2020/4/24　北大医学科5年・臨床統合講義

因果と予測

北海道大学　医学統計学
横田　勲

今回の内容 2

- 予測モデル
  - 生存時間アウトカムとCox回帰
  - 感度・特異度、ROC曲線
- 因果モデル
  - 交絡
  - 反事実アウトカムモデル、因果DAG

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

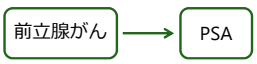
因果と予測 4

- 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.

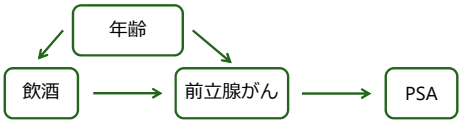
前立腺がんとPSA 6

- 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



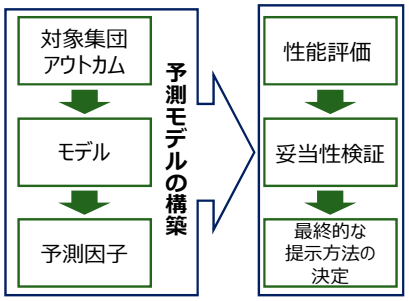
前立腺がんの予測をしたい 7

- 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



予測モデル構築の流れ

9



DLBCLの新規予後予測モデル

11

- 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

Time-to-event アウトカム

12

- 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

打ち切りのあるデータ

13

- 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

Kaplan-Meier法①

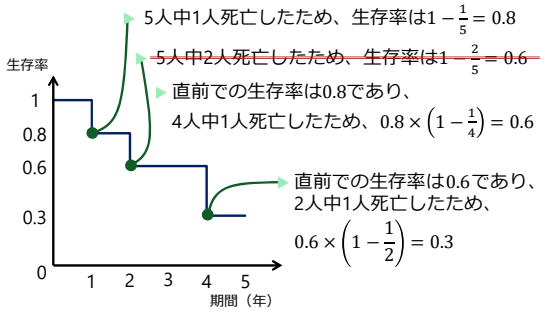
14

- 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）

Kaplan-Meier法②

15

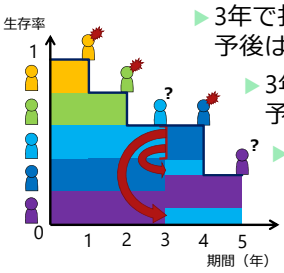


打ち切り例の扱い 16

- 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

追跡開始時は5人でスタート 17

- 1人あたり、20%の確率をもつ



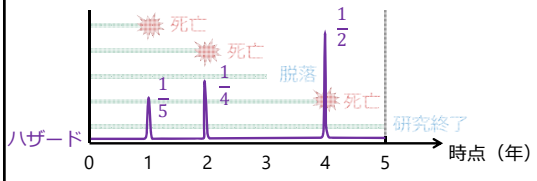
- 3年で打ち切りとなった人の予後は分からない
  - 3年でat riskな人の予後で置き換えよう
  - 4年では、 ＋ ＝30%だけ生存率が低下

無情報な打ち切り noninformative censoring 18

- 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

ハザード hazard 19

- 直前まで生存している下で微小時間あたりのイベント発生
  - 単位は1／時間

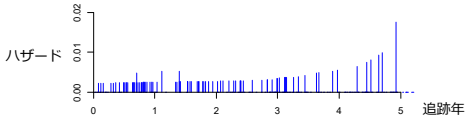


時点とともに変化するハザード 20

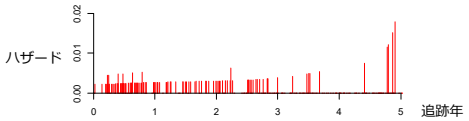
- 前スライドでは、1年時点で1/5、2年時点で1/4、4年時点で1/2、それ以外の時点では0
  - 要約指標には向かない
- 時が経つにつれ、発生しやすさが変化することを柔軟に捉えられる
  - 術後すぐは再発は少ないが、しばらくしてから再発が起こりうる
  - ある期間経過後は再発がまれ（治癒する）

