数理解析研究所講究録 1273 巻 2002 年 78-86

臨床試験における統計的諸問題

明星大学・理工学部　広津　千尋(Chihiro Hirotsu)
Faculty of Science and Technology, Department of Mathematics
Meisei University

1．序論

臨床試験，とりわけ新薬臨床評価には統計的方法の適用や統計的考察を要する場面がいろいろある．広く応用統計学として見た場合，例えば品質管理における方法などと比べて臨床試験が特別異なっているというわけでもないが，いくつか特徴的なこともあるので以下に箇条書きで示す．

(1) 実験の構造はむしろ単純で通常 2 群比較が多い．

(2) 例えば，性，年令，重症度，起尖菌，急性・慢性，初発・再燃など，多くの共変量が存在し，これらの共変量に関し薬剤との交互作用解析や層別解析が求められる．

(3) 時間経過に従って割り付けが行われ，総例数も未定なことが多い．実際上，十分な例数の確保が難しい場合も多く，また試験期間中の脱落の多いことも特徴である．また，試験期間が長期にわたり，その間に技術および環境の変化が生じることがある．

(4) 個体差（サンプルの不均一性）が大きく，純粋な繰返しは皆無であり，偏りを防ぐための無作為化が必須である．

(5) Response(Endpoint，薬効)が多様であり，しかも線形な多変量解析はそぐわない．この項は§2 の多重性の項でも触れる．

(6) 新薬の有効性を検証するのに，優越性だけでなく，同等性（非劣性）検証が求められることがある．それは十分有用な薬剤を対照とする試験の場合，他に何らかの特長があれば，有効性に関しては同等，もしくは劣っていないことを証明すればよいという思想に基づく．ここで同等性を示す適切な統計的方法が従来考えられていないことが問題である．

2．多重性の諸問題

臨床試験では種々推論の多重性により擬陽性を招く場合が多くあり，そもそも筆者がこの分野に深くかかわるようになったのもこのトピックスによる．以下にその諸問題を箇条書きで示した上，概説する．

(1) 多種評価変数 (Multiple endpoints)

(2) 層別解析 (Subgroup analysis)

(3) 多種検定 (Different methods of analysis simultaneously applied to one set of data)

(4) 多重比較 (Multiple comparisons)

(5) 尺度合せ (Repeated χ^2 tests for an ordered $2\times K$ contingency table)

(6) 経時測定データ (Repeated measurements)

(7) 中間解析 (Interim analysis)

多種評価変数とは薬効には種々の側面があり，どの側面を見るかで結論が異なる場合を指す．とくに事後的な議論は擬陽性を生じるので避けなければならない．この問題は極く少数の主要評価変数 (primary endpoint) をその解析方法と共にプロトコルに明記するということで決着している．

層別解析とは，事後解析によって，見掛け上治験薬の有効性のとくに優れる集団を選択したり，性・年令などの層別要因と薬剤の交互作用を主張しがちなことを示す．現在では予想される交互作用をプロトコルに明記すること，また事後解析によって明らかとなった交互作用は，新たな試験で証明すべき仮説が示唆されたとする考え方が徹底してきている．例えば，痴呆患者を対象とする試験では重症度が重要な予後因子であることが分かっており，あらかじめその程度を指定した対象患者に対して試験を計画することなどは容認される．

多種検定とは，例えば順序分類データに対して Wilcoxon 検定や max χ^2 などの傾向性仮説検定や，あらかじめ定めたカットポイントで 2 分してできる分割表に対する χ^2 検定など，性能の異なる複数の検定方法（各一つひとつは論理的に正しい）を適用して，最も有意性の高かった検定法を事後的に採択することをいう．現在これは，前述の通り，主たる評価変数に対する解析手法をプロトコルに明記することで防御されている．

