

統計モデルの作り方

北海道大学 医学統計学
横田 勲

1

今回の内容

- ▶ 因果モデル
- ▶ 予測モデル
 - ▶ モデルのあてはまりの指標
 - ▶ 診断法の統計学
 - ▶ ROC解析

2

医学研究での目的

- ▶ "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

前立腺がんとPSA

- ▶ 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

前立腺がん → PSA

5

前立腺がんのリスク因子を検討

- ▶ 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

飲酒 → 前立腺がん → PSA

6

交絡の影響を解析で除去

- ▶ 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

年齢 → 飲酒 → 前立腺がん → PSA

7

因果推論に必要なもの

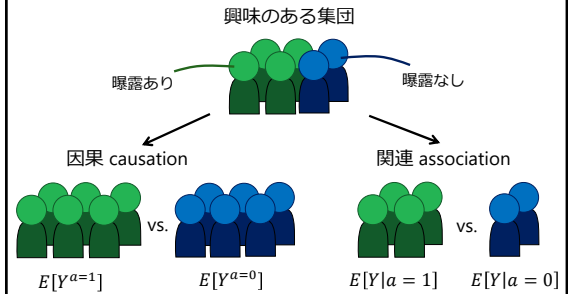
8

- ▶ 因果ネットワークに関する専門家の意見と検証不能な仮定
- ▶ 因果ダイアグラム causal diagram
 - ▶ 因果関係を仮定、図示化
 - ▶ 生じるバイアスを整理
 - ▶ 交絡バイアス、選択バイアス、情報バイアス
 - ▶ 因果効果の分離
 - ▶ 直接効果・間接効果

8

Association is not causation

9



Hernán MA, Robins JM (2020). Causal Inference: What if. Chapman & Hall/CRC. を基に作成

9

交絡 confounding

10

- ▶ 実際の曝露群での結果と集団全体が曝露した場合が違う
- and / or
- ▶ 実際の非曝露群での結果と集団全体が曝露しなかった場合が違う

10

潜在アウトカム potential outcome

11

- ▶ $\gamma^{a=1}$
 - ▶ 曝露 $a = 1$ を受けた場合のアウトカム
- ▶ $\gamma^{a=0}$
 - ▶ 曝露 $a = 0$ を受けた場合のアウトカム
- ▶ アウトカムも2値(0,1)の場合

	$\gamma^{a=1}$	$\gamma^{a=0}$
Doomed	1	1
Helped	1	0
Hurt	0	1
Immune	0	0

11

興味のある集団

12

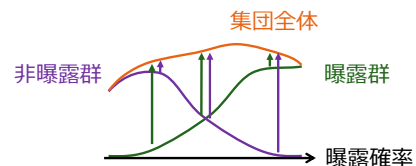
- ▶ 研究対象集団が全員曝露を受けた場合と全員曝露を受けなかった場合
- ▶ 実際に曝露を受けた集団がもし曝露を受けなかった場合
- ▶
- ▶ 潜在アウトカムの分布が異なるはず

12

重み付け解析での擬似集団①

13

- ▶ Average treatment effect
 - ▶ 曝露群では曝露確率の逆数
 - ▶ 非曝露群では非曝露確率の逆数

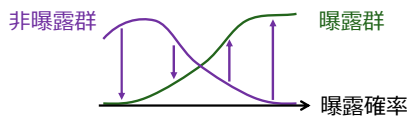


13

重み付け解析での擬似集団②

14

- ▶ Average treatment effect as treated (exposed)
 - ▶ 曝露群はそのまま
 - ▶ 非曝露群では曝露確率のオッズ



14

回帰分析での結果

15

- ▶ 条件付きの効果の大きさ
 - ▶ $E(Y|A = a)$
 - ▶ A の水準が a である場合の Y の期待値
 - ▶ 集団全体で平均化した効果の大きさとは違う

