多変量解析 II.

～ 予測を目的とする解析
（予後予測/予後因子解析）

京都大学医学部附属病院

探索医療センター検証部

吉村健一

多変量解析を用いる研究

- 因果推論を行う研究　　etiologic research
  - 原因 cause を探求する研究
  - 前回「多変量解析I」の内容
- 予後因子解析研究　　prognostic research
  - 本日「多変量解析II」の内容

2

前回のスライド（再掲）：

同種の多変量解析法を用いていても予後因子解析、

予後予測（次回）とは全く異なる解析と考えたほうが良い？

- 因果推論では「曝露と結果の間の関連」に興味
  - この関連を適切に評価することが統計解析の目的
  - 曝露以外の変数は交絡の調整のために用いたものである

- 予後因子解析では「結果の予測」に興味
  - 予測性能をあげることが統計解析の目的
    - 極端に言えば・・・どの変数を用いようが予測が当たれば良い
  - *詳しくは次回*

3

本日の内容

- 予後因子解析研究の一例をみてみよう
- 予後因子解析
- 予後因子解析　vs.　因果推論　← 特に重要
- 予測モデルの開発
  - モデル構築
  - 妥当性の検証

4

例.　予後因子解析研究：Maione *et al.* JCO 2005

VOLUME 23 · NUMBER 28 · OCTOBER 1 2005

JOURNAL OF CLINICAL ONCOLOGY　ORIGINAL REPORT

Pretreatment Quality of Life and Functional Status Assessment Significantly Predict Survival of Elderly Patients With Advanced Non–Small-Cell Lung Cancer Receiving Chemotherapy: A Prognostic Analysis of the Multicenter Italian Lung Cancer in the Elderly Study

Paolo Maione, Francesco Perrone, Ciro Gallo, Luigi Manzione, FrancoVito Piantedosi, Santi Barbera, Silvio Cigolari, Francesco Rosetti, Elena Piazza, Sergio Federico Robbiati, Oscar Bertetto, Silvia Novello, Maria Rita Migliorino, Adolfo Favaretto, Mario Spatafora, Francesco Ferraù, Luciano Frontini, Alessandra Bearz, Lazzaro Repetto, and Cesare Gridelli

From the S Giuseppe Moscati Hospital, Avellino; National Cancer Institute; Medical Statistics, Second University of Napoli; Pneumology V, Monaldi Hospital, Napoli; S Carlo Hospital, Potenza; Mariano Santo Hospital, Cosenza; S Giovanni di Dio e Ruggi d'Aragona, Salerno; Civil Hospital, Noale, Venezia; Sacco Hospital; S Gerardo Hospital, Monza, Milano; Civil Hospital, Rovereto,

A B S T R A C T

Purpose

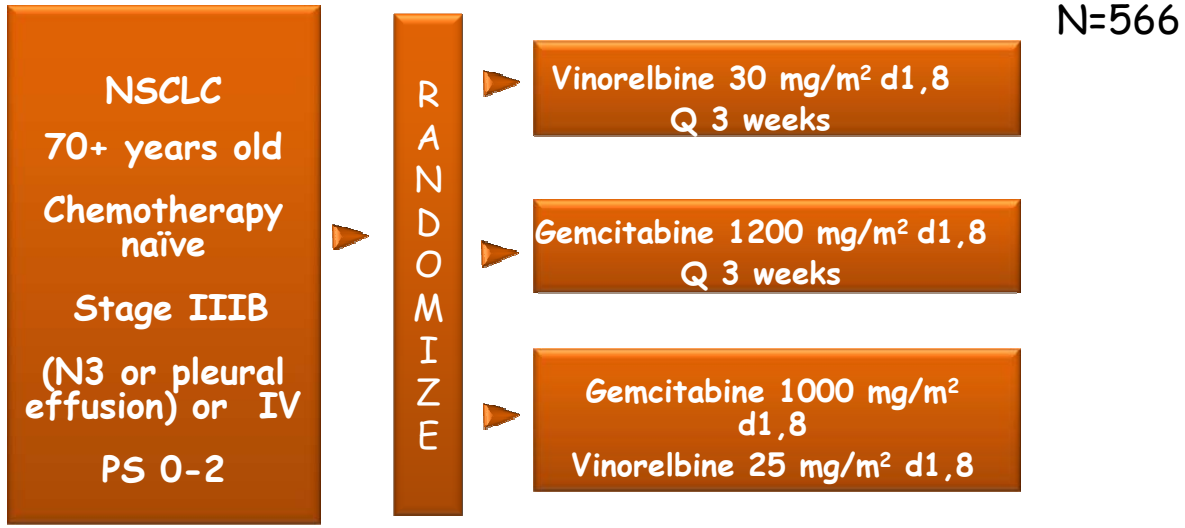
To study the prognostic value for overall survival of baseline assessment of functional status, comorbidity, and quality of life (QoL) in elderly patients with advanced non–small-cell lung cancer treated with chemotherapy.

J Clin Oncol 23:6865-6872.

5

例.　予後因子解析研究：Maione *et al.* JCO 2005

- 対象者：高齢者非小細胞肺癌を対象としたMILES第III相試験に登録



ASCO 2001 Abstr #1230; JNCI 2003, 95:362-372.

- Endpoint：生存期間
- 予後因子の候補（治療前）：
  - 併存疾患、QOL、Functional Status（機能状態）

J Clin Oncol 23:6865-6872. 6

例．Maione *et al.* JCO 2005，*cont.*

多変量解析 **Table 3.** Multivariate Analysis

Variable	HR	95% CL Upper	95% CL Lower	*P**
Sex				.07
Male (n = 465)	Ref			
Female (n = 101)	0.78	0.59	1.02	
Age, years				.69
< 75 (n = 337)	Ref			
75-79 (n = 210)	1.09	0.89	1.32	
≥ 80 (n = 19)	0.96	0.57	1.64	
Performance status				.006
0-1 (n = 460)	Ref			
2 (n = 106)	1.46	1.12	1.88	
Charlson score				.66
0 (n = 237)	Ref			
1 (n = 210)	1.06	0.85	1.32	
2 (n = 92)	1.12	0.85	1.48	
≥ 3 (n = 27)	0.84	0.52	1.36	
ADL				.44
No dependence (n = 482)	Ref			
One or more dependence (n = 84)	1.12	0.85	1.47	
IADL				.04
Better (n = 188)	Ref			
Intermediate (n = 217)	0.97	0.76	1.22	
Worse (n = 161)	1.31	1.00	1.71	
Quality of Life				.0003
Better (n = 119)	Ref			
Intermediate (n = 294)	1.62	1.24	2.10	
Worse (n = 153)	1.76	1.29	2.39	
Stage				.71

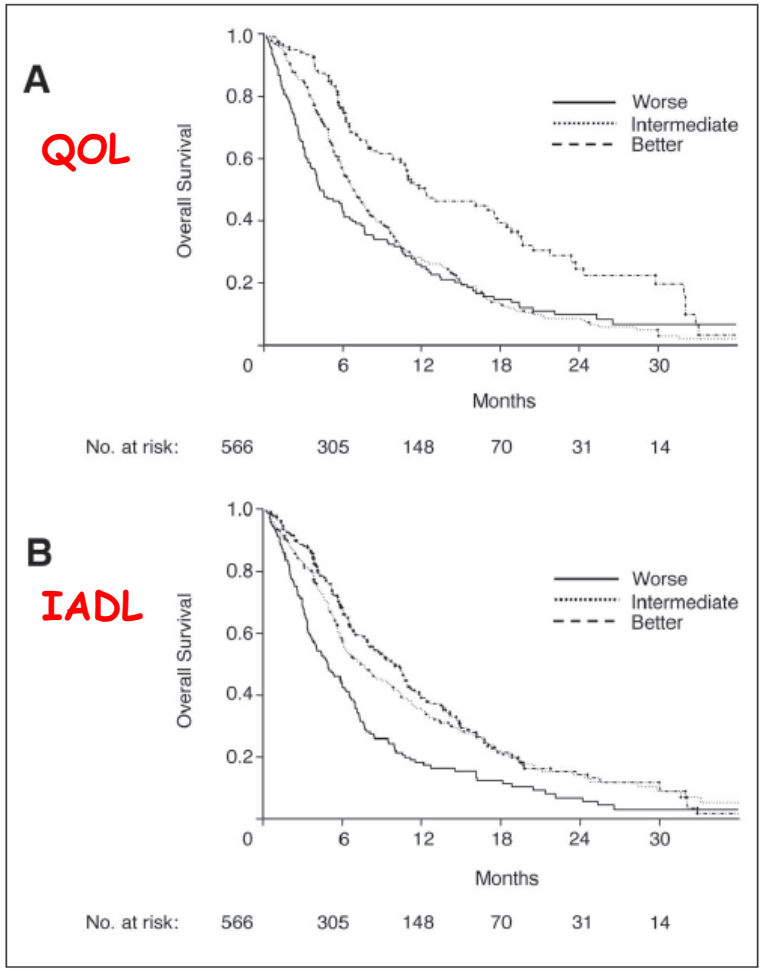
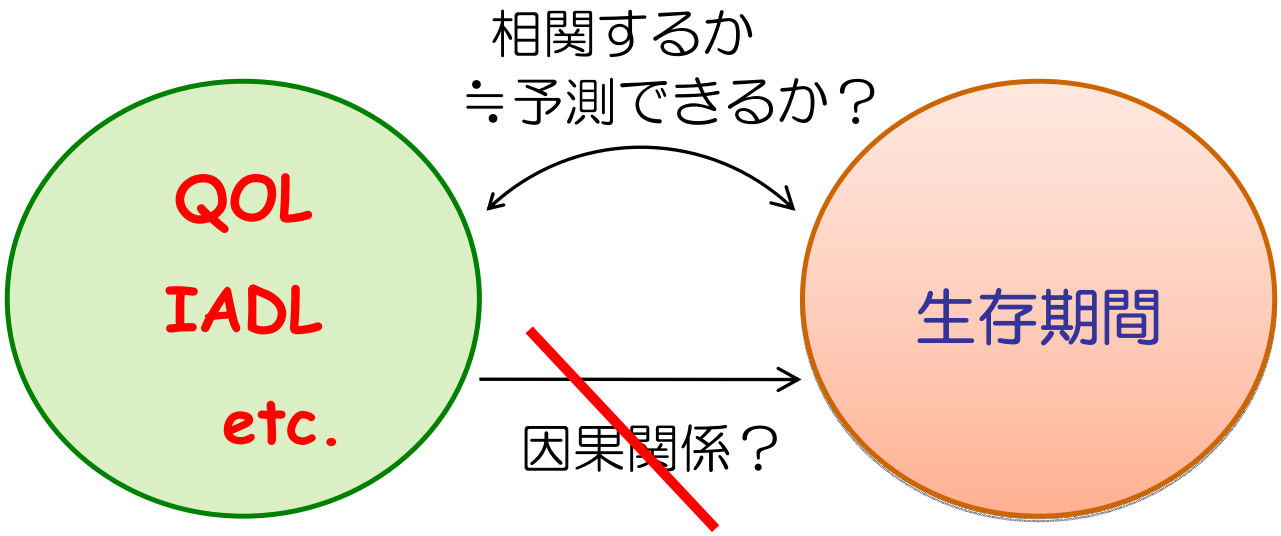


Fig 2. Kaplan-Meier–estimated overall survival curves according to pretreatment (A) quality of life (QoL) and (B) intermediate Activities of Daily Living (IADL) categories.

J Clin Oncol 23:6865-6872.

