ATLAS
ASIAN CLINICAL TRIALS NETWORK FOR CANCERS PROJECT

National Cancer Center Japan

Multivariate Analysis 2

National Cancer Center, Japan

Biostatistics Division

Shogo Nomura

JCOG
Japan Clinical Oncology Group

Outline

- Review of the Fifth Lecture
- Multivariate Analysis for Prediction
- Necessity of Validation

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 × area + Error
  - The outcomes of interest are called "response and outcome variables."
  - The variables that explain the outcomes are called
  - "explanatory variables, causal variables, and covariates."

■ Most statistical models have a "linear equation" structure

  - Linear equations: Equations represented by addition and subtraction of multiplication terms.
- Response variable = $\boldsymbol{\beta_0}$ + $\boldsymbol{\beta_1}$ × Explanatory variable 1 + $\boldsymbol{\beta_2}$ × Explanatory variable 2 + ... + Error
    - Bolded parts are called "parameters."
    - Model assumes "additive effects of explanatory variables"

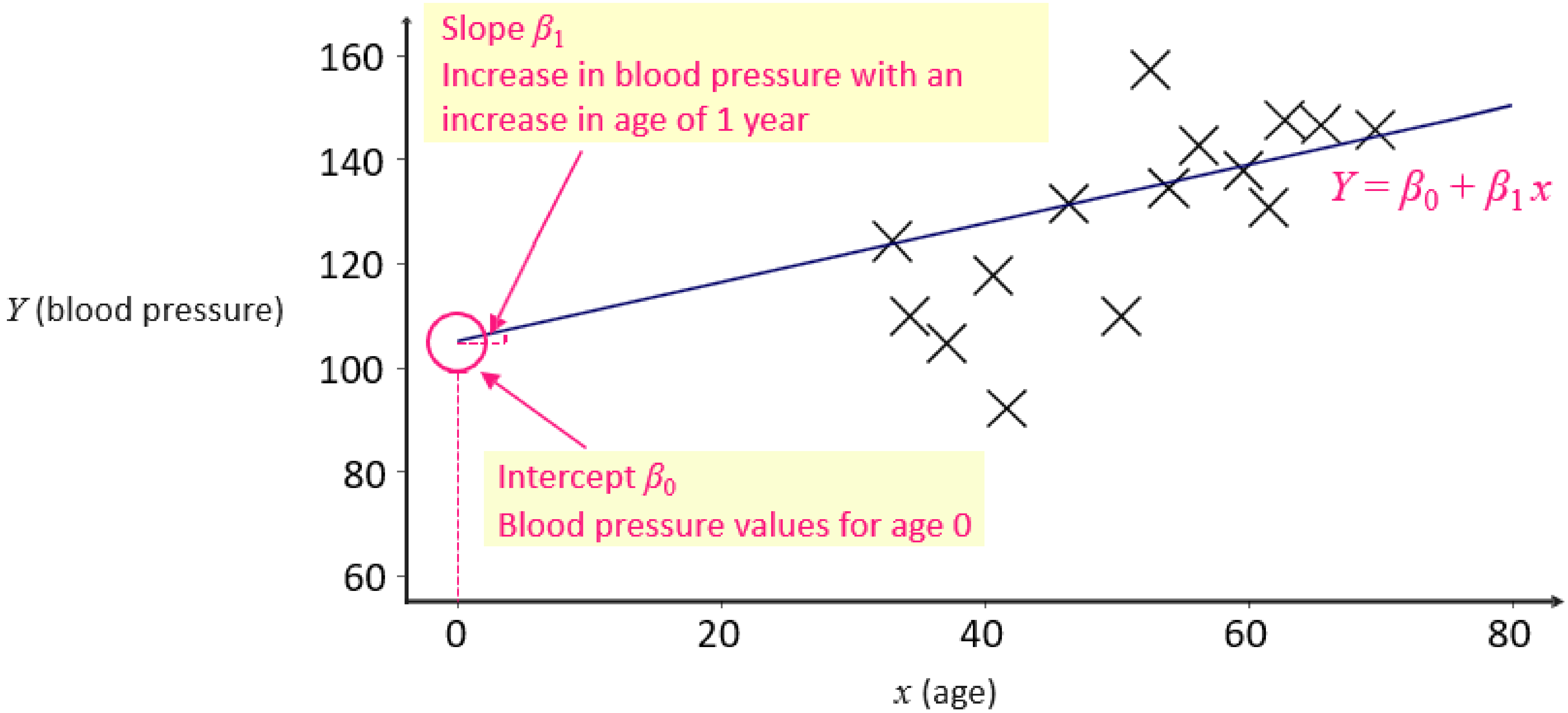
Simple Regression Model

- $Y = \beta_0 + \beta_1 x + \text{Error}$
  - 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) $= \beta_0 + \beta_1 \times \text{Age}\,(x) + \text{Error}$
    - The blood pressure value is the systolic blood pressure
    - A linear model of the relationship between blood pressure and age

Note: Hypothetical example

Fit a straight line to the scatter plot (n = 15)



Slope β_1
Increase in blood pressure with an increase in age of 1 year

Intercept β_0
Blood pressure values for age 0

Statistical Model: Summary

- "Linear equation" to account for variability
  - Response variable = $\boldsymbol{\beta_0}$ + $\boldsymbol{\beta_1}$ × Explanatory variable1 + $\boldsymbol{\beta_2}$ × Explanatory variable2 + ... + Error
  - For univariate models, we assume a linear relationship (intercept: $\boldsymbol{\beta_0}$, slope: $\boldsymbol{\beta_1}$)

- Logistic regression model

$$\log\left[\frac{p}{1-p}\right] = \beta_0 + \beta_1 x$$

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

- Cox proportional hazards model

$$h(t) = h_0(t) \times \exp(\beta x)$$

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

Outline

- Review of the Fifth Lecture
- Multivariate Analysis for Prediction
- Necessity of Validation

Royal Marsden Hospital Index

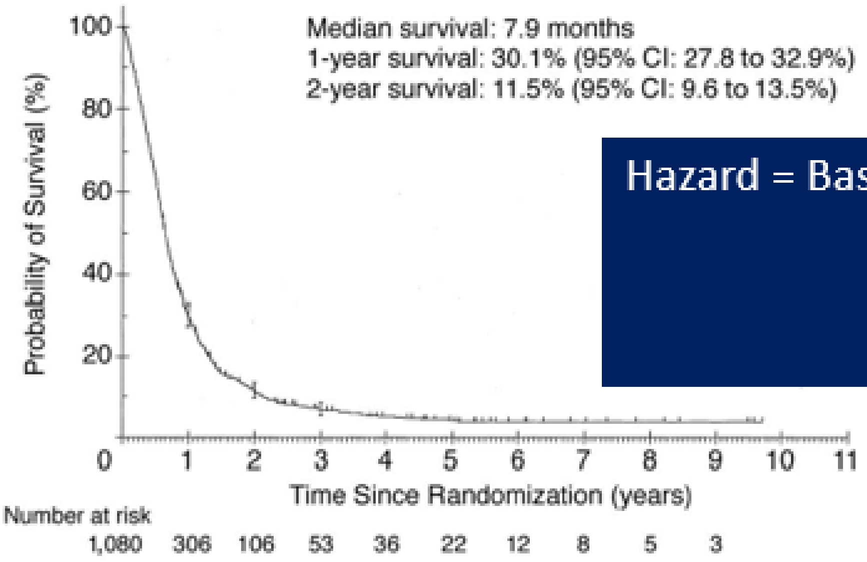
Chau et al. J Clin Oncol. 2004.

