

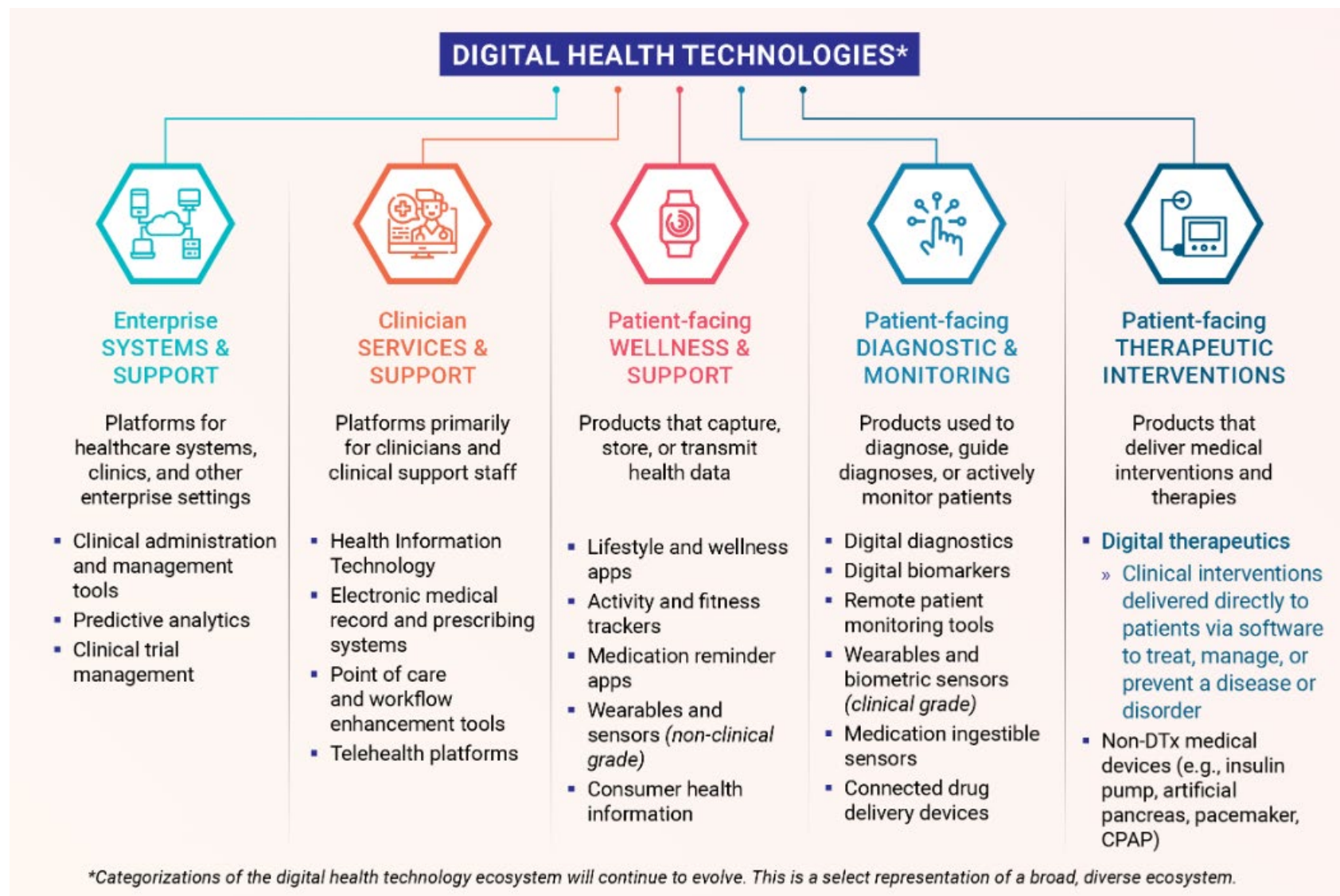
診断能の評価に必要な統計の基礎知識 ～プログラム医療機器開発における注意点～

**研究支援センター 生物統計部
/ 東病院 医薬品開発推進部門**

若林 将史

2023. 11.25 (土)

プログラム医療機器への期待



臨床的仮説の定式化

プログラム医療機器開発特有の統計的手法はあまり聞かない

- 試験をデザインする上でのポイント

- 臨床的仮説を定式化すること

→ （必要に応じて統計家と相談し）統計的手法を決定

	連続量	2値(0 / 1)	生存時間
アウトカムの例	血圧値、 臨床検査値	奏効割合 有害事象発生割合	全生存期間 無増悪生存期間
データの要約法	ヒストグラム	分割表	Kaplan-Meier曲線
群間比較(検定)	t検定、 Wilcoxon検定	カイ2乗検定、 Fisherの 直接確率検定	ログランク 検定
モデルのあてはめ	重回帰分析	ロジスティック回帰	Cox回帰

診断能の評価

あるプログラム医療機器（検査法）の診断能を評価したい

①真の分類結果（病理判定や疾患の発症等）に対する検査法の評価

		真の分類		計
		陽性	陰性	
検査	陽性	a	c	$a+b$
	陰性	b	d	$c+d$
計		N_1	N_2	N

②複数の検査法間の評価

		検査2		計
		陽性	陰性	
検査1	陽性	a	c	$a+b$
	陰性	b	d	$c+d$
計		N_1	N_2	N

①と②で評価方法にどのように違いがあるのか？

本日の講義の流れ

- 診断法評価における指標
 - 感度, 特異度
 - 陽性的中度, 陰性的中度
 - ROC曲線, c統計量
- 閾値（カットオフ値）設定における注意点
- 診断法間の一致度の指標
 - マクネマー検定 (McNemar test)
 - カッパー係数 (κ)