例. Maione et al. JCO 2005：予後因子解析 or 因果推論？



- 因果推論（原因を調べること）が目的ではない
  - 原因：介入するターゲット？、そうは考えていない
- 治療前のQOLやFunctional Statusが治療後の生存期間と相関するか
  ≒予後予測に有用かを評価することが目的

8

予後予測を目的とした解析

例．Maione *et al.* JCO 2005，*cont.*

多変量解析 **Table 3.** Multivariate Analysis

Variable	HR	95% CL Upper	95% CL Lower	*P**
Sex				.07
Male (n = 465)	Ref			
Female (n = 101)	0.78	0.59	1.02	
Age, years				.69
< 75 (n = 337)	Ref			
75-79 (n = 210)	1.09	0.89	1.32	
≥ 80 (n = 19)	0.96	0.57	1.64	
Performance status				.006
0-1 (n = 460)	Ref			
2 (n = 106)	1.46	1.12	1.88	
Charlson score				.66
0 (n = 237)	Ref			
1 (n = 210)	1.06	0.85	1.32	
2 (n = 92)	1.12	0.85	1.48	
≥ 3 (n = 27)	0.84	0.52	1.36	
ADL				.44
No dependence (n = 482)	Ref			
One or more dependence (n = 84)	1.12	0.85	1.47	
IADL				.04
Better (n = 188)	Ref			
Intermediate (n = 217)	0.97	0.76	1.22	
Worse (n = 161)	1.31	1.00	1.71	
Quality of Life				.0003
Better (n = 119)	Ref			
Intermediate (n = 294)	1.62	1.24	2.10	
Worse (n = 153)	1.76	1.29	2.39	
Stage				.71

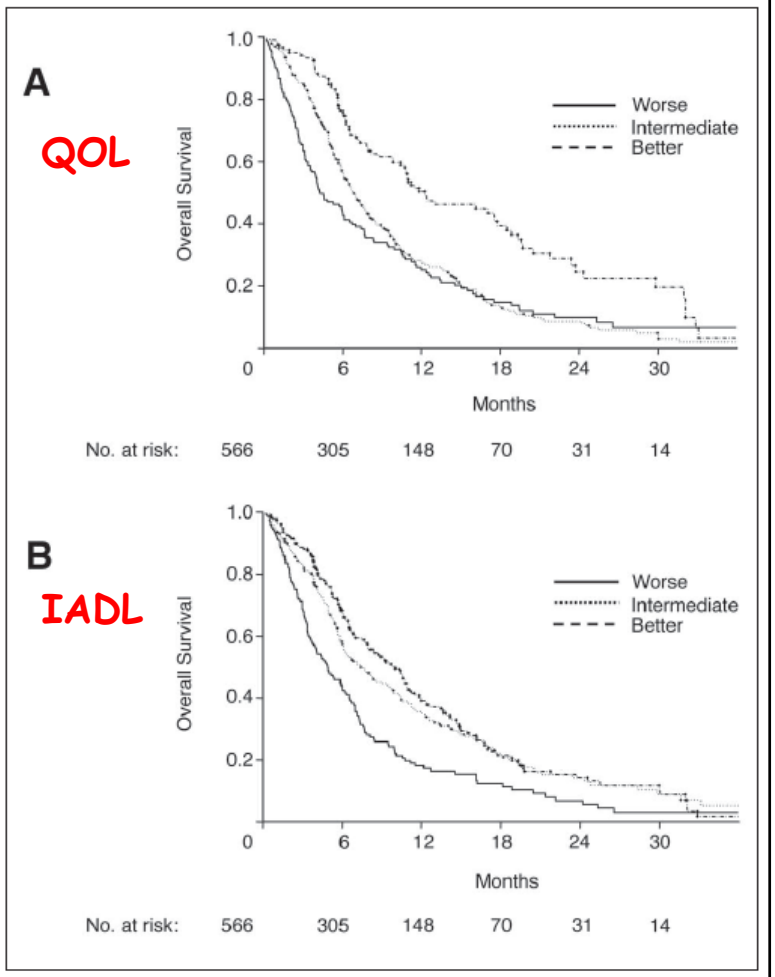


Fig 2. Kaplan-Meier–estimated overall survival curves according to pre-treatment (A) quality of life (QoL) and (B) intermediate Activities of Daily Living (IADL) categories.

J Clin Oncol 23:6865-6872.

予後予測を目的とした解析

例．Maione *et al.* JCO 2005，*cont.*

多変量解析 **Table 3.** Multivariate Analysis

Variable	HR	95% CL Upper	95% CL Lower	*P**
Sex				.07
Male (n = 465)	Ref			
Female (n = 101)	0.78	0.59	1.02	
Age, years				
< 75 (n				
75-79				
≥ 80				
Perform				
0-1 (n				
2 (n				
Charlso				
0 (n				
1 (n				
2 (n				
≥ 3 (				
ADL				
No de				
One or				
IADL				
Better (n = 188)	Ref			
Intermediate (n = 217)	0.97	0.76	1.22	
Worse (n = 161)	1.31	1.00	1.71	
Quality of Life				.0003
Better (n = 119)	Ref			
Intermediate (n = 294)	1.62	1.24	2.10	
Worse (n = 153)	1.76	1.29	2.39	
Stage				.71

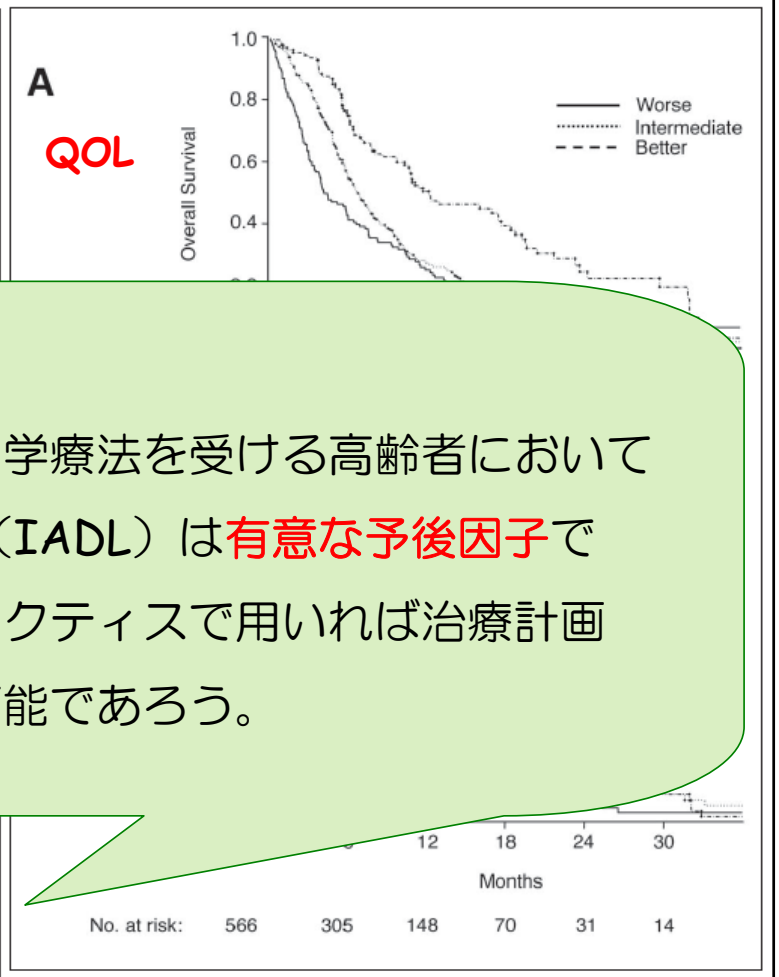


Fig 2. Kaplan-Meier–estimated overall survival curves according to pretreatment (A) quality of life (QoL) and (B) intermediate Activities of Daily Living (IADL) categories.

J Clin Oncol 23:6865-6872.

Maioneらの結論

進行非小細胞肺癌に対して化学療法を受ける高齢者において治療前のQOLや機能スコア（IADL）は有意な予後因子であった。これらスコアをプラクティスで用いれば治療計画する際の予後予測の改善が可能であろう。

※ 予後因子解析としてよく行われる典型例

- 後向きにデータを収集する（後向き研究）
- 変数ごとに予後との関連をみる（単変量解析）
- 単変量解析で有意な(複数の)変数と予後の関連を同時に評価
  （"多変量解析"）
- 連続変数はp値が最小になるカットオフ値を用いて二値化（最小p値法）
- 多変量解析で計算されたP値に基づいて変数選択
- 最終モデルに基づくリスク分類の作成
  - 例.　low- / intermediate- / high-risk group
- リスク分類ごとの生存曲線の差の比較（モデルの性能評価）

これで良いのか？

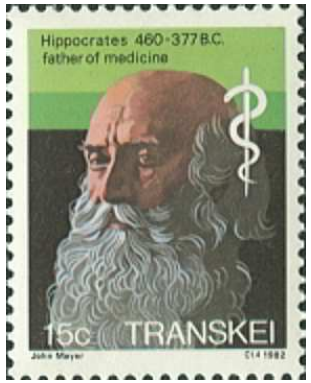
11

本日の内容

- 予後因子解析研究の一例をみてみよう
- 予後因子解析
- 予後因子解析　vs.　因果推論
- 予測モデルの開発
  - モデル構築
  - 妥当性の検証

12

予後因子解析　　prognostic analysis



- 医学の父 Hippocrates (460BC - 377BC)
  - 患者の経過情報に基づく正確な"予後"の予測を重視
- **予後　prognosis**
  - 「死亡」、増悪、症状の軽減など特定の結果を生じるリスク
  - 予後因子（年齢、性別、経過、症状、検査結果など）に基づいて予測されうるもの
    - 予後因子解析

Hippocrates "On airs, waters and places"

予後因子解析の主たる目的

1. 将来のリスク情報を医師と患者にもたらし
   この情報に基づいて個々の介入/治療を決定する（実臨床）
   - 新生児におけるApgar score
   - Framingham Cardiovascular Risk Score と抗コレステロール剤/降圧剤
   - 乳がんにおけるNottingham Prognostic Index と術後補助化学療法

2. 臨床試験の対象集団を決定する（臨床試験）
   - 罹患リスクの高い集団を対象として予防試験
   - 再発リスクの低い集団を対象として縮小手術を評価する試験
     - 介入が必要な or 予後が均一となる集団の特定

3. 施設の機能評価を行う（施設評価）、現状ではminor
   - 新生児集中治療室の評価：Clinical Risk Index for Babies

14

予後予測と多変量解析

- 背景要因/治療効果における大きな個体差
- 疾患メカニズムの複雑さ、多様性
- 1つの予後因子に基づく予測では不十分な状況が多い
  - 複数の予後因子の組み合わせによる予後予測
    - 多変数アプローチ　multivariable approach　が自然
      - いわゆる“多変量解析”
      - 予後/リスク指標（スコア）、予後/予測モデル
        prognostic/risk index (score), prognostic/prediction model
      - 予測ルール
        prediction rule

15

例.　遺伝子発現解析でも予後因子解析

- Class comparison
  - comparison of gene expression in different groups of specimens
- **Class prediction**
  - derivation of predictors of prognosis, response to therapy, or any phenotype or genotype defined independently of the gene expression profile
    - to predict accurately the class membership of a new sample
- Class discovery
  - defining previously unrecognized tumor subtypes
    - no classes are predefined

Science 1999, 286:531-7; JNCI 2003, 95:14-18;　16

例.　遺伝子発現解析でも予後因子解析

- Class comparison
  - comparison of gene expression in different groups of specimens
- **Class prediction**
  - derivation of predictors of prognosis, therapy

遺伝子発現プロファイルとは独立に定義されたクラス（予後、治療への奏効、表現型など）を予測できる因子を得ること。
新しいサンプルが属するクラスの正確な予測が目的。

- Class discovery
  - defining previously unrecognized tumor subtypes
    - no classes are predefined

予後因子解析

Science 1999, 286:531-7; JNCI 2003, 95:14-18;

本日の内容

- 予後因子解析研究の一例をみてみよう
- 予後因子解析
- 予後因子解析　vs. 因果推論
- 予測モデルの開発
  - モデル構築
  - 妥当性の検証

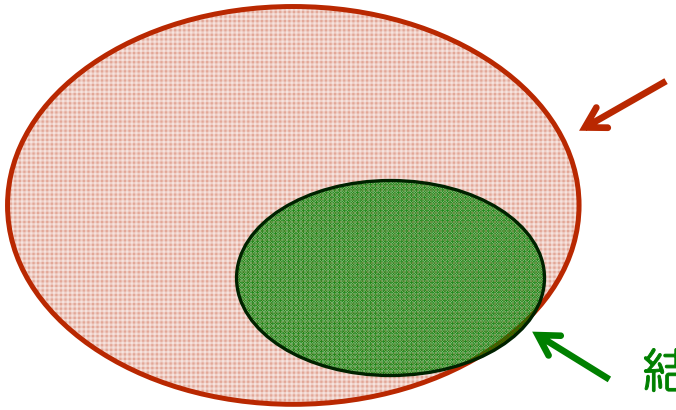
18

予後因子解析 vs. 因果推論（etiologic research）

- 何れも "多変量解析" を頻用、しかしその位置づけが異なる
- 予後因子解析の目的は「結果の予測」
  - 「死亡」するリスクを可能な限り正確に予測する方法として多変量解析を用いる
  - 予測性能をより向上させるために複数の変数を用いる
- 因果推論の目的は「曝露と結果の間の関連」の評価
  - 他の変数を調整した下でも「死亡」が曝露に起因しているかを評価する方法として多変量解析を用いる
  - 交絡を調整するために（曝露以外にも）複数の変数を用いる

19

予後因子解析 vs. 因果推論，cont.