- Exploration of Prognostic Factors in Gastric-/Esophageal Cancer
  - Integrated analysis of three studies conducted between 1992 and 2001

- The purpose was to explore prognostic factors
  - *To **identify baseline prognostic factors** and (to) assess whether pretreatment quality of life (QoL) predicts survival in patients with locally advanced or metastatic esophago-gastric cancer.*

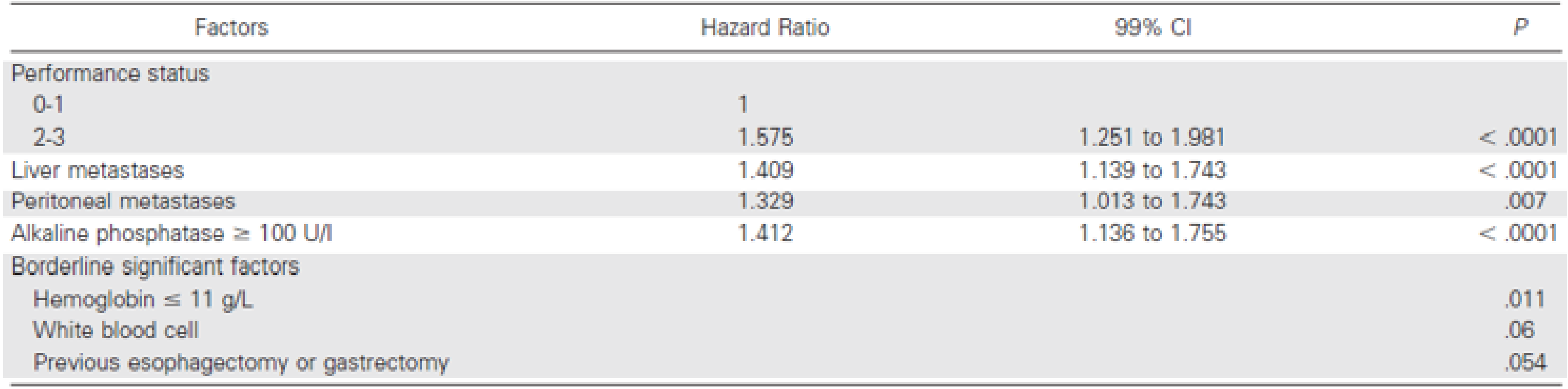
Royal Marsden Hospital Index

Chau et al. J Clin Oncol. 2004.



$$
\begin{aligned}
\text{Hazard} = \text{Base hazard} \times \exp\{0.7149 &\times (\text{PS: 0-1 / 2}) \\
+ 0.3873 &\times (\text{liver meta}) \\
+ 0.4357 &\times (\text{peritoneal meta}) \\
+ 0.1080 &\times (\text{ALP})\}
\end{aligned}
$$

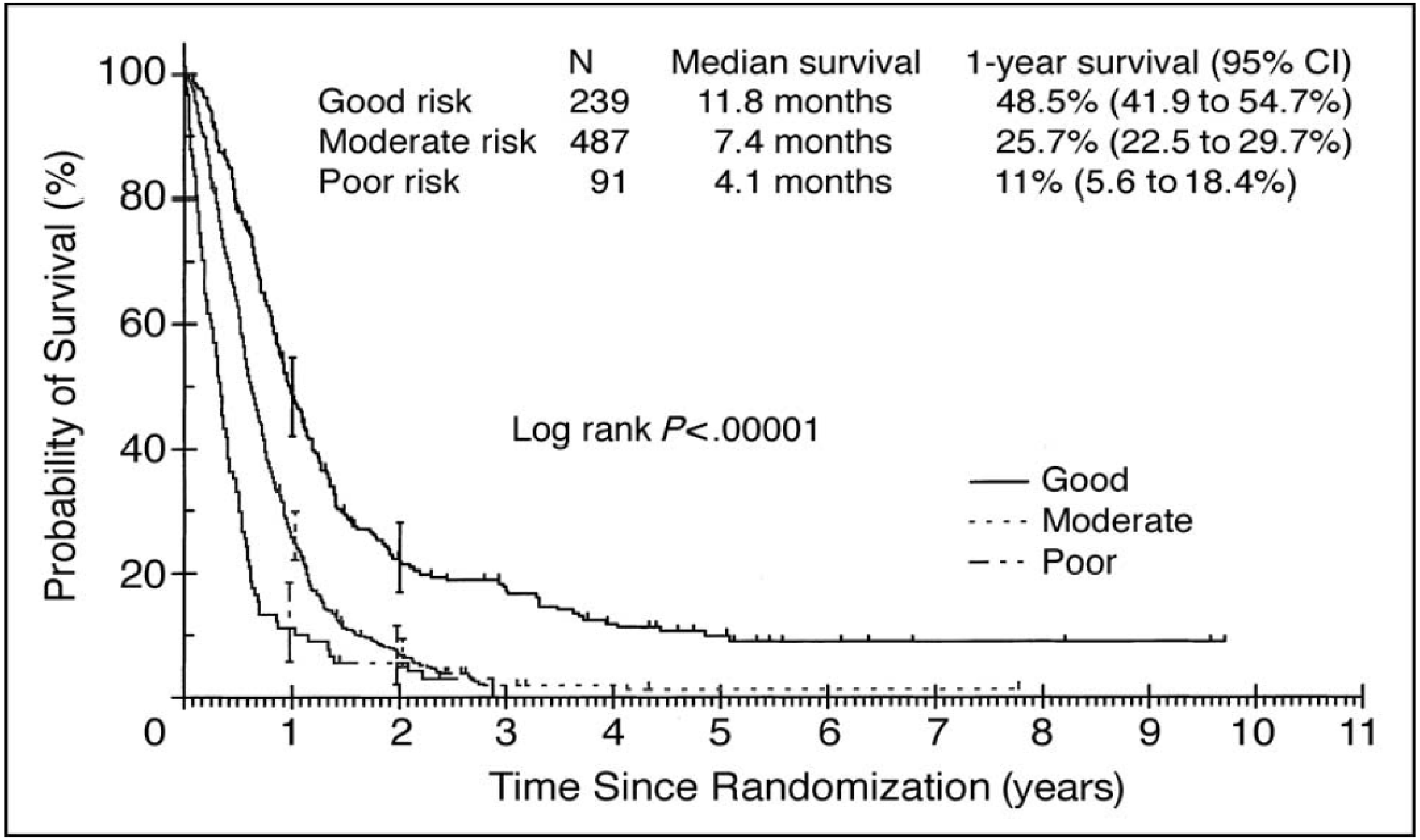
Table 3. Multivariate Baseline Prognostic Model



Factors	Hazard Ratio	99% CI	*P*
Performance status			
0-1	1		
2-3	1.575	1.251 to 1.981	< .0001
Liver metastases	1.409	1.139 to 1.743	< .0001
Peritoneal metastases	1.329	1.013 to 1.743	.007
Alkaline phosphatase ≥ 100 U/l	1.412	1.136 to 1.755	< .0001
Borderline significant factors			
Hemoglobin ≤ 11 g/L			.011
White blood cell			.06
Previous esophagectomy or gastrectomy			.054

Fig. 2 Overall survival by prognostic index (n = 817)

Chau et al. J Clin Oncol. 2004.



	N	Median survival	1-year survival (95% CI)
Good risk	239	11.8 months	48.5% (41.9 to 54.7%)
Moderate risk	487	7.4 months	25.7% (22.5 to 29.7%)
Poor risk	91	4.1 months	11% (5.6 to 18.4%)

Oncotype DX® Breast Cancer Assay

- Background of development
  - N(−)/ER(+) / Patients with breast cancer previously treated with tamoxifen
    - In the NSABP B-14 and B-20 trials,
      the effectiveness of tamoxifen + chemotherapy as postoperative adjuvant chemotherapy was verified.
    - However, distant metastasis was still low at approximately 15% even after 10 years.
      Chemotherapy for the remaining 85% was considered excessive treatment.
- Development and purpose of *Oncotype* DX®
  - 250 genes were extracted from gene databases and other sources, and 16 genes related to breast cancer recurrence were selected from them.
  - The recurrence score was calculated by combining the 16 genes and the 5 reference genes.

