生物統計基礎セミナー 第6回

2014年3月28日

多変量解析 2

国立がん研究センター 生物統計部門

野村 尚吾

OUTLINE

- 第5回セミナーの復習
- 予測のための多変量解析
- バリデーションの必要性

統計モデルとは？

■ バラツキ（誤差）を考慮した数学モデル

家賃 = 地域の相場家賃 + 1.5 × 階数 + 2.5 × 広さ + 誤差

– 関心のあるアウトカムを「目的変数、結果変数」と言う
– アウトカムを説明する変数を
「説明変数、原因変数、共変量」と言う

■ 多くの統計モデルは”線形式”の構造をとる

– 線形式：掛け算同士の加減で表現される式

目的変数 = β_0 + β_1 × 説明変数1 + β_2 × 説明変数2 + … + 誤差

- 太字部分を“パラメータ”と呼ぶ
- 「説明変数の効果が加法的」と仮定したモデル

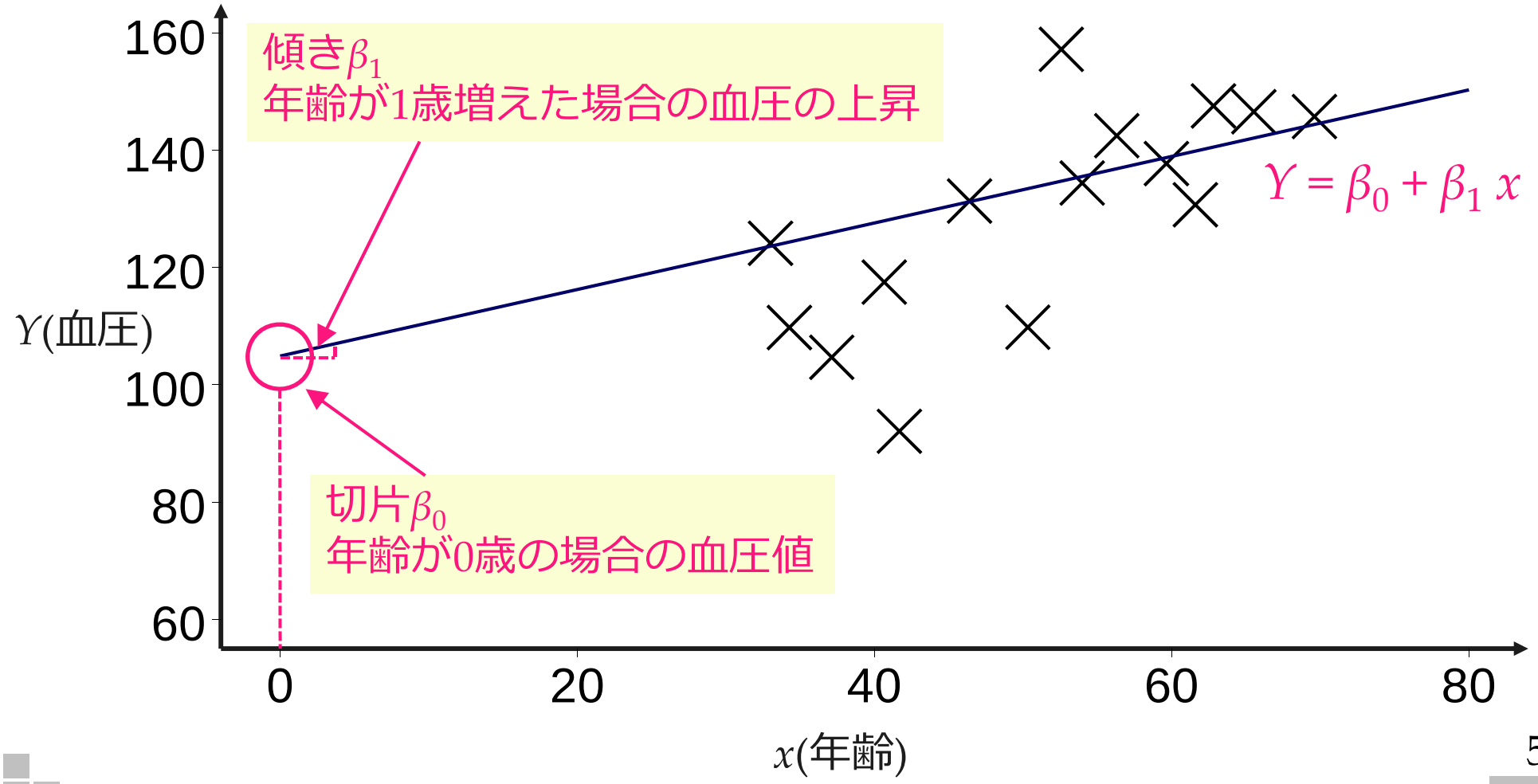
単回帰モデル

- $Y = \beta_0 + \beta_1 x +$ 誤差
  - Y：目的変数、x：説明変数（**1つだけ**）
  - β_0：切片、β_1：傾き

- 血圧(Y)と年齢(x)の関係をモデル化
  - モデル：血圧$(Y) = \beta_0 + \beta_1 \times$年齢$(x) +$誤差
    - 血圧値は収縮期血圧とする
    - 血圧と年齢の関係を直線で表したモデル

注：仮想例

散布図(n=15)に直線を当てはめ



統計モデル ～まとめ～

- バラツキを考慮した“線形式”

  目的変数 = β_0 + β_1 × 説明変数1 + β_2 × 説明変数2 + … + 誤差

  - 単変量モデルのときは直線(切片：β_0、傾き： β_1)の関係を仮定する

- ロジスティック回帰モデル

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

  - 二値アウトカムに対する統計モデル
  - 推定したパラメータからオッズ比が推定できる

- Coxの比例ハザードモデル

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

  - 生存時間アウトカムに対する統計モデル
    - 比例ハザード性(時点によらずハザード比が一定)を仮定したモデル
  - 推定したパラメータからハザード比が推定できる

OUTLINE

- 第5回セミナーの復習
- 予測のための多変量解析
- バリデーションの必要性

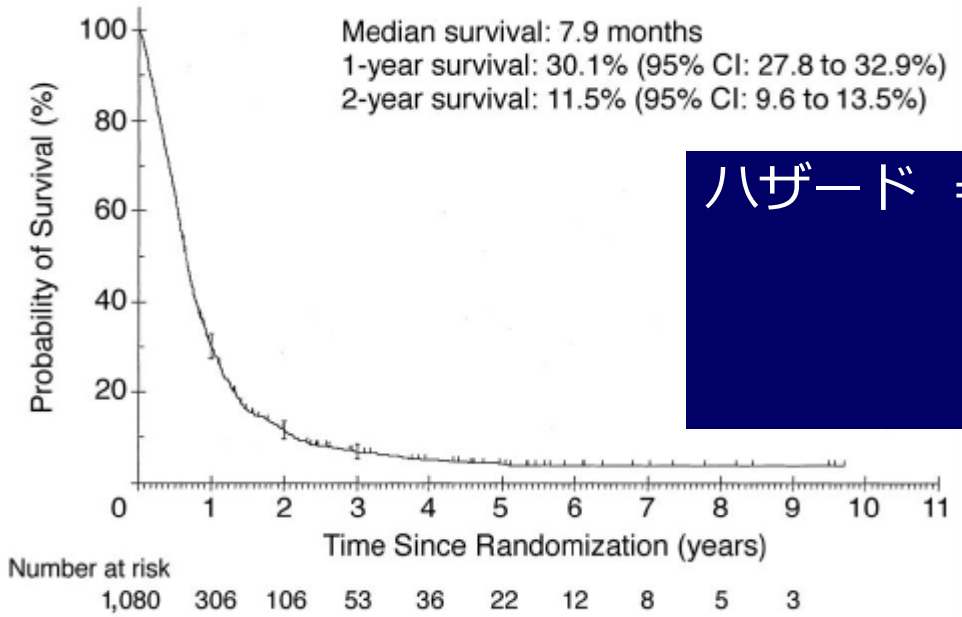
Chau, et al. J Clin Oncol 2004.

Royal Marsden Hospital Index

- 胃癌と食道癌の予後因子探索
  - 1992～2001年に実施された3試験の統合解析

- 予後因子の探索が目的
  - *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.*

Chau, et al. J Clin Oncol 2004.