診断法評価における指標

2×2分割表

- 検査データの要約

		真の分類		計
		陽性	陰性	
検査	陽性	a	b	a+b
	陰性	c	d	c+d
計		N_1	N_2	N

- 「行：検査(診断)の結果, 列：真の分類」で統一

がんの遺伝子検査の例

- ある遺伝子変異(変異割合5%)を1000例で検査

		真の分類		
		陽性	陰性	計
検査	陽性	45	50	95
	陰性	5	900	905
	計	50	950	1000

- この検査の診断能を評価する方法は？

正診率（Accuracy rate）

- 定義：真の分類と一致した割合
 - 「1 - 誤分類率」を正診率（正しく診断できた割合）
- シンプルな評価指標

		真の分類		計
		陽性	陰性	
検査	陽性	45	50	95
	陰性	5	900	905
計		50	950	1000

$$\text{正診率} = \frac{45 + 900}{1000} = 94.5\%$$

正診率の問題点

- 真に陽性の450例を追加検査した場合
(同じ検査性能が再現されるとする)

ここが単純に10倍になるだけ

		真の分類		
		陽性	陰性	計
検査	陽性	450	50	500
	陰性	50	900	950
	計	500	950	1450

正診率= $\frac{450+900}{1450} = 93.1\%$

- 検査対象集団によって結果が変わってしまう
- 陽性例・陰性例の検査を個別に評価できない

感度（sensitivity）と特異度（specificity）

- 真の分類を正しく予測できた割合
 - － 感度：真に陽性の人のうち、検査でも陽性となった人の割合
 - － 特異度：真に陰性の人のうち、検査でも陰性となった人の割合

表を縦にみる ↓↓↓↓↓↓↓↓

		真の分類		計
		陽性	陰性	
検査	陽性	45	50	95
	陰性	5	900	905
計		50	950	1000

$$\text{感度} = \frac{45}{45 + 5} = 90.0\%, \quad \text{特異度} = \frac{900}{50 + 900} = 94.7\%$$

追加検査後のデータ

- 真に陽性の450例を追加検査した場合
(同じ検査性能が再現されるとする)

		真の分類		計
		陽性	陰性	
検査	陽性	450	50	500
	陰性	50	900	950
	計	500	950	1450

$$\text{感度} = \frac{450}{450 + 50} = 90.0\%, \quad \text{特異度} = \frac{900}{50 + 900} = 94.7\%$$

- 検査対象集団（真の陽性割合）に依存しない

陽性的中度と陰性的中度

- 検査結果が的中したか否かを測る指標
 - 陽性的中度 (PPV; Positive Predictive Value)
 - 検査陽性例のうち、真に陽性だった人の割合
 - 陰性的中度 (NPV; Negative Predictive Value)
 - 検査陰性例のうち、真に陰性だった人の割合

表を横にみる



		真の分類		計
		陽性	陰性	
検査	陽性	45	50	95
	陰性	5	900	905
計		50	950	1000

$$\text{PPV} = \frac{45}{45 + 50} = 47.4\%, \quad \text{NPV} = \frac{900}{5 + 900} = 99.4\%$$

真の陽性割合とPPV・NPVの関係

- 感度90.0%・特異度94.7%の2×2表データ

(1) 真の陽性割合5%

		真の分類		計
		陽性	陰性	
検査	陽性	45	50	95
	陰性	5	900	905
計		50(5%)	950	1000

$$\text{PPV} = \frac{45}{45 + 50} = 47.4\%$$
$$\text{NPV} = \frac{900}{5 + 900} = 99.4\%$$

(2) 真の陽性割合20%

		真の分類		計
		陽性	陰性	
検査	陽性	180	43	223
	陰性	20	757	777
計		200(20%)	800	1000

$$\text{PPV} = \frac{180}{180 + 43} = 80.7\%$$
$$\text{NPV} = \frac{757}{20 + 757} = 97.4\%$$

ここまでのまとめ

- 感度/特異度
 - 検査対象集団の真の陽性割合に依存しない
 - 検査方法の特性
- PPV/NPV
 - どのような対象集団に検査を適応するかによって異なる
 - 検査対象集団の真の陽性割合に依存
 - 一般集団に対するスクリーニング検査か確定診断の検査か
(後述)