Selected Genes and Recurrence Scores

HER2
GRB7
HER2

INVASION
Stromelysin 3
Cathepsin L2

ESTROGEN ER
PR
Bcl2
SCUBE2

CD68

GSTM1

BAG1

PROLIFERATION Ki-67 STK15
Survivin
Cyclin B1
MYBL2

REFERENCE
Beta-actin
GAPDH
RPLPO
GUS TFRC

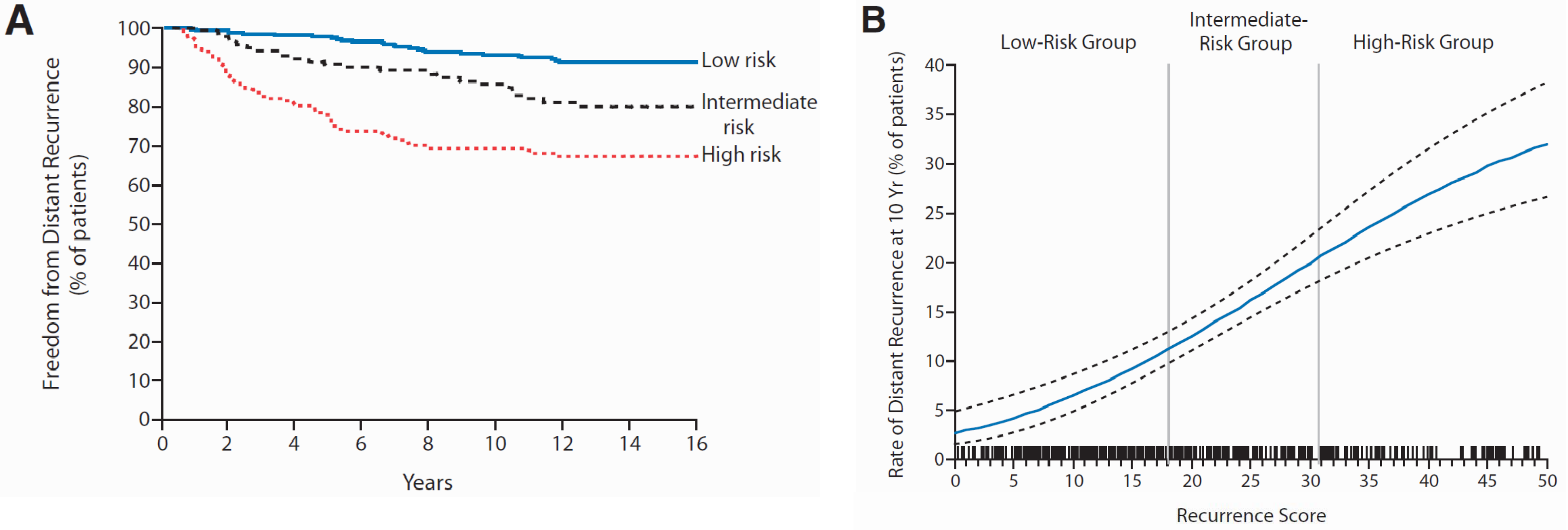
A score calculation formula was developed using multivariate analysis techniques.

$$\begin{aligned} \text{RS} = \; & + 0.47 \times \text{HER2 Group Score} \\ & - 0.34 \times \text{ER Group Score} \\ & + 1.04 \times \text{Proliferation Group Score} \\ & + 0.10 \times \text{Invasion Group Score} \\ & + 0.05 \times \text{CD68} \\ & - 0.08 \times \text{GSTM1} \\ & - 0.07 \times \text{BAG1} \end{aligned}$$

Category	RS (0–100)
Low risk	RS < 18
Int risk	RS 18–30
High risk	RS ≥ 31

Paik et al. N Engl J Med. 2004. Sparano and Paik. J Clin Oncol. 2008.

Results of Recurrence Prediction



Statistical models can be used even when
you are not interested in exposures and outcomes.

Prediction of Postoperative Complications

- Example: Complications after hepatectomy (Liver Resection)
  - If high-risk groups for complications can be predicted immediately after surgery, frequent monitoring can be conducted.
  - Identify high-risk groups for complications based on the resected specimen’s preoperative, postoperative, surgical technique, and pathological characteristics.

Jargalin et al. Ann Surg. 2002.

Uses of Multivariate Analysis

1. Confounding adjustment (causal inference)
   - Adjusting for bias of background factors when comparing with and without exposure

Today's Theme

2. Prognosis and risk prediction
   - Search for prognostic or risk factors
   - Stratification of patients by risk group

3. Consideration of interactions
   - Examine differences in effectiveness by subgroups
   - Patient stratification based on interactions is also possible

Differences in Variables to be included in the Model

- Multivariate analysis for causal inference
  - The primary objective is to estimate the association between exposure and outcome.
  - The secondary objective is to evaluate associations between variables other than exposure and outcomes.
  - Variables other than exposure are used for confounding adjustment.

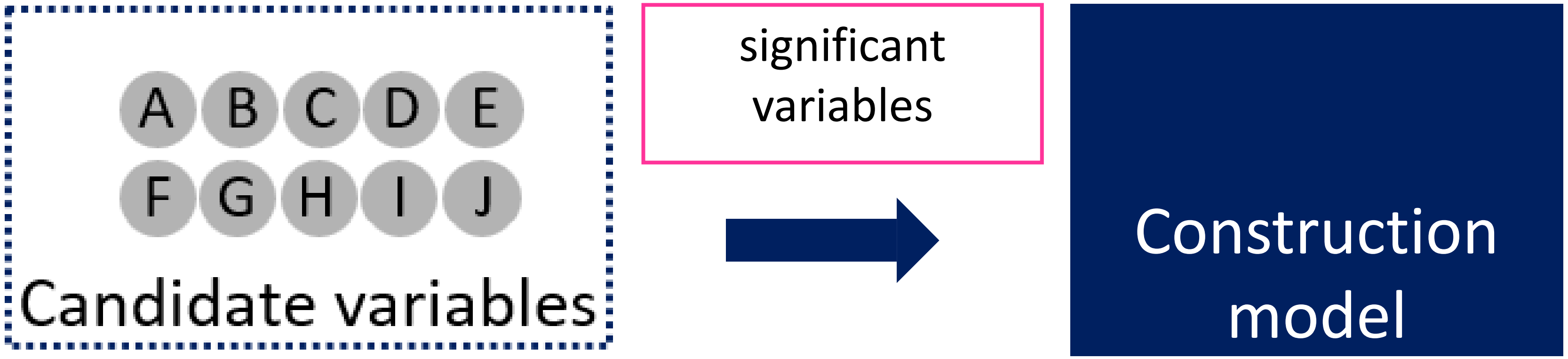
- Multivariate analysis for prediction
  - The objective is to build a model with high predictive ability.
  - Multiple **candidate variables** need to be **selected.**
  - However, **care must be taken to avoid overfitting.**

Variable Selection Method (When the prediction is the objective)

- Mechanical algorithms are common
  - Select only variables that were significant in univariate analysis
  - Forward procedure
  - Backward procedure
  - Stepwise procedure
  - Best-subset selection procedure
  - Others (such as neural networks and regression trees)

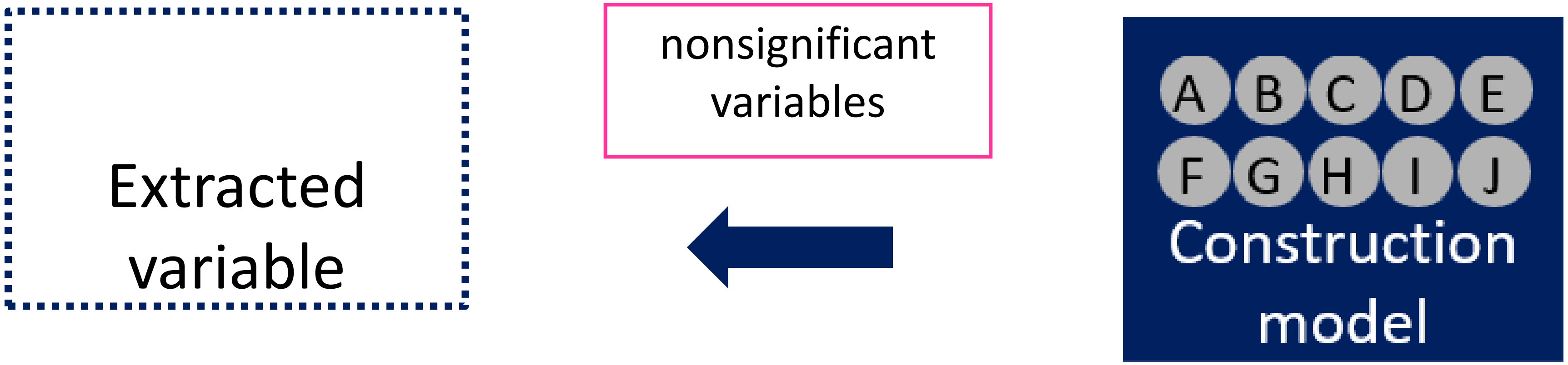
- Note: There is no definitive method for variable selection

Forward Procedure



- Considering only variable inputs
  - Start with an empty construction model
    - Evaluate candidate variables A–J for inclusion in the model one by one
    - Include the variable with the lowest *P* value that is significant
    - Repeat until there are no more significant variables
  - Transition of variables in the constructed model
- “Zero → One → Two → ...”

