生物統計基礎セミナー 第5回

2014年3月6日

# 多変量解析 1

国立がん研究センター 生物統計部門

野村 尚吾

# OUTLINE

- 第4回セミナーの復習
- （統計）モデルとは？
- 交絡調整のための多変量解析
- 回帰モデルを用いる上での注意点
- 多変量解析の用途

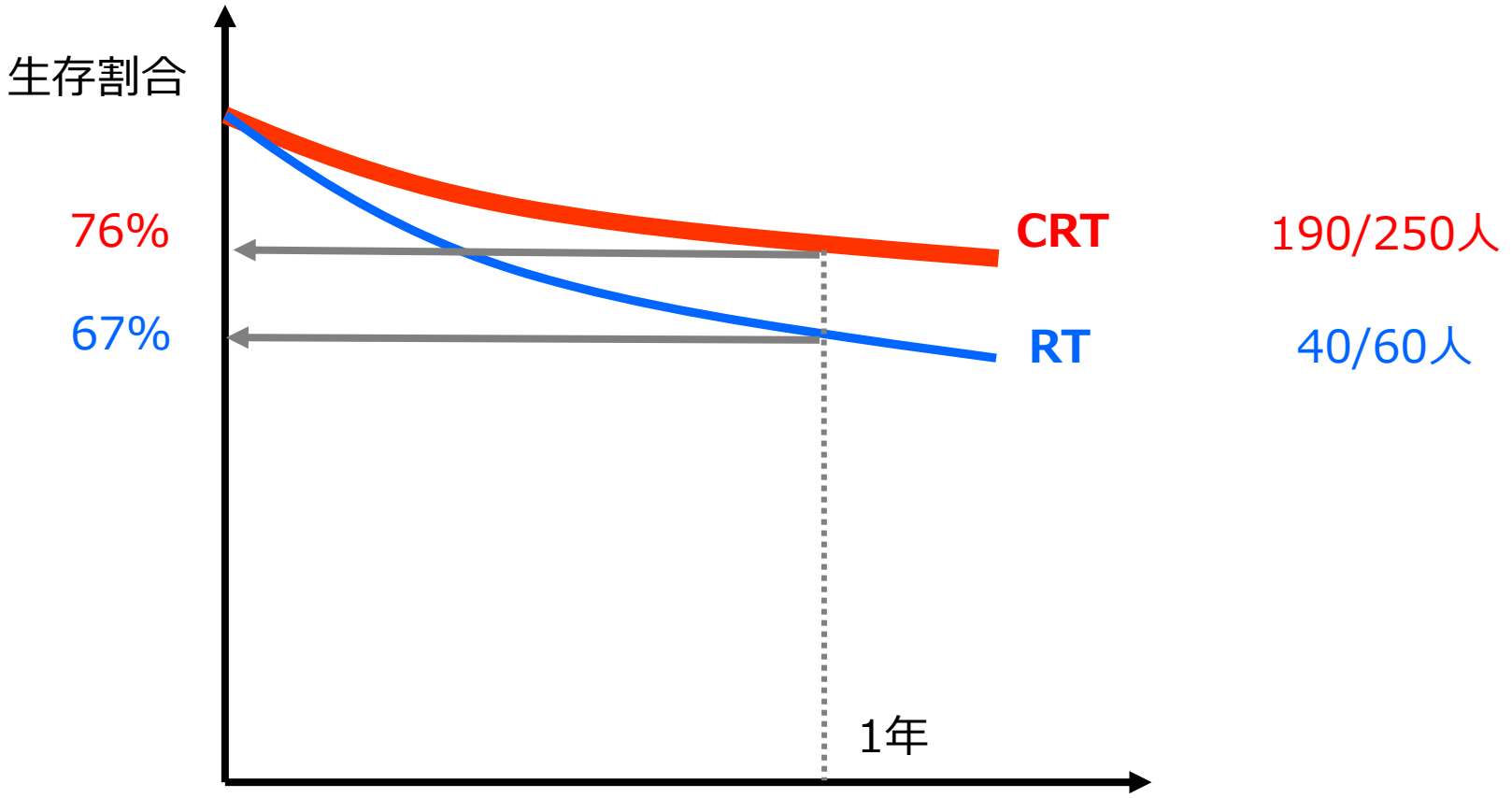
# 第4回セミナーの復習

- 「交絡」とは？
- ランダム化とは？
- 交絡を除去するための方法
  - デザイン段階での工夫：ランダム化
  - 解析段階での工夫

# 仮想例

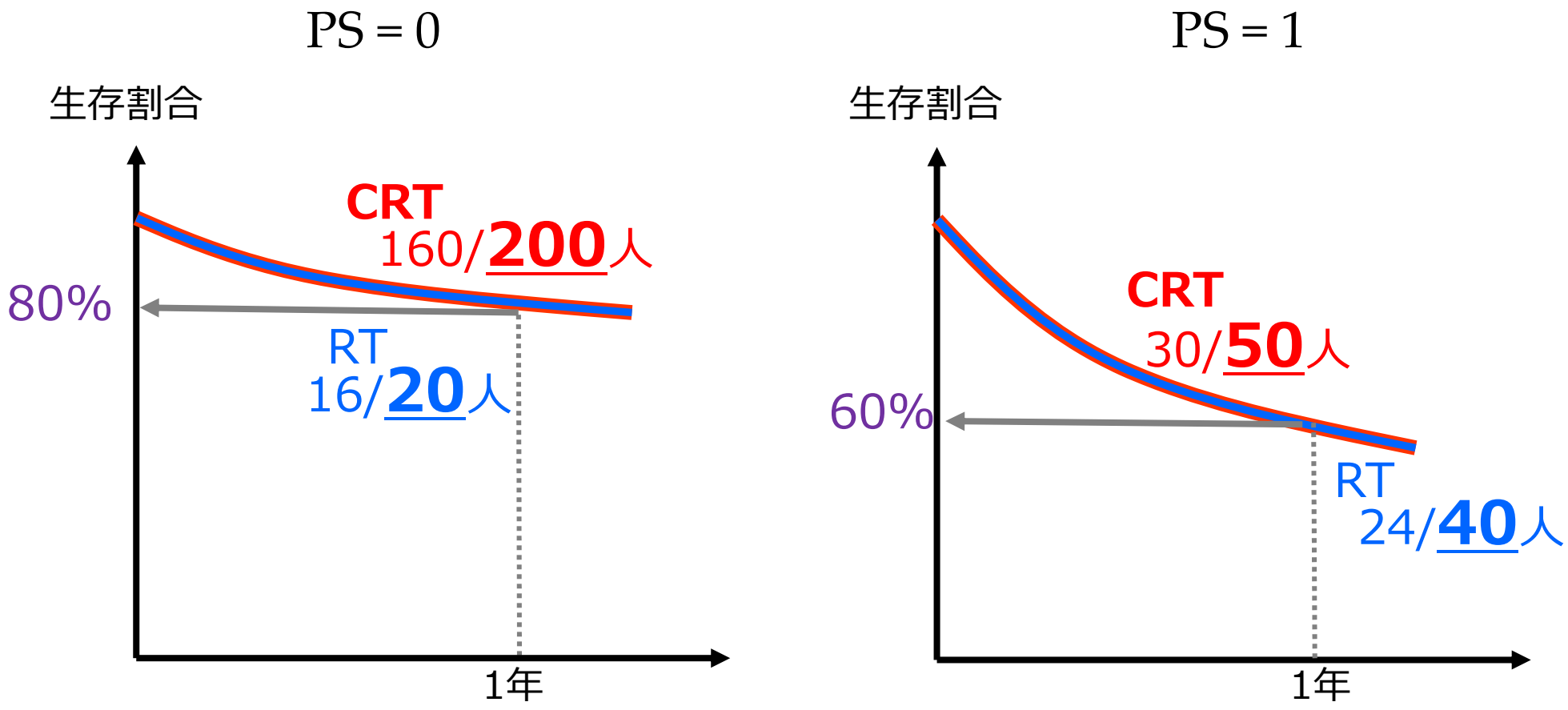
**CRT ：化学放射線療法**
**RT ：放射線療法**

- **CRT群**(250例)は**RT群**(60例) と比較し予後良好であった。
- この対象に**CRT**をすることが推奨される？



注：仮想例

# PSで分けた場合の予後




PSによらずCRTとRTの予後は変わらない

注：仮想例

# なぜCRTが上回ったのか？

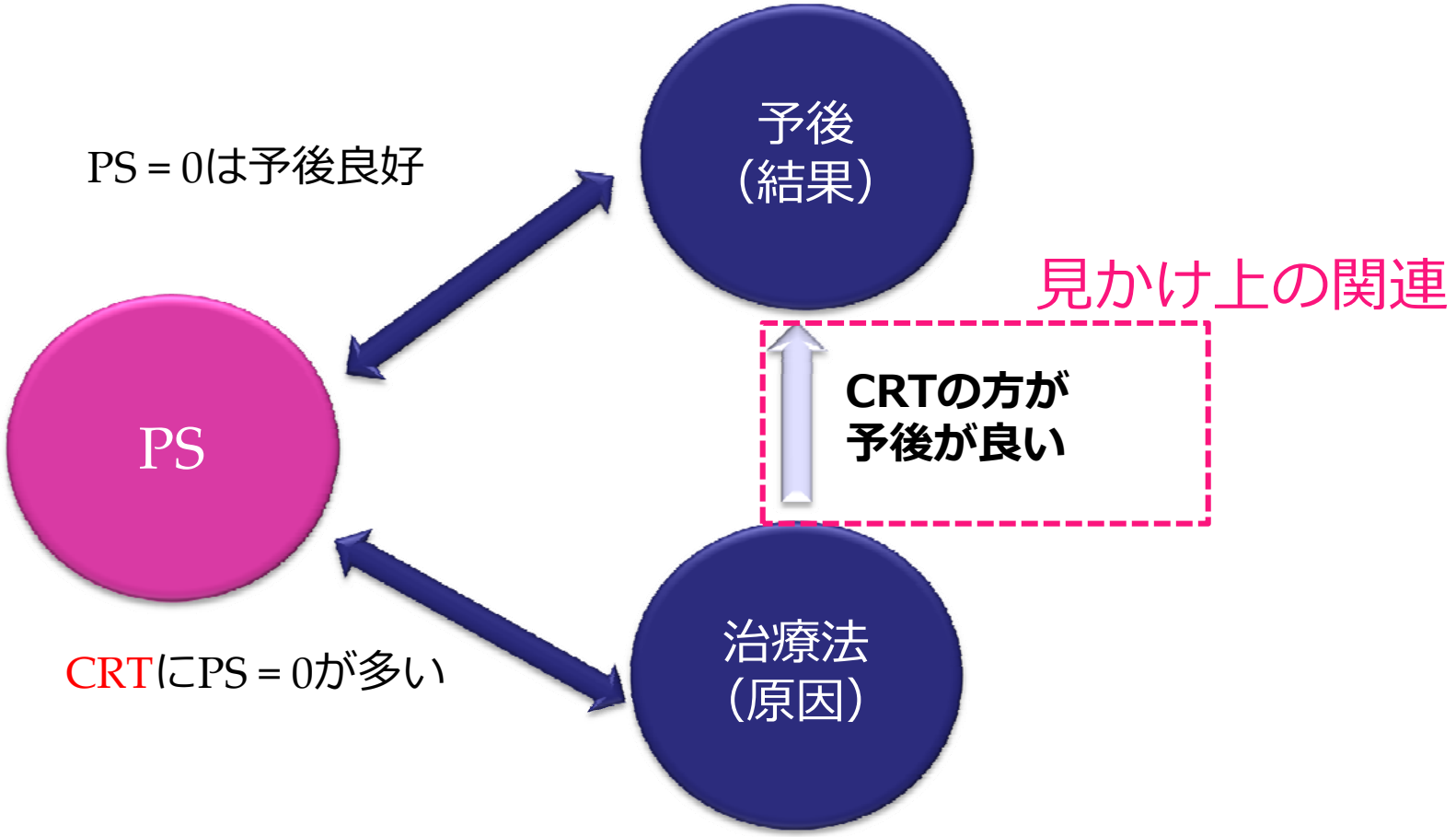
| 治療 | PS = 0 | | PS = 1 | 合計 |