閾値（カットオフ値）設定における注意点

連続量マーカーによる検査

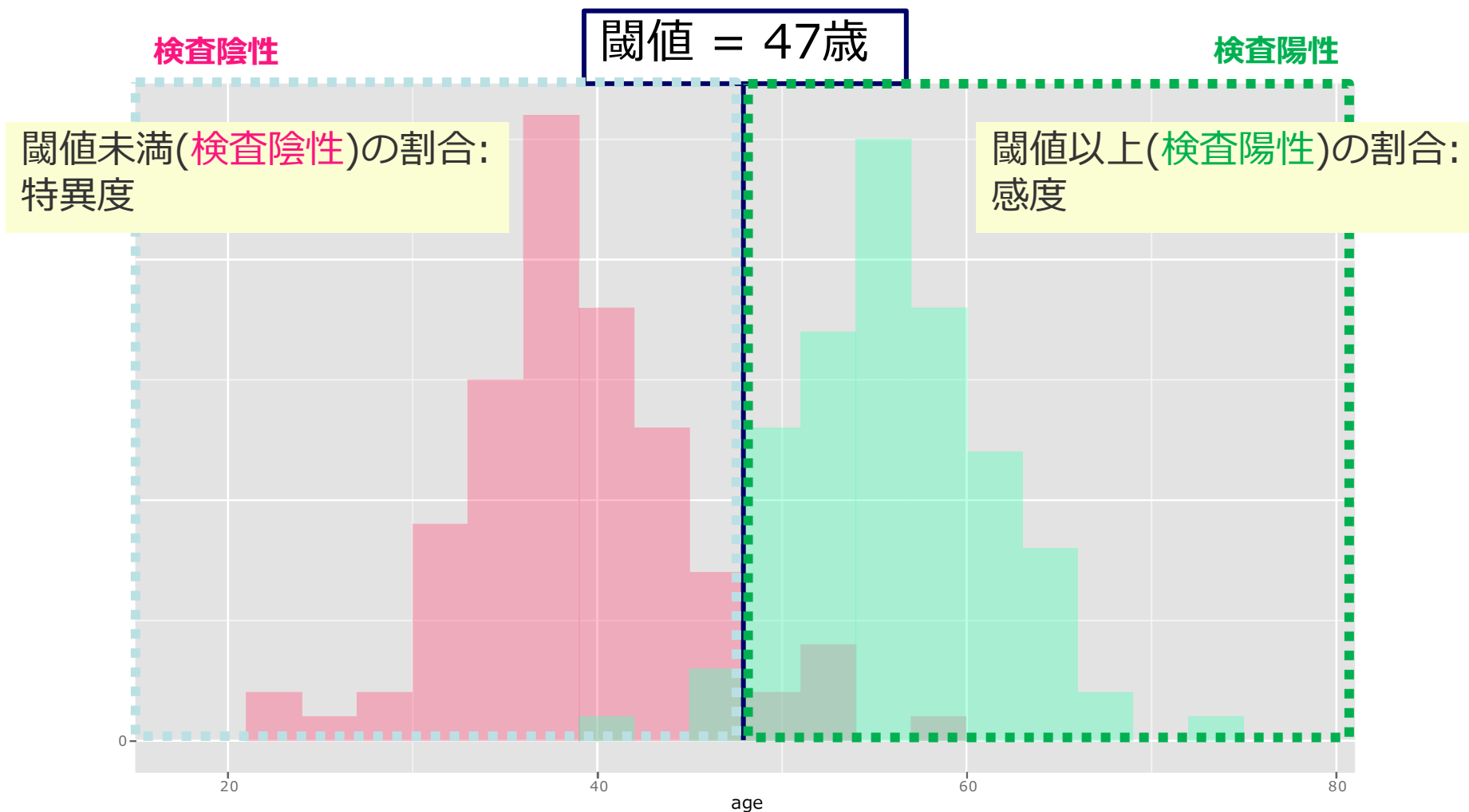
- 年齢が高いと合併症の発生頻度が高い？

ID	年齢	合併症	200例のデータ（年齢の昇順）	
1	22	なし	合併症低リスク群 (検査陰性)	
2	24	なし		
3	32	なし		
...				
100	46	あり	合併症高リスク群 (検査陽性)	閾値 = 47歳
101	48	なし		
102	49	あり		
103	50	あり		
...				
199	68	あり		
200	74	あり		