ハザード ＝ 基準ハザード ×exp{ 0.7149 × (PS: 0-1 / 2)
+ 0.3873 × (肝meta)
+ 0.4357 × (腹膜meta)
+ 0.1080 × (ALP) }

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

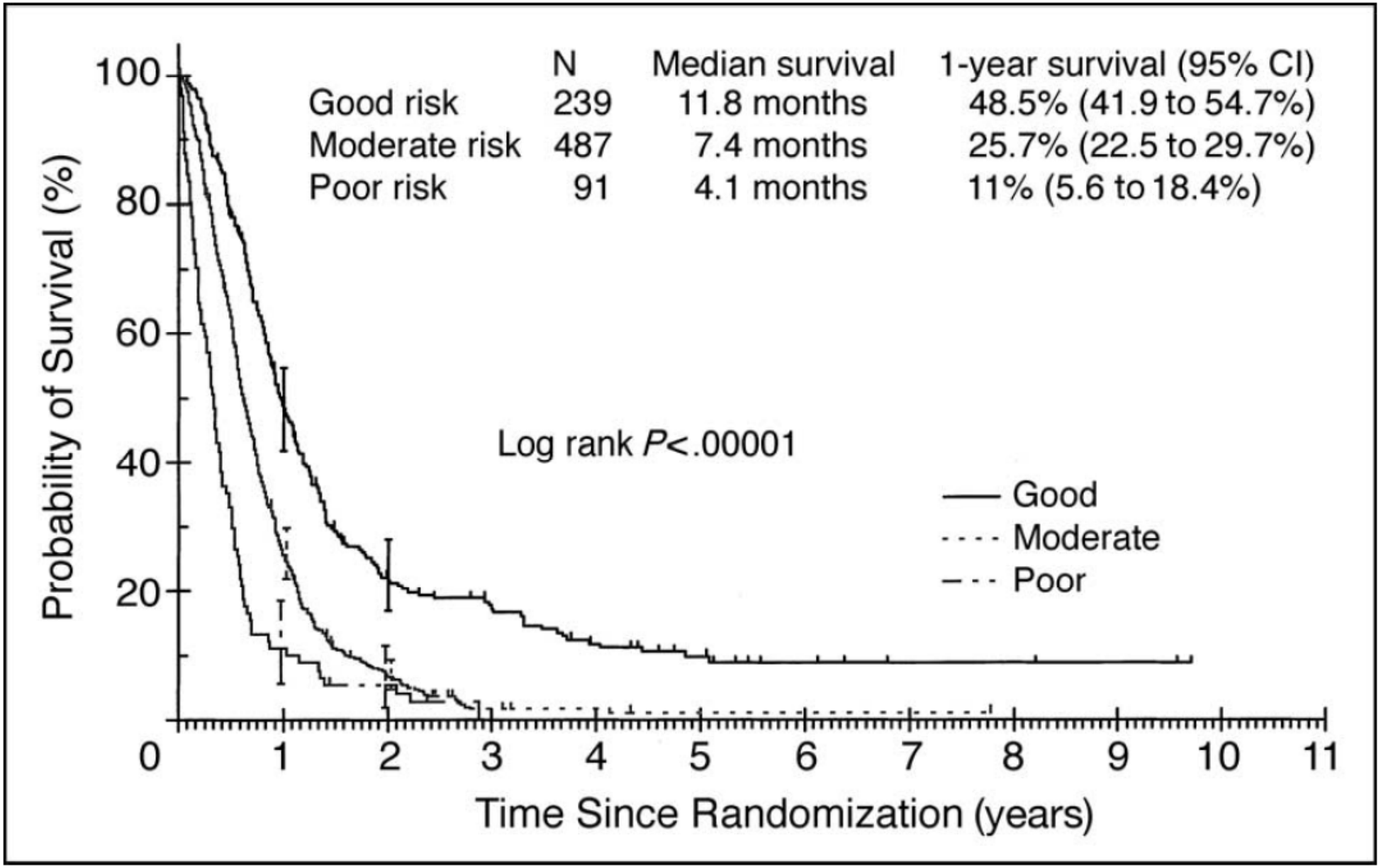


Fig 2. Overall survival by prognostic index (n = 817).

Oncotype DX® Breast Cancer Assay

■ 開発の背景

- N(-)/ER(+)/タモキシフェン既治療例の乳がん患者
  - NSABP B-14, B-20試験で術後補助化学療法として
    タモキシフェン＋化学療法の有効性が検証されていた
  - しかし遠隔転移は10年時点でも約15%と低く、
    残りの85%に対する化学療法は過剰治療と捉えられていた

■ Oncotype DX®の開発と目的

- 遺伝子データベースなどから250遺伝子を抽出し、
  その中から乳癌の再発に関連する16遺伝子を選択
- 16遺伝子とreferenceとなる5遺伝子を併せて
  再発スコアを計算

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

選択された遺伝子と再発スコア

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

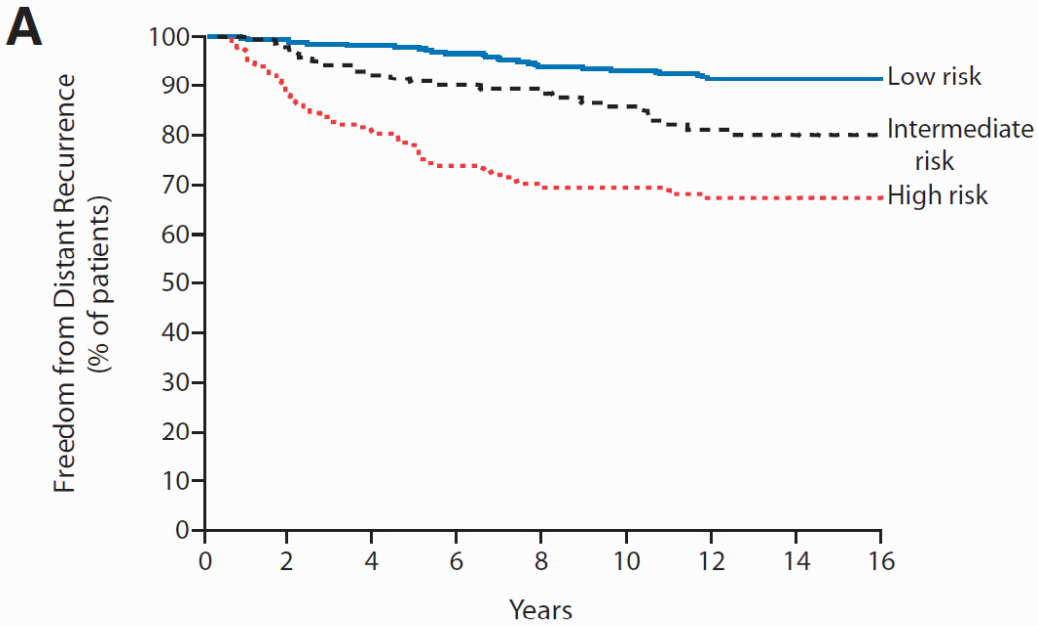
多変量解析の手法を用いてスコア計算式を構築

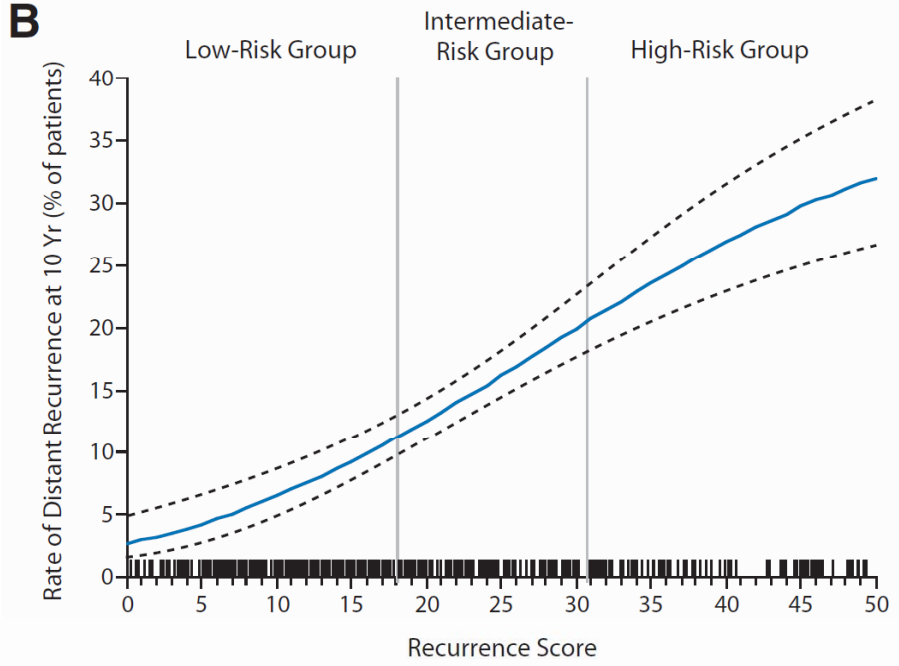
RS = + 0.47 x HER2 Group Score
- 0.34 x ER Group Score
+ 1.04 x Proliferation Group Score
+ 0.10 x Invasion Group Score
+ 0.05 x CD68
- 0.08 x GSTM1
- 0.07 x BAG1

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

Sparano and Paik. J Clin Oncol 2008.

再発予測の結果





曝露と結果に興味がある場合以外でも
統計モデルが活用できる

Jargalin, et al. Ann Surg 2002.