|---|---|---|---|---|
| CRT | **200人 (80%)** | ＞＞ | 50人 | 250人 |
| RT | 20人 (33.3%) | ＜＜ | **40人** | 60人 |

- CRTはPS = 0の症例が多い
- （一般的に）PS = 0だと予後が良い
- 治療法以外の予後に影響する要因の条件が同じでなければ、”比較”にならない！！

# 交絡とは？

- 原因と結果に関連する第3の因子によって見かけ上の関連が生じてしまう現象のこと
  - 交絡を引き起こす因子を交絡因子という

- 仮想例の交絡因子 = PS

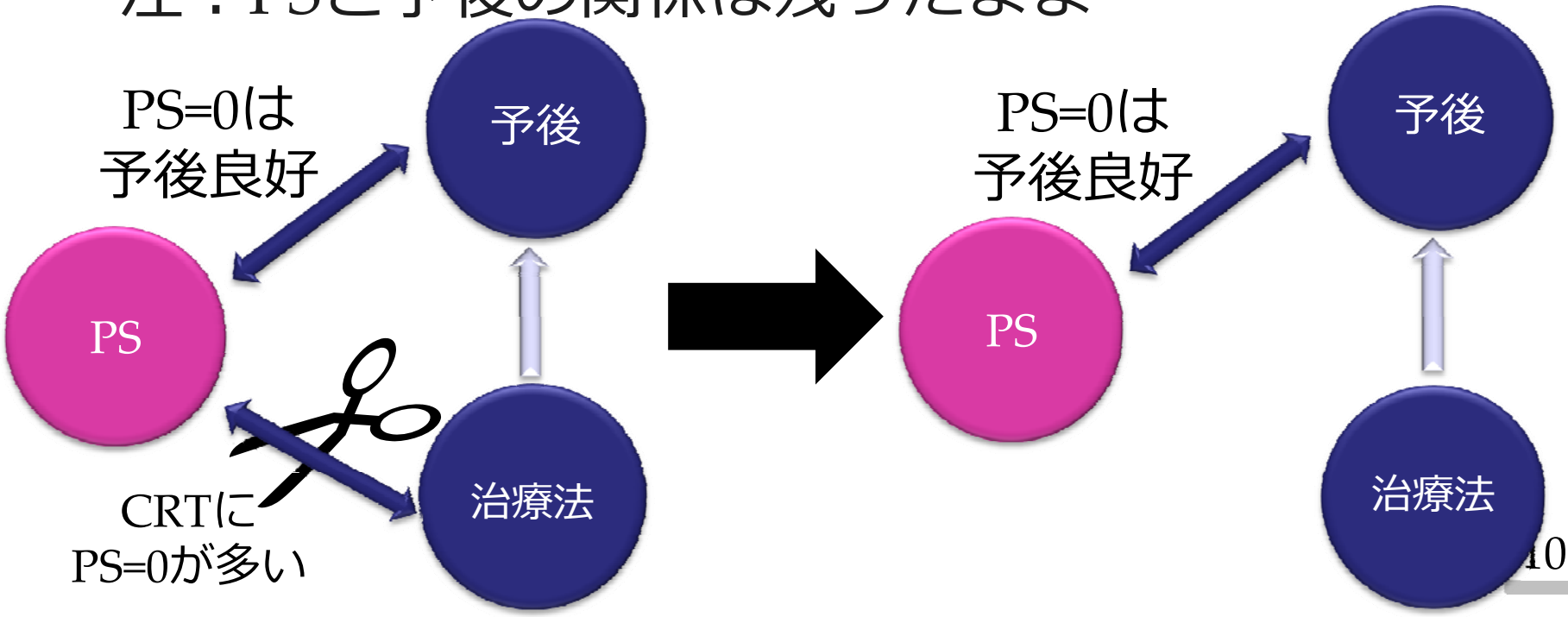


# デザイン段階で交絡を除去する方法

- ランダム化
  - 医師あるいは患者の意思によらず、確率に基づいて各治療群に患者を割り付ける
  - 治療法以外は等しい集団
    → 効果に差があれば治療法の違い

# ランダム化の意義

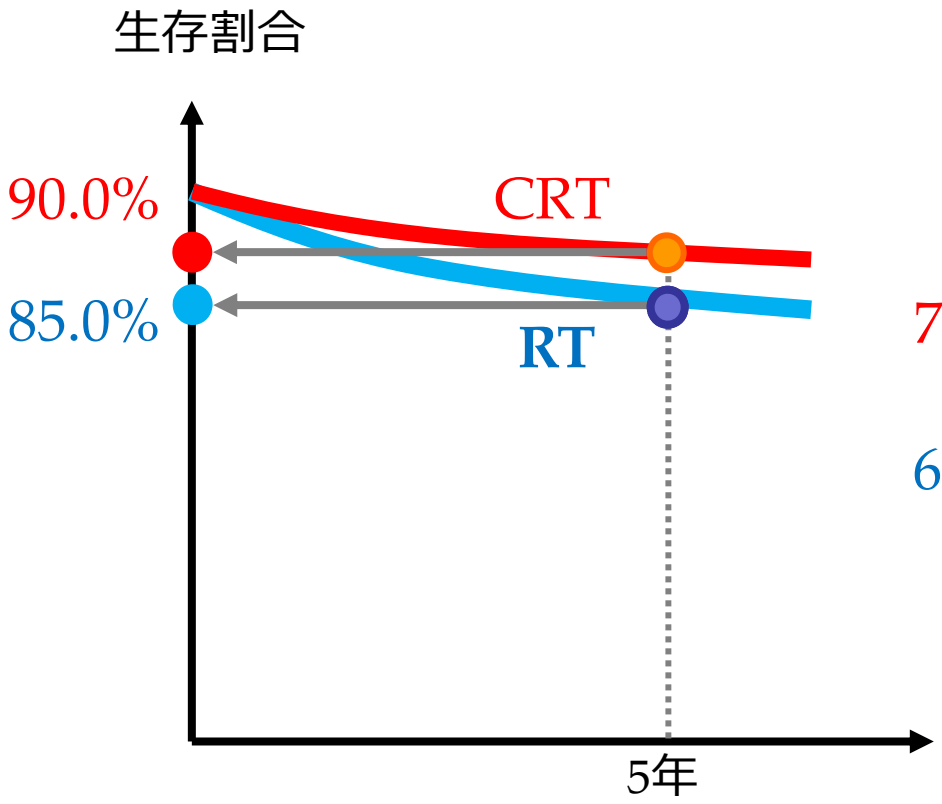
- 治療法とPSの関連を除去できる
  - PSによる交絡は除去されるので、治療法と予後の関係を評価できる
  - 注：PSと予後の関係は残ったまま



