Lecture
- What Is a (Statistical) Model?
- Multivariate Analysis for Confounding Adjustment
- Notes on the Use of Regression Models
- Uses of Multivariate Analysis

Hypothetical Example

	5-year survival	Death	Total
Radiation	47 (**78.3%**)	13	60
Surgery	185 (74.0%)	65	250

	5-year survival	Death	Total
Radiation	34 (85%)	6	40
Surgery	45 (**90%**)	5	50

PS 0 cases

	5-year survival	Death	Total
Radiation	13 (65%)	7	20
Surgery	140 (**70%**)	60	200

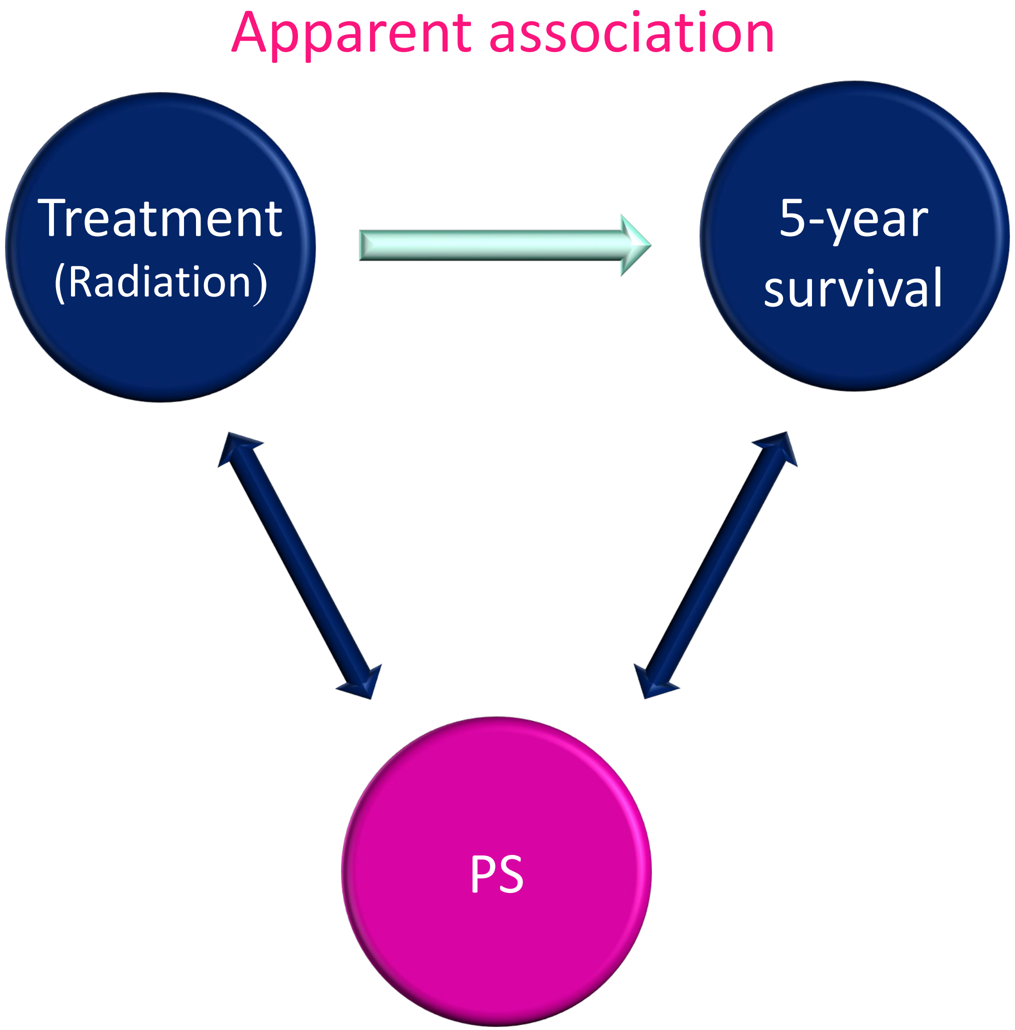
PS 1 cases

Re-Review of Confounding

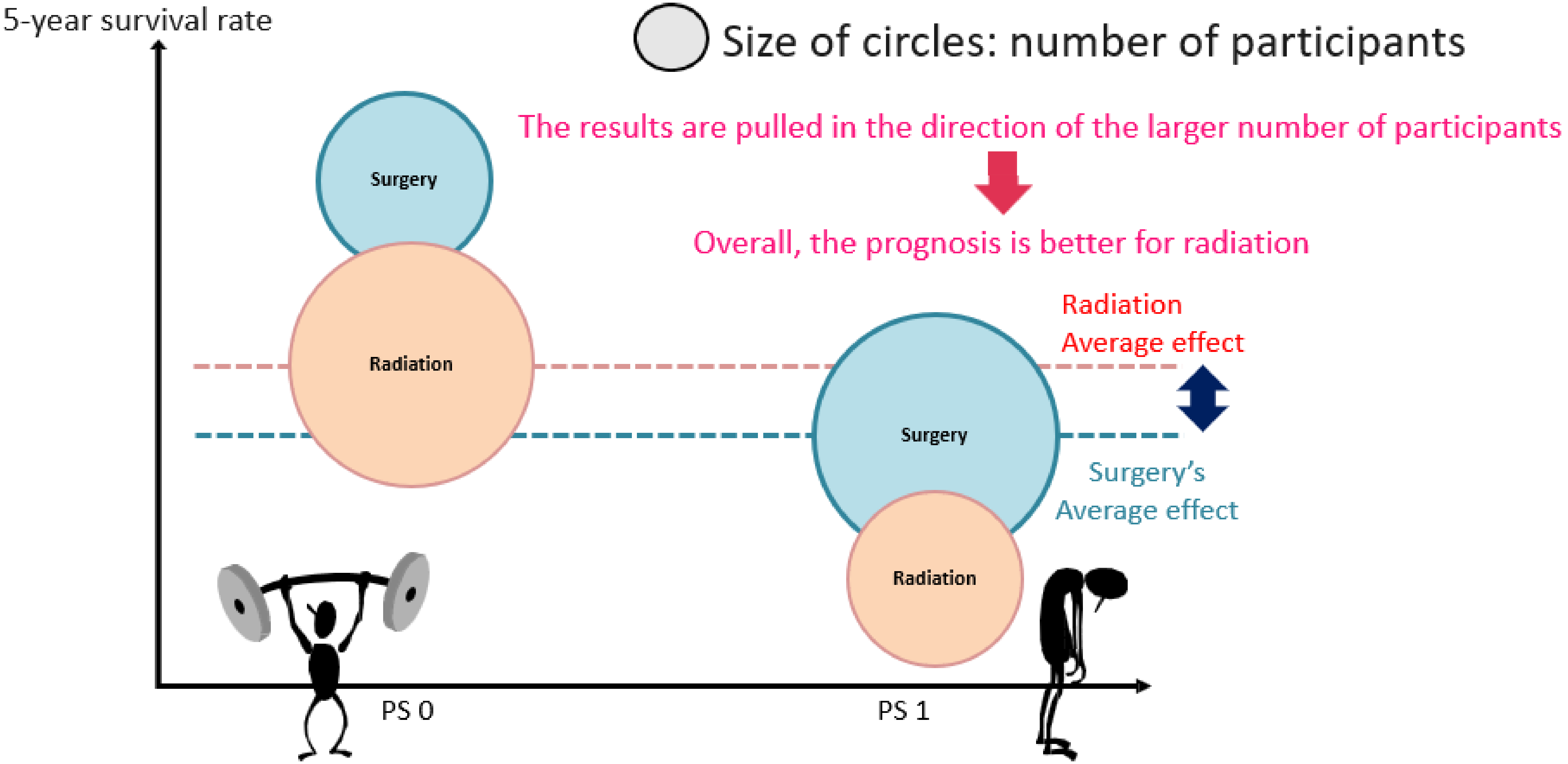
<u>Cause of confounding</u>

Because the component ratios of confounders (PS) associated with prognosis were biased among the groups