予測のために有用な変数

・・・結果と相関を有するもの

結果と因果関係を有する曝露（十分条件）

・・・理論的に因果連鎖に位置づけられるもの

- 予測のために有用であれば、因果関係は必要条件でない
  - 例.　仮死とApgar scoreの皮膚色　※ 血中酸素濃度が「原因」
- 予後予測によって因果関係やbiologyに関する示唆も得られうる
- 主たる目的ではない、必要条件でもない
  - 反対に、予後因子であることは「原因」であることの必要条件

20

目的に対応して指標も異なる：因果推論

相対リスク（ハザード比）

- 例．Cox回帰
  - ハザード ＝ ベースライン×曝露の効果
    ×交絡因子の効果×・・・×・・・
- 因果推論における指標は相対リスクが主
- 曝露なしに比べた曝露ありの相対リスクに興味がある
  - 交絡因子をモデルに含めるのは交絡調整のため

21

目的に対応して指標も異なる：予後因子解析

相対リスク（ハザード比）

- 例．Cox回帰
  - ハザード ＝ ベースライン× 予後因子1の効果 ×予後因子2の効果 ×・・・×・・・

絶対リスクの予測式

- 因果推論における指標は 相対リスク が主
  - 曝露なしに比べた曝露ありの 相対リスク に興味
    - 交絡因子をモデルに含めるのは交絡調整のため
- 予後予測における指標は 絶対リスク が主
  - 個々の患者の予後（絶対リスク）予測の正確さに興味
    - 予後因子ごとの 相対リスク は従
      - 絶対リスク を導くため「予測式」の一部

22

予測の正確さを優先≒あたれば良い、で良いのか？

- そもそもの研究目的に対応して考えると「あたれば良い」で良い
  - 因果推論が主ではない
  - 因果関係は問わない、予測が正確であればOK
- 勿論、生物学的にも正しい/因果関係があるに越したことはない
  - 以下を全て満たすならば、むしろ積極的にモデルに取り込むべき
    - 生物学的に明らかな因果関係
    - 測定誤差が小さい、高精度での測定が可能
    - 臨床応用する際にも容易に測定できる

      →データ依存が減れば Overfitting（後述）の程度も減少
  - 医学的尤もらしさ≒face validityもバランスを考慮すべき？

23

本日の内容

- 予後因子解析研究の一例をみてみよう
- 予後因子解析
- 予後因子解析vs. 因果推論
- 予測モデルの開発
  - モデル構築
  - 妥当性の検証

24

構築：目的と対象者

- 目的
  - 「興味のある集団」において結果が生じるリスクを
    複数の予後因子の組み合わせにより予測すること
- 対象者
  - 「興味のある集団」にあてはまるように適格規準を設定
    - 興味のある結果が生じるリスクを有する
  - モデル構築のためのtraining (data-)set

25

構築：モデル構築の方法

- 研究デザインの決定
- 予後因子の候補の選択
- 変数の取り扱い方の決定
- 予測モデルの構築
  - データに基づくモデル選択
- （リスク分類の作成が目的の場合）モデルに基づく分類の作成

26

構築：研究デザイン

- 観察研究ならば、前向きコホート研究が最適
  - 臨床試験に比べると治療が不均一
- 後向きコホート研究：別の目的で集積されたデータを事後に収集
  - 現状、よく用いられる
  - 利点：追跡期間が長い　　限界：データの質が劣る
- ランダム化比較試験
  - 前向きコホート研究と同様
    - 群間差あり → 群も予後因子にする
    - 　　なし → 群をプール
  - 限界：厳しい適格規準、一般化可能性が劣る
- ケース・コントロール研究
  - 限界：絶対リスクの推定が困難であるため注意

27

構築：予後因子の候補の選択

- 先行研究で既に報告されたものは優先的に
  - がん患者の予後予測におけるリンパ節転移、遠隔転移
  - 冠動脈疾患発症リスクに対する年齢、TC、HDL、血圧、喫煙
- 予測モデルを実際に用いる場面でも測定可能なもの
  - 予測モデルを用いる場面に依存、侵襲性、測定系の問題
- 手元のデータではなく、臨床的観点を優先した方が好ましい
  - 「単変量解析で有意でない」といって安易に除かない
    - 「予後因子として有用でない」とは限らない

28

構築：予後因子の候補の選択

- 他の予後因子との**相関が高いもの**は除く候補
  - 相関が高い ≒ 他の予後因子でも十分に説明できる
  - 追加的な情報をほとんどもたない
- **測定誤差が大きい変数**は除く候補
  - 重要な因子であろうとも測定誤差大であると予測性能が低下
- **欠測が多い変数**は除く候補
  - 欠測値のある対象者を除くとバイアスがかかる
  - 補完法 imputation もあるが、検証不能な強い仮定が必要

29

構築：変数の取り扱い方

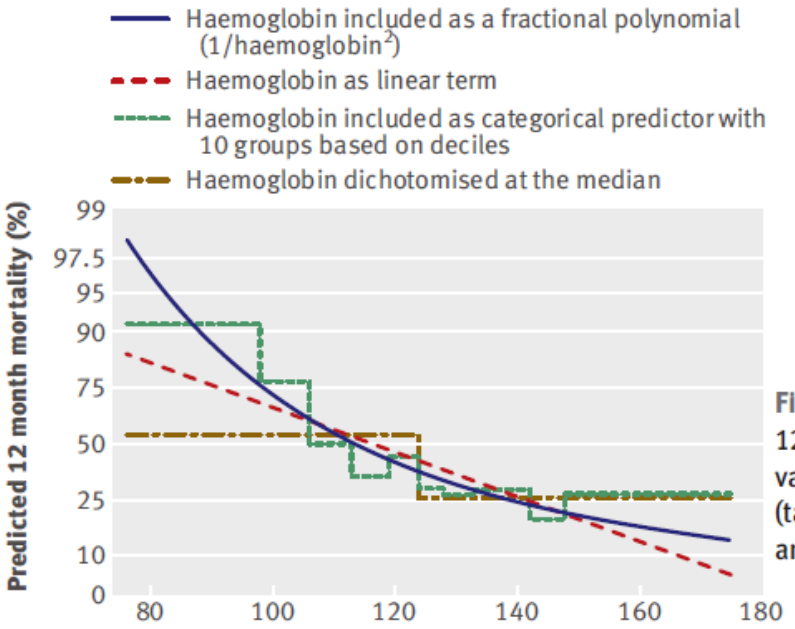
- 合成変数の作成
  - 拡張期血圧と収縮期血圧：何れか一方 or「平均」？
- 順序変数
  - e.g. 病期：水準の併合 or 適切なコーディングは必要か？
- 連続変数　e.g. 年齢
  - 先行知見・臨床的根拠に基づく：二値化 or カテゴリ化？
    - Data-driven or Not data-driven？
  - そのまま連続量、線形？
  - 非線形（多項モデル etc.）？

30

構築：連続変数の取り扱い方

- 主目的は予測・・・二値化や線形よりも非線形がベター？

  例．Hbによる腎がん患者の予後予測



多項

線形（直線）

10等分

2等分（二値化）

Fig 3 | Estimates of the association between haemoglobin and 12 month mortality in the kidney cancer data, adjusted for the variables in the Akaike information criterion derived model (table 1). The vertical scale is linear in the log odds of mortality and is therefore non-linear in relation to mortality

1年死亡率

Hb

BMJ 2009, 338:1373-7.

31

構築：“医学的尤もらしさ face validity”と予測性能のバランス

- 極端な非線形
  - たとえフィットが良いとしても生物学的な説明はしばしば困難

32

構築：連続変数の二値化が必要ならば

- データに基づかずに決定するのが最善
- Median で分割：根拠に乏しい、再現性にも乏しい
- 最小p値法：データに基づいてp値最小→カットオフ値
  - 例. "Optimal" cutpoints in S-phase fraction

Table 2. Optimal cutpoints derived by the minimum *P*-value approach in the Freiburg DNA breast cancer study*

	Population 1, all (n = 207)	Population 2, diploid (n = 119)	Population 3, node negative (n = 98)	Population 4, node positive (n = 109)	Population 5, node positive, diploid (n = 59)	Population 6, node positive, aneuploid (n = 50)
Optimal cutpoint	5.4	5.4	9.0-9.1	10.7-10.9	3.7	10.7-11.2
P value	.037	.051	.084	.007	.068	.003
Relative risk using optimal cutpoint	1.58	1.87	0.28	2.37	1.94	3.30
95% confidence interval	1.03, 2.44	1.00, 3.49	0.07, 1.19	1.27, 4.44	0.95, 3.96	1.49, 7.29
Corrected *P* value	.403	>.5	>.5	.123	>.5	.063
P value from Cox model	.340	.276	>.5	.061	>.5	.031

*Estimated relative risks with 95% confidence intervals (upper part); corrected *P* values (for explanation *see* text) and *P* values from a Cox model including SPF as a continuous covariate (lower part). n = number of patients.

- 検定の多重性も可能、調整してみると　p値 ＞＞ 有意水準

Journal of the National Cancer Institute, Vol. 86, No. 11, June 1, 1994 33

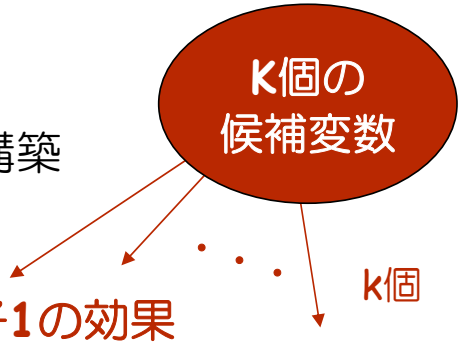
※ 予後因子解析としてよく行われる典型例

- 後向きにデータを収集する（後向き研究）
  - 前向き研究がbetter
- 変数ごとに予後との関連をみる（単変量解析の結果を優先）
  - 単変量解析の結果に基づくことは必ずしも適切でない
- ~~単変量解析で有意な~~(複数の)変数と予後の関連を"多変量解析"
- 連続変数はp値が最小になるカットオフ値を用いて二値化（最小p値法）
  - 安易に二値化しない、最小p値法は危険
- 多変量解析で計算されたP値に基づいて変数選択
- 最終モデルに基づくリスク分類の作成
- リスク分類ごとの生存曲線の差の比較（モデルの性能評価）

34

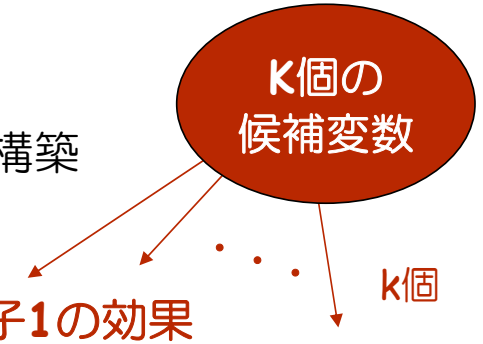
構築：予測モデルの構築

- 候補K個のうち、k個の変数により最終モデル構築
  - 例．Cox回帰
    - ハザード ＝ ベースライン × 予後因子1の効果
      × ・・・ × 予後因子kの効果



35

構築：予測モデルの構築



- 候補K個のうち、k個の変数により最終モデル構築
  - 例．Cox回帰
    - ハザード ＝ ベースライン × 予後因子1の効果
      × ・・・ × 予後因子kの効果

- 候補を全て含めたモデル
  - 利点：データに基づく変数選択によるバイアス（後述）がない
  - 限界：精度が劣る
  - 変数＜＜対象者数 or イベント数でない限り推定が不安定
  - 変数選択しなくても良い状況に限定

→　変数選択の必要性

36

構築：モデル選択・変数選択の必要性

- 変数の多いモデル：データに対するあてはまりが良い
- 変数の少ないモデル：臨床応用を考えると好ましい
  - parsimony：良いモデルの特徴の1つでもある
- 現実的には、候補数が多すぎる状況が多い
  - 多くのシミュレーション研究により、サンプル数/変数≧10
    - 生存時間解析（e.g. Cox回帰）の場合、イベント数/変数≧10
  - 重要な変数/適切なモデルを選択する必要性
  - モデル選択する際にはデータ依存が避けられない
    - data-dependent, data-driven

37

構築：データに基づくモデル選択

- 決め手はない！
- 現状、最もよく用いられるのは変数減少法/変数増加法
  - 例. 候補が10個
    - 変数減少法：10→9→8→・・・　有意でないものを逐次除いていく
      - 10個含めた解析でp値が最大の変数を除く、続いて残り9個含めた…
    - 変数増加法：1→2→3→・・・　有意なもの逐次含めていく
  - 何れかといえば、変数減少法がベター
  - 有意水準「大」→変数「多」、有意水準「小」→変数「少」
  - 利点：単に便利（殆どそれのみ）
    - 先行知見など外的情報を必要としない"自動変数選択法"
      - 先行知見は積極的に用いるべき

38

構築：データに基づく変数選択，cont.