# 層別解析

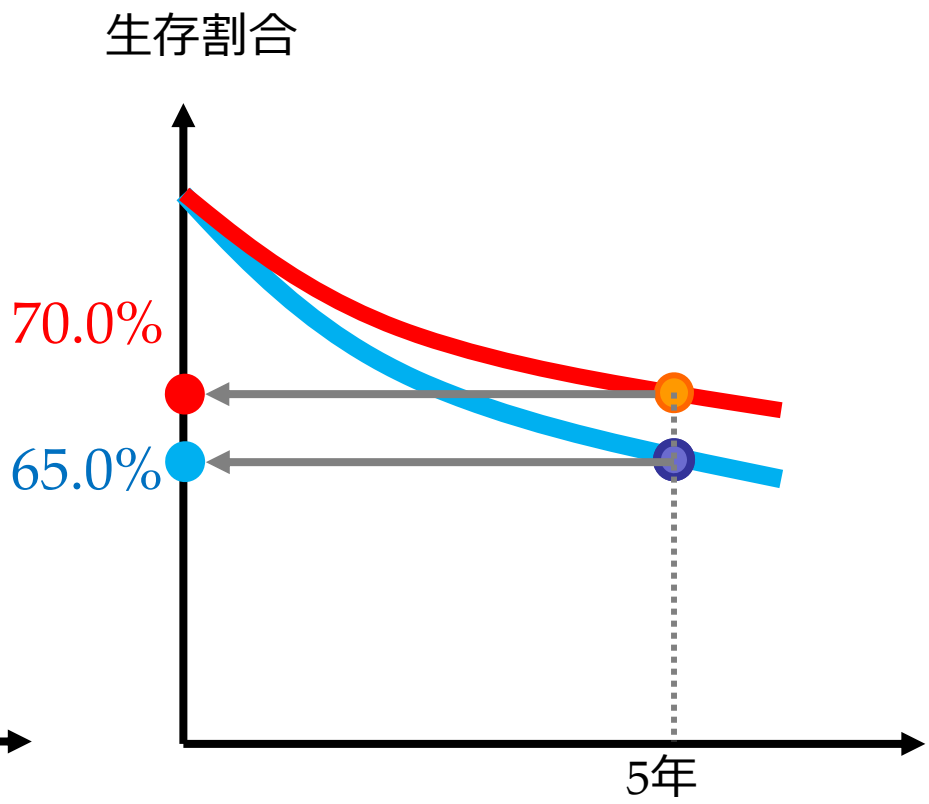
PS 0/1で治療効果が共通であるという仮定の下、治療効果を併合する解析

PS 0 の症例




PS0における HR=0.65

PS 1 の症例




PS1における HR=0.83

**統合したHR=0.78**

# 層別解析の利点と欠点

- 利点
  - 集団全体の治療効果を求めることができる
  - (モデルを用いた解析と比較して) 仮定が少ない
- 欠点
  - サブグループが多すぎると、
    各サブグループのサンプルサイズが小さくなりすぎる
    - 交絡因子が5個あれば、少なくとも $2^5$=32個のサブグループ
    - 交絡因子が連続量の場合は、カテゴリー化してからでなければ
      サブグループ解析できない
  - 交絡因子自体の効果の大きさがわからない
    - PS 0に対するPS 1の予後への影響を評価することができない

# 交絡への対処法

- デザイン段階
  - ランダム化(randomization)
  - マッチング(matching)

- 解析段階
  - サブグループ解析(subgroup analysis)
  - 層別解析(stratified analysis)
  - **モデルを用いた多変量解析(multivariate analysis)**
    - **ロジスティック回帰やCox回帰など**

# OUTLINE

- 第4回セミナーの復習
- （統計）モデルとは？
- 交絡調整のための多変量解析
- 回帰モデルを用いる上での注意点
- 多変量解析の用途

# 仮想例

- 賃貸物件の家賃を数式で表せないか？
  - おそらく関連している因子
    - 階数、広さなど
    - 家賃がこれらの重み付き和で表現できるとする

$$\text{家賃} = \text{地域の相場家賃} + 1.5 \times \text{階数} + 2.5 \times \text{広さ}$$

- この数式はすべての物件に成立しない
  - バラツキが存在するから

# 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 = $\beta_0$ + $\beta_1$ × 説明変数1 + $\beta_2$ × 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