- More PS 0 in radiation
- Good PS = good prognosis



Diagram



The results are pulled in the direction of the larger number of participants

Overall, the prognosis is better for radiation

Fitting a Logistic Regression Model

※*p*: 5-year survival rate, treatment: 1 for radiation, 0 for surgery

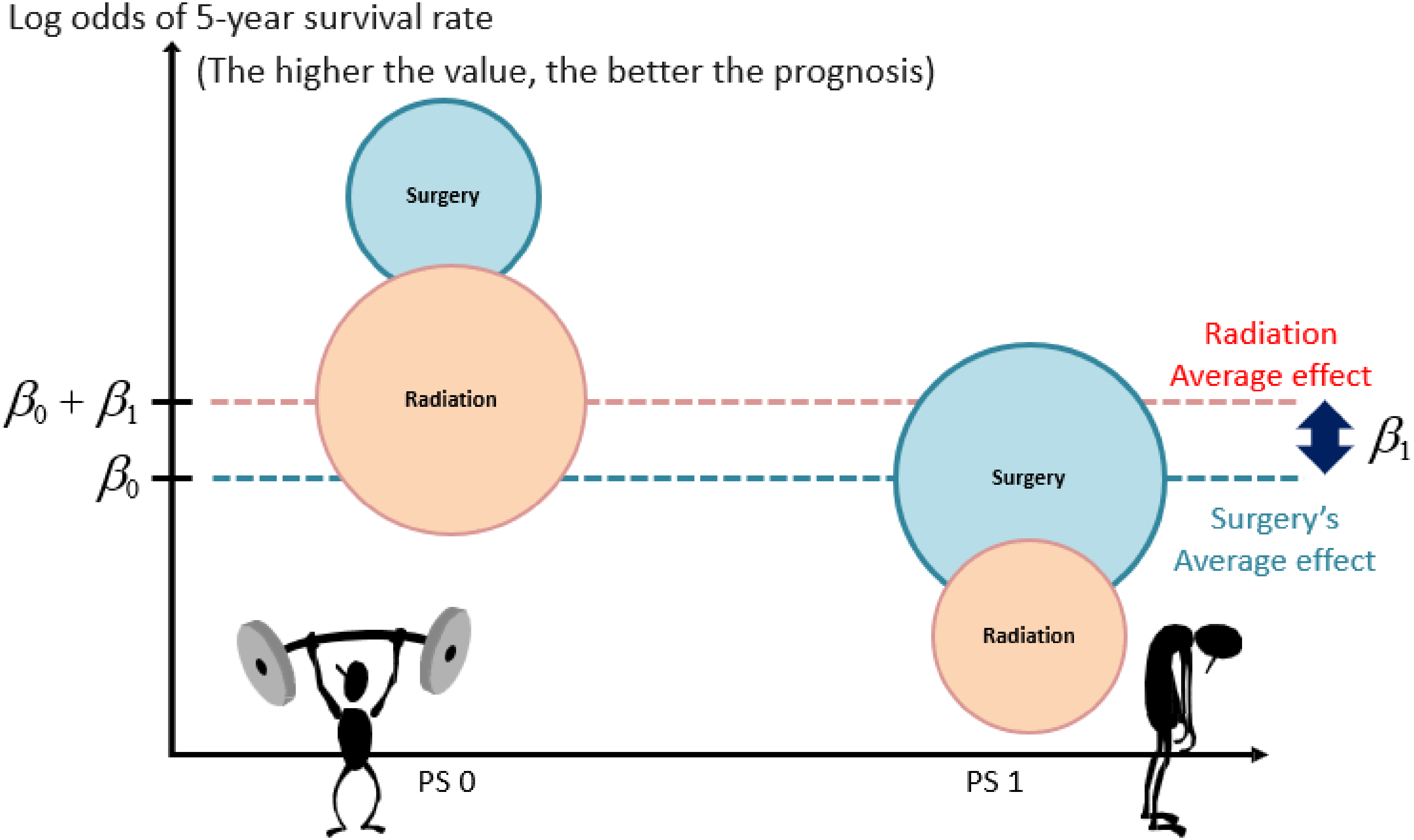
- Univariate model

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \times \text{treatment}$$

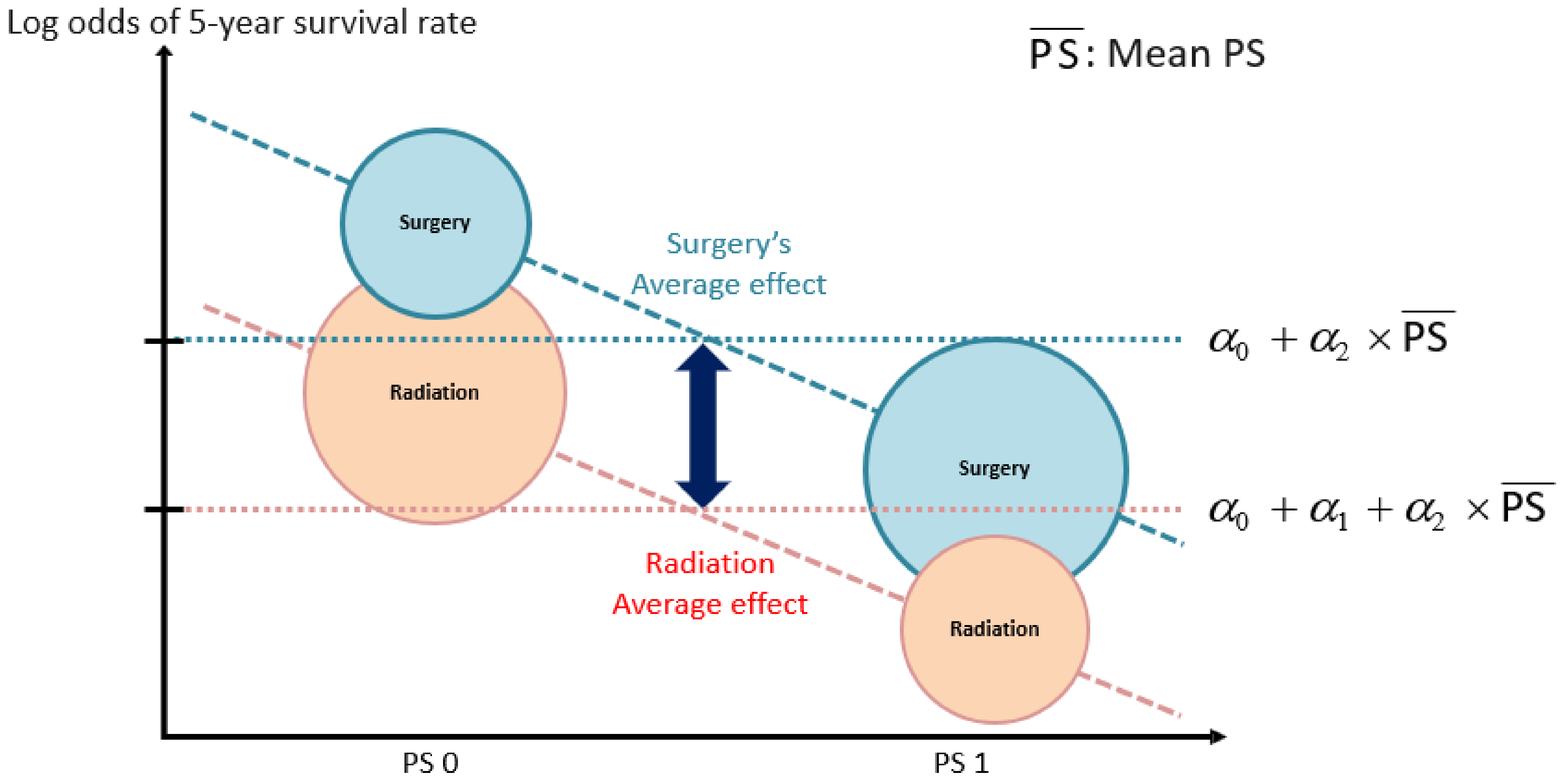
- Multivariate (bivariate) model

$$\log\left(\frac{p}{1-p}\right) = \alpha_0 + \alpha_1 \times \text{treatment} + \alpha_2 \times \text{PS}$$

Univariate model $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \times$ treatment



Multivariate Model $\log\left(\frac{p}{1-p}\right) = \alpha_0 + \alpha_1 \times \text{treatment} + \alpha_2 \times \text{PS}$



Difference Between β_1 and α_1

- β_1: Log odds ratio <u>not adjusted by</u> PS
  - Interpretation: "The radiation group has a higher 5-year survival rate."
  - Confounding distorts the truth.