多重比較の問題は多群比較で生じる．複数の薬剤の同時比較試験の解析において，有意水準の調整なしに対比較を繰り返すと事実上の危険率が増大し，擬陽性が生じる．この問題にはきちんとした対応策（統計的手法）が存在し，例えばすべての対比較を行う場合は Tukey 法，標準薬と複数の治験薬に関する対比較を行うなら Dunnett 法が論理的に正しい標準的方法を与える．とくに用量反応解析のように群間に自然な順序があり，反応の単調性が仮定できる場合には，Williams 検定，その Marcus による改良法，あるいは max t 検定などを用いることができる．用量反応解析に関しては，至適用量選択を目的とする場合と，単に用量反応性（トレンド）が示されれば良いとする立場の違いからの議論もある．

尺度合せとは奇妙な表現で語源は明確でないが，先に述べた著明改善，改善，･･･，悪化のような順序分類データに基づく群比較で，事後的に χ^2 検定の有意性が最も高くなるカットポイントを選択した 2×2 分割表により有効性を主張することを指す．この問題に対しては，多重性を考慮した正確法として max χ^2 検定が導入され，解決している（広津，1992，参照）．

経時測定データとは，例えば抗高脂血症の臨床試験で，主要評価変数を体内コレステロール量として，1 ヶ月ごとに半年間 6 ポイントのデータが得られるような場合を指す．かつて，このようなデータに対して各ポイントのデータと初期値の間の対比較を繰り返すことが行われていた．その方法は多大の擬陽性を生じることが自明な上，さらに各ポイントの解析結果をどう総合評価するかという難点を孕んでいる．経時測定データに関してはその後，経時プロファイルを全体として比較するためのいろいろな統計的方法が提案されている（例えば，広津，1989；Hirotsu, 1991；丹後，1989 など）．

最後に中間解析は，癌の生存時間を対象とした試験などで日本でも用いられるようになってきた．中間解析のプロセスを明確にして正当な統計的方法を用いないと，早期に生じた偶然の好成績で誤った判断を犯す可能性のあることが問題点である．

以上のように（4），(5)，(6)，(7) は数理統計学の観点からも，解決すべき多くの興味ある問題を提供している．

3．多施設臨床試験

一般に臨床試験は複数の施設によって行われる．とくに日本の場合，非常に多くの施設を用い，各施設当たりの例数が極めて少ないことが特徴とされる．これに対しては諸外国から，ノイズが大きく薬効の検証が難しい，あるいは薬剤と施設の交互作用の大きさが推測できないとの批判がある．他方，少数の厳選された施設に例数を集中すると，得られた結論の一般化可能性に問題が生じる．これは臨床試験固有の困難な問題といえるが，この辺りの定量的な議論をきちんと行っている論文は意外に少ない．そこで表 1 に，最近 5〜6 年以内の精神疾患に関する試験報告に基づいた，施設間差の定量的分析の結果を示す．ここで，分析のためのモデルとして薬効を母数，施設を変量とする混合模型

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk},$$

を仮定する．ただし，$\alpha_1,\ldots,\alpha_a$ は処理（薬剤）効果を表し，治験薬と対照薬の 2 群比較の場合なら $a = 2$ である．$\beta_1,\ldots,\beta_b$ は施設効果を表し，b は試験に組み込まれた施設の数である．$(\alpha\beta)_{ij}$ は薬剤×施設交互作用であり，施設によって有効な薬剤が異なる程度を表す．施設効果 β_j との違いは，それが薬剤間の相対的な差は一定のまま，施設によって平均的な有効率が変動する程度を表すのに対し，交互作用は相対的な差が施設間で変動する程度を表すことにある．

