メタアナリシスの原理と
医療機器メタアナリシス論文の読み方

株式会社バイオスタティスティカル　リサーチ
古川敏仁

2008年5月17日

メタアナリシスはどのような場合に活用されるか
医療機器分野における過去2年間の経験

- 審査資料(STED)における機器の性能評価
  →エビデンスの確立、競合他品目との比較

- 審査当局（PMDA）から提示されたメタアナリス論文への回答
  →機器の性能への疑問への回答

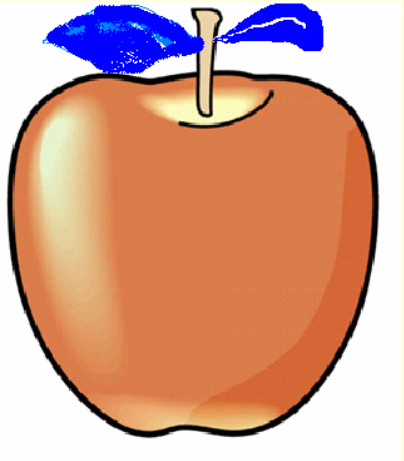
- マーケッティング部門からの競合製品との性能比較
  →エビデンスとしての信頼性

- 医師が主催する試験のコンサルティング
  例数設計、
  試験を実施する価値があるかどうかの事前情報

- 企業内勉強会、セミナー要望

第1部：
メタアナリシス
とは何か

- 複数の比較臨床試験の結果を、
- 合わせて解析し
- 治療法の比較における1試験よりも精度の高い結果を得る
- 個別試験の結果と合わせてエビデンスを検討
  - ・治療成績のばらつきがない→証拠の強さ
  - ・時代的変化における治療効果の差の検討



参考文献（Albert論文）

A Meta-Analysis of 16 randomized Trials of Sirolimus-Eluting tents Versus Paclitaxel-Eluting Stents in Patients With Coronary Artery Disease

Albert Schömig, MD, Alban Dibra, MD, Stephan Windecker, MD, Julinda Mehilli, MD,José Suarez de Lezo, MD, Christoph Kaiser, MD, Seung-Jung Park, MD, Jean-Jacque Goy, MD,Jae-Hwan Lee, MD, Emilio Di Lorenzo, MD,Jinjin Wu, MD, Peter Jüni, MD, Matthias E. Pfisterer, MD, Bernhard Meier, MD, Adnan Kastrati, MD

Journal of the American College of Cardiology Vol. 50, No. 14, 2007

今日はこの論文から、例を採用いたします

メタアナリシス

- 現在、最も証拠能力の高い結果（事実？）
- 原理：比較したい効果の差の重み付け平均
- 材料：無作為化比較試験
  - システマチックレビューにより質の高い試験を選択
  - 比較可能性のある効果の比較差しか重み付け平均する価値がない

