

2021/1/20 医療統計学（3年）①

因果と予測



北海道大学 医学統計学
横田 勲

1

今回の内容

2

- ▶ 因果モデルと予測モデル
 - ▶ 交絡
 - ▶ 結果の結果での条件付け
- ▶ 回帰モデルにおける変数選択
- ▶ 性能評価
 - ▶ 感度・特異度、ROC曲線

2

医学研究での目的

3

- ▶ "X" は "疾病Y" と 関連 がある
 - ▶ X: 健康状態マーカーや
疾病Yを引き起こす疾患など



- ▶ "X" は "疾病Y" の 原因 となる
- ▶ "X" は "疾病Y" を 予測 する

より目的を明確に

3

因果と予測

4

- ▶ 回帰分析から、X-Y間の「関連」を検討
- ▶ Xが原因となり、Yという結果が導かれる
 - ▶ 回帰モデルは因果モデル ('do' model)
 - ▶ 交絡因子は制御すべきもの
- ▶ Xの値を与えて、Yという結果を当てる
 - ▶ 回帰モデルは予測モデル ('see' model)
 - ▶ 予測精度を高めるためにXを選ぶ

Allison PD. 1998(Book).
vanHouwelingen JC. The President's speech in ISCB34.

4

前立腺がんとPSA

6

- ▶ 前立腺がんの発見・病勢と強い関連
 - ▶ スクリーニングにも用いられる
- ▶ がんの細胞壁が壊れやすいため、
がんのvolumeに応じてPSAが血液中に漏出

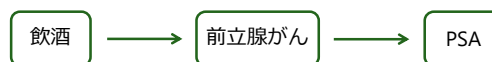


6

前立腺がんのリスク因子を検討

7

- ▶ 明らかなリスク因子は、年齢、家族歴
- ▶ 他にもリスク因子はあるに違いない！
 - ▶ 例えば、飲酒の影響を調べてみる
 - ▶ 因果関係を知りたい



7

交絡の影響を解析で除去

8

- ▶ 以下の条件を満たすことで、飲酒と前立腺がんの関係を歪めてしまう
 - ▶ 年齢が高いほど前立腺がんは増える
 - ▶ 年齢と飲酒には関係がある
 - ▶ 飲酒をすれば年齢が増えるわけではない

```

    graph LR
      Age[年齢] --> Alcohol[飲酒]
      Age --> ProstateCancer[前立腺がん]
      Alcohol --> ProstateCancer
      ProstateCancer --> PSA[PSA]
    
```

8

交絡を防ぐには

9

理想の対照
曝露群が
曝露を受けなかった場合の
結果

異なれば
バイアス

現実の対照
実際に
曝露を受けなかった群の
結果

曝露群で観察される結果

なんとかして理想的な対照を作りたい

9

個人では基本的にムリ

10

- ▶ 曝露を受けた場合の結果を知った後に曝露を受けさせないことはできない
- ▶ 集団で議論することはできないだろうか

10

ランダム化 randomisation

11

ランダムに対象者を
曝露群と非曝露群に割付

実際に観察される内容

曝露を受ける

曝露を受けない

半分の人の結果で代用

曝露群で観察される結果

理想的には知りたい内容

曝露群が
曝露を受けなかった場合の
結果

11

脚気論争

松田誠, 慈恵医大誌, 2002, 2003. 12

- ▶ ビタミンB1欠乏症
 - ▶ 白米ばかりの食生活
 - ▶ 大正時代は結核と並ぶ二大国民病
- ▶ 高木兼寛 (1849-1920)
 - ▶ William Willisよりイギリス医学を学ぶ
 - ▶ 海軍軍医になり、イギリス留学中(1875-80)疫学の方法論を学ぶ
 - ▶ 帰国後、脚気調査に

12

栄養欠陥説

13

- ▶ 龍驤艦 明治15(1882)~16年
 - ▶ 378名中169名が脚気に罹患、25名死亡
 - ▶ 外国で停泊中は脚気患者が減る
 - ▶ 寄港したハワイで食料を全部入れ替えて以降、誰も罹患しなかった
 - ▶ タンパク質を多くすれば予防できるのでは
- ▶ 筑波艦 明治17年
 - ▶ 食事を洋食に変え、龍驤艦と同一航路
 - ▶ 333名中14名が罹患、死亡者ゼロ
 - ▶ 14名のうち12名は洋食を拒んだ