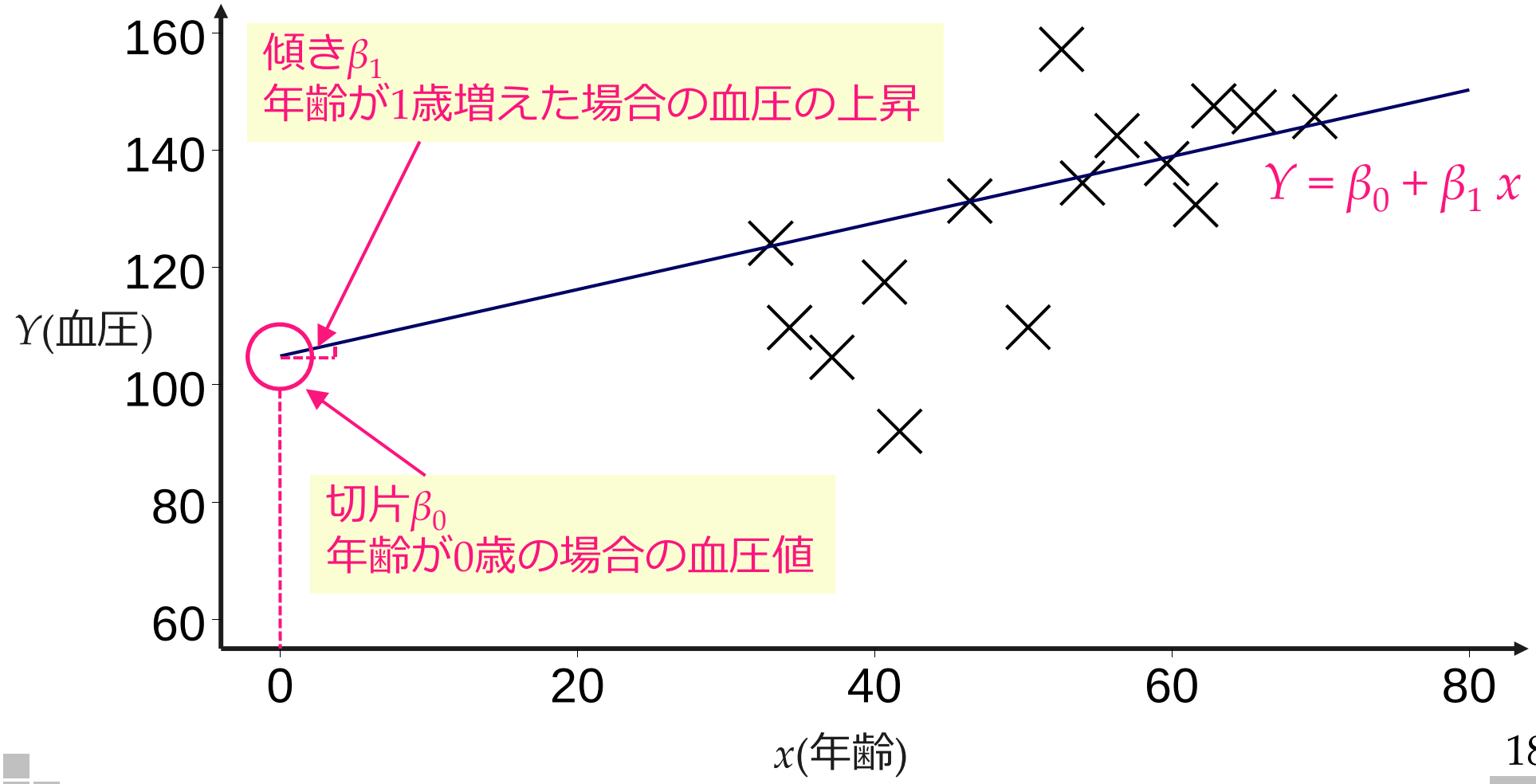
# 単回帰モデル

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

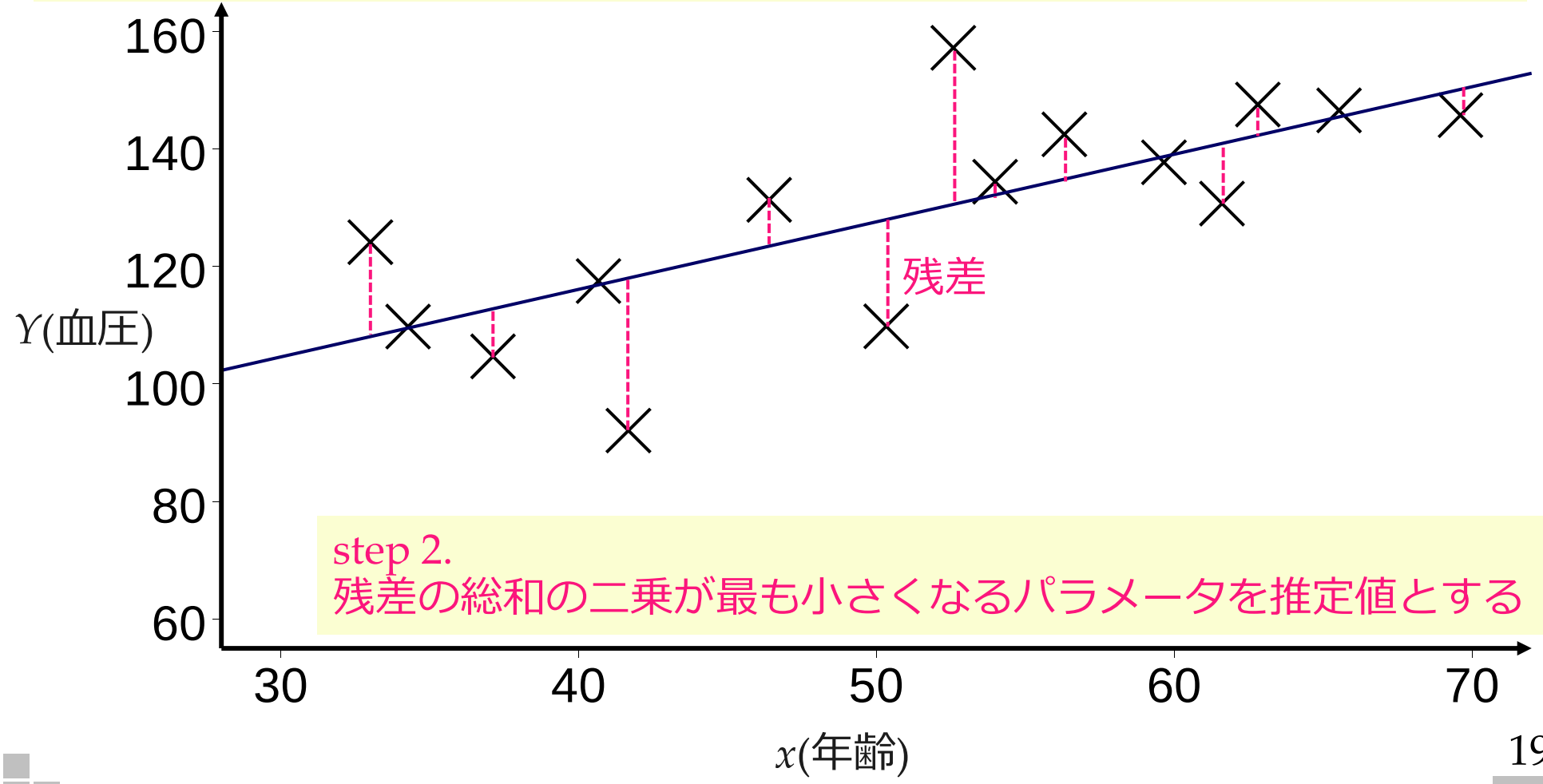
# 散布図(n=15)に直線を当てはめ



注：仮想例

# パラメータはどうやって決める？

step 1.
パラメータに適当な値を代入して残差(=予測値 - 実測値)の二乗の総和を計算




step 2.
残差の総和の二乗が最も小さくなるパラメータを推定値とする

# ロジスティック回帰モデル

■2値変数のアウトカムに対する統計モデル

– 例えば奏効割合に関心があるとする

■ロジスティック回帰モデル（説明変数が1個の場合）

$p$ : 奏効割合

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

対数オッズ　　切片　傾き　線形式

# オッズ比とは

＝ オッズという（イベント発生数 ÷ イベント非発生数） の比

| 治療群 | 奏効あり | 奏効なし | 合計 |
|---|---|---|---|
| 標準 | 20 | 80 | 100 |
| 試験 | 40 | 60 | 100 |
| 合計 | 60 | 140 | 200 |

標準群のオッズ $= \frac{20}{100} \Big/ \frac{80}{100} = \frac{20}{80} = \frac{1}{4}$

試験群のオッズ $= \frac{40}{100} \Big/ \frac{60}{100} = \frac{40}{60} = \frac{2}{3}$

$$\text{オッズ比} = \frac{2}{3} \Big/ \frac{1}{4} = \frac{8}{3} \approx 2.67$$

**試験群のオッズは標準群のオッズの2.67倍**

# ロジスティック回帰モデルとオッズ比の関係

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

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x \qquad x = \begin{cases} 0 & \text{標準群} \\ 1 & \text{試験群} \end{cases}$$

$$\log(\text{試験群オッズ}) = \beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1$$

$$\log(\text{標準群オッズ}) = \beta_0 + \beta_1 \times 0 = \beta_0$$

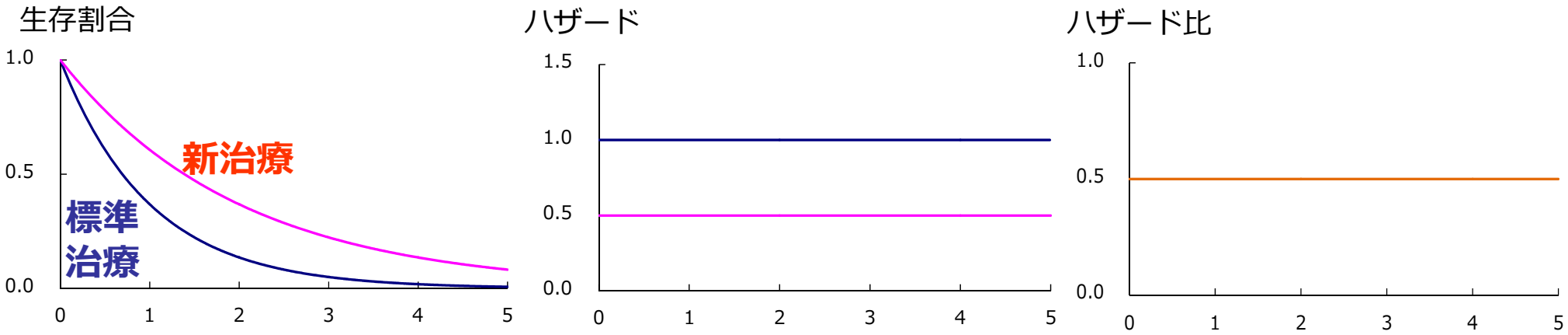
$$\text{対数オッズ比} = \log\left(\frac{\text{試験群オッズ}}{\text{標準群オッズ}}\right) = \log(\text{試験群オッズ}) - \log(\text{標準群オッズ}) = \beta_1$$

$$\text{オッズ比} = \exp(\beta_1)$$

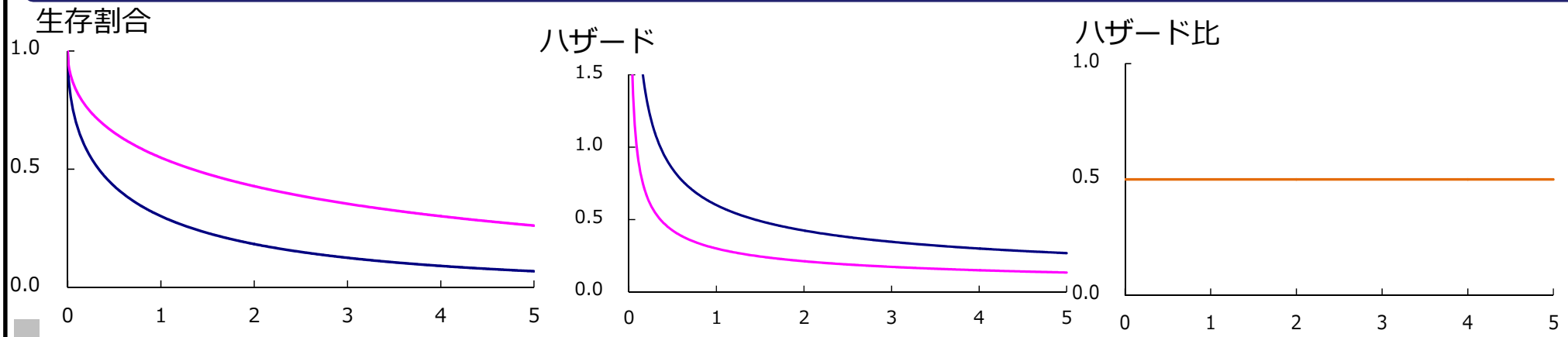
※ exp(β)とはe^βのことです

# 復習：比例ハザード性

## 指数分布：時点を通してハザードが一定。両群指数分布ならHRは一定




## ワイブル分布：時点を通してハザードが変化。両群同じ形状のワイブル分布ならHRは一定



# Coxの比例ハザードモデル

- 生存時間アウトカムに対する統計モデル
  - 例えば全生存期間に関心があるとする

- Coxの比例ハザードモデル（説明変数が1個の場合）

$h_0(t)$：ベースラインのハザード(時間の関数として表現したもの)