ハザードの比較 21

- 試験群のハザード

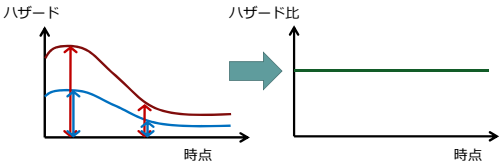


- 対照群のハザード



ハザード比 22

- ハザード自体はとびとびの値をとるので、期間全体を通して、何倍の違いであるか
  - 時点によらずハザード比は一定、という仮定（比例ハザード性）
  - ハザードは時間経過とともに変化してもよい



Cox回帰 23

- ハザード比を推定する回帰分析

パラメータ	自由度	パラメータ推定値	標準誤差	カイ２乗値	Pr > ChiSq	ハザード比	95% ハザード比信頼限界	
alloc	1	-0.35502	0.14017	6.4155	0.0113	0.701	0.533	0.923

対数ハザード比

$e^{-0.355} \approx 0.701$

ログランク検定のp値と同じ
（設定等によってちょっと違うこともある）

変数選択 24

- 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
  - 常に行うべき解析方法ではない！
- 目的やデータの特徴に応じた使い分け
  - 変数増加法 forward
  - 変数減少法 backward
  - ステップワイズ法 stepwise
  - LASSO法、elastic net法

モデル構築のアプローチ 25

Collins GS, et al. *Ann Intern Med*. 2015. TRIPOD guideline

ランダム分割を採用 26

- 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

最終モデル 27

- 変数減少ステップワイズ法を利用

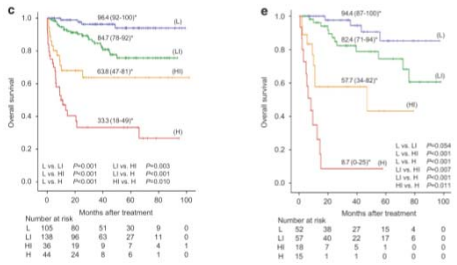
因子	ハザード比	95%信頼区間	回帰係数	スコア
LDH ≤ 1×ULN	1	-	0	
LDH > 1×ULN,≤3×ULN	2.47	1.20-5.08	0.91	1点
LDH > 3×ULN	3.68	1.57-8.66	1.31	2点
ECOG-PS ≥ 2	2.50	1.40-4.45	0.91	1点
ALB < 3.5mg/dL	2.52	1.36-4.69	0.93	1点
特定部位への節外病変	1.71	1.03-2.84	0.54	1点

- 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

各コホートでのリスク分類 28

トレーニング　　バリデーション



予測性能の評価方法 29

- Brierスコア（予測誤差）
- c-index（判別能力）
- キャリブレーション（較正）

} 本日扱う話題

- Net Reclassification Improvement
- Integrated Discrimination Index
- Net Benefit

Brierスコア 30

- 死亡有無と死亡確率の差（ズレ）の2乗
  - 死亡確率：1ー生存確率
  - {(Aさん死亡:1／生存:0) ー (Aさんの死亡確率) }^{2}
- Brierスコアの平均値は0から0.25をとる
  - 0に近づくほどズレが小さいことを表し、よい予測モデル

平均Brierスコアの数値例 31

- 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

相対Brierスコア減少 32

- 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」

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

生存時間解析でのBrierスコア 33

- ある時点までのイベント有無で判定

Graf E, *et al. Stat Med.* 1999. 2529-2545.

- さらに期間全体で平均化した integrated Brierスコア

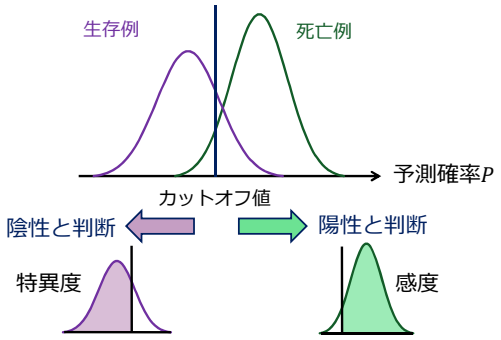
判別 discrimination

34

- アウトカムの異なる対象者を分ける予測
- 予測確率Pを用いて、生存／死亡を診断
  - 感度：死亡例において
    Pがカットオフ値以上（陽性である）
  - 特異度：生存例において
    Pがカットオフ値以下（陰性である）

