

被験者あるいは動物内に繰り返しのあるエンドポイントの試験デザインと統計解析

古川敏仁
株式会社バイオスタティスティカル
リサーチ
2009/6/13

はじめに

- 動物実験で複数のデバイスを1匹の動物に埋植する場合の評価の仕方
動物単位で集計すべきか、デバイス単位で集計すべきか
- アンケート調査 A、B2つの機器は本当はどちらが優秀なのか
多施設調査の評価の仕方の問題
- 第10回医療機器臨床試験研究会 2008年5月17日 コンタクトレンズ有
害事象 1つの体に目は二つ、はたして集計方法は
- 心血管インターベンション
複数血管部位にステントを使用する場合の注意事項
- 連続して複数発生する心発作の制御の成功の有無の評価の仕方は
発作単位、人間単位それとも個人差を考慮した発作単位の解析

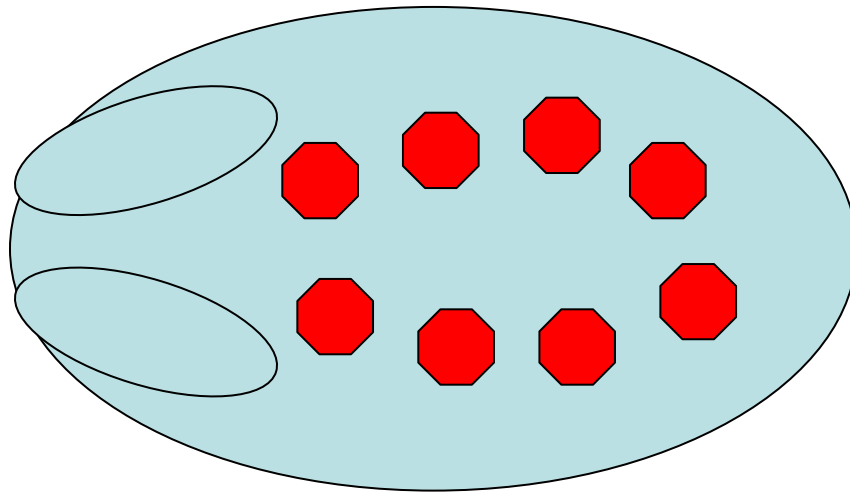


動物実験で複数のデバイスを
1匹の動物に埋植する場合の
評価の仕方は



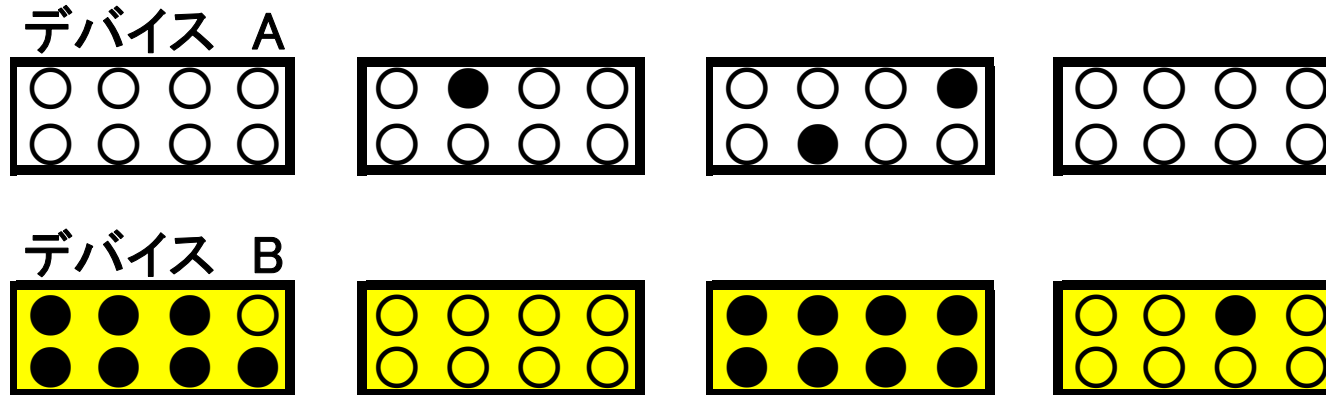
例①：動物実験データ

- ウサギの背中にデバイスを埋植し、炎症の有無を見る実験



- 1匹のウサギに8個のデバイスが埋植され、それぞれのデバイスの周辺で炎症の有無が観察されていました。
- Aデバイスを4匹のウサギに
- Bデバイスを4匹に使用し、
- 炎症の発生率に差があるかを確認しました。

例1: 実験結果



- デバイス単位

- 炎症あり Aデバイス 3/32 $p=0.0004$ (χ^2 検定)
 - Bデバイス 16/32

- ウサギ単位

- 炎症あり Aデバイス 0/4 $p=0.1025$ (χ^2 検定)
 - Bデバイス 2/4

あなたは、どちらの結果を支持しますか？

- デバイス単位
 - 炎症あり Aデバイス 3/32 $p=0.0004$ (χ^2 検定)
 - Bデバイス 16/32
- ウサギ単位
 - 炎症あり Aデバイス 0/4 $p=0.1025$ (χ^2 検定)
 - Bデバイス 2/4