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

イベントを死亡としたハザード関数

$\log\left\{\frac{h(t)}{h_0(t)}\right\} = \beta x$ とすればロジスティック回帰モデルと構造は同じ

# Cox比例ハザードモデルとハザード比の関係

$$h(t) = h_0(t) \times \exp(\beta x) \quad x = \begin{cases} 0 & \text{標準群} \\ 1 & \text{試験群} \end{cases}$$

$$\text{標準群のハザード} = h_0(t) \times \exp(\beta \times 0) = h_0(t)$$

$$\text{試験群のハザード} = h_0(t) \times \exp(\beta \times 1) = h_0(t) \times \exp(\beta)$$

$$\text{ハザード比} = \frac{\text{試験群のハザード}}{\text{標準群のハザード}} = \frac{\cancel{h_0(t)} \times \exp(\beta)}{\cancel{h_0(t)}} = \exp(\beta)$$

# 統計モデル ～まとめ～

- バラツキを考慮した“線形式”

  目的変数 = $\beta_0$ + $\beta_1$ × 説明変数1 + $\beta_2$ × 説明変数2 + … + 誤差

  - 単変量モデルのときは直線(切片：$\beta_0$、傾き： $\beta_1$)の関係を仮定する

- ロジスティック回帰モデル
  - 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
  - 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4回セミナーの復習
- （統計）モデルとは？
- 交絡調整のための多変量解析
- 回帰モデルを用いる上での注意点
- 多変量解析の用途

# 仮想例

| | 5年生存 | 死亡 | 合計 |
|---|---|---|---|
| 放射線 | 47(78.3%) | 13 | 60 |
| 手術 | 185(74.0%) | 65 | 250 |

| | 5年生存 | 死亡 | 合計 |
|---|---|---|---|
| 放射線 | 34(85%) | 6 | 40 |
| 手術 | 45(90%) | 5 | 50 |

PS 0 の症例

| | 5年生存 | 死亡 | 合計 |
|---|---|---|---|
| 放射線 | 13(65%) | 7 | 20 |
| 手術 | 140(70%) | 60 | 200 |