各試験のデータは上段が FGIR(Final Global Improvement Rate)，下段が HAM-A，HAM-D などのスコアである．解析では FGIR の 7 段階評価に 1, 2,...,7 の数値を付与しているが，この特定の数量化が結論に大きく影響を与えることはない．表 1 には分散成分の推定値 $\widehat{m\sigma^2_{\alpha\beta} + \sigma^2_\varepsilon}$ の，誤差分散の推定値 $\hat{\sigma}^2_\varepsilon$ に対する比，および帰無仮説

$$H_0:\ \sigma^2_{\alpha\beta} = 0$$

の検定の p 値が示されている．ただし m は第 i 薬剤第 j 施設の症例数 n_{ij} の調和平均である．調和平均を採る理由は，繰返し数 n_{ij} が不揃いのモデルを，繰返し数が一定 m のモデルで近似するのに，m は算術平均ではなく調和平均とする方が良いことが知られているからである（Hirotsu, 1966，参照）．表 1 には $(m\sigma^2_{\alpha\beta} + \sigma^2_e)/\sigma^2_e$ と $(\sigma^2_{\alpha\beta} + \sigma^2_e)/\sigma^2_e$ の上側 90%信頼上限 UC1，UC2 も与えられているが，とくに後者は $n_{ij} \equiv 1$（1 施設 1 群 1 例）という極端な場合に，薬剤×施設の

表1　臨床評価の施設間変動

疾患名 (評価変数)	a	b	n	m	$m\sigma^2_{\alpha\beta}+\sigma^2_e$ $/\sigma^2_e$	p	UC1	UC2
1.*Neurosis	3	9	47	1.38	1.20	0.35	2.26	1.92
(HAM-A)	3	9	47	1.38	0.82	0.65	1.56	1.41
2.*Neurosis	3	28	156	1.56	1.47	0.06	2.06	1.68
(HAM-A)	3	29	162	1.57	1.01	0.48	1.40	1.26
3.Depression	2	36	137	1.53	0.62	0.94	0.93	0.95
(HAM-D)	2	36	137	1.53	0.62	0.94	0.93	0.95
4.*Depression	3	20	96	1.34	1.40	0.16	2.14	1.85
(HAM-D)	3	20	95	1.34	1.02	0.48	1.56	1.42
5.Depression	2	45	192	1.63	1.05	0.41	1.49	1.30
(HAM-D)	2	42	178	1.63	1.42	0.09	2.03	1.63
6.Schizophrenia	2	25	132	2.08	1.44	0.11	2.29	1.62
(BPRS)	2	25	133	2.09	1.58	0.07	2.51	1.72
7.Schizophrenia	2	32	141	1.55	0.91	0.60	1.38	1.24
(BPRS)	2	32	141	1.55	0.50	0.98	0.76	0.85

UC1：$\left(\sigma^2_{\alpha\beta}+\sigma^2\right)/\sigma^2$ に対する90%信頼上限

UC2：$\left(\sigma^2_{\alpha\beta}+\sigma^2_e\right)/\sigma^2_e$ に対する90%信頼上限

p ：帰無仮説 $H_0:\sigma^2_{\alpha\beta}=0$　検定のための p 値

* ：Ⅱ相試験

交互作用に依る誤差分散の増大の程度を与える．表1の例でそれは高々1.9程度であり，それは患者スコアの母集団分布範囲が±10の時，交互作用によるばらつきの増大が高々±14程度であることを示している．また，$(m\sigma^2_{\alpha\beta}+\sigma^2_e)/\sigma^2_e$ の点推定値が1.5程度であることは，これらの試験では極めて均質な施設による試験に対し総症例数を50%増やすことにより，薬効差検出力低下を補うことが出来ることを示す．降圧剤や，抗アレルギー剤，抗炎症剤などではこれより少なく総症例数10～20%増が示唆される．ここで交互作用の解釈が母数模型とは異なることに注意を要する．すなわち，混合模型による解析は交互作用を越えた主効果（薬効差）の検証という実践的なアプローチに相当する．この時，20～50%の例数増は見合う代償とは言えないだろうか．

一方，国際的なガイドラインでは，母数模型を仮定し，しかも交互作用より先に主効果を検定するという通常とは異なる手順が述べられている．母数模型の場合，主効果は交互作用に依存し，交互作用が存在する時には主効果の定義に恣意性が入る．従って，母数模型の場合は交互作用を先に検定し，それが検出されたならその解釈に進み，主効果の検定は行わないというのが本来の

手順である.