参考文献：「メタアナリシス入門」 丹後俊郎　朝倉書店 2002

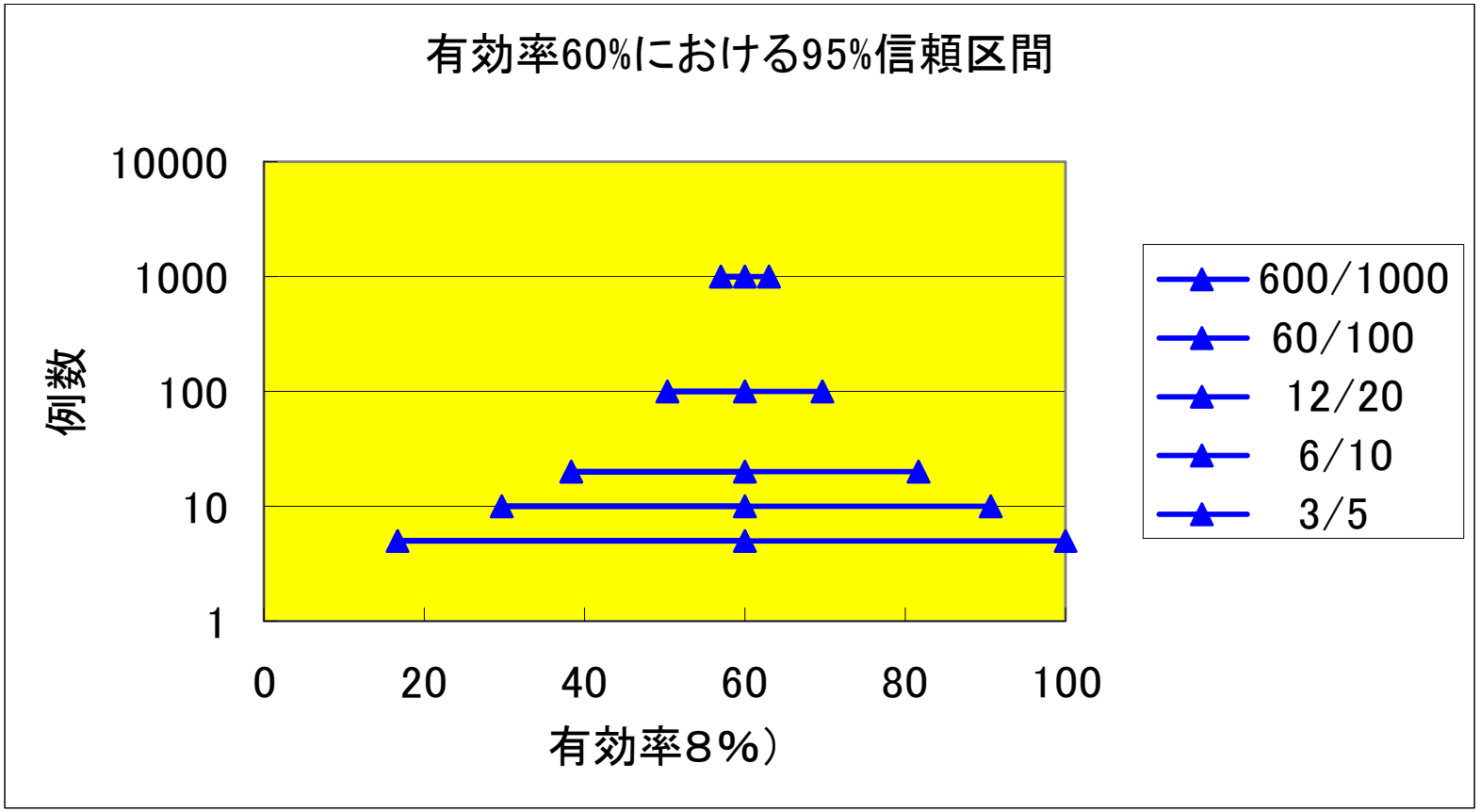
メタアナリシスの原理の説明

① なぜ、メタアナリシスすると治療群の差の精度が上がるのか？

②なぜ、重み付け平均を使用するのか？
- ②-1 個々の試験成績の単純比較がいけない理由
- ②-2 試験の症例すべてを合わせて解析してはいけない理由
- ②-3 重み付け平均の原理とは