Backward Procedure

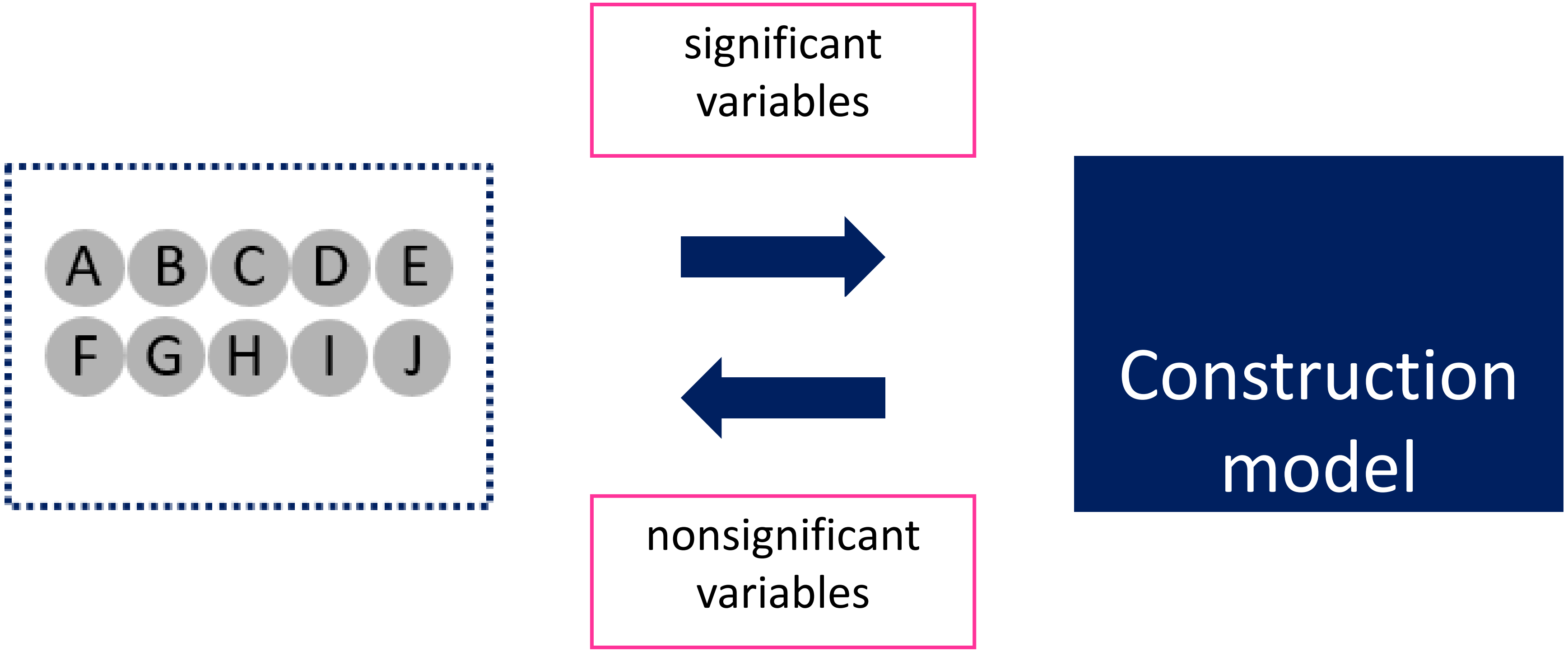


- Consider only removing variables
  - Start from a state where the constructed model is full
    - The variable with the largest and nonsignificant *P* value is removed
    - Repeat until there are no more nonsignificant variables
  - Transition of variables in the constructed model
- “Ten → Nine → Eight → ...”

Stepwise Procedure

※Inclusion and removal criteria may be different
(for example, inclusion significance level 0.20, removal significance level 0.05)

- Sequential increase/decrease of variables



Best-Subset Selection Procedure

Variable	A	B	C	D	E	F	G
Include?	YES	YES	YES	YES	YES	YES	YES
	NO	NO	NO	NO	NO	NO	NO

- Consider all patterns (*2K* patterns if there are *K* variables)
  - Calculate the goodness of fit to the data for each number of variables (*K)* to be included
  - Extract the model with the best fit for each *K*
    - Suppose that including *K** or more *variables* results in a small increase in goodness of fit
  - Select the model with the highest goodness of fit that contains *K** variables
- If there is no difference in the degree of fit between the models, you can choose the one that is clinically reasonable.

Generalized Variable Selection: Advantages and Disadvantages

Technique	Advantage	Disadvantage
(1) Forward procedure	Can be implemented with a small number of examples	• Depends on the order of inclusion • Likely to miss important variables
(2) Backward procedure	Less likely to miss important variables	• Difficult to implement with a small number of cases • Depends on the removal order • Once a variable is removed, it cannot be included again
(3) Stepwise procedure	Shortcomings of (2) can be addressed	Can be influenced by the order
(4) Best-subset selection procedure	Not affected by the order	High computation load

Supplementation for Variable Selection Method

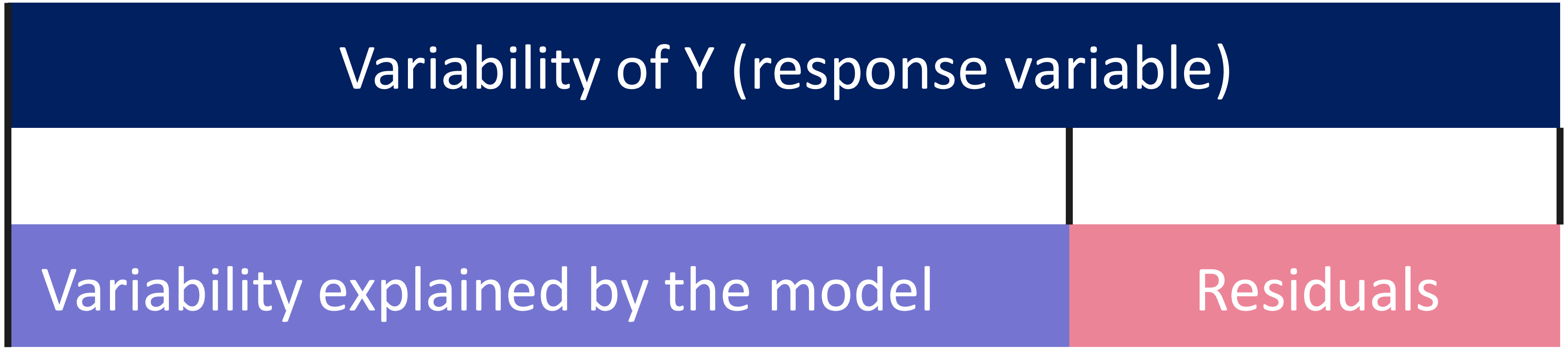
- How to select only significant variables in univariate analysis
  - This is not necessarily wrong, but there is a high risk of missing variables.
    - Because a variable that is not significant in univariate analysis can become significant in multivariate analysis
  - Frequently used in gene expression analysis
- Important variables can be forced in
  - The goal is to build a model with high predictive ability
  - There is a risk that the data in hand may not be selected
- Variable selection is arbitrary
  - Results may vary depending on inclusion/removal criteria and goodness of fit indicators.

Review: 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.
  - The contribution ratio increases as the number of explanatory variables is increased.
    - Even if factors that are completely unrelated to the outcome are included, the contribution rate increases.
  - If you use a complex model (for example, a model with additional quadratic terms), **the contribution ratio can be as close to 1 as possible.**

- Overfitting models
  - Models with a higher-than-necessary fit to the data in hand
  - **Low extrapolation and cannot be generalized**

Image of Contribution Ratio


Variability of Y (response variable)
Variability explained by the model
Residuals

$$\text{Contribution ratio } (R^2) = \frac{\text{Variability explained by the model}}{\text{Variability of } Y \text{ (response variable)}} = 1 - \frac{\text{Residuals}}{\text{Variability of } Y \text{ (response variable)}}$$

The smaller the residuals, the higher the contribution ratio.

Examples of Overfitting

- Model the relationship between blood pressure *(Y)* and age *(x)*
  - – Model: Blood pressure $(Y) = \alpha_0 + \alpha_1 \times$ Age $(x) +$ Error
    - The blood pressure value is the systolic blood pressure
    - A linear model of the relationship between blood pressure and age
- Include variables unrelated to blood pressure in the model
  - – A random variable (random number) is generated that takes values between 0 and 1
  - – It is random and not related to blood pressure
  - – Does the result remain the same?