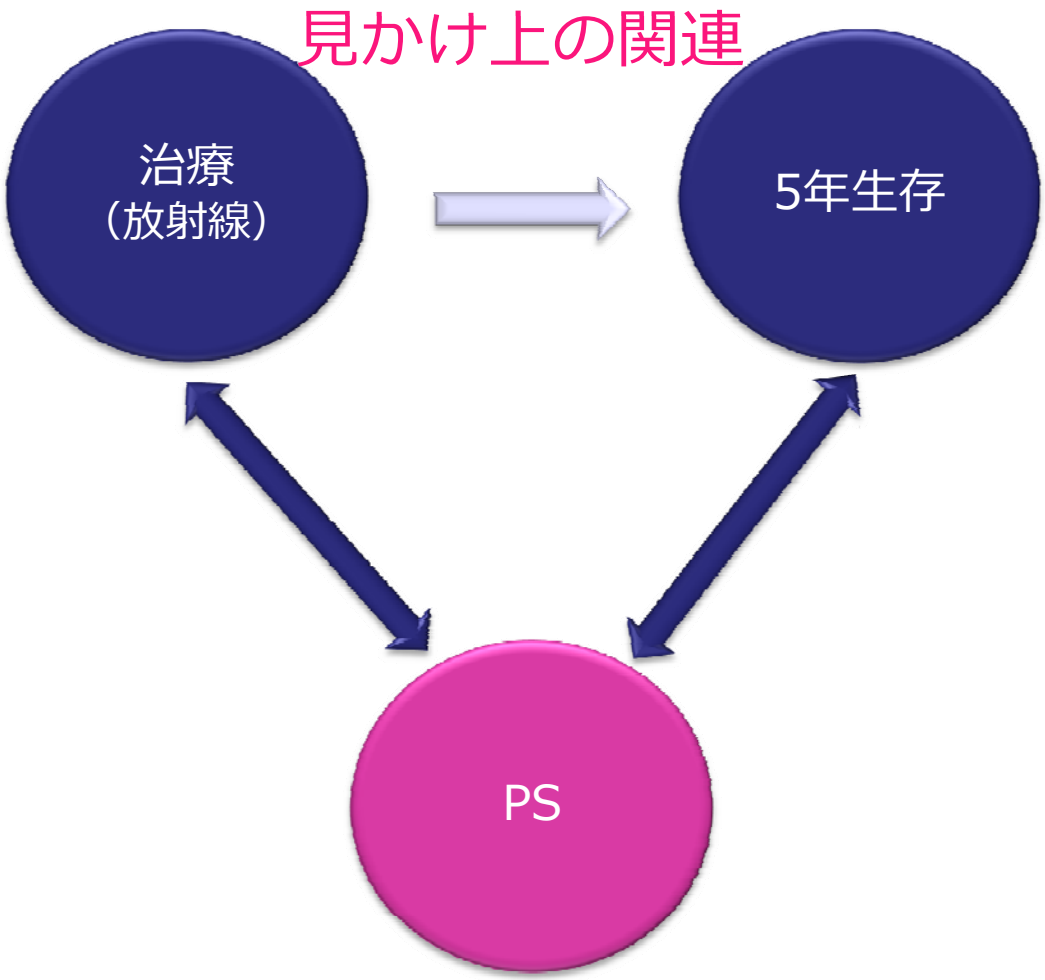
PS 1 の症例

# 交絡の再復習

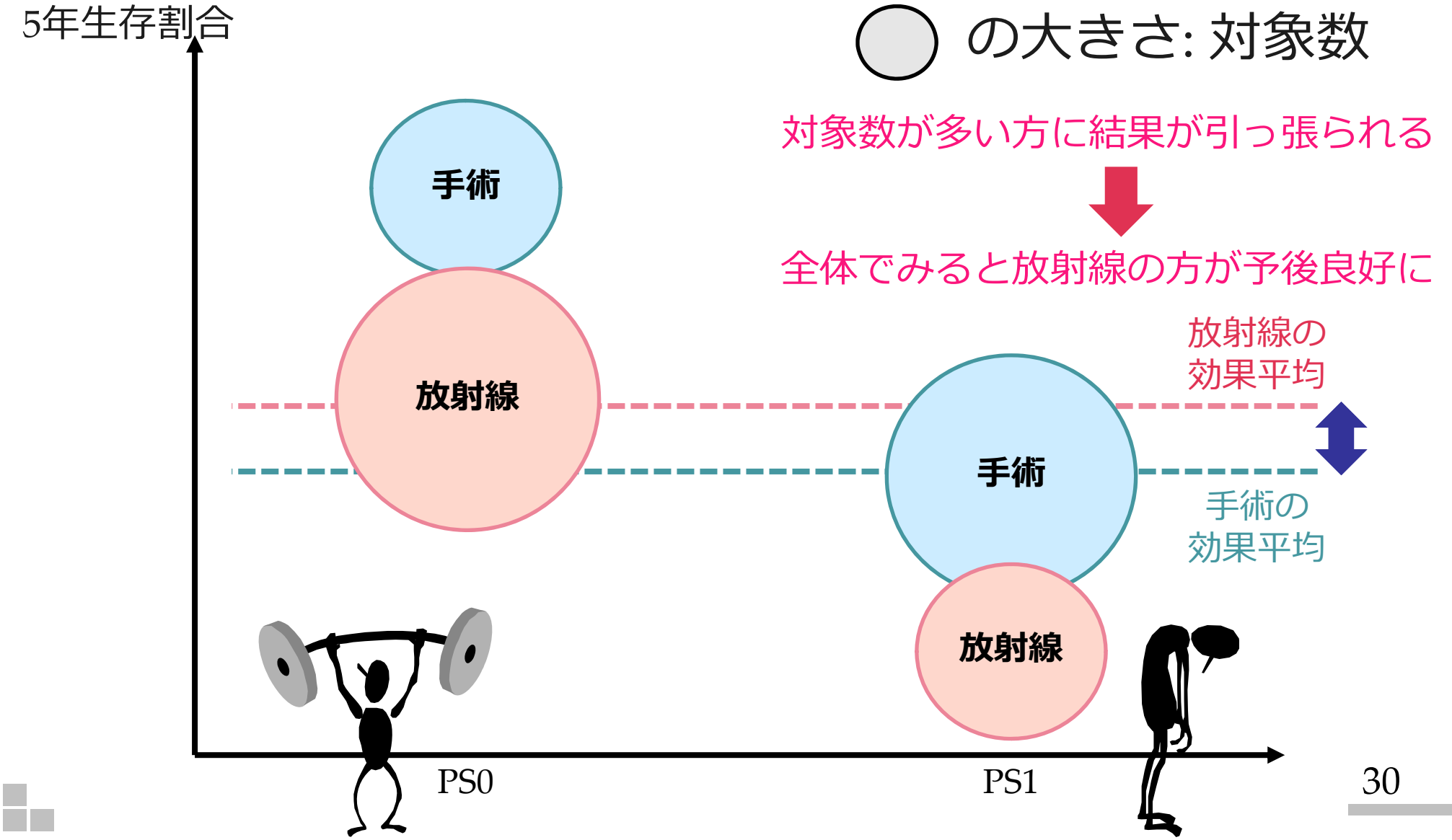
## 交絡の原因

予後と関連のある
交絡因子(PS)の構成比が
群間で偏っていたため

- 放射線にPS 0が多い
- PSが良いと予後が良い



# イメージ図



# ロジスティック回帰モデルの当てはめ

※ $p$ : 5年生存割合, treatment：放射線なら1、手術なら0

■単変量モデル

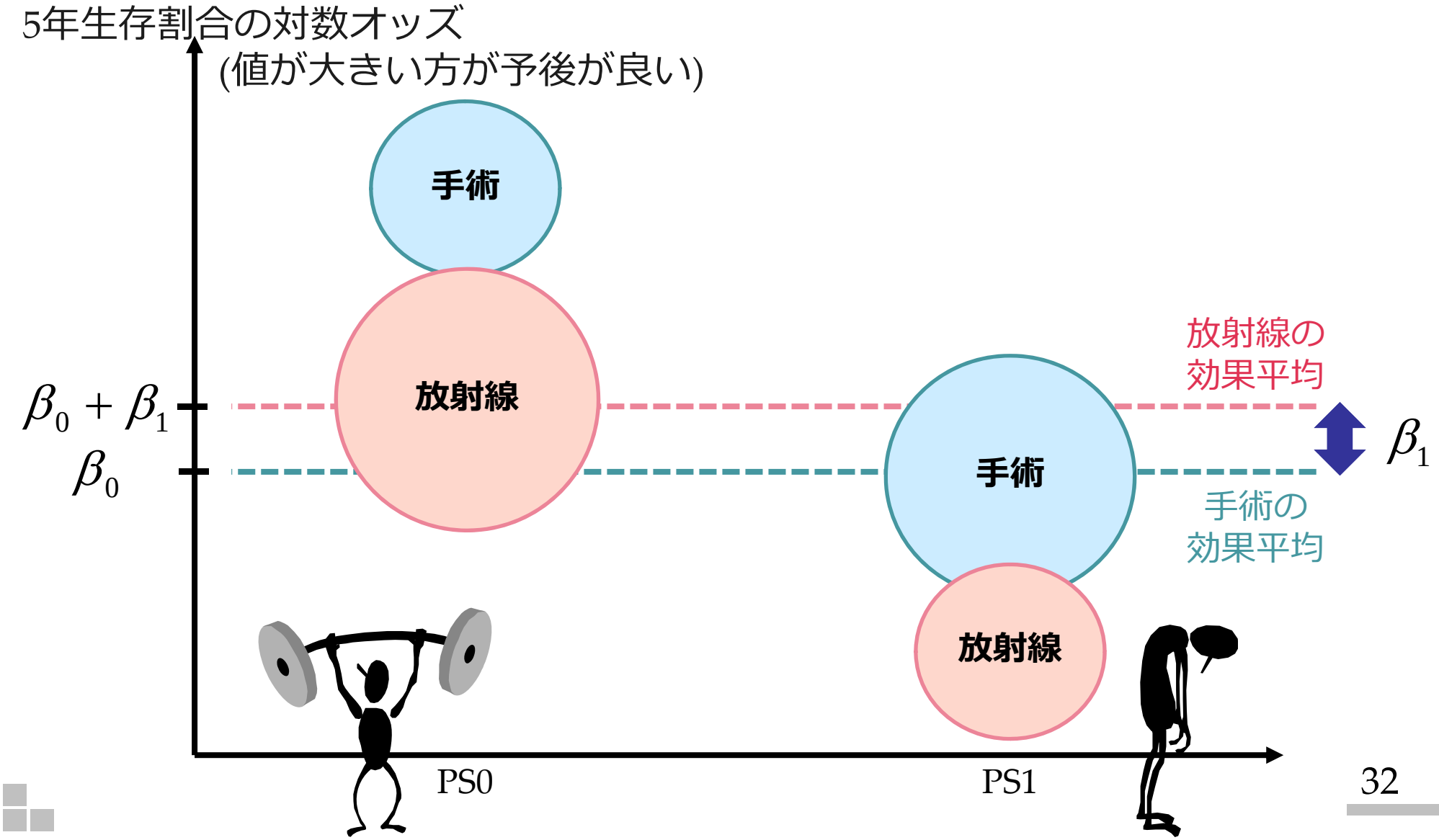
$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \times \text{treatment}$$

■多変量(2変量)モデル

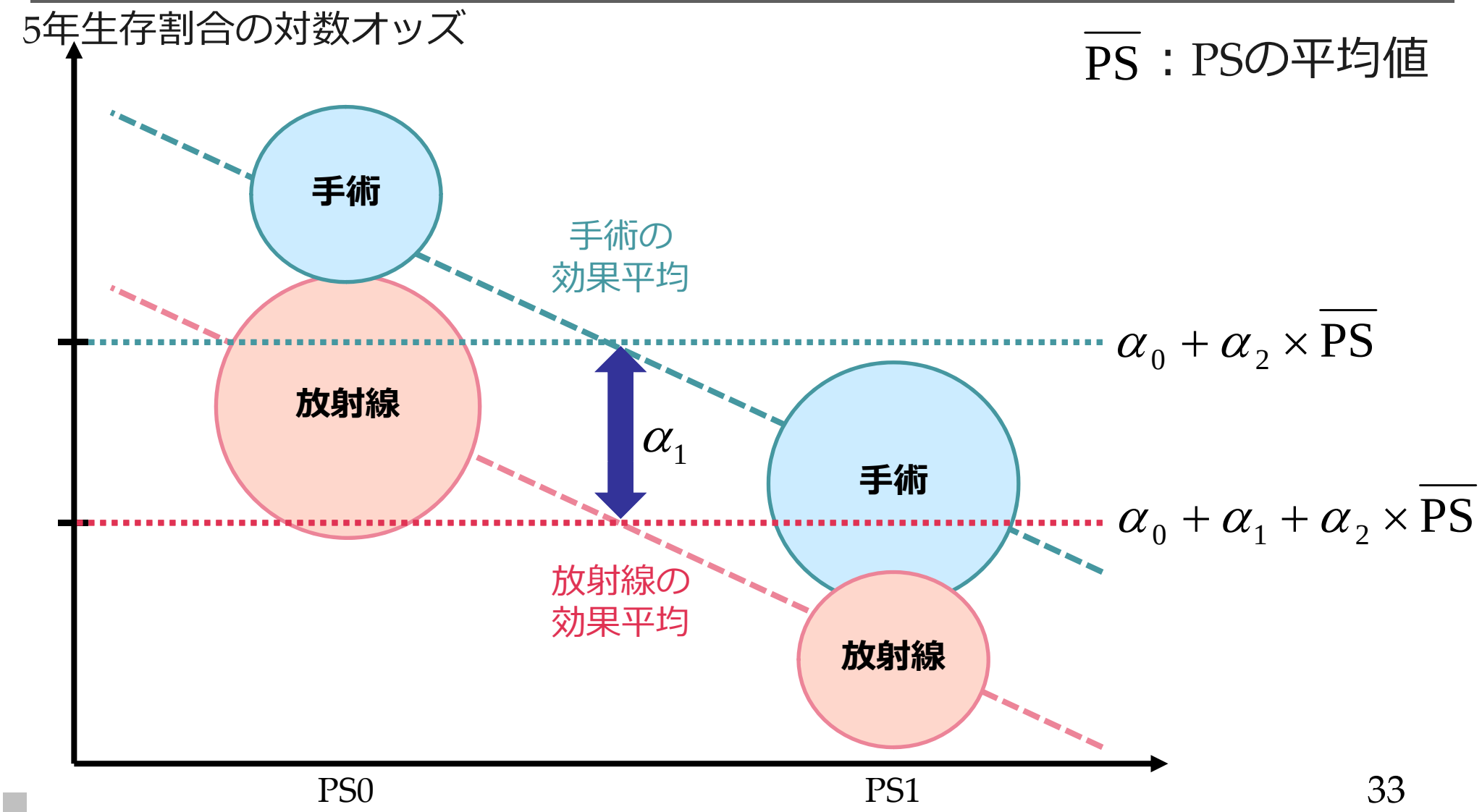
$$\log\left(\frac{p}{1-p}\right) = \alpha_0 + \alpha_1 \times \text{treatment} + \alpha_2 \times \text{PS}$$