- α_1: Log odds ratio <u>adjusted by</u> PS
  - Interpretation: "The surgery group has a higher 5-year survival rate."
  - The effects of confounding can be ruled out.
  - The effect of exposure when confounders are included in the explanatory variables is called an "adjusted effect."

What These Results Tell Us

- If you are interested in the relationship between exposure and outcome

  (If you are interested in inferring the association of cause and outcome [causal inference])

  - The presence of confounding distorts the effects of exposure.
    - Except when exposures are completely randomly assigned
    - Nonrandomized exposures (e.g., smoking, drinking, and eating) need to be adjusted for confounding using appropriate methods.

- **Including confounders in multivariate models allows appropriate estimation of exposure effects.**

Confounder Selection

- To properly infer causality, confounders need to be fully adjusted
  - How can we select confounders?
    - List variables that match the <u>three confounder conditions</u>
    - (1) Correlated with outcome
    - (2) Correlated with exposure
    - (3) Not an intermediate variable (a variable between exposure and outcome)

- If the accuracy of parameter estimation appears to drop, **select variables**

Basic Principles for Variable Selection

- Data alone does not lead to an optimal model.
- Information on known risk/prognostic factors should be used.
- The objective is to estimate the effect of exposure.
  - – Choice of variable should be examined based on the perspective of "variables other than exposure are used for confounding adjustment."

Deciding Whether to Add a Certain Confounder C to the Model

- Change in estimate standard
  - If the effect of exposure is unchanged before and after the inclusion of C, it is not worth complicating the model by adding C.
  - It is time-consuming because of the need to consider all possible combinations.

- Mechanical algorithms (backward/forward procedure, etc.)
  - Decide based on the strength of the association

    between C and outcome when including or excluding C
  - Use backward procedure or force the entry of known prognostic factors so that confounders with low association with the results are not excluded.

Points to Remember When Making Variable Selections

- Significant in univariate or known prognostic factors are important candidates for confounders, but **they do not necessarily need to be adjusted.**
- The constructed model does not necessarily have a correct answer.
  - If confounding is sufficiently adjusted, there is no problem

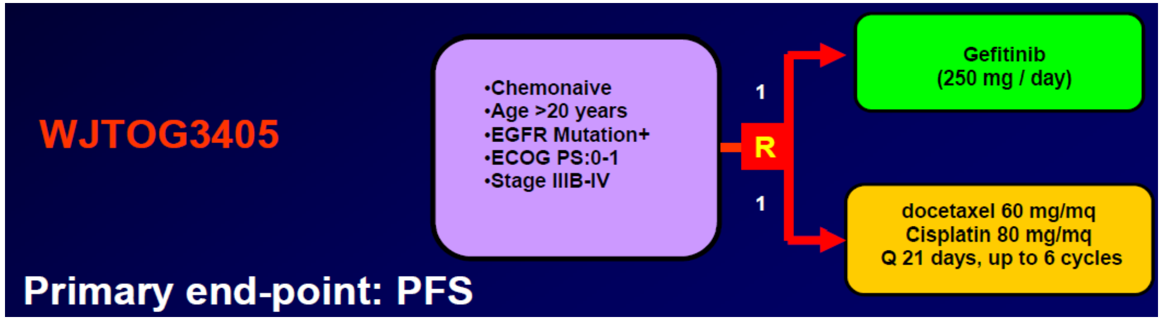
Advantages and Disadvantages of Regression Models

- Advantages
  - Adjustment is possible even if the number of confounders increases
    - However, if the number of participants or events is small,

      the estimation accuracy will drop if the number of confounders to be adjusted is large.

  The impact of confounders on outcomes can be evaluated
    - Because confounding adjustment is the first priority, evaluation of the impact of confounders should be considered secondary.
  - There is no need to assume that "treatment effects are common across strata."
    - By using a statistical model that is not a linear model assuming additive effects, it is also possible to evaluate the strength of interactions.
- Disadvantages
  - It relies on stronger assumptions than stratified analysis (discussed later)

