臨床予測モデルにおける変数選択と判別・較正の方法

野間 久史
統計数理研究所
2018年12月7日
日本計量生物学会 2018年度計量生物セミナー
e-mail: noma@ism.ac.jp
URL: http://www.ism.ac.jp/~noma/

Research

Original Investigation

Clinical Risk Score for Persistent Postconcussion Symptoms Among Children With Acute Concussion in the ED

Roger Zemek, MD; Nick Barrowman, PhD; Stephen B. Freedman, MDCM, MSc; Jocelyn Gravel, MD; Isabelle Gagnon, PhD; Candice McGahern, BA; Mary Aglipay, MSc; Gurinder Sangha, MD; Kathy Boutis, MD; Darcy Beer, MD; William Craig, MDCM; Emma Burns, MD; Ken J. Farion, MD; Angelo Mikrogianakis, MD; Karen Barlow, MD; Alexander S. Dubrovsky, MDCM, MSc; Willem Meeuwisse, MD, PhD; Gerard Gioia, PhD; William P. Meehan III, MD; Miriam H. Beauchamp, PhD; Yael Kamil, BSc; Anne M. Grool, MD, PhD, MSc; Blaine Hoshizaki, PhD; Peter Anderson, PhD; Brian L. Brooks, PhD; Keith Owen Yeates, PhD; Michael Vassilyadi, MDCM, MSc; Terry Klassen, MD; Michelle Keightley, PhD; Lawrence Richer, MD; Carol DeMatteo, MSc; Martin H. Osmond, MDCM; for the Pediatric Emergency Research Canada (PERC) Concussion Team

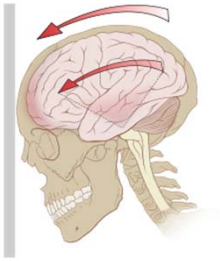
IMPORTANCE Approximately one-third of children experiencing acute concussion experience ongoing somatic, cognitive, and psychological or behavioral symptoms, referred to as persistent postconcussion symptoms (PPCS). However, validated and pragmatic tools enabling clinicians to identify patients at risk for PPCS do not exist.

Editorial page 987
JAMA Report Video at jama.com
Supplemental content at jama.com

JAMA 2016; 315(10): 1014-25.

持続的な脳震盪後症候群

- 脳震盪の多くは、一過性のものであるが、小児の患者の３分の１ほどは、持続的な脳震盪後の症状に悩まされる
  - Persistent Postconcussion Symptoms (PPCS)
- しかし、このPPCSが起こるか否かを予測するための、Validateされた実用的な予測ツールは開発されていなかった
- Zemek et al. (2016) では、カナダのオタワ大学のグループを中心に、脳震盪で救急を受診した子どもを対象として、PPCSの発症を予測するためのリスクスコアを開発した



https://ja.wikipedia.org/wiki/%E8%84%B3%E9%9C%87%E7%9B%AA

3

Clinical Review & Education

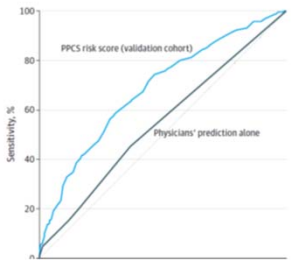
JAMA Guide to Statistics and Methods

Logistic Regression Diagnostics
Understanding How Well a Model Predicts Outcomes

William J. Meurer, MD, MS; Juliana Tolles, MD, MHS

In the March 8, 2016, issue of *JAMA*, Zemek et al[1] used logistic regression to develop a clinical risk score for identifying which pediatric patients with concussion will experience prolonged postconcussion symptoms (PPCS). The authors prospectively recorded the initial values of 46 potential predictor variables, or risk factors—selected based on expert opinion and previous research—in a cohort of patients and then followed those patients to determine who developed the primary outcome of PPCS. In the first part of the study, the authors created a logistic regression model to estimate the probability of PPCS using a subset of the variables; in the second part of the study, a separate set of data was used to assess the validity of the model, with the degree of success quantified using regression model diagnostics. The rationale for using logistic regression to develop predictive models was summarized in an earlier JAMA Guide to Statistics and Methods article.[2] In this article, we discuss how well a model performs once it is defined.

Figure. Receiver Operating Characteristic Curves



Meurer and Tolles (2017)

4

JAMA Guide to Statistics and Methods から

The authors prospectively recorded the initial values of 46 potential predictor variables, or risk factors—selected based on expert opinion and previous research…

In the first part of the study, the authors created a logistic regression model to estimate the probability of PPCS using a subset of the variables…

…in the second part of the study, a separate set of data was used to assess the validity of the model, with the degree of success quantified using regression model diagnostics.

Meurer and Tolles (2017)

5

本日のお話

- 臨床予測モデル　Clinical Prediction Models
  - Binary outcome: ロジスティック回帰モデル
  - Time-to-event outcome: Cox回帰モデル
- モデルをどのように構築し、どのように評価すればよいのか？
  - 変数選択　Variable Selection
  - 判別　Discrimination
  - 較正　Calibration
- 実例をもとに、これらの方法論について、平易な解説を行います

6

PPCSのリスクスコアの作製

- 研究デザイン：前向きの多施設共同コホート研究
- 対象者：5歳から18歳の小児で、受傷後48時間以内に救急を受診した患者
- プライマリアウトカム：受傷後、28日以内のPPCSの発症

- あらかじめ先行研究によるエビデンスと専門家による意見をもとにして選定した、46個の予後因子の候補を収集していた
- ロジスティック回帰モデルによって、PPCSの発症を予測するためのリスクスコアを開発し、評価を行った

Zemek et al. (2016)　7

ロジスティック回帰による臨床予測モデル

- ロジスティック回帰モデル

$$\Pr(Y=1|\boldsymbol{x}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}$$

- 説明変数 $x_1, x_2, \ldots, x_p$ によって、結果変数 Y $(= 0,1)$ を予測するモデル
- 数理統計学の多変量解析では、ロジスティック判別分析（logistic discriminant analysis）として知られている

McLachlan (2004)　8

臨床予測モデルの構築

学習データ（Training Data）
$(Y_1, \boldsymbol{x}_1), (Y_2, \boldsymbol{x}_2), \dots, (Y_n, \boldsymbol{x}_n)$

結果変数、説明変数の双方が測定された n 人のデータ（正解がわかっているデータ）で、予測モデルを作る

$$\Pr(Y = 1|\boldsymbol{x}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}$$

２値結果変数の予測モデルとして、ロジスティック回帰モデルを用いる

最尤推定

$$\widehat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p)^T$$

9

臨床予測モデルによる予測

将来の患者のデータ
$\boldsymbol{x}_j = (x_{j,1}, \dots, x_{jp})$

説明変数のみが測定された将来の患者のデータ（正解がわかっていないデータ）の結果変数を予測したい！

$$\widehat{\Pr}(Y_j = 1|\boldsymbol{x}_j) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{j,1} + \hat{\beta}_2 x_{j,2} + \cdots + \hat{\beta}_p x_{j,p})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{j,1} + \hat{\beta}_2 x_{j,2} + \cdots + \hat{\beta}_p x_{j,p})}$$