15

観察研究の例

16

▶ 集団全体

治療法	イベントあり	イベントなし	合計
試験治療	22 (2.2%)	978	1000
標準治療	20 (2.5%)	780	800

▶ 75歳以上

治療法	イベントあり	イベントなし	合計
試験治療	18 (3.0%)	582	600
標準治療	6 (6.0%)	94	100

▶ 75歳未満

治療法	イベントあり	イベントなし	合計
試験治療	4 (1.0%)	396	400
標準治療	14 (2.0%)	686	700

16

集団全体での効果

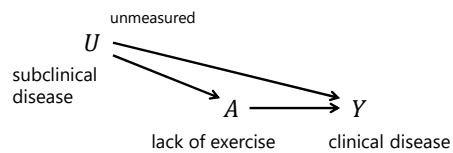
17

- ▶ 興味のある集団は集団全体
 - ▶ 試験治療：
 $(700 \times 3.0\% + 1100 \times 1.0\%) / 1800 = 1.78\%$
 - ▶ 標準治療：
 $(700 \times 6.0\% + 1100 \times 2.0\%) / 1800 = 3.56\%$
- ▶ 興味のある集団は試験治療を受けた集団
 - ▶ 試験治療： $22/1000 = 2.20\%$
 - ▶ 標準治療：
 $(600 \times 6.0\% + 400 \times 2.0\%) / 1000 = 4.40\%$

17

因果DAG ; 交絡の例①

18

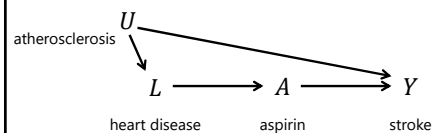


- ▶ 逆因果 reverse causation
- ▶ $A \leftarrow U \rightarrow Y$ のバックドアパスが存在
 - ▶ U をもし観察できれば、条件付けることでバックドアパスをブロック

18

因果DAG ; 交絡の例②

19

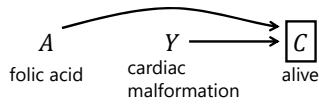


- ▶ 適応による交絡 confounding by indication
- ▶ U か L を条件付ければ $A \leftarrow L \leftarrow U \rightarrow Y$ をブロック
 - ▶ L は U を介して Y に影響するため、経験的同定基準でも交絡因子として定義される

19

因果DAG ; 選択バイアスの例①

20

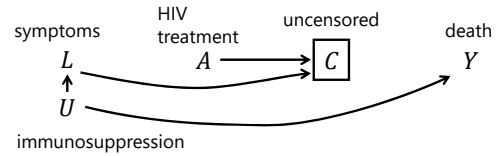


- ▶ 周産期の疫学研究でよくある例
 - ▶ 流産、死産例を無視することで生じる

20

因果DAG ; 選択バイアスの例②

21



- ▶ Cを条件付けると
 - A → C ← L ← U → Y のパスが開く

21

前立腺がんの予測をしたい

22

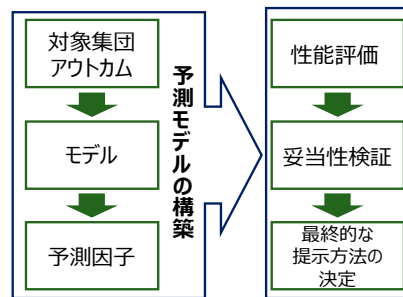
- ▶ 前立腺がん発生を精度良く当てたい
 - ▶ どのような因子を用いてもよい
 - ▶ 年齢のようなリスク因子
 - ▶ 前立腺がんの"結果"であるPSA



22

予測モデル構築の流れ

24



24

DLBCLの新規予後予測モデル

26

- ▶ 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

26

予測性能指標による評価

27

- ▶ 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
 - ▶ ノイズまでモデルをあてはめてしまい、将来の対象者への予測性能が悪くなること

ところで「予測性能がよい」とはどういうこと？

27

モデルのよさ、精度の測り方

28

- ▶ 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

28

決定係数 R^2

29

$$R^2 = 1 - \frac{\text{残差平方和}}{\text{全体の平方和}} = \frac{\text{モデルで説明した平方和}}{\text{全体の平方和}}$$

