

# 医学統計学事始め：がん臨床研究で 用いられる統計的方法の基礎

下川敏雄

和歌山県立医科大学 医学部  
和歌山県立医科大学附属病院 臨床研究センター

# 本日のお話の流れ

- がん臨床試験の区分
- 奏効割合などの2値アウトカムの統計的評価
- 全生存期間など生存期間アウトカムの評価
- いわゆる多変量解析について
- すこしおまけ：統計ソフトウェアについて

# がん臨床試験の区分

---

# がん臨床試験のフェーズ(相)と主なアウトカム

Phase I/II

## Phase I

用量反応毒性の検討

最大耐用量(MTD)がわかる

## Single-arm phase II

Historical Controlとの比較

非比較の範疇で治療効果(effect size)がわかる

## Randomized phase II

複数の試験治療アームの比較

Selection biasを除外したもとの治療効果がわかる(かも)

Seamless phase II/III

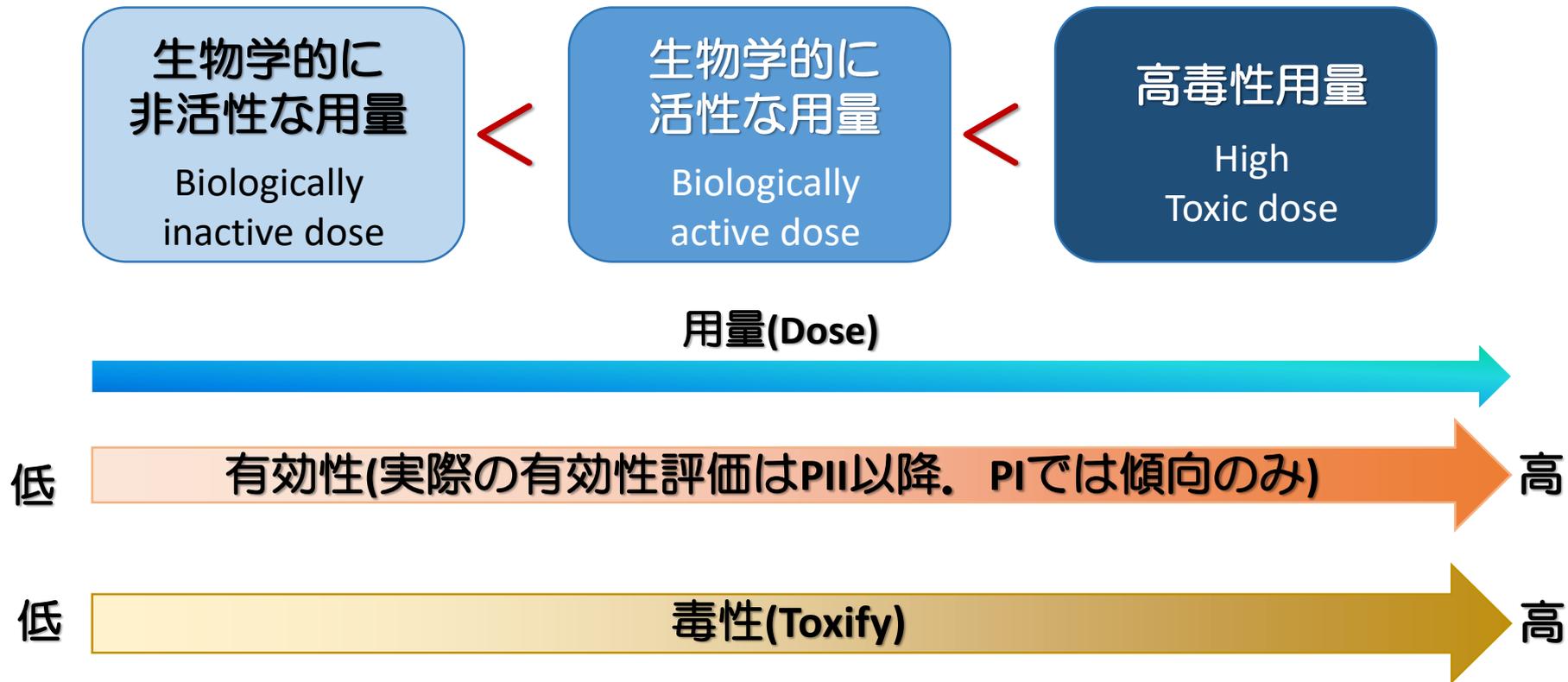
## Phase III

標準治療との比較

標準治療に対する有効性・安全性が検証できる

# 第I相試験

第I相試験の主たる目標は、新規薬剤(あるいは治療レジメン)のMTDを決定することである。その理由は、「多すぎる」用量が患者に対して、思い毒性が出るることが出ないようにするためである(Green, Benedetti & Crowley, 2003)



## 試験デザインの諸型

- アルゴリズム・ベース[非統計的]：3+3デザイン, Best of five design
- Bayes流推測ベース[統計的]：CRM(mCRM), EWOC

# 第II相試験(Phase II study)

第I相試験より決定された用法・用量に従って、対象とする癌腫における治験薬の臨床的意義のある治療効果，及び安全性を評価する(抗悪性腫瘍薬の臨床評価方法に関するガイドライン(厚生労働省 H17.11)の一部抜粋)

第II相試験における臨床的意義のある治療効果とは，主に**腫瘍縮小効果(奏効割合)**を指すが，標的病変がない場合や手術・手技の場合には，OS等を用いることもある。

## 試験デザインの諸型 (奏効割合の場合)

- 通常のデザイン：母比率の検定 (カイ2乗検定と同じ)
- 2段階デザイン：Simonの2段階デザイン, Green & Dahlbergの2段階デザイン

### 早期有効中止に対するSWOGの立場(Green et al., 2012)

がんの第II相試験において「有効性」が示されたために，試験を早期に中止することが重要であるという状況はまずお目にかからない。試験の途中で有効性が見出された薬剤の評価をさらに続けるかわりに，そこで第II相試験を終了することが妥当と判断されるほど高い有効性が示されることは，先ずなく，また有望な薬剤に対する同じプロトコルの試験に，有効性の程度をさらに詳しく調べるために，患者を追加して登録することは，殆どの場合正しいと見做される(略)。

# 無作為化第II相試験(Randomized Phase II study)

## 単群試験

- 信頼できる標準治療に対するヒストリカルコントロールが存在する場合  
Taylor et al. (2006)<sup>1)</sup>はRCTにおいて問題になるのは試験群のバイアスであることを指摘している(つまり、信頼がおけるヒストリカルコントロールがある場合には単群試験のほうが効率が良い)。
- 新治療の治療完遂率を評価する場合
- 少数例の患者を研究対象とする場合

## 無作為化比較試験

- 標準治療が存在しない場合。
- 標準治療が急激に変化したことで信頼できるヒストリカルコントロールが存在しない場合
- 対象患者が変化した場合(分子標的薬で行われるサブセット等)

Mandrekar and Sargent(2010)は、無作為化比較第II相試験のタイプとして、以下を提示している。

- 並行非比較レジメン (Parallel Noncomparative Regimens)
- ランダム化選択デザイン (Randomized Selection Designs)
- ランダム化スクリーニングデザイン (Randomized screening Designs)

## 第III相試験(Phase III study)

第III相試験は、より優れた標準的治療法を確立するために行われる臨床試験である。第II相試験において安全性と腫瘍縮小効果、又は何らかのメリット（症状緩和効果等）が確認された新規抗悪性腫瘍薬の単独又は併用療法と適切な対照群との比較試験である。この比較試験では、新規抗悪性腫瘍薬の臨床的有用性が明確に検証できるように試験を計画しなければならない。従って、第III相試験では、**生存率、生存期間等をプライマリーエンドポイント**とし、他の適切なエンドポイントとして安全性、妥当性の評価された方法による症状緩和効果やQOL（Quality of Life）等に関する評価を行い、これらに対し何らかの有用性が示される必要がある。そのためには対象とする腫瘍を指定し、明確に規定された患者集団において、重要な予後因子等を考慮した適切な割付を行い、適切なデータ管理を実施して試験を遂行する必要がある。（厚生労働省 H17.11)の一部抜粋)

### 無作為化比較 第III相試験で検討しなければならない課題

- **エンドポイント：OS or surrogate**
- **割付：割付方法・割付調整因子・盲検化**
- **ITT, FASの取扱い**
- **試験の目標 (優越性, 非劣性)**

詳しくはBKUPに記載

# 癌臨床研究における主なアウトカム

## ■ 割合に関するアウトカム

奏効割合, 病勢コントロール割合, 治療完遂割合, 手術移行割合 etc...

## ■ 生存期間に関するアウトカム

全生存期間, 無増悪生存期間, 無再発生存期間, 治療成功期間 etc...

癌臨床研究では, 割合 and/or 生存期間のアウトカムが多い(いわゆる, 連続データと呼ばれるデータを用いることは少ない)



ここでのお話は, 割合と生存期間に焦点を絞って議論する.

# 奏効割合などの2値アウトカムの統計的評価

---

# 単群試験における評価：抗腫瘍効果

抗腫瘍効果に対する治療前の標的病変を100としたときのざっくりとした評価の意味(RESIST)

## 完全奏効(CR)



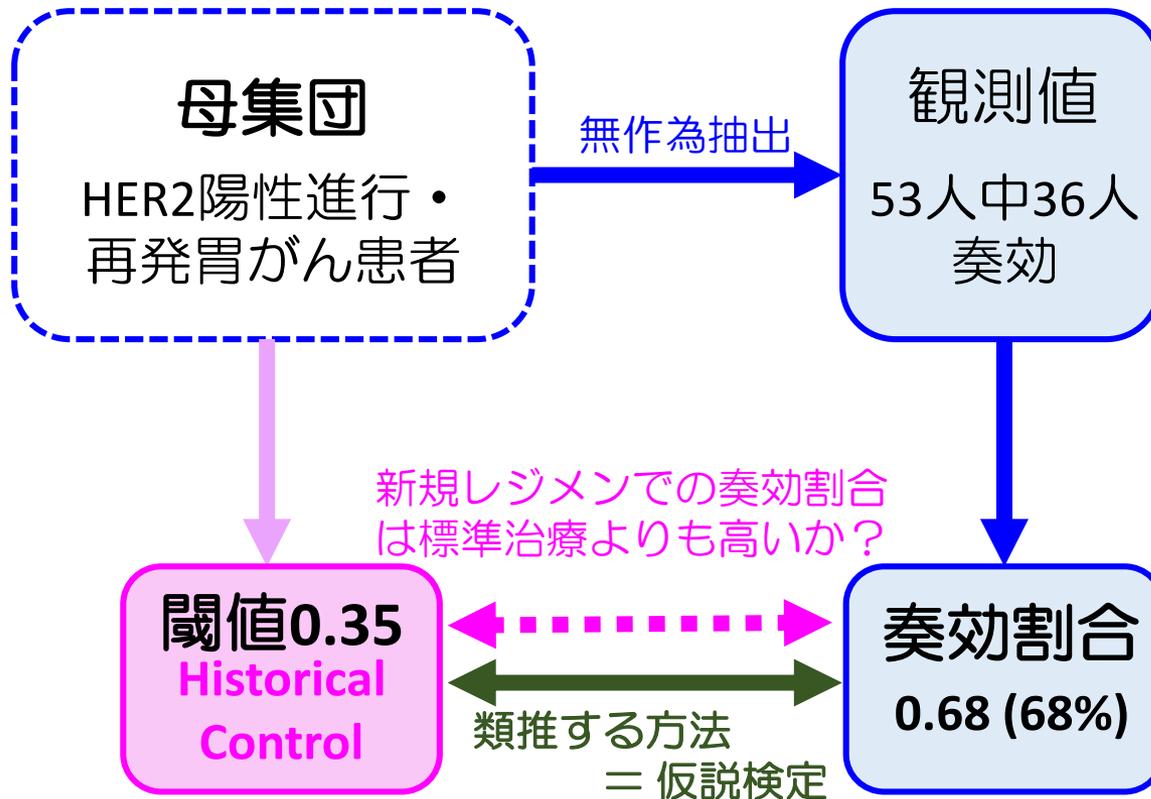
また、何らかの理由で評価できなかった症例は、NE(Not Evaluate)となる、

$$\text{奏効率} = \frac{\text{CR(完全奏効)} + \text{PR(部分奏効)}}{\text{全登録症例}}$$

因みに、増悪しなかった症例の割合(CR+PR+SDの割合)を病勢コントロール割合という。

# 単群試験の例示

HER2陽性の進行・再発胃癌患者に対するフッ化ピリミジン系抗がん剤とCDDPを併用したレジメンでの奏効割合が35%であると報告されている。新しい臨床試験では、S1+CDDP+Tmabの併用療法による新たなレジメンの有効性を単アーム試験で実施した。その結果、53例中36例の奏効例が認められた(奏効割合=68%) (Kurokawa et al., *BJC*, 2014)



この臨床試験の奏効割合は68%だった。この結果に基づいて、母集団(本邦のすべてのHER2陽性進行・再発胃癌患者)に新規レジメンを実施したと仮定した場合に、その奏効割合が閾値奏効割合35%よりも高いと統計的に判断する。

母比率の検定  
(binomial test)

# 母比率の検定：帰無仮説と対立仮説

## Clinical Question

新規治療レジメンでの奏効割合は、標準治療(閾値奏効割合)35%よりも高いか？

## 帰無仮説 $H_0$ (Null hypothesis)

新規治療レジメンでの奏効割合が標準治療と同じである，つまり  
(新規治療レジメンでの奏効割合) = 35% (標準治療の奏効割合)  
である。

唯一の数値で設定できるこちら側を評価の基準とする  
(ただし研究の目的と逆)

## 対立仮説 $H_1$ (Alternative hypothesis)

新規治療レジメンでの奏効割合が標準治療よりも大きい，つまり，  
(新規治療レジメンでの奏効割合) > 35% (標準治療の奏効割合)  
である。

上の仮説が誤っていることを示すことで，逆の仮説である  
こちらが正しいと判断する。

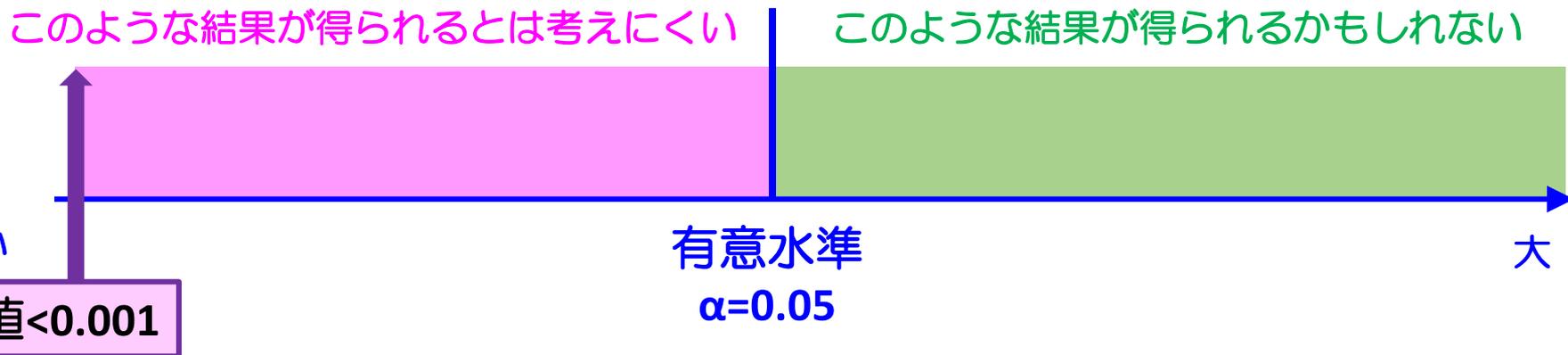
# 母比率の検定：どのように「誤っている」と判断するのか？

帰無仮説 $H_0$ が正しいときに、データから得られた差が偶然に生じたものであるかどうかを表す確率のようなものが**P値**である。

**p値が小さいとは、帰無仮説 $H_0$ が正しいとはいえないほどに差があることを意味する。つまり、対立仮説 $H_1$ が正しい状況にある。**

➡ 対立仮説が支持されることを「**有意である(significance)**」という。

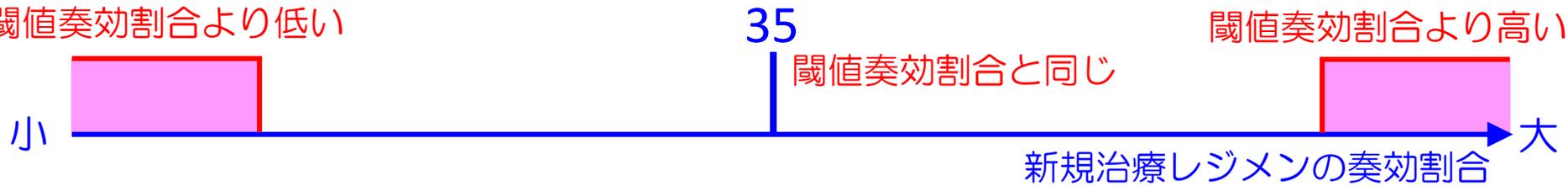
**有意水準 $\alpha$** とは、この結果が得られる状況が考えられ得るものなのか、そうでないのかを判断するカットオフ値のようなもの。



新規治療レジメンでの真の奏効割合が35%とすると、本試験の結果が得られる確率(p値)は0.1%未満(有意水準5%未満である)。したがって、帰無仮説が棄却され(有意である)、逆仮説である新規治療レジメンの奏効割合は35%以上である(対立仮説)と判断する。

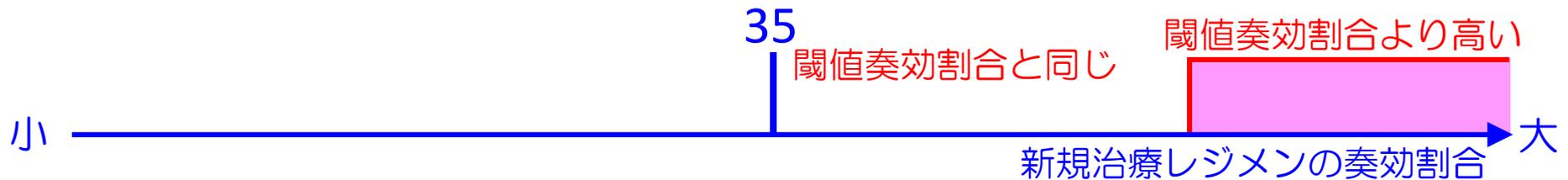
# 片側対立仮説と両側対立仮説

- 両側対立仮説：新規治療レジメンの奏効割合と閾値奏効割合に違いがある。



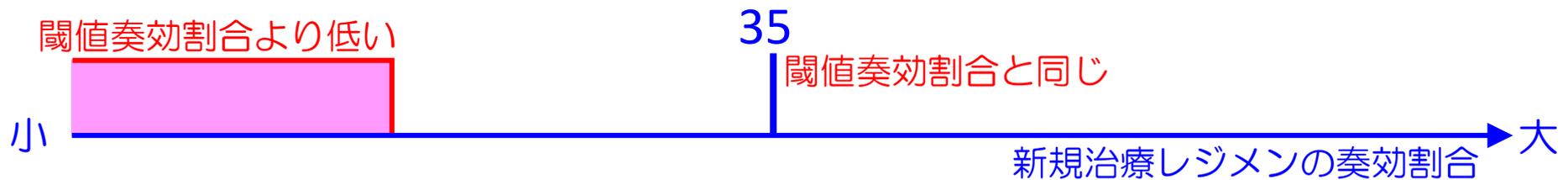
有意水準は「閾値奏効割合の35%よりも高い確率」「閾値奏効割合の35%よりも低い確率」の両方で設定されるので、片側では $\alpha/2$ になる。