イベントを起こす確率の推定値。リスクスコア（Risk score）といわれる。このスコアの大小によって「イベントあり」「なし」の予測の結果を決める。

学習データから構築された予測モデル（最尤推定値をPlug-inしたロジスティック回帰モデル）にデータをあてはめる

10

変数選択　Variable Selection

- 46もの候補の変数から、どのように予測に有用な変数の選択を行う？
- All reliable variables associated with PPCS ($P < .20$) were entered into a multivariable model using **forward stepwise binary logistic regression analysis ($P = .05$ included but $P = .10$ removed).** Zemek et al. (2016)
- Zemekらは、”Forward Stepwise” 法によって、ロジスティック回帰モデルの変数の選択を行っている

11

モデル構築の原理

- 交絡調整のための回帰モデル
  - 主たる関心のある治療（曝露）についての偏りのない因果効果を推定することが目的
  - 「因果効果の推定の正確性」という基準から「良いモデル」を構築する
- 臨床予測モデル
  - 将来の患者の結果変数を、より正確に予測することが目的
  - 「将来のデータの予測の正確性」という基準から「良いモデル」を構築する

12

“Parsimony” の原理

- 測定された変数すべてを説明変数としてモデルに組み込む解析
- “Parsimony” の原理
  - 重要性の低い、無駄な情報は、加えても予測の性能を向上させないだけではなく、悪化させる可能性もある
  - 少数の高い予測能力のある変数を用いることが、オーバーフィッティングを防ぎ、予測性能も高くなる傾向がある
  - 予測モデルの実用化の際も、測定するべき必要な情報が絞られる

Ockham chooses a razor

https://www.google.co.jp/url?sa=i&source=images&cd=&cad=rja&uact=8&ved=2ahUKEwj8muvk6aXeAhXGzLwKHd1zBxYQjRx6BAgBEAU&url=https%3A%2F%2Fwww.safalniveshak.com%2Flatticework-mental-models-occams-razor%2F&psig=AOvVaw0fzAf-z8CDpgRJZs-hAc6S&ust=1540702375243042

13

Stepwise法

- Stepwise法は、最も有意性の大きな変数の組を選択するためのアルゴリズム
- 選択の基準となる有意性は、線形回帰モデルでは F検定、ロジスティック回帰やCox回帰では、尤度比・Wald・スコア検定などで評価される
  - 有意性の指標以外にも、Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Mallow’s C_p なども

Akaike (1973), Mallows (1973), Schwarz (1978)

14

Stepwise法のアルゴリズム

- Forward法：なにも説明変数を加えていないモデルからスタートして、有意性の大きな候補の変数を１つ１つ加えていく
- Backward法：すべての候補変数を含むFull modelからスタートして、有意性の小さな変数を１つ１つ除外していく
- 一般的には、Backward法のほうが好まれる
  - Full Modelからスタートするので、すべての変数の関連を同時に評価することができる
  - 相関のある予測性能の高い変数の組が同時にモデルに残る可能性がある（Forward法ではすべてドロップする可能性もある）

Derksen and Keselman (1992), Steyerberg (2009) 15

Stepwise法の Stopping Criterion

- 有意性の指標によって判定を行う際には、有意水準 5%が用いられることが一般的
- 有意水準は緩めに設定してもよい（e.g., 20%，50%など）
  - サンプルサイズが小さい条件下では、有意水準を厳しくしすぎると（5%など）、予測性能は悪くなる
  - ただし、弱い関連しか持たない変数が残ることもある
- AICよりBICのほうが小さいモデルを選択する傾向がある
  - Ambler, Brady and Rosyton (2002) は、AIC の使用を勧めている

Steyerberg et al. (2000), Steyerberg (2009) 16

Stepwise法の問題点

- 変数選択の不安定性
  - 選ばれる「変数の組」が安定しない
  - 同じコホートから、一部の少し違う対象者のサブセットを除外してStepwise法にかけると、違う変数の組が残るなど
- 回帰係数の推定量は不偏性を失う（バイアスが生じる）
- 標準誤差，信頼区間，P値も妥当性を失う
  - 最終モデルは、何度も何度も回帰モデルの分析を行い、有意性の高い変数が選ばれた後のモデルであるため
  - モデルが１つに特定されたもとで妥当性が担保される方法であるため

Steyerberg et al. (1999, 2000), Steyerberg (2009) 17

“Noise Variables” の影響

- 実際にはアウトカムと関連のない “Noise Variables” は、Stepwise法でも、かなりのものが選択されることも
  - Derksen and Keselman (1992) のシミュレーションでは、30-70%の選ばれた変数がノイズであったとも
- 少数の Noise Variables が誤って含まれてしまうと、当然ながら、予測性能も下がる
- また、少数の Noise Variables の混入はそれほど大幅に予測性能を下げることはなく、重要な予後因子が含まれないことのほうが予測性能への影響は大きい

Steyerberg (2009) 18

Univariate Screening

- 単変量解析を行って、P < 0.05, 0.1, 0.2, 0.5 などの基準に合致した変数を残して、多変量モデルを構築する、または、そこから、Backword Stepwise法にかけるというアプローチ
- ほとんど関連の認められなかった変数をあらかじめ除いて検討を行うことができる
- 大規模なデータベースを利用した、共変量の多い研究などでは、計算コストを節減することもできる
- 実践的には、よく採用される方法であるが、Stepwise法全般と同じ欠点を有する方法である

Steyerberg (2009) 19

External Information の利用

- 既存のエビデンスや専門家の “expert opinions” を利用して、変数の選択を行う
  - 既存の文献情報が蓄積されていないことも
  - 過去の研究が “false positive” の結果を報告していても、興味深い結果だと捉えられると、誤った変数が選ばれる傾向がある
  - 専門家の意見に、バイアスが含まれることも

Greenland (2008), Walter and Tiemeier (2009), Steyerberg (2009) 20

Bootstrap法を用いた選択

- Bootstrapサンプルに対して（e.g., 1000 sets）、Backward Stepwise法での変数選択を行い、最終モデルに残った頻度によって変数をランキングする
- 一定の頻度以上（e.g., >50%）残った変数を最終的に採用する
- ただし、この方法は、Heuristicなアプローチであり、まだその「手法の良さ」を保証する理論的根拠は与えられていない
- 最終的なモデルに残る変数の組は、Stepwise法と似通ったものになることが知られている

Austin and Tu (2004), Chen and George (1985), Sauerbrei and Schumacher (1992) 21

Bagging (bootstrap aggregating)

- Bootstrapサンプルに対して（e.g., 1000 sets）、予測モデルを作成し、それぞれの予測の結果を “aggregate” した予測モデルを構築する機械学習の方法
- それぞれの予測モデルによる予測の結果を「平均」して、最終的な予測の結果とする
- 機械学習のアンザンブル学習（ensemble learning）といわれる方法
  - Boostingといわれる方法も（判別のうまくいかない対象の重みを大きくする適合的な予測アルゴリズム）

Breiman (1996), Schapire (1990) 22

Bayesian Model Averaging

- p 個の説明変数をもつ観測データ D に対して、すべての考え得るモデル $M = \{M_1, \ldots, M_m\}$ を考える（$m = 2^p$）。このとき、関心のあるパラメータ δ の「平均化」された事後分布

$$\Pr(\delta|D) = \sum_{i=1}^{m} \underbrace{\Pr(\delta|M_i, D)}_{\text{モデル } M_i \text{ のもとでの事後分布}} \underbrace{\Pr(M_i|D)}_{\text{モデル } M_i \text{ の事後確率（寄与率を調整する重みに！）}}$$