- 総当たり法：全ての組み合わせを評価する
  - K個の候補があるならば 2^K 個のモデルを比較
  - 比較する指標：Akaike Information Criterion (AIC) など
    - AIC: 変数が増えることにペナルティを課したあてはまり指標
    - 有意水準15.7%の変数減少法≒AICを用いた総当たり法
- Regression tree，neural network，…
  - 一般の"多変量解析"とは異なる（複雑な）アルゴリズムを用いて、データにあてはまるモデルを探索する方法
  - 良いモデルを構築可能だからというよりも目新しさが先行
    - 事実として、あまり有用でないという報告も多くある
- 共通する限界：データに基づく　→　Overfittingの問題（後述）

39

構築：予測モデルに基づくリスク分類

- リスク分類の作成が目的であるならば、更に分類ルールを作成
  - 最終モデルから得られる絶対リスクの予測値に基づいて分類
  - 例. 2分類
    - 絶対リスクの予測値＞閾値　→ 高リスク群
    - それ以外　→ 低リスク群
  - 閾値の決め方
    - Data-driven：誤分類率（感度や特異度）に関して最適化
    - 臨床的観点：リスク分類を実際に用いる状況を想定
      - e.g. 絶対リスクが10%以上となる集団を特定したい

40

構築：Over-fittingの問題

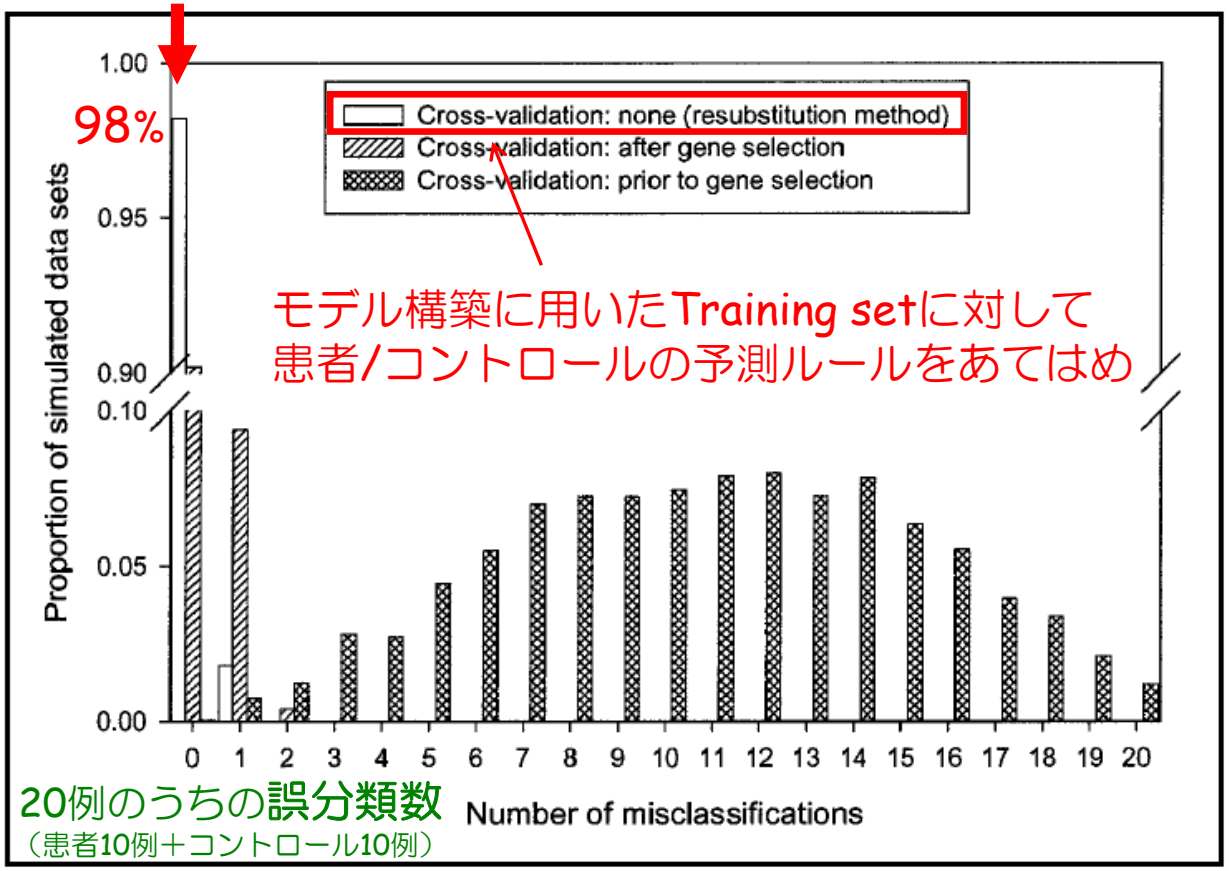
- Training set に過剰に適応してしまうこと
  - 真の構造でなく、誤差的ばらつきにフィットしてしまう
    - 内弁慶、training setを除いては予測性能が劣る
    - 再現性 reproducibility がない結果の主たる原因
- 例.　癌患者/コントロールを予測するモデルを作成する状況
  （Simonらによるシミュレーション研究）
  - Training set：癌患者10例、コントロール10例の2クラス
    - 600遺伝子・・・出鱈目に発生させた遺伝子データ
      - 真には「遺伝子」と「患者/コントロール」間に関連なし
      - 予測するという目的に対して全くもって有用でない
        - これを用いて予測モデルを作っても意味がないはず

41

Simonらのシミュレーション

- 出鱈目の遺伝子なのに98%の確率で
  患者かコントロールかを完全に予測可能　→Over-fittingという現象

縦軸
頻度
（割合）



Journal of the National Cancer Institute, Vol. 95, No. 1, January 1, 2003

42

※ 予後因子解析としてよく行われる典型例

- 後向きにデータを収集する（後向き研究）
  - 前向き研究がbetter
- 変数ごとに予後との関連をみる（単変量解析の結果を優先）
  - 単変量解析の結果に基づくことは必ずしも適切でない
- ~~単変量解析で有意な~~(複数の)変数と予後の関連を"多変量解析"
- 連続変数はp値が最小になるカットオフ値を用いて二値化（最小p値法）
  - 安易に二値化しない、最小p値法は危険
- 多変量解析で計算されたP値に基づいて変数選択
- 最終モデルに基づくリスク分類の作成
  - モデル選択法として決め手はないが、妥当性の検証は必須
- リスク分類ごとの生存曲線の差の比較（モデルの性能評価）

43

本日の内容

- 予後因子解析研究の一例をみてみよう
- 予後因子解析
- 予後因子解析　vs.　因果推論
- 予測モデルの開発
  - モデル構築
  - 妥当性の検証
    - 妥当性の検証のデザイン
    - Internal validationの方法
    - 妥当性の検証をおこなう対象
    - 予測性能の指標

44

妥当性の検証　Validaton

- Feinstein 曰く
  - "Validation is one of those words — like health, normal, probability, and disease — that is constantly used and seldom defined. We can ... simply say that, in data analysis, validation consists of efforts made to confirm the accuracy, precision, or effectiveness of the results."

Feinstein AR (1996). "Multivariable Analysis: An Introduction " 45

妥当性の検証　Validaton

- Feinstein 曰く
  - "Validation"は、健康、正常、確率、疾患といった単語と同様に絶えず使用されるもののめったに定義が為されない単語である。データ分析においては、結果の正確度、精度、効果を確認するために行われる活動によって構成される。

- 「予測モデル」に対する妥当性の検証
  - 予測モデルがモデル構築に用いたものとは"異なる" データに対しても十分な予測性能をもつか否か

Feinstein AR (1996). "Multivariable Analysis: An Introduction " 46

「妥当性の検証」が何故必要か？

- 構築法の欠点
  - 優れた構築法/モデルとして標準的なものがない（恣意性）
  - モデル/変数の選択がデータに基づくことによる over-fitting
- 研究デザインの欠点
  - 適格規準や除外規準が不明瞭なことが多い（偏った集団）
  - 欠測値を理由とした対象者を除外することが多い（バイアス）
  - 不十分なサンプルサイズ（不十分な精度）
- 一般化可能性の本質：時間・場所などを変えないと検証できない
  - "case-mix"のばらつきにより異なる場所での予測性能は容易に低下
- 重要な予後因子に漏れがあるほど

47

妥当性の検証のデザイン： internal or external

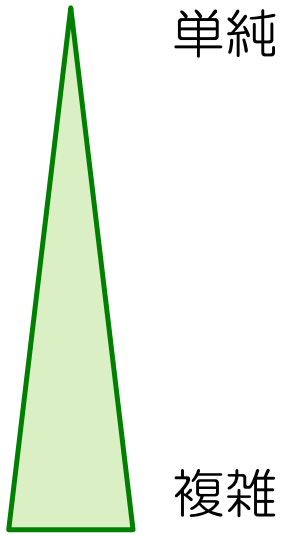
- **Internal validation**
  - 一つのデータを構築にも妥当性の検証にも用いる
  - 用いた「モデル構築法」の（内的）妥当性の検証が主

- **External validation**
  - 独立の"異なる"データを validation set として用意する
    - 時間 and/or 場所が異なる
  - 一般化可能性の評価が主
    - 異なる対象集団：大人→子供、欧米→日本

48

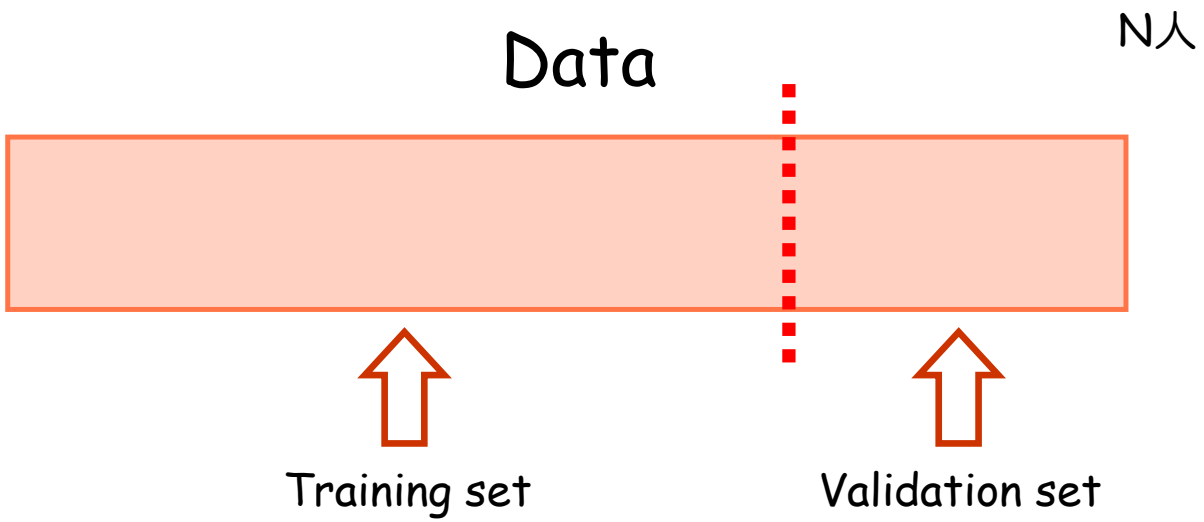
妥当性の検証： internal validationの方法

- Split sample 法
- Cross-validation (CV) 法
  - Split-half CV 法
  - Leave-one-out 法（LOO-CV法）
  - K-fold CV 法



49

典型例：Split-sample 法



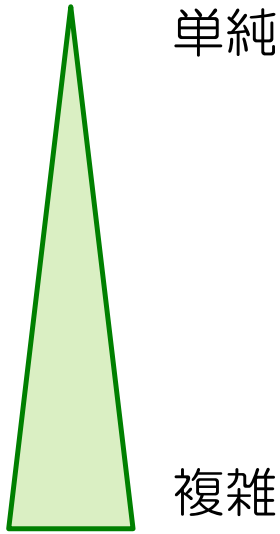
- 予測性能評価：モデル構築に用いたtraining set以外のデータを用いる
  - Random or Non-random
- モデル構築には training set のみを利用
  - training set と test set は役割入れ替えなし、no “crossing”
  - 限界：十分な大きさのNが必要