- 片側対立仮説(1)：新薬のほうがプラセボの罹患率よりも低い。



有意水準は「閾値奏効割合の35%よりも高い確率」で設定されるので、 $\alpha$ になる。

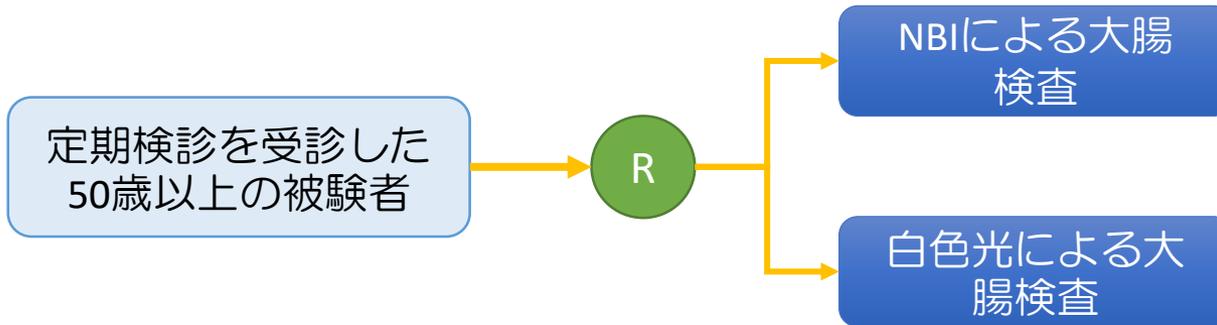
- 片側対立仮説(2)：新薬のほうがプラセボの罹患率よりも高い。



有意水準は「閾値奏効割合の35%よりも低い確率」で設定されるので、 $\alpha$ になる。

# 2値アウトカムにおける比較

## ■ 2種類の大腸内視鏡検査における診断能の比較



### Outcome

1個以上近位結腸の鋸歯(きよし)状病変が存在した被験者割合

	病変あり	病変なし	計
NBI	204 (51.1%)	195 (48.9%)	399
白色光	158 (39.4%)	243 (60.6%)	401
計	362 (45.2%)	438 (54.2%)	800



Clinical  
Question

大腸内視鏡検査において、NBI検査と白色光検査で鋸歯状病変の検出の有無に違いがあるか？

このような場面で良く用いられる方法：カイ2乗検定

# カイ2乗検定

帰無仮説：内視鏡検査(NBI, 白色光)によって鋸歯状病変の検出能に違いがない。

対立仮説：内視鏡検査(NBI, 白色光)によって鋸歯状病変の検出能に違いがある。

仮説検定：帰無仮説の位置からどれだけ離れているか(どれだけ違っているか)を確率(p値)で表す

	病変あり	病変なし	計
NBI	204	195	399
白色光	158	243	401
計	362	438	800

帰無仮説でのクロス集計表を作らないといけない。

## 臨床試験の結果

	病変あり	病変なし
NBI	204	195
白色光	158	243

## 期待度数(帰無仮説での位置)

病変あり	病変なし
$\frac{399 \times 362}{800} = 180.5$	$\frac{399 \times 438}{800} = 218.5$
$\frac{401 \times 362}{800} = 181.5$	$\frac{401 \times 438}{800} = 219.5$

乖離

各セルに対して、度数の差の2乗値を計算して、その和をとったものが帰無仮説の位置からの離れ度合いである(カイ2乗値という)。

その結果、**p値は0.001**なので、有意であった。つまり、1個以上近位結腸の鋸歯状病変が存在した被験者割合において違いが認められた。

## 2値アウトカムにおいて重要な指標

	病変あり	病変なし	病変割合
NBI	204 (a)	195 (b)	$p_1=0.511$
白色光	158 (c)	243 (d)	$p_2=0.394$

オッズとは、ある結果が生じる比率とある結果が生じない比率との比である。

$$\text{NBI} = \frac{0.511}{1-0.511} = 1.045 \qquad \text{白色光} = \frac{0.394}{1-0.394} = 0.650$$

オッズ比は、結果(病変の有無)に対してどれぐらい要因(内視鏡検査)が寄与しているかを表す。

$$\text{オッズ比} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{a \times d}{b \times c} = \frac{204 \times 243}{195 \times 158} = 1.609$$

つまり、NBIは白色光に比べて**1.609倍の病変の検出(所見)の上昇**を与えることができる。

# 少数例の場合の比較：Fisherの正確検定

ある病院において、大腸内視鏡検査のNBI検査と白色光検査で鋸歯(きよし)状病変の検出の有無に違いが調査された(仮想例)。

	病変あり	病変なし	計
NBI	12 (54.5%)	10 (45.5%)	22
白色光	3 (20.0%)	12 (80.0%)	15
計	15 (40.5%)	22 (59.5%)	37

	病変あり	病変なし	計
NBI	12	10	22
白色光	3	12	15
計	15	22	37

Diagram illustrating the relationship between cells in a 2x2 cross-tabulation table. The cell containing '12' (NBI, Disease Present) is highlighted in red. Purple arrows show the calculation of other cells based on marginal totals:  $22 - 12 = 10$  (NBI, Disease Absent),  $15 - 3 = 12$  (White Light, Disease Present), and  $22 - 10 = 12$  (White Light, Disease Absent). The marginal totals (15, 22, 37) are enclosed in a blue box.

2×2クロス集計表では、周辺度数(青色の部分)を固定してしまうと、1個のセル(赤色)が決まれば、残りのセルは全て決まる。



クロス集計表におけるすべてのパターンを作ることができる。

こうなる確率 = 0.0000

$$a \times d - b \times c = -330$$

	あり	なし
NBI	0	22
白色光	15	0

こうなる確率 = 0.0000

$$a \times d - b \times c = -293$$

	あり	なし
NBI	1	21
白色光	14	1

こうなる確率 = 0.0000

$$a \times d - b \times c = -256$$

	あり	なし
NBI	2	20
白色光	13	2

こうなる確率 = 0.0001

$$a \times d - b \times c = -219$$

	あり	なし
NBI	3	19
白色光	12	3

こうなる確率 = 0.0011

$$a \times d - b \times c = -182$$

	あり	なし
NBI	4	18
白色光	11	4

こうなる確率 = 0.0084

$$a \times d - b \times c = -145$$

	あり	なし
NBI	5	17
白色光	10	5

こうなる確率 = 0.0399

$$a \times d - b \times c = -108$$

	あり	なし
NBI	6	16
白色光	9	6

こうなる確率 = 0.1172

$$a \times d - b \times c = -71$$

	あり	なし
NBI	7	15
白色光	8	7

こうなる確率 = 0.2197

$$a \times d - b \times c = -34$$

	あり	なし
NBI	8	14
白色光	7	8

こうなる確率 = 0.2659

$$a \times d - b \times c = 3$$

	あり	なし
NBI	9	13
白色光	6	9

こうなる確率 = 0.2074

$$a \times d - b \times c = 40$$

	あり	なし
NBI	10	12
白色光	5	10

こうなる確率 = 0.1028

$$a \times d - b \times c = 77$$

	あり	なし
NBI	11	11
白色光	4	11

こうなる確率 = 0.0314

$$a \times d - b \times c = 114$$

	あり	なし
NBI	12	10
白色光	3	12

こうなる確率 = 0.0056

$$a \times d - b \times c = 151$$

	あり	なし
NBI	13	9
白色光	2	13

こうなる確率 = 0.0005

$$a \times d - b \times c = 188$$

	あり	なし
NBI	14	8
白色光	1	14

こうなる確率 = 0.000

$$a \times d - b \times c = 225$$

	あり	なし
NBI	15	7
白色光	0	15

こうなる確率 = 0.0000

$$a \times d - b \times c = -330$$

	あり	なし
NBI	0	22
白色光	15	0

こうなる確率 = 0.0000

$$a \times d - b \times c = -293$$

	あり	なし
NBI	1	21
白色光	14	1

こうなる確率 = 0.0000

$$a \times d - b \times c = -256$$

	あり	なし
NBI	2	20
白色光	13	2

こうなる確率 = 0.0001

$$a \times d - b \times c = -219$$

	あり	なし
NBI	3	19
白色光	12	3

こうなる確率 = 0.0011

$$a \times d - b \times c = -182$$

	あり	なし

こうなる確率 = 0.0084

$$a \times d - b \times c = -145$$

	あり	なし

こうなる確率 = 0.0399

$$a \times d - b \times c = -108$$

	あり	なし

こうなる確率 = 0.1172

$$a \times d - b \times c = -71$$

	あり	なし

試験で得られた状況よりも極端な状況のクロス集計表になる確率の総和を計算する。その結果、p値は

$$0.0000 + 0.0000 + 0.0000 + 0.0001 + 0.0011 + 0.0084 +$$

$$0.0314 + 0.0056 + 0.0005 + 0.0000 = \mathbf{0.0471}$$

であるので、有意水準 $\alpha=0.05$ のもとで有意である。したがって、NBIと白色光には病変の検出に差異が認められた。

白色光	7	8
-----	---	---

白色光	6	9
-----	---	---

白色光	5	10
-----	---	----

白色光	4	11
-----	---	----

こうなる確率 = 0.0314

$$a \times d - b \times c = 114$$

	あり	なし
NBI	12	10
白色光	3	12

こうなる確率 = 0.0056

$$a \times d - b \times c = 151$$

	あり	なし
NBI	13	9
白色光	2	13

こうなる確率 = 0.0005

$$a \times d - b \times c = 188$$

	あり	なし
NBI	14	8
白色光	1	14

こうなる確率 = 0.000

$$a \times d - b \times c = 225$$

	あり	なし
NBI	15	7
白色光	0	15

# カイ2乗検定とFisherの正確検定の使い分け

	病変あり	病変なし	計
NBI	12 (54.5%)	10 (45.5%)	22
白色光	3 (20.0%)	12 (80.0%)	15
計	15 (40.5%)	22 (59.5%)	37

## カイ2乗検定

p値 = 0.0783

有意差が認められない

## Fisherの正確検定

p値 = 0.0471

有意差が認められる



カイ2乗検定では、p値を**近似**で計算する。そのため、

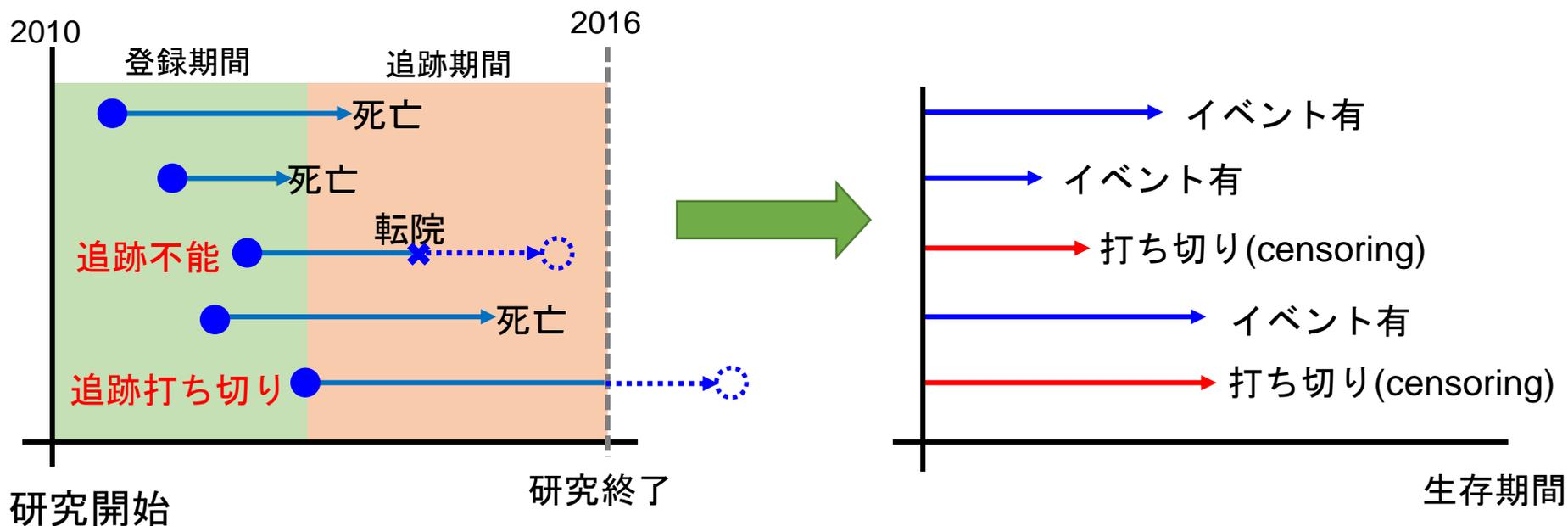
- (1) 標本サイズが少ない場合 (50未満)
- (2) どこかのセルの度数が小さい場合

には、近似精度が悪くなる。そのため、Fisherの正確検定を用いるほうが良い。一方で、標本サイズが大きくなると、Fisherの正確検定は計算できなくなる(パソコンがフリーズする)ので、カイ2乗検定を推奨する。なお、標本サイズが増えれば、カイ2乗検定のp値とFisherの正確検定のp値は一致する。

# 全生存期間など生存期間アウトカムの評価

---

# 生存期間データのメカニズム：打ち切り(censoring)



生存時間データは，個体の死亡(あるいは故障)までの時間を観測して得られる．観測される生存時間は生存時間(確率変数) $T$ の実現値 $t$ であるか，あるいは中途打ち切りが生じたときには，その中途打ち切り(censoring)時点の値である．

本報告でとり扱う右側中途打ち切りとは，個体の登録時点はわかるものの，追跡期間中に死亡(event)が起こらなかつたことを表す．

# 生存期間を取り扱ううえで重要な測度(関数)

## ■ 生存時間関数

個体が少なくとも時点  $t$  で生存する確率

時点 $t$ まで生存した割合。

年次生存割合などは、生存時間関数を用いて計算する

## ■ ハザード関数

個体が時点 $t$ まで生存しているという仮定のもとで微小区間

$t \leq T < t + \Delta t$ で死亡する確率

時点 $t$ における死亡リスク。

症例数設計などは、ハザード関数を用いて計算する

## ■ 確率密度関数 (probability density function)

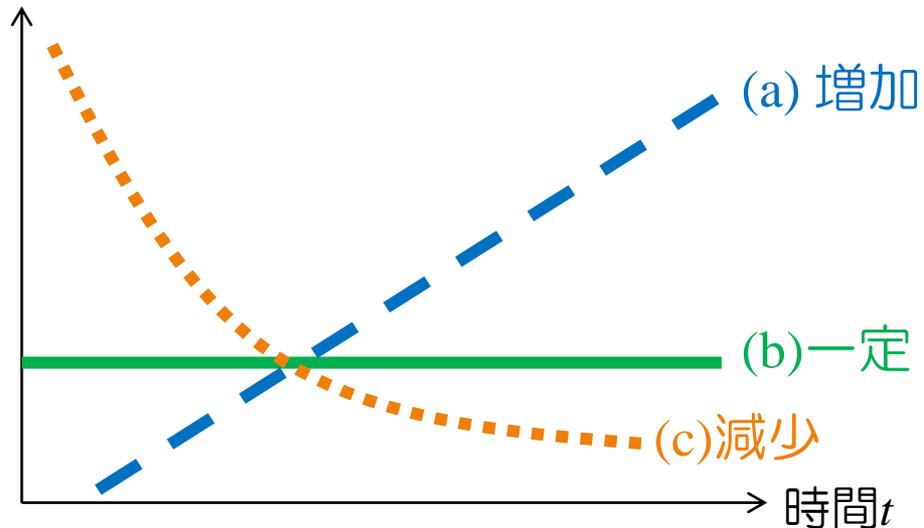
個体が微小区間  $t \leq T < t + \Delta t$ 内で死亡する確率