Example: WJTOG3405 Test



WJTOG3405

- Chemonaive
- Age >20 years
- EGFR Mutation+
- ECOG PS:0-1
- Stage IIIB-IV

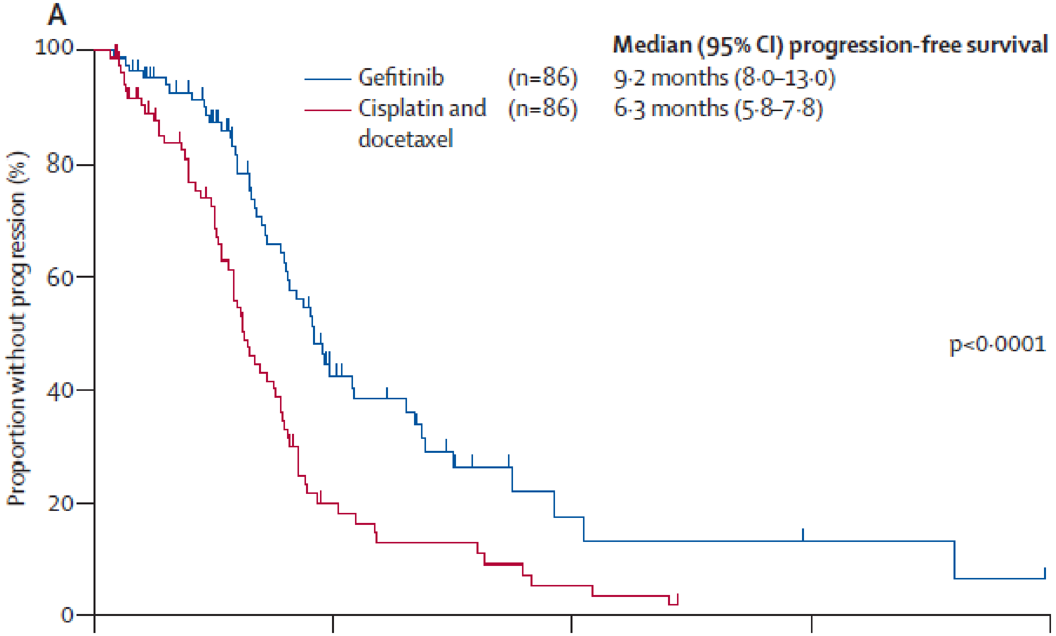
R

1 Gefitinib (250 mg / day)

1 docetaxel 60 mg/mq Cisplatin 80 mg/mq Q 21 days, up to 6 cycles

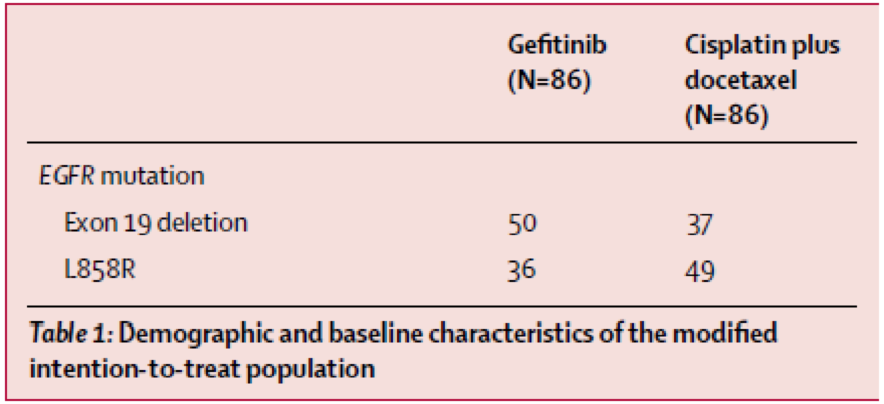
Primary end-point: PFS

Mitsudomi T, et al. Lancet Oncol. 2010.



Number at risk									
Gefitinib	86	63	22	11	4	3	2	2	0
Cisplatin and docetaxel	86	49	11	7	3	0	0	0	0