4．同等性検証

序論で述べたように，固有な問題として同等性検証があるが，実際には治験薬が標準薬に対してある程度（以下，Δで表す）以上劣っていないという非劣性検証が行われている．また，優越性検証は両側，非劣性検証は片側対立仮説を想定し，前者は有意水準0.05，後者は0.025で行うという，論理的に説明のつかないことが前述のガイドラインに述べられており，やや混乱している．実は優越性検証と非劣性検証，また両側仮説と片側仮説を一つの統計的推測にまとめる興味ある統計的論理が存在するので以下に紹介する．

いま，治験薬の有効率をp_t，対照薬のそれをp_cとしてp_t-p_cのパラメータ空間を次のように分割する，

$$H_1 : p_t - p_c > 0,$$
$$H_2 : p_t - p_c = 0,$$
$$H_3 : p_t - p_c < 0.$$

H_1は治験薬の対照薬に対する優越性，H_2は同等性，そしてH_3は対照薬の優越性を表す．
さらに治験薬の対照薬に対する許容範囲Δを越えた劣性に対応する領域をH_4とする，

$$H_4 : p_t - p_c < -\Delta .$$

このとき，H_4がH_3に含まれることから，

$$H_4 \cap H_3 = H_4,\ H_4 \cap H_1 = \phi,\ H_4 \cap H_2 = \phi$$

が成り立つ．ただし，ϕは空集合を表す．ここで，H_1，H_2，H_3，H_4を帰無仮説とする検定をすべて同一の有意水準αで行うが，閉手順(Marcus et al., 1976)に従い，まずH_4を検定する．これは通常の許容範囲Δの非劣性試験の片側検定に当たる．H_4が棄却されなければ，治験薬剤の対照薬に対する非劣性は主張できないことになる．H_4が棄却された場合のみ次の検定に進むが，この方式を要約すると次のようになる．

(1) H_4が棄却されなければ非劣性の主張はできない．検定の手順はここで終わる．

(2) H_4が棄却され，H_3（優越性の片側検定）が棄却されなければ，許容範囲Δの非劣性$p_t-p_c \geqq -\Delta$までが主張できる．

(3)もしH_3が棄却され，H_2（通常の優越性両側検定）が棄却されなければ，同等以上$p_t-p_c \geqq 0$が主張できる．

(4)もし，H_2も棄却されれば，優越性$p_t-p_c > 0$が主張できる．

以上の検定をすべて同一有意水準αで行った場合，上記一連の推論の危険率はαで押さえられている．言い換えると，主張の信頼係数が$1-\alpha$であり，これは極めて明解といえる．すなわち，あらかじめ検定する仮説を決めておく必要がなく，事後的な手続きとして正当化されるのである．もちろん，検定の替りに，90%および95%信頼区間の組合せで議論することも可能である．なお，これ以外の検定の組合せがMorikawa and Yoshida(1995)に述べられている．

最後に，非劣性検証では倫理性も問題となる．従来の厚生省ガイドライン(1992)は，非劣性検証は有効性以外の他の側面で治験薬に特別な長所のあることを前提としている．ところが，ICH

E9 ガイドラインではそれが有効性検証の枠内のみで論じられているため，当然試験の倫理性を問題視する声も上っている（広津他，1999）．しかしながら，ここで対する優越性試験の困難さが問題となる．例えば，p_t=0.75，p_c=0.70 の場合に，検出力 0.8 で優越性試験を計画すると α=0.05（両側）に対し，1 群必要症例数は 1251 となり，やや過度の要求に見える．一方，Δ=0.1 の非劣性試験とすると片側 α0.05 および 0.025 それぞれの場合に，必要症例数は 110 および 139 と合理的な数になる．従って現実的には，$p_t - p_c$ のそう大きくない実薬対照試験としては何らかの非劣性試験を考えざるを得ないだろう．その場合，倫理性に対する一つの弁解は，きちんとした非劣性検証なら，薬剤の平均出検品質（品質管理用語，実際に試験をパスして市場に出ていく薬剤の有効率）は $p_c - \Delta$ より相当大きくなるはずだということであろう．