術後合併症予測

- 例：肝切除後の合併症
  - 合併症の高リスク群が術後すぐに予測できれば、頻回なモニタリングなどの対策をとることが可能
  - 術前・術後・手術法・切除検体の病理的特徴などから合併症の高リスクグループを同定

多変量解析の用途

1. 交絡の調整(因果推論)
   - 曝露の有無を比較する際に背景因子の偏りを調整

本日のテーマ

2. 予後やリスクの予測
   - 予後因子 or リスク因子の探索
   - リスクグループによる患者の層別化

3. 交互作用の検討
   - サブグループによる効果の違いを検討
   - 交互作用に基づく患者の層別化も可能

モデルに含める変数の違い

■因果推論のための多変量解析
- 曝露と結果の関連を推定することが主目的
- 副次的に曝露以外の変数と結果の関連を評価
- 曝露以外の変数は交絡調整のために用いる

■予測のための多変量解析
- 予測能の高いモデルを構築することが目的
- 複数の**候補変数の取捨選択**が必要
- ただし、**過剰適合に注意しなければならない**

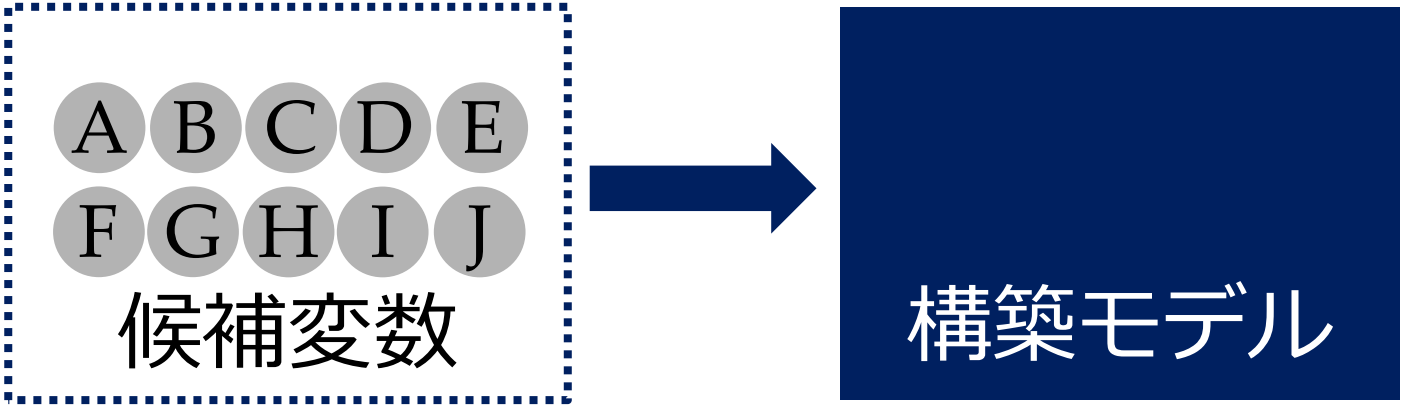
変数選択法(予測が目的の場合)

- 機械的なアルゴリズムが一般的
  - 単変量解析で有意な変数のみ選択
  - 変数増加法（forward procedure）
  - 変数減少法（backward procedure）
  - 変数増減法
  - 総当たり法
  - その他(neural networkやregression tree等)

- 注意：変数選択法に決め手はない

変数増加法

有意な変数



■変数の取入れだけを考える

– 構築モデルが空の状態からスタート

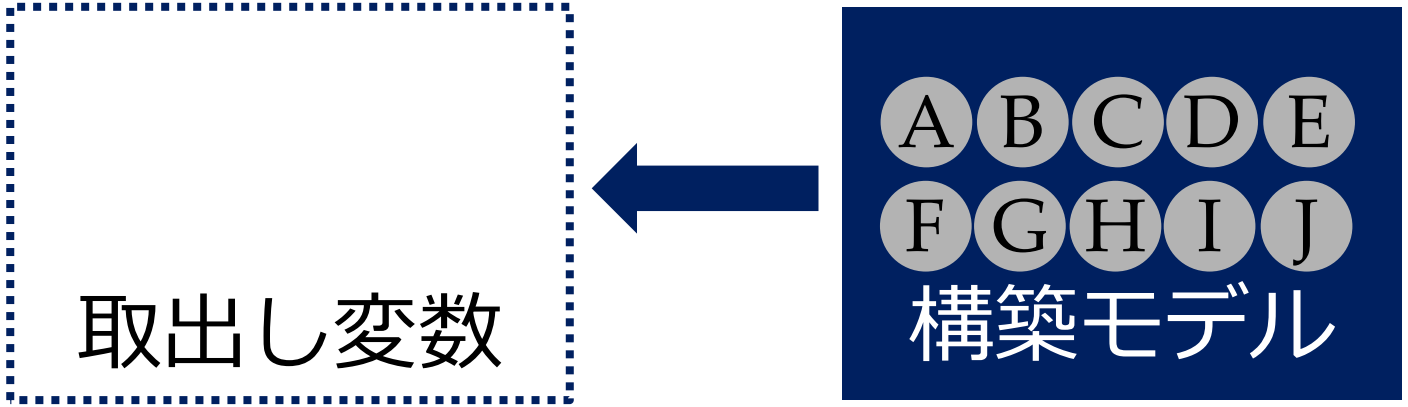
- 候補変数A~Jを1つずつモデルに含めるか評価
- P値が有意かつ最小の変数を取り込む
- 有意な変数がなくなるまで繰り返す

– 構築モデル内の変数の推移

「0個 ⇒ 1個 ⇒ 2個 ⇒ ・・・」

変数減少法

有意でない変数



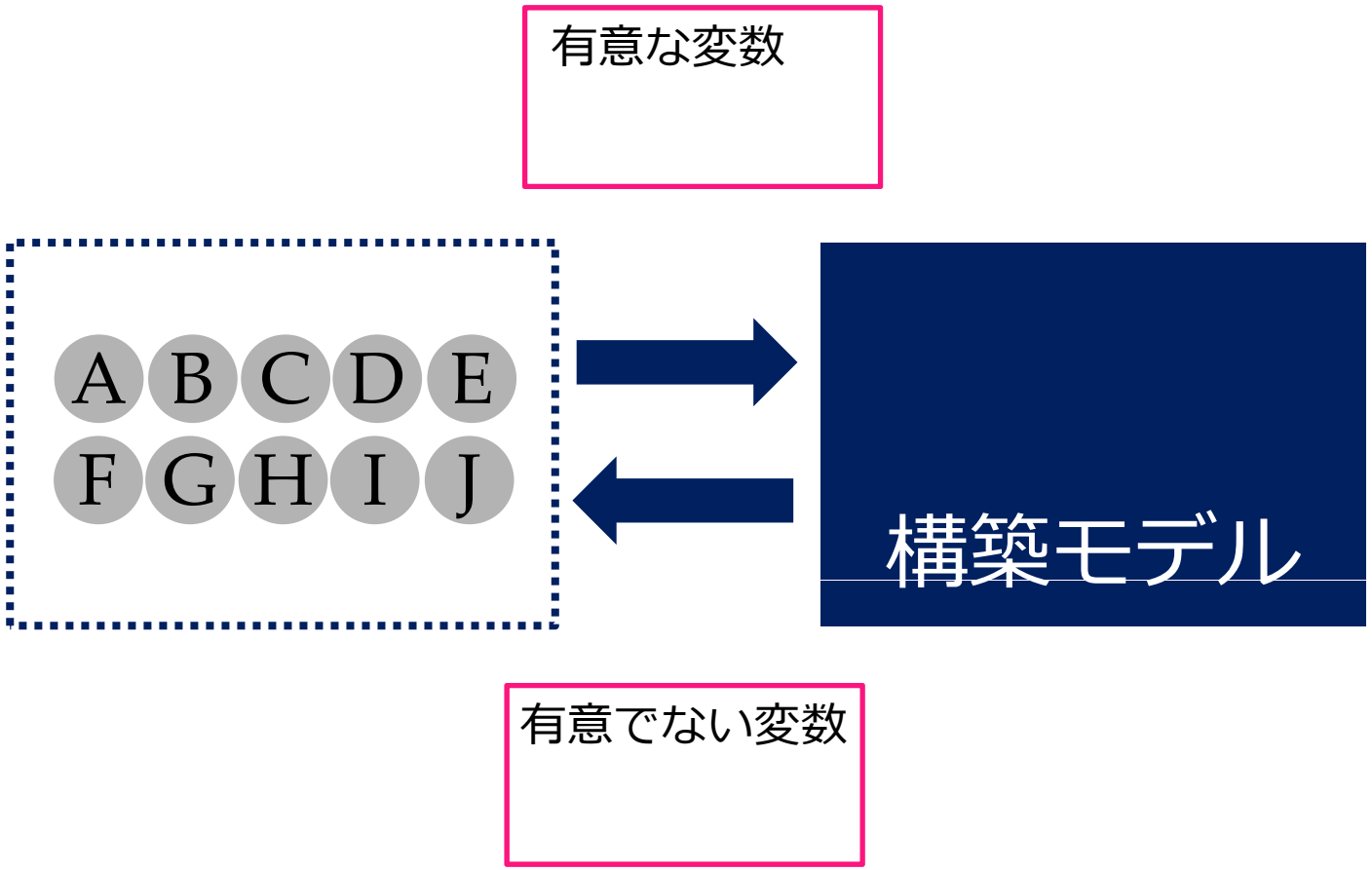
■変数の取出しだけを考える

– 構築モデルが満杯の状態からスタート
  • P値が有意でなくかつ最大の変数を取出し
  • 有意でない変数がなくなるまで繰り返す
– 構築モデル内の変数の推移
  「10個 ⇒ 9個 ⇒ 8個 ⇒ ・・・」

変数増減法

※ 取入れ基準と取出し基準は異なっていて良い(例：取入れ有意水準0.20,取出し有意水準0.05)

■変数増加・減少を逐次的に繰り返す



総当たり法

変数	A	B	C	D	E	F	G
含める？	YES	YES	YES	YES	YES	YES	YES
	NO	NO	NO	NO	NO	NO	NO

- 全パターン(変数がK個なら2^K通り)を検討
  - 含める変数の数(K)毎にデータへの適合度を計算
  - K 毎に適合度が最大のモデルを抽出
    - 変数をK^*個以上含めても適合度の増加が僅かだとする
  - 変数をK^*個含む適合度最大のモデルを選択
- 適合度に差がないモデル同士なら、
  臨床的にreasonableな方を選択しても良い

汎用的な変数選択法：利点と欠点

手法	利点	欠点
① 変数増加法	少ない例数でも実施可能	・取入れ順に依存 ・重要な変数を逃す可能性が高い
② 変数減少法	重要な変数を逃す可能性が低い	・少ない例数だと実施困難 ・取出し順に依存 ・一度取出されると取入れできない
③ 変数増減法	②の欠点に対処可能	順番に影響されうる
④ 総当たり法	順番には影響されない	計算の負荷が高い