## ■ 累積ハザード関数 (cumulative hazard function)

ハザード関数の累積値

# ハザードの意味

ハザード関数 $h(t)$



## (a) の状況

登録時点では死亡(あるいはイベント)のリスクは低いものの時間経過とともに増大することを意味する。

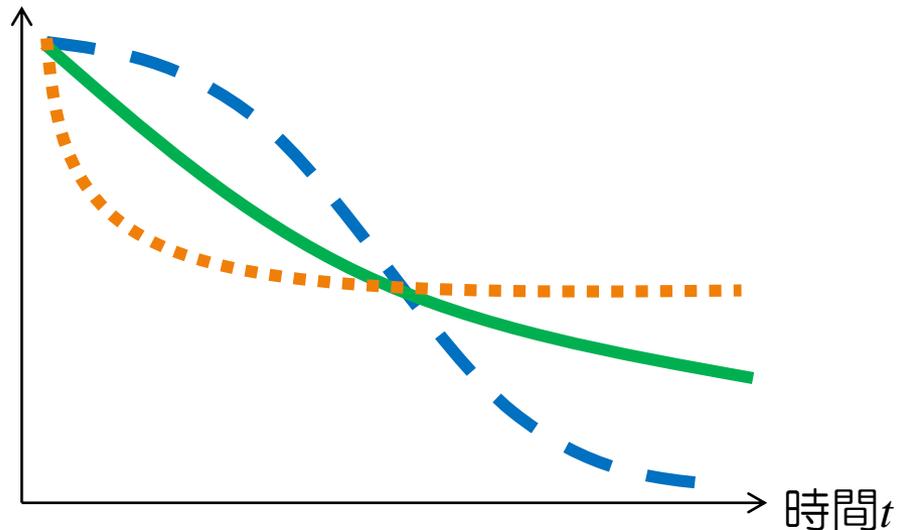
## (b) の状況

全時間上でハザード(すなわちリスク)が一定であることを意味する。

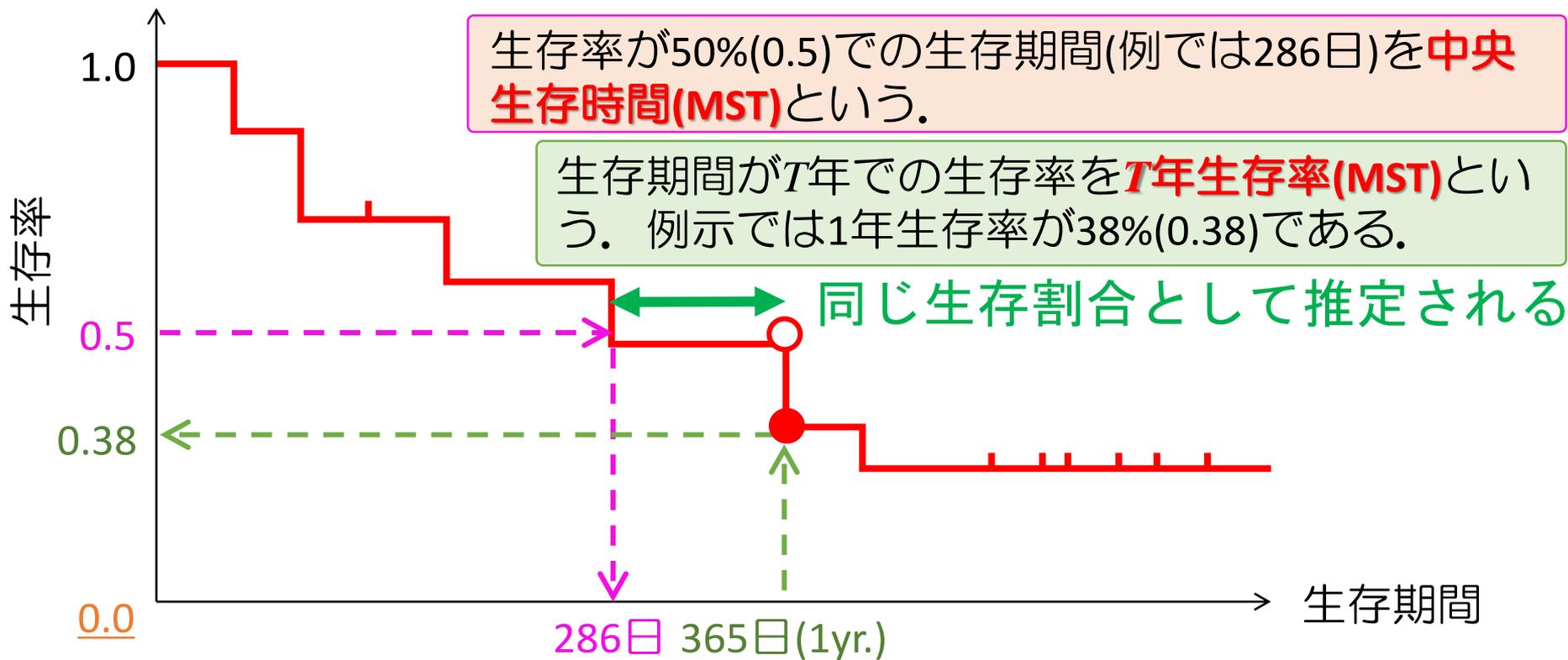
## (c) の状況

登録時点では死亡(あるいはイベント)のリスクは高いものの時間経過とともに減少することを意味する。一般に、進行がんに対して見られる。

生存確率 $S(t)$

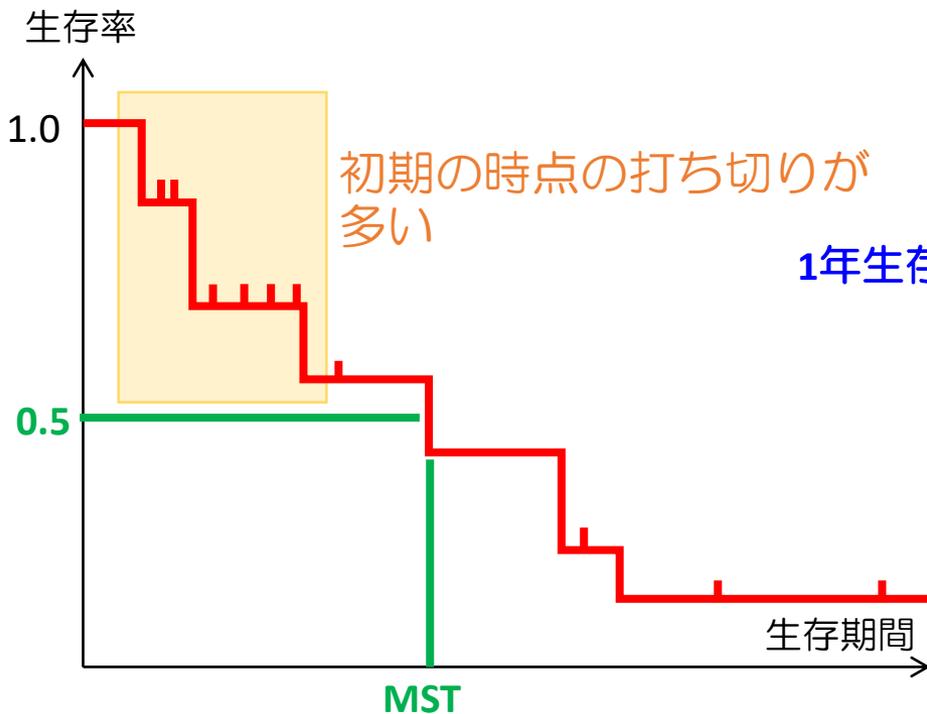


# 生存曲線の描写：Kaplan-Meier法



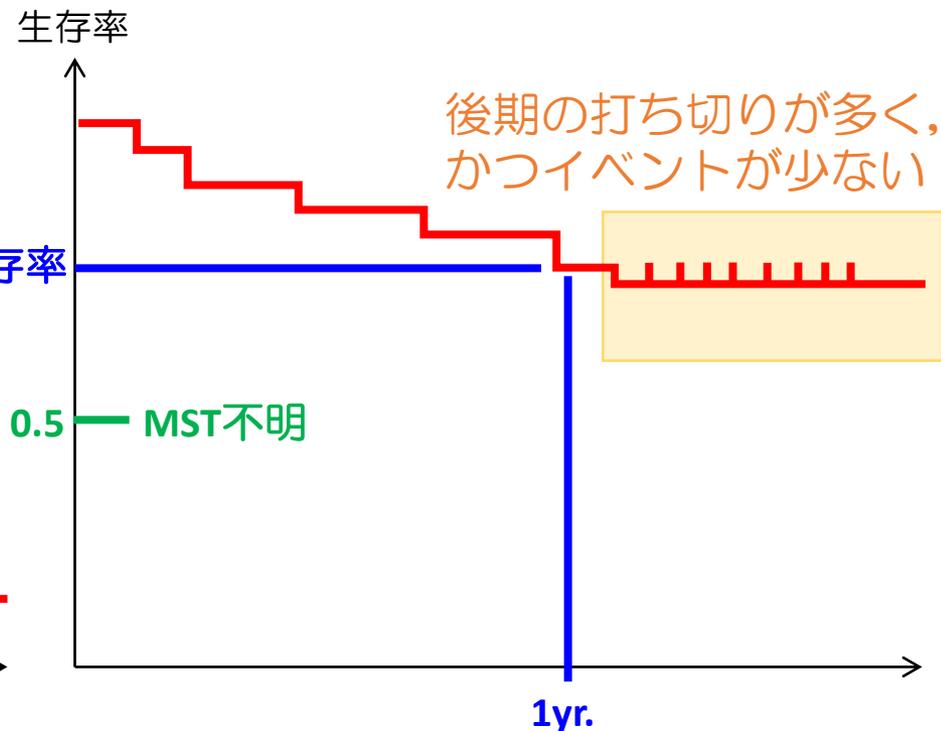
ここで注意点は、365日のところで、矩形の上側を見るのか下側を見るのかという点である。Kaplan-Meier曲線の矩形は上側ではなく下側での生存率を見なければならない。

# Kaplan-Meier法の見方での注意点



初期の時点で打ち切りが多いということは、きちりとしたフォローアップがなされていない、あるいは、研究実施体制に問題があると判断される可能性がある。

また、この場合のMSTは不完全である。



後期の時点で打ち切りが多く、かつフォローアップ期間中にイベントが少ない理由：

- ① 対象疾患の予後が良好な疾患の場合
- ② フォローアップ期間が十分でない場合がある。

この場合には、MSTは計算できないため、年次生存率のみを計算

# 2つの生存曲線の比較：ハザード比

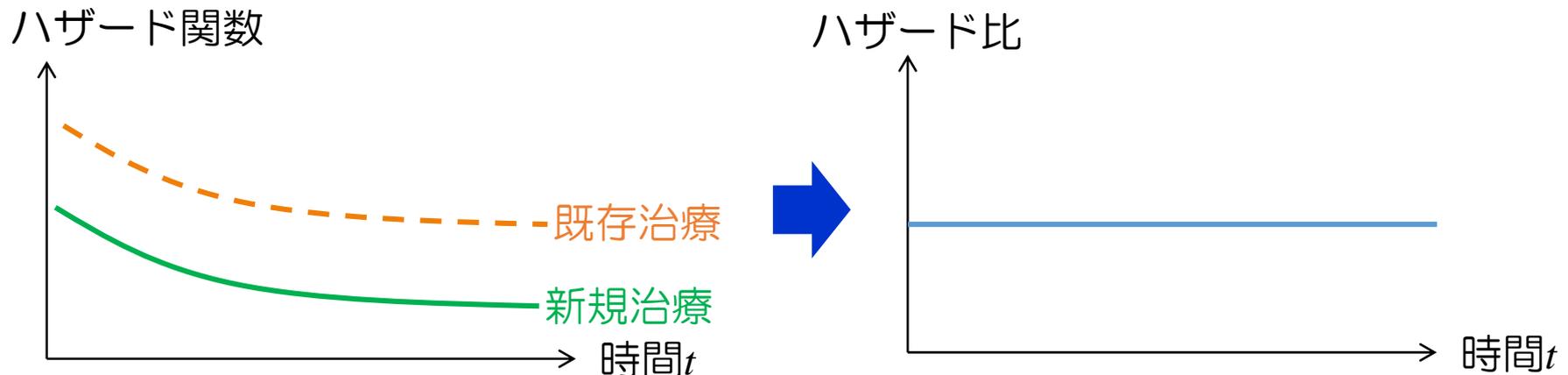
ハザード比は、

$$\text{時間}t\text{における}HR = \frac{\text{時間}t\text{における新規治療のハザード}}{\text{時間}t\text{における既存治療のハザード}}$$

で表される。HRは新規治療は既存治療に比べて**何倍程度のイベント(OSの場合、死亡)のリスク**があるかを表している。

ハザード比が1.0超の場合には新規治療のほうが既存治療よりNegativeであり、1.0未満の場合はPositiveである。

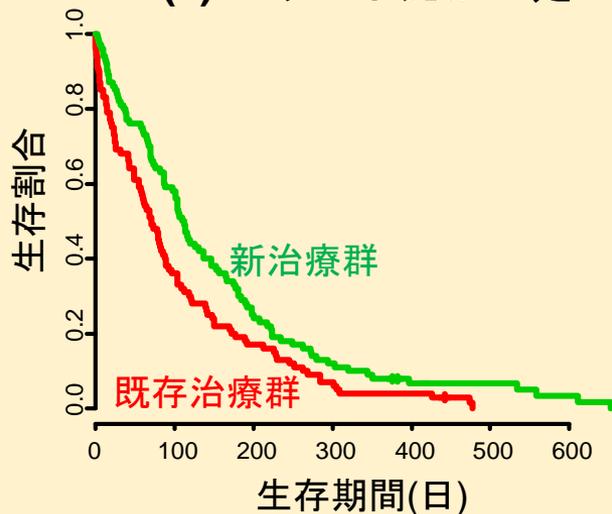
生存期間解析では、ハザードの比例性を仮定することが多い



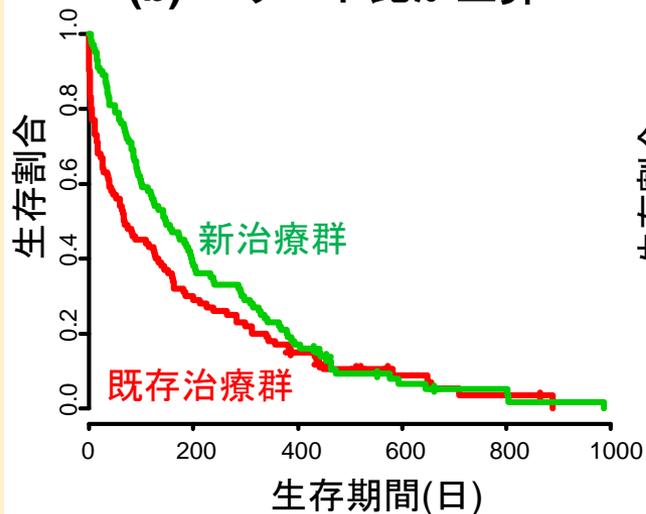
- 比例ハザード性とは、時間 $t$ に関して、同じようなハザードをもつことをいう。この仮定があることで、ハザード比は、時間 $t$ に関して**一定**になる。
- いいかえれば、薬剤の有効性は、患者の生存期間に依らず一定であることを仮定している。

# 2つの生存曲線の比較：生存曲線とハザード比

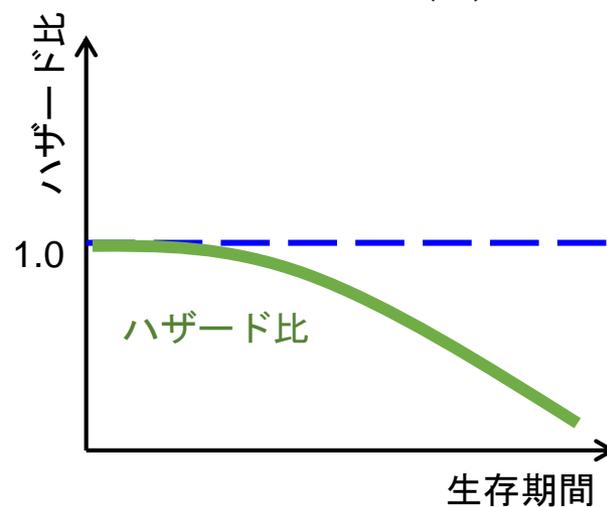
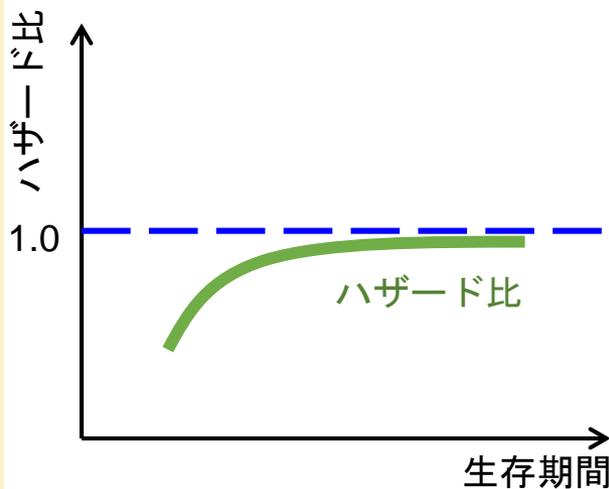
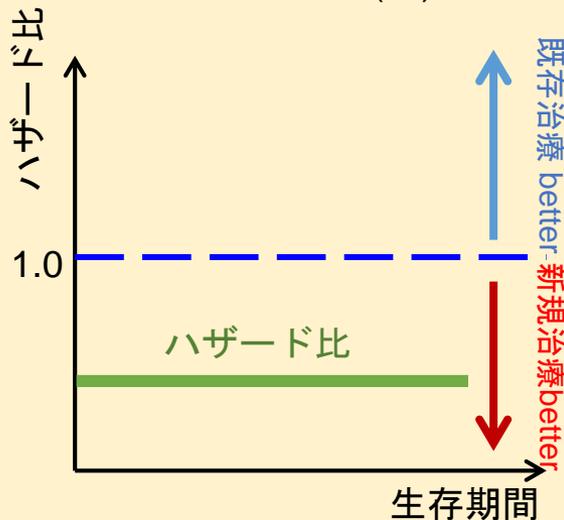
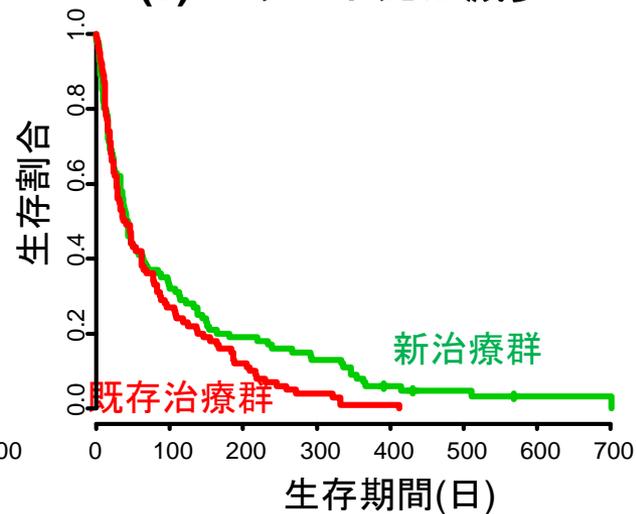
(a) ハザード比が一定