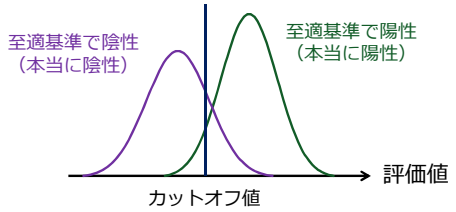
カットオフ値をもって評価

35



感度と特異度はトレードオフ

36

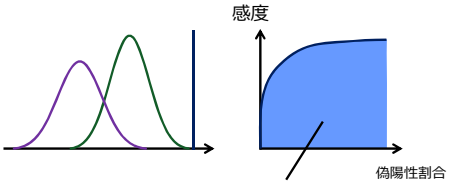


- 感度を上げれば、特異度は下がる
- 特異度を上げれば、感度は下がる

ROC曲線

37

- 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)

(ROC-)AUC

38

- 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、判別能力なし
  - AUCが1に近づくほど、判別能力がよい
  - 絶対値的な解釈は困難

c-index

39

- ROC-AUCは、
  死亡例と生存例を1人ずつ注目した時、
  死亡例の予測確率のほうが
  大きな値をとる確率に同じ
  - 大小関係が揃っていることを「一致」

c-indexの数値例① 40

- 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

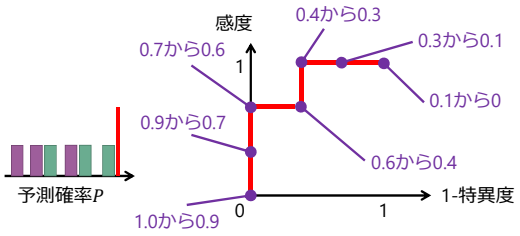
		生存例		
		0.6	0.3	0.1
死亡例	0.4	×	○	○
	0.7	○	○	○
	0.9	○	○	○

○：一致　×：不一致

c-indexは8/9=0.89

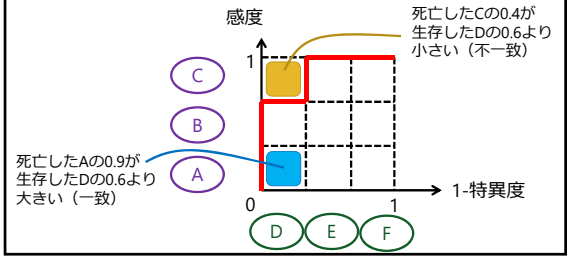
c-indexの数値例② 41

- 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



c-indexの数値例③ 42

- 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



生存時間解析のc-index 43

- 生存時間の短長関係と死亡確率の高低がそろえば一致
- 生存時間が長いのに死亡確率が高い場合は不一致
- 打ち切りを考慮した、Uno's c-indexを利用

Uno H, *et al. Stat Med.* 2011. 1105-1117.

DLBCL予測モデル研究 44

	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

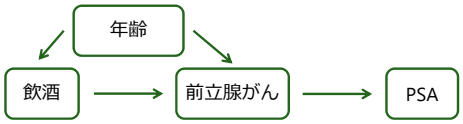
前立腺がんのリスク因子を検討 45

- 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

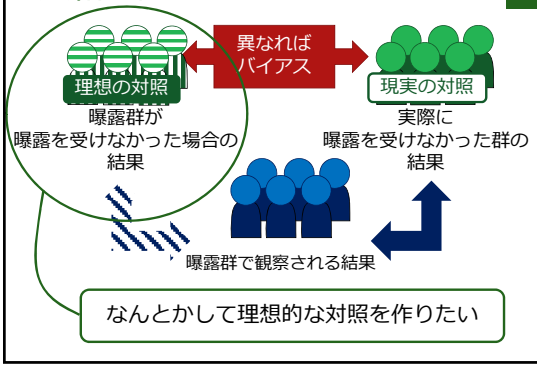
飲酒 → 前立腺がん → PSA

交絡の影響を解析で除去 46

- 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



交絡を防ぐには 47



反事実アウトカム 48

- 事実が観察されたら、観察されないアウトカム
- 事実データ factual data
  - 曝露を受けて（$a = 1$）、死亡した（$Y = 1$）
- 反事実データ counterfactual data
  - 曝露を受けなかったら（$a = 0$）、予後は？（Y =?）