変数選択法の補足

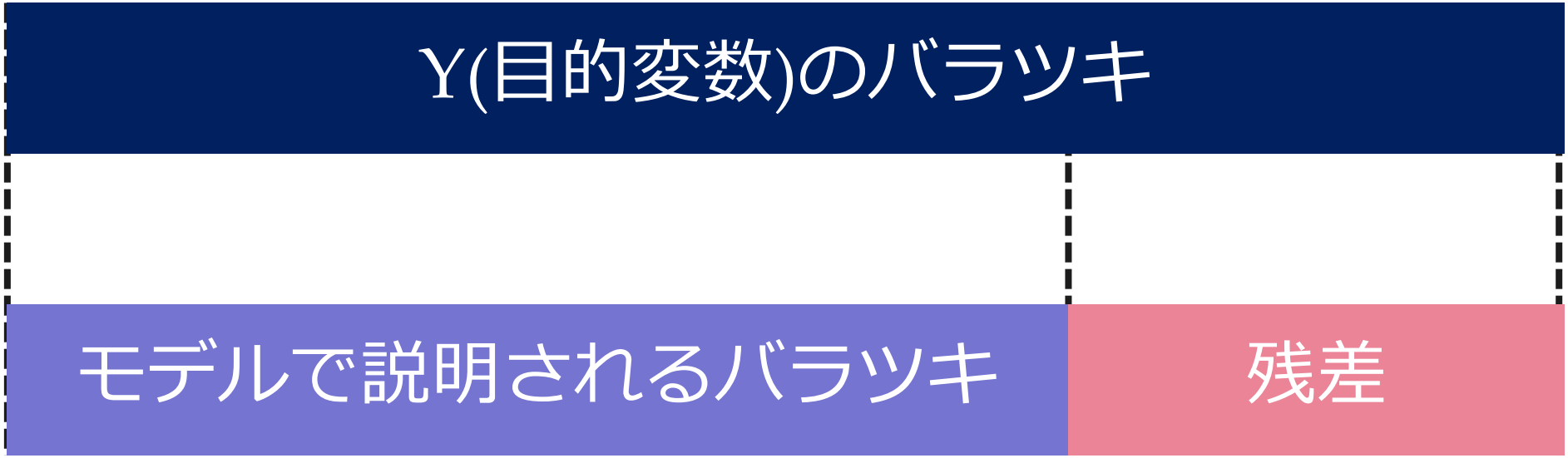
- 単変量解析で有意な変数のみ選択する方法について
  - 決して間違いではないが変数の取り逃しリスク大
    - 単変量で有意でなくても多変量で有意になりうるため
  - 遺伝子発現解析では頻用されている
- 重要な変数は強制投入して構わない
  - 予測能が高いモデルを構築することが目的
  - たまたま手持ちデータで選択されないリスクもある
- 変数選択は恣意的である
  - 取入れ・取出しの基準、適合度の指標によって結果が変わりうる

復習：過剰適合(over-fitting)

- 寄与率R^2(Coefficient of determination)
  - 回帰モデルのデータへの当てはまりを表す指標
    - 1に近づけば近づくほど当てはまりが良い
  - 説明変数の数を増やすと寄与率は上昇する
    - アウトカムとまったく関係のない因子を含めても、寄与率は上昇してしまう
  - 複雑なモデル(二次の項を追加したモデルなど)を用いれば**寄与率は限りなく1に近づけられる**

- 過剰適合したモデル
  - 手持ちデータへの当てはまりが必要以上に高いモデル
  - **外挿性が低く一般化できない**

寄与率のイメージ



$$\text{寄与率}\ (R^2) = \frac{\text{モデルで説明されるバ　ラツキ}}{Y\text{(目的変数)のバラツキ}} = 1 - \frac{\text{残差}}{Y\text{(目的変数)のバラツキ}}$$

残差が小さくなると寄与率が高くなる

過剰適合の例

- 血圧(Y)と年齢(x)の関係をモデル化
  - モデル：$\text{血圧}(Y) = \alpha_0 + \alpha_1 \times \text{年齢}(x) + \text{誤差}$
    - 血圧値は収縮期血圧とする
    - 血圧と年齢の関係を直線で表したモデル
- 血圧と無関係の変数をモデルに含める
  - 0から1の値をとるデタラメ変数(乱数)を生成
  - デタラメなので血圧とはまったく無関係
  - 結果は変わらない？

当てはめるモデル

■デ(d)タラメ変数を10個まで含める

モデル1： $\text{血圧}(Y) = \alpha_0 + \alpha_1 \times \text{年齢}(x) + \beta_1 \times d_1 +$ 誤差

モデル2： $Y = \alpha_0 + \alpha_1 \times x + \beta_1 \times d_1 + \beta_2 \times d_2 +$ 誤差

モデル3： $Y = \alpha_0 + \alpha_1 \times x + \beta_1 \times d_1 + \beta_2 \times d_2 + \beta_3 \times d_3 +$ 誤差

・・・

モデル10：$Y = \alpha_0 + \alpha_1 \times x + \beta_1 \times d_1 + \cdots + \beta_{10} \times d_{10} +$ 誤差

■それぞれのモデルで寄与率を計算する

注：仮想例

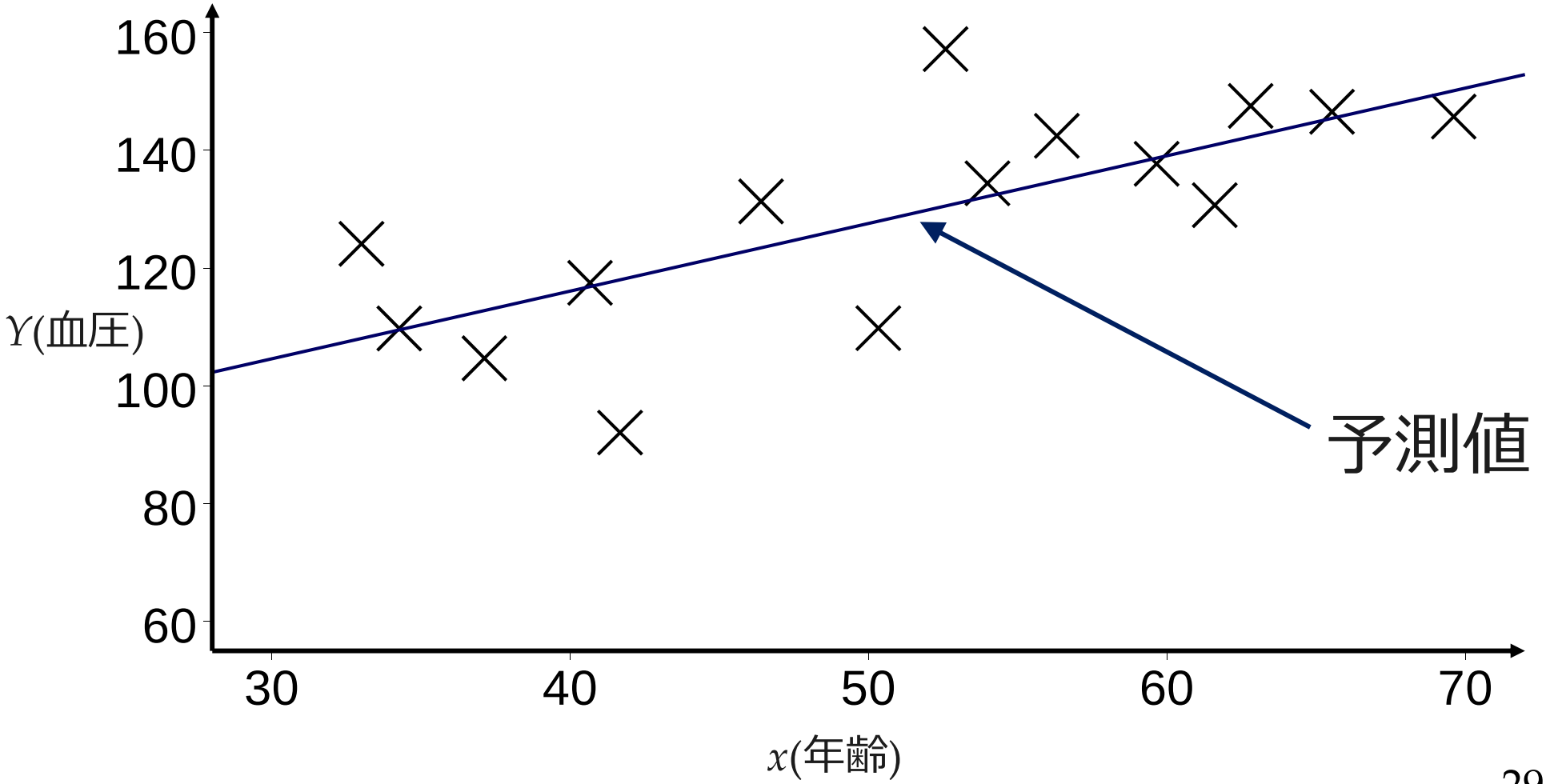
寄与率の変化

- デタラメ変数を含める前：$R^2 = \mathbf{0.525}$
  - 解釈：モデル 血圧$(Y) = \alpha_0 + \alpha_1 \times$ 年齢(x) + 誤差 は、データの52.5%を説明している
- デタラメ変数を含めた後