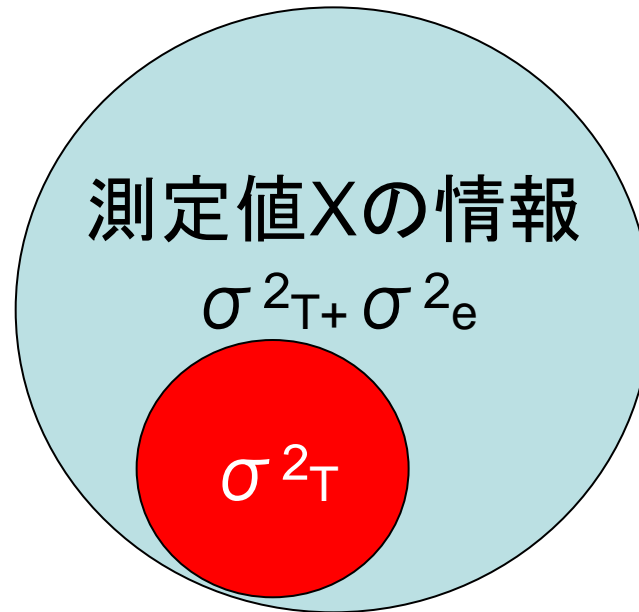
私の出した答え

- デバイス単位
 - 炎症あり Aデバイス 3/32 $p=0.0004$ (χ^2 検定)
Bデバイス 16/32
- ウサギ単位
 - 炎症あり Aデバイス 0/4 $p=0.1025$ (χ^2 検定)
Bデバイス 2/4
- GEE (ウサギ内のデバイス評価の類似性を考慮した方法)
 - $p=0.0346$ (GEE)
 - 対象内相関 $R=0.65$ ウサギ1匹に対して8個のデバイスがありますが、実は 3.3個の対象(例数)が存在することになります。
 - (GEEの解析では $R=0.3674$ ですが)

相関係数 Rとは

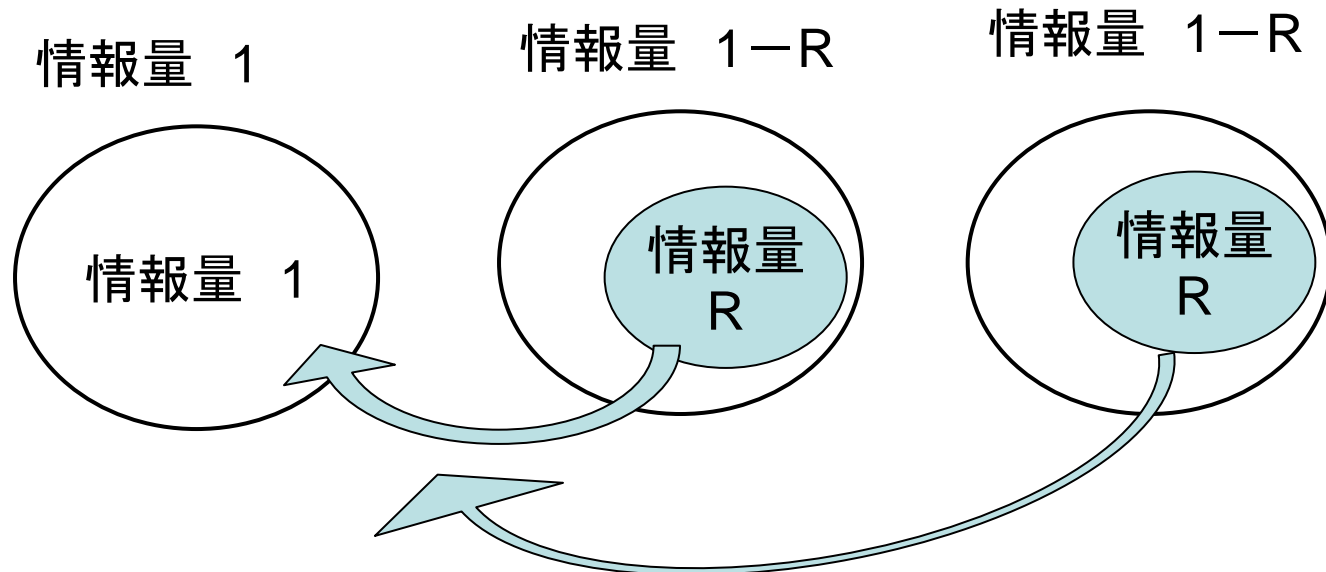
相関係数 $R = \frac{\sigma^2_T}{\sigma^2_T + \sigma^2_e}$ は、

測定イベント、機器が持っている情報を1とした場合の別のイベント、機器で説明されてしまう情報 σ^2_T : の割合
そのイベント、機器だけが持つ情報 σ^2_e が重要 (1-R)