閾値と感度・特異度の関係

注：仮想例

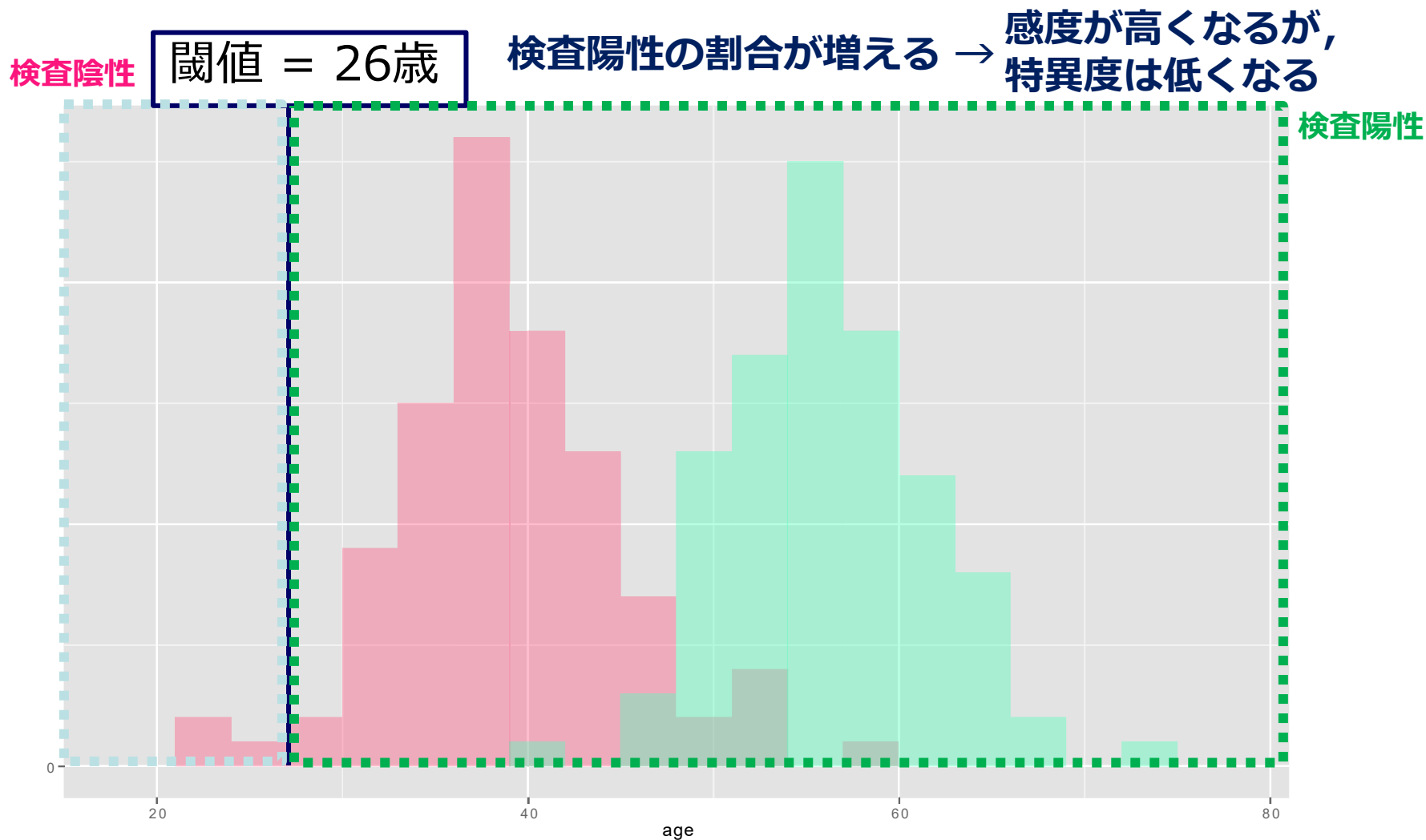
合併症なし
合併症あり



閾値を低くすると・・・

注：仮想例

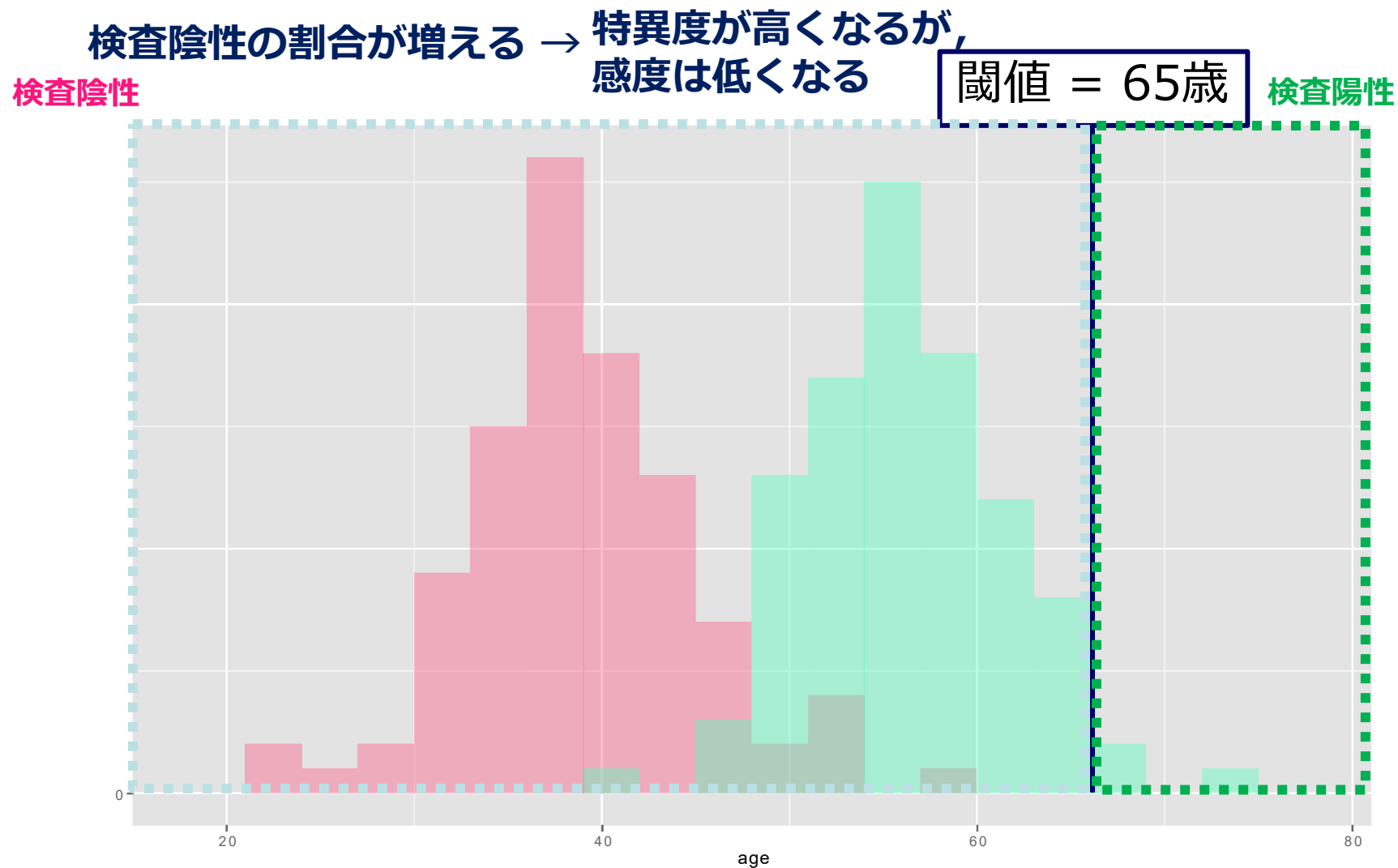
合併症なし
合併症あり



閾値を高くすると・・・

注：仮想例

合併症なし
合併症あり



感度・特異度はトレードオフの関係

- 連続マーカーの場合は閾値の選択によって、診断能の評価が変わってしまう
- 全体的な診断能の要約にはROC曲線がよく用いられる
 - 閾値の探索にも用いられる

感度/特異度の高い閾値の探索法

- 閾値を「c」と決めたときの2×2分割表

		合併症あり	合併症なし	
検査	年齢 $\geq c$ (陽性)	30	10	40
	年齢 $< c$ (陰性)	70	90	160
	計	100	100	200