50

妥当性の検証： internal validationの方法

- Split sample 法
- Cross-validation (CV) 法
  - Split-half CV 法
  - Leave-one-out 法（LOO-CV法）
  - K-fold CV 法

効率的な方法

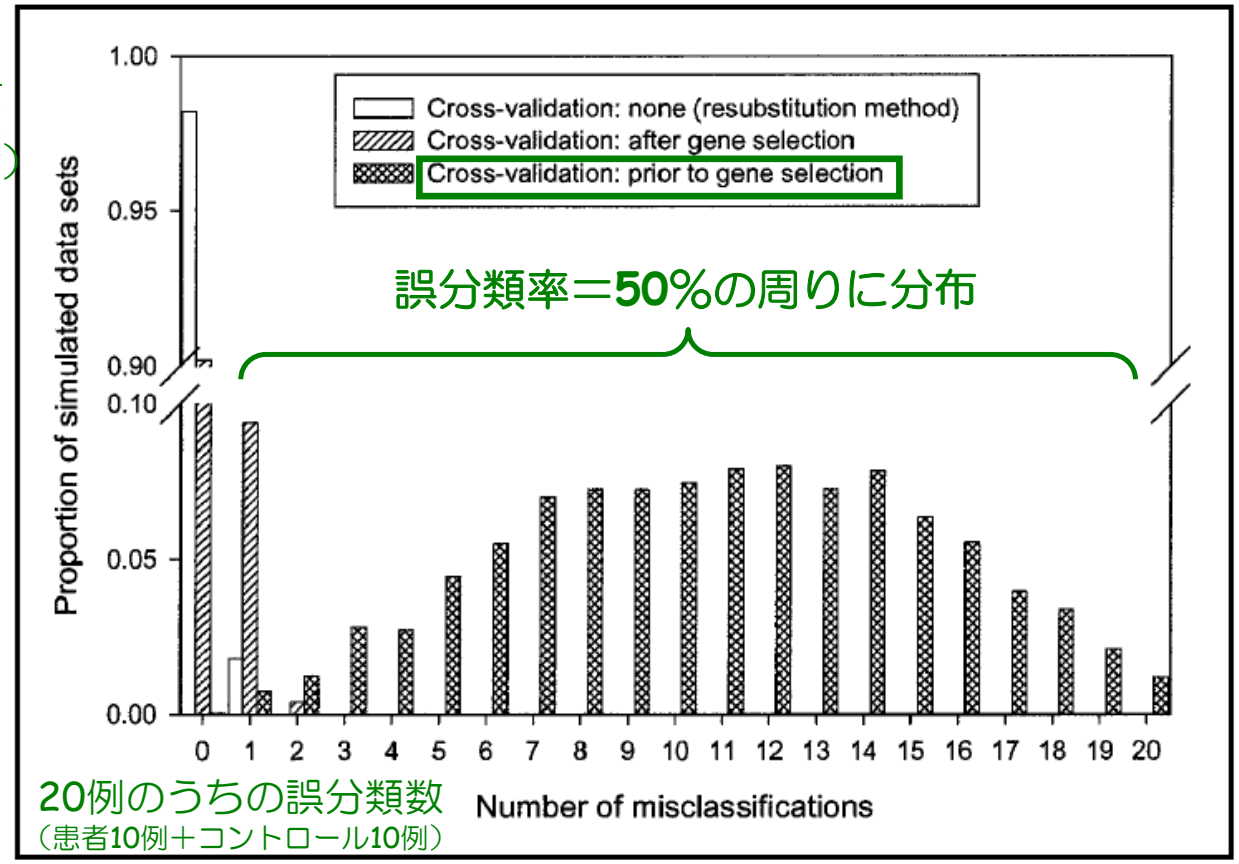


51

再掲.　Simonらのシミュレーション

- 適切な方法を用いて実施した妥当性の検証結果

縦軸
頻度
（割合）



Journal of the National Cancer Institute, Vol. 95, No. 1, January 1, 2003

52

妥当性の検証：予測性能の指標

- リスク分類が可能か？　　識別能　discrimination
- 予測値が観測値に近いか？　　較正/校正　calibration

53

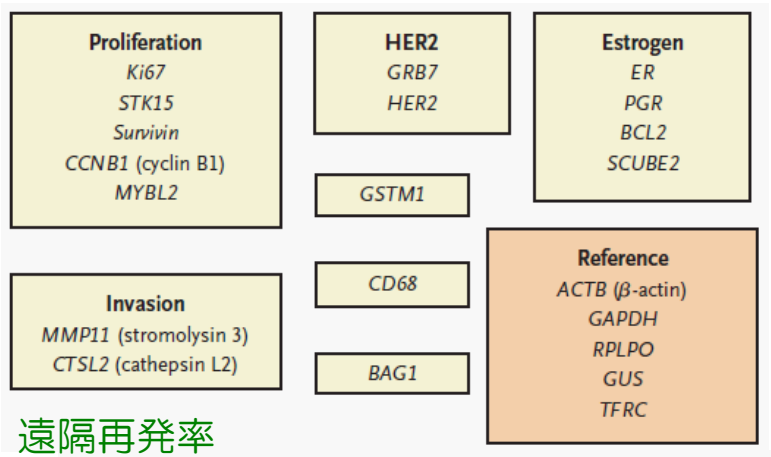
妥当性の検証：予測性能の指標

- リスク分類が可能か？　　識別能　discrimination
  - 例．　リスク分類間の再発率の差

- 予測値が観測値に近いか？　　較正/校正　calibration

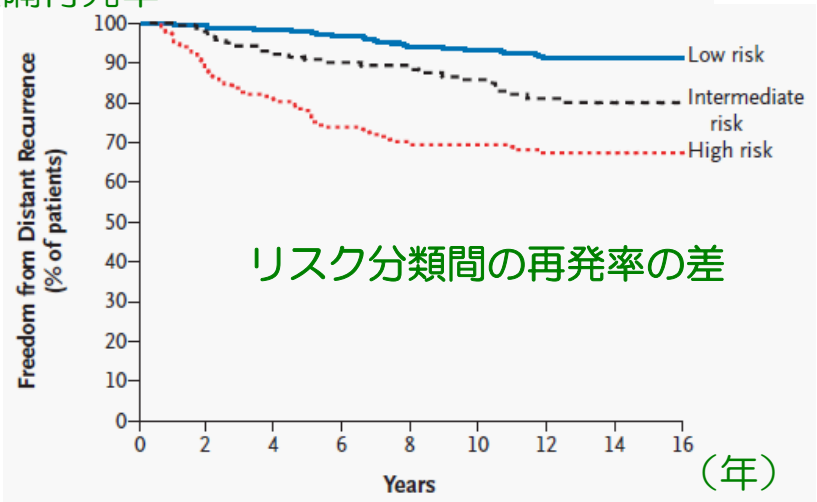
54

Molecular Classification (Class Prediction) -Oncotype DX Assay



再発リスクスコア（加重合計点）＝
+ 0.47 × HER2 Score
- 0.34 × ER Score
+ 1.04 × Proliferation Score
+ 0.10 × Invasion Score
+ 0.05 × CD68
- 0.08 × GSTM1
- 0.07 × BAG1

遠隔再発率



リスク分類間の再発率の差

（年）

- The recurrence score has been **validated** as quantifying the likelihood of distant recurrence in tamoxifen-treated patients

NEJM 2004, 351:2817-2826.

妥当性の検証：予測性能の指標

- リスク分類が可能か？　　識別能　discrimination
  - 例．　リスク分類間の再発率の差
  - 例．　誤分類とROC曲線（附録）
- 予測値が観測値に近いか？　　較正/校正　calibration
  - 例．　絶対リスクについて観察値×予測値のプロット

56

例．Calibration

ELSEVIER

European Journal of Cardio-thoracic Surgery 29 (2006) 441–446

EUROPEAN JOURNAL OF CARDIO-THORACIC SURGERY

www.elsevier.com/locate/ejcts

EuroSCORE（心臓手術の手術死亡率の予測モデル）の妥当性研究

Validation of the EuroSCORE model in Australia☆

Cheng-Hon Yap[a,1,*], Christopher Reid[b,1], Michael Yii[a,1], Michael A. Rowland[c,1], Morteza Mohajeri[d,1], Peter D. Skillington[e,1], Siven Seevanayagam[f,1], Julian A. Smith[g,1]

Table 4
Predicted and observed mortality by EuroSCORE risk level for isolated CABG cohort

観測　　予測

	Patients (deaths)	Observed mortality (95% CI)		Predicted mortality (95% CI)
EuroSCORE additive				
0–2 (low risk)	1955 (8)	0.41% (0.18–0.80)	poor	1.03% (0.99–1.06)
3–5 (medium risk)	1996(17)	0.85% (0.50–1.36)	⟷	3.90% (3.87–3.94)
6+ (high risk)	1641 (87)	5.30% (4.27–6.50%)		8.52% (8.39–8.65)
Total	5592 (112)	2.00% (1.65–2.40)		4.25% (4.16–4.34)
EuroSCORE logistic				
Low risk	1864 (8)	0.43% (0.19–0.84)	poor	1.24% (1.22–1.25)
Medium risk	1864 (14)	0.75% (0.41–1.26)	⟷	2.91% (2.88–2.94)
High risk	1864 (90)	4.83% (3.90–5.90)		14.43% (13.79–15.07)
Total	5592 (112)	2.00% (1.65–2.40)		6.19% (5.93–6.46)

Eur J Cardio-thorac Surg 2006, 29:441–446.

57

例．Calibration

Anaesthesia, 2005, **60**, pages 323–331

Assessing the applicability of scoring systems for predicting postoperative nausea and vomiting

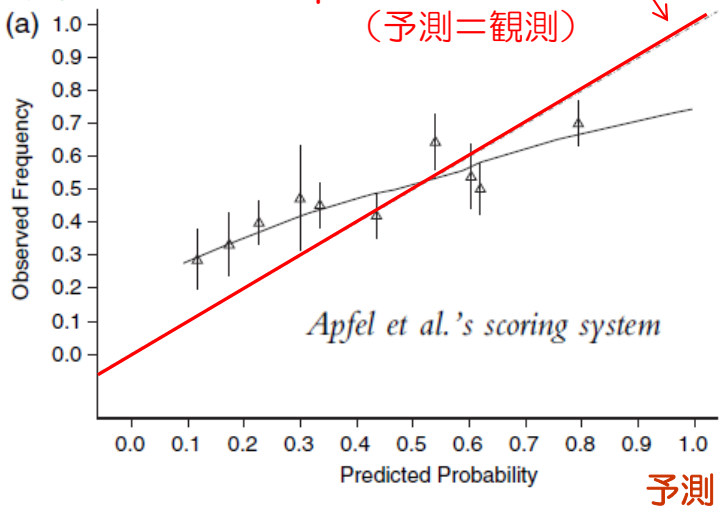
術後悪心・嘔吐を予測するスコアリングシステムに関する妥当性研究

J. E. van den Bosch,[1] C. J. Kalkman,[2] Y. Vergouwe,[3] W. A. Van Klei,[4] G. J. Bonsel,[5] D. E. Grobbee[6] and K. G. M. Moons[7]

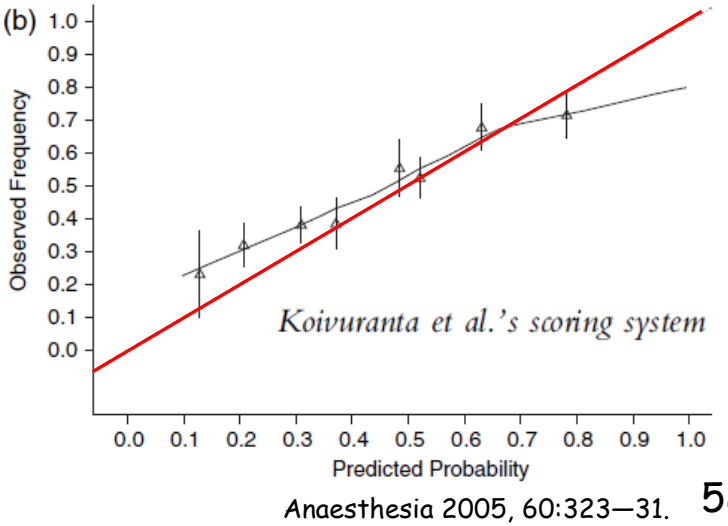
1 PhD Student, 2 Professor, 4 Resident in Anaesthesiology, and 7 Associate Professor, Department of Perioperative Care and Emergency Medicine, University Medical Centre Utrecht, the Netherlands
1 PhD Student, 3 Assistant Professor, 6 Professor and 7 Associate Professor, Julius Centre for Health Sciences and Primary Care, University Medical Centre Utrecht, Utrecht, the Netherlands
5 Professor, Department of Public Health, University Medical Centre Amsterdam, Amsterdam, the Netherlands

観察

perfect calibration
（予測＝観測）



予測



Anaesthesia 2005, 60:323—31.