潜在アウトカム potential outcome 49

- $Y^{a=1}$
  - 曝露$a = 1$を受けた場合のアウトカム
- $Y^{a=0}$
  - 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

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

潜在アウトカムと観察アウトカム 50

- 受けた曝露に応じて、潜在アウトカムのいずれかが観察される

	A	$Y^{a=1}$	$Y^{a=0}$	Y
Doomed	1	1	1	1
Helped	1	1	0	1
Hurt	1	0	1	0
Immune	1	0	0	0
Doomed	0	1	1	1
Helped	0	1	0	0
Hurt	0	0	1	1
Immune	0	0	0	0

個人での因果効果 51

	$Y^{a=1}$	$Y^{a=0}$	Causal effect $Y^{a=1} - Y^{a=0}$
Doomed	1	1	$1-1=0$
Helped	1	0	$1-0=1$
Hurt	0	1	$0-1=-1$
Immune	0	0	$0-0=0$

- データとして観察はできない
  - 反事実アウトカムとの比較で定義可能
- Sharp causal null hypothesis
  - Doomed, Immuneな人しかいない

個人では基本的にムリ 52

- 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

平均因果効果 Average Causal Effects 53

- $E[Y^{a=1}] - E[Y^{a=0}]$
  - 集団全員が曝露を受けた場合と集団全員が曝露を受けなかった場合の差
- Null hypothesis of no average causal effect
  - $E[Y^{a=1}] = E[Y^{a=0}]$
  - Sharp causal null hypothesisに加え、Helpedな人とHurtな人が同数いる場合も成立

因果効果の指標 54

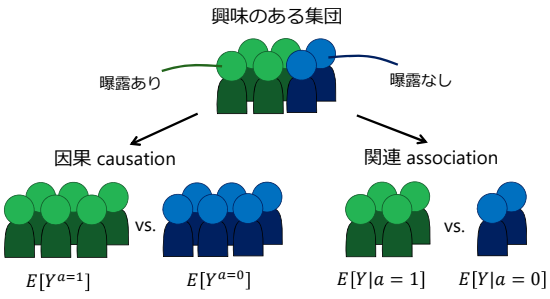
- 因果リスク差
  - $\Pr[Y^{a=1} = 1] - \Pr[Y^{a=0} = 1]$
- 因果リスク比
  - $\frac{\Pr[Y^{a=1}=1]}{\Pr[Y^{a=0}=1]}$
- 因果オッズ比
  - $\frac{\Pr[Y^{a=1}=1]/\Pr[Y^{a=1}=0]}{\Pr[Y^{a=0}=1]/\Pr[Y^{a=0}=0]}$

練習① 以下のACEは？ 55

ID	$Y^{a=1}$	$Y^{a=0}$
1	0	1
2	1	0
3	0	0
4	0	0
5	0	0
6	1	0
7	0	0
8	0	1
9	1	1
10	1	0

ID	$Y^{a=1}$	$Y^{a=0}$
11	0	1
12	1	1
13	1	1
14	0	1
15	0	1
16	0	1
17	1	1
18	1	0
19	1	0
20	1	0

Association is not causation 56



Hernán MA, Robins JM (2019). Causal Inference. Chapman & Hall/CRC, forthcoming. を基に作成

関連効果の指標 57

- 関連リスク差
  - $\Pr[Y = 1|A = 1] - \Pr[Y = 1|A = 0]$
- 関連リスク比
  - $\frac{\Pr[Y=1|A=1]}{\Pr[Y=1|A=0]}$
- 関連オッズ比
  - $\frac{\Pr[Y=1|A=1]/\Pr[Y=0|A=1]}{\Pr[Y=1|A=0]/\Pr[Y=0|A=0]}$