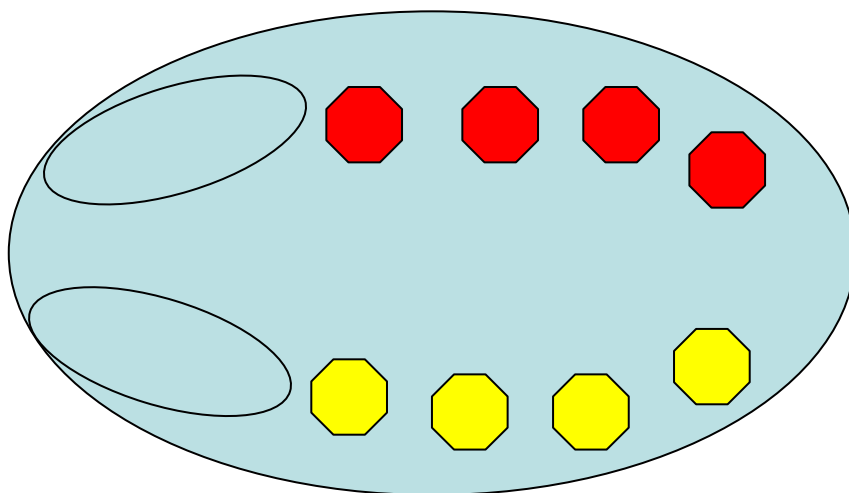


クラスター(対象内)相関

- 1つのクラスター内にk個の観察があるとし、
- 対象内に相関(R)があると、実質症例数は $1+(K-1)(1-R)$ になります。

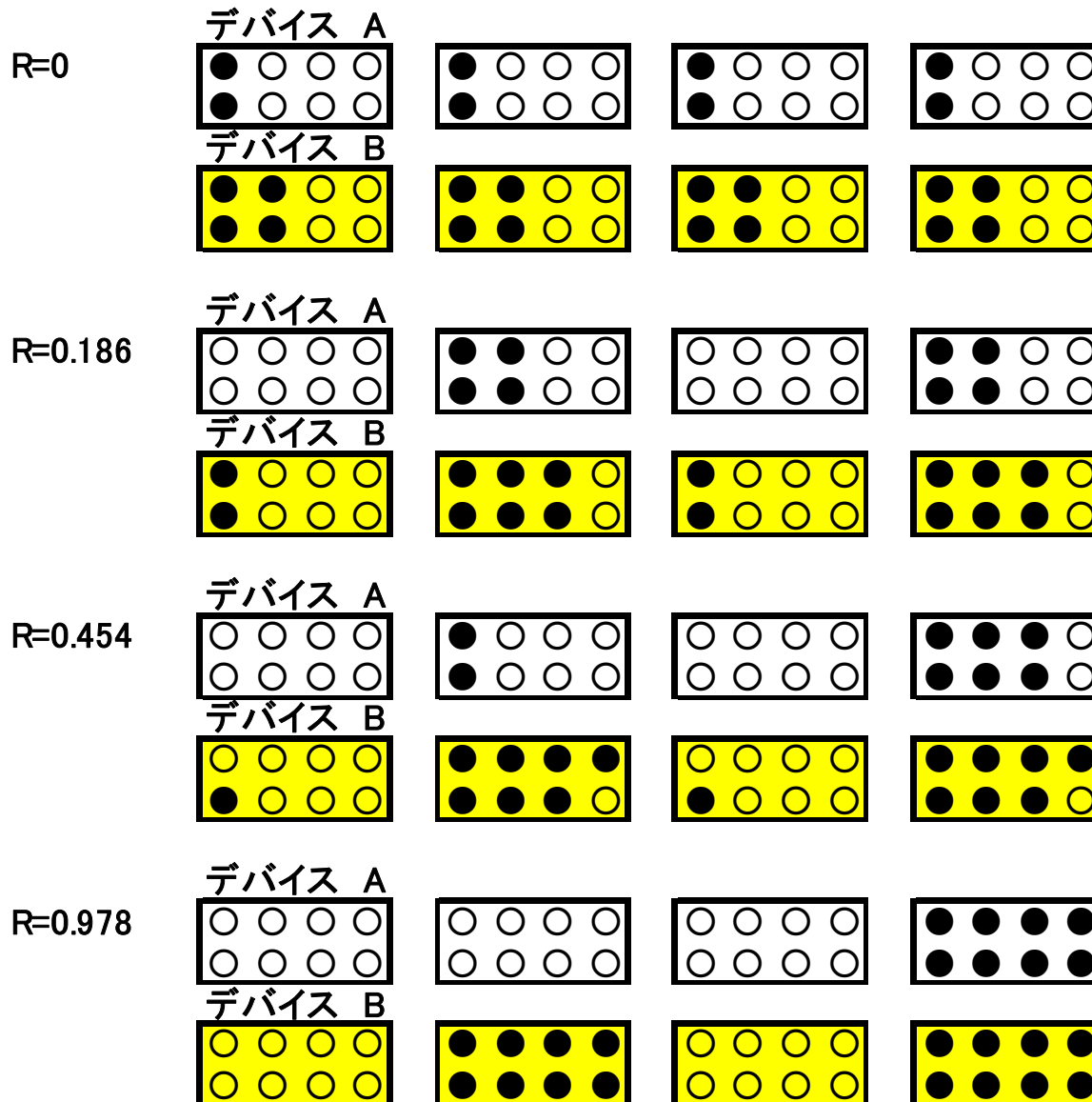


正しい実験の仕方とは何か



- 統計的には1匹のウサギに3.3個分の情報があることが分かりました。
- しかし、ウサギ個体の体調や性質により炎症の程度は違ってきます。統計的には3.3個の情報ですが、やはり、この場合はウサギ単位で解析すべきでしょう。つまり、この実験からだけ判断するのなら、A機器とB機器は有意差はないと考えるのが妥当だと思います。
- また、ウサギ単位でも、ウサギに対する無作為化割付や、十分な盲検化がなければ評価自体無意味でしょう
- 1匹のウサギの情報を3.3倍であるか本当に確認するためには、上記のように、1匹のウサギの中に対照機器を入れる必要があります。