- この表から感度・特異度が計算できる
- 閾値 (c) を変えると水色のセルの値も変わる

- ROC (Receiver Operating Characteristics) 曲線

- 取り得るすべての閾値 (c) で感度・特異度を計算
- 横軸を1-特異度, 縦軸を感度とした曲線

例：先ほどの例から10例だけ抽出

年齢	合併症	
24	なし	閾値①
28	なし	閾値②
32	あり	閾値③
35	なし	閾値④
40	あり	閾値⑤
45	なし	閾値⑥
50	あり	閾値⑦
52	なし	閾値⑧
55	あり	閾値⑨
57	あり	閾値⑩
		閾値⑪

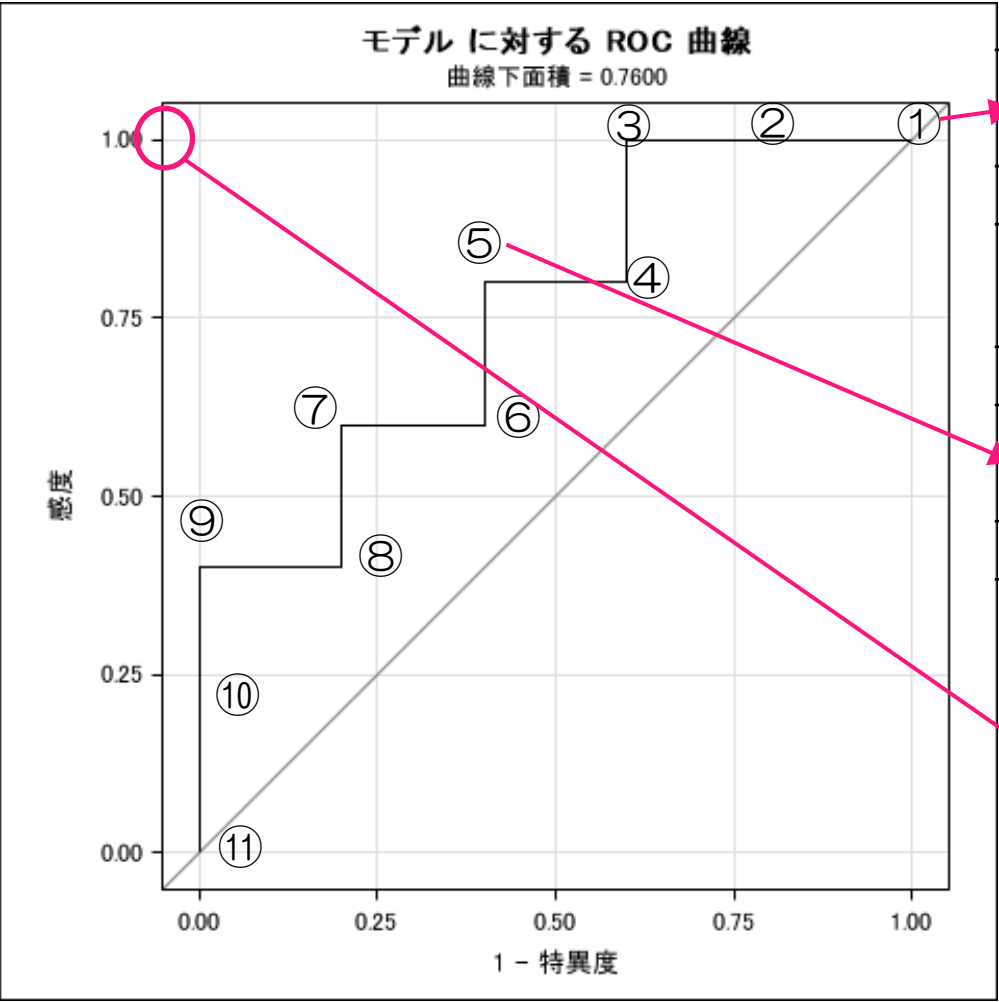
● 閾値①の場合(感度：100%，特異度：0%)

		合併症あり		合併症なし
検査	24歳以上	5	5	10
	24歳未満	0	0	0
	計	5	5	10

● 閾値⑤の場合(感度：80%，特異度：60%)

		合併症あり		合併症なし
検査	40歳以上	4	2	6
	40歳未満	1	3	4
	計	5	5	10

作成したROC曲線



閾値①	合併症あり 合併症なし	
24歳以上	5	5
24歳未満	0	0
計	5	10

(1-特異度：100%，感度：100%)