③なぜ、メタアナリシスはエビデンスが高いのか
=エビデンスは質の高い試験の選択から

① なぜ、メタアナリシスすると治療群の精度が上がるのか？
例数と精度は比例の関係



① なぜ、メタアナリシスすると治療群の精度が上がるのか？

- 複数の試験→総合的例数が増える→精度向上

精度は推定値（差）の信頼区間の幅と考える
信頼区間幅は推定値（差）のseに比例

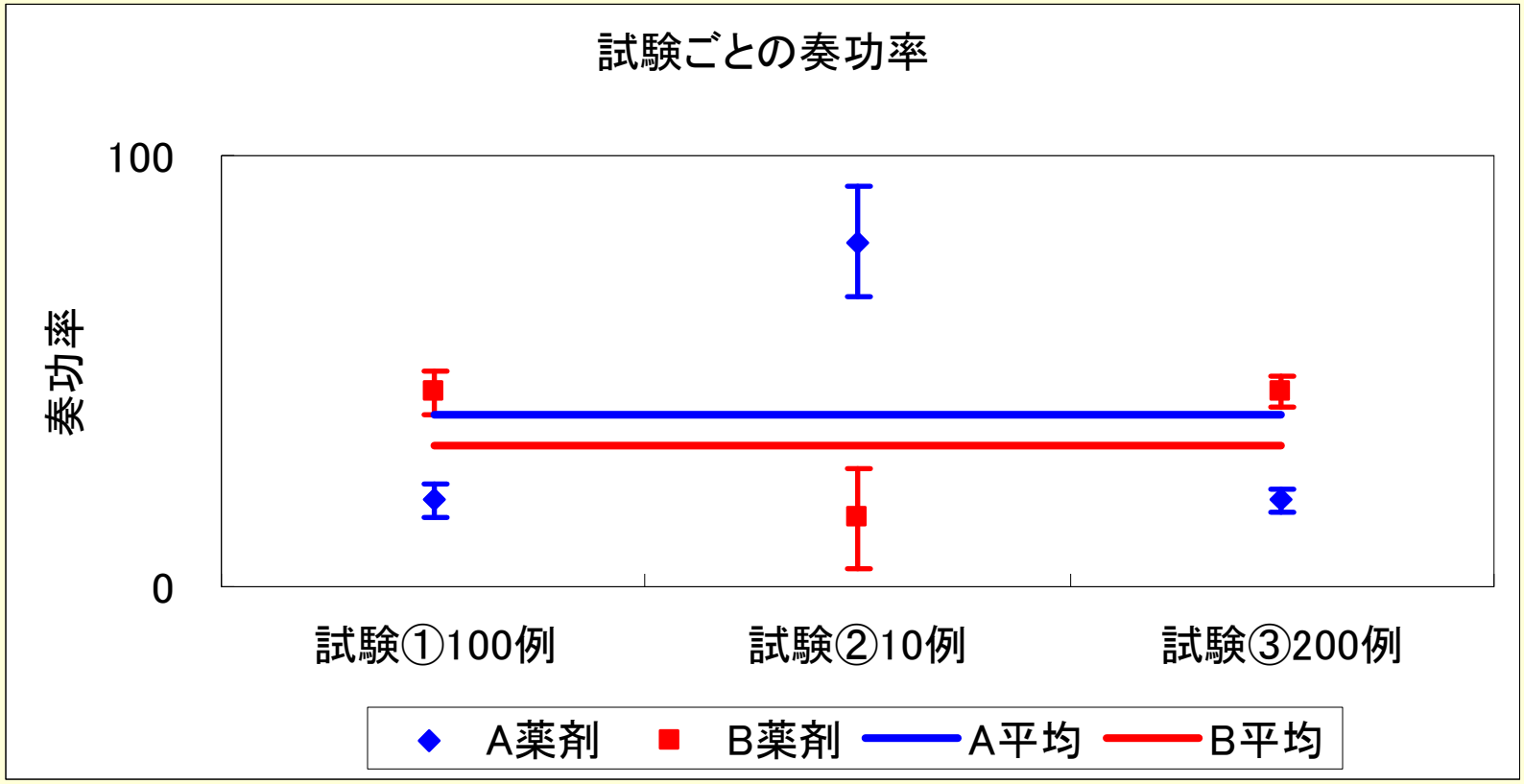
信頼区間幅＝推定値（差）±1.96･se

$$\text{se}\ (\text{標準誤差}) \propto \sqrt{\frac{1}{n_{\text{比較群}}} + \frac{1}{n_{\text{対照群}}}}$$

②なぜ、重み付け平均を使用するのか？
②-1 試験結果を単純平均するのはNo！

	A薬剤		B薬剤		B−A
	奏功率	例数	奏功率	例数	奏功率
試験①	20	20/100	45	40/100	25
試験②	80	8/10	16	2/10	−64
試験③	20	40/200	45	80/200	25
単純平均	40		35		−5

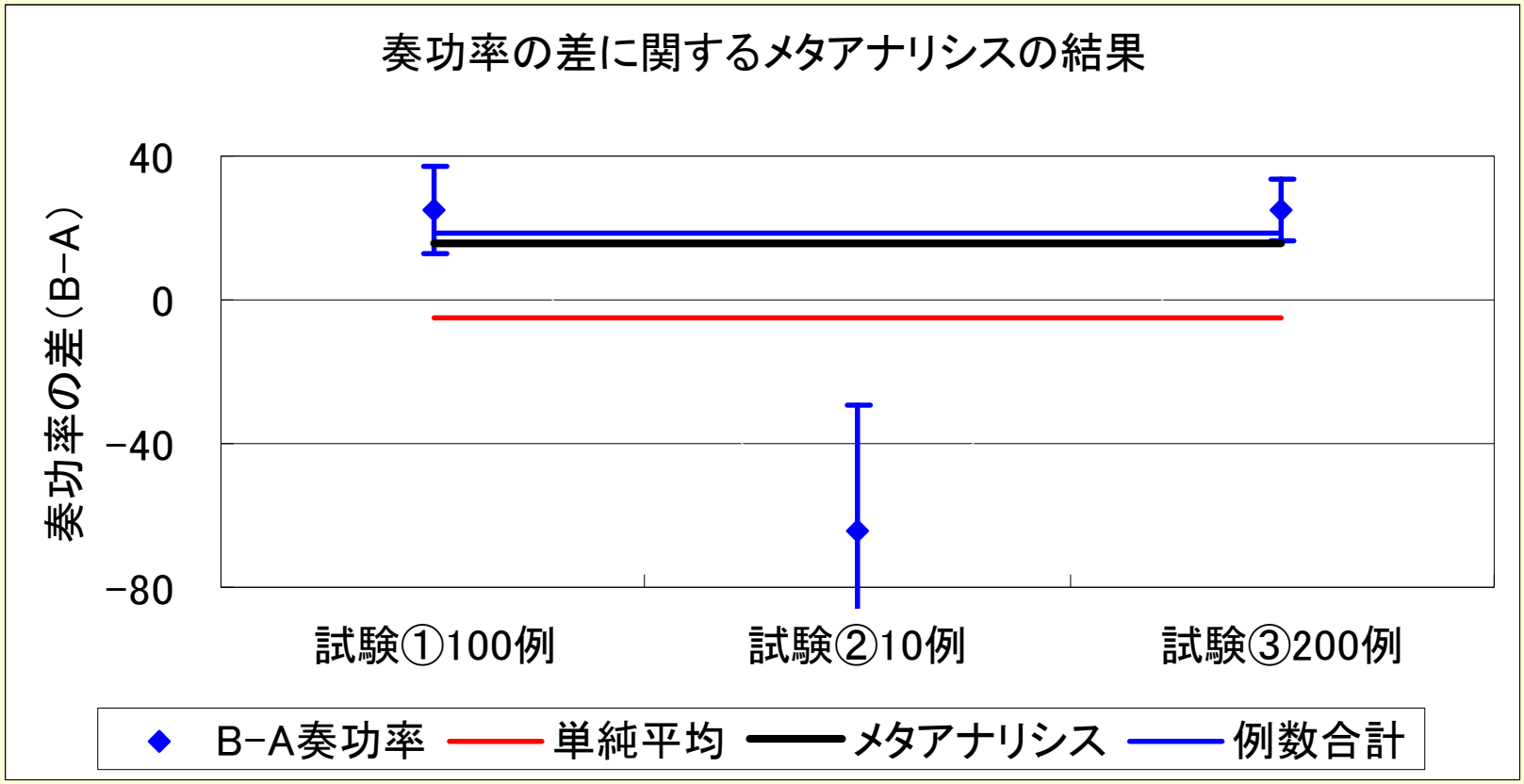
メタアナリスの原理
何が問題か？



重み付け平均計算例

	A薬剤		B薬剤		B-A			重み×
	奏功率	評価例数	奏功率	評価例数	奏功率	分散	重み	奏功率
試験①	20	100	45	100	25	40	0.025	0.625
試験②	80	10	16	10	-64	320	0.003125	-0.2
試験③	20	200	45	200	25	20	0.05	1.25
単純平均	40		35		-4.7			
メタアナリシス					16.8			21.44
例数合計差	22		39		17.4			