(b) ハザード比が上昇

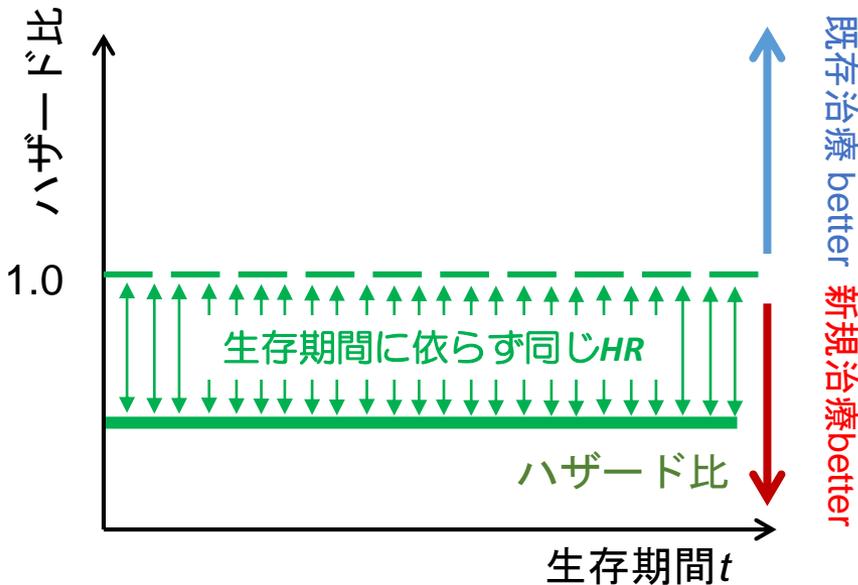


(c) ハザード比が減少



生存曲線の比較で最も用いられているログランク検定は、ハザード比が一定であることを仮定している。

# ログランク検定

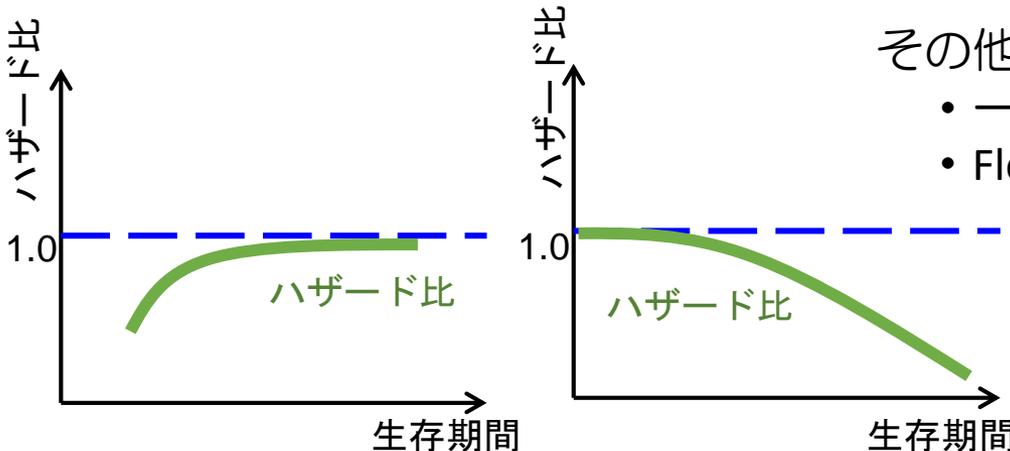


## ログランク検定における仮説

- 帰無仮説 $H_0$  :  $HR=1$  (ハザード比は1である)
- 対立仮説 $H_1$  :  $HR \neq 1$  (ハザード比は1でない)

言い換えれば、ログランク検定は、イベント・リスク(死亡リスク)が群間(薬剤)で異なるか否かを評価する検定方法である。

## 比例ハザード性が満たされない場合

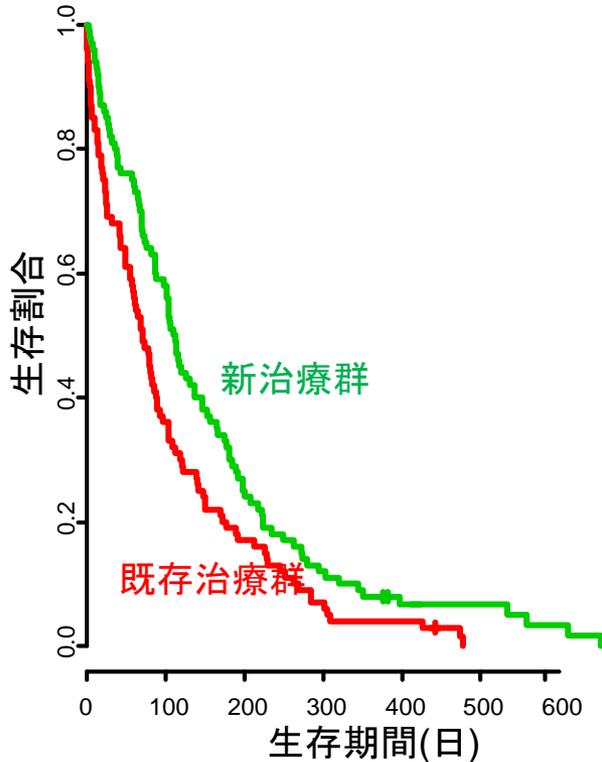


## その他の方法

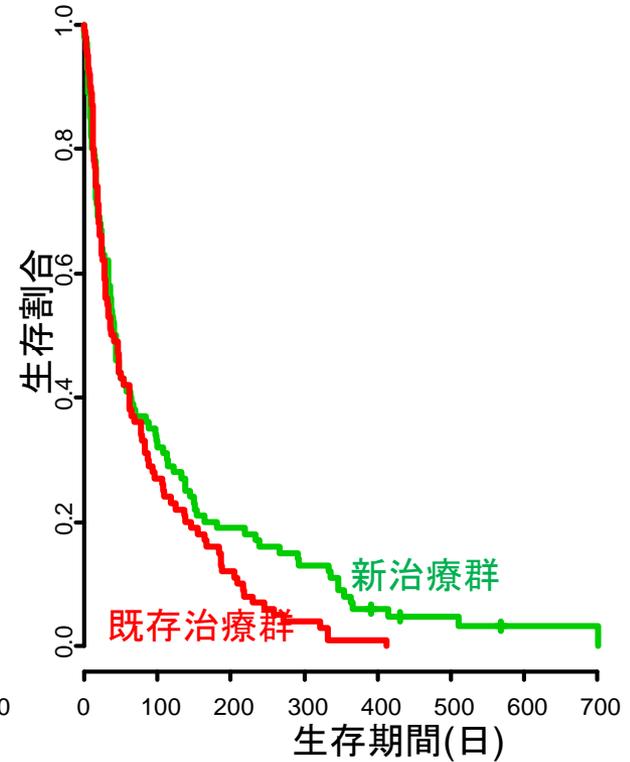
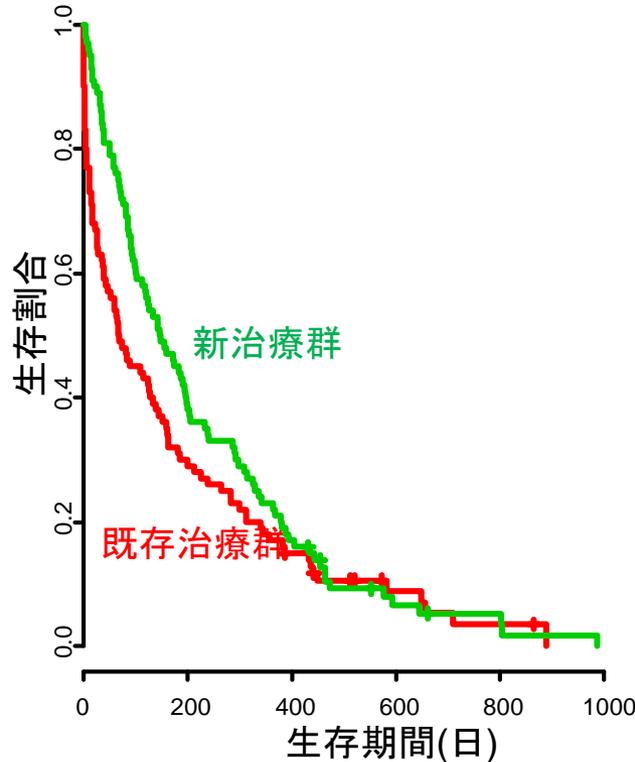
- 一般化Wilcoxon検定
  - Fleming-Hamirton検定(一般化順位検定)
- などの方法があるが、実際の論文では、あまり用いられていない。

# 例示

(a) 全観察期間に差がある場合  
比例ハザード性 ○



(b) 観察期間前期に差がある場合 (c) 観察期間後期に差がある場合  
比例ハザード性 ×



■ ログランク検定 ( $p=0, q=0$ )  
: p値=0.008

■ 一般化Wilcoxon検定 ( $p=1, q=0$ )  
: p値=0.003

■ Fleming & Harrington検定  
( $p=0, q=1$ )  
: p値=0.003

■ ログランク検定 ( $p=0, q=0$ )  
: p値=0.094

■ 一般化Wilcoxon検定 ( $p=1, q=0$ )  
: p値=0.002

■ Fleming & Harrington検定  
( $p=0, q=1$ )  
: p値=0.698

■ ログランク検定 ( $p=0, q=0$ )  
: p値=0.073

■ 一般化Wilcoxon検定 ( $p=1, q=0$ )  
: p値=0.585

■ Fleming & Harrington検定  
( $p=0, q=1$ )  
: p値=0.006

# いわゆる多変量解析について

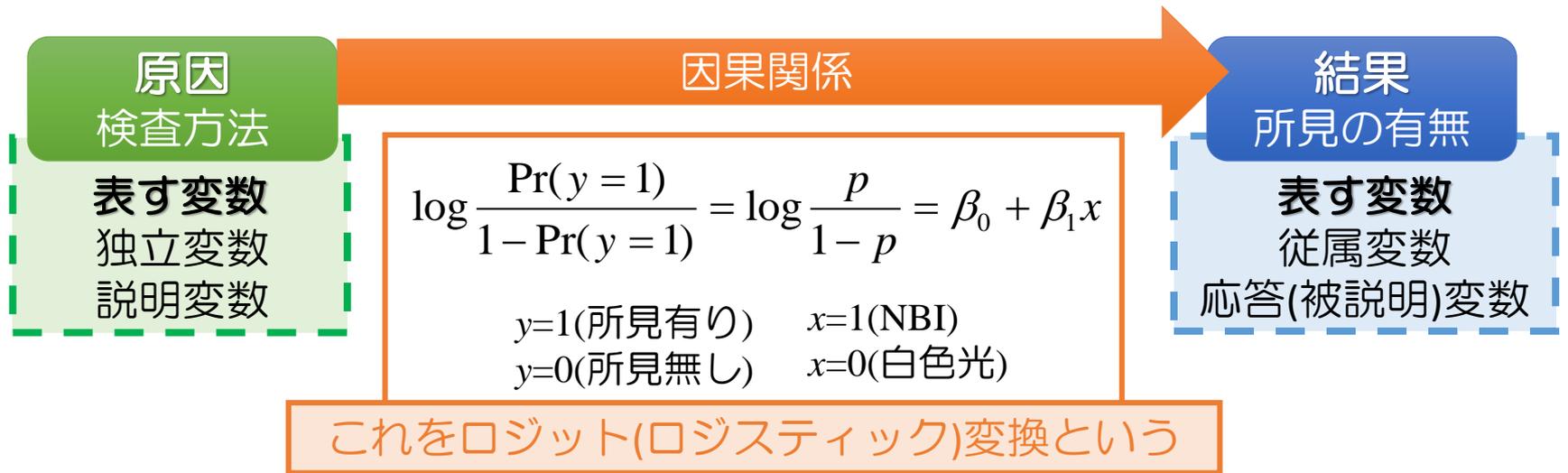
---

# いわゆる多変量解析の種類

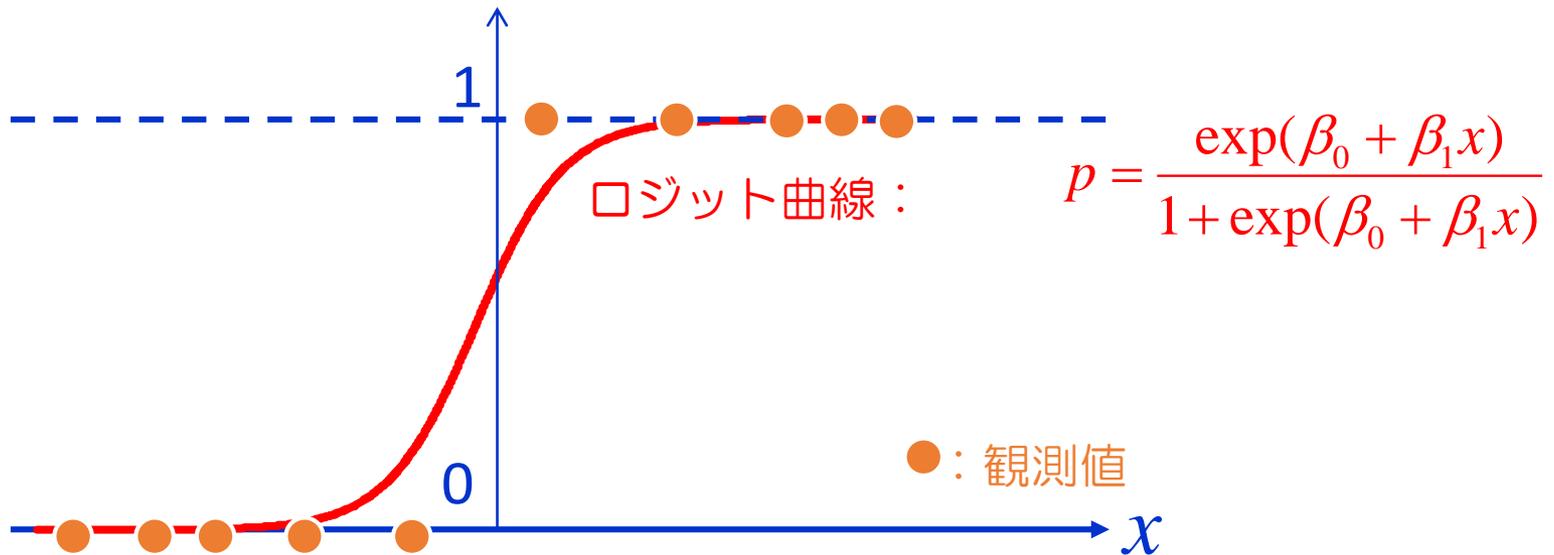
名前	応答の形式	例	係数の解釈
重回帰分析	量的	体脂肪率	回帰係数 (標準回帰係数)
ロジスティック回帰分析	2値	治療の成功/失敗	オッズ比
— 名義ロジスティック	名義	疾患の種類	オッズ比
— 比例オッズモデル	順序	疾患の進行程度	オッズ比
Poisson回帰分析	計数	ポリープの検出個数	率比
Cox比例ハザード・モデル	生存時間	がん患者の生存期間	ハザード比

- 回帰係数は説明変数の尺度に依存するため、すべての変数を標準化したもとで計算する回帰モデルの係数は標準回帰係数(標準化係数)と呼ばれる。標準化係数の絶対値の大きさを利用することで、応答に対する各説明変数の影響を評価できる。
- ロジスティック回帰、Poisson回帰、Cox比例ハザード・モデルでは、指数関数  $A = \exp(\beta)$  を計算することで、それぞれ、オッズ比、率比、ハザード比を計算できる。それぞれの解釈は下記のとおり：
  - ・ **オッズ比**：変数Xが1上がるとA倍y=1になる(例：A倍治療が成功する)。
  - ・ **率比**：変数Xが1上がるとA倍計数が上がる(例：A倍ポリープが検出される)。
  - ・ **ハザード比**：変数Xが1上がるとA倍イベントリスクが高まる(例：A倍死亡リスクがあがる)

# ロジスティック回帰分析とは



ロジスティック回帰モデルは、対数オッズに対するモデル



# ロジスティック回帰分析とオッズ比の関係

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

y=1(所見有り)      x=1(NBI)  
y=0(所見無し)    x=0(白色光)

## ■ 白色光のオッズ $odds_{白}$ を計算

$$\log \frac{p_{白}}{1-p_{白}} = \beta_0 + \beta_1 \times 0 = \beta_0 \quad \Rightarrow \quad odds_{白} = \frac{p_{白}}{1-p_{白}} = \exp(\beta_0)$$

## ■ NBIのオッズ $odds_N$ を計算

$$\log \frac{p_N}{1-p_N} = \beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1 \quad \Rightarrow \quad odds_N = \frac{p_N}{1-p_N} = \exp(\beta_0 + \beta_1)$$

## ■ NBI/白色光のオッズ比 $OR$ を計算

$$OR = \frac{odds_N}{odds_{白}} = \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\beta_0 + \beta_1 - \beta_0) = \exp(\beta_1)$$

つまり、ロジスティック回帰では、オッズ比が $\exp(\beta_1)$ で表される。

# ロジスティック回帰分析における調整オッズ比

下の表は、ある疾患に対する治療法(既存治療：0，新規治療：1)と年齢(65歳未満：0，65歳以上：1)とその成功の有無(成功：1，失敗：0)に対する2個のクロス集計表である。

年齢	治療	成功/失敗		
		成功(1)	失敗(0)	合計
65歳以上(1)	新規(1)	444	567	1011
	既存(0)	56	150	206
	合計	500	717	1217
65歳未満(0)	新規(1)	286	324	610
	既存(0)	171	254	425
	合計	457	578	1035

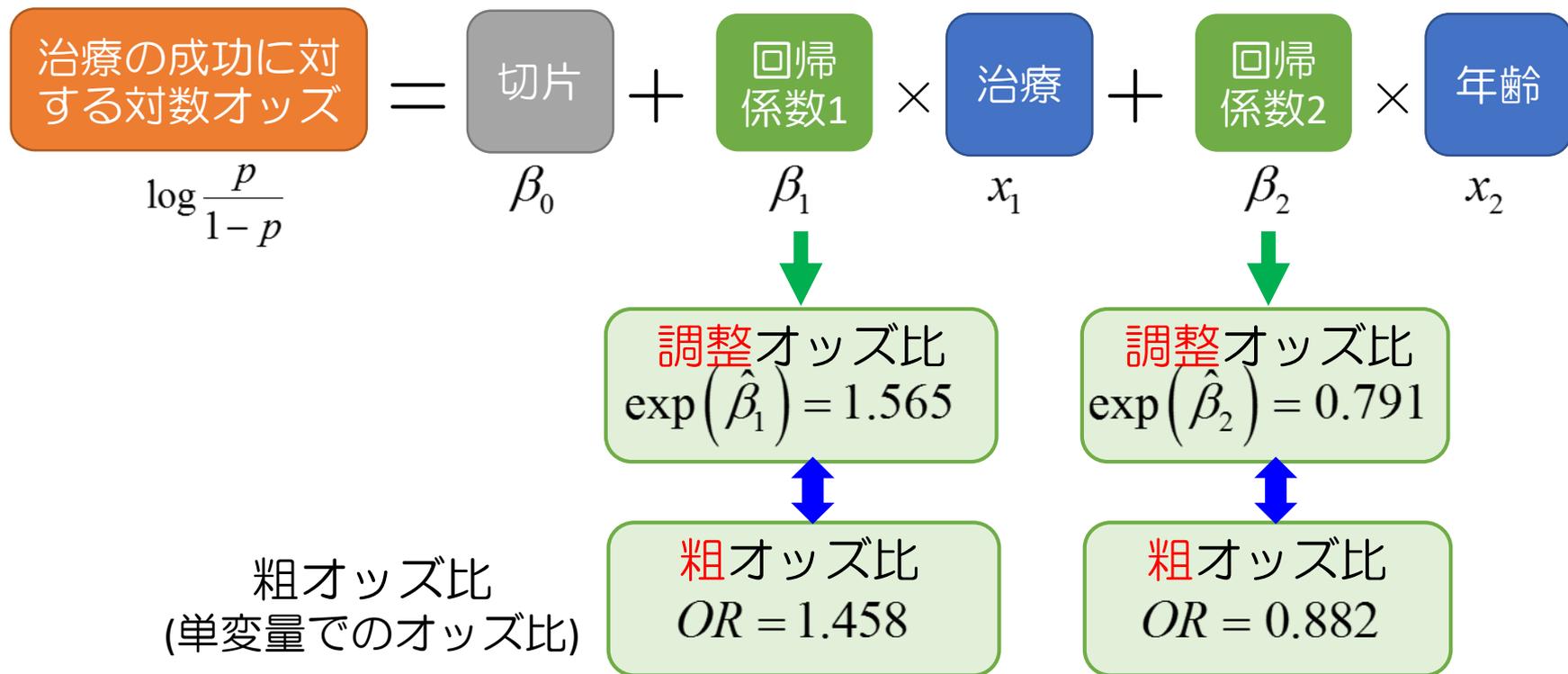
(多重)ロジスティック回帰分析のモデルは

$$\begin{array}{c} \text{治療の成功に対する対数オッズ} \\ \log \frac{p}{1-p} \end{array} = \begin{array}{c} \text{切片} \\ \beta_0 \end{array} + \begin{array}{c} \text{回帰係数1} \\ \beta_1 \end{array} \times \begin{array}{c} \text{治療} \\ x_1 \end{array} + \begin{array}{c} \text{回帰係数2} \\ \beta_2 \end{array} \times \begin{array}{c} \text{年齢} \\ x_2 \end{array}$$