5．医師による総合評価(主観評価)の問題

ある一人の患者の臨床評価を考えた時に，それは本質的に多変量的であり，単一の，それも計量的な主要評価変数の得られることは稀である．そこで日本では従来，医師による総合判定が主要評価変数として用いられてきた．しかし，最近では，より客観的に測れる生化学的測定値や HAM-D，BPRS，DIEPSS など世界的に容認されている評価尺度を用いるべきという主張が強く，計量値に対する信仰すら感ぜられることがある．そこで，以下に客観指標，主観指標の長所・欠点を論じておく．

5－1 客観的指標の問題点

(1) 線形性・加法性・正規性

臨床試験に用いられる変数は，例え計量値であっても線形性・加法性・正規性を満たさない場合が多く，適切な非線形変換を施さないと解釈を誤り易い．そして，何が適切な非線形変換であるかは一般に自明でない．代表的な計量値と思われる血圧ですら，同じ 30mmHg を初期値 280mmHg から下げるのと 180mmHg から下げるのでは意味が異なるし，(5)で述べるような J-shape 問題も存在する．さらに，統計的手法の中には，臨床評価ではおなじみの外れ値に対しあまり頑健でなく，誤った結論を導き易いものもある．そもそも，臨床評価変数は対数正規分布に従うことが多く，正規分布を基準として見ると外れ値とみなされるものでも，決して外れ値ではないことが多い．比較的少数例の試験で，一見外れ値と見えるデータを除外した議論をよく見かけるが，これはミスリーディングであり，むしろ変動係数の大きな対数正規分布を前提とした計画を立て，解析すべきである．一見外れ値と見えるようなデータが，実は例数を増やしていくと結構頻繁に現れるということが珍しくないのである．

(2) 測定誤差

計量値といえどもいろいろな種類の誤差要因に曝されている．まず患者の個人変動として，例えば，血圧の日内変動，白衣高血圧，食事によるコレステロール変動などがある．次に血圧や骨量のように本来正確な測定器に依る場合でも，実験者や測定器具，測定施設に依る変動は予想外

に大きい．従って，これら計量値を信頼性あるものにするために，測定の標準化や訓練は欠かせない．

(3) 総合化

計量測定値は一般に多変量データとして得られる．例えば典型的な計量値である血圧ですら，収縮期血圧，拡張期血圧，両者の平均，朝・昼または夜間血圧，血圧の変化量または変化率，結果として正常値に戻せたか否かの 2 値情報など多彩である．かつて主要評価変数として trough-peak ratio が推奨されたことがあったが，本質的な問題点が指摘され，今はむしろ単純な変化量が降圧効果として採られることが多いように思われる．しかし，それも(1)項および(5)項に述べるような問題点を避けられず，今なお混乱が見られる．他の例として，高脂血症でも，LDL，non HDL=TC-HDL あるいは TC/HDL などいくつかの提案があり意見が分かれている．抗鬱病の評価尺度にしても，異なった反応性の評価点数を単に合計した総点に関しては意義が唱えられている（石郷岡，1999）．このように，今，各疾患分野ごとに多変量データをどう総合化して主要評価変数とするかの議論と合意が求められている．

(4) 経時測定データに基づく比較

血圧，コレステロール量，骨量などは数ヶ月あるいは数年の経時データとして得られる．しかるに通常よく見られるのは初期値と最終値のみに基づく解析であり，より精緻なプロファイル解析が望まれる（例えば，Hirotsu, 1991, 参照）．

(5) 臨床目的との整合性

計量値に価値があるとしたらそれは，それが客観指