  を用いた推測を行う方法
- モデル選択の不確実性を推測・予測に取り込んだ方法

23

Bayesian Model Averaging

- p が大きいと候補モデルの数が膨大に！（e.g., $2^{10=}1024$）
- Occam's window：最良のモデルに比べて、大幅にモデルの事後確率が小さいモデルをあらかじめ除外する（e.g., 事後確率が、"Best model" よりも20分の1未満のモデルは除外する）
- シミュレーション実験によって、Stepwise法よりも正確に、真の関連がある変数を採用し、ノイズ変数を除外する傾向があることが示されている
- Framingham研究への応用などでは、Stepwise法よりも高い予測性能を示すことも示されている

Raftery et al. (1997), Wang et al. (2003) 24

Lasso (least absolute shrinkage and selection operator)

- 対数（部分）尤度関数に、回帰パラメータの絶対値の和が一定の値以下になるように制約をつけた罰則化推定を行う方法
- 制約の数理的な性質により、全体的に回帰係数は小さめの値をとる
- 一部の変数は、強く縮小され、点推定値が「0」になるため、自動的に変数選択のアルゴリズムにもなる
- シミュレーション実験・事例研究によって、Stepwise法などの従来のスタンダードな方法よりも、優れた性能を示したという報告も

Vach et al. (2001), Steyerberg (2000, 2001) 25

補足：変数選択・モデリング

- どの方法にも一長一短があり、「真のモデル」が未知である以上、推定量の不偏性・推測の妥当性はまず成り立たない
- しかし、予測モデルの目的は「高い予測精度を達成すること」であり、モデルは、高い予測精度を達成するための「作業モデル（working model）」であるという考え方も
  - 正しいモデルを特定することは必要条件ではない
  - 推定量の一致性や CI, P値の妥当性も同様
- 機械学習の手法などを活用することも検討されており、今後、さらなる発展も期待される

Uno et al. (2007, 2011), Bagherzadeh-Khiabani et al. (2016) 26

JAMAの事例より

Zemek et al. (2016)

- Emphasizing clinical relevance and face validity, predictors with continuous outcomes were categorized or dichotomized.
- All reliable variables associated with PPCS ($P < .20$) were entered into a multivariable model using forward stepwise binary logistic regression analysis ($P = .05$ included but $P = .10$ removed).
- 臨床的有用性と表面的妥当性を重視して、連続アウトカムに対する予測因子は、カテゴリ化もしくは二値化された（Sullivan et al. (2004) の方法による）。
- PPCSと相応の関連があったもの（単変量解析で P < .20）から、Forward Stepwise法によって、多変量モデルに含める変数の選択を行った ($P = .05$ included but $P = .10$ removed)。

- The final multivariable model included (1) age, (2) sex, (3) prior concussion with symptom duration of longer than 1 week, (4) physician-diagnosed migraine history, (5) headache, (6) sensitivity to noise, (7) fatigue, (8) answering questions slowly, and (9) abnormal tandem stance (Table 5).
- 最終的な多変量モデルには、上記の９つの変数が選ばれた。

27

Table 5. Selected Predictor Variables for Multivariable Model of Persistent Postconcussive Symptoms (PPCS) at 28 Days in the Derivation Cohort[a]

	No. of Risk Points for PPCS	Odds Ratio (95%CI)	P Value
Age group, y			
5-7	0	1 [Reference]	
8-12	1	1.54 (1.09-2.19)	<.001
13-<18	2	2.31 (1.62-3.32)	
Sex			
Male	0	1 [Reference]	<.001
Female	2	2.24 (1.78-2.82)	
Prior concussion and symptom duration			
No prior concussion; symptom duration <1 wk	0	1 [Reference]	.01
Prior concussion; symptom duration ≥1 wk	1	1.53 (1.10-2.13)	
Physician-diagnosed migraine history			
No	0	1 [Reference]	.001
Yes	1	1.73 (1.24-2.43)	
Answering questions slowly			
No	0	1 [Reference]	.008
Yes	1	1.37 (1.08-1.74)	
Balance Error Scoring System tandem stance No. of errors			
0-3	0	1 [Reference]	.02
≥4 or Physically unable to undergo testing	1	1.31 (1.04-1.66)	
Headache			
No	0	1 [Reference]	.01
Yes	1	1.66 (1.11-2.48)	
Sensitivity to noise			
No	0	1 [Reference]	.002
Yes	1	1.47 (1.15-1.87)	
Fatigue			
No	0	1 [Reference]	<.001
Yes	2	1.84 (1.37-2.46)	

[a] There were 1701 patients in the derivation cohort included in the primary analysis.

リスクスコアの計算・解釈のしやすさのために、連続変数をカテゴリ化して、予測モデルを構築する方針もよく採られる

この研究では、左記のカテゴリに割り振られたスコアごとに、整数値のスコア（0-12点）を計算し、リスクスコアを計算することに（Sullivan et al. (2004) の方法による）

Zemek et al. (2016) 28

臨床予測モデルの評価：判別と較正

- 判別（Discrimination）
  - モデルが正しく高リスクの患者を高リスクと分類できる能力
- 較正（Calibration）
  - 個々人におけるイベントを起こす確率の推定値を正確に各患者・グループに割りふれる能力

Meurer and Tolles (2017)

29

判別　Discrimination

- 予測モデルによって、イベントの発生の有無（{0,1} の結果）をどの程度正確に予測できたかを評価する指標
  - 感度　Sensitivity，特異度　Specificity
  - 診断精度　Diagnostic Accuracy
  - 陽性的中率　Positive Predicative Value
  - 陰性的中率　Negative Predicative Value
  - 陽性尤度比　Positive Likelihood Ratio
  - 陰性尤度比　Negative Likelihood Ratio
- 個々の指標の詳細については、BACK UPをご参照ください

30

感度，特異度

- 感度　Sensitivity
  - 将来、実際にイベントを起こす患者が、予測モデルによって「イベントを起こす」と判定される確率
- 特異度　Specificity
  - 将来、実際にはイベントを起こさない患者が、予測モデルによって「イベントを起こさない」と判定される確率

$$\mathrm{Se} = \frac{\#\ \text{True positives}}{\#\ \text{True positives} + \#\ \text{False negatives}}$$

$$\mathrm{Sp} = \frac{\#\ \text{True negatives}}{\#\ \text{True negatives} + \#\ \text{False positives}}$$

31

臨床予測モデルの感度・特異度

$$\widehat{\Pr}(Y = 1|\boldsymbol{x}) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p)}$$

推定されたロジスティック回帰モデルによるイベント発生確率の推定値（もしくは、その単調な変換となる線形予測子 $\hat{\boldsymbol{\beta}}^T\boldsymbol{x}$）を判別の基準スコアとして用いる