既存治療：0  
新規治療：1

65歳未満：0  
65歳以上：1

# 調整オッズ比と粗オッズ比



粗オッズ比は、他の要因(共変量)を何も考慮しない。これに対して、調整オッズ比では、

- 治療のオッズ比では年齢による違いを考慮
- 年齢のオッズ比では治療による違いを考慮してオッズ比を計算している。

# 変数選択

多変量解析では、多数の説明変数を入れたほうが良いというわけではない(意味のない(応答に影響のない)変数を入れることは好ましくない)。



意味のない説明変数を削除して、最適な予測式(回帰モデル)を作る方法が**変数選択(variable selection)**である。

## (1) 変数選択の評価基準

- 検定を用いる場合(おススメできない)
- 情報量規準を用いる場合
  - 赤池の情報量規準(AIC; Akaike's Information Criteria)
  - Bayes流情報量規準(BIC; Bayesian Information Criteria)  
どちらがよいというわけではないが、BICのほうが選択される変数が少なくなる傾向にある。

## (2) 変数選択のアルゴリズム

- 総当たり法  
すべての候補を総当たりで選択する方法
- 変数増加法 (前進ステップワイズ法)  
切片のみのモデルから出発し、1個ずつ説明変数をモデルに加える方法
- 変数減少法 (後退ステップワイズ法)  
全ての説明変数を含むモデルから出発し、1個ずつ説明変数をモデルから除外する方法
- 変数増減法  
全ての説明変数を含むモデルから出発し、1個づつ説明変数を加えるのか除外するのかを評価・実施する方法

# 事例：女性の骨粗鬆症に関する世界的な研究(Hosmer et al., 2013)

## 説明変数

骨折既往歴(PRIORFRAC)

登録時の年齢(AGE)

登録時の身長(HEIGHT)

登録時のBMI(BMI)

45歳未満での閉経の有無(PREMENO)

母親の股関節骨折歴の有無(MOMFRAC)

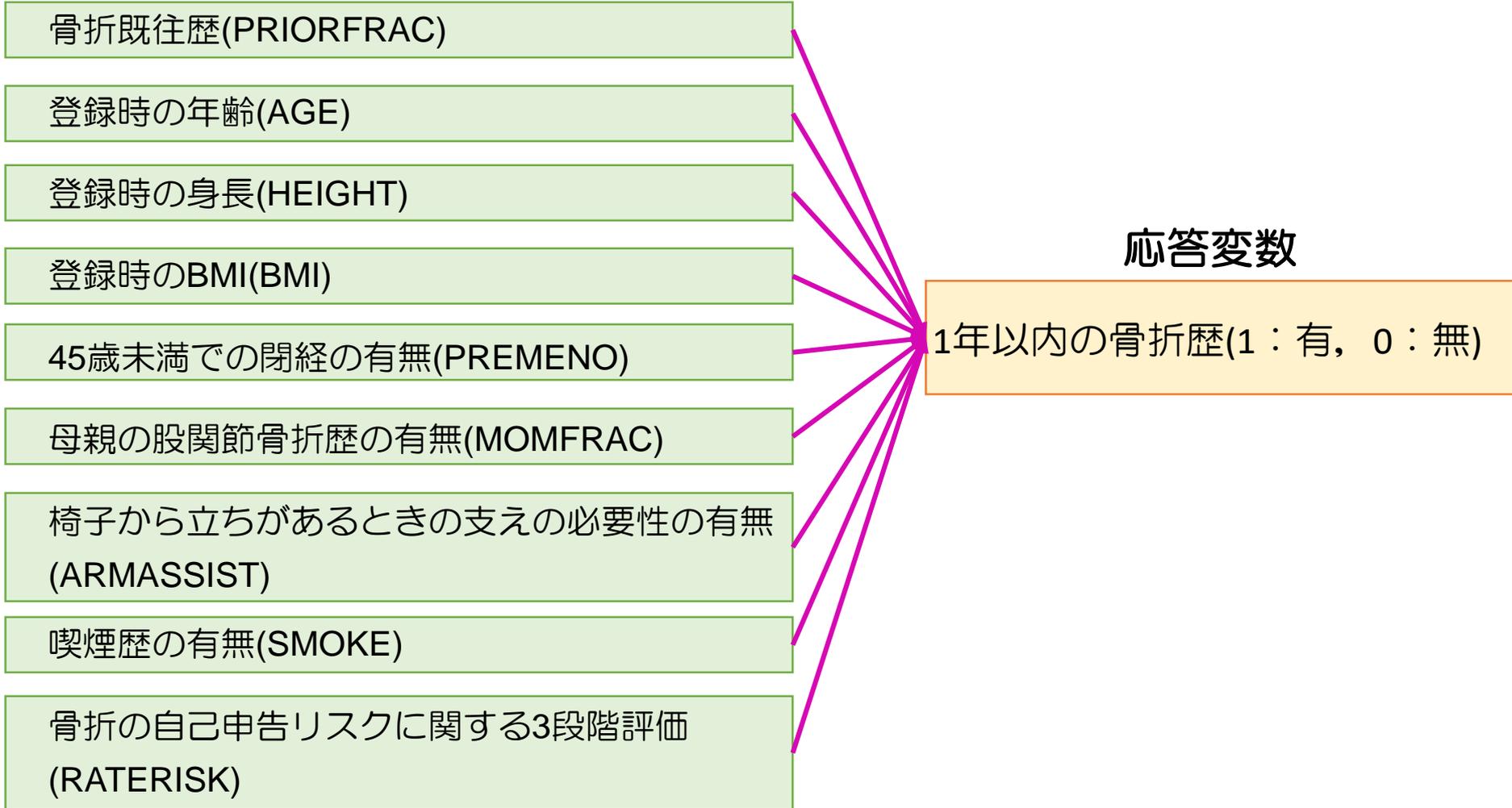
椅子から立ちがあるときの支えの必要性の有無  
(ARMASSIST)

喫煙歴の有無(SMOKE)

骨折の自己申告リスクに関する3段階評価  
(RATERISK)

## 応答変数

1年以内の骨折歴(1：有, 0：無)



# ロジスティック回帰分析の結果

	全変数を利用			変数選択		
	回帰係数	オッズ比	p値	回帰係数	オッズ比	p値
PRIORFRAC	0.611	1.843	0.011	0.649	1.913	0.004
AGE	0.011	1.011	0.457	—	—	—
HEIGHT	-0.029	0.972	0.115	—	—	—
BMI	-0.052	0.949	0.022	—	—	—
PREMENO	-0.246	0.782	0.403	—	—	—
MOMFRAC	0.627	1.871	0.032	—	—	—
ARMASSIST	0.728	2.072	0.003	0.618	1.855	0.004
SMOKE	0.275	1.316	0.525	—	—	—
RATERISK	0.314	1.368	0.039	0.406	1.500	0.005
Bayes流情報量規準	578.141			555.912		

# 変数選択の要件 (1/2)

## (1) 評価したい要因は変数選択に強制的に加える

ランダム化比較試験の結果を評価する場合、治療群を表す共変量を含まなければ意味をもたない。つまり、このような場合には、背景因子などの他の共変量を調整したうえで治療群(評価変数)を調べることに意義がある。

## (2) 変数増加法の落とし穴

標本サイズが小さい場合に、変数増加法を用いて変数選択を行う場合、結果の解釈が困難なモデルを選択することがしばしばある。また、本当は必要な共変量に取り込まれる前に変数選択が終了する場合がある。

## (3) 多数の共変量(項目)がある場合の留意点

医学系研究では、多数の調査項目(共変量)を評価に用いることは少なくない。このような場合には、全ての共変量を用いて変数選択を行うのではなく、事前スクリーニングを行うことが推奨される。事前スクリーニングでは、共変量毎に単変量解析(1個の共変量による回帰モデルを推定する)を実施し、その回帰係数に対する検定(回帰係数が0であるか否かを評価する検定)のp値や回帰係数(オッズ比, ハザード比)を用いる。

## 変数選択の要件 (2/2)

### (4) 欠測が多い共変量(項目)には注意が必要である

多変量解析では、共変量のなかで1個でも欠測があれば、その被験者を削除しなければならない。そのため、欠測が多い共変量をモデルに含めると、多くの被験者を削除することになる。また、観測方法が煩雑な場合には、欠測が多くなる傾向にある。そのため、このような共変量は、予め変数選択の候補から覗いておくことが望ましい。

### (5) 可能であれば総当たり法を用いる

変数増加法や変数減少法が必ずしも最適なモデルに到達するとは限らない。最適なモデルを選択できる唯一の方法は、すべての候補モデルを評価する総当たり法のみである。共変量の数が10個の場合、候補となるモデルの数は1,023個である。最近のコンピューターであれば実現不可能な数ではない(共変量の数が20個の場合には1,048,575個となり、不可能に近い数値となる)。そのため、臨床的知見あるいは、事前スクリーニングなどを用いて変数選択に用いる共変量を可能な限り少なくし、そのもとで、総当たり法によって変数選択を実施することが考えられる。

すこしおまけ：統計ソフトウェアについて

---

# 統計ソフトウェア



**EZR**

RのGUI版(医学統計主体)

GUI



**SPSS**

人文社会学系では主流



**SAS JMP**

最近導入が増えている

無償

有償



**R**

デファクトスタンダード

CUI



**SAS**

治験では必須



**STATA**

手法が豊富



**S-Plus**

Rの源流

# 統計ソフトウェアについて

## 統計ソフトウェアは非常に高価!!

統計ソフトウェアは、数十万円から数百万円の価格で販売されており、非常に高価である。

一方で、EZRというソフトウェアを使えば、無料で医学統計学を実施することができる。

The screenshot shows the homepage for EZR software. The header includes the logo of Jichi Medical University Saitama Medical Center and the text '自治医科大学附属さいたま医療センター 血液科のホームページへようこそ。'. The main content area has a navigation menu on the left and a central text area with several announcements. The right sidebar contains information about the software's availability on Amazon.co.jp, including a price of ¥3,888 and a 'Buy now' button. The main text area includes the following announcements:

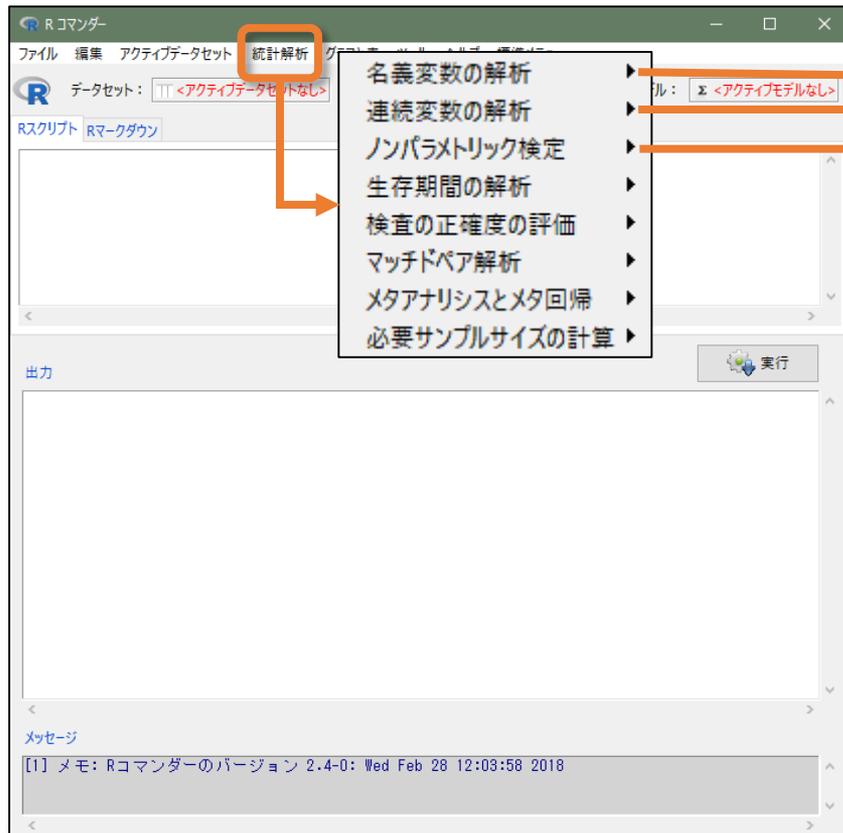
- \*\*\*\*\* Click here for English version \*\*\*\*\*
- 2012年6月11日のページで訂正後のアクセス致す。
- 2017年9月1日 EZR version 1.36公開
- 傾向スコア変数を自動的に作成するオプションを追加!
- 2014年11月初心者向けマニュアル刊行
- 2015年4月EZRMニュアル第2版刊行
- (クリックすると立ち読みすることができます)
- 2016年10月みんなのEBMと臨床研究刊行
- (EBMの基礎、新倫理指針に適合した臨床研究の進め方、論文の書き方をまとめています)
- 1,234サンプルの背景データの一覧表(Table 1)を自動作成する新機能を搭載!
- 動作確認系OS Windows XP~10, Mac OS X Snow Leopard~El Captain, Ubuntu 11.10~15.10
- EZRを使用した学術論文を発表される場合は  
New Metrics Translational 2013; 4(1): 433-438  
を参考文献として引用ください。よろしく申し上げます。  
(EZRMの使い方、変更履歴)を参照。2017年4月の時点で英文論文での引用回数が1000回を超えました。
- 国際的に有名なSoftware software databaseに  
"100% CLEAN" Software awardで掲載されました。

自治医科大学附属さいたま医療センター 血液科の神田善伸教授が統計解析環境RのGUI環境であるRコマンダーをカスタマイズしたものがEZRである。

EZRでは、医学統計学における諸種の統計的方法をマウスのみで実行できる。

ダウンロードは、ブラウザで「EZR」と入力すれば、左図のホームページがトップにくる。

# EZRでできること(1/2)



頻度分布  
比率の信頼区間の計算  
1標本の比率の検定  
2群の比率の差の信頼区間の計算  
2群の比率の比の信頼区間の計算

分割表の直接入力と解析  
分割表の作成と群間の比率の比較(Fisherの正確検定)  
対応のある比率の比較(二分割表の対称性の検定、McNemar検定)  
対応のある3群以上の比率の比較(Cochran Q検定)  
比率の傾向の検定(Cochran-Armitage検定)

二値変数に対する多変量解析(ロジスティック回帰)

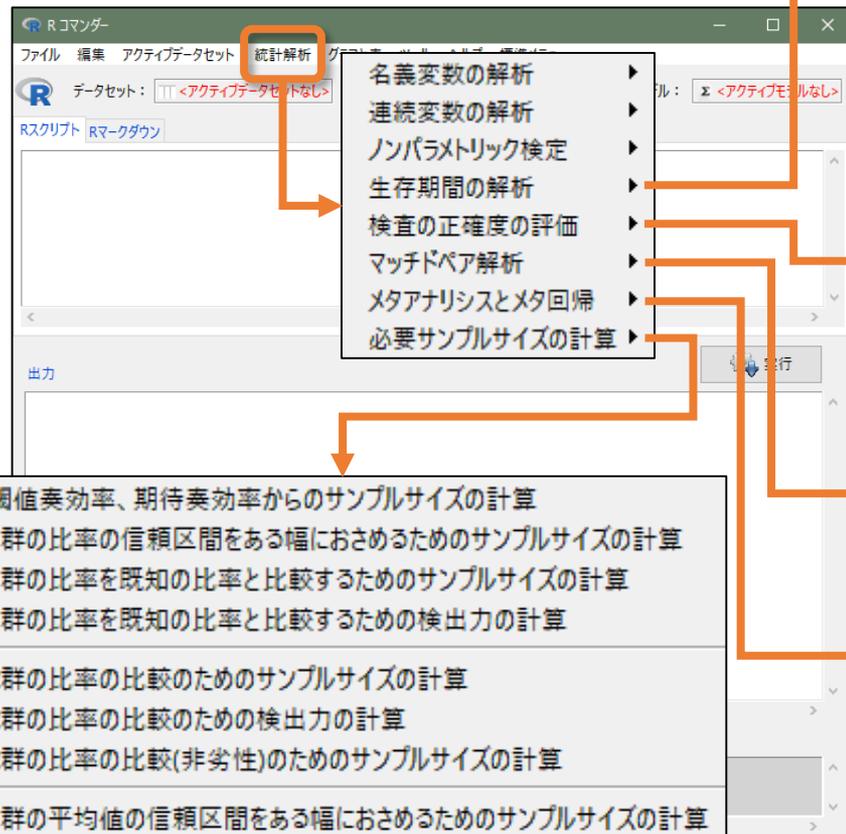
連続変数の要約  
外れ値の検定と除外(Smirnov-Grubbs検定)  
正規性の検定(Kolmogorov-Smirnov検定)  
平均値の信頼区間の計算  
1標本の平均値のt検定

2群の等分散性の検定(F検定)  
2群間の平均値の比較(t検定)  
対応のある2群間の平均値の比較(paired t検定)  
3群以上の等分散性の検定(Bartlett検定)  
3群以上の間の平均値の比較(一元配置分散分析one-way ANOVA)  
対応のある2群以上の間の平均値の比較(反復[経時]測定分散分析)  
複数の因子での平均値の比較(多元配置分散分析multi-way ANOVA)  
連続変数で補正した2群以上の間の平均値の比較(共分散分析ANCOVA)