メタアナリシスの結果


奏功率の差に関するメタアナリシスの結果
奏功率の差（B−A）
40
0
−40
−80
試験①100例
試験②10例
試験③200例
B−A奏功率
単純平均
メタアナリシス
例数合計

メタアナリスの原理
②-1　なぜ、試験結果を単純に平均してはならないのか？回答

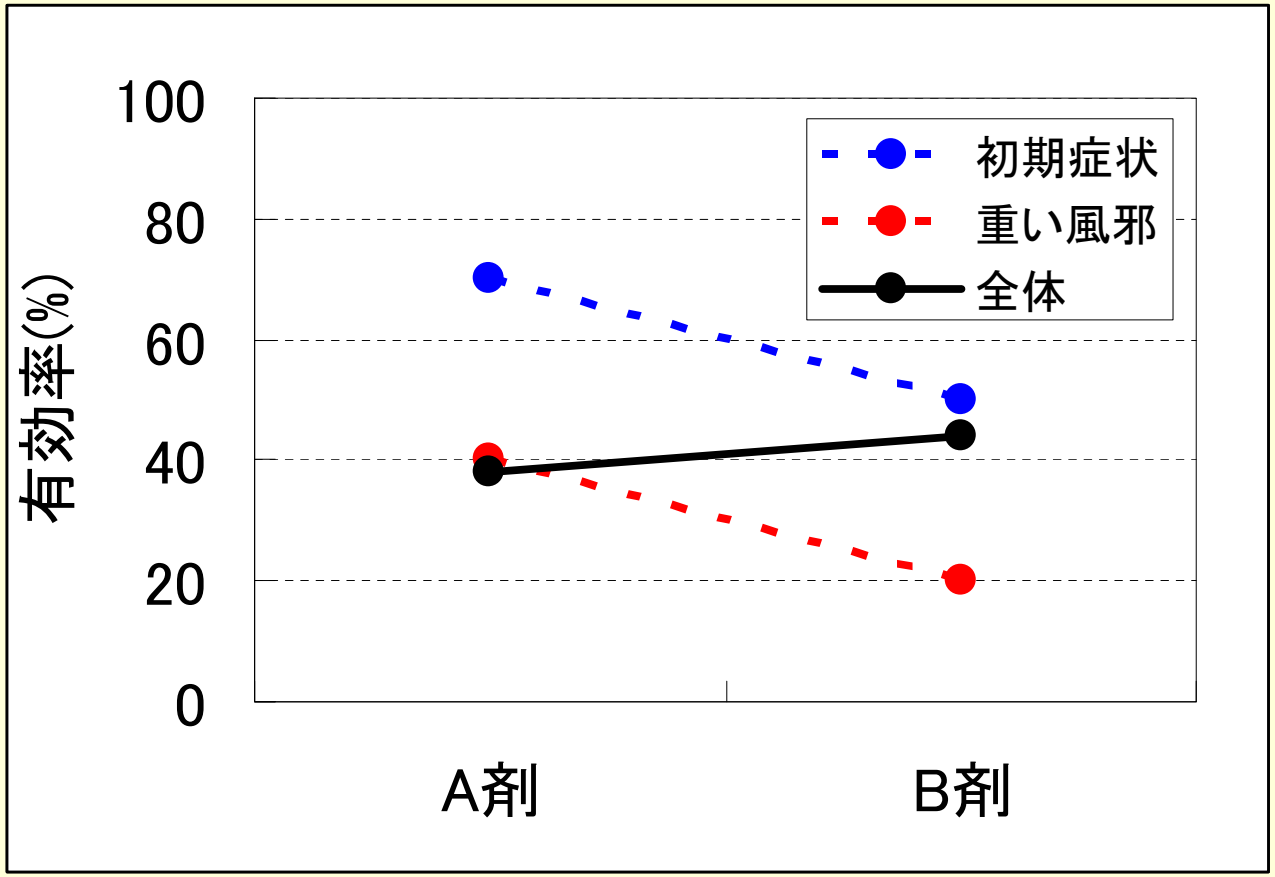
- 例数が少ない試験は信頼性が乏しい
- 信頼性が乏しい試験と、高い（例数が多い）試験を単純平均するということは、それらを平等に扱うということ
- すると、信頼性の乏しい試験の比重が重くなり、最終的に得られたメタアナリシスの結果も信頼性が低くなる。

- 通常の実験でも動物数が違う結果の単純平均など、重み付け平均しないで単純平均している例が多く見られます。重み付け平均が必要です。

②-2 なぜ、複数試験の結果を単純にまとめて解析してはいけないのか？
シンプソン・パラドックス
（Simpson's Paradox）

	風邪薬の有効率(%)				
	A剤		B剤		A−B
初期症状中心の試験	70%	(14/20)	50%	(40/80)	20
重い風邪が中心の試験	30%	(24/80)	10%	(2/20)	20
全体(合計例による解析)	38%	(38/100)	42%	(42/100)	−4
差の重み付け平均					20