Model to Fit

■ Include up to 10 random variables (d)

- Model 1: Blood pressure (Y) = $\alpha_0 + \alpha_1 \times$ age (x) $+ \boxed{\beta_1 \times d_1}$ + error
- Model 2: $Y = \alpha_0 + \alpha_1 \times x + \boxed{\beta_1 \times d_1 + \beta_2 \times d_2}$ + error
- Model 3: $Y = \alpha_0 + \alpha_1 \times x + \boxed{\mathrm{\beta}_1 \times d_1 + \beta_2 \times d_2 + \beta_3 \times d_3}$ + error

●

●

●

- Model 10: $Y = \alpha_0 + \alpha_1 \times x + \boxed{\beta_1 \times d_1 + \ldots + \beta_{10} \times d_{10}}$ + error

■ Calculate the contribution ratio in each model

Note: Hypothetical example

Change in Contribution Ratio

- Before including the random variable: R^2 = **0.525**
  - – Interpretation:

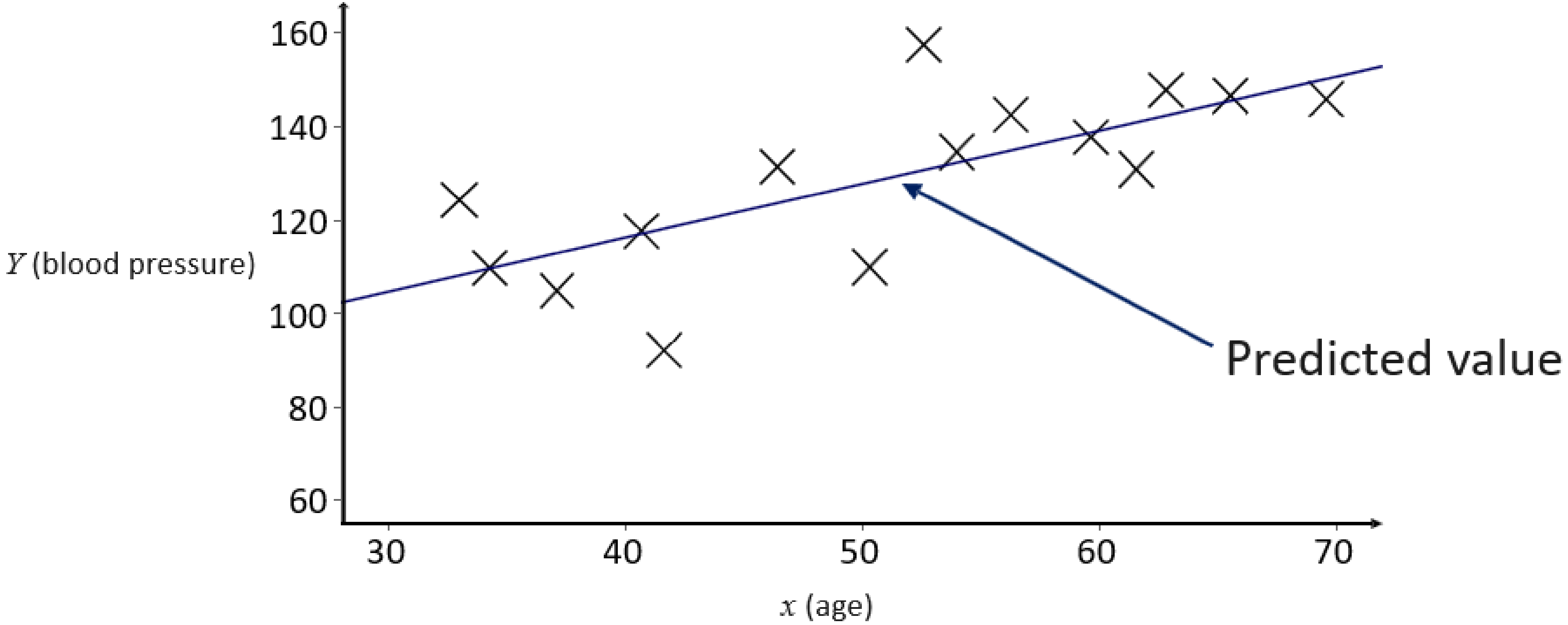
    Model Blood pressure $(Y) = \alpha_0 + \alpha_1 \times$ Age (x) + Error explains 52.5% of the data

- After including the random variables

Model 1	Model 2	Model 3	Model 4	Model 5
0.525	0.560	0.656	0.656	**0.731**
Model 6	**Model 7**	**Model 8**	**Model 9**	**Model 10**
0.731	0.797	0.806	0.807	**0.865**

Note: Hypothetical example

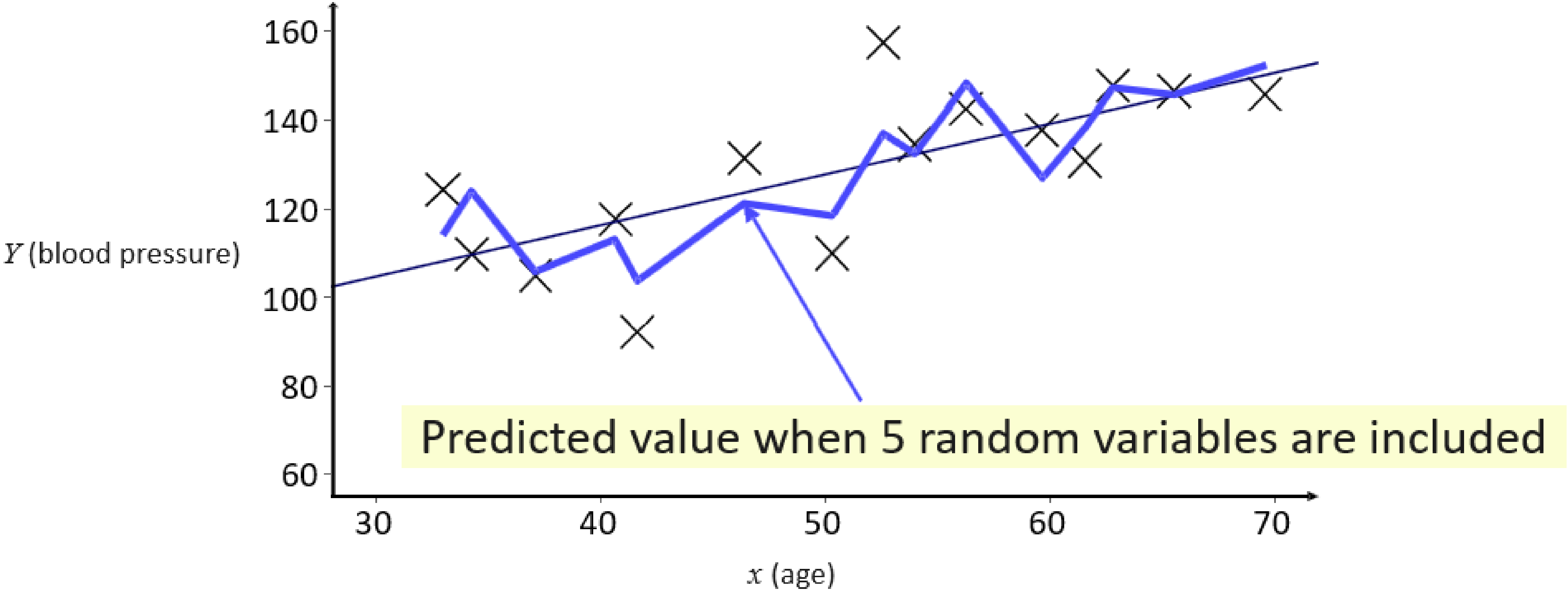
Prediction before Inclusion (R^2 = 0.525)



Note: Hypothetical example

Prediction for Model 5 (R^2 = 0.731)

The residual becomes smaller, and the line turns into a curve.