実践的には、イベント発生確率の推定値を基準スコアとして用いることが多い

アウトカム予測の際のカットオフをどこに設定するかで、判別の精度（e.g., 感度・特異度）は異なる

Wilson et al. (1998)　32

カットオフと感度・特異度の関係

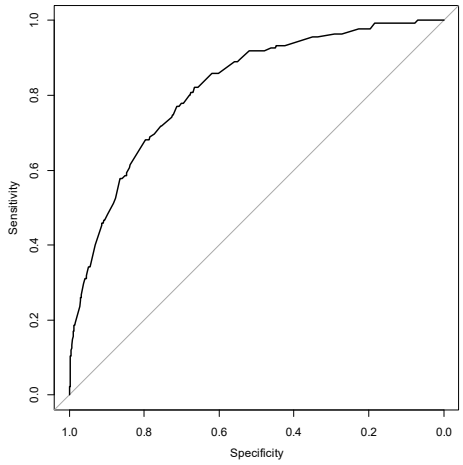
Discrimination Rule	感度	特異度
$\widehat{\Pr}(Y=1\|\boldsymbol{x})>0.05$	95.0%	2.0%
$\widehat{\Pr}(Y=1\|\boldsymbol{x})>0.10$	90.0%	22.5%
$\widehat{\Pr}(Y=1\|\boldsymbol{x})>0.20$	75.0%	50.0%
$\widehat{\Pr}(Y=1\|\boldsymbol{x})>0.50$	70.0%	50.0%
$\widehat{\Pr}(Y=1\|\boldsymbol{x})>0.75$	50.0%	82.0%
$\widehat{\Pr}(Y=1\|\boldsymbol{x})>0.80$	40.0%	88.0%
$\widehat{\Pr}(Y=1\|\boldsymbol{x})>0.90$	30.0%	95.0%

リスクスコアの分布が決まれば、カットオフごとに感度・特異度が決まる

感度・特異度は、トレードオフの関係にあり、どちらか一方が大きくなれば、もう一方は小さくなる

33

ROC (receiver operating characteristic) 曲線



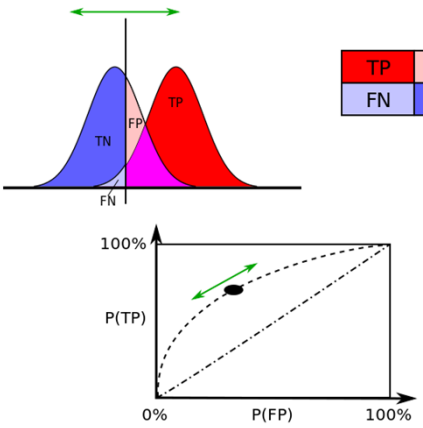
縦軸に感度、横軸に（１－特異度）をとった２次元平面上の曲線

連続量の診断マーカー（スコア）に対して、陽性・陰性を決めるカットオフを小刻みに変えていったときの感度，特異度の値を連続的にプロットした折れ線グラフ

予測スコアの総合的な予測性能を評価するためのグラフとして用いられる

34

ROC曲線の解釈



有病者・非有病者での診断マーカー（スコア）の分布が左のような２つの一峰性の山の分布になっていたとする。カットオフを左から右に大きくしていくと、TNの患者は増えていき（特異度が上がり）、逆にTPの患者は減っていく（感度は下がっていく）。

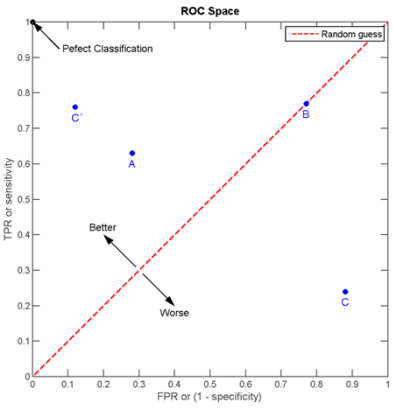
マーカーの分布が決まってしまえば、カットオフを変えても、感度・特異度の両方を同時に大きくすることはできない（トレード・オフの関係）

ROC曲線とは、両者の分布が Fix されたもとで、カットオフをさまざまに変えたもとでの総合的な診断精度を１枚のグラフに要約したものとなる。

https://en.wikipedia.org/wiki/Receiver_operating_characteristic

35

ROC曲線の解釈



まったくのでたらめの診断（e.g., コインを投げて表か裏かで陽性・陰性を決める）でも、半分は正しい診断が可能。この結果に対応しているのが、平面上の45度の斜め線（y=xの関数）になる。これより右下にくるROC曲線は、でたらめな診断よりも、診断性能が悪いことに。

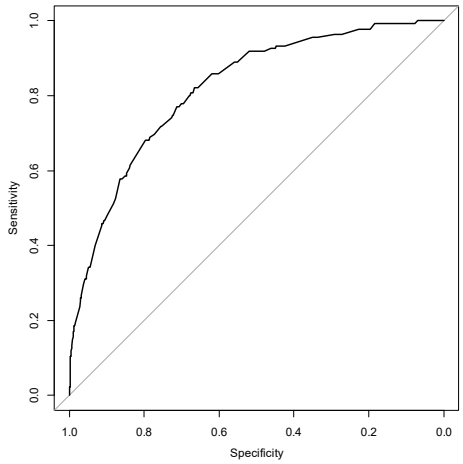
一方で、カットオフの値に関わらず、感度・特異度がともに常に100%だったとすると、ROC曲線は、平面の左上の端を通ることに（Perfect Classification!!）。

AよりCは、感度・特異度ともに高いことに。Bはでたらめの診断と同程度の診断精度です。Cはでたらめの診断より診断精度は悪いことになる。

https://en.wikipedia.org/wiki/Receiver_operating_characteristic

36

AUC (area under the curve)



ROC曲線の下側面積のことをAUC（AUROCとも）という。ひとつの辺の長さが１の正方形の箱の中に定義される曲線なので、理論上、0から1の間に値をとる指標になる。

でたらめな診断であれば、45度の直線になるので、AUCは0.50に。完璧な診断であれば、左上にピッタリくっつくROC曲線になるはずなので、AUCは1になる。

AUCは、ROC曲線をもとに、カットオフによらない診断マーカー（スコア）の総合的な診断性能を１つのスコアに要約した指標。C 統計量（concordance statistics）ともいう。

37

AUCと診断性能の目安

AUC	結果の解釈
1.0	Perfect Classification!!
0.90-1.0	Excellent
0.80-0.90	Good
0.70-0.80	Fair
0.60-0.70	Poor
0.50-0.60	Fail
0.50	Equivalent to Random Classification…
< 0.50	Worse than Random Classification…