58

※ 予後因子解析としてよく行われる典型例

- 後向きにデータを収集する（後向き研究）
  - 前向き研究がbetter
- 変数ごとに予後との関連をみる（単変量解析の結果を優先）
  - 単変量解析の結果に基づくことは必ずしも適切でない
- ~~単変量解析で有意な~~(複数の)変数と予後の関連を“多変量解析”
- 連続変数はp値が最小になるカットオフ値を用いて二値化（最小p値法）
  - 安易に二値化しない、最小p値法は危険
- 多変量解析で計算されたP値に基づいて変数選択
- 最終モデルに基づくリスク分類の作成
  - モデル選択法として決め手はないが、妥当性の検証は必須
- リスク分類ごとの生存曲線の差の比較（モデルの性能評価）
  - モデル構築に用いたデータそのものでは妥当性検証できない

59

結果を公表する際に有用な指針

- REMARKガイドライン（McShane *et al.* 2005）
  - 公表の質を上げることを目的として要点を整理

COMMENTARY

Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK)

Lisa M. McShane, Douglas G. Altman, Willi Sauerbrei, Sheila E. Taube, Massimo Gion, Gary M. Clark for the Statistics Subcommittee of the NCI-EORTC Working Group on Cancer Diagnostics

Despite years of research and hundreds of reports on tumor markers in oncology, the number of markers that have emerged as clinically useful is pitifully small. Often, initially reported studies of a marker show great promise, but subsequent studies on the same or related markers yield inconsistent conclusions or stand in direct contradiction to the promising results. It is imperative that we attempt to understand the reasons that multiple studies of the same marker lead to differing conclusions. A variety of methodologic problems have been cited to explain these discrepancies. Unfortunately, many tumor marker studies have not been reported in a rigorous fashion, and published articles often lack sufficient information to allow adequate assessment of the quality of the study or the generalizability of study results. The development

Specimen availability may be related to tumor size and patient outcome *(12)*, and the quantity, quality, and preservation method of the specimen may affect feasibility of conducting certain assays. There can also be biases or large variability inherent in the assay results, depending on the particular assay methods used *(13–17)*. Statistical problems are commonplace. These problems include underpowered studies or overly optimistic reporting of effect sizes and significance levels due to multiple testing, subset analyses, and cutpoint optimization *(18)*.

Unfortunately, many tumor marker studies have not been reported in a rigorous fashion, and published articles often lack sufficient information to allow adequate assessment of the quality of the study or the generalizability of study results. Such reporting deficiencies are increasingly being highlighted by systematic reviews of the published literature on particular markers or cancers *(19–25)*.

Journal of the National Cancer Institute, Vol. 97, No. 16, August 17, 2005

60

Table 1. Reporting recommendations for tumor marker prognostic studies (REMARK)

Journal of the National Cancer Institute, Vol. 97, No. 16, August 17, 2005

INTRODUCTION

1. State the marker examined, the study objectives, and any prespecified hypotheses.

MATERIALS AND METHODS

Patients

2. Describe the characteristics (e.g., disease stage or comorbidities) of the study patients, including their source and inclusion and exclusion criteria.
3. Describe treatments received and how chosen (e.g., randomized or rule-based).

Specimen characteristics

4. Describe type of biological material used (including control samples) and methods of preservation and storage.

Assay methods

5. Specify the assay method used and provide (or reference) a detailed protocol, including specific reagents or kits used, quality control procedures, reproducibility assessments, quantitation methods, and scoring and reporting protocols. Specify whether and how assays were performed blinded to the study endpoint.

Study design

6. State the method of case selection, including whether prospective or retrospective and whether stratification or matching (e.g., by stage of disease or age) was used. Specify the time period from which cases were taken, the end of the follow-up period, and the median follow-up time.
7. Precisely define all clinical endpoints examined.
8. List all candidate variables initially examined or considered for inclusion in models.
9. Give rationale for sample size; if the study was designed to detect a specified effect size, give the target power and effect size.

Statistical analysis methods

10. Specify all statistical methods, including details of any variable selection procedures and other model-building issues, how model assumptions were verified, and how missing data were handled.
11. Clarify how marker values were handled in the analyses; if relevant, describe methods used for cutpoint determination.

RESULTS

Data

12. Describe the flow of patients through the study, including the number of patients included in each stage of the analysis (a diagram may be helpful) and reasons for dropout. Specifically, both overall and for each subgroup extensively examined report the numbers of patients and the number of events.
13. Report distributions of basic demographic characteristics (at least age and sex), standard (disease-specific) prognostic variables, and tumor marker, including numbers of missing values.

Analysis and presentation

14. Show the relation of the marker to standard prognostic variables.
15. Present univariate analyses showing the relation between the marker and outcome, with the estimated effect (e.g., hazard ratio and survival probability). Preferably provide similar analyses for all other variables being analyzed. For the effect of a tumor marker on a time-to-event outcome, a Kaplan–Meier plot is recommended.
16. For key multivariable analyses, report estimated effects (e.g., hazard ratio) with confidence intervals for the marker and, at least for the final model, all other variables in the model.
17. Among reported results, provide estimated effects with confidence intervals from an analysis in which the marker and standard prognostic variables are included, regardless of their statistical significance.
18. If done, report results of further investigations, such as checking assumptions, sensitivity analyses, and internal validation.

DISCUSSION

19. Interpret the results in the context of the prespecified hypotheses and other relevant studies; include a discussion of limitations of the study.
20. Discuss implications for future research and clinical value.

61

Take Home Messages（まとめ）

- 予後因子解析と因果推論は異なる
  - みかけ上は同じ多変量解析モデルを用いて解析しようとも別
- 予後予測/予後因子解析の目的は結果を予測すること
  - 予測が当たっている保証が必要
  - 臨床的に役に立たないとモデルは使われない
- 予測が当たるモデルを作るには
  - データに基づいてモデル構築すると Overfitting の可能性あり
  - 構築したモデルに対する妥当性の検証が不可欠

62

附録

63

附録.　Simonらのシミュレーションの注釈

Fig. 1. Effect of various levels of cross-validation on the estimated error rate of a predictor derived from 2000 simulated datasets. Class labels were arbitrarily assigned to the specimens within each dataset, so poor classification accuracy is expected. Class prediction was performed on each dataset as described in the supplemental information (http://jncicancerspectrum.oupjournals.org/jnci/content/vol95/issue1/index.shtml and http://linus.nci.nih.gov/~brb), varying the level of leave-one-out cross-validation used in the prediction. **Vertical bars** indicate the proportion of simulated datasets (of 2000) resulting in a given number of misclassifications for a specified cross-validation strategy.

クラスを出鱈目に割り付けた＝真には遺伝子とクラスの間に関連がない

LOOCV法

横軸は誤分類数

Journal of the National Cancer Institute, Vol. 95, No. 1, January 1, 2003　64

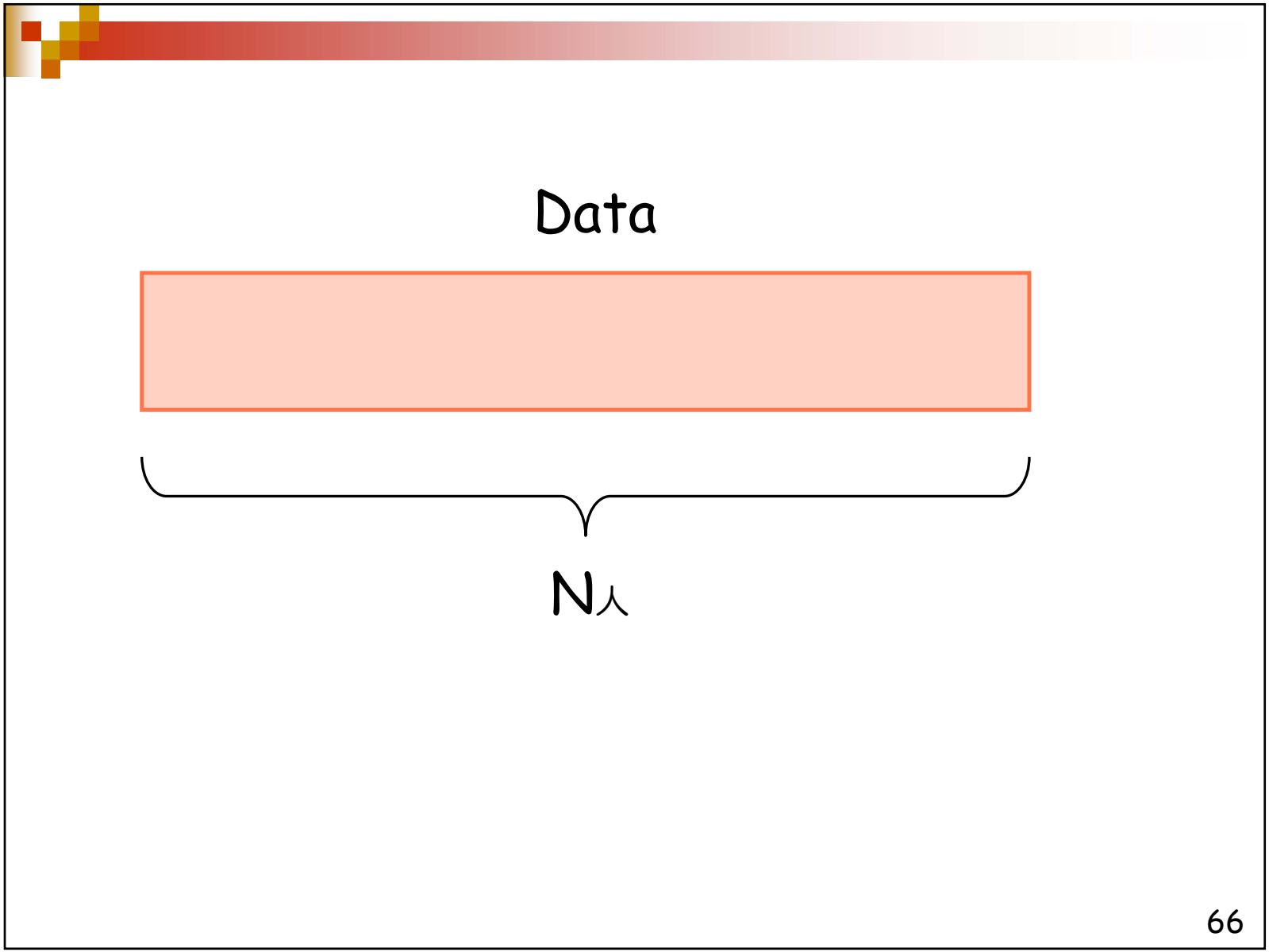
附録：S-phase fractionのカットオフ値の決め方とp値

- データに基づくCutpointの決定
  - 研究間で整合性がない
  - Optimal法（minimum-p法）を用いると更に多重性の問題

Table 1. Cutpoints for SPF used in the literature and *P* values for different patient populations in the Freiburg DNA breast cancer study