シンプソン・パラドックス
（Simpson's Paradox）



メタアナリスの原理
②-2　なぜ、複数試験の結果を単純にまとめて解析してはいけないのか？　？回答

- 複数の試験の個々の症例を合わせると、比較治療群の中で集団が不均一になり、治療群間の「比較可能性」が低くなる可能性が生じる。

- 前述のシンプソン・パラドックスに代表される、応答（有効率）と治療群との関係に、他の因子（風邪の重症度）の交絡やバイアスが生じる可能性がある。

- まとめて解析すると、差の推定値の標準誤差に、試験間差の変動も含まれることになり、検定の有意性が低下する（信頼区間が広くなる）。

比較可能性

- 二群を比較しても問題がないかどうかのこと
- 比較可能性が理想な状態とは、比較したいA群、B群には、なんら**系統的**な要因の違いはなく、**偶然**な事象による違いしかないこと
- 比較可能性の中でも**時間**というのは最も重要な要因
- 治療年代が違えば、医療技術の進歩による補助的な薬剤や医療機器の違い、患者の生活環境の違い、医師のその病気に対して持っている知識の進歩

選択バイアス

- 選択バイアス
- 恣意的あるいは潜在的に治療法の評価に影響を与えるような偏りを症例選択時に与えること

- **所属集団バイアス**
- 特定の集団、例えば、非常に重症な集団、あるいは逆に軽症な集団、あるいは例外的な集団は、一般的な集団とは違った応答（outcome）を示すことにより評価が偏ること。

選択バイアス・交絡を避ける
無作為化割付

- 例数さえ多ければ、理論的に治療効果意外の背景因子のバランスが、実験群、対照群で等しくなる。

メタアナリシス

- メタアナリシスでRCT（無作為化比較試験）しか、対象としない理由はこれ
- メタアナリシスの個々の試験内でのみ、無作為化割付によって比較可能性は保たれる。
- ゆえに、試験ごとに比較可能性が保たれた治療群間の差を求め、その差を重み付け平均する（単純に症例を合わせることはできない）

②-3　重み付け平均とは　手順

① 個々の試験ごとにまず、治療群間の差とその分散（SEの2乗）を求める

②重み＝1/分散　とする。

③個々の治療群間の差に重みをかけて、総和する。

④上記の総和を、　重みの総和で割る

試験 iごとの奏功率pa_i(A群)とpb_i(B群)の差D_i を求める。

$$D_i = pb_i - pa_i$$

$$\because\ pa_i = \frac{ra_i}{na_i}, \qquad pb_i = \frac{rb_i}{nb_i}$$

i試験の重みを分散の逆数から求める

$$w_i = \left(\frac{ra_i(na_i - ra_i)}{{na_i}^3} + \frac{rb_i(nb_i - rb_i)}{{nb_i}^3} \right)$$

重み付け平均　$\overline{D}$　は下記となる。

$$\overline{D} = \left(\frac{\sum_{i=1}^{K} w_i d_i}{\sum_{i=1}^{K} w_i} \right)$$

$\overline{D}$　の分散は下記となる。

$$se(\overline{D}) = \sqrt{\frac{1}{\sum_{i=1}^{k} w_i}}$$

試験i　1～K　試験

na_i, nb_i: 試験iの各群の例数

ra_i, rb_i: 試験iの各群の有効例数

重み付け平均　計算例

	A薬剤		B薬剤		B-A		重み	重み×
	奏功率	評価例数	奏功率	評価例数	奏功率	分散	1/分散	奏功率
試験①100例	20	100	45	100	25	40	0.025	0.625
試験②10例	80	10	16	10	-64	320	0.003125	-0.2
試験③200例	20	200	45	200	25	20	0.05	1.25
合計							0.078125	1.675
メタアナリシス							0.078125÷1.675=	21.44

メタアナリシスの計算方法

- メタアナリシスの計算方法はいろいろあります　大きく大別すると
  - ○　従来の頻度論的な原理に基づく方法
    petoの方法が有名
  - ○　ベイジアン的な手法
    混合効果モデルに基づく手法とも呼ばれる

- しかし、原理はすべて重み付け平均です
- 結果もどの手法を使ってもそんなに違いません

- 重要なのは、手法よりも、臨床試験選択の問題です。

③ エビデンスの種類

- Ⅰａランダム化比較試験のメタアナリシスによる
- Ⅰｂ 少なくとも一つのランダム化比較試験による

- Ⅱａ 少なくとも一つのよくデザインされた非ランダム化比較試験による
- Ⅱｂ 少なくとも一つの他のタイプのよくデザインされた準実験的研究による

- Ⅲよくデザインされた非実験的記述的研究による。比較試験、相関研究、ケースコントロール研究など

- Ⅳ専門家委員会のレポートや意見あるいは、かつ権威者の臨床試験

- Centre of Evidence-Based Medicine 1998
  津谷喜一郎 訳

③なぜ、メタアナリシスはエビデンスが高いのか

- 当然、たくさんの試験を集めて解析するからです。
- でも、単に試験を合わせるのではだめ

- 比較可能性が保たれた質の高い試験を集めることが必要
  ①同じ評価項目　outcome（評価時期等も）
  ②無作為化比較臨床試験（比較可能性）
  ③できるだけ同じ選択除外基準
  ④できるだけ、すべての試験を（選択バイアス）
  ⑤一定の質の試験を
  - 年代
  - 試験デザイン