Example: WJTOG3405 Test



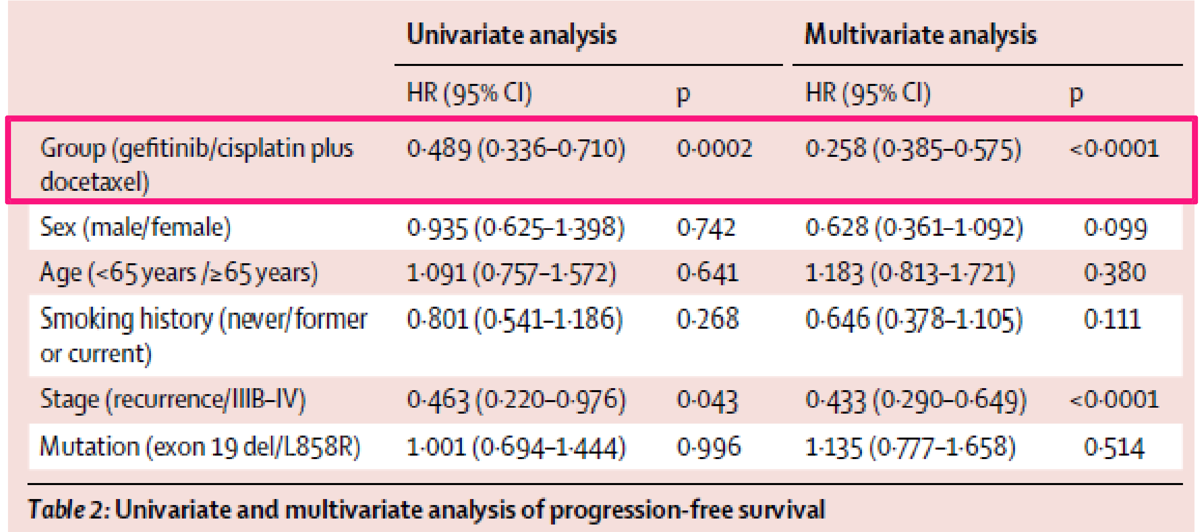
	Gefitinib (N=86)	Cisplatin plus docetaxel (N=86)
EGFR mutation		
Exon 19 deletion	50	37
L858R	36	49

Table 1: **Demographic and baseline characteristics of the modified intention-to-treat population**

Mitsudomi et al. Lancet Oncol. 2010. Excerpt from Table 1

- Because it was a randomized controlled trial of a small number of cases,
  EGFR gene mutation types were biased in both groups.
- Multivariate analysis adjusted for this effect

Example: WJTOG3405 Test



	Univariate analysis		Multivariate analysis	
	HR (95% CI)	p	HR (95% CI)	p
Group (gefitinib/cisplatin plus docetaxel)	0·489 (0·336–0·710)	0·0002	0·258 (0·385–0·575)	<0·0001
Sex (male/female)	0·935 (0·625–1·398)	0·742	0·628 (0·361–1·092)	0·099
Age (<65 years /≥65 years)	1·091 (0·757–1·572)	0·641	1·183 (0·813–1·721)	0·380
Smoking history (never/former or current)	0·801 (0·541–1·186)	0·268	0·646 (0·378–1·105)	0·111
Stage (recurrence/IIIB–IV)	0·463 (0·220–0·976)	0·043	0·433 (0·290–0·649)	<0·0001
Mutation (exon 19 del/L858R)	1·001 (0·694–1·444)	0·996	1·135 (0·777–1·658)	0·514

***Table 2:* Univariate and multivariate analysis of progression-free survival**

←Hazard ratio adjusted for confounding
0.258

Mitsudomi T, et al. Lancet Oncol. 2010.

$$\text{Hazard} = {}_{h0}(t) \times \boxed{\text{gefitinib's effects}}$$

←primary interest

confounders

× gender effects

× ● ● ● ●

× Effects of types of EGFR mutations

Outline

- Review of the Fourth Lecture
- What Is a (Statistical) Model?
- Multivariate Analysis for Confounding Adjustment
- Notes on the Use of Regression Models
- Uses of Multivariate Analysis

Drawbacks of Regression Model: Strong Assumptions

- Assumptions to perform causal inference
  - Unmeasured or unknown confounders do not exist
    - This assumption holds for large randomized controlled trials

- Mathematical assumptions
  - A linear trend exists between the factors included in the model and the results.
  - Assumptions for the model: proportional hazard property (Cox regression), etc.
  - Not an over-fitting model to the data
  - Assumptions for proper parameter estimation
    - Multicollinearity does not exist

      (Variables showing strong correlation are not included in the model)
    - The number of targets and events is sufficient for the number of variables to be included in the model.