13

海軍は麦食に

14

▶ 洋食はウケが悪かった

年次	食事	兵員数	脚気患者数	死亡者数
明治11(1878)年	米食	4,528	1,485	32
明治12年	米食	5,030	1,978	57
明治13年	米食	4,956	1,725	27
明治14年	米食	4,641	1,163	30
明治15年	米食	4,769	1,929	51
明治16年	米食	5,346	1,236	49
明治17年	洋食	5,638	718	8
明治18年	麦食	6,918	41	0
明治19年	麦食	8,475	3	0
明治20年	麦食	9,016	0	0
明治21年	麦食	9,184	0	0

14

栄養欠陥説への批判

15

- ▶ 緒方正規「脚気病菌発見」(1885)
 - ▶ 北里柴三郎により実験不備を指摘され消える
- ▶ 大沢謙二「食物消化の試験」(1887)
 - ▶ 栄養吸収は米のタンパクのほうがよい
- ▶ 森林太郎「統計二就テノ分疏」(1889)
 - ▶ 患者の減少時期と麦食への切り替え時期が偶然一致しただけ
 - ▶ スタチスチックは科学でなく方法であり、原因を探り法則を知り得るものではない

15

ランダム化比較試験の示唆

16

- ▶ 一つの兵団を二分して、一方に麦食を、もう一方に米食を与え、両者を同一の地に住まわせ、他の生活条件も同じにすべき
- ▶ もし米食者のみが脚気に罹り、麦食者は罹らなかったら、米食が原因だ

森林太郎「統計二就テノ分疏」(1889)

16

日清戦争から日露戦争

17

- ▶ 日清戦争(1894-95)
 - ▶ 陸軍：40,000+の罹患、4,000+の死亡
 - ▶ 戦死者は約300
 - ▶ 海軍：罹患ゼロ
- ▶ 義和団の乱(1900)
 - ▶ 第五師団20,000+中、2,351の罹患
- ▶ 日露戦争(1904-05)
 - ▶ 陸軍：250,000+の罹患、約28,000の死亡
 - ▶ 戦死者は約47,000
 - ▶ 海軍：若干の罹患、死亡ゼロ

17

疫学的観点から

18

- ▶ 脚気論争
 - ▶ 高木の栄養欠陥説自体は間違っていた
 - ▶ 食事に問題があることは正しかった
 - ▶ 疫学研究によって特定できた
 - ▶ 航海実験から約30年後の1911年、鈴木梅太郎によるビタミンB1発見

18

ランダム化比較試験の歴史

19

Randomised Controll Trial; RCT

- ▶ 1946年よりイギリスで計画・実施された結核薬ストレプトマイシンの試験
 - ▶ 初めてのヒトを対象としたRCT
- ▶ 1962年より米国食品医薬品局ではRCTを2回行い有効性評価(two-trials rule)
 - ▶ 背景にサリドマイド事件
- ▶ 1990年以降Evidence Based Medicine普及により多くのRCTが実施

19

交絡への対処

20

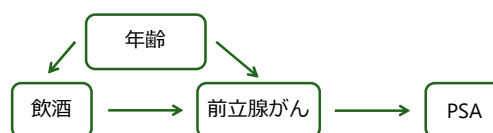
- ▶ デザインで
 - ▶ ランダム化
 - ▶ マッチング
 - ▶ 限定
- ▶ 解析で
 - ▶ 層別解析
 - ▶ 回帰分析
 - ▶ 重み付け解析

20

前立腺がんの予測をしたい

21

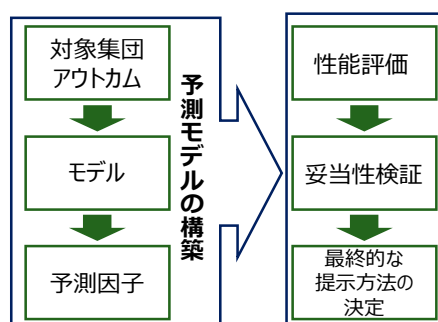
- ▶ 前立腺がん発生を精度良く当てたい
 - ▶ どのような因子を用いてもよい
 - ▶ 年齢のようなリスク因子
 - ▶ 前立腺がんの"結果"であるPSA