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

注：仮想例

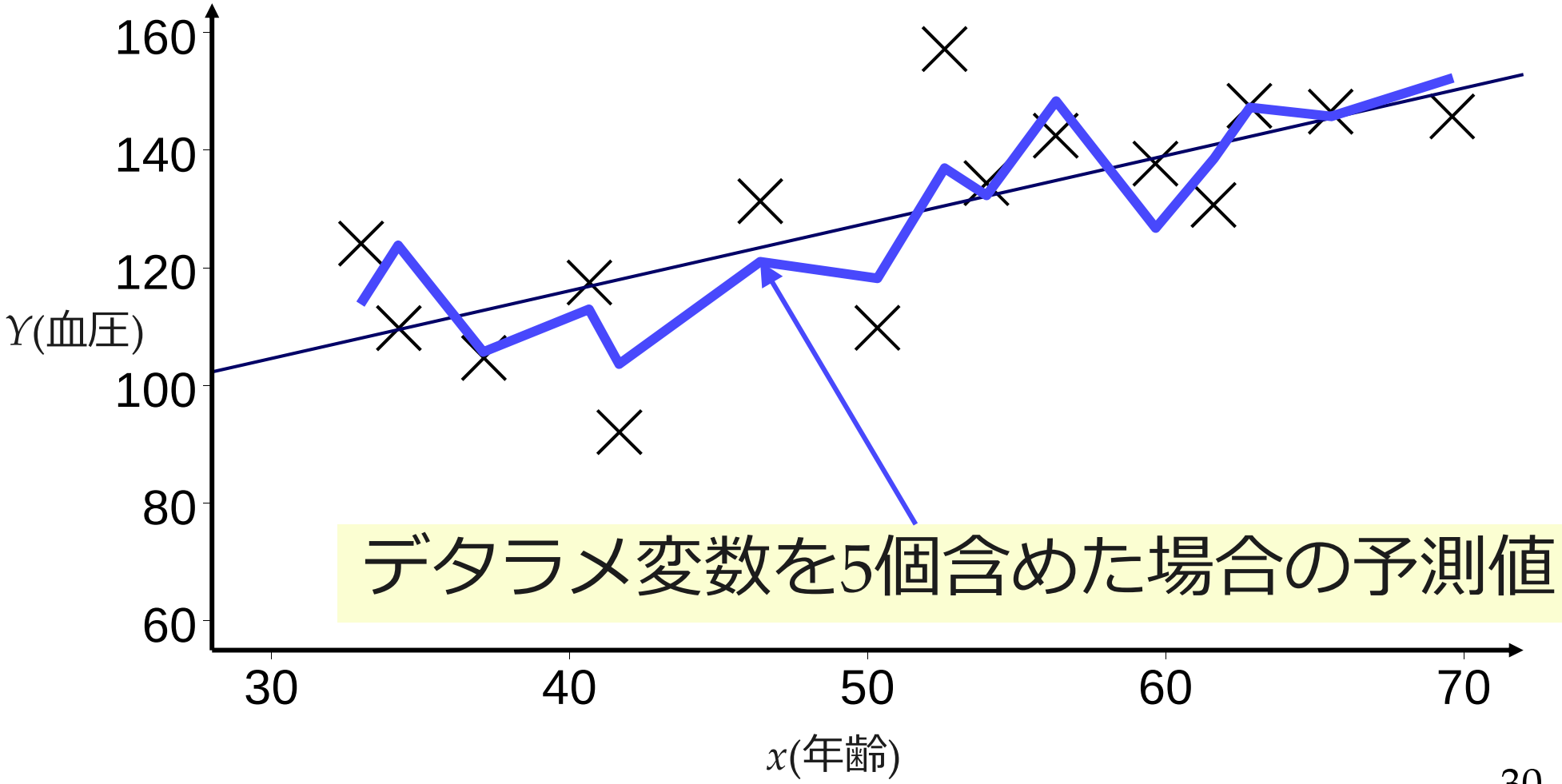
含める前 (R^2=0.525)の予測



注：仮想例

モデル5 (R^2=0.731)の予測

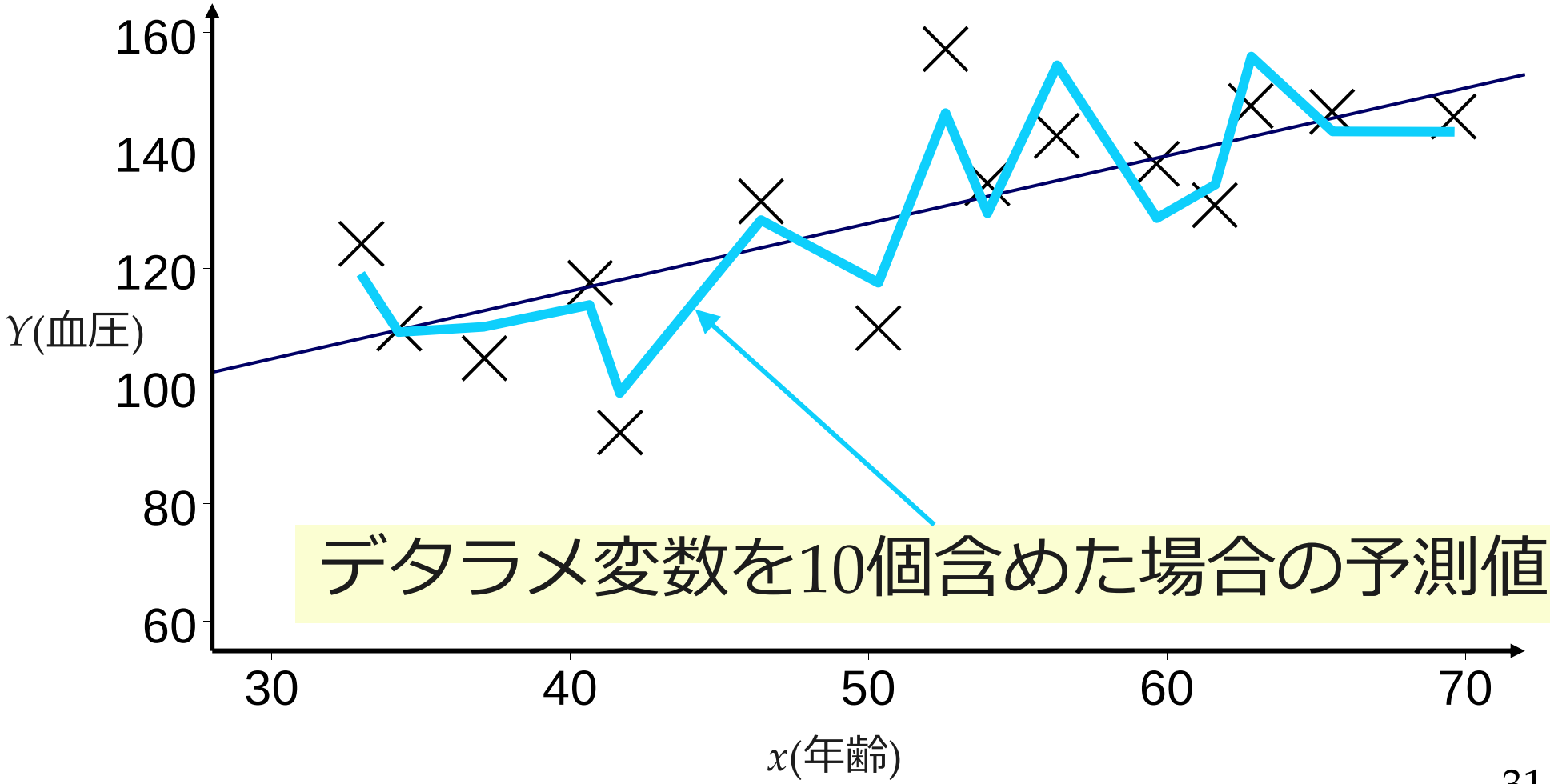
残差が小さくなり、直線から曲線へ・・・



注：仮想例

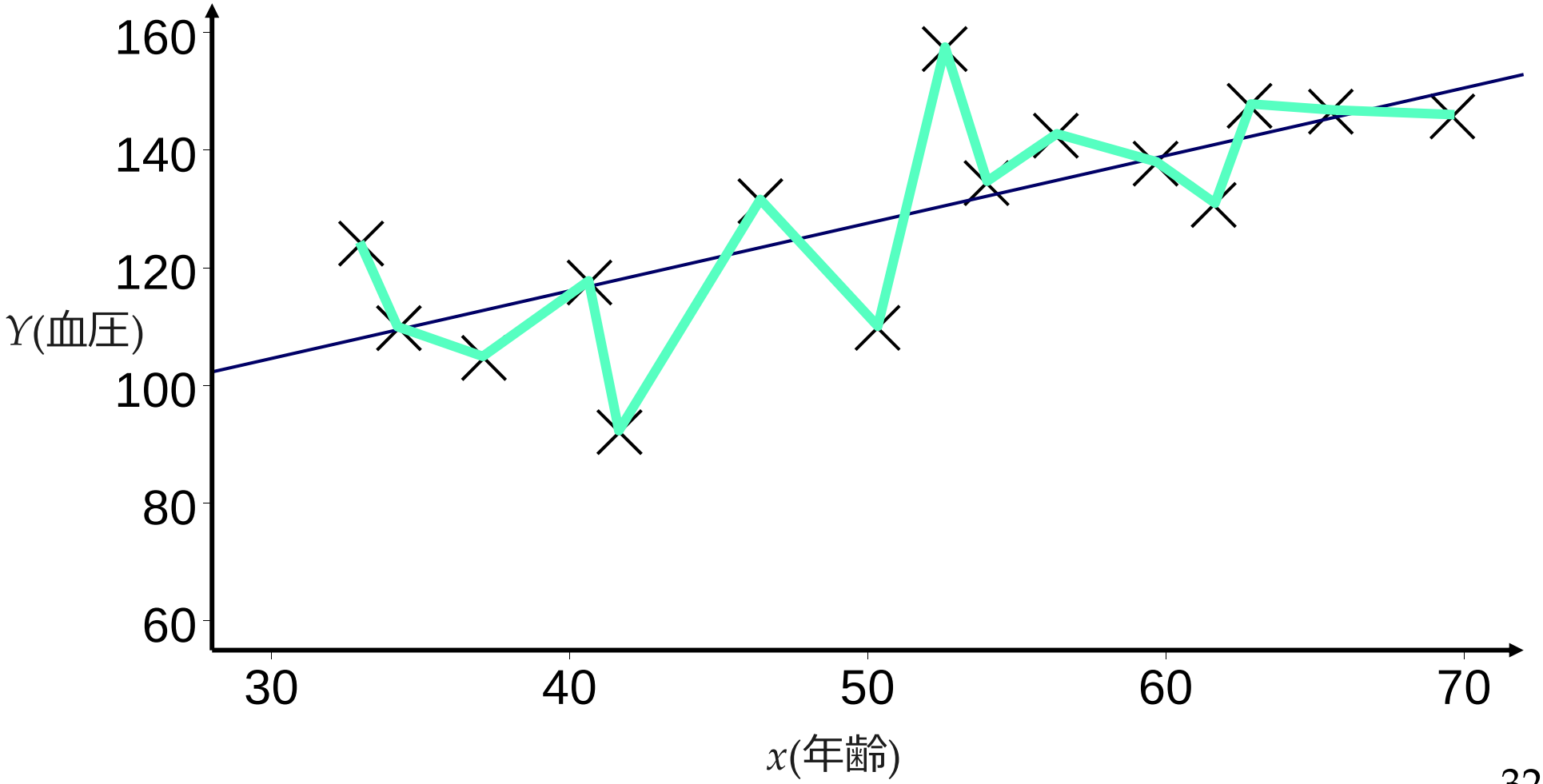
モデル10 (R^2=0.865)の予測

全データを通るような曲線になっていく



より複雑なモデルを構築した場合

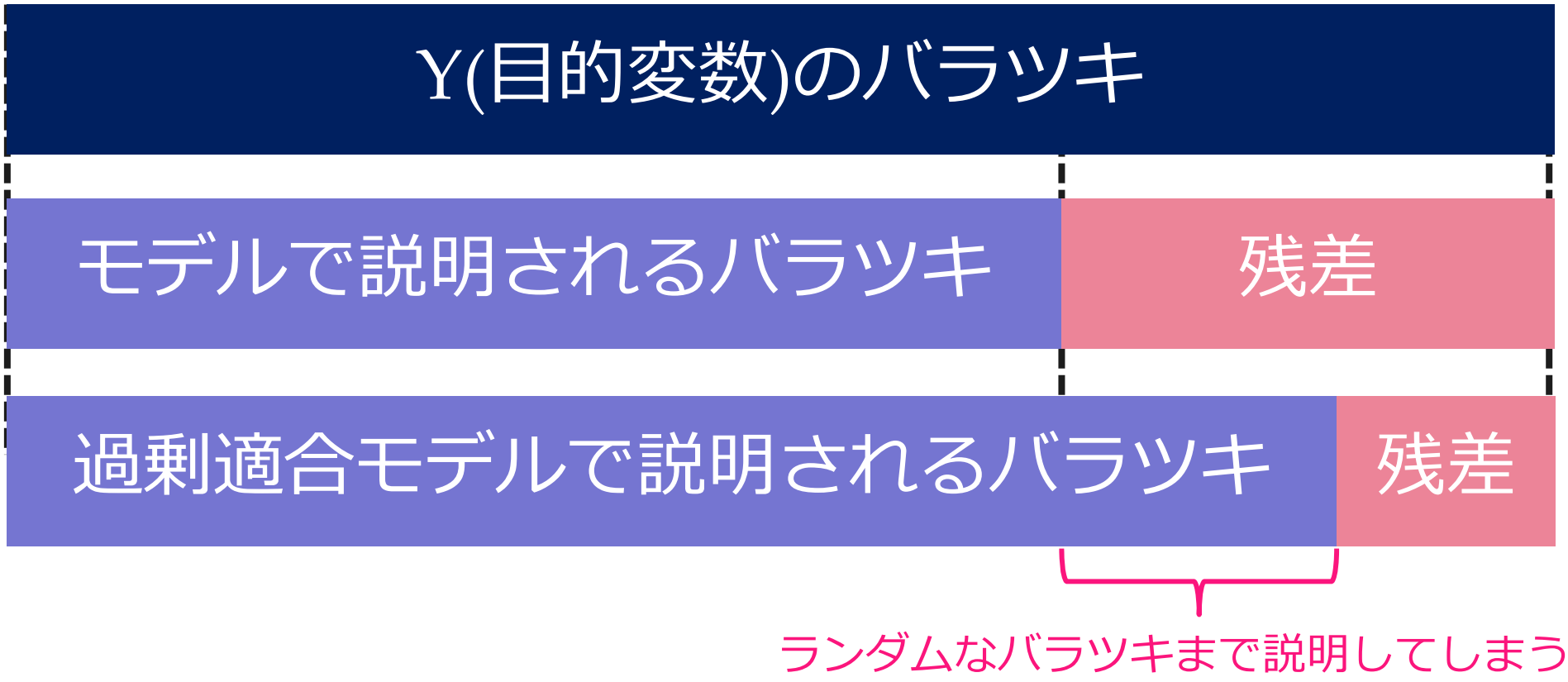
$R^2=1$



なぜ寄与率が向上したのか？

- デタラメ変数を多数含めると言うことは・・
  - 出身地や職業など、被験者を識別する変数をモデルに多数含めることに等しい
  - 全データに適合するように予測できてしまう

- 結果と関連がない変数を含めても、「モデルで説明されるバラツキ」の占める割合は向上してしまう