クラスター間に相関があるとは



GEEと χ^2 p値の比較

Working Correlation	GEE p値	デバイス	ウサギ*
		n=32	n=8
		χ^2 p値	
0	—	0.0389	
0.047	0.0675	0.0389	
0.186	0.1874	0.0389	
0.454	0.3217	0.0389	
0.978	0.4720	0.0389	0.4652

*Aデバイス n=1/4、Bデバイス=2/4として計



アンケート調査 A、B2つの機器
は本当はどちらが優秀なのか
多施設調査の評価の仕方の問題



調査内容

- 患者さんのQOLが問題となる機器です。
- A、B2つの機器があり、交換することが可能です。
- 5つの病院（実際にはもっと多数ですが）の
- A、B両方の機器を使用したことのある患者さんに
- どちらの機器が良かったかを質問しました。

A社の調査結果

施設名	調査人数	Aが良い 回答者	Aが良い 回答率(%)
①	2	2	100
②	5	4	80
③	10	6	60
④	50	24	48
⑤	100	35	35
平均			64.6

- 施設ごとのAが良い評価者の割合の平均を取ったら64.6%と50%を超えていたので、
- A社の製品が優れています。

B社の調査結果

施設名	調査人数	Aが良い 回答者	Aが良い 回答率(%)
①	2	2	
②	5	4	
③	10	6	
④	50	24	
⑤	100	35	
合計	167	71	42.5

- 全体の人数のAが良い評価者の割合を取ったら42.5%と50%を下回っていたので、
- B社の製品が優れています。

あなたは、A社、B社のどちらの主張を 支持するか

施設名	調査人数	Aが良い 回答者	B社主張 全体 率 (%)	A社主張 施設ごと 率 (%)
①	2	2		100
②	5	4		80
③	10	6		60
④	50	24		48
⑤	100	35		35
合計	167	71	42.5	64.6

古川のやり取り

施設名	調査人数	Aが良い 回答者	B社主張 全体 率(%)	A社主張 施設ごと 率(%)
①	2	2		100
②	5	4		80
③	10	6		60
④	50	24		48
⑤	100	35		35
合計	167	71	42.5	64.6

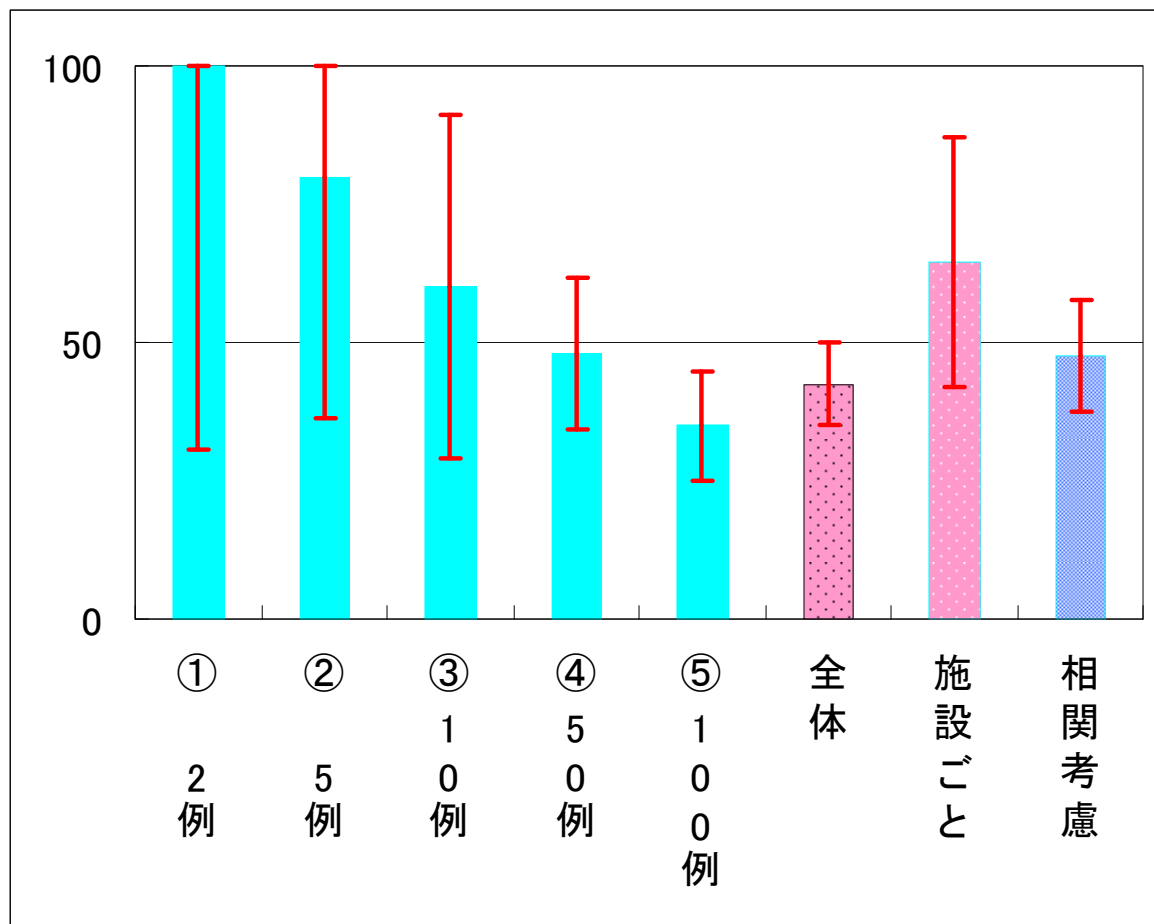
- 客観的に言えば、A社のやり方では、①の施設2名、②の施設5名と言った極端に例数の少ない結果の値を他の結果と平等に扱うのはおかしいと思いますので、A社のやり方は間違っていると思います。
- B社の集計も、B社が優れていると明確に示されているのは、⑤施設の結果のみであり、⑤施設特有の問題があるかもしれません。