相関係数の検定(Pearsonの積率相関係数)  
線形回帰(単回帰、重回帰)

2群間の比較(Mann-Whitney U検定)  
対応のある2群間の比較(Wilcoxon符号付順位和検定)  
3群以上の間の比較(Kruskal-Wallis検定)  
対応のある3群以上の間の比較(Friedman検定)  
連続変数の傾向の検定(Jonckheere-Terpstra検定)  
相関係数の検定(Spearmanの順位相関係数)

# EZRでできること(2/2)



- 生存曲線の記述と群間の比較(Logrank検定)
- 生存期間の傾向の検定(Logrank trend検定)
- 生存期間に対する多変量解析(Cox比例ハザード回帰)
- 時間依存性変数を含む生存期間に対する多変量解析(Cox比例ハザード回帰)
- 累積発生率(競合イベントを含む)の記述と群間の比較(Gray検定)
- 累積発生率(競合イベントを含む)に対する多変量解析(Fine-Gray比例ハザード回帰)

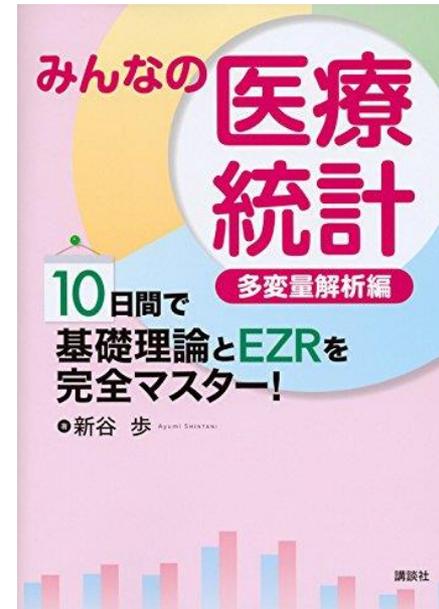
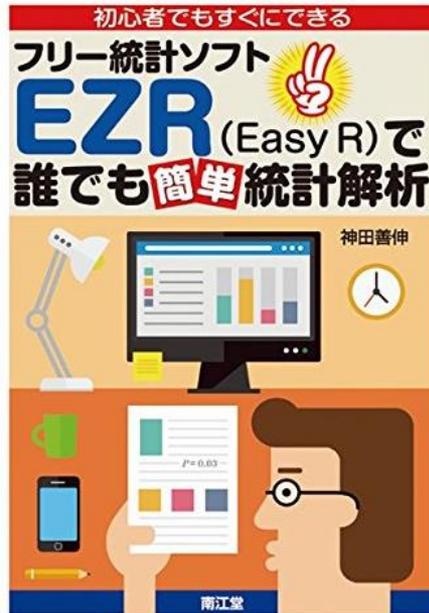
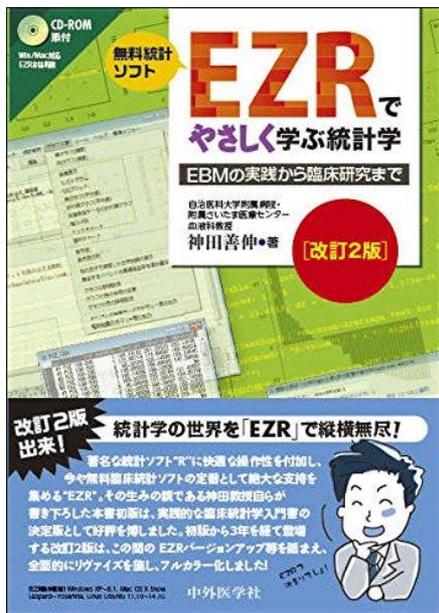
- 定性検査の診断への正確度の評価
- 2つの定性検査の一致度の評価(Kappa係数)
- 陽性的中率、陰性的中率の計算
- 定量検査の診断への正確度の評価(ROC曲線)
- 2つのROC曲線のAUCの比較
- 質問項目の信頼性の評価(Cronbachの $\alpha$ 信頼性係数)

- マッチさせたコントロールの抽出
- マッチさせたサンプルの比率の比較(Mantel-Haenzel検定)
- マッチさせたサンプルの比率の多変量(条件付ロジスティック回帰)
- マッチさせたサンプルの生存率の多変(層別化比例ハザード回帰)

- 比率の比較のメタアナリシスとメタ回帰
- 平均値の比較のメタアナリシスとメタ回帰
- ハザード比のメタアナリシスとメタ回帰

- 閾値奏効率、期待奏効率からのサンプルサイズの計算
- 1群の比率の信頼区間がある幅におさめるためのサンプルサイズの計算
- 1群の比率を既知の比率と比較するためのサンプルサイズの計算
- 1群の比率を既知の比率と比較するための検出力の計算
- 2群の比率の比較のためのサンプルサイズの計算
- 2群の比率の比較のための検出力の計算
- 2群の比率の比較(非劣性)のためのサンプルサイズの計算
- 1群の平均値の信頼区間がある幅におさめるためのサンプルサイズの計算
- 2群の平均値の比較のためのサンプルサイズの計算
- 2群の平均値の比較のための検出力の計算
- 2群の平均の比較(非劣性)のためのサンプルサイズの計算
- 対応のある2群の平均値の比較のためのサンプルサイズの計算
- 対応のある2群の平均値の比較のための検出力の計算
- 2群の生存曲線の比較のためのサンプルサイズの計算
- 2群の生存曲線の比較のための検出力の計算
- 2群の生存曲線の比較(非劣性)のためのサンプルサイズの計算

# EZRの関連書籍



EZRに関する書籍としては、製作者の神田先生、大阪市立大学の新谷歩先生が出版している。いずれの書籍もマニュアルとして利用しやすく、また、医学統計での利用を想定している点も利点の一つである。

# もっと詳しいお話を聴きたい場合は...

医学統計セミナー（18：00～19：00）

場所：和歌山県立医科大学附属病院 臨床講堂Ⅱ

日程	タイトル	内容
11/8（木）	一般的なアウトカムの評価： 量的変数・質的変数の評価	<ul style="list-style-type: none"><li>• t検定、Wilcoxon検定（対応あり／なし）</li><li>• カイ2乗検定、Fisherの正確検定</li></ul>
12/21（金）	特殊なアウトカムの評価：生 存期間・臨床検査値の評価	<ul style="list-style-type: none"><li>• Kaplan-Meier法、ログランク検定</li><li>• 感度、特異度、ROC曲線</li></ul>
1/17（木）	多変量解析と統計的因果推論	<ul style="list-style-type: none"><li>• 重回帰分析、ロジスティック回帰分析</li><li>• Cox比例ハザードモデル</li></ul>
2/5（火）	観察研究と傾向スコア分析	傾向スコア分析の方法と統計ソフトによる解析の紹介
3/26（火）	メタアナリシス	メタアナリシスの方法と統計ソフトによる解析の紹介

どなたでも参加できます。

連絡先： [crc@wakayama-med.ac.jp](mailto:crc@wakayama-med.ac.jp)

ご清聴ありがとうございました



# バックアップ

---

- 癌臨床研究における生存期間に関する考え方
- 無作為化比較試験における割り付けについて
- ITT(Intent-To-Treat)の原則とPPS(Per Protocol Set)
- 優越性, 非劣性について
- 交互作用とは
- 多変量解析における留意点

# 全生存期間とサロゲートエンドポイント

がん臨床研究での真のエンドポイントは、全生存期間である。一方で、全生存期間には、

- 予後良好な場合、打ち切り(censoring)が多くなり、不完全データになりやすい。
- 長期のフォローアップ期間が必要になる。
- 後続ラインの影響を受ける。とくに、治療法がクロスオーバーする場合は深刻である。

のような問題があることから、代替エンドポイント(サロゲートエンドポイント)が用いられる。その代表的なものの一つが無増悪生存期間(PFS)である。

Petrelli et al.(2015)およびHamada et al.(2016)は、膵癌に対する臨床試験において、PFSがOSのサロゲートエンドポイントになり得ることを指摘している。

# 全生存期間は治療効果を「常に」評価し得るものか？

## JCOGにおけるエンドポイントの定義

エンドポイント	イベント (いずれか早いもの)			打ち切り日
全生存期間 Overall Survival (OS)	あらゆる死亡	—	—	最終生存確認日
無増悪生存期間 Progression-Free Survival (PFS)	あらゆる死亡	増悪／再発	—	最終無増悪確認日
無発生存期間 Relapse-Free survival (RFS)	あらゆる死亡	再発	—	最終生存確認日 (最終無再発確認日)
無病生存期間 Disease-Free survival (DFS)	あらゆる死亡	再発	2次がん	最終生存確認日 (最終無再発確認日)
治療生存期間(治療完了なし) Time to Treatment Failure (TTF)	あらゆる死亡	治療中止	—	最終治療継続確認日
治療生存期間(治療完了あり) Time to Treatment Failure (TTF)	あらゆる死亡	治療中止	プロトコル治療完了後の増悪／再発	最終治療継続確認日 (プロトコル治療中) 最終無増悪確認日 (プロトコル治療後)

- 術後補助化学療法において、PFSを利用することは多いように思われる。
- 術前補助化学療法、手術については、DFSを利用することが多いように思われる。
- 術後補助化学療法の継続性を評価することを意図して、TTFを副次的に評価することも増えてきている。

# ランダム化比較試験の「ランダム化」とは

ランダム化は、比較試験において最も重要である。これにより、担当医師による**選択バイアス**を回避できるだけでなく、盲検化、層別因子による割付けなどを通して、バイアスの回避を試みる。

統計的データ解析が適切であっても、データが不適切であれば何にもならない。

## ■ 用語の解説：層別因子(Stratification factors)

ランダム化のみでは、群間の患者のバイアスを除外することはできない。

そのため、予後に影響を与える最も重要な因子について、適度に群間でバランスをとる必要がある。バランスをとるための患者背景因子を「**層別因子**」と呼ぶ。

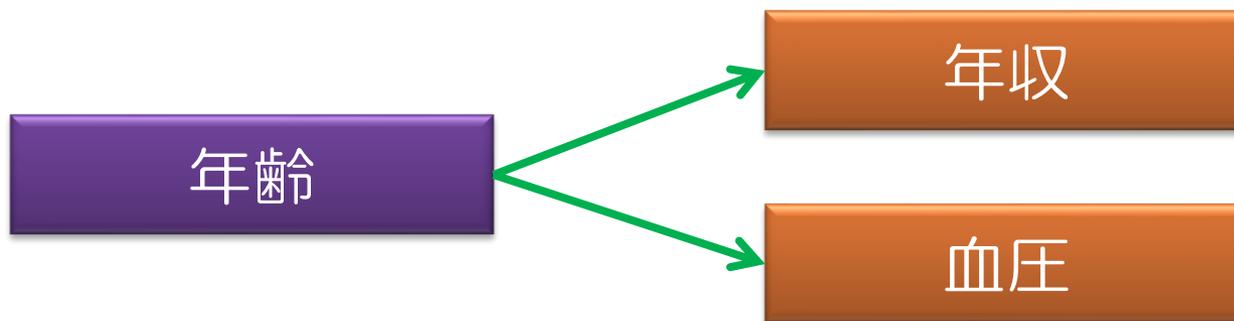
ただし、多くの層別因子を用いることは、ランダム化に対するバランスがとれなくなる恐れがある。そのため、Green, Benedetti & Crowley (2003)は層別因子としてせいぜい3個程度が妥当であると指摘している。

## 割付因子での注意：交絡に注意

ある誤った疫学研究で、「高血圧症の患者ほど年収が高い」という報告がなされた。果たして本当だろうか

一般に、年齢が高くなるほど

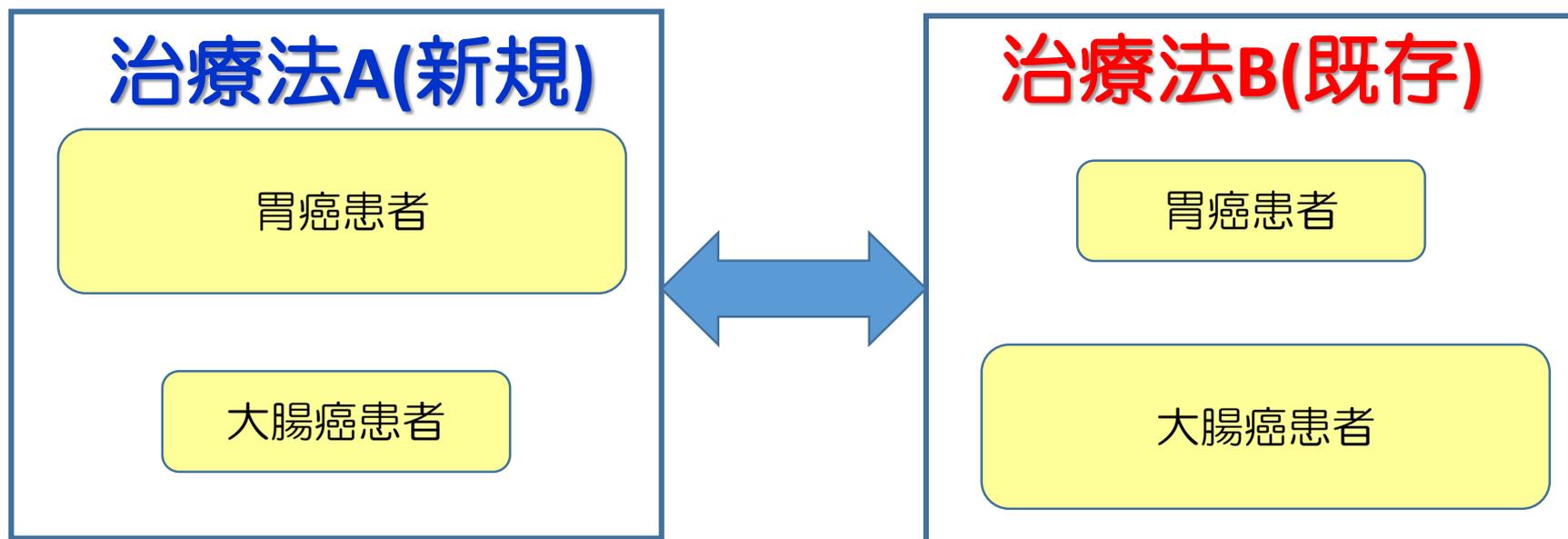
- 血圧が上昇する(血管および増加の老化およびメタボリックなどが原因)
- 年収が増加する(給料は年齢が上昇すれば増加する)ことは知られている。したがって、この事実は



となり、実際には年収と血圧には関係がない。このとき、年齢は**交絡因子**と呼ばれる。

# 割付に関する経験則

胃癌, 大腸癌患者に対する外科的手術での縫合方法に関する  
無作為化比較第III試験



割付因子(エンドポイントに影響を与える因子)である手術部位において、割付結果のアンバランスが出てしまった。この治療法は、その後の結果で、大腸癌において新規治療法が既存治療法よりも顕著な有効性が認められていることがわかった。しかしながら、割付の失敗で治療法Aには有効性が高かった大腸癌患者が少ないため、Overallでの有効性が認められなかった。このように、「統計的評価以前に割付の失敗があると、試験の品質にもかかわる大きな損失に繋がる」ので注意が必要である。

# 割付の方法

割付方法	
静的割付 (登録前に予め割付コード表を規定する)	単純無作為化法
	ブロック無作為化(置換ブロック法)
	層別無作為化法
動的割付 (層別因子毎に登録のたびに割付を行う)	最小化法 拡張最小化法(層別因子に連続量を追加)

- 静的割付の3種類にはランダム化が含まれているため、確率的な割付が行われる。他方、最小化法では、層別因子により決定論的に決定される。ICH-E9では、完全な決定論的な割付に対して、盲検性を保障するための十分な注意が喚起されている。
- そのため、完全に決定するのではなく、ランダム化を加味するようなデザインが研究されている。

# 静的割付の方法：層別無作為化法

## 割付の例示

群：2群(治療群A, 治療群B)

層別因子：層別因子(年齢( $\leq 65$ ,  $> 65$ ), PS(0,1))

## 層別無作為化法(層別ブロック法)

症例番号

年齢( $\leq 65$ ), PS(0)	A B A B	B A B A	B B A A	•	•	•
年齢( $\leq 65$ ), PS(1)	A B B A	A B A B	A A B B	•	•	•
年齢( $> 65$ ), PS(0)	A A B B	B B A A	B A B A	•	•	•
年齢( $> 65$ ), PS(1)	B A A B	A B A B	A B B A	•	•	•

層別因子の組み合わせ毎に無作為化が行われる(ブロック無作為化)

# 動的割付の方法：最小化法

- 患者が登録される毎に(最初の登録患者は完全無作為化する), それぞれの層別因子毎に症例数の均等化を図る方法である.
- 無作為性を犠牲にする一方で, 層別因子の均等化を登録時点で図りたいというデザインである.

## 例示：前立腺がん患者に対する例(丹後,2003)

年齢

層別因子	A剤	B剤	合計
65歳未満	3	3	6
65歳以上	12	13	25
合計	15	16	31

分化度

層別因子	A剤	B剤	合計
高分化	5	4	9
中分化	8	9	17
低分化	2	3	5
合計	15	16	31

疾病

層別因子	A剤	B剤	合計
A	3	3	6
B	4	4	8
C	8	9	17
合計	15	16	31

層別因子：年齢，病期分類，分化度

登録された症例の層別因子で分けたときの割り付け症例数の合計を比較する(同数の場合はランダム) 参考：丹後俊郎：無作為化臨床試験(第2版), 朝倉書店, 2017.

### 「65歳以上, 疾病分類=D, 分化度=高」の場合

	A剤	B剤
年齢	12	13
疾病分類	8	9
分化度	5	4
合計	<b>25</b>	26



A剤に割り付けられる

実際の適用場面では, 決定論的には割り付けを決めるのではなく, 一定の確率(例えば80%)でA剤に割り付けといった形式をとる(Biased Coin design).