- ▶ データの持つ全ばらつきのうち、モデルで説明した割合
 - ▶ 単回帰の場合、相関係数の2乗に一致

29

Akaike's Information Criterion; AIC

30

- ▶ $-2 \log L(\boldsymbol{\beta}; x) + 2K$
 - ▶ $\log L(\boldsymbol{\beta}; x)$: 対数尤度関数
 - ▶ パラメータ $\boldsymbol{\beta}$ の確からしさ、あてはまりのよさ
 - ▶ K : $\boldsymbol{\beta}$ 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

30

古典的な変数選択法

31

- ▶ 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
 - ▶ 変数候補から最もp値の小さなものを加え、モデルから最もp値の大きな変数を除く

31

他の変数選択法

32

- ▶ p値の代わりに用いる基準
 - ▶ AIC
 - ▶ 平均二乗誤差、Brierスコア
 - ▶ c-index
 - ▶ . . .
- ▶ 総当たり法
 - ▶ 変数の組合せ全パターン調べる

32

平均二乗誤差 Mean Squared Error

33

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

33

Brier スコア

34

- ▶ イベント有無と生存確率のズレ
 - ▶ 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
- ▶ Brierスコア
 - ▶ $\{I(y = 1) - \hat{y}\}^2$
 - ▶ $\{I(T > t) - \hat{S}(t|X)\}^2$
 - ▶ $I(\cdot)$: かつこ内が真のときに1、それ以外は0
 - ▶ $\hat{S}(\cdot)$: 生存関数の予測値

34

平均Brierスコアの数値例

35

- ▶ 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

35

相対Brierスコア減少

36

- ▶ 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」

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

36

感度と特異度

37

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

37

陽性的中度、陰性的中度との違い

38

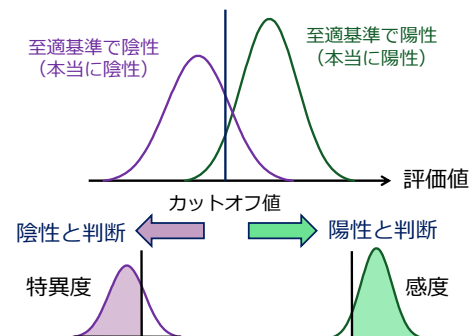
至適基準	評価値	
	陽性	陰性
陽性	a	b
陰性	c	d

- ▶ 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

38

カットオフ値をもって評価

39



39

感度と特異度はトレードオフ

40

- ▶ 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

40

ROC曲線

41

- ▶ 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

AUC (Area Under the Curve)

41

(ROC-)AUC

42

- ▶ 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

42

c-index

43

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

- ▶ \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)

43

c-indexの数値例①

44

- ▶ 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

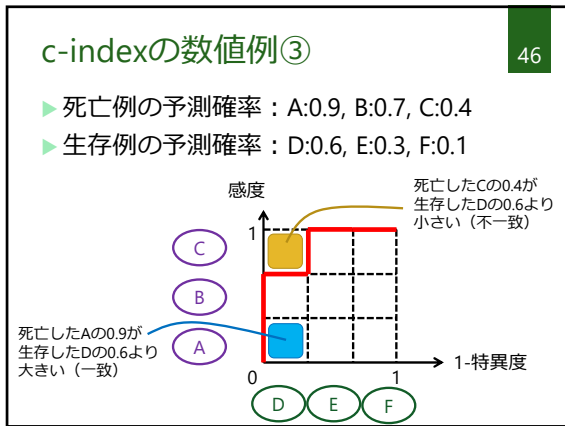
44

c-indexの数値例②

45

- ▶ 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

45



46

DLBCL予測モデル研究

	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

47

- ### まとめ
- ▶ 因果モデル
 - ▶ 潜在アウトカム
 - ▶ 条件付きの効果と集団全体での効果
 - ▶ 因果DAG
 - ▶ 関連から因果、予測へ
 - ▶ 予測モデルの構築のあと、妥当性の検証
 - ▶ 決定係数
 - ▶ 平均二乗誤差、Brierスコア
 - ▶ 感度、特異度、ROC曲線、c-index

48