Overfitting

- Contribution ratio R^2 (Coefficient of determination)
  - An index of the fit of the regression model to the data
    - The closer to 1, the better the fit.
  - Increasing the number of explanatory variables increases the contribution rate.
    - Even if factors that are completely unrelated to the outcome are included, the contribution rate increases.
  - If a complex model (such as a model with a quadratic term) is used, the contribution rate can approach 1.
- Overfitting models
  - Models with a higher-than-necessary fit to the data in hand
  - Low extrapolation and cannot be generalized (details will be explained in the 6th lecture)

Multicollinearity

- A phenomenon that parameter estimation becomes unstable due to variables showing strong correlation

- Inevitable unless we increase the number of targets and events
  - Variables that are clearly highly correlated are of greater interest or only those aspects that are easy to interpret should be included in the model
    - e.g.: BMI and weight/height, stage of disease and T/N factor

Notes on Explanatory Variables

- How much information is needed per variable?
- Should it be a continuous variable or a category?
- What if there are measurement errors or gaps?

How much information is needed per variable?

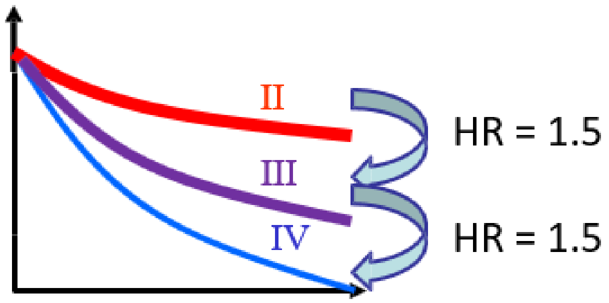
- It is said that there are more than 10–20 cases.
  - 10–20 events for Cox regression

- Problems can occur even with a sufficient amount of information
  - Example: Is a certain genetic variant associated with a specific response rate?
    - Cases where even an estimate is not possible
      - 0 mutations in 1000 cases of data
    - Cases of unstable estimation
      - There is a layer consisting of a combination of confounders to be adjusted, in which the number of cases is minimal (e.g., less than 5 cases).

Continuous? Categorical?

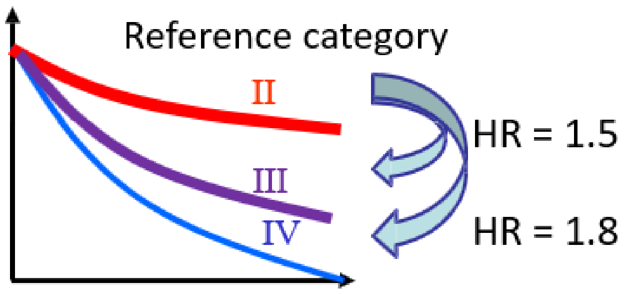
Example: Should the clinical stage be treated as a continuous or categorical variable?

- Continuous variable: estimated hazard ratio is one



HR of III against II and HR of IV against II:
an assumption that these are common
→ Estimate HR for an increase of 1 unit

- Categorical variables: estimating multiple hazard ratios



HR of III against II and HR of IV against II:
estimating these separately
→ Estimating HRs for a reference category

When You Cannot Decide...

- When a clinically reasonable cutoff exists
  - Decide by "Which treatment effect do you want to know about?" (clinical interpretation first)
    - "HR by increments of one year of age" vs. "HR by older adults or not"
    - Categorize using an easily interpretable cutoff (or a reference value for test values)
- When no clinically appropriate cutoff exists
  - It can also be decided depending on the data (e.g., median)
  - No need to force categorization
- Beware of the information lost due to continuous → categorical
  - When there is a tendency to be non-linear, e.g., the association between BMI and cancer deaths
    - A group with too high or too low BMI is at a higher risk of death.
    - It is not necessarily appropriate to separate BMI by the reference value (25)

Summary

- What is a Statistical Model?
  - It is expressed as a linear equation with error tolerance (in general).
  - The odds ratio from the logistic regression model and

    hazard ratio from the Cox regression model can be estimated.

- Multivariate analysis for adjusting confounding
  - Use when interested in causal inference
  - Fit a statistical model consisting of confounders except for exposure
  - Estimating the association between exposure and cause by adjusting for imbalances in the amount of information due to confounding under various assumptions