38

C統計量　C Statistics

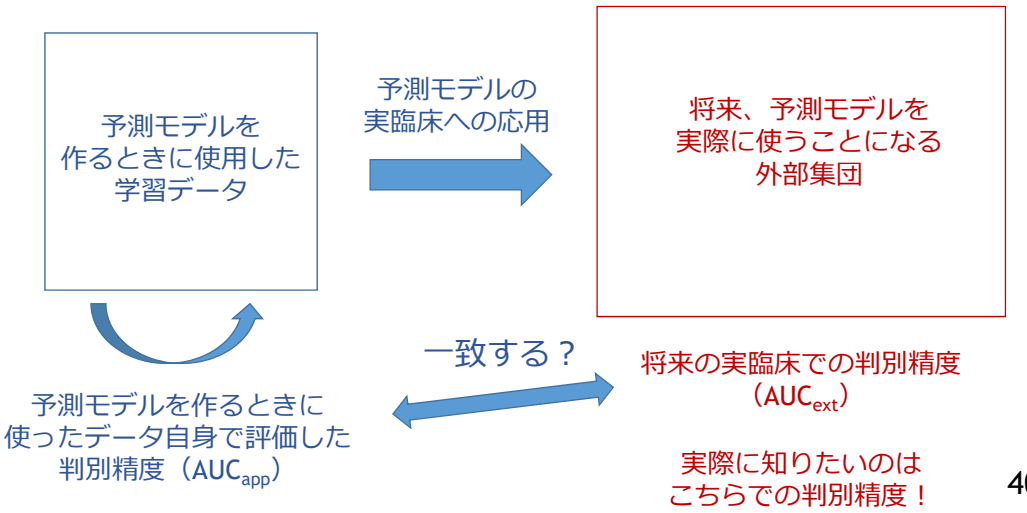
- ROC曲線のAUCは「イベントを起こした患者とイベントを起こさなかった患者から、ランダムに１名ずつのペアを取り出してきたとき、前者のリスクスコアのほうが大きくなる確率」に一致する

$$\mathrm{AUC} = \Pr(\lambda_i > \lambda_j | D_i = 1, D_j = 0)$$

- 標本推定量は、一致度（concordance）の程度を表すMann-Whitney統計量に一致する（標本分散の推定量は、Mann-Whitney統計量の分散から求めることができる）
- Concordanceの頭文字から、C統計量といわれる

DeLong et al. (1988)　39

予測モデルの “optimism” の問題



40

“optimism” の調整の必要性

- 学習データに対する感度・特異度、ROC曲線やAUCは、一般的には、外部集団での予測に使った場合よりも、結果が「良く」出過ぎてしまう！
  - 一般的に「$AUC_{app} \geq AUC_{ext}$」
  - 予測モデルは、学習データのデータを最も正確に分類するように回帰パラメータを推定しており、AUC_{app}はその学習データに対する判別精度の指標であるため（予測モデルは、学習データにオーバーフィッティングする！）
- 予測精度の過大評価のバイアス（’optimism’ といわれる）

41

Split-Sample法

対象集団をランダムに、予測モデル構築のための学習データ（Training Data）と、その性能評価のためのテストデータ（Test Data）の２つに分割する

学習データで作製した予測モデルによって、テストデータがどの程度正確に判別できるかを判定する

「予測モデルを作るのに使ったデータ」と「評価に用いるデータ」が重複する場合、予測精度の過大評価のバイアスが生じるため、両者を別々のものになるように設定することで、バイアスのない推定値を得る

42

K-fold Cross-Validation (CV) 法

Training Data	Training Data	Training Data	Test Data	Training Data

対象集団をランダムに、*K* 組の均等なサブセットに分けて、順繰りに、(*K* – 1) 組のデータを学習データに、残りの 1組のデータをテストデータにする

K 通りのTraining-Testの組み合わせすべてで、テストデータによる予測モデルの予測精度を評価して、その総計をとったものを最終的な予測精度の推定値とする

K = *N* (sample size) のCV法を leave-one-out (LOO) CVという（1個抜きのCV）。LOOCVは、実践的にはよく用いられるが、データセットにかなり依存した予測精度の推定値が得られるので、*K* は、5 or 10 程度に設定することが推奨されている（Hastie et al., 2008）

43

Bootstrapバイアス補正法による補正①

- 1. オリジナル集団における判別指標 θ の推定量を θ_{app}とする
  - θ は、C統計量など、推測を行いたい指標
- 2. オリジナル集団からB回のリサンプリングを行い、B 組のBootstrapサンプルを作成する。それぞれのBootstrapサンプルに対して、評価指標 θ の推定量 $\theta_{1,\mathrm{boot}}, \dots, \theta_{B,\mathrm{boot}}$を求める。同様にして、 B組のBootstrapサンプルで構成されたそれぞれの予測モデルによって、オリジナル集団のアウトカムの予測を行い、得られる推定量を、$\theta_{1,\mathrm{orig}}, \dots, \theta_{B,\mathrm{orig}}$ とする。

Harrell et al. (1996)　44

Bootstrapバイアス補正法による補正②

- 3. optimismの Bootstrap 推定量は、

$$\Lambda = \frac{1}{B}\sum_{b=1}^{B}\left(\theta_{b,\mathrm{boot}} - \theta_{b,\mathrm{orig}}\right)$$

  となる
- 4. オリジナル集団ではじめに求めた θ_{app} から optimism Λ を差し引いて、optimismを補正した θ の推定量は、$\theta_{\mathrm{app}} - \Lambda$ となる

Harrell et al. (1996)　45

Bootstrap法によるその他の補正方法

- Efron の .632 推定量，.632+ 推定量
- Cross-Validation法の考え方から、リサンプリングごとのサンプル内・サンプル外のデータでの予測指標の推定量を用いて求めた optimism を補正した推定量

Efron (1983), Efron and Tibshirani (1997)　46

シミュレーション実験：どの手法がよい？

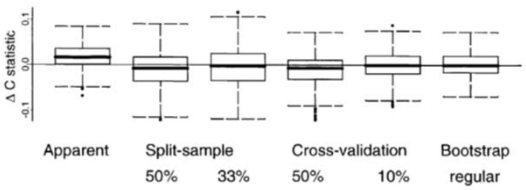


Fig. 2. Differences between estimated and test performance in data sets with EPV = 10, according to some of the studied internal validation procedures: apparent performance, split-sample validation on 50% or 33% of the sample, cross-validation on 50% or 10% of the sample, and regular bootstrapping.

Steyerberg et al. (2001)

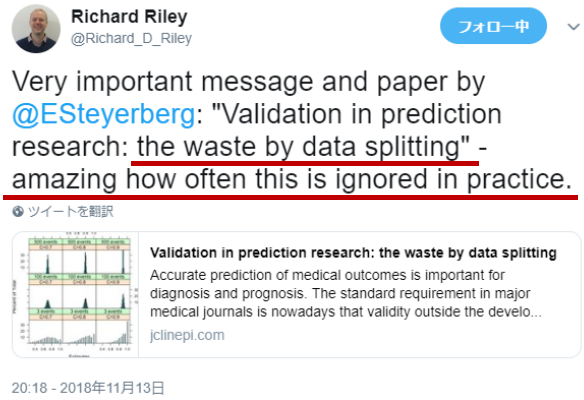
Apparent：”optimism” を調整しない推定量には過大推定のバイアスがある

Split-sample：対象者集団をTraining-Testに分割するため、予測精度の推定にバイアスはないが、ばらつきが大きい（わかりやすく、単純なので、よく使われているが、**推奨されない**！）

Cross-validation, Bootstrap：バイアス，標準誤差ともに同程度。.632, .632+ も、概ね同程度であり、この中のいずれかが推奨される

47

“The Waste by Data Splitting”



https://twitter.com/Richard_D_Riley

48

- In the absence of sufficient sample size, independent validation is misleading and should be dropped as a model evaluation step.
- We should accept that small size studies on prediction are exploratory in nature, at best show potential of new biological insights, and cannot be expected to provide clinically applicable tests, prediction models or classifiers.
- Validation studies should have at least 100 events to be meaningful. In Big Data, heterogeneity in model performance should be quantified rather than average performance.