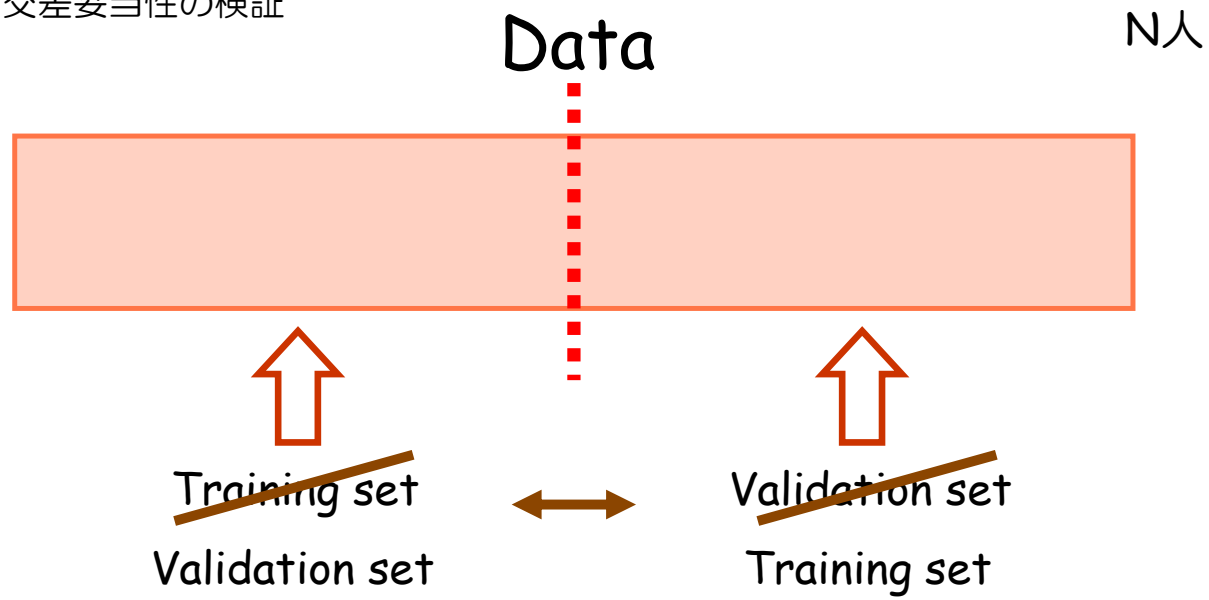
Cutpoint	Investigators, y (ref. No.)	Method	Population 1, all*	Population 2, diploid†	Population 3, node negative‡	Population 4, node positive§	Population 5, node positive, diploid‖	Population 6, node positive, aneuploid¶
2.6	Dressler et al., 1988 (*21*)	Median	.369	.338	.459	.953	.704	.333
3.0	Fisher et al., 1991 (*8*)	Median	.479	.520	.485	.928	.805	.692
4.0	Hatschek et al., 1990 (*22*)	#	.108	.181	.572	.135	.307	.170
5.0	Arnerloev et al., 1990 (*10*)	Not given	.178	.272	.884	.129	.403	**.094**
6.0	Hatschek et al., 1989 (*23*)	Median	. 160	.232	.920	**.088**	.381	.115
6.7	Clark et al., 1989 (*12*)	"Optimal"	. 170	.280	.884	.104	.565	**.084**
7.0	Baak et al., 1991 (*24*)	Not given	. 345	.452	.670	.133	.836	**.066**
7.1	O'Reilly et al., 1990 (*15*)	Median	. 540	.566	.330	.158	.836	**.091**
7.3	Ewers et al., 1992 (*25*)	Median	. 540	.566	.330	.158	.836	**.091**
7.5	Sigurdsson et al., 1990 (*9*)	Median	. 739	.860	.330	.232	.802	**.074**
8.0	Kute et al., 1990 (*11*)	Median	. 524	.996	.407	.106	.719	**.023**
9.0	Witzig et al., 1993 (*26*)	Median	. 999	.852	.838	**.044**	.962	**.023**
10.0	O'Reilly et al., 1990 (*14*)	"Optimal"	.483	.581	.164	**.012**	.962	**.006**
10.3	Dressler et al., 1988 (*21*)	Median	.316	.581	.229	**.012**	.962	**.006**
12.0	Sigurdsson et al., 1990 (*9*)	"Optimal"	.603	.613	.463	.144	.994	**.060**
12.3	Witzig et al., 1993 (*26*)	**	.755	.631	.406	.133	.994	**.047**
12.5	Muss et al., 1989 (*27*)	Median	.911	.631	.463	.233	.994	**.094**
14.0	Joensuu et al., 1990 (*16*)	"Optimal"	.616	.835	.490	.953	.994	.736
15.0	Joensuu and Toikkanen, 1991 (*28*)	"Optimal"	.220	.835	.490	.352	.994	.530

Journal of the National Cancer Institute, Vol. 86, No. 11, June 1, 1994 65


Data
N人
66

Cross-validation: Split-half 法

交差妥当性の検証



- 予測性能は平均をとって評価
- 二分の仕方はほかにもある、繰り返して二分を行えばよい
- 全てのデータがモデル構築に寄与するが、各々は半分ずつが寄与
  - 効率向上の余地

67

Cross-validation: leave-one-out 法（LOO-CV法）

Data　　N人

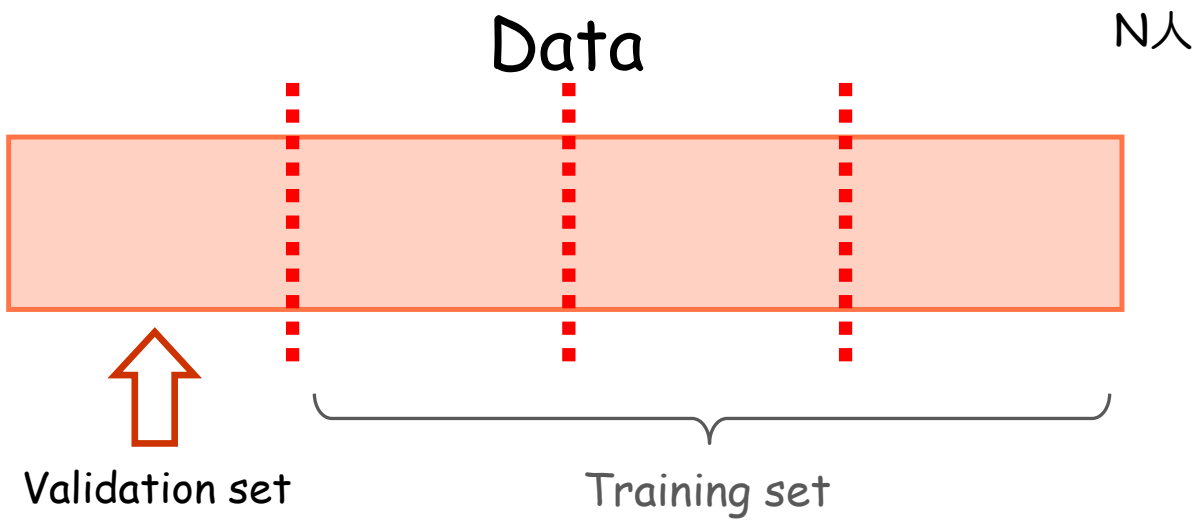
Validation set は 1患者　“one”

Training set は “one” を除く（N―1）人

- N回繰り返し、予測性能はN回の平均をとって評価
- （N―1）人でモデル構築、効率が良い
- 限界：Nがあまりにも多いと問題がある
  - 計算時間：多
  - 数理的性質（一致性）：難

Statistica Sinica 1997, 7:221-264.　68

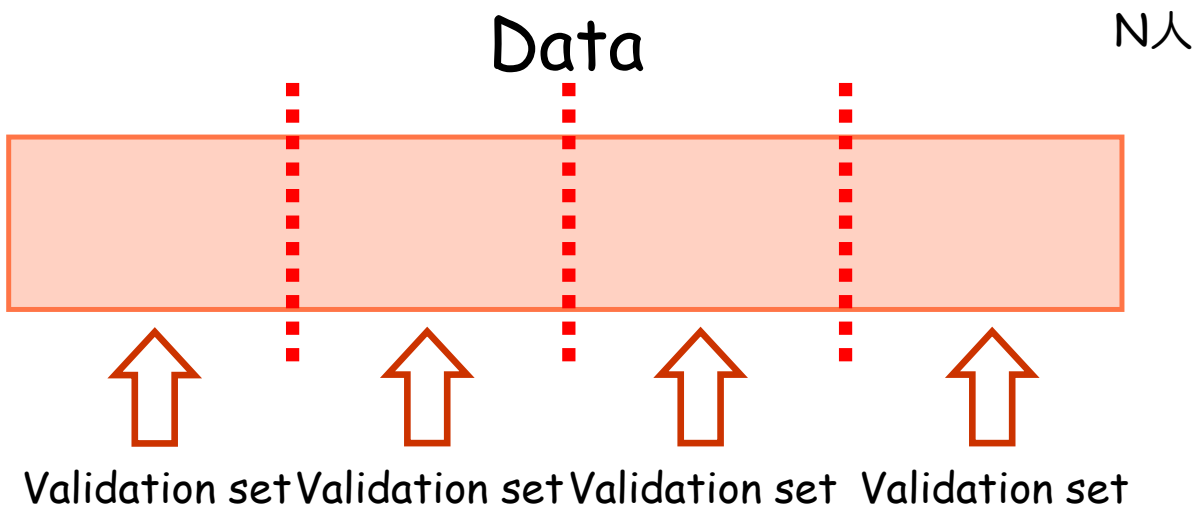
Cross-validation: 4-fold 法（一般に K-fold 法）



- Validation set を除いた残りを Training set

69

Cross-validation: 4-fold 法（一般に K-fold 法）



- Validation set を除いた残りを Training set
- 予測性能は平均をとって評価
- Leave-d-out 法
  - Leave-one-out法に比べれば容易