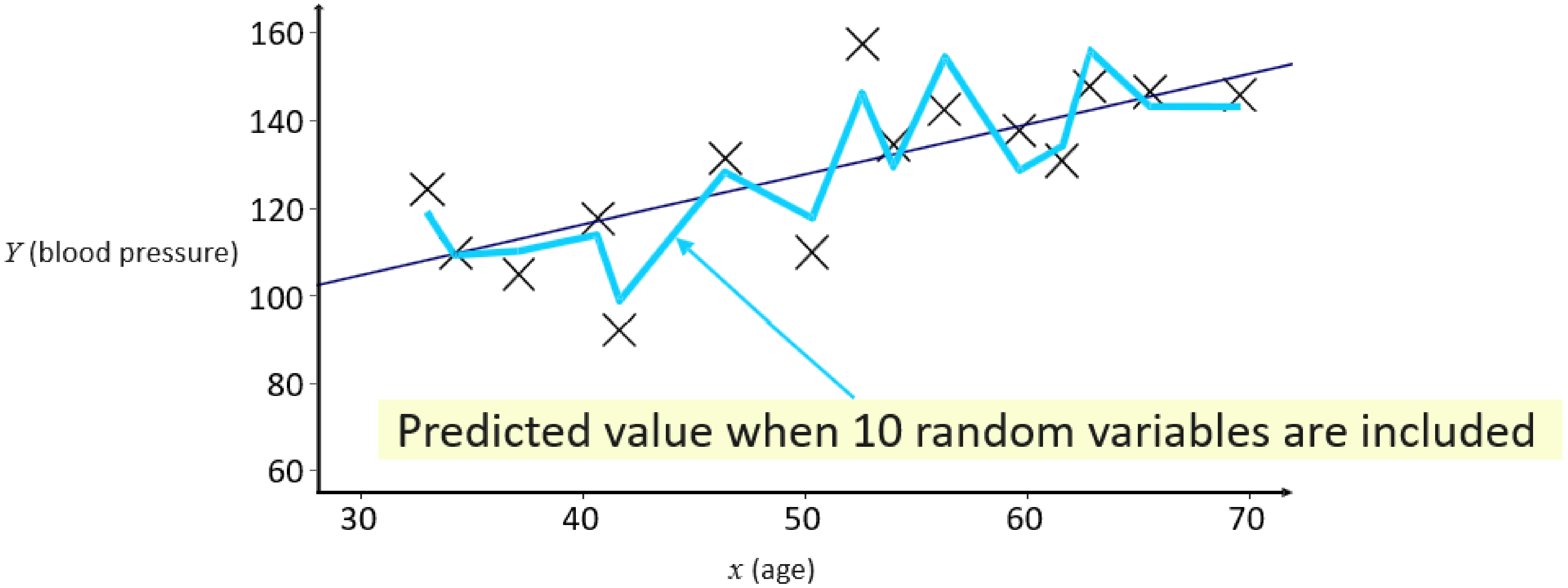


Predicted value when 5 random variables are included

Note: Hypothetical example

Prediction for Model 10 ($R^2 = 0.865$)

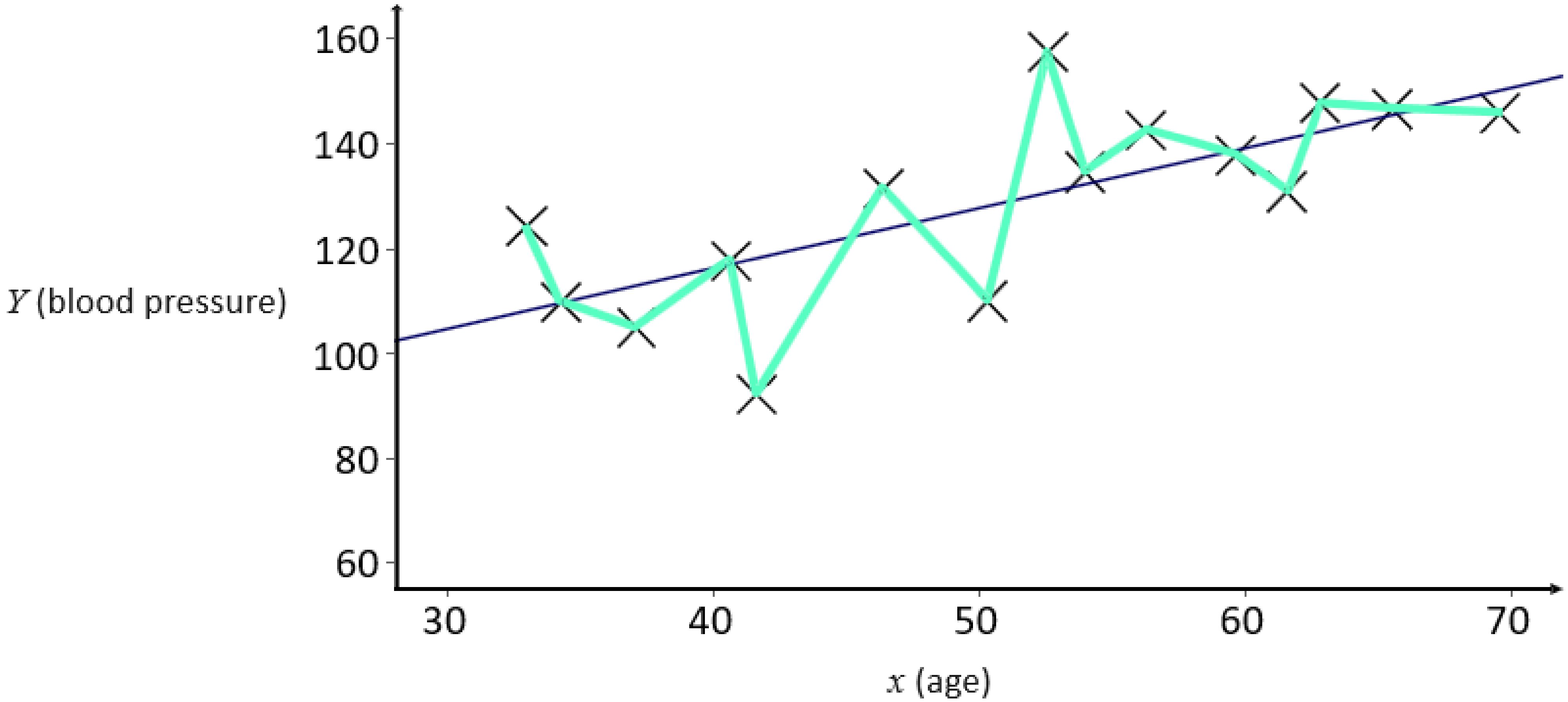
It becomes a curve that passes through all the data points.



Predicted value when 10 random variables are included

When More Complex Models Are Built

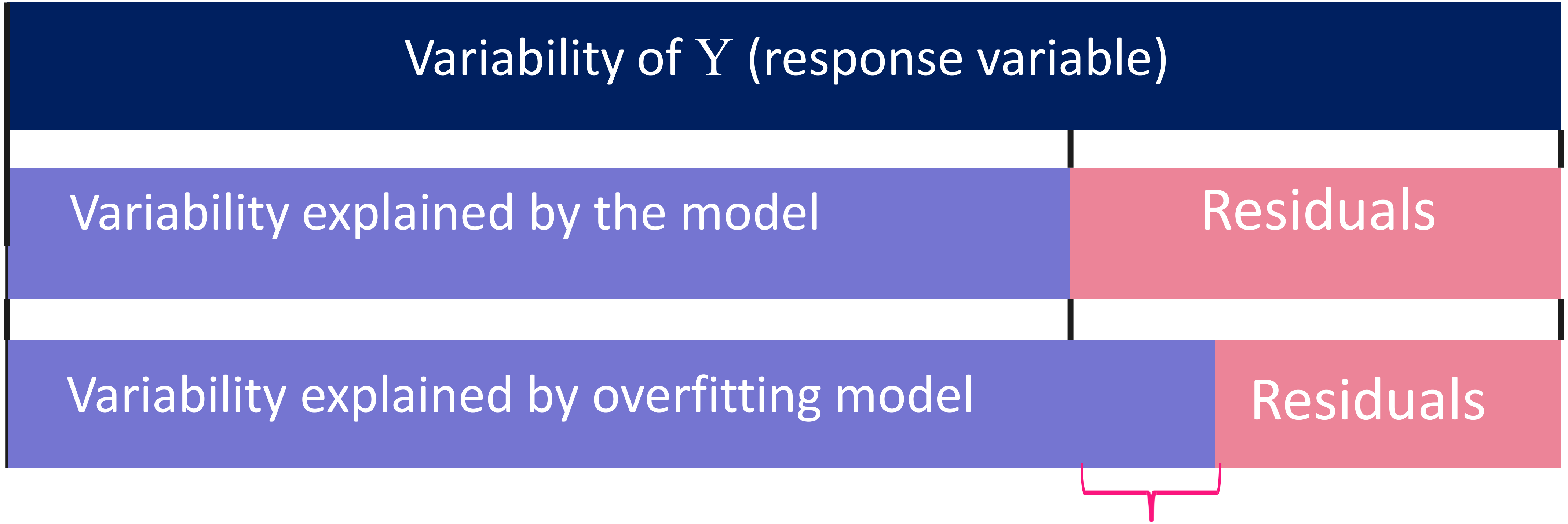
$R^2 = 1$



Why Did the Contribution Ratio Improve?

- Including many random variables means that...
  - This is equivalent to including many variables that identify participants, such as place of origin or occupation, in the model.
  - It could be predicted to fit all data.
- Even if we include variables that are not related to the outcome, the proportion of "variability explained by the model" improves.

Why Did the Contribution Ratio Improve?