施設内被験者の関連を考慮した解析


施設名	調査人数	Aが良い 回答者	B社主張 全体 率(%)	A社主張 施設ごと 率(%)	施設内 相関を 考慮
①	2	2		100	
②	5	4		80	
③	10	6		60	
④	50	24		48	
⑤	100	35		35	
合計	167	71	42.5	64.6	47.5

- 施設ごとの平均ではなく
- ⑤施設の結果の偏りをなくす
- 施設内被験者の関連を考慮した解析
- 結果47.5% B社調査42.5%ほどBに有利ではないが、多少Bに有利な結果


施設内被験者の関連を考慮した解析



- 残る問題
- 施設数の増加とともにB製品が有利となる傾向は存在するののか
- 施設内被験者の関連を考慮した解析は正しいとは限らない



第10回医療機器臨床試験研究会
2008年5月17日 コンタクトレンズ
有害事象 1つの体に目は二つ、
相関を考慮した集計方法は



角膜潰瘍の発生率の比較

あなたは、①、②どちら

	Aコンタクトレンズ			Bコンタクトレンズ		
	評価例数	発生人数	発生率	評価例数	発生人数	発生率
症例単位	100	40	40	100	30	30

① 人単位の集計 A、B2つのコンタクトレンズの角膜潰瘍の発生率に差はない
 発生率の差 10%
 χ^2 p値 0.1392

	Aコンタクトレンズ			Bコンタクトレンズ		
	評価目数	発生目数	発生率	評価目数	発生目数	発生率
右目	100	30	30	100	20	20
左目	100	30	30	100	20	20
合計	200	60	30	200	40	20

② 目単位の集計 A、B2つのコンタクトレンズの角膜潰瘍の発生率に差がある
 発生率の差 10%
 χ^2 p値 0.0211

角膜潰瘍の発生率の比較

正解は①？

- ①の症例単位の集計はFDAの採用する方法です。
- ②の目単位の方法は間違っています。
- 目に発生する有害事象には、右目で発生すれば左目でも発生しやすいという「個人内相関」があります。
- 有害事象の発生する環境 すなわちレンズの装用習慣、洗浄の方法、免疫力、菌感染環境は目単位ではなく、個人単位だからです。



目の有害事象と症例単位の有害事象 発生率の比較（相関を考慮）

	AコンタクトL		BコンタクトL	
	評価目数	発生目数	評価目数	発生目数
右目	100	30	100	20
左目	100	30	100	20
左右合計	200	60	200	40

相関係数	AコンタクトL		BコンタクトL	
	評価例数	発生人数	評価例数	発生人数
0.0	100	51	100	36
0.3	100	46	100	33
0.5	100	41	100	28
0.8	100	35	100	24
1.0	100	30	100	20

個人内相関を考慮した目単位の解析 GEE解析

	AコンタクトL		BコンタクトL	
	評価目数	発生目数	評価目数	発生目数
右目	100	30	100	20
左目	100	30	100	20
左右合計	200	60	200	40

個人内相関を考慮した解析方法 GEE

相関係数	χ^2	p値
0.0	0.0211	目単位の結果と同じ
0.3	0.0335	
0.5	0.0459	
0.8	0.1324	
1.0	0.1392	症例単位と同じ

目に発生する有害事象の解析 まとめ

- 1症例に対する目は2つあり、目に発生する有害事象には個人内でかなりの相関がある。
 - 目間の相関係数の大きさに、検定結果は大きく変わってしまう。
 - 有害事象ごとに個人内相関係数の大きさは違うが、1有害事象ならGEEは可能(主要評価項目であれば)
 - 複数の項目ですべて相関を考慮して解析するのは難しい(相関の大きさは事象ごとに違うので)
-
- 主要評価項目の解析: GEE
 - 一覧形式の集計: FDAタイプの症例単位。