（Albert論文の読み方）
臨床試験の集め方

- Methods　clinical trial selection
- 選択条件：SESとPESを直接比較した臨床試験
（keyを論文に書くのが普通）

- データベース検索(2007　4月まで）

PubMed, NIH database, Cochrane Central Register of Controlled Trial, American Heart Association, American College of cardiology, European Society of Cardiology

Internet検索、参考文献リスト

論文以外でのポイント：その分野の情報を個人的に知っていることが、臨床試験収集には重要

（一方で論文収集にバイアス？）

メタアナリシスで臨床試験を収集する人の資質がすべて

スポンサー環境、個人の問題点が論文を読み取る上で重要

公表バイアス

- 臨床試験が終わり、論文かされるときに、実験治療法について有効な結果は発表されるが無効な結果は公表されないことが多い。あるいは、治験などでは依頼者の意向に沿わない結果は発表されにくい。
- たまたま自分の目的に合致した偶然の結果のみが公表され、不都合な事実は公表されないことから、メタアナリシス結果の真実性が偏る可能性

臨床試験の事前登録制

- 臨床試験の事前登録制の開始
- 対象：平成17年7月以降に開始された臨床試験、医学研究
- 臨床試験概要を公的な組織に事前登録しておかないと有力医学雑誌は受理ない。
- 受理機関：WHOが中心
  日本はUMIN、JAPIC他関連団体ごとに機関

（Albert論文の読み方）
臨床試験の選び方

- Methods　Data collection and assessment of quality
- 選択条件：
  ①　無作為化割付
  ②　outcome：死亡、MI、TLR、ステント血栓症、が揃っているもの
  ③　最終　フォローアップ期間　定義がない（論文から9ヶ月以上か）
  ④　生存時間に関しては最終確認があるもの

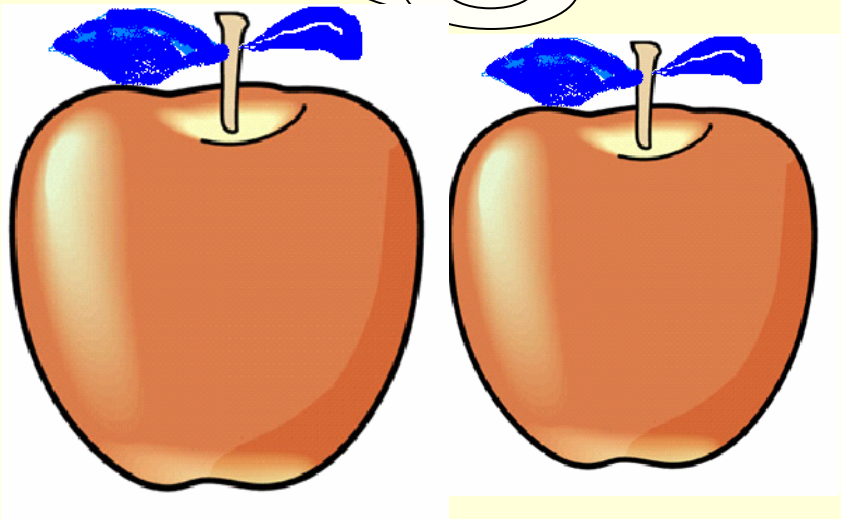
- 16の試験中　11試験で生データを試験責任医師から得た

- QC：
  ①割付の秘匿性　（封筒法は×？）
  ②解析ノ妥当性　ITT集団に対して解析されているか
  ③outcomeの第三者評価　ブラインド性

論文を読む上でのポイント：
通常、QCは数名のレビュー者が品質を得点化、質の悪い試験は採用しない
今回は、全試験採用　なぜ？
また、QCに関する記述が弱い

Q&A

第２部：ここまでで、メタアナリシスに関する基礎知識は揃いました。
次は実際に結果を読んで見ましょう。



（Albert論文の読み方）
結果を読む

- 森林プロット　forest plot

  　が一番分かりやすい

  個別試験の差の推定値（ハザード比）と信頼区間が、メタアナリシス　結果とともに表示

- メタアナリシス　結果

  ハザード比　0.74(0.63～0.87)

  $p<0.001$

  結果：SES　は　PESよりも有意に優れる。

I2情報量について

- I2情報量は個々の臨床試験結果の不均一性を測る尺度で、I^2とも記述されます。
- 試験結果の差異が偶然的な変動によるものか、それとも試験の異質性（系統的な要因）による変動なのかを表す指標です。

- I2は0を超えると、偶然ではない試験の異質性が存在する可能性を示す。
- 25%前後で小さな不均一性、I2=50%前後の場合は、中等度の不均一性、75%を超えると大きな不均一性と呼ぶ人もいる。

- I2の検定は、帰無仮説　H0：I2＝0に対する検定、すなわち治療間差の試験ごとのばらつきが偶然以外の要因によるものなのかの検定になっています。

- 今回の結果　I2=0.06　p=0.39から、試験結果は

I2情報量からいえること

- 今回の結果　I2=0.06　p=0.39から、試験結果は、
- 試験ごとに結果のばらつきはあまりなく、どの試験もSES＞PESであることが推測され（森林プロットからも裏付けられますが）