21

予測モデル構築の流れ

23



23

DLBCLの新規予後予測モデル

25

- ▶ びまん性大細胞型B細胞リンパ腫
- ▶ 全生存予後を予測したい
- ▶ 臨床で簡単に利用できるスコアを作りたい
 - ▶ 年齢、血清LDH、Ann Arborステージ、ECOG-Performance Status、血清CRP、低アルブミン血症、節外（骨髄、骨、皮膚、肺/胸膜）病変
 - ▶ 変数選択により、予測に用いる因子を決定

25

Time-to-event アウトカム

26

- ▶ 連続量、カテゴリカルのほか、医学研究でよく登場するアウトカム
- ▶ あらかじめ定義した「イベント」が起こるまでの時間
 - ▶ 死亡、再発、入院、ある基準の達成、など
 - ▶ at risk : まだイベントを起こしていない状態

26

打ち切りのあるデータ

27

- ▶ ある時点までイベントを起こしていない
- ▶ その先で起こるはずのイベントの正確な時点が分からない
 - ▶ 脱落や研究終了等による
- ▶ 適切に考慮する解析方法が生存時間解析
 - ▶ 単に除外すると有病率を過大評価しがち
 - ▶ イベントなしとすると有病率を過小評価
 - ▶ 無情報な打ち切りの仮定

27

Kaplan-Meier法①

28

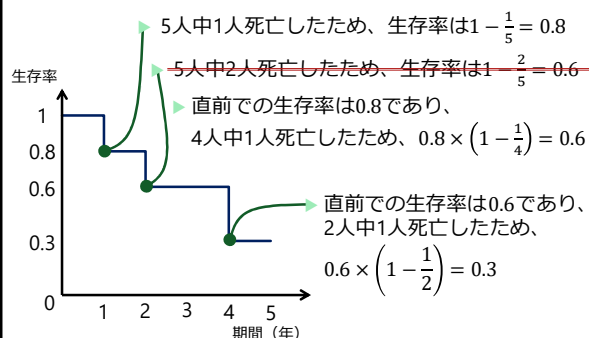
- ▶ 直前までat riskである人について、イベントを起こさなかった確率を乗じる
 - ▶ 生存例は、それまでの間、常に生存してきた
- ▶ 以下のデータセットを想定

イベント発生時点 (年)	内容
1	死亡 (イベント)
2	死亡 (イベント)
3	脱落 (打ち切り)
4	死亡 (イベント)
5	研究終了 (打ち切り)

28

Kaplan-Meier法②

29



29

打ち切り例の扱い

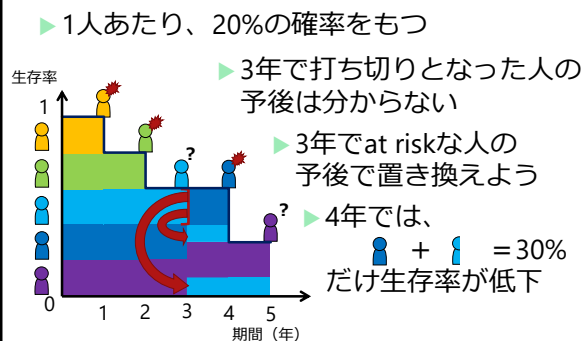
30

- ▶ 3年で打ち切りとなった対象者
 - ▶ 1年、2年での生存率を計算する際には、at riskであった人として解析に寄与
 - ▶ 4年、5年での生存率計算では分母に入らず
 - ▶ 生存率の計算自体には反映されている

30

追跡開始時は5人でスタート

31



31

無情報な打ち切り noninformative censoring

32

- ▶ 代表的な生存時間解析法で置かれる仮定
 - ▶ ランダムな打ち切り、とも
- ▶ 打ち切りとイベント発生が無関係
 - ▶ 研究終了時の生存
 - ▶ 偶然の事故による追跡不能
- ▶ 打ち切り例の予後を、at risk例で置き換えるため