過剰適合モデルの問題点

$$\text{血圧}(Y) = \alpha_0 + \alpha_1 \times \text{年齢}(x) + \text{誤差}$$

- **過剰適合モデルは手持ちデータにしかない ランダムな誤差まで「説明」してしまう**
- **手持ちデータと同じ誤差を持つデータでない限り、予測結果が再現されない**

OUTLINE

- 第5回セミナーの復習
- 予測のための多変量解析
- バリデーションの必要性

過剰適合への対処

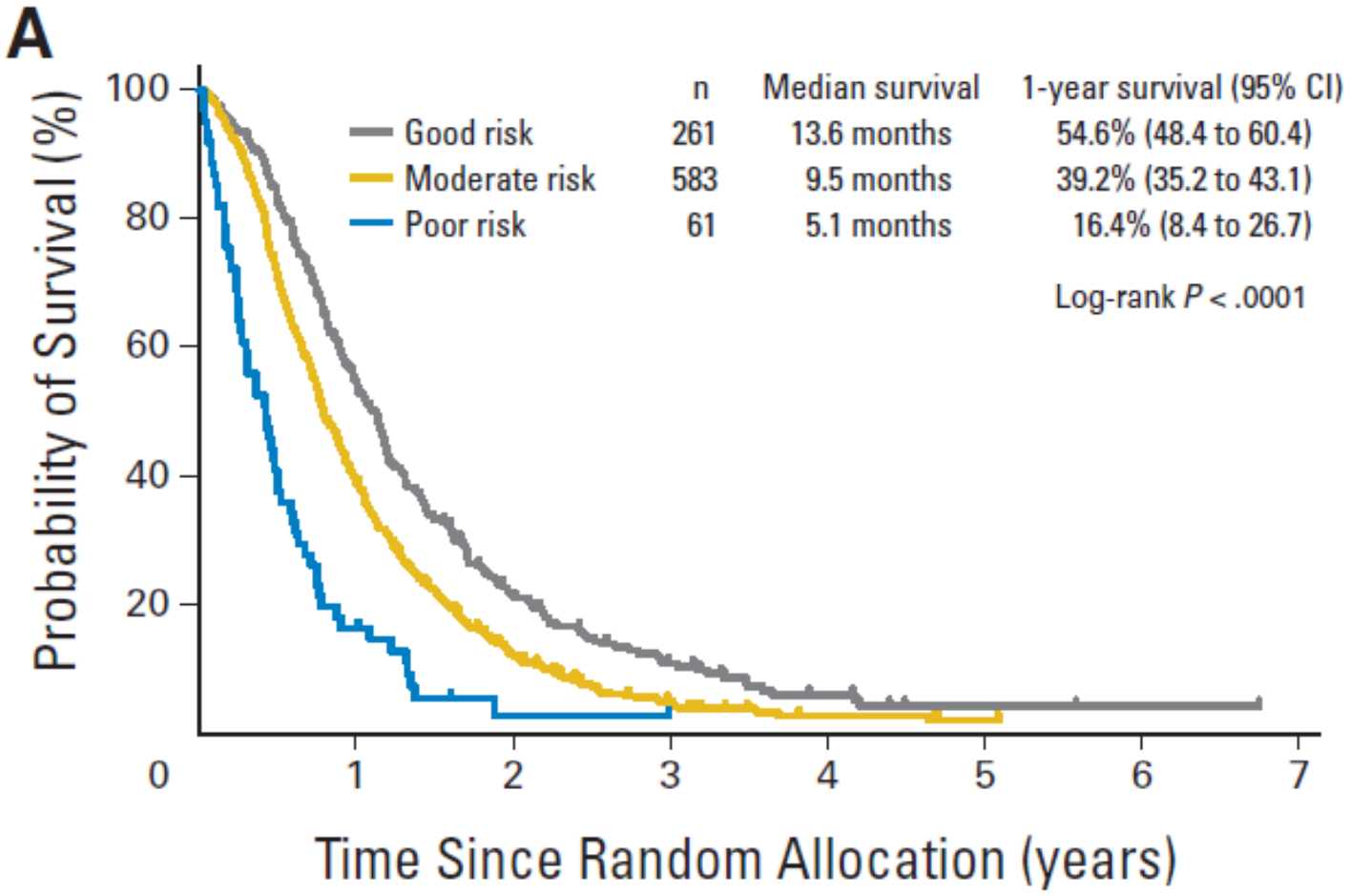
- 将来のデータに適合するためには？
- 構築アルゴリズムを工夫
  - 医学的に説明のつかない変数は含めない
  - 厳しい変数選択基準でスクリーニングする
- モデルの外挿性を検証(validation)
  - 構築アルゴリズムの妥当性を検証する
  - 構築用データ(training data)に対して
    検証用データ(test/validation data)を用意する

Chau, et al. J Clin Oncol 2009.