閾値⑤	合併症あり 合併症なし	
40歳以上	4	2
40歳未満	1	3
計	5	5

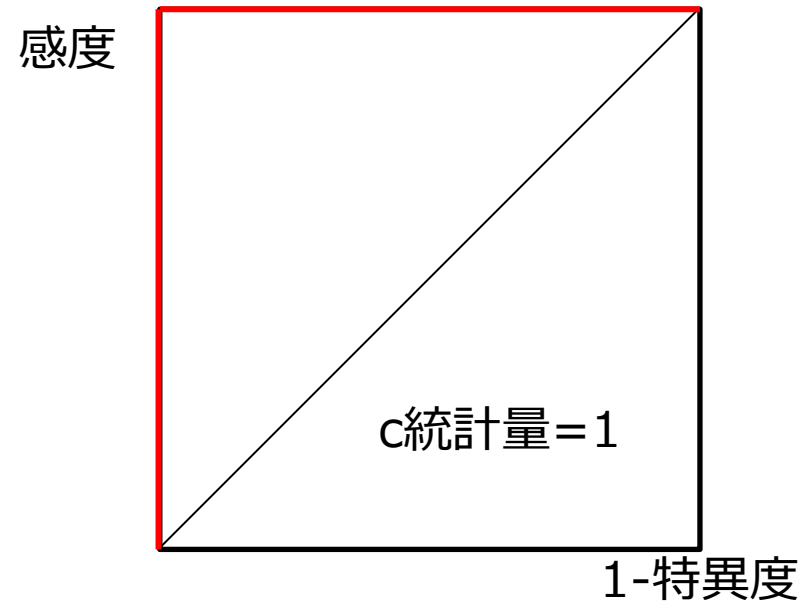
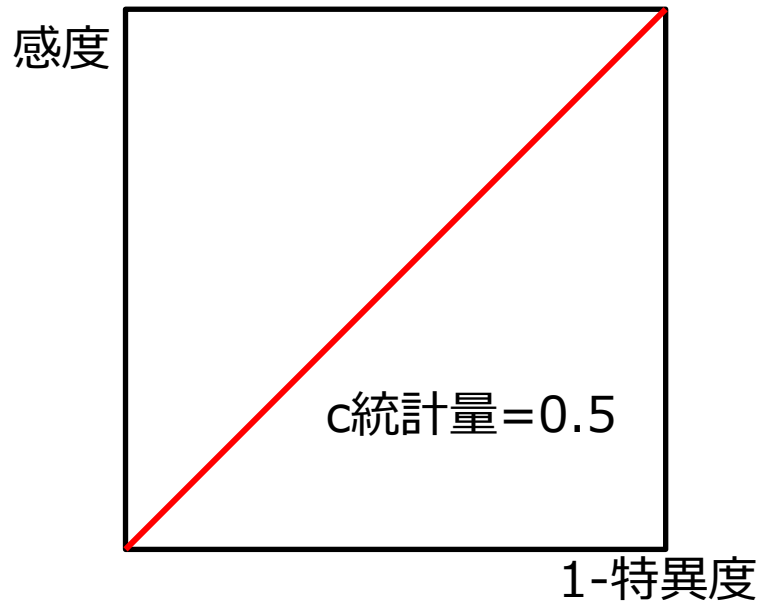
(1-特異度：40%，感度：80%)

左上の点 (感度100%，特異度100%)
に近いほど良い閾値

連続マーカーの診断能

- ROC曲線の曲線下面積 (AUC: Area Under the Curve)

無意味なマーカー: ランダムな診断 完璧なマーカー: どの閾値でも感度100%



- c統計量と等価
 - concordance statistics, c-index, concordance probability

c統計量

- 連続マーカで閾値以上を陽性とする場合
 - 「陽性例のマーカ値が陰性例のマーカ値を上回る確率」と定義される
 - 検査結果と真の状態との一致の程度
- 取り得るすべてのペアについて、一致/不一致を検討

年齢	24	28	32	35	40	45	50	52	55	57
合併症	なし	なし	あり	なし	あり	なし	あり	なし	あり	あり



一致/不一致のペア

注：仮想例

-  (35歳) <  (50歳)
 - 年齢が高いと合併症あり → 「一致」
-  (32) <  (35)
 - 年齢が高いと合併症なし → 「不一致」
-  (30) =  (30)
 - 年齢が同じ → 「タイデータ」
-  (32) <  (50)
 - 両者ともに合併症あり（または合併症なし）
 - 判断がつかないため「情報なし」

 : 合併症なし
 : 合併症あり

「情報あり」

c統計量の計算

年齢	24	28	32	35	40	45	50	52	55	57
合併症	なし	なし	あり	なし	あり	なし	あり	なし	あり	あり

- 全ての情報のあるペア=25通り

- 24歳 - 32,40,50,55,57
- 28歳 - 32,40,50,55,57
- 32歳 - 35,45,52
- 35歳 - 40,50,55,57
- 40歳 - 45,52
- 45歳 - 50,55,57
- 50歳 - 52
- 52歳 - 55,57
- 55歳 - なし

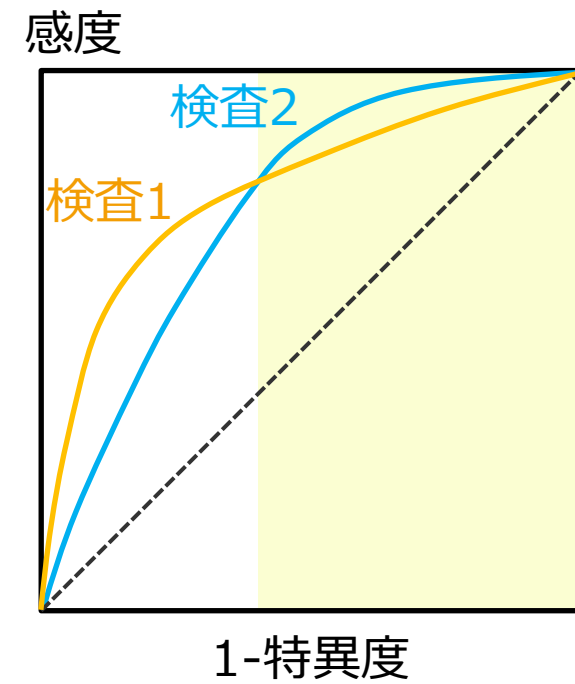
赤：一致19 , 青：不一致6

$$\begin{aligned}
 \text{c統計量} &= \frac{\text{一致ペア数} + 0.5 \times \text{タイデータ数}}{\text{一致ペア数} + \text{不一致ペア数} + \text{タイデータ数}} \\
 &= 19 / 25 = 0.76
 \end{aligned}$$