32

Cox回帰を利用

33

Cox DR. JRSS B. 1972

- ▶ ハザード $\lambda(t)$ に対する回帰分析

$$\lambda(t) = \lambda_0(t) \exp(x^T \beta)$$
- ▶ ハザード
 - ▶ ある時点における瞬間的な死亡しやすさ

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{\Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$
 - ▶ T : イベント時点を表す確率変数

33

変数選択

34

- ▶ 回帰分析において、複数の因子候補から、関連の強そうなものだけに絞る方法
 - ▶ 予測モデルをシンプルにするためには便利
- ▶ 予測モデルを作るため
 - ▶ 少ない変数で当たりのよいモデルを
 - ▶ 特別な測定を要する変数で作ったモデルは使われづらい
- ▶ 因果関係を調べるためには使わない
 - ▶ 交絡調整が目的ゆえ、利用可能なすべての変数を用いる

34

ランダム分割

35

- ▶ 465例のデータ
 - ▶ 323例(70%)をトレーニングコホート
 - ▶ 142例(30%)をバリデーションコホート
- ▶ トレーニングコホートで予測モデルを構築
- ▶ バリデーションコホートで他の予測モデルとの性能を比較
 - ▶ モデル構築に用いていないデータであるため、公平な性能比較を行えるだろう

35

最終モデル

36

- ▶ 変数減少ステップワイズ法を利用

因子	ハザード比	95%信頼区間	回帰係数	スコア
LDH ≤ 1×ULN	1	-	0	
LDH > 1×ULN, ≤3×ULN	2.47	1.20-5.08	0.91	1点
LDH > 3×ULN	3.68	1.57-8.66	1.31	2点
ECOG-PS ≥ 2	2.50	1.40-4.45	0.91	1点
ALB < 3.5mg/dL	2.52	1.36-4.69	0.93	1点
特定部位への節外病変	1.71	1.03-2.84	0.54	1点

- ▶ 合計点を基にさらにリスク分類

合計点	0点	1-2点	3点	4-5点
リスク分類	低	低中間	高中間	高

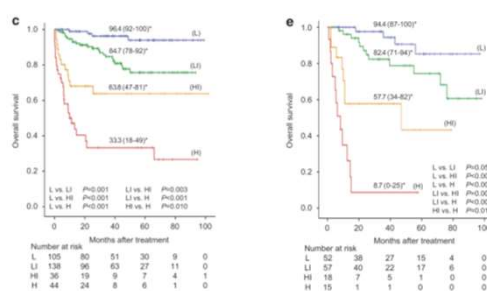
36

各コホートでのリスク分類

37

トレーニング

バリデーション



37

モデルの複雑さとoverfitting

38

- ▶ モデルを複雑にするとより細かな違いまで捉えられる
 - ▶ 関数形を高次にする
 - ▶ 説明変数を増やす
 - ▶ それが無意味な説明変数であっても!
- ▶ 無意味な説明変数をモデル化すると、他のデータへのあてはまりが悪くなる
 - ▶ モデルを活用する際に使いづらい
 - ▶ Overfitting (過適合) という

38

予測性能指標による評価

39

- ▶ 予後の悪い対象者を特定するための予測モデルがどれだけ有用かを知りたい
- ▶ 他の予測モデルと比較したい
- ▶ 予測モデルを構築する上で、overfittingを避けたい

ところで「予測性能がよい」とはどういうこと？

39

モデルのよさ、精度の測り方

40

- ▶ モデルのあてはまり
 - ▶ 決定係数 R^2
 - ▶ 尤度とAIC
- ▶ 予測精度、予測結果のよさ
 - ▶ 平均二乗誤差、Brierスコア
 - ▶ ROC曲線、c 統計量

40

確率(密度)関数と尤度

41

- ▶ データは確率変数の実現値
 - ▶ 確率分布 (パラメータ β を持つモデル) を仮定すれば、当該データが得られる確からしさを定義
 - ▶ $f(x_1, \dots, x_n; \beta) = \prod_i^n f(x_i; \beta)$
- ▶ 尤度 $L(\beta; x)$
 - ▶ $f(x; \beta)$ を β の関数としてみたもの