RMHI*のバリデーション

* RMHI：Royal Marsden Hospital Index

■REAL-2試験による予測結果の再現



Sparano and Paik. J Clin Oncol 2008.

Oncotype DX® のバリデーション

- モデル構築(n=447)
- 外部データのバリデーション
  - 臨床試験のデータ(n=668)

Paik et al. N Engl J Med 2004.

- その他のデータでもバリデーション
  - ケースコントロール研究のデータ(n=790)
    (Habel, et al. Breast Cancer Res 2005)
  - 日本人データ(n=280)
    (Masuda, et al. ASCO Breast Cancer Symposium 2009: #68.
    / Toi, et al. Cancer 2010.)

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

参考：JCO *statistical guidelines*

- *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.*

バリデーションの種類

- Resubstitution(構築データへの適用)は禁忌
- External validation
  - 独立した外部データで構築モデルの一般化可能性を評価
    - 例：RMHIやOncotype DX®のバリデーション
- Internal validation
  - 手持ちデータで構築・評価を同時に実施
    - 構築アルゴリズム自体の妥当性評価
    - 完全な外部データでないため、
      構築モデルの一般化可能性が保証されるとは限らない
      （時期や施設/国が異なるデータでも再現される保証はない）

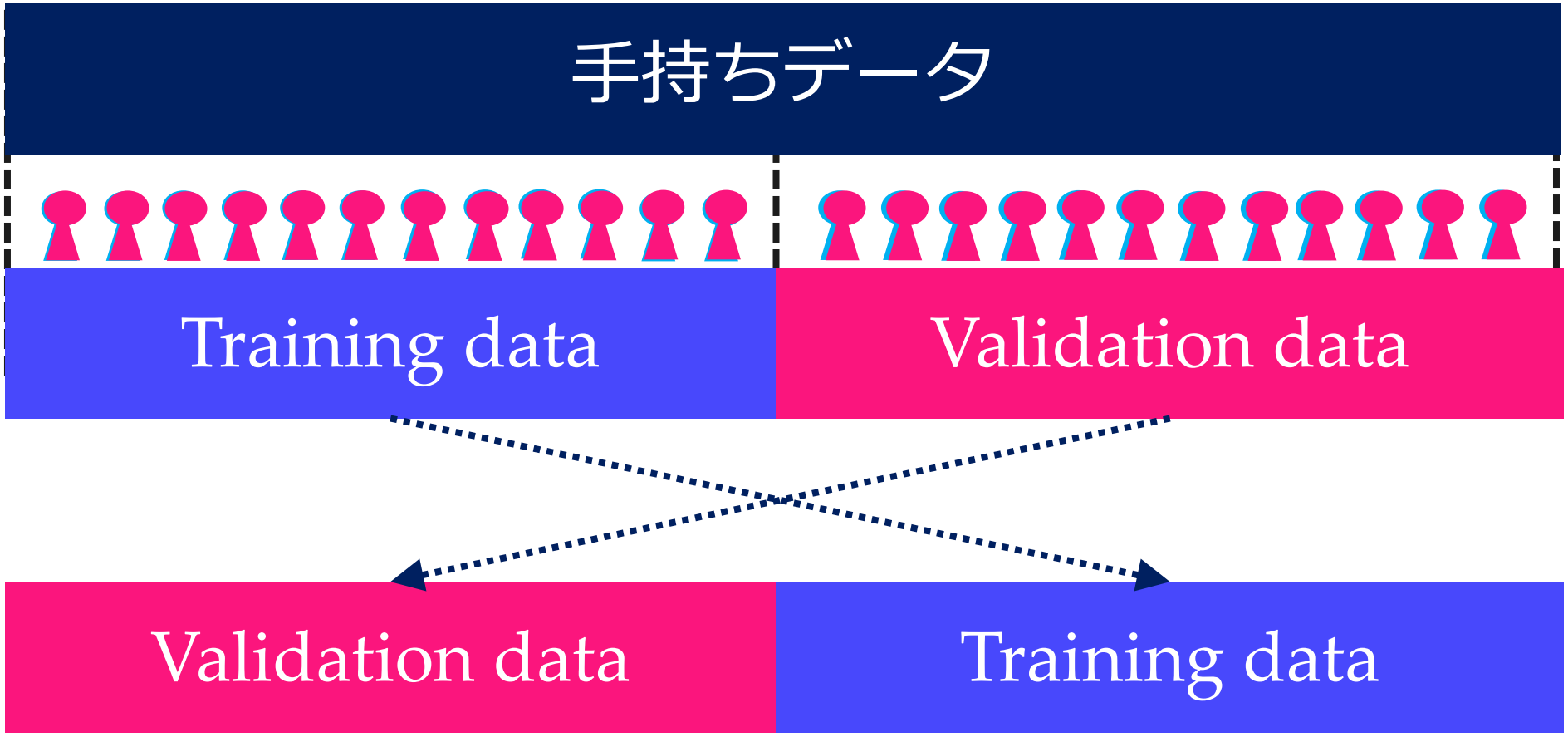
代表的なinternal validation

- Split-group validation
  - 手持ちデータを以下の2つに分割
    - Training data：モデルの構築用データ
    - Validation (Test) data：モデルの検証用データ
  - 十分大きな症例数が必要となる

- Cross-validation
  - 全データをモデルの検証に用いる方法
  - 症例数が少ない場合はSplit-groupより効率が良い

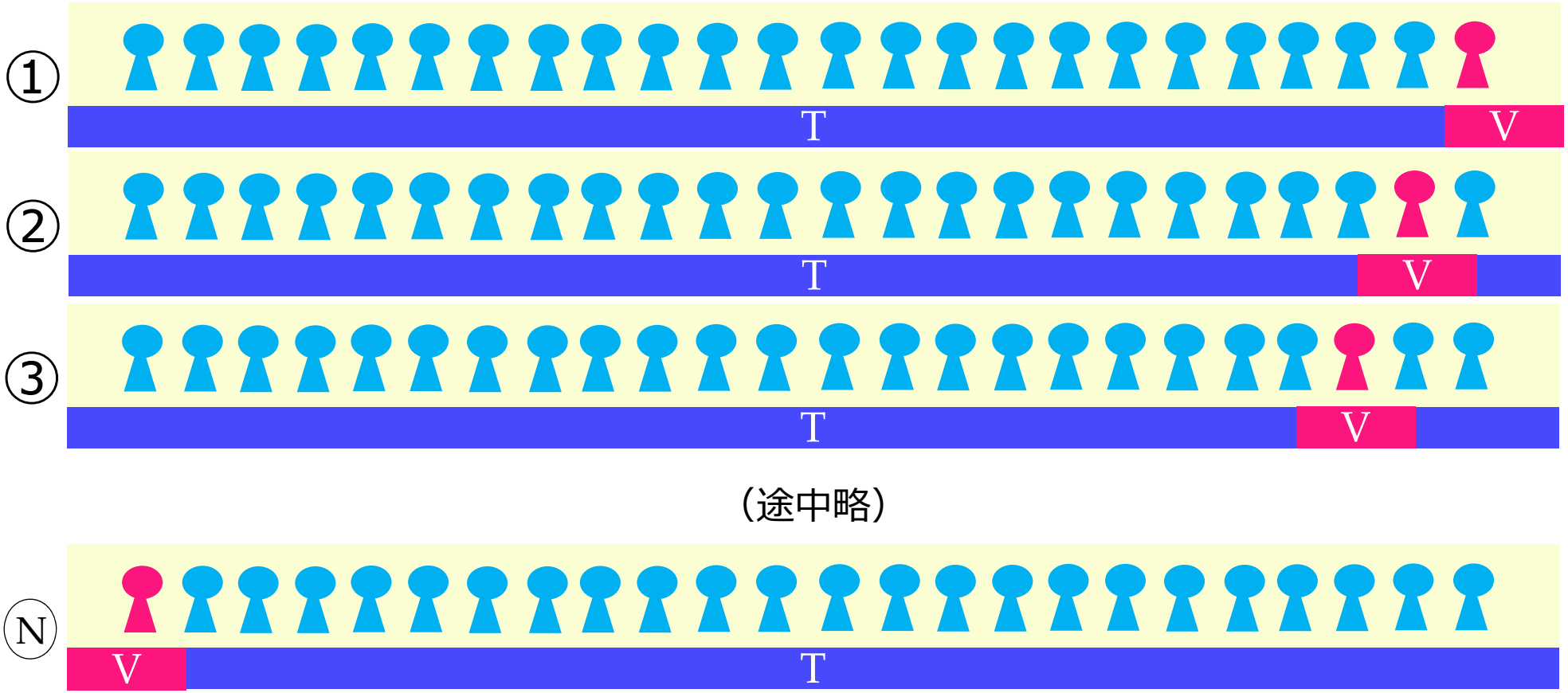
単純なクロスバリデーション

※ データはランダムに分割しても良いし、非ランダムに分割しても良い



Leave-one-out cross-validation (LOOCV)