# 単変量モデル

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



# 多変量モデル $\log\left(\frac{p}{1-p}\right) = \alpha_0 + \alpha_1 \times \text{treatment} + \alpha_2 \times \text{PS}$



# $\beta_1$ と $\alpha_1$ の違い

- $\beta_1$：PSで調整していない対数オッズ比
  - 解釈：「放射線群の5年生存割合が高い」
  - 交絡により真実が歪められている

- $\alpha_1$：PSで調整した対数オッズ比
  - 解釈：「手術群の5年生存割合が高い」
  - 交絡の影響を除外することができる
  - 説明変数に交絡因子を含めた場合の曝露の効果を「調整済みの効果」と呼ぶ

# この結果から分かること

- 曝露と結果の関連に興味がある場合
  （原因と結果の関連を推論すること(因果推論)に興味がある場合）
  - 交絡が存在すると曝露の効果が歪められてしまう
    - 曝露が完全にランダム割付けされている場合を除く
    - ランダム化できない曝露(例：喫煙、飲酒、食事など)は適切な手法を用いて交絡を調整する必要がある

- **多変量モデルに交絡因子を含めることで曝露の効果を適切に推定できる**

# 交絡因子の選択

- 適切に因果推論するためには
  交絡因子を漏れなく調整する必要がある
  - 交絡因子をどうやって選択すれば良いか？
    - 交絡因子の3条件に合致する変数をリストアップする
      ① アウトカムと関連がある
      ② 曝露と関連がある
      ③ 中間変数（曝露とアウトカムの中間にある変数）でない

- パラメータの推定精度が落ちるようであれば
  **変数を取捨選択**する

# 変数選択の基本的な方針

- データだけでは最適なモデルにならない
- 既知のリスク/予後因子という情報は使うべき
- あくまで目的は曝露の効果を推定すること
  - 「曝露以外の変数は交絡調整のために用いる」という観点で変数の取捨選択を検討すべき

# ある交絡因子Cをモデルに追加するかの判断

- change in estimate基準
  - Cを含める前と含めた後で曝露の効果が不変なら、モデルを複雑にしてまでCを追加する意義はない
  - あらゆる組合わせの検討が必要で手間がかかる

- 機械的なアルゴリズム(変数減少/増加法など)
  - Cを含めた or 除いた場合のCとアウトカムの関連の強さに基づいて決定する
  - 結果と関連が低い交絡因子が除かれないよう変数減少法を用いたり、既知の予後因子を強制投入したりする

# 変数選択を行う上での注意点

- 単変量で有意 or 既知の予後因子は
  交絡因子の重要な候補だが、
  **必ず調整しなければならない訳ではない**

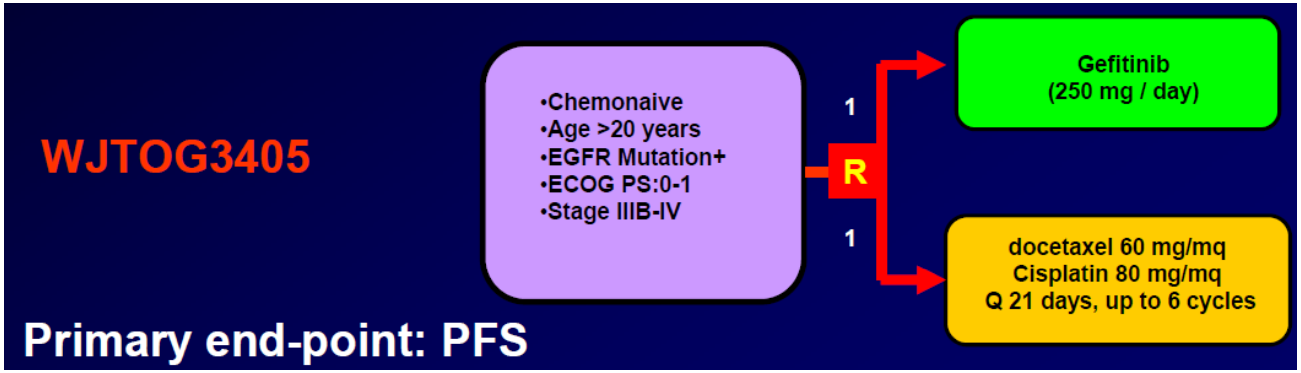
- 構築したモデルに正解がある訳ではない
  - 交絡が十分調整できていれば問題ない

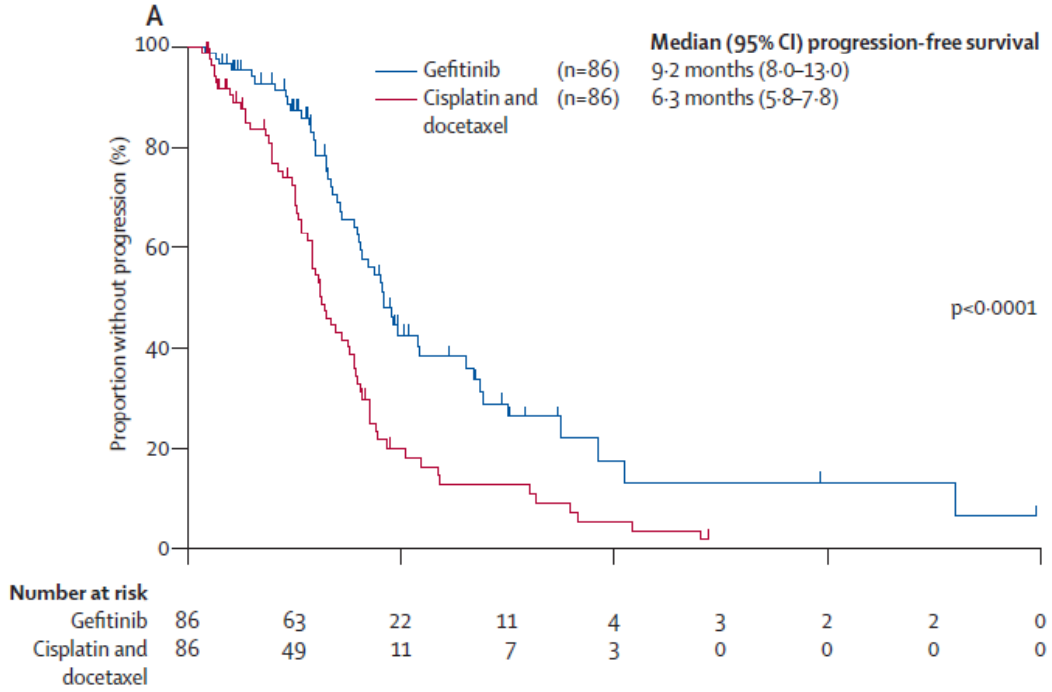
# 回帰モデルの利点と欠点

- 利点
  - 交絡因子の数が増えても調整できる
    - ただし、対象数やイベント数が少ない場合に
      調整する交絡因子の数が多いと推定精度が落ちてしまう
  - 交絡因子がアウトカムに与える影響を評価できる
    - あくまで交絡調整を第一に考えているので
      交絡因子の影響評価は二次的と考えるべき
  - 「層毎に治療効果が共通である」とする仮定は不要
    - 加法的な効果を仮定する線形モデルでない統計モデルを使えば、
      交互作用の強さについても評価することができる
- 欠点
  - 層別解析より仮定が強い(後述)

Mitsudomi T, et al. Lancet Oncology 2010.

# 実例：WJTOG3405試験



| | 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**

- 小数例のランダム化比較試験だったため EGFR遺伝子変異の種類が両群で偏っていた
- 多変量解析でこの影響を調整した

Mitsudomi T, et al. Lancet Oncology 2010.

| | 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**

**交絡を調整したハザード比**
**0.258**

ハザード ＝ $h_0(t)$ × gefitinibの効果 ← primary interest

交絡因子

× 性別の効果
× ・ ・ ・ ・
× EGFR変異の種類の効果

# OUTLINE

- 第4回セミナーの復習
- （統計）モデルとは？
- 交絡調整のための多変量解析
- 回帰モデルを用いる上での注意点
- 多変量解析の用途

# 回帰モデルの欠点：強い仮定

- 因果推論を行う上での仮定
  - 未測定 or 未知の交絡因子が存在しない
    - 大規模なランダム化比較試験ならこの仮定は成立する
- 数学的な仮定
  - モデルに含めた因子と結果に線形の傾向がある
  - モデルに対する仮定：比例ハザード性（Cox回帰）など
  - データに過剰に適合(over-fitting)するモデルでない
  - パラメータがきちんと推定されるための仮定
    - 多重共線性が存在しない
      （強い相関を示す変数がモデルに含まれていない）
    - モデルに含める変数の数に対して対象数・イベント数が十分ある

# 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（詳細は第6回セミナーで説明します）

# 多重共線性

- 強い相関を示す変数によりパラメータの推定が不安定になってしまう現象

- 対象数・イベント数を増やさない限り避けられない
  - 相関が高いことが自明な変数は関心が高い or 解釈が容易な方だけモデルに含めるべき
    - 例）BMIと体重/身長、病期とT/N因子