- ROC曲線のAUCと一致

c統計量による比較の注意点

- 検査同士の優劣はc統計量の大小だけでは不十分
 - ROC曲線が完全に上回っているとは限らない
 - 網掛け部分より左では検査1の方が検査2よりも感度が高い
 - 実際は閾値を固定した検査を行う
 - 感度/特異度のバランスが良い検査が選択されるべき
 - c統計量の大小では判断できない



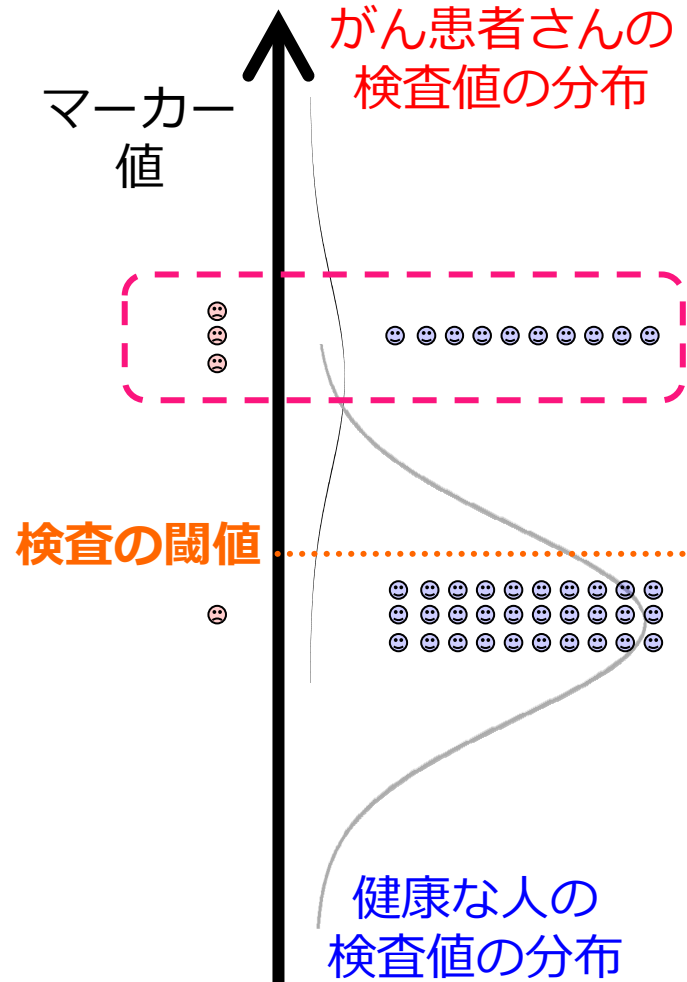
事例の補足：閾値の決定

- 臨床的観点から感度と特異度に優先順位を設定
 - 「優先する特異度（or 感度）が〇〇%以上の閾値のうち、感度（or 特異度）が最大」という基準は有用
- 感度と特異度に優先順位が特にならない
 - 統計的（幾何学的）に決めることは可能
 - 「左上点からの距離が最も近い閾値」
 - 「対角線からの距離が最も長い閾値」