Variability of Y (response variable)

Variability explained by the model

Residuals

Variability explained by overfitting model

Residuals

It even explains random variabilities

Problems with Overfitting Models

$$\text{Blood pressure } (Y) = \alpha_0 + \alpha_1 \times \text{age } (\mathrm{x}) + \boxed{\text{error}}$$

- **Overfitting models "explain" even random errors that only exist in the data in hand.**

- **Unless the data has the same error as the data in hand, the prediction results will not be reproduced.**

Outline

- Review of the Fifth Lecture
- Multivariate Analysis for Prediction
- Necessity of Validation

Dealing with Overfitting

- How to fit future data?

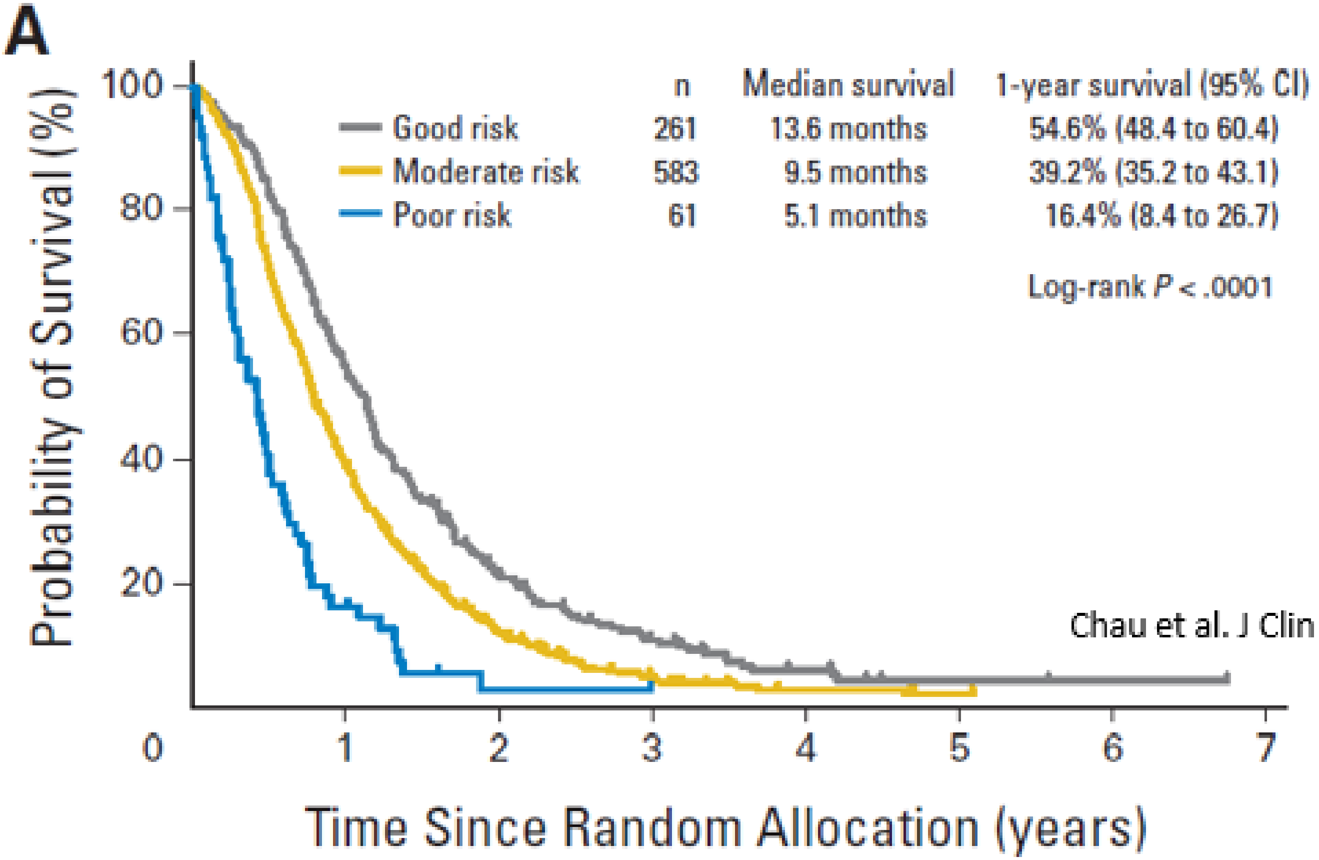
- Developing a construction algorithm
  - Not to include medically unexplainable variables
  - Screening with strict variable selection criteria

- Validation of model extrapolation
  - Validate the construction algorithm
  - For training data, prepare validation data

Validation of RMHI*

*RMHI: Royal Marsden Hospital Index

- Reproduction of predicted results from REAL-2 test



Chau et al. J Clin Oncol. 2009.

Validation of Onco*type* DX®

Sparano and Paik. J Clin Oncol. 2008.

- Model construction (n = 447)
- External Data Validation
  - Data from clinical studies (n = 668)

Paik et al. N Engl J Med. 2004.

- Validation on other data
  - Data from case–control studies (n = 790)
    (Habel et al. Breast Cancer Res. 2005)
  - Japanese data (n = 280)
    (Masuda et al. ASCO Breast Cancer Symposium 2009: #68.
    / Toi et al. Cancer 2010.)

Reference: JCO *statistical guidelines*

http://jco.ascopubs.org/site/ifc/stats.xhtml

- *When studies report prognostic or predictive markers (based on clinical, etiologic, or genetic factors), JCO gives highest priority to studies in which an algorithm obtained in a training set is applied exactly the same way in the validation set as in the training set. That is, the factors included and the cutoffs must be determined in the training set and applied to each patient in the validation set. Authors should clearly identify further statistical analyses as exploratory.*

- *JCO gives lower priority to prognostic studies that report the results of an algorithm from the same data that were used to fit the algorithm. However, inclusion of cross-validation analyses and a clear statement of the limitations of the study will enhance the article's priority.*

Types of Validation

- Re-substitution (application to constructed data) is contraindicated
- External validation
  - Evaluating the generalizability of the constructed model with independent external data
    - Example: Validation of RMHI and Onco*type* DX®.
- Internal validation
  - Simultaneous construction and evaluation with data in hand
    - Validation of the construction algorithm itself
    - Because it is not completely external data, generalizability of the construction model is not always guaranteed.

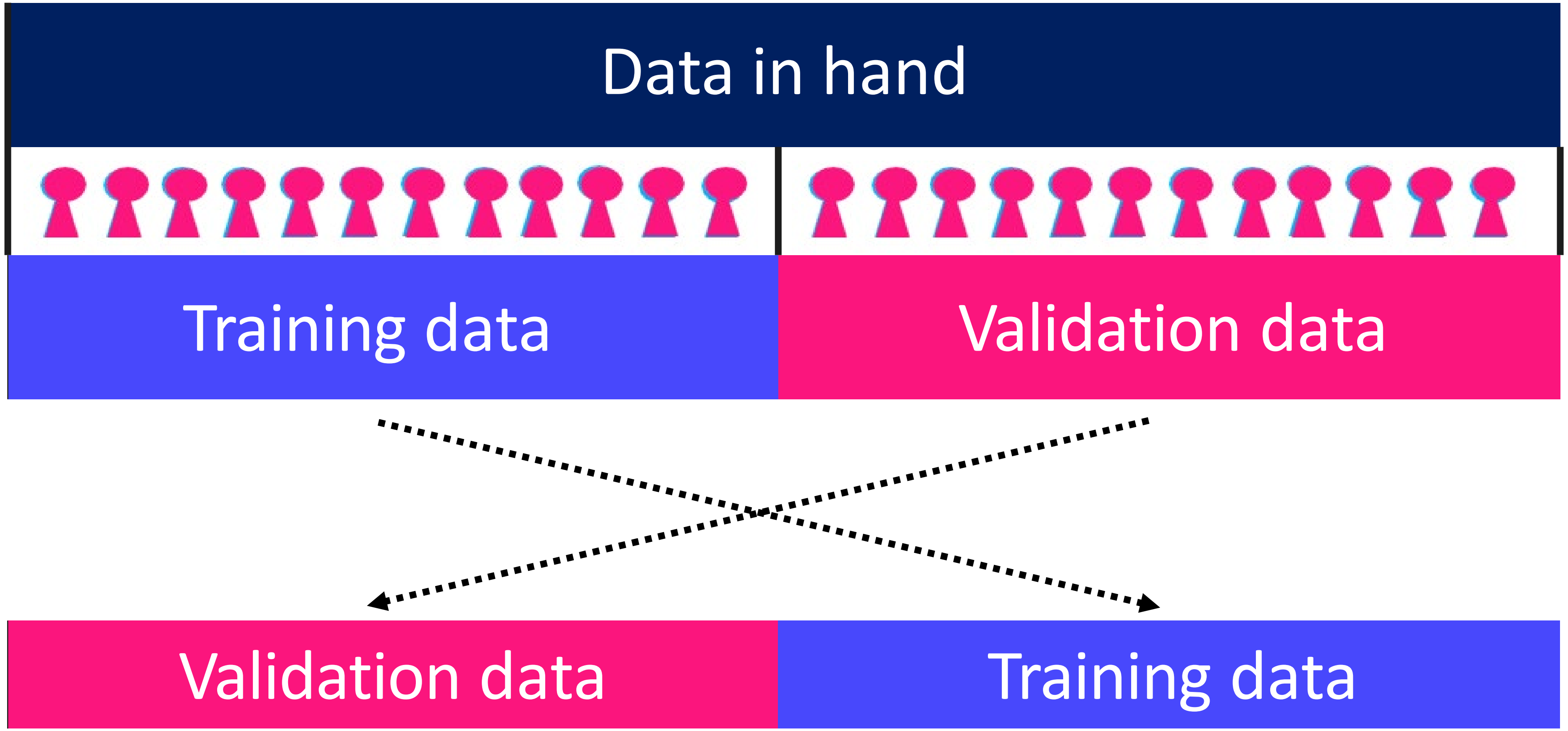
    (There is no guarantee that the data will be reproduced at different times or in different facilities/countries)