# 説明変数に関する注意点

- 1変数あたりどの程度の情報量が必要か？
- 連続変数とカテゴリーのどちらにすべき？
- 測定誤差や欠測がある場合は？

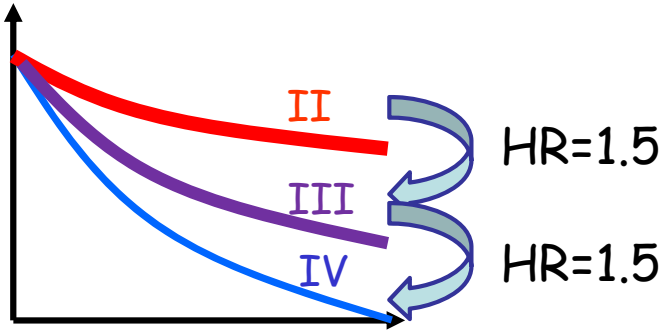
# 1変数あたりどの程度の情報量が必要か？

- 10～20例以上と言われている
  - Cox回帰の場合は10～20イベント
- 十分な情報量があっても問題が生じうる
  - 例：ある遺伝子変異が奏効割合に関連するか？
    - 推定すらできないケース
      - 1000例のデータでも変異が0例
    - 推定が不安定になってしまうケース
      - 調整する交絡因子の組合わせで構成される層の中に例数が極端に少ない(例：5例未満)層が存在する

# 連続？カテゴリー？

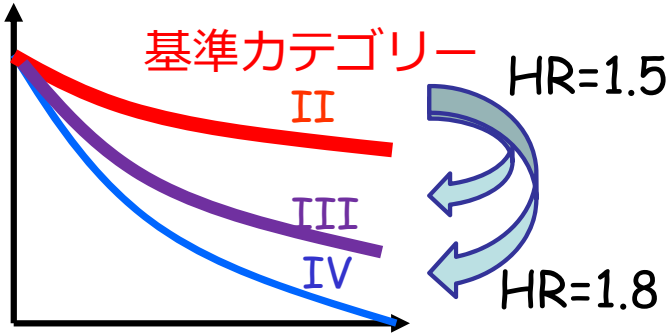
例）臨床病期を連続変数と扱うかカテゴリー変数と扱うか？

## ■連続変数：推定されるハザード比は1つ




IIに対するIIIのHR、IIに対するIVのHR
は共通である、とする仮定
→ 1単位増加した場合のHRを推定する

## ■カテゴリー変数：複数のハザード比を推定




IIに対するIIIのHR、IIに対するIVのHR
を別々に推定する
→ 基準カテゴリーに対するHRを推定する

# 迷ったときは・・・

- 臨床的に妥当なcutoffが存在する場合
  - 「どちらの治療効果が知りたいか？」で判断（臨床的な解釈を優先）
    - 「年齢1歳刻みのHR」vs.「高齢者 or notのHR」
    - 解釈しやすいcutoff (検査値なら基準値)でカテゴリー化する
- 臨床的に妥当なcutoffが存在しない場合
  - データに依存して決める(中央値など)ことも可能
  - 無理にカテゴリー化する必要はない
- 連続→カテゴリーにより失われる情報に注意
  - 非線形な傾向にある場合：例）BMIと癌死の関連
    - BMIが低すぎる集団、高い集団は死亡のリスクが高い
    - BMIを基準値(25)で区切ることは必ずしも適切でない

# 測定誤差や欠測がある場合は？

- 測定誤差があると治療効果にバイアスが入る
  - 極力、測定誤差の少ない変数を選択すべき
- 欠測があると解析から除外する必要がある
  - 欠測をカテゴリーの1つと扱うことはできるが解釈が容易でない
  - 重要な交絡因子はできる限り漏れなく収集すべき

# OUTLINE

- 第4回セミナーの復習
- （統計）モデルとは？
- 交絡調整のための多変量解析
- 回帰モデルを用いる上での注意点
- 多変量解析の用途

# 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

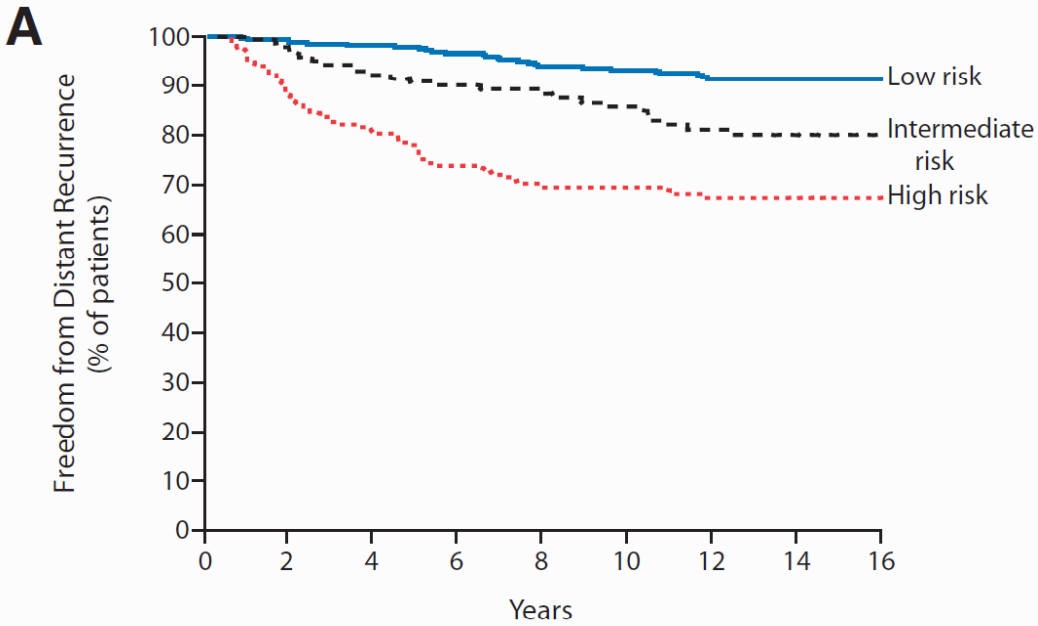
多変量解析の手法を用いてスコア計算式を構築

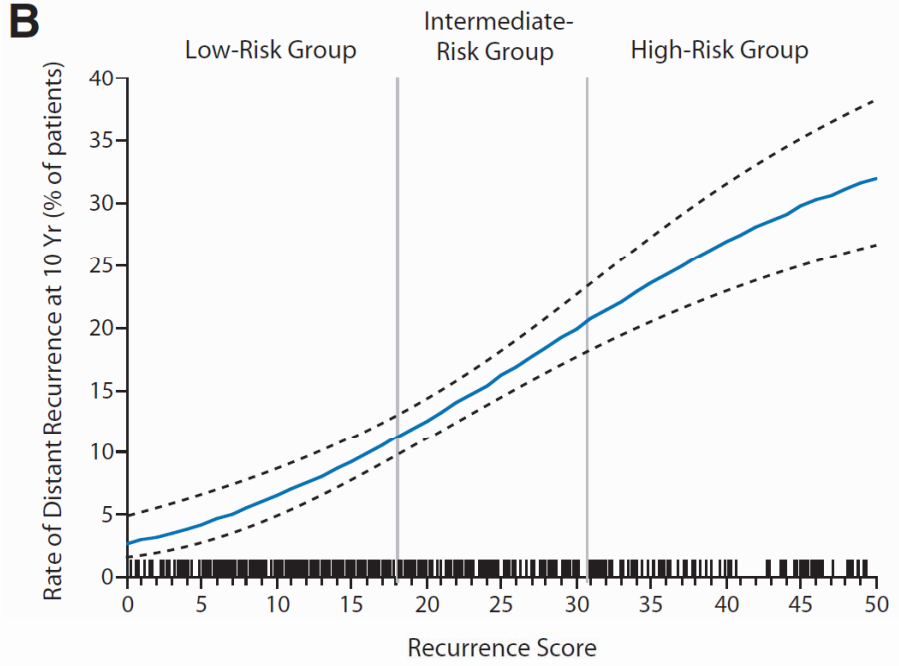
RS = + 0.47 x HER2 Group Score
- 0.34 x ER Group Score
+ 1.04 x Proliferation Group Score
+ 0.10 x Invasion Group Score
+ 0.05 x CD68
- 0.08 x GSTM1
- 0.07 x BAG1

| 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

# 多変量解析の用途

本日(第5回セミナー)のテーマ

- 因果推論アプローチ
  - 曝露と結果の関連に最も興味がある
  - 曝露以外はすべて交絡を調整する因子

- **予後予測アプローチ**
  - 結果の予測に興味がある
  - 予測能が高ければモデルに含める変数は問わない
  - 第6回セミナーで詳しく説明します

# まとめ

- 統計モデルとは？
  - （一般的には）誤差を許容した線形式で表される
  - ロジスティック回帰モデルからはオッズ比が、Cox回帰モデルからはハザード比が推定できる
- 交絡調整を目的とした多変量解析
  - 因果推論に関心がある場合に用いる
  - 曝露以外は交絡因子で構成される統計モデルを当てはめる
  - 種々の仮定の下で交絡による情報量の不均衡を調整し、曝露と原因の関連を推定する

# 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