70

妥当性検証をおこなう対象

- 構築された「予測モデル」or「リスク分類ルール」そのもの

  validation set に基づいて予測モデルを再度みつける

  ✕ validation set で予後因子ごとの有意性を再評価

  ✕　　✓妥当性の検証：「予測式を再推定」することではない

- 予測モデル＝予測「式」

  ＝予後因子の組み合わせ　＋　係数（重み）

  ○ 既構築の予測モデルをvalidation setにあてはめて評価

- 構築プロセスの全て、データ依存したもの全て

- リスク分類が主たる興味である場合

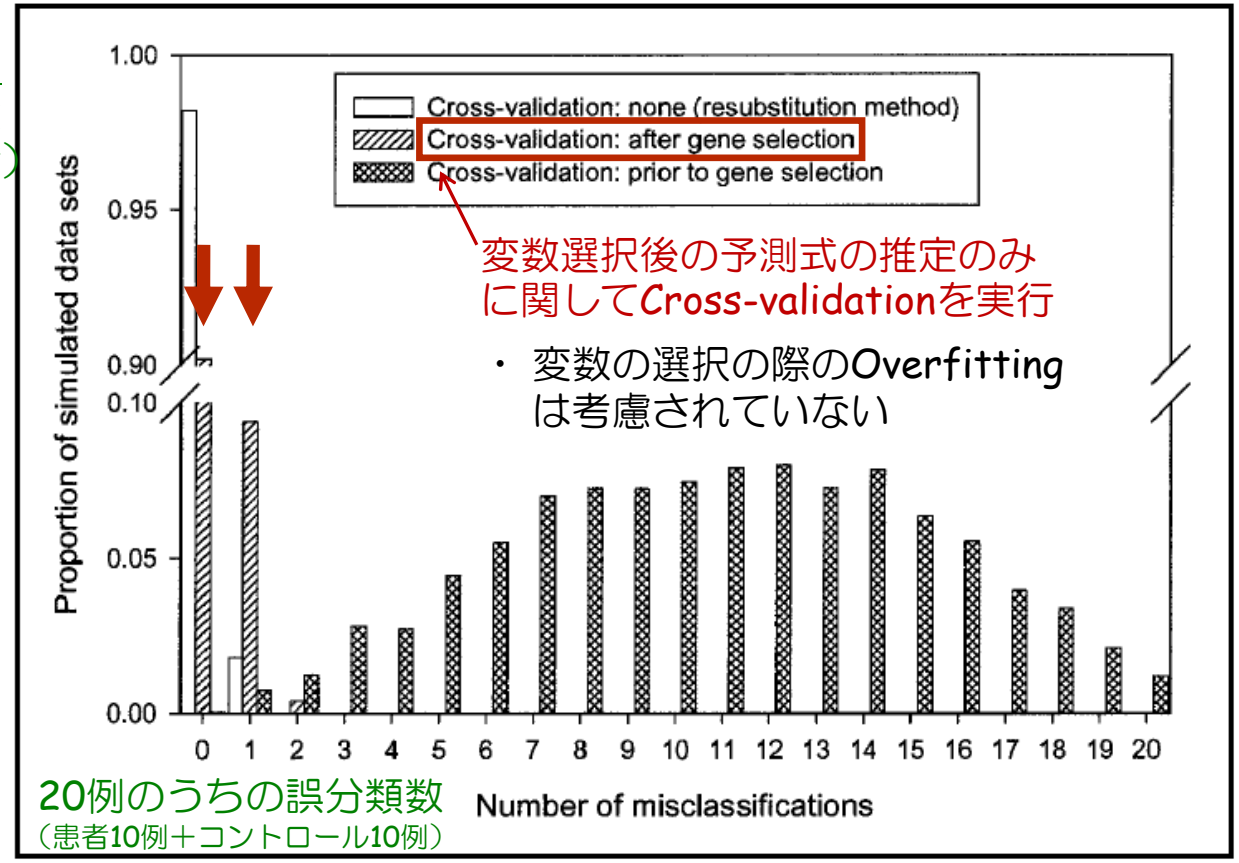
  ○ 予測モデル＋分類ルールをセットでvalidation setにあてはめ

71

再掲.　Simonらのシミュレーション

- 全「構築プロセス」を妥当性検証の対象とすべき

縦軸
頻度
（割合）



変数選択後の予測式の推定のみ
に関してCross-validationを実行

- 変数の選択の際のOverfitting
は考慮されていない

20例のうちの誤分類数
（患者10例＋コントロール10例）

Journal of the National Cancer Institute, Vol. 95, No. 1, January 1, 2003

72

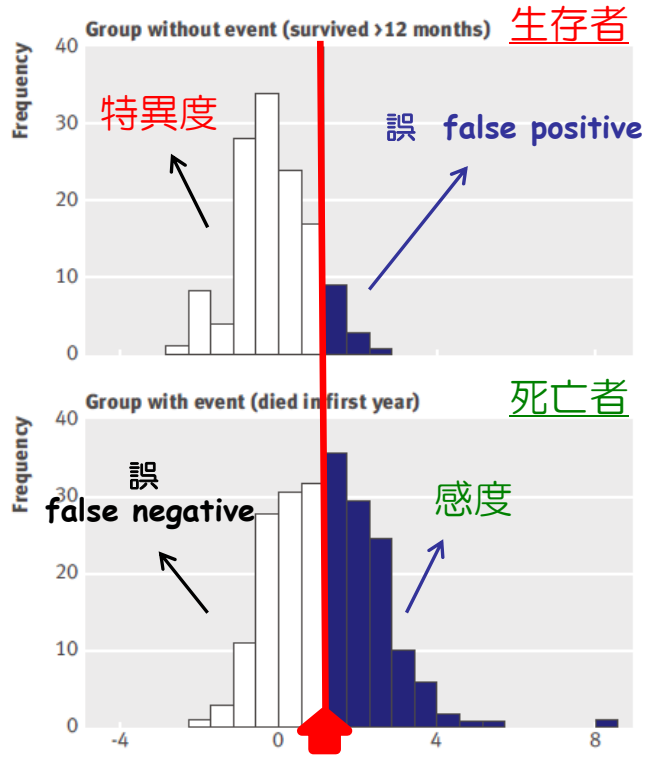


Fig 2 | Distribution of risk scores from full model for patients who survived at least 12 months or died within 12 months. The vertical line represents a risk score of 1.4, corresponding to an estimated death risk of 80%. The specificity (90%) is the proportion of patients in the upper panel whose risk is below 1.4. The sensitivity (46%) is the proportion of patients in the lower panel whose risk score is above 1.4

例. discrimination

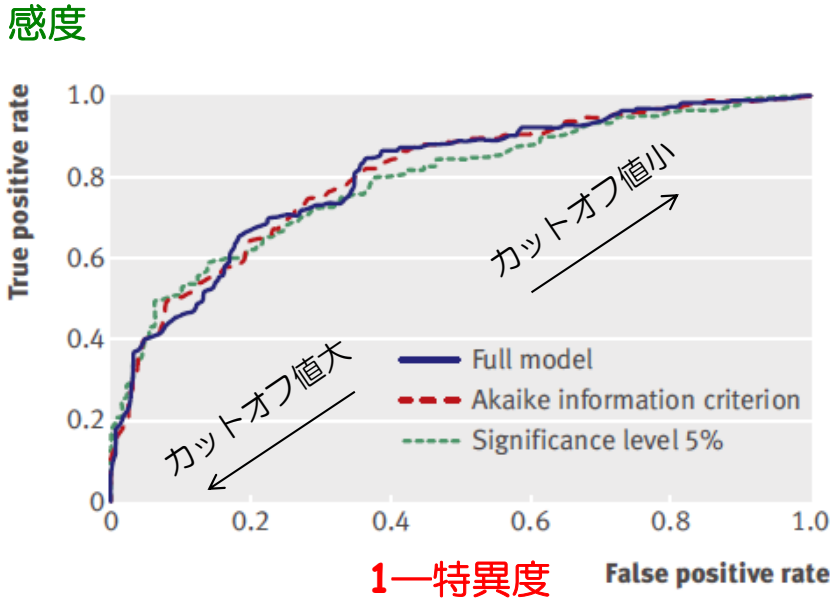
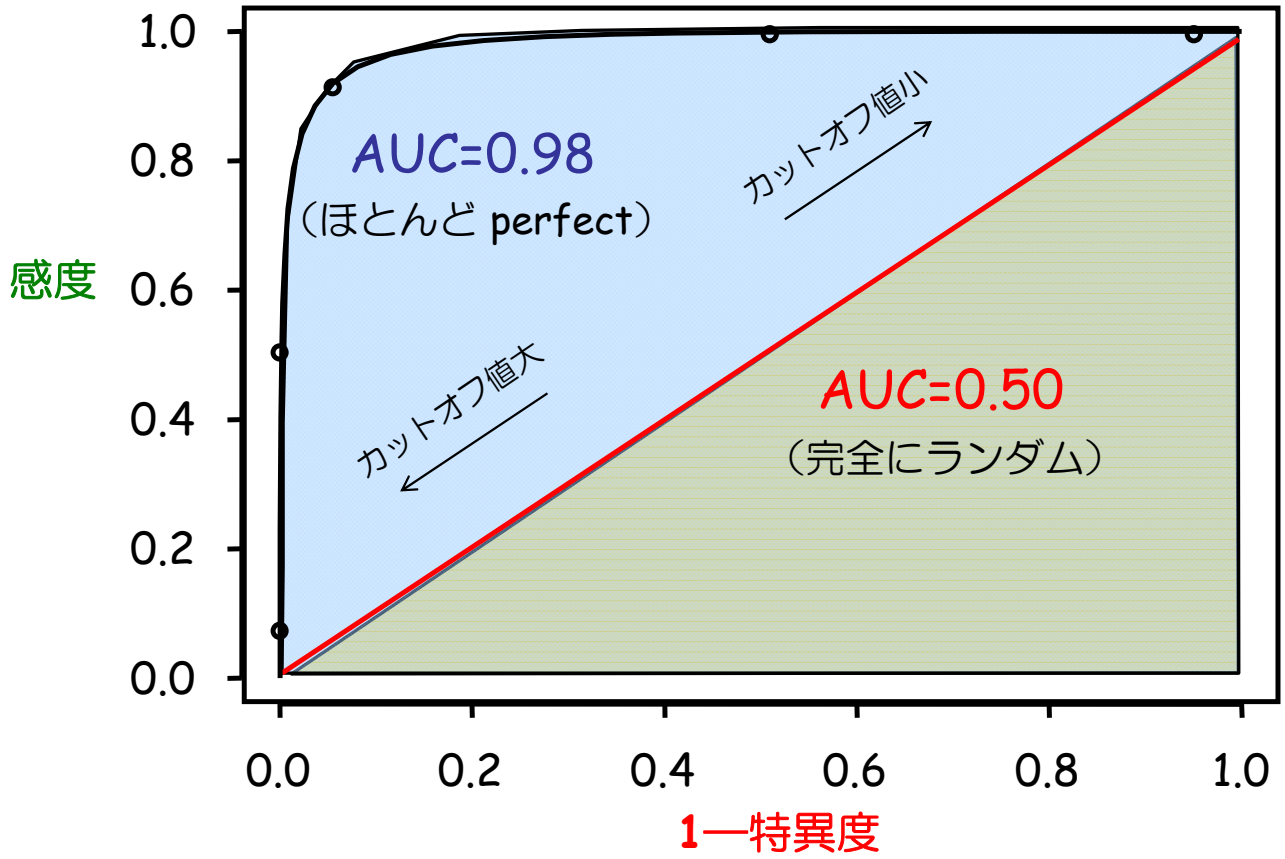


Fig 1 | Receiver operating characteristic (ROC) curves for three multivariable models of survival with kidney cancer

BMJ 2009, 338:1373-7.

ROC曲線と曲線下面積（AUC）による評価



Yapらによる妥当性研究

- ROC曲線により discrimination を評価

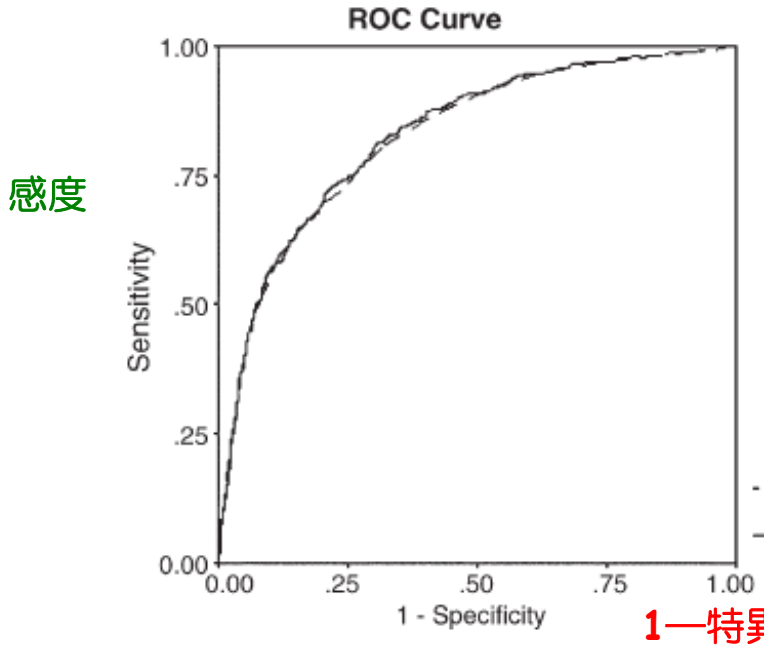


Fig. 1. Receiver operating characteristic (ROC) curves. Area under the curve is 0.83 for the additive and logistic EuroSCORE models when applied to the entire study cohort.

Eur J Cardio-thorac Surg 2006, 29:441—446.

何をもって予測性能を「良し」とするか

- Contexts に強く依存
  - 切れ味の良い "統計的規準" は難しい
  - 臨床現場で、如何なる目的で用いるのか
  - 既存の予測モデルが存在するのか
    - 存在する場合、有用な情報を追加するもの/超えるものか

- 例.　Framingham Cardiovascular Risk Score
  - 多くの報告で AUC of ROC=0.70程度
    - それでも十分に予測として役立つ、広く使用されている
    - Face validity も明らかに優れる

76

Framingham Risk Score in men

構築された分類（予測）ルールに基づいて10年間のCHD発症リスクを算出できる

Step 1: Sum of Points

Age
Total Cholesterol
HDL-C
Systolic Blood Pressure
Smoking Status
Point Total

Step 2: 10-year CHD Risk

Point Total	10-year Risk	Point Total	10-year Risk	Point Total	10-year Risk
<0	**<1%**	6	**2%**	13	**12%**
0	**1%**	7	**3%**	14	**16%**
1	**1%**	8	**4%**	15	**20%**
2	**1%**	9	**5%**	16	**25%**
3	**1%**	10	**6%**	≥17	**>30%**
4	**1%**	11	**8%**		
5	**2%**	12	**10%**		

Smoking Status Points

	Age 20-39	Age 40-49	Age 50-59	Age 60-69	Age 70-79
Nonsmoker	0	0	0	0	0
Smoker	8	5	3	1	1

Age Points

Years	Points
20-34	-9
35-39	-4
40-44	0
45-49	3
50-54	6
55-59	8
60-64	10
65-69	11
70-74	12
75-79	13

HDL-C Points

HDL-C (mg/dl)	Points
≥60	-1
50-59	0
40-49	1
<40	2

SBP Points

SBP (mmHg)	If treated	If untreated
<120	0	0
120-129	0	1
130-139	1	2
140-159	1	2
≥160	2	3

Total Cholesterol Points

TC (mg/dl)	Age 20-39	Age 40-49	Age 50-59	Age 60-69	Age 70-79
<160	0	0	0	0	0
160-199	4	3	2	1	0
200-239	7	5	3	1	0
240-279	9	6	4	2	1
≥280	11	8	5	3	1