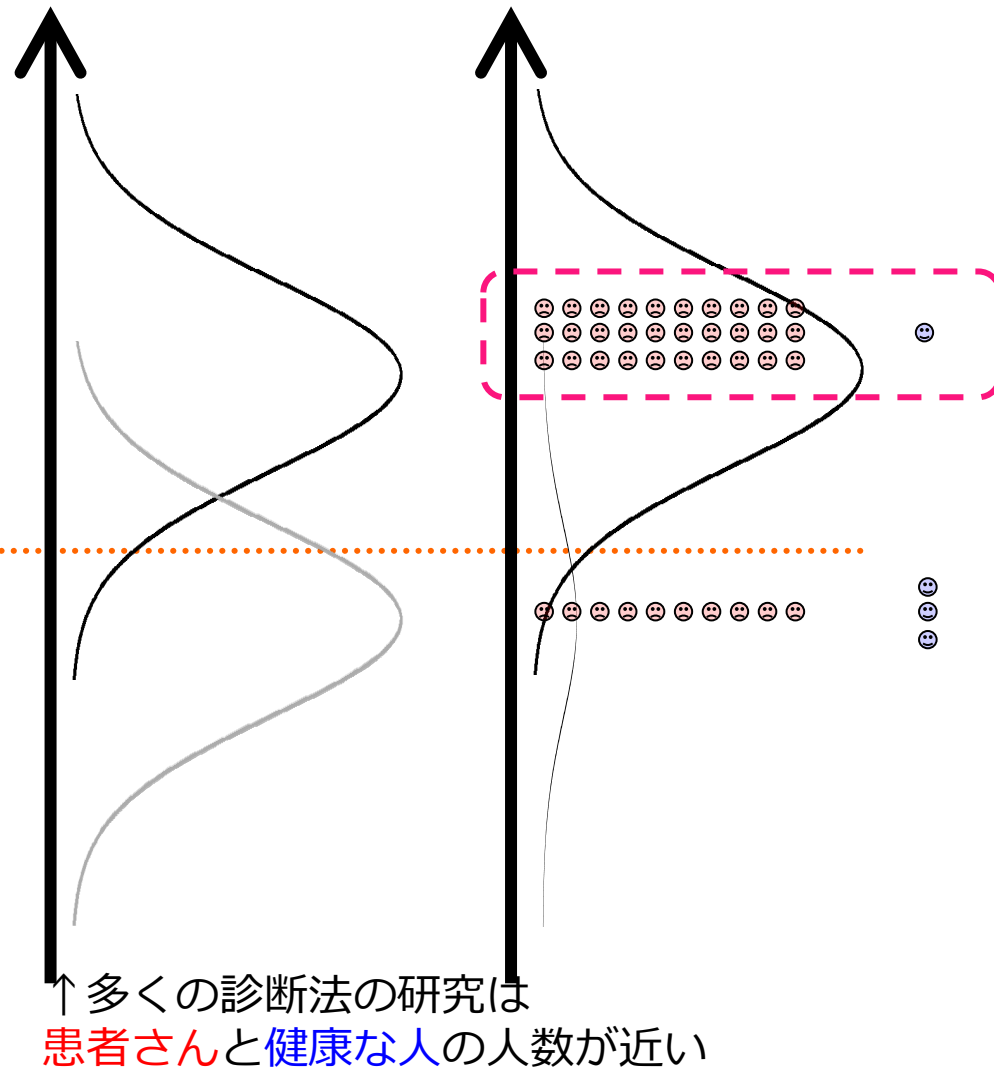
検診の診断能評価

- 診断目的、対象集団によって考え方が異なる
- スクリーニング時：罹患割合が低い
 - PPVは低くなる
 - NPVが高いほどよい
 - 検査が陰性であれば, 安心
 - 感度が高いことが必要
- 追加検査・確定診断時：罹患割合が高い
 - PPVが高いほどよい
 - 特異度が高いことが必要

スクリーニング（がん検診）の場合：
検査陽性の多くは健康な人



確定診断に使う場合：
検査陽性の多くはがんの人



ここまでのまとめ

- 診断能の要約にROC曲線, c統計量
- c統計量が大きい検査が良いとは限らない
 - 感度/特異度のバランスが重要
 - 検査によって過小/過剰治療となってしまう場合, 感度/特異度に優先順位がつけられることがある
- 検査目的、対象集団に応じた診断能の評価が必要
 - PPV/NPVは検査結果の解釈に重要
 - 診断法を評価した対象と実際の検査対象の罹患割合が異なるとPPV/NPVも変わるため、診断能を見誤る恐れがある

診断法間の一致度の指標

診断法間の一致

- 新しい診断法（ツール）を評価する状況において、
（真の分類との一致ではなく）**既存の診断法との一致を評価**
 - 評価者間での一致を評価する場合も同じ
 - 感度・特異度といった指標は計算できない
- 同一症例に複数の検査（評価）
 - Pairedデータ

注：仮想例

		検査法2		計
		陽性	陰性	
検査法1	陽性	n_{11}	n_{12}	n_{1+}
	陰性	n_{21}	n_{22}	n_{2+}
計		n_{+1}	n_{+2}	n_{++}

Paired データに対する一致の評価

- McNemar検定
 - 「検査法1の陽性割合 = 検査法2の陽性割合」に対する検定
 - 有意差あり：検査間で陽性割合に差がある
 - 検定結果（p値）は症例数に依存するため評価が難しい
 - 傾向スコアマッチング等、マッチングした集団におけるペアを考慮した割合の検定として用いられることもある
- κ 係数
 - 検査間の一致度を定量的に示す指標
 - 「偶然の一致」を除いた上で一致度を評価
 - 正診率とは値が異なるので注意が必要

McNemar検定

		検査法2		計
		陽性	陰性	
検査法1	陽性	n_{11}	n_{12}	n_{1+}
	陰性	n_{21}	n_{22}	n_{2+}
計		n_{+1}	n_{+2}	n_{++}

- 「 H_0 : 検査法1の陽性割合 = 検査法2の陽性割合」に対する検定
- 非対角成分における割合の差の検定（カイ二乗検定）

$$\chi^2 = \frac{(n_{12} + n_{21})^2}{n_{12} + n_{21}}$$

κ係数

		検査法2		
		陽性	陰性	計
検査法1	陽性	p_{11}	p_{12}	p_{1+}
	陰性	p_{21}	p_{22}	p_{2+}
計		p_{+1}	p_{+2}	1

$$\hat{\kappa} = \frac{p_o - p_e}{1 - p_e}$$

$$p_o = \sum_{i=1}^2 p_{ii}, \quad p_e = \sum_{i=1}^2 p_{i+} \cdot p_{+i}$$

• $\hat{\kappa} = 1$: 完全に一致

• $\hat{\kappa} = 0$: 独立

Landis & Kochの分類



*Landis & Koch以外にも
評価分類は複数存在

まとめ

- 診断能を評価する指標の紹介
 - 感度/特異度/PPV/NPV/ROC曲線/c統計量
 - 真の陽性割合に影響される指標には注意が必要
 - 検査目的、対象集団に応じた指標の使い分け
- 閾値の設定は、統計的観点と臨床的観点のバランスが重要
- 検査法の一致を評価する方法
 - McNemar検定
 - κ 係数