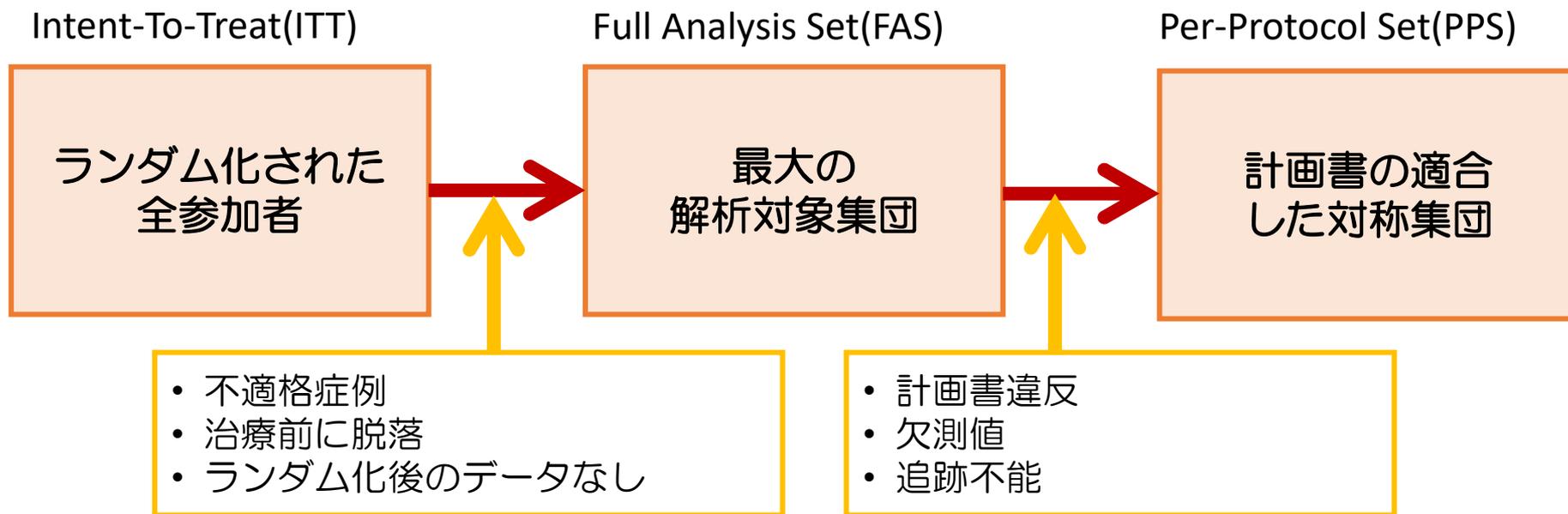
### 「65歳未満, 疾病分類=D, 分化度=高」の場合

	A剤	B剤
年齢	3	3
病気分類	8	9
層化因子	5	4
合計	16	16



乱数によってランダム(確率1/2)に割りつけ

# ITT(Intent-To-Treat)の原則とPPS(Per Protocol Set)



- 選択バイアスの観点から、原則的には、ITT症例(or FAS)での解析が行われる(ICH-E9ガイドラインにも記載).
- 純粋に治療法(あるいは薬剤)の評価を行う場合には、PPSを主解析とする場合がある。例えば、安全性が主要評価項目の場合である。
- また、非劣性試験では、研究の質が悪いと群間差が小さくなることや、プロトコルのアドヒアランスを評価するために、PPSで実施することがある。

# 例示

## ■ ToGA STUDY

ToGA Studyは、切除不能な局所進行性、再発 and/or 転移性のHER2陽性の患者に対する化学療法単独群（5-FU /CDDP あるいはCapecitabine /CDDP）とTrastuzumab併用群を比較する多施設共同無作為化非盲検試験

対照群 Capecitabine or 5-FU + CDDPを3週間毎に6サイクル

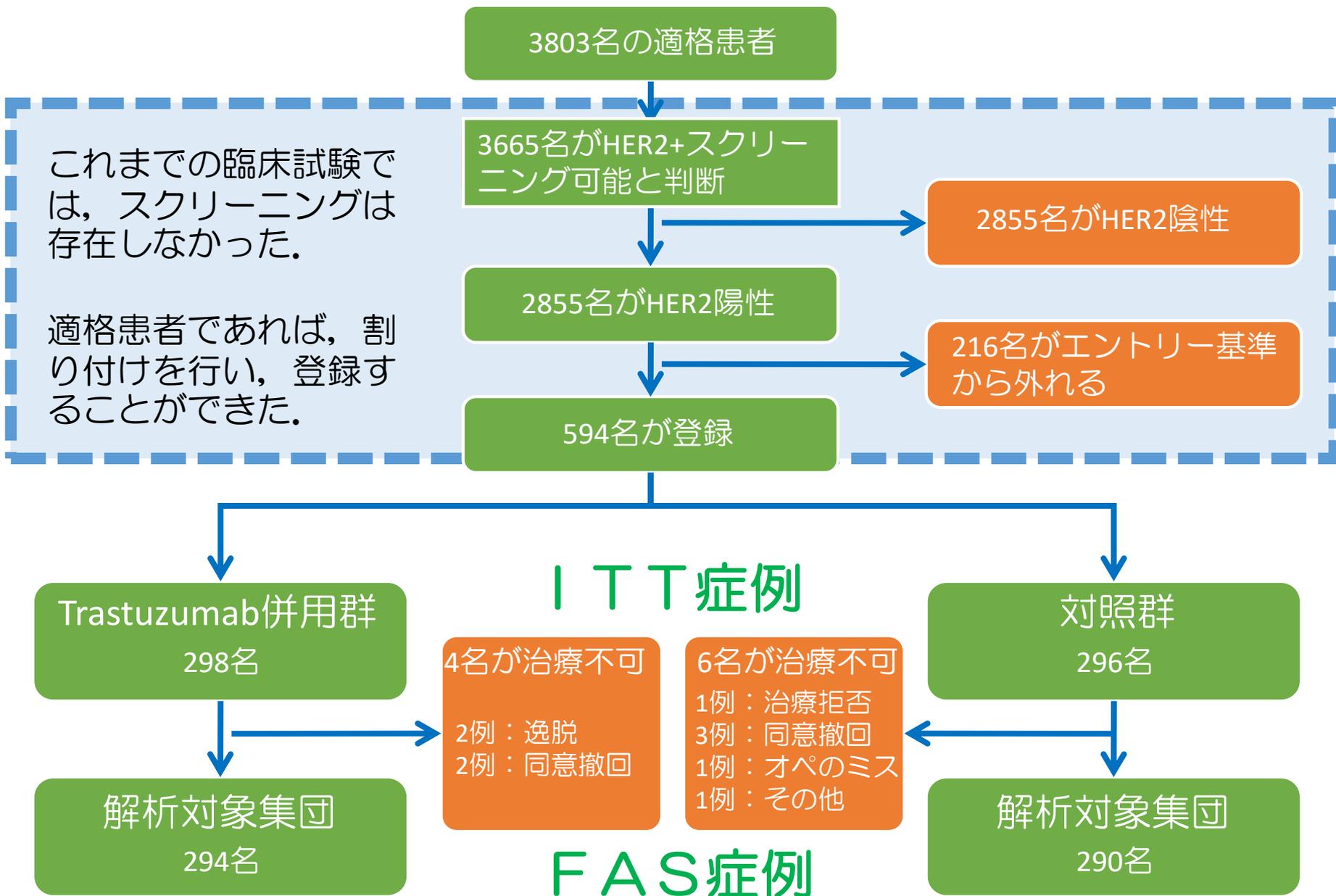
VS

Trastuzumab 併用群 対照群の治療 + Trastuzumab 6mg/kgを3週毎にPDまで

主要評価項目は、全生存期間(OS)である。

Y-J Bang, et al. Trastuzumab in combination with chemotherapy versus chemotherapy alone for treatment of HER2-positive advanced gastric or gastro-oesophageal junction cancer (ToGA): a phase 3, open-label, randomised controlled trial, *The Lancet*, 28, 687 - 697, 2010.

# 症例の整理



# 臨床試験のデザイン

第III相試験の目標：新規治療の有効性は期待できるが，toxicityが悪そうな場合(Toxic New)

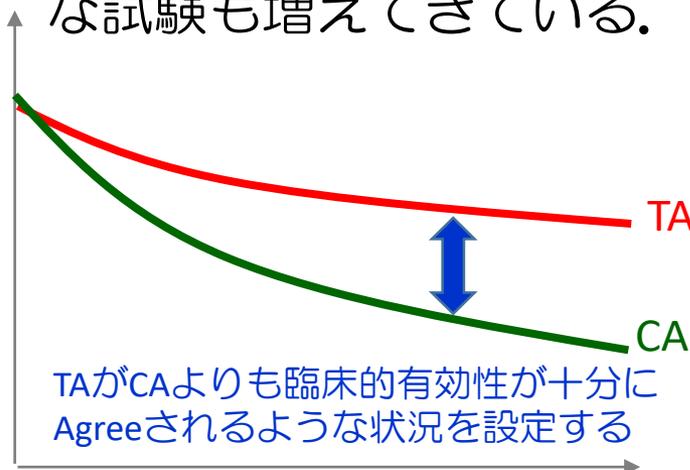
比較

既存治療  
(CA:Control Arm)

有効性：CA < TA  
安全性：CA > TA

試験治療  
(TA:Treatment Arm)

- 上記の状況を証明するには，**優越性試験**のみが許容される。
- つまり，試験治療が既存治療に対して臨床的に意味がある差異として認識されなければ，安全性を担保できない。
- 最近では，分子標的薬や免疫チェックポイント阻害剤の併用に伴うToxic Newな試験も増えてきている。



優越勢デザインでは，TAのハザード $H_T$ がCAのハザード $H_C$ に比べて臨床的に有効であることを統計的に証明しなければならない，次の仮説

帰無仮説： $H_T = H_C$

対立仮説： $H_T < H_C$  (片側 or 両側対立仮説)

を検定する(両側の場合には”≠”を証明)．このときの臨床的に有効な差異は，標本サイズの選定のときに考慮される。

# 臨床試験のデザイン

## 第III相試験の目標：新規治療の安全性(など)のBenefitがある場合

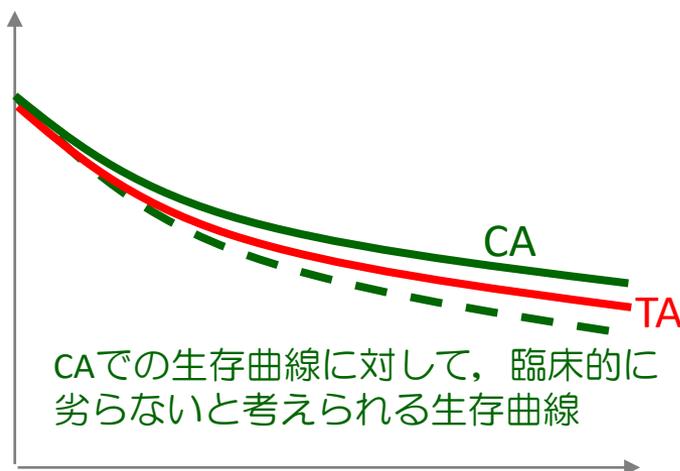
比較

既存治療  
(CA:Control Arm)

有効性：CA  $\doteq$  TA  
安全性：CA < TA

試験治療  
(TA:Treatment Arm)

- 上記の状況を証明するには、**非劣勢試験**として行えることがある。
- つまり、試験治療が既存治療に対して有意に劣らなければ、ひとつの治療アームとしての価値があることを示すことができる。
- 安全性だけでなく、たとえば、既存治療が入院だったのに対して、新規治療が外来である、あるいは治療コストが安価になるなどのBenefitも考えられる。



非劣勢デザインでは、TAのハザード $H_T$ がCAのハザード $H_C$ に比べて臨床的に劣らない範囲 $\Delta$ を決定し、次の仮説

帰無仮説：  $H_T = H_C + \Delta$

対立仮説：  $H_T < H_C + \Delta$  (片側対立仮説)

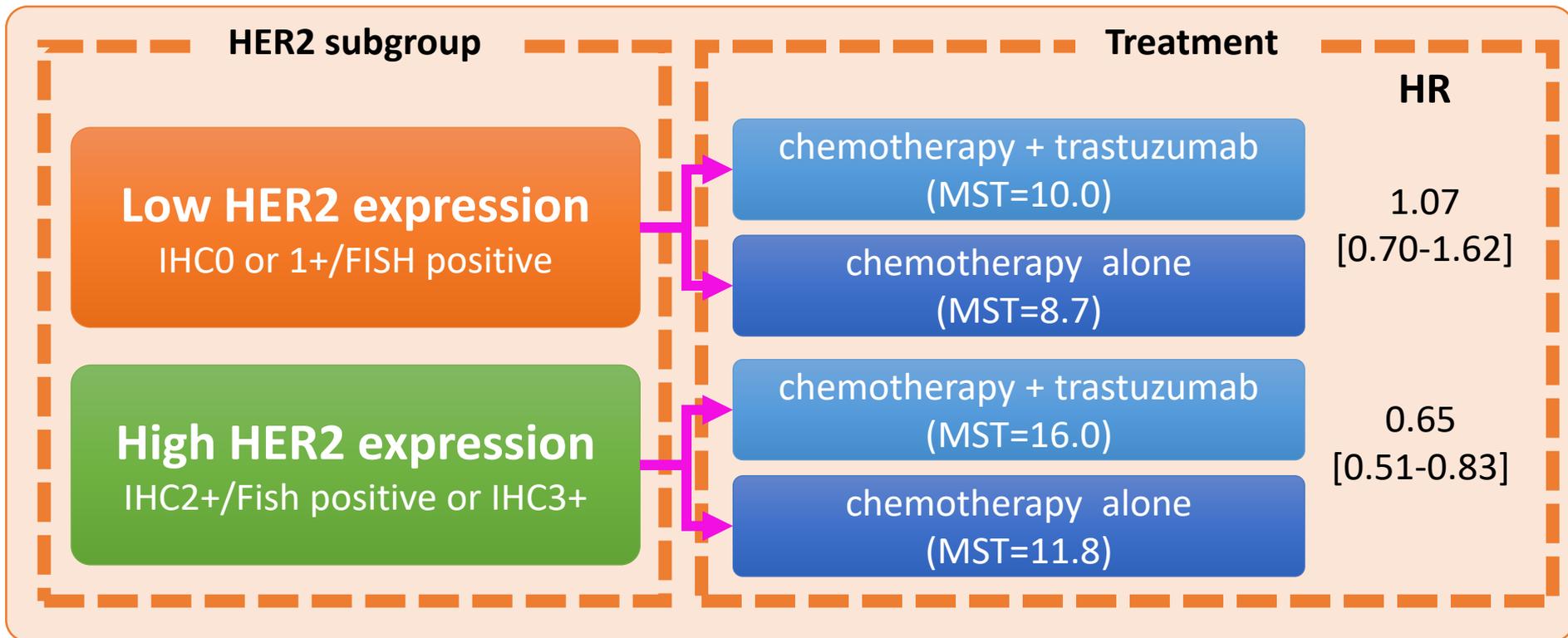
を検定する。すなわち、TAの治療は、CAに比べて、 $\Delta$ (臨床的に許容される下限値)まで劣ることはない治療であることを証明している。

# 交互作用とは(1/2) : ToGA study

## ToGA study(Bang et al.: *The Lancet*, 376, 687-697, 2010)の一文

There was evidence of a **significant interaction test** (p=0.036) between treatment and the two HER2 subgroups (high HER2 expression vs low HER2 expression).

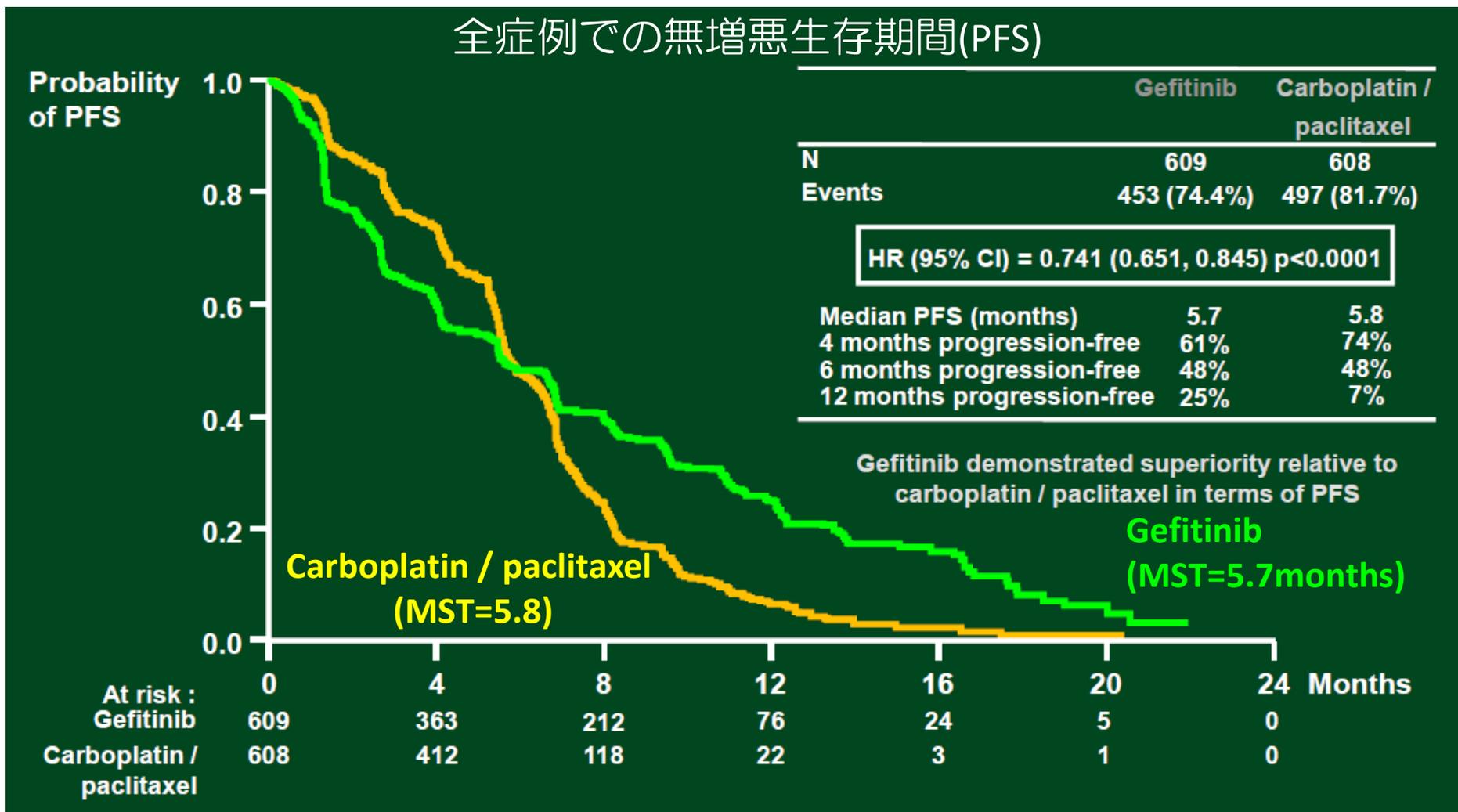
切除不能な局所進行性、再発 and/or 転移性のHER2陽性の患者に対する化学療法単独群（5-FU+CDDP (FP療法)あるいはCapecitabine +CDDP(XP療法)）と化学療法+Trastuzumab併用群を比較する国際共同無作為化非盲検試験



Low HER2ではハザード比(HR)が有意でないが, High HER2では有意に chemotherapy + trastuzumabのハザード比が低い.