心血管インターベンション 複数血管部位にステントを使用す る場合の注意事項



例 : Drug-Eluting Stent 比較試験

- レジストリー試験です
- 標的血管にA、B2つのステントのどちらかが埋植されています。
- エンドポイントは再狭窄までの時間です
- 再狭窄の発生ならびに発生までの時間をA、B2つのステントで比較します。
- ただし、約1割ほどの症例が1症例に複数のステントが埋植されていました。
- 同一症例にDESを埋植する場合、必ず同じ種類のステントを埋植します。

例：Drug-Eluting Stent 比較試験 あなたはどちら

- この再狭窄までの時間をA、B2種類のDESで比較する場合、あなたは、どちらの方法を使いますか。
- ① 複数ステントを埋植した症例に関しては、代表的なステント1個を選択し、他は削除して解析する。
- ②複数ステントを埋植した症例に関しては、ステント単位で集計する。すなわち、ステントが2本埋植された場合は2症例として取り扱う

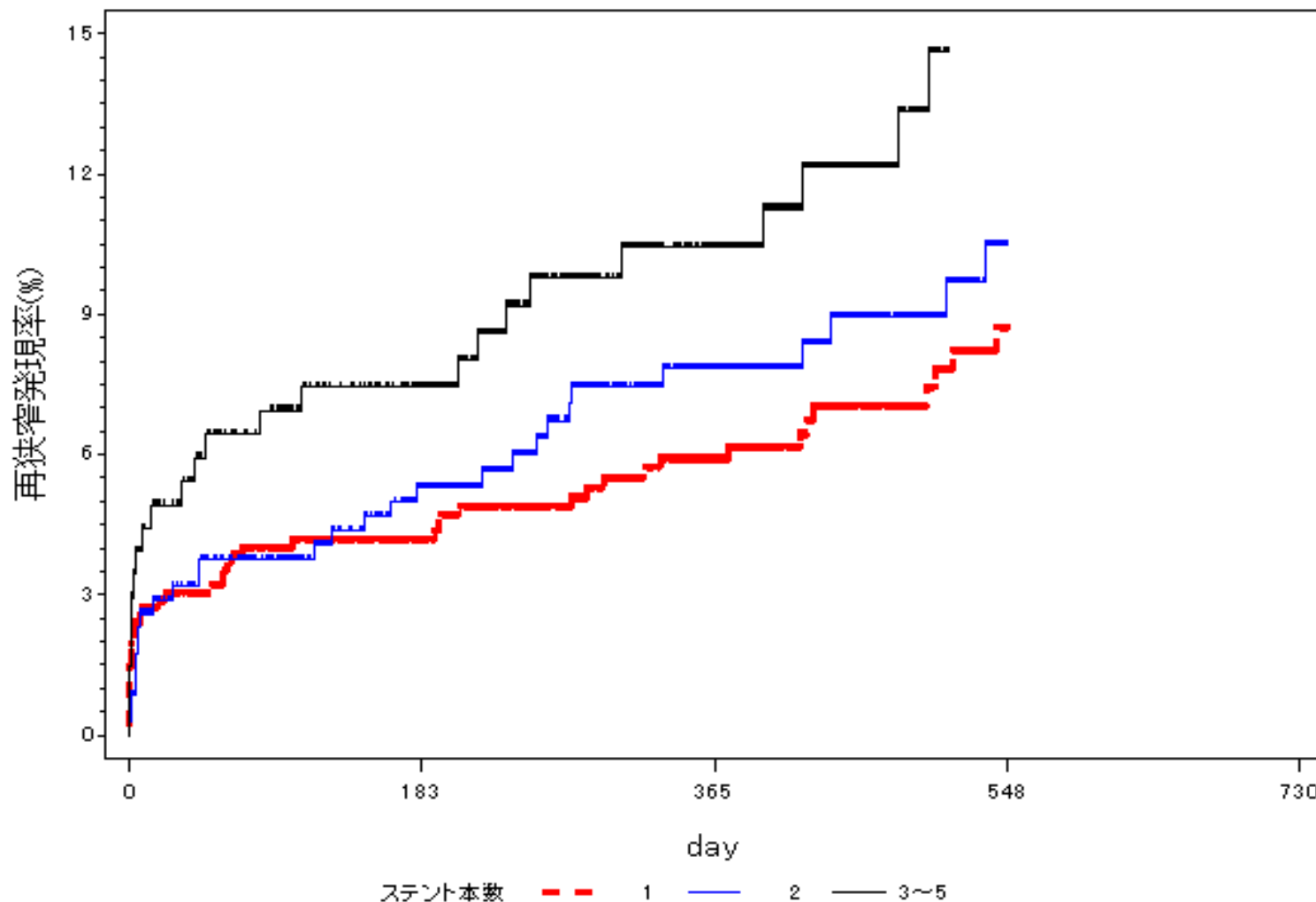
例 : Drug-Eluting Stent 比較試験

正解は③

- ① 複数ステントを埋植した症例に関しては、代表的なステント1個を選択し、他は削除して解析する。
 - ステントを選択する(削除する時点でバイアスが入り、公平な評価はできない)
 - 削除されたステントは情報を持っていないのか？
 - 複数ステントという、重要な情報を無視することによる問題