Steyerberg (2018) 49

JAMAの事例より

Zemek et al. (2016)

- Bootstrapping analysis (ie, resampling the model 1000 times) revealed a mean overoptimism value of 0.01 (95% CI, -0.02 to 0.03) and a corrected AUC of 0.70. In the final derivation model, 94.3% (1604/1701) of the participants with primary outcome data had complete data on all 9 predictor variables included in the multivariable model. The PPCS risk score derived from the multivariable model (score range, 0 to 12) linearly corresponded to risk estimate. Three cutoff points were selected to stratify PPCS risk (low risk: ≤3 points; medium risk: 4-8 points; and high risk: ≥9 points; Table 6).

- ブートストラップ法による解析（1000回のリサンプリングによる）によって、AUCの過大推定バイアス は 0.01 (95%CI: -0.02, 0.03) であり、修正後のAUCの推定値は 0.70であった。
- 最終的なDerivationモデルでは、94.3% (1604/1701) の患者が、すべての9つの変数が測定されており、モデルに含めることができた。PPCSのリスクスコアは、0から12にスコア化された。
- 3つのカットオフが、PPCSリスクの層別化のために選ばれた (low risk: ≤3 points; medium risk: 4-8 points; and high risk: ≥9 points)。

Zemek et al. (2016) 50

Table 6. Risk Categories for Persistent Postconcussive Symptoms (PPCS) in the Derivation Cohort[a]

PPCS Risk Category	Total No. of Risk Points	Estimated Risk of PPCS, % (95% CI)	No. With PPCS/ Total No. of Patients (%)
Low risk	0	4.1 (2.4-6.7)	0/6 (0)
	1	5.8 (3.9-9.5)	6/37 (16.2)
	2	8.3 (6.0-13.2)	11/98 (11.2)
	3	11.8 (8.5-17.8)	15/165 (9.1)
Medium risk	4	16.4 (11.9-22.4)	41/239 (17.2)
	5	22.3 (16.7-29.7)	71/289 (24.6)
	6	29.7 (22.7-37.9)	90/299 (30.1)
	7	38.2 (30.1-46.9)	96/243 (39.5)
	8	47.6 (38.9-57.1)	80/172 (46.5)
High risk	9	57.1 (48.2-65.6)	58/103 (56.3)
	10	66.1 (57.2-74.4)	30/43 (69.8)
	11	74.1 (65.8-81.5)	9/13 (69.2)
	12	80.8 (74.6-88.3)	3/3 (100)

[a] There were 1701 patients in the derivation cohort included in the primary analysis.

Zemek et al. (2016)

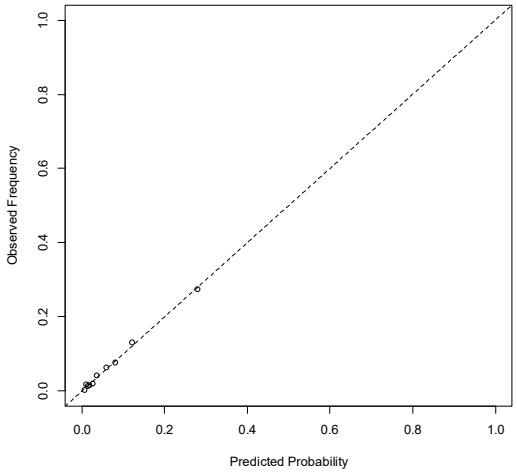
51

較正　Calibration

- 判別の性能が高いモデルでも、較正（イベント発生確率そのものの推定）の性能は悪いということもあり得る
- 例えば、とある予測モデルが、長期的な脳損傷のリスクが高い患者に「99%の確率でイベントが起こる」と、高い確率を正しく割り振っていたとする。判別は正しくできていることに。しかし、その患者のイベント発生確率が、実際には 80% であれば、「リスクの絶対値の推定」はうまくできていないことになる。
- この「リスクの絶対値の推定」の評価を較正という

52

Calibration Plot



予測スコアの大きさによって、データセットを J のグループに分けて、それぞれの
① 観測頻度：実際のイベント発生割合
② 期待頻度：ロジスティック回帰モデルから予測されたイベントの発生割合
を２次元平面上にプロットしたもの

２つの指標が近い値をとっているほど、モデルの「較正」はうまくできていることになる

斜め 45度の y = x の直線上に、 J のデータのプロットがどの程度近いかを見ます

J = 10 が最も一般的に用いられている

53

Hosmer-Lemeshow検定

	cutyhat	Obs	Exp
1	[0.00611,0.00793]	0.002849003	0.007127305
2	(0.00793,0.0116]	0.018867925	0.010751633
3	(0.0116,0.0151]	0.012448133	0.014511585
4	(0.0151,0.0183]	0.016483516	0.017894785
5	(0.0183,0.0306]	0.019607843	0.026319190
6	(0.0306,0.0395]	0.042553191	0.036340677
7	(0.0395,0.0686]	0.064516129	0.058820127
8	(0.0686,0.0874]	0.076502732	0.080337083
9	(0.0874,0.151]	0.131221719	0.122330135
10	(0.151,0.872]	0.274111675	0.278438755

J のサブグループごとのObserved FrequencyとExpected Frequencyの「ずれ」を総合的に評価して、全体としての「実際の観測頻度」と「モデルから予測された頻度」の分布が異なるかどうかを評価する検定

$$\chi^2 = \sum_{j=1}^{J} \frac{\left(X_j - n_j \hat{P}_j\right)^2}{n_j \hat{P}_j (1 - \hat{P}_j)}$$

帰無仮説が「両者の頻度の分布が同じ」なので、P < 0.05 となれば、「両者の頻度の分布が異なる（較正がうまくいっていない！）」と判定します。

Hosmer-Lemeshow Test:
P = 0.951291

Hosmer and Lemeshow (2000)

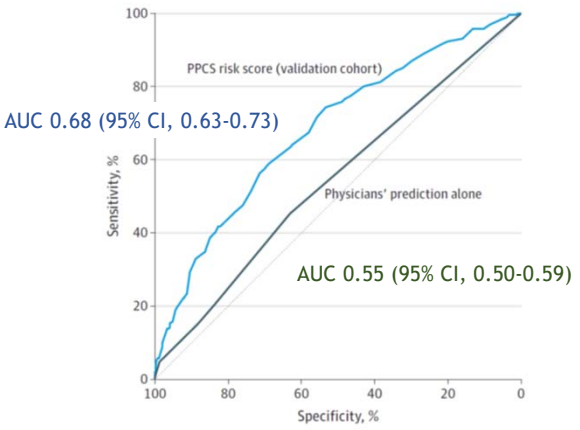
54

注意！Hosmer-Lemeshow検定

- サンプルサイズが大きいときには、Hosmer-Lemeshow統計量は、偽陽性の結果を生むことがある（精度が上がるため）。
- リスクグループをいくつにするかによって結果も変わる。一般的には「１０」に設定することが多いが、「正確」なグループの数を定める理論的根拠はない。また、サンプルサイズが500未満の場合、Hosmer-Lemeshow検定は、検出力が低く、較正の悪いモデルを検出し損なうことがある。

Meurer and Tolles (2017)　55

Figure 2. Receiver Operating Characteristic Curves



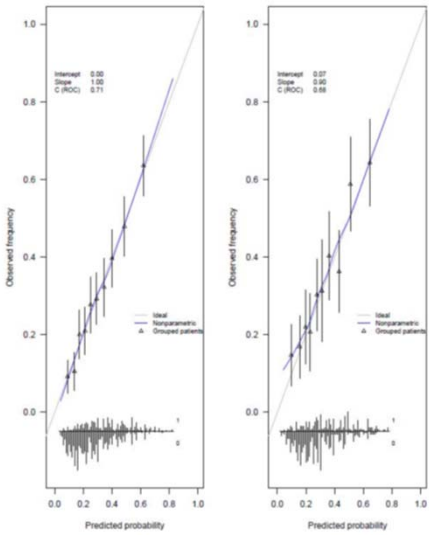
PPCS indicates persistent postconcussive symptoms. The area under the curve was 0.71 (95% CI, 0.69-0.74) for the derivation cohort and 0.68 (95% CI, 0.65-0.72) for the validation cohort.