  メタアナリスの結果の妥当性を裏付けています。

I2情報量について

- I2情報量はコクランのQ統計量をもとに計算されます。
- Q統計量は個々の試験の治療間差di、メタアナリシスによって推定された治療間差dmとの差の自乗に試験の重みwiをかけた総和になっています(1)。
- 試験が多いほどこのQは大きくなってしまいますので、Q自体は解釈が難しい量です。そこで、(2)式のように変形しますと、不均一性を示す指標となります。

$$\text{Cochran's Q} = \sum_{i=1}^{K} w_i (d_i - d_m)^2 \qquad \sim \chi^2_{K-1} : \quad (1)$$

$$I^2 = 100\% \times (Q - (K - 1))/Q \qquad (2)$$

∵ K : 臨床試験の数

試験間の成績に異質性が認められた例

- この例では、I2=0.88　p<0.0001

- 試験間で結果の異質性がかなり高い
- 解釈：メタアナリシスの結果の妥当性は低い
  =今回の解析では治療群の優劣のを結論付けれない
- 原因：もともと、エンドポイントが均質ではない、対照治療が均質ではない、評価期間が均質ではない　などの問題があり比較には　限界がある。
- ただ、弱くてもよいから証拠（evidence）を求めるためこのような解析
- 解釈する人の力量が必要

- Hylan Versus Hyaluronic Acid for Osteoarthritis of the Knee: A Systematic Review and Meta-Analysis
- STEPHAN REICHENBACH, SACHA BLANK, ANNE W. S. RUTJES, AIJING SHANG,ELIZABETH A. KING, PAUL A. DIEPPE, PETER JU¨ NI, AND SVEN TRELLE

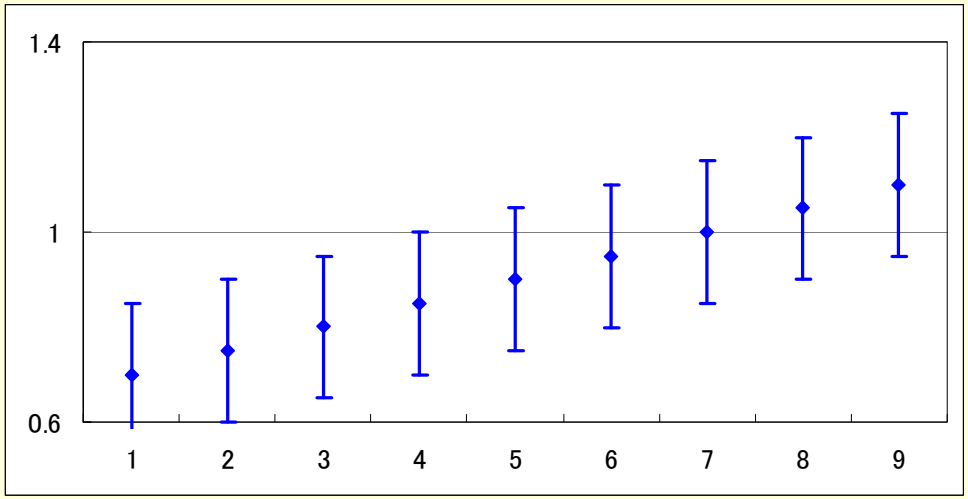
（Albert論文の読み方）
古川の解釈

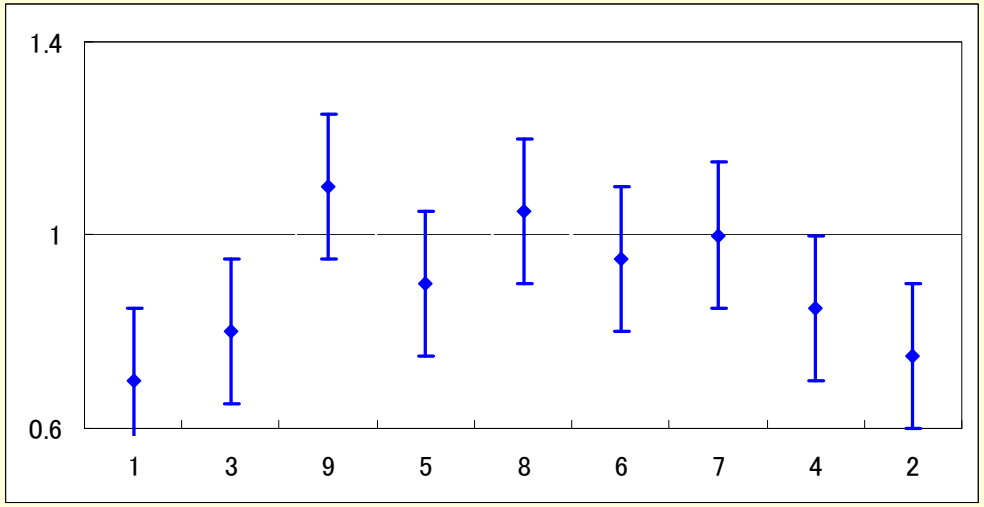
- 疑問点
- なぜ、森林プロットを試験名のアルファベット順に　年代順に並べてくれないと、時代による治療成績の推移が見れない
- また、臨床試験の年代に関する情報がない

  治療効果は、時代によってかなり変化するのに

ハザード比　0.74(0.63～0.87)　p<0.001 I2=0.06　p=0.39　試験の並び替えで結果の解釈は違う

上図：年代順　下図：作為的に並び替え　（想定）





（Albert論文の読み方）
古川の解釈

- Limitation（重要）
- 試験の選択に問題はないのか＝解析者の恣意性は排除されているか
- 試験の質に関する記述が少ないのは好意的に判断すべきか→質による試験の選択と、恣意的な選択は区別がしにくい