# 交互作用とは (2/2) : IPASS Study

IRESSA Pan-Asia Study(IPASS)試験は、アジア地域における非盲検ランダム化並行群間比較試験である。そこでは、進行非小細胞肺癌患者1,217例を対象に1次治療としてのGefitinib(イレッサ)の有効性・安全性および忍容性をCarboplatin+Paclitaxel併用化学療法と比較した試験である。

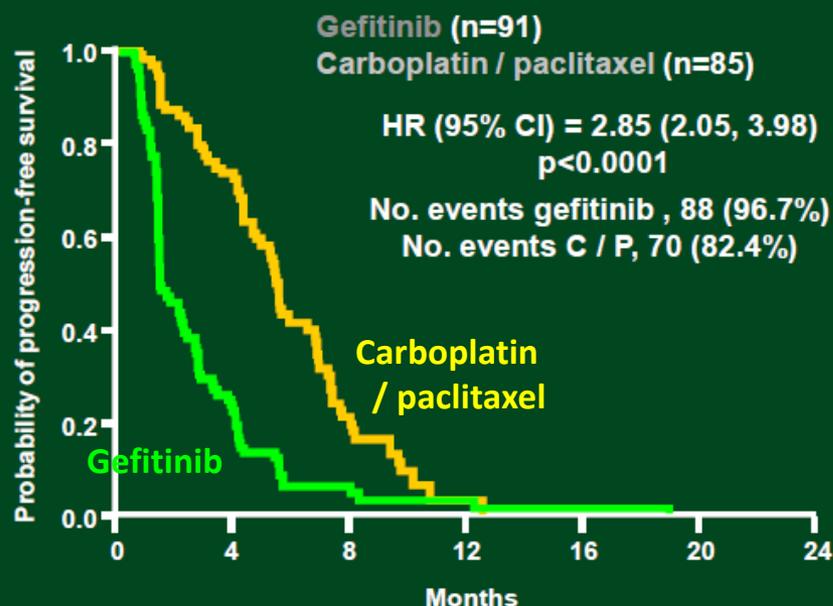
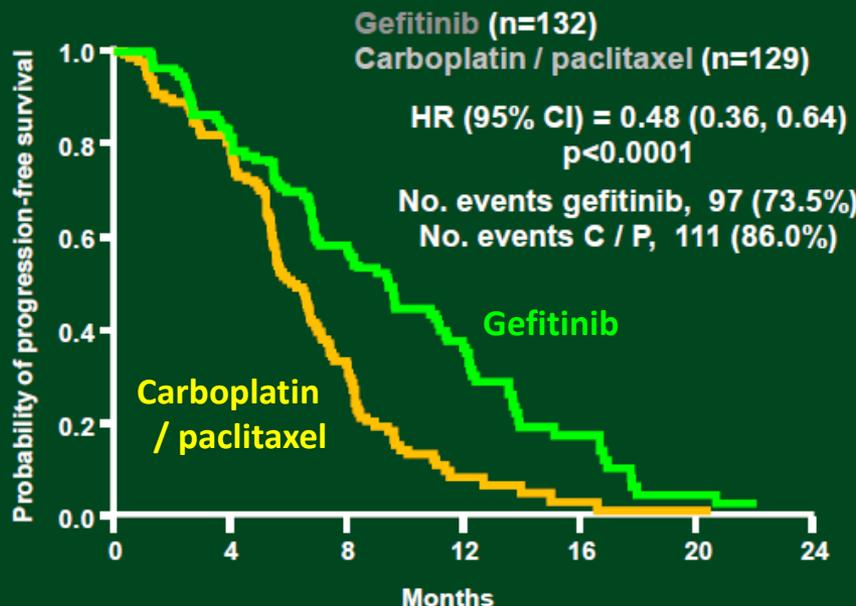


Mok, T.S. et al. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma, N.Engl.J Med., 361(10), 947-57, 2009.

EGFRで分けると...

EGFR mutation+

EGFR mutation-



At risk :

Gefitinib	132	108	71	31	11	3	0
C / P	129	103	37	7	2	1	0

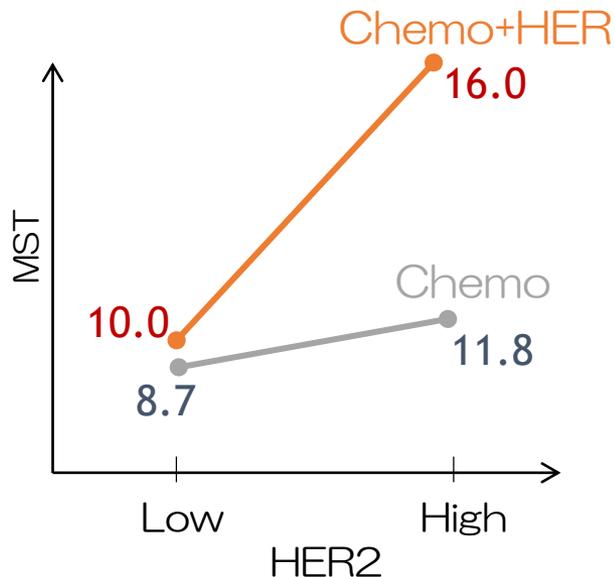
Gefitinib	91	21	4	2	1	0	0
C / P	85	58	14	1	0	0	0

Treatment by subgroup interaction test, p < 0.0001

# 交互作用とは？

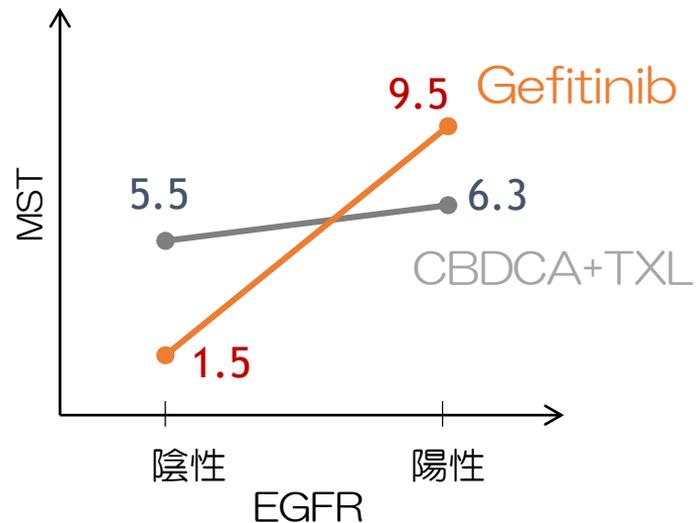
## ToGA Study

(a) MSTが上昇



## IPASS Study

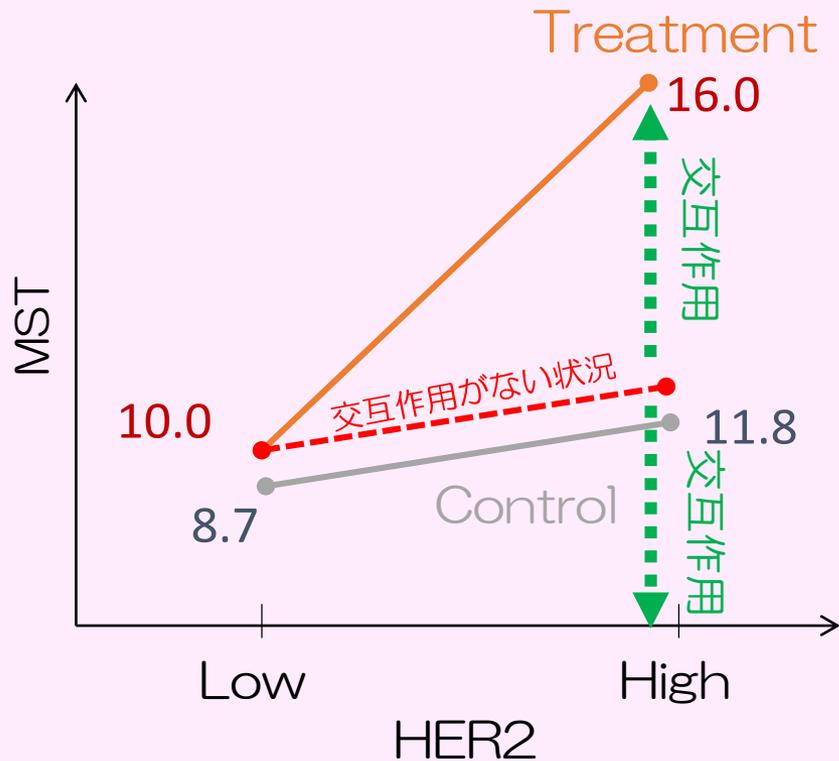
b) MSTが交差



何らかの要因によって、抗がん剤の効果が逆転したり、相乗効果によって、より効果が上昇するような現象を交互作用(interaction)という。交互作用は、治療方法の一般可能性に繋がる。

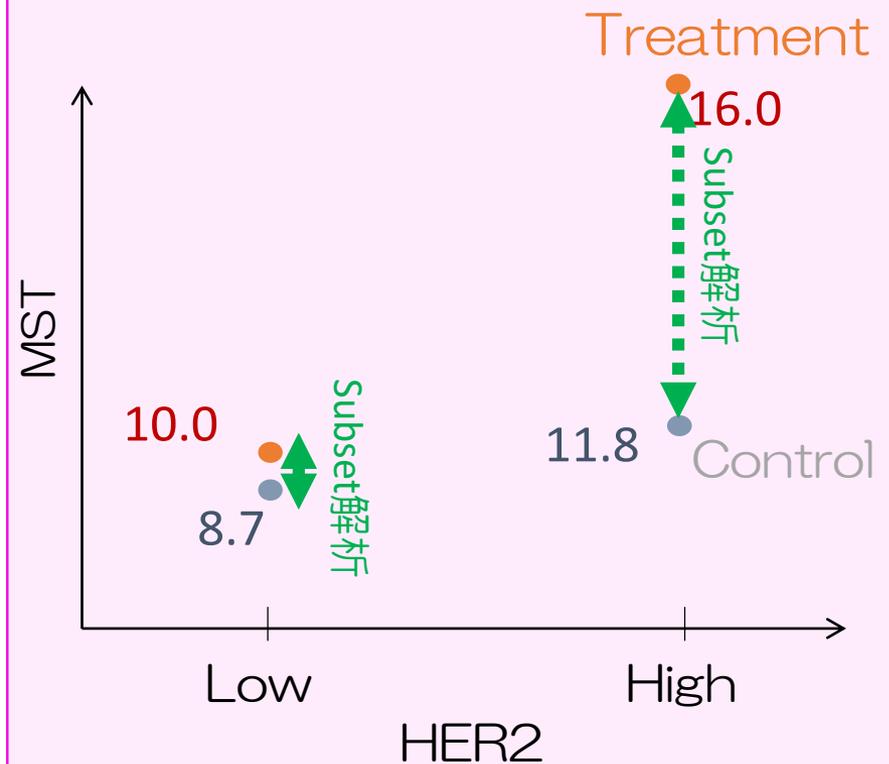
# 交互作用の評価とサブセット解析の違い

## 交互作用



HER2と治療による相乗効果(交互作用)が認められることを検定によって明らかにした。しかし、個々のサブセットで差異があるかはわからない。

## Subset解析



HER2Lowでは差異がないものの、HER2Highでは有意差が認められた。つまり、Her2Highの患者にのみ治療法が有効であった。

# 共変量が順序カテゴリカル変数の場合

ECOG Performance Status(PS)による全生存期間への影響を評価する場合(被験者の適格基準がPSが0~2であるとする)

$$\lambda = \lambda_0(t) \cdot \exp\{\beta_1 \times (\text{PS1}) + \beta_2 \times (\text{PS2})\}$$

$$\exp\{\beta_1\} : \text{PS1/PS0のHR}, \quad \exp\{\beta_2\} : \text{PS2/PS0のHR}$$



ハザード比は、基準となるカテゴリあるいはペアワイズに計算しなければならない。  
一方で、PSが全生存期間に影響するか否かは、回帰パラメータ・ベクトル $\beta$ が有意であるか否か、すなわち

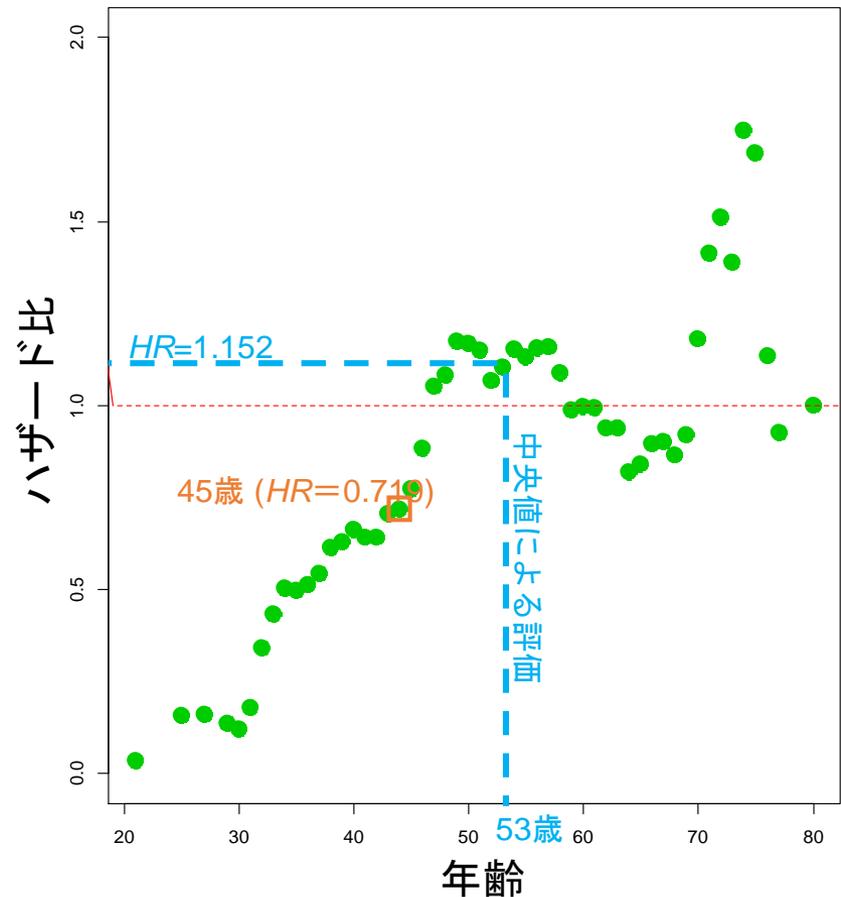
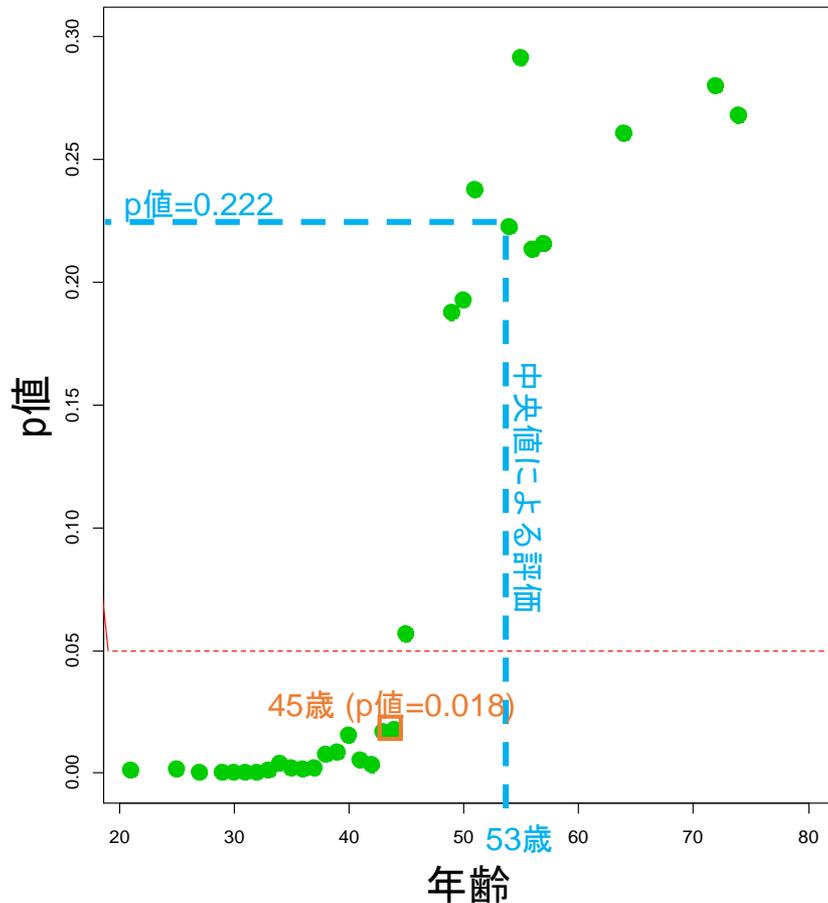
$$H_0 : (\beta_1, \beta_2)^T = (0, 0)^T, \quad H_1 : H_0 \text{ではない}$$

を検定すればよい(JMPでは、計算できる)。

この検定では、共変量全体では有意でないにも関わらず、個々のカテゴリで見ると有意になる場合がある(解釈が困難になる)。

# 共変量が量的尺度の場合の2値化の留意点

ドイツ乳がん研究グループによって実施されたランダム化比較試験のデータ(Schumacher et al., 1994)



変量の2値化において用いられることが多い中央値(53歳以上, 53歳未満)をカットオフ値にした場合, p値は0.222である(有意でなかった). 一方で, 45歳(45歳以上, 45歳未満)をカットオフ値にした場合, p値は0.018である(有意だった).

# 共変量が量的尺度の場合の2値化の留意点

## 45歳(<45/≥45)の場合

高年齢群(45歳以上)のほうが  
低年齢群(45歳未満)よりも死  
亡リスクが**低い** (HR=0.719)

解釈が逆



## 53歳(<53/≥53)の場合

高年齢群(53歳以上)のほうが  
低年齢群(53歳未満)よりも死  
亡リスクが**低い** (HR=1.152)

## カットオフ値の選定に関するRule of Thumb

- (1) 中央値を用いる
- (2) 臨床的にリーズナブルなカットオフ値を選定する

臨床的にリーズナブルとは、今回の事例の場合には、例えば疫学的調査などから得られた乳癌患者の平均年齢を用いたり、あるいは、閉経年齢を用いることを意味する。いずれにしても、カットオフ値選定におけるゴールド・スタンダードは存在しないため、試行的に決定する必要がある。