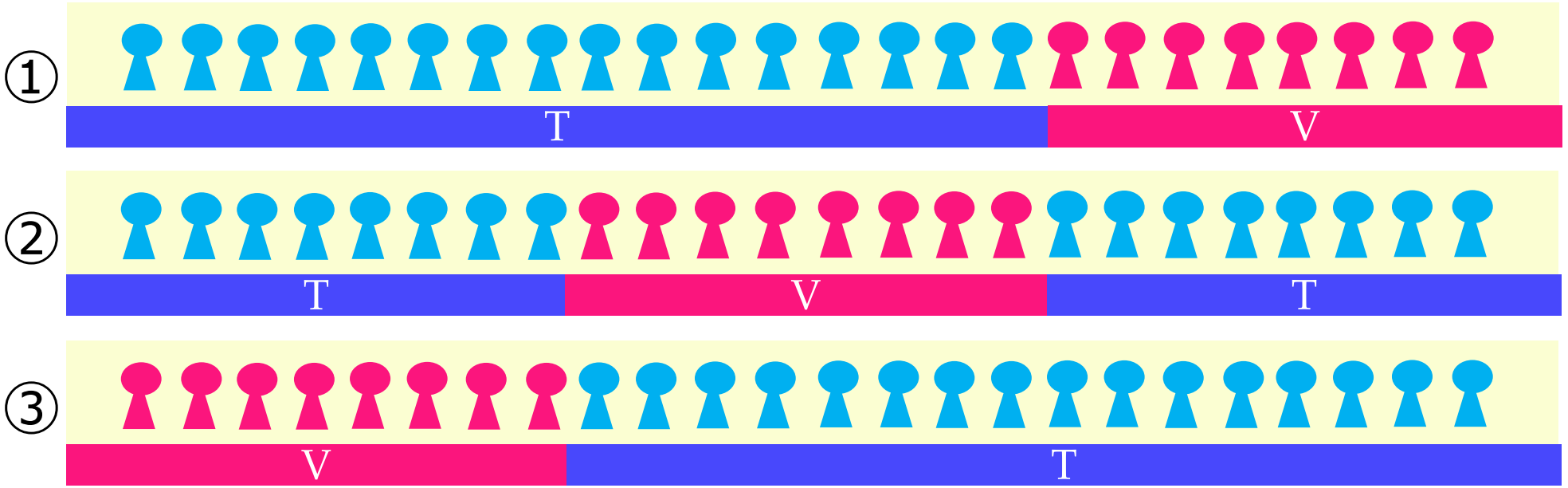
※ N：症例数、T：Training data、V：Validation data



- 計算に時間はかかるがバリデーションをN回(N人分)実施できる
- 予測が外れる割合(エラー割合)を評価できる

K-fold cross validation

※ 例えばK=3の場合（T：Training data、V：Validation data）



- LOOCVより短時間で実行できる
- Kの値が大きければエラー割合を評価できる

まとめ

- 予測を目的とした多変量解析
  - 過剰適合に配慮したモデル構築が必要
  - 構築したモデルはバリデーションが必要
- 構築と検証は独立に実施する必要がある
- 第5・6回セミナーを通して
  - 目的に応じて適切な手法を選択しましょう
  - 迷ったら統計家に相談しましょう

謝辞

- スライドの一部を提供頂きました
  - JCOGデータセンター 水澤純基先生

告知

- 第7回セミナー　「観察研究・レトロ研究1」
  - 4/9(水) 18:30-20:00
  - 築地キャンパス：特別会議室
  - 柏キャンパス：セミナー1,2 （中継）

ご清聴ありがとうございました。
アンケートにご協力下さい。

BACKUP

- 構築と検証を独立に行う理由
  - Simonらによるシミュレーション
- 予測マーカーの臨床応用
  - Onco*type* DX®, MammaPrint®
- 統計モデルによる交互作用検定
  - NSCLCにおけるgefitinibとEGFR遺伝子
  - 大腸癌におけるRAS検査

Simon, et al. J Natl Cancer I 2003.

Simonらによるシミュレーション研究

遺伝子	cancer patients			non-cancer patients		
	1	…	10	1	…	10
1	88.9		77.8	23.1		35.8
2	4.5		15.6	8.6		10.2
:	:					
600	12.4		4.5	34.2		28.9

■デタラメな遺伝子発現量データ(n=20)

– すべての遺伝子で癌患者と非癌患者で差がない

- 癌/非癌かが全く予想できない状況
- つまり、予測が外れとなる割合*は50%になるはず

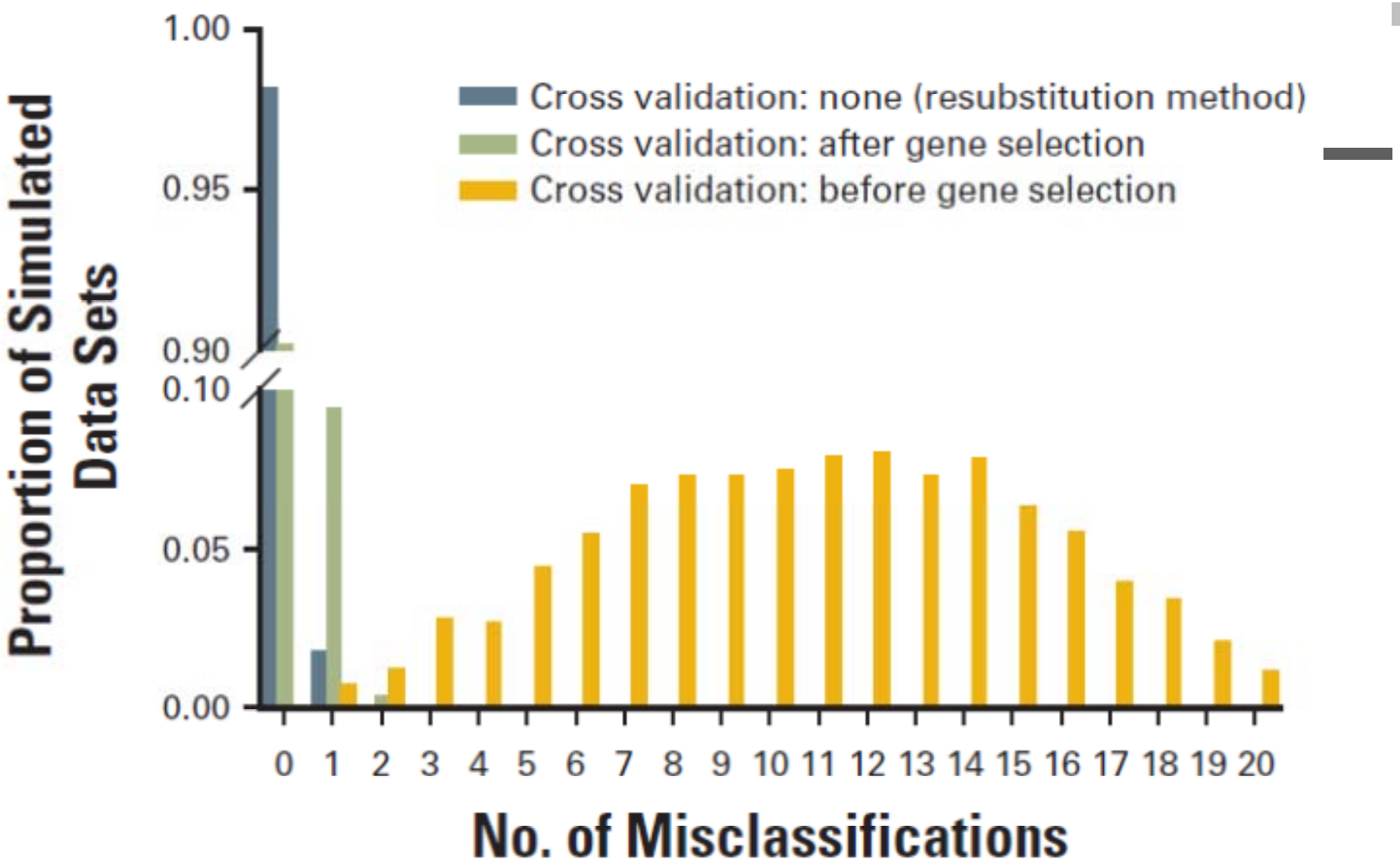
* 以降、「エラー割合」と呼ぶ

Simon, et al. J Natl Cancer I 2003.

比較した方法

- Resubstitution
  - 20例(全体)で10遺伝子をスクリーニング→予測スコア作成
  - 同じデータに当てはめてエラー割合を計算
- Cross-validation after gene selection
  - 20例で10遺伝子をスクリーニング→予測スコア作成
  - (19例,1例)という組合わせを20パターン作成
  - 1例のvalidationを20回繰り返してエラー割合を計算
- Cross-validation prior to gene selection
  - (19例,1例)という組合わせを20パターン作成
  - 19例で10遺伝子をスクリーニング→予測スコア作成
  - 残りの1例でvalidationする作業を20回繰り返して
    エラー割合を計算

Simon, et al. J Natl Cancer I 2003. Simon R. J. Clin Oncol 2006.



構築・検証を独立に行わないと正しく予測できない