- ② 複数ステントを埋植した症例に関しては、ステント単位で集計する。すなわち、ステントが2本埋植された場合は2症例として取り扱う
 - シングルステント症例と複数ステント症例のリスクは同じではない。
 - 複数ステント症例が多いステント(A、orB)のリスクが相対的に高くなる。
 - 被験者背景等により、背景因子を調整する(統計的にそろえる:傾向スコア解析等)場合、複数ステント症例は同じ背景情報が使われることになり、複数ステントを持つ症例の情報に全体的な評価が引っ張られる(複数ステントの症例の重みが増しすぎる)

ステントを複数使用した場合の 1本当たりの再狭窄のリスク(fiction)



例 : Drug-Eluting Stent 比較試験 まとめ

- 同一症例に使用した複数のステントから1本選ぶことや、ステント単位で単純に解析するということは、バイアスの入った正しくない結果になる。
- ステント本数に起因する再狭窄のリスクを評価できるような統計手法が必要
 - ・ 共変量による調整
 - ・ ステント本数による層化
- 多変量解析で披験者で調整したり、傾向スコアを用いて解析する場合、複数ステントの症例は同じ情報が何度も使われることになり、調整がすべての症例で同じ重さにならない。この場合、同一症例内のステントの再狭窄のリスクの類似性を考慮した、ある意味症例単位に近い背景因子の取り扱いも必要になる。



連続して複数発生する心発作の制御
の成功の有無の評価の仕方は



心室頻拍の電気刺激による正常化に対する 2種類の機器の性能評価 あなたはどちら

- イベントは一定期間に同一症例内で複数おき、それぞれのイベントごとに「成功」、「失敗」の結果がある。
- →○→○○→●→○○○→●→
- →○→○ ● ○→→○→ ● ● ○○→
- →○○○→ ○ →
- 無作為化割付でA、B2機種に症例を割付、その結果を比較する場合

- ①個人単位で成功、不成功の定義として、症例単位の成功率でA、B2機種を比較

- ②イベント単位の成功、不成功成功率でA、B2機種を比較

心室頻拍の電気刺激による正常化に対する2 種類の機器の性能評価 回答③

- ①個人単位で成功、不成功の定義として、症例単位の成功率でA、B2機種を比較
 - 症例単位の成功、不成功の定義は可能か
 - イベントの情報を全て使ってはいない

- ②イベント単位の成功、不成功成功率でA、B2機種を比較
 - イベントの成功、不成功は個人によって、成功しやすい症例と成功しにくい症例と言う傾向がある。
 - イベント数は個人によって0から100とばらつきがあり、イベント数が多い症例を含む群とそうでない群で評価例数が異なることになる

心室頻拍の電気刺激による正常化に対する2種類の 機器の性能評価 正しい統計解析

- →○→○○→●→○○○→●→
- →○→○●○→→○→●●○○→
- →○○○→○→
- イベント単位の解析は情報量が最も多い
- ただし、症例内の類似性を考慮しなければならない
- 実質的なイベント数は、個人内相関係数 r で調整
- GEE(個人内相関を考慮した解析方法)で、イベント単位の解析

統計解析と症例数との兼ね合

- 平均的に6ヶ月間に20イベントが発生
- 被験者内の「成功」の有無に関する過去の論文から被験者内相関 $R=0.74$
- すると $1+(K-1)(1-R)=1+(20-1)(1-0.74)=4$
- つまり、被験者内相関を考慮したイベントベースの解析を行えば、1症例で4症例分の情報が得られる
- 試験必要症例数は症例単位の評価に比べ、クラスター内相関を考慮すれば1/4になる

降圧効果臨床試験での間違い

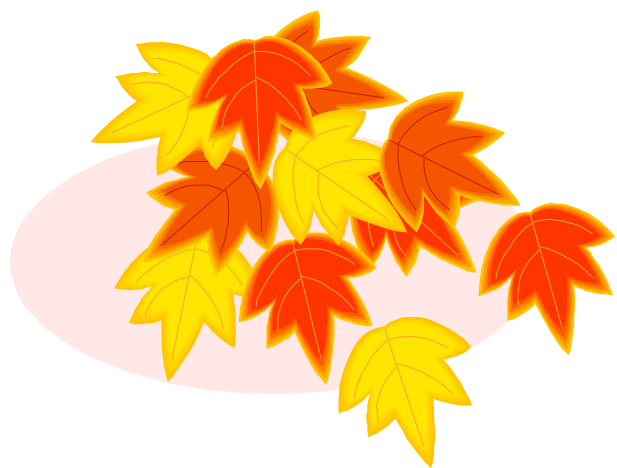
一定期間の観察で、血圧が正常化した割合を薬剤間で比較するのは間違い

GEEなど適切な評価が必要

							評価数	成功数	成功率
A群	①	○				○	2	2	100.0
A群	②	○	○	●	○	●	5	3	60.0
A群	③		○	○	○	●	4	3	75.0
A群	④		●		●		2	0	0.0
例数							4	平均	58.8
								GEE	62.0