因果効果指標と関連効果指標 58

- 因果効果指標は定義、概念的なもの
  - 反事実アウトカムを用いて定義されるため
- 関連効果指標は観察データから求まる

- 関連効果指標をもって
  因果効果指標を求めるには？
  - どのような条件が成立すれば？
  - どのような解析を行えば？

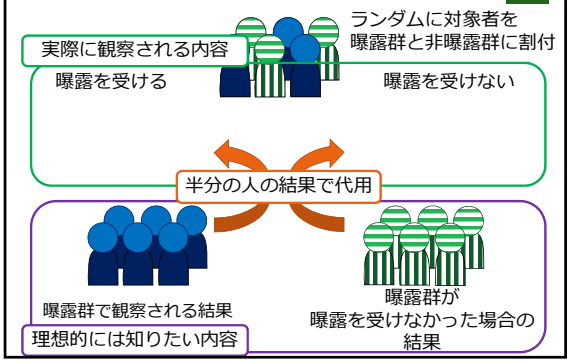
交絡 confounding 59

- 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
  - $E[Y^{a=1}|A=1] \neq E[Y^{a=1}]$

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
  - $E[Y^{a=0}|A=0] \neq E[Y^{a=0}]$

ランダム化 randomisation 60



ランダム化による交換可能性の成立 61

exchangeability

- 曝露群での結果と
  非曝露群が、仮に曝露を受けた場合の結果
  が一致（その逆も）
  - $\Pr[Y^{a=1}|A=1] = \Pr[Y^{a=1}|A=0]$
    $\Pr[Y^{a=0}|A=0] = \Pr[Y^{a=0}|A=1]$
  - $Y^a \perp\!\!\!\perp A$ for all a
- 片方の集団と全体集団での結果と一致
  - $\Pr[Y^{a=1}|A=1] = \Pr[Y^{a=1}|A=0] = \Pr[Y^{a=1}]$
    $\Pr[Y^{a=0}|A=0] = \Pr[Y^{a=0}|A=1] = \Pr[Y^{a=0}]$

交換可能性の意味 62

- 実際の曝露と反事実アウトカムが独立
  - 曝露とアウトカムの関連ナシではない！
- 交絡が生じる状況では交換可能性が不成立
  - 曝露群には実はdoomedな人だらけ
    非曝露群には実はimmuneな人だらけ

条件付き交換可能性 63

- 予後因子Lが同じ値を持つ集団（層内）
  では交換可能性が成立
  - $\Pr[Y^{a=1}|A=1,L=1] = \Pr[Y^{a=1}|A=0,L=1]$
    $\Pr[Y^{a=0}|A=0,L=1] = \Pr[Y^{a=0}|A=1,L=1]$
  - $\Pr[Y^{a=1}|A=1,L=0] = \Pr[Y^{a=1}|A=0,L=0]$
    $\Pr[Y^{a=0}|A=0,L=0] = \Pr[Y^{a=0}|A=1,L=0]$
  - $Y^a \perp\!\!\!\perp A|L$ for all a
- No unmeasured confounding
  - 残差交絡 residual confounding がない

回帰モデルによる効果推定

64

- 例：ロジスティック回帰モデル

$$\log(\text{オッズ}) = \beta_0 + \beta_1 \times A + \beta_2 \times L$$

オッズ	$L=0$	$L=1$
$A=1$	$\exp(\beta_0+\beta_1)$	$\exp(\beta_0+\beta_1+\beta_2)$
$A=0$	$\exp(\beta_0)$	$\exp(\beta_0+\beta_2)$

$L=0$でのオッズ比

$$\frac{\exp(\beta_0+\beta_1)}{\exp(\beta_0)} = \exp(\beta_1)$$

$L=1$でのオッズ比

$$\frac{\exp(\beta_0+\beta_1+\beta_2)}{\exp(\beta_0+\beta_2)} = \exp(\beta_1)$$

練習②－１　練習①データの続き

65

ID	L	A	Y
1	0	0	0
2	0	0	1
3	0	0	0
4	0	0	0
5	0	1	0
6	0	1	0
7	0	1	0
8	0	1	1
9	1	0	1
10	1	0	1

ID	L	A	Y
11	1	0	0
12	1	1	1
13	1	1	1
14	1	1	1
15	1	1	1
16	1	1	1
17	1	1	1
18	1	1	0
19	1	1	0
20	1	1	0

練習②－２

66

- 粗リスク差を求めてみよう
  - Lを無視して、20名全体での
    曝露あり（$A=1$）での発生リスクから
    曝露なし（$A=0$）での発生リスクを引く
- 条件付きリスク差を求めてみよう
  - $L=0$であったグループと
    $L=1$であったグループで
    それぞれリスク差を計算