Zemekらの研究では、予測モデルを作製した Derivation Cohortとは別に、独立な Validation Cohort（N=883）による予測性能の評価を行っている

新たに作製されたリスクスコアは、臨床医の所見によるリスク予測よりも、PPCS発症の予測精度が有意に高かった。

ただし、Validation Cohortは、同じ施設からリクルートを行った集団であり、施設の特性に応じて、似通った背景の患者が対象となっている。外部妥当性については、情報不足である可能性がある。

Zemek et al. (2016)　56



Derivation Cohort　　Validation Cohort

Validation Cohortにおける Hosmer-Lemeshow検定は、モデルの相応の当てはまりを示唆しており (*P* = .50)、較正プロットの切片は 0.07、傾きは 0.90であった。

Zemek et al. (2016)

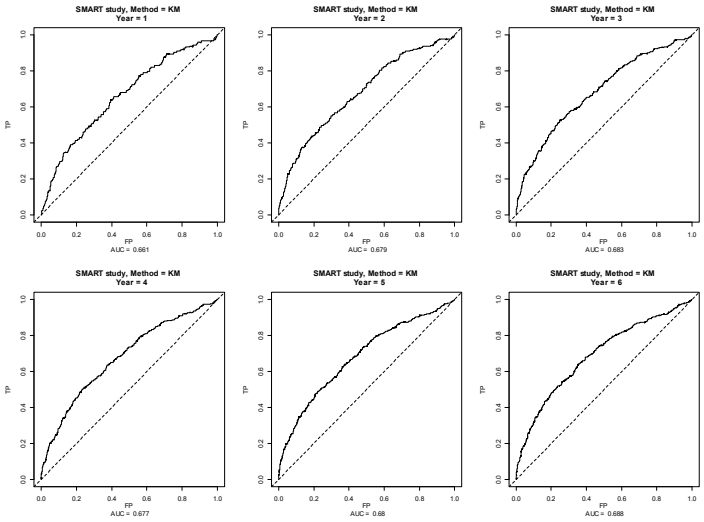
57

Censored Time-to-event Data への拡張

- 累積ハザードに対するCox回帰モデル
  - $H(t) = H_0(t)\exp(\beta_1 x_1 + \cdots + \beta_p x_p)$
  - 最大部分尤度推定量：$\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_p\right)^T$
- 時点 t におけるリスクスコア
  - $\eta_t = 1 - \exp\{-\hat{H}_0(t)\exp(\hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p)\}$
- η_t もしくは線形予測子 $\hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$を予測スコアとした予測方式を用いる

58

Time-dependent ROC curve



「ある時点 t までにイベントを起こすか否か (=0, 1) 」に対して打ち切りを考慮して定義されるROC曲線

図は、Heargerty et al. (2000) のKaplan-Meier式の方法による時間依存性ROC曲線の推定結果（R: survivalROCで実行できる）

59

生存時間解析における C 統計量

- Harrell's Concordance
  - $C_H = \Pr(\boldsymbol{\beta}^T \boldsymbol{x}_1 > \boldsymbol{\beta}^T \boldsymbol{x}_2 | T_1 < T_2, T_1 < \min(C_1, C_2))$
  - $\boldsymbol{x}_1, \boldsymbol{x}_2$：集団内のペアにおける共変量ベクトル，
    T_1, T_2：イベントまでの時間，C_1, C_2：打ち切りまでの時間
- Uno's Concordance
  - $C_U = \Pr(\boldsymbol{\beta}^T \boldsymbol{x}_1 > \boldsymbol{\beta}^T \boldsymbol{x}_2 | T_1 < T_2, T_1 < \tau)$
  - τ：a specified time point within the support of the censoring variable

Harrell (1996), Uno et al. (2011)

60

Harrell の C 統計量

```
Call:
coxph(formula = Surv1 ~ AGE + BMI + HDL + DIABETES + CREAT +
    albumin + STENOSIS + IMT)

  n= 3444, number of events= 395

          exp(coef) exp(-coef) lower .95 upper .95
AGE         1.0439     0.9579    1.0319    1.0560
BMI         0.9740     1.0267    0.9453    1.0035
HDL         0.6168     1.6212    0.4506    0.8444
DIABETES    1.1494     0.8700    0.9113    1.4498
CREAT       1.0019     0.9981    1.0009    1.0030
albumin     1.4543     0.6876    1.2149    1.7409
STENOSIS    1.2545     0.7971    1.0053    1.5655
IMT         1.6329     0.6124    1.2282    2.1711

Concordance= 0.676  (se = 0.016 )
Rsquare= 0.053   (max possible= 0.822 )
Likelihood ratio test= 188.3  on 8 df,   p=0
Wald test            = 197.5  on 8 df,   p=0
Score (logrank) test = 217.1  on 8 df,   p=0
```

Harrell's C-Statistic

R では coxph の summary の出力に、C統計量も含まれています

UnoのC統計量は、R: survC1で計算可能

61

R example codes

- 以下のURLから、R の事例コードをダウンロードすることできます
  - http://www.ism.ac.jp/~noma/JBS2018_logistic.r
  - http://www.ism.ac.jp/~noma/JBS2018_coxph.r

62

文献

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory*, B. N. Petrov, and F. Csaki (eds), 267-281. Budapest: Akademiai Kiado.
- Ambler, G., Brady, A. R., and Royston, P. (2002). Simplifying a prognostic model: a simulation study based on clinical data. *Statistics in Medicine* **21**, 3803-3822.
- Austin, P. C., and Tu, J. V. (2004). Bootstrap methods for developing predictive models in cardiovascular research. *American Statistician* **58**, 131-137.
- Bagherzadeh-Khiabani, F., Ramezankhani, A., Azizi, F., Hadaegh, F., Steyerberg, E. W., and Khalili, D. (2016). A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results. *Journal of Clinical Epidemiology* **71**, 76-85.
- Breiman, L. (1996). Bagging predictors. *Machine Learning* **24**, 123-140.
- Chen, C. H., and George, S. L. (1985). The bootstrap and identification of prognostic factors via Cox's proportional hazards regression model. *Statistics in Medicine* **4**, 39-46.

63

- DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837-845.
- Derksen, S., and Keselman, H. (1992). Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology* **45**, 265-282.
- Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association* **78**, 316-331.
- Efron, B., and Tibshirani, R. (1997). Improvements on cross-validation: the .632+ bootstrap method. *Journal of the American Statistical Association* **92**, 548-560.
- Greenland, S. (2008). Invited commentary: variable selection versus shrinkage in the control of multiple confounders. *American Journal of Epidemiology* **167**, 523-529; discussion 530-521.
- Harrell, F. E., Jr., Lee, K. L., and Mark, D. B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* **15**, 361-387.