							評価数	成功数	成功率
B群	①	○	○	○	○	○	5	5	100.0
B群	②	○	○	●	○	●	5	3	60.0
B群	③	○	○	○	○	●	5	4	80.0
B群	④	○	●	○	●	○	5	3	60.0
例数							4	平均	75.0
								GEE	75.0

②標本t 0.2231
GEE 0.4176



このような試験を計画するときは
プロトコール作成時の注意事項



情報単位

- この、被験者内相関の話は、臨床試験のタイプ、試験例数を検討する基本的な部分では、共通点があります。
- それは、情報をどう捉えるかです。
- つまり、症例を1情報単位とするのか
- 症例内の機器(イベント)を情報単位とするのか
- それとも、個人内相関を考慮した症例と機器(イベント)の中間的な情報を情報単位にするのかです。

例数設計では

- 例数設計はこの情報単位をもとに例数設計をします。
- 1症例内に複数の機器や評価イベントが存在する場合、適切な統計解析を使用すれば、必要症例数は症例単位の評価よりも少数例ですむこととなります
例(1症例に20イベントが発生する心発作制御の機器では 相関が0.74と高いにも関わらず症例数が1/4ですむこととなります)

試験実施計画書での記載注意

- 試験実施計画書(プロトコール)は臨床試験の設計図です。
- 科学的な結論を得るためには、計画時の論理的な妥当性が明確となる記述が必要になります。(決して、治験をするためにしぶしぶ作成する文章ではありません。)
- 被験者内相関を考慮する内容を以下のポイントについて記載する必要があります
 - 試験の目的
 - 主要評価項目
 - 例数設計
 - 統計解析

主要評価項目

評価項目とその評価の仕方を記載

例:

DESに発生する10ヶ月までの再狭窄の発生ならびに発生までの時間を、ステントごとに個人内相関を考慮した上で比較する。

例: 一定の観察ポイントごとに観察される降圧効果の達成の有無を、個人内相関を考慮した(症例特性を考慮した)イベント単位に降圧薬A、Bで比較する

試験実施計画書での記載注意

- 症例数設定

統計解析手法

検定の有意水準 α 、

検出力 $1-\beta$

予想される効果差 Δ

予想される評価項目の標準偏差 σ

被験者内相関 r

特にその推定方法とその妥当性

試験実施計画書での記載注意

- 統計解析 主要評価項目

主要評価項目の評価の方針

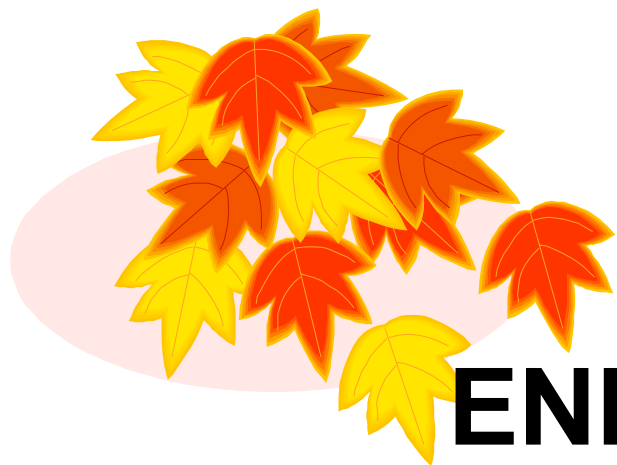
その方針のもとで適切と判断される統計手法

- ・例 個人内に繰返されるイベント治療成功の有無を個人内の相関を考慮した上でA、B2機器間で比較する。そのために、GEE手法を用いてA、B機器の③被験者特性を考慮したイベント治療成功率を推定し比較検定する。①なお、相関構造はExchangeableを仮定するが②個人内相関がマイナスとなるような場合は一般化線形混合効果モデルなど他の手法を選択する。

- ・①計算のオプションの指定によっても結果が異なることが予想される場合は、オプション名
- ・②想定する統計手法にデータに起因する問題が発生することが予想される場合は、その回避方法
- ・③表示された結果が個人内相関を考慮した上でイベント成功率の平均値であることが分かるようにすることが重要

最後に

- 今回はGEEという統計手法の話をしていてではありません。
- 個人内で繰り返しイベントを観察する場合や、個人内に複数の機器を埋植する場合は、個人内相関という現象を理解し正しく解釈しないと妥当な結論は得られないということです。
- GEE、一般化線形混合効果モデル、混合効果モデル、変量モデル、ベイジアン解析などいろいろな手法がこの問題を解決するために使われますが目的は皆同じです。統計手法の名前を聞くと頭が痛くなりますが、
- 目的を理解していれば、逆にこのような手法は複数イベントから妥当な結論を得るために使うのだと分かれば、迷うことなく試験計画や論文が作成できます。
- 細かいこと(統計手法)は専門家にまかせれば良いのです。



END Q&A time
質問、文句は古川まで
お便り待ってます
(furukawa@biostar.co.jp)