因果推論に必要なもの

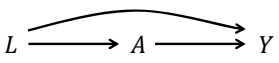
67

- 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
  - 生じうるバイアスを整理
    - 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

有向非循環グラフ

68

- Directed Acyclic Graphs; DAGs
- Directed 有向
  - ノード node 間の矢線 arrow で順序性をいう
  - LがAの原因
- Acyclic 非循環
  - 自分自身の原因となることがない



DAGで出てくる用語

69

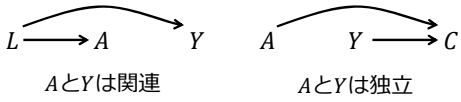
- ノード、節点 node、点 vertex
  - 各変数をノードにおく
- 矢線 arrow、辺 edge
  - 一般に、辺は方向によらず使える言葉
- パス、経路、道 path
  - あるノードから異なるノードまでの行き方
- 親 parent
  - 祖先 ancestor：親の親、その親・・を含める
- 子 child
  - 子孫 descendant：子の子、その子・・を含める

因果DAG 70

- 以下のようなDAG
  - ノード間を直接結ぶ矢線がない場合、直接(因果)効果がない
    - あるかもしれない、なら矢線を示しておく
  - ある変数達に共通する原因は、観察できないとしても、同じグラフ上に示す
  - いかなる変数もその子孫に対し原因となる
- 因果DAGは背景にある反事実モデルを表現

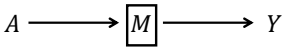
周辺独立 marginally independent 71

- 因果DAGにおける2変数間の特徴
- 以下のいずれかを満たせば"(周辺)関連"
  - 一方がもう一方の原因
  - 共通の原因（親）をもつ
- 関連しない場合、(周辺)独立



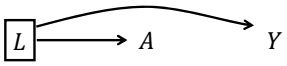
AとYは関連　　AとYは独立

条件付き独立 conditional independence 72



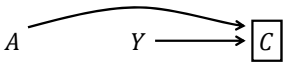
- AとYに周辺関連がある
  - Mは中間変数、媒介変数 mediator
- Mの水準を限定したら？
  - 条件付ける conditional on
  - □で囲う
- Mで条件付けることで、関連のあったパス$A \rightarrow M \rightarrow Y$をブロック
  - 条件付き独立にした

共通原因をBlocked 73



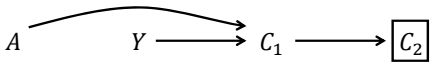
- AとYに周辺関連がある
  - Lが共通原因
- Lを条件付け
- 関連のあったパス$A \leftarrow L \rightarrow Y$をブロック
  - 条件付き独立にした

合流点 collider をブロック 74



- AとYは周辺独立
  - $A \rightarrow C \leftarrow Y$というパスは関連を生まない
  - Cが合流点 collider
- Cを条件付け
- $A \rightarrow C \leftarrow Y$という
  ブロックされていたパスをオープンに
  - 関連が生じる

合流点の子孫をブロック 75



- 合流点のみならず、その子孫についても
  AとYは原因となっていた
- C_2で条件つけても、
  $A \rightarrow C_1 \leftarrow Y$をオープンに
  - 直接の合流点C_1で条件付けることと同様

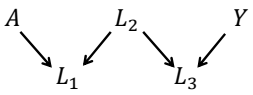
blockedかopenか 76

- パスがblockedな状況は以下のいずれか
  - 非合流点で条件付け
    - 中間変数や共通原因で条件付け
  - 合流点とその子孫は条件付けない
- blockedでないパスがopen path