64

- Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd edition. New York: Springer.
- Heagerty, P. J., Lumley, T., and Pepe, M. S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56**, 337-344.
- Hosmer, D. W., and Lemeshow, S. (2000). *Applied Logistic Regression*, 2nd edition. New York: Wiley.
- Mallows, C. L. (1973). Some comments on C_p. *Technometrics* **15**, 661-675.
- McLachlan, G. J. (2004). *Discriminant Analysis and Statistical Pattern Recognition*. Hoboken, NJ: John Wiley & Sons.
- Meurer, W. J., and Tolles, J. (2017). Logistic regression diagnostics: Understanding how well a model predicts outcomes. *JAMA* **317**, 1068-1069.
- Raftery, A. E., Madigan, D., and Hoeting, J. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* **92**, 179-191.
- Sauerbrei, W., and Schumacher, M. (1992). A bootstrap resampling procedure for model building: application to the Cox regression model. *Statistics in Medicine* **11**, 2093-2109.

65

- Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning* **5**, 197-227.
- Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**, 461-464.
- Steyerberg, E. W. (2009). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. New York: Springer.
- Steyerberg, E. W., Eijkemans, M. J., and Habbema, J. D. (1999). Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *Journal of Clinical Epidemiology* **52**, 935-942.
- Steyerberg, E. W., Eijkemans, M. J., and Habbema, J. D. (2001a). Application of shrinkage techniques in logistic regression analysis: a case study. *Statistica Neerlandica* **55**, 76-88.
- Steyerberg, E. W., Eijkemans, M. J., Harrell, F. E., Jr., and Habbema, J. D. (2000). Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Statistics in Medicine* **19**, 1059-1079.

66

- Steyerberg, E. W., Harrell, F. E., Jr., Borsboom, G. J., Eijkemans, M. J., Vergouwe, Y., and Habbema, J. D. (2001b). Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology* **54**, 774-781.
- Sullivan, L. M., Massaro, J. M., and D'Agostino, R. B., Sr. (2004). Presentation of multivariate data for clinical use: The Framingham Study risk score functions. *Statistics in Medicine* **23**, 1631-1660.
- Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L. J. (2011). On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine* **30**, 1105-1117.
- Uno, H., Cai, T., Tian, L., and Wei, L. J. (2007). Evaluating prediction rules for *t*-year survivors with censored regression models. *Journal of the American Statistical Association* **102**, 527-537.
- Vach, W., Sauerbrei, W., and Schumacher, M. (2001). Variable selection and shrinkage: comparison of some approaches. *Statistica Neerlandica* **55**, 53-75.

67

- Walter, S., and Tiemeier, H. (2009). Variable selection: current practice in epidemiological studies. *European Journal of Epidemiology* **24**, 733-736.
- Wang, D., Lertsithichai, P., Nanchahal, K., and Yousufuddin, M. (2003). Risk factors of coronary heart disease: a Bayesian model averaging approach. *Journal of Applied Statistics* **30**, 813-826.
- Wilson, P. W., D'Agostino, R. B., Levy, D., Belanger, A. M., Silbershatz, H., and Kannel, W. B. (1998). Prediction of coronary heart disease using risk factor categories. *Circulation* **97**, 1837-1847.
- Zemek, R., Barrowman, N., Freedman, S. B., *et al.* (2016). Clinical risk score for persistent postconcussion symptoms among children with acute concussion in the ED. *JAMA* **315**, 1014-1025.

68

BACK UP

69

診断検査の正確性に関する統計

	True Status	
	Has the disease	Does not have the disease
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

70

診断の正確性の指標

- 感度（Sensitivity）
  - 疾患を有する患者が正しく陽性に判定される確率
  - Se = TP / (TP + FN)
- 特異度（Specificity）
  - 疾患を有していない患者が正しく陰性に判定される確率
  - Sp = TN / (FP + TN)

71

陽性的中率・陰性的中率

- 陽性的中率（Positive Predictive Value）
  - 陽性と判定された人が実際に疾患を有する確率
  - PPV = TP / (TP + FP)
- 陰性的中率（Negative Predictive Value）
  - 陰性と判定された人が実際に疾患を有さない確率
  - NPV = TN / (FN + TN)

72

診断精度

- 診断精度（Diagnostic Accuracy）
  - 感度・特異度を併せた、全体としての正判別率
  - DA = (TP + TN) / (TP + FP + TN + FN)

73

陽性尤度比

- 陽性尤度比（Positive likelihood ratio）
  - 「有病者が陽性と判定される確率」が「非有病者が陽性と判定される確率」の何倍になるかを表す指標
  - $LR_+ = Se/(1 - Sp) = \Pr(T_+|D_+)/\Pr(T_+|D_-)$
- 診断法の「陽性」の判定結果の起こりやすさの違いを、相対的に測る指標
  - １であれば有病者・非有病者ともに同程度、１より大きければ有病者の検出率が高く、１より小さければ逆に非有病者の偽陽性率のほうが高い

74

陰性尤度比

- 陰性尤度比（Negative likelihood ratio）
  - 「非有病者が陰性と判定される確率」が「有病者が陰性と判定される確率」の何倍になるかを表す指標
  - $LR_- = (1 - Se)/Sp = \Pr(T_-|D_+)/\Pr(T_-|D_-)$
- 診断法の「陰性」の判定結果の起こりやすさの違いを、相対的に測る指標
  - １であれば有病者・非有病者ともに同程度、１より大きければ有病者の偽陰性率が高く、１より小さければ逆に非有病者の正判別率のほうが高い

75

尤度比の解釈：ベイズの定理

- 尤度比は、診断前の “Prior Odds” が、診断を受けることによって、診断後の “Posterior Odds” に、何倍変わるかを表す指標になる（有病率には依存しない指標）

$$\frac{\Pr(D_+|T_+)}{\Pr(D_-|T_+)} = \frac{\Pr(T_+|D_+)}{\Pr(T_+|D_-)} \times \frac{\Pr(D_+)}{\Pr(D_-)}$$

$$\frac{\Pr(D_+|T_-)}{\Pr(D_-|T_-)} = \frac{\Pr(T_-|D_+)}{\Pr(T_-|D_-)} \times \frac{\Pr(D_+)}{\Pr(D_-)}$$

Posterior Odds　　Prior Odds

76

陽性・陰性尤度比の解釈

Estimation Table [edit]

This table provide examples of how changes in the likelihood ratio affects post-test probability of disease.

Likelihood Ratio	Approximate* Change in Probability[11]	Effect on Posttest Probability of disease[12]
Values between 0 and 1 *decrease* the probability of disease (-LR)		
0.1	- 45%	Large decrease
0.2	- 30%	Moderate decrease
0.5	- 15%	Slight decrease
1	- 0%	None
Values greater than 1 *increase* the probability of disease (+LR)		
1	+ 0%	None
2	+ 15%	Slight increase
5	+ 30%	Moderate increase
10	+ 45%	Large increase

https://en.wikipedia.org/wiki/Likelihood_ratios_in_diagnostic_testing

77