41

Akaike's Information Criterion; AIC

42

- ▶ 尤度を基にしたあてはまりの指標
 - ▶ 様々なタイプのアウトカムに対応
- ▶ $-2 \log L(\beta; x) + 2K$
 - ▶ K : β のパラメータ数
- ▶ AICが最小となるモデルがよいモデル
 - ▶ パラメータを増やすことへのペナルティを与えた指標
 - ▶ 自由度調整済み決定係数も同様

42

古典的な変数選択法

43

- ▶ 基準に至るまで以下の操作を繰り返す
 - ▶ 変数増加法
 - ▶ 変数候補から最もp値の小さなものを加える
 - ▶ 変数減少法
 - ▶ 変数候補をすべて含めたモデルから最もp値の大きな変数を除く
 - ▶ ステップワイズ法
 - ▶ 変数増加法と減少法を同時に行う
- ▶ 総当り法
 - ▶ 変数の組み合わせを全パターン調べる

43

平均二乗誤差 Mean Squared Error

44

- ▶ $\frac{1}{n} (\hat{y}_i - y_i)^2$
 - ▶ 予測値と実測値の差を評価
 - ▶ 平方根をとって、Root MSE; RMSE

44

Brier スコア

45

- ▶ イベント有無と死亡確率のズレ
 - ▶ 生存時間アウトカムの場合、ある時点 t でのイベント有無と確率のズレ
 - ▶ $\{I(y = 1) - \hat{y}\}^2$
 - ▶ $\{I(T > t) - \hat{S}(t|\mathbf{X})\}^2$
 - ▶ $I(\cdot)$: かつこ内が真のときに1、それ以外は0
 - ▶ T : イベント発生時点
 - ▶ $\hat{S}(\cdot)$: 生存関数の予測値

45

平均Brierスコアの数值例

46

▶ 2人死亡、2人生存という仮想例

▶ 無情報モデル

ID	生存/死亡	予測確率	Brierスコア
1	死亡	0.5	$(1 - 0.5)^2 = 0.25$
2	死亡	0.5	$(1 - 0.5)^2 = 0.25$
3	生存	0.5	$(0 - 0.5)^2 = 0.25$
4	生存	0.5	$(0 - 0.5)^2 = 0.25$

平均Brierスコア
0.25

▶ 予測モデル

ID	生存/死亡	予測確率	Brierスコア
1	死亡	0.9	$(1 - 0.9)^2 = 0.01$
2	死亡	0.6	$(1 - 0.6)^2 = 0.16$
3	生存	0.3	$(0 - 0.3)^2 = 0.09$
4	生存	0.2	$(0 - 0.2)^2 = 0.04$

平均Brierスコア
0.075

46

相対Brierスコア減少

47

▶ 期待Brierスコアのとりうる範囲は0から0.25

▶ しかも0に近いほど「予測性能がよい」

▶ 集団全体の生存確率によって、上限が変化

▶ 無情報モデルに対する、予測モデルでの期待Brierスコアを小さくした割合

▶ 0から1をとり、1に近いほど「予測性能がよい」

$$\frac{\text{Brier}_{\text{無情報モデル}} - \text{Brier}_{\text{予測モデル}}}{\text{Brier}_{\text{無情報モデル}}}$$