有向分離 d-separation 77

- 次の条件のいずれかを満たすとき、
  $\{A, Y\}$と排反な変数集合Sが
  $A-Y$間を有向分離する
  - $A-Y$間のすべてのパスにおける合流点で、
    その合流点と子孫がSに含まれないものがある
  - $A-Y$間のすべてのパスに非合流点で、
    Sに含まれるものがある
- Sで条件付ければ、
  $A-Y$間をつなぐパスをすべてblocked
  - open pathが含まれる場合をd-connected

練習③　有向分離するSは？ 78

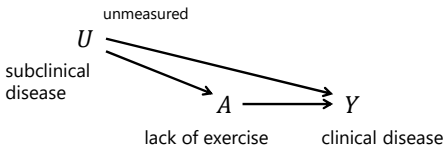


1) $\{\phi\}$　（空集合）
2) $\{L_1\}$
3) $\{L_1, L_2\}$
4) $\{L_1, \quad L_3\}$
5) $\{L_2\}$
6) $\{L_2, L_3\}$
7) $\{L_3\}$
8) $\{L_1, L_2, L_3\}$

バックドア基準 back-door criterion 79

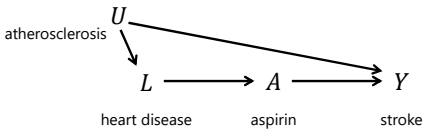
- AはYの非子孫
- 次の2条件を満たす頂点集合Sは
  $A-Y$についてバックドア基準を満たす
  - AからSの任意の要素へ有向道がない
  - Aから出る矢線をすべて除いたグラフにおいて、
    SがAとYを有向分離する
- S, A, Yが観察されていれば、
  AからYへの因果効果は識別可能

交絡の例① 80



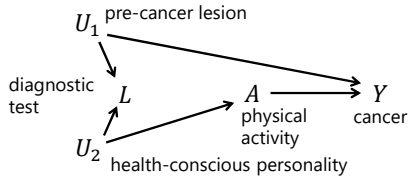
- 逆因果 reverse causation
- $A \leftarrow U \rightarrow Y$のバックドアパスが存在
  - Uをもし観察できれば、条件付けることで
    バックドアパスをブロック

交絡の例② 81



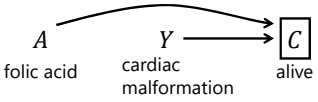
- 適応による交絡 confounding by indication
- UかLを条件付ければ$A \leftarrow L \leftarrow U \rightarrow Y$をブロック
  - LはUを介してYに影響するため、
    経験的同定基準でも交絡因子として定義される

交絡の例③ 82



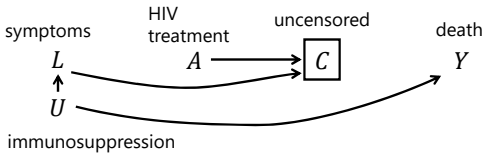
- $A \leftarrow U_2 \rightarrow L \leftarrow U_1 \rightarrow Y$は既にブロック
  - Lが合流点ゆえ
  - しかしLは経験的同定基準では交絡因子
- Lで条件付けるとバックドアパスが開く

選択バイアスの例① 83



- 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

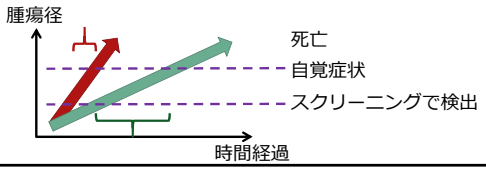
選択バイアスの例② 84



- Cを条件付けると
  $A \rightarrow C \leftarrow L \leftarrow U \rightarrow Y$のパスが開く

Length-time bias 85

- 疾病の経過速度の違いによるバイアス
  - スクリーニングで発見された患者は予後がよい
    - 進行の遅く、予後の良い集団ほど
      スクリーニングにて発見されやすいから
    - 進行の早いがん患者は罹病期間が短いため、
      集団全体では進行の遅いがん患者の割合が増加



練習④　DAGを描いてみよう 86

- がんのスクリーニングを行った際に
  生じうるlength-time biasを表現する
  因果DAGを描け
  - Y：死亡・生存
  - A：がんの悪性度（進行のはやさ）
  - L：スクリーニング

まとめ 87

- 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
- 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