- 時代（前述）の問題

- メカニズム的な解釈から説明されるか　なぜ、シロリムスは血栓症が少ないのか　との整合性はあるか（勉強不足で分からない）

- TLRと血栓症は有意　死亡、MIは治療群間に差がない　合理的説明のようにも思えるが、TLRと血栓症の測定には問題がないか？

- などを結果とともに総合的に判断する

Albert論文　vs 「The Emperor's New Clothes」論文　VS　古川

① メタアナリシスに絶対はない。

② いつも、試験の収集は論議の的

③ 「優れている」、「同等」は統計学的なp値ではなく、ハザード比　0.74と相対リスク比0.89の大きさで

④ FDAは恐らく、臨床的立場から、両者は同等といっている。ただし、真の臨床的意義は、この結果から患者が決める問題

①メタ、RCT、レジストリーそれぞれの功罪
メタアナリシス(meta-analysis)

証拠能力が最も高い→EBMとして採用

→治療ガイドライン採用

→健康保険の対象

con：

① 多数のRCTが存在することが条件（時間、費用）

② 出版バイアス　着目する治療に関し、良い治療効果のみが発表され、治療効果は見かけ上実際よりも良くなる。

臨床試験登録制：出版バイアスを回避？

①メタ、RCT、レジストリーそれぞれの功罪

RCT（Randomized Controlled Trial)

メタアナリシスほどではないが証拠能力が最高い

pros：

① 比較群の間には、選択バイアスが介入しない

② 比較に対して、比較可能性の保証：制御できない背景因子も例数さえ多ければ、比較群間で均一にすることができる。

③ 治療効果の証明はこれ以外の方法では難しい

con：

① 1試験のみの結果　どのようなp値がついても偶然はありえる。

② 資源（症例数、時間、設備）が必要

③ 複数の仮説を同時に検証することはできない

④ 注意事項：無作為化の方法によって試験の証拠能力は左右される（中央登録＞封筒法）

①メタ、RCT、レジストリーそれぞれの功罪

レジストリー試験

よほど条件が整わないと製品間の性能比較はできない。希少な合併症の発見には適している。

pros：

① RCTほど資源を必要としない
② 大規模データを得ることができる
③ 日常診療に近い環境のデータ
④ 希少な合併症の発見
⑤ パイロット比較、RCTの準備のための探索的な試験

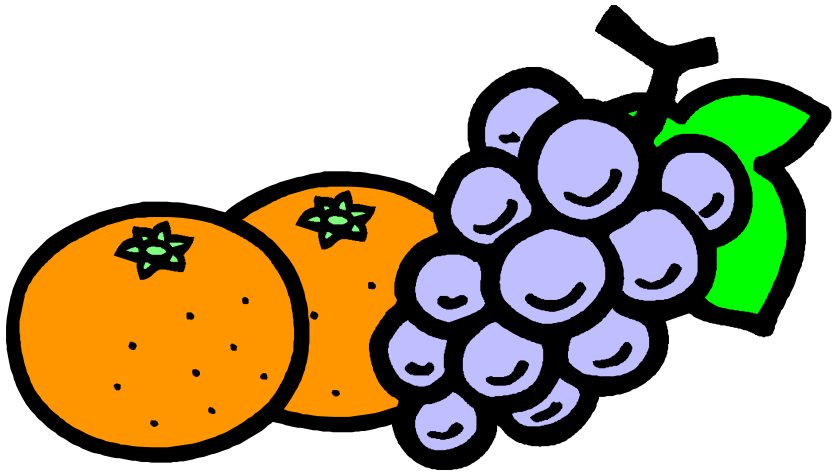
con：

① 確証論的な治療効果の証明はほとんどできない
② どのような患者集団でも、同じ治療成績が期待でき、また、評価が絶対的尺度のとき（例：OPC）しか

証拠能力が高いとは考えられない

① 対象集団全員登録であっても、症例の選択バイアスは避けられない→比較可能性が低い、一般化可能性が低く、保証が難しい

Q&Aと最後の興味ある話題



ネットワークメタアナリシス

- 個々の試験に発表されているOdds比とその信頼区間を利用して、論文中で直接比較していない治療群間の差を推定する方法
- 間接的推定-メタアナリシスとも呼ぶ

- 参考文献：Thomas Lumley　‘Network meta-analysis for indirect treatment comparisons’ Statist. Med. 2002; 21:2313–2324 (DOI: 10.1002/sim.1201)

ネットワークメタアナリシス

- 例：

試験1　A治療有効率　40%
　　B治療有効率　60%
　　B-A=20%
試験2　B治療有効率　55%
　　C治療有効率　65%
　　C-B=10%

　　C-A=（C-B）+（B-A）＝30%

　　このように、各効果の距離をモデル化、複数試験のOdds比の信頼区間を重みとしてネットワーク地図を作る手法

ネットワークメタアナリシスの可能性

- ベアメタル vs DES 、ベアメタル VS ベアメタル、DES vs DES等の試験のすべてを合わせたデータベースを作成

- すると、DES vs DES 試験から、基盤のベアメタルステントの性能が推測できる
- 少数例の比較試験でも、その信頼区間に応じたそのステントの相対性能が推測できる
- 海外試験のデータの利用、あるいは、日本独自試験の症例数を少なくすることが可能