47

感度と特異度

48

至適基準	評価値	
	陽性	陰性
陽性	a	b
陰性	c	d

▶ 感度： $\frac{a}{a+b}$

▶ 本当に陽性であるものを陽性といえたか

▶ 特異度： $\frac{d}{c+d}$

▶ 本当に陰性であるものを陰性といえたか

48

陽性的中度、陰性的中度との違い

49

至適基準	評価値	
	陽性	陰性
陽性	a	b
陰性	c	d

▶ 陽性的中度： $\frac{a}{a+c}$ 、陰性的中度： $\frac{d}{b+d}$

▶ 評価した結果が本当はどうであったか？

▶ 真の陽性、陰性者の分布によって変わってしまう指標

49

陽性尤度比

50

▶ 本当は陰性である人より本当に陽性の人、何倍陽性と判定されやすいか

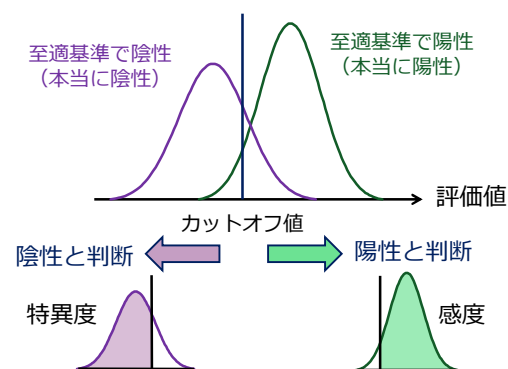
▶ 感度を偽陽性割合で除す

$$\frac{\frac{a}{a+b}}{\frac{c}{c+d}}$$

50

カットオフ値をもって評価

51



51

感度と特異度はトレードオフ

52

- ▶ 感度を上げれば、特異度は下がる
- ▶ 特異度を上げれば、感度は下がる

52

ROC曲線

53

- ▶ 縦軸に感度、横軸に偽陽性率 (1-特異度)
- ▶ カットオフ値を全範囲で動かした場合の感度と偽陽性率をプロット

53

(ROC-)AUC

54

- ▶ ROC曲線の要約指標
 - ▶ 判別能力を表す指標として解釈
- ▶ AUC自体はモデルに依らずに計算される
 - ▶ AUC=0.5であれば、no discriminative ability
 - ▶ AUCが1に近づくほど、判別能力がよい
 - ▶ 絶対的な解釈は困難
- ▶ c (concordance) indexとも呼ばれる

54

c-index

55

$$\frac{\sum_{i=1}^n \sum_{j=1}^m \{I(\hat{p}_i > \hat{p}_j) + 0.5 \times I(\hat{p}_i = \hat{p}_j)\}}{nm}$$

- ▶ \hat{p}_i : 実際にイベント発生した*i*のイベント確率の推定値
- ▶ \hat{p}_j : 実際にイベント発生しなかった*j*のイベント確率の推定値
- ▶ イベント有無と予測変数の
 - ▶ 大小関係が一致していれば1点
 - ▶ 大小関係が一致していなければ0点
 - ▶ 値が等しければ0.5点 (引き分け)

55

c-indexの数値例①

56

- ▶ 死亡例の予測確率: A:0.9, B:0.7, C:0.4
- ▶ 生存例の予測確率: D:0.6, E:0.3, F:0.1
- ▶ 総当たり表

		生存例			c-indexは 8/9=0.89
		0.6	0.3	0.1	
死亡例	0.4	×	○	○	
	0.7	○	○	○	
	0.9	○	○	○	

○: 一致 ×: 不一致

56

c-indexの数値例②

57

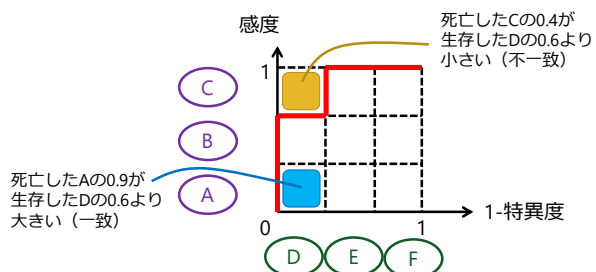
- ▶ 死亡例の予測確率: A:0.9, B:0.7, C:0.4
- ▶ 生存例の予測確率: D:0.6, E:0.3, F:0.1

57

c-indexの数値例③

58

- ▶ 死亡例の予測確率：A:0.9, B:0.7, C:0.4
- ▶ 生存例の予測確率：D:0.6, E:0.3, F:0.1



58

DLBCL予測モデル研究

59

	PFS		OS	
	c-index	RBSR	c-index	RBSR
R-IPI	0.668	0.122	0.642	0.135
NCCN-IPI	0.749	0.172	0.736	0.251
提案スコア(4段階)	0.703	0.183	0.740	0.305
元の0-5点スコア	0.711	0.215	0.754	0.356

RBSR：相対Brierスコア減少

- ▶ 提案スコアが従来スコアより概ね性能がよいことを示した

59

まとめ

60

- ▶ 因果モデル
 - ▶ 変数間の因果関係を仮定して、適切な条件付けを考える
- ▶ 予測モデル
 - ▶ Overfittingを防ぎながら予測性能のよいモデルを選択
- ▶ ROC曲線
 - ▶ 古典的に使われる判別性能の評価方法

60