Typical Internal Validation

- ■ Split-group validation
  - – Split the data in hand into the following two parts
    - Training data: Data for building the model
    - Validation (Test) data : Data for model validation
  - – Requires a sufficiently large possible number of cases.
- ■ Cross-validation
  - – Using all data for model validation
  - – More efficient than split-group when the number of cases is small

Simple Cross-Validation

*Data may be split randomly or nonrandomly



Data in hand

Training data

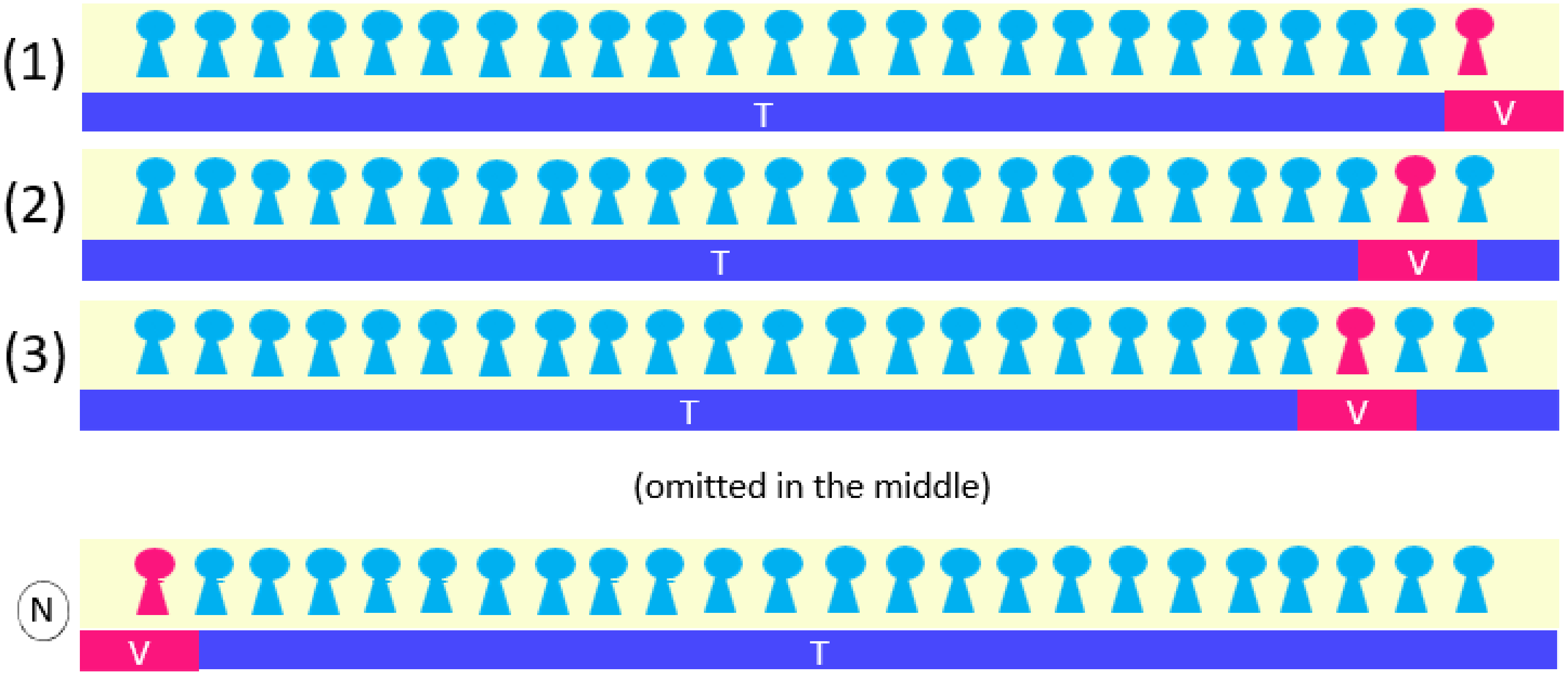
Validation data

Validation data

Training data

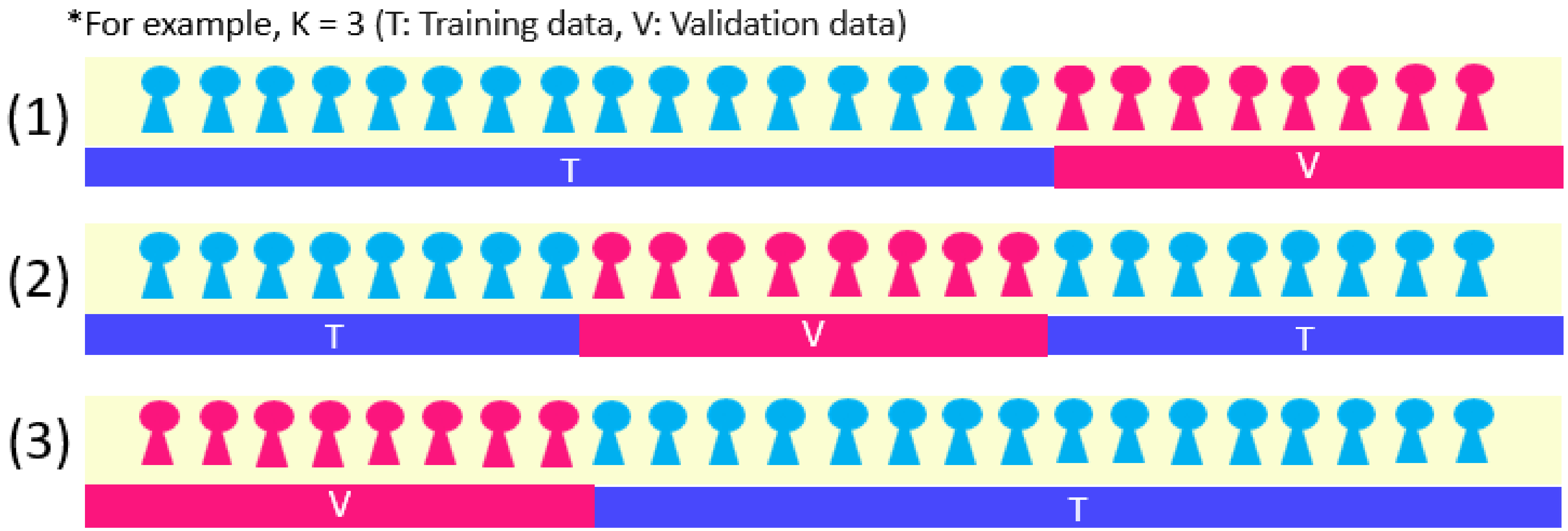
Leave-One-Out Cross-Validation (LOOCV)

*N: Number of cases, T: Training data, V: Validation data



- It takes time for calculation, but validation can be performed N times (for N people)
- The rate at which predictions are wrong (error rate) can be evaluated

K-fold Cross Validation



- Can be executed in a shorter time than LOOCV
- If the value of K is large, the error proportion can be evaluated

Summary

- Multivariate analysis for prediction
  - It is necessary to build a model with consideration for overfitting
  - The built model needs to be validated
- Construction and validation must be carried out independently
- Through the 5th and 6th lectures,
  - select the appropriate method for your objective.

    If you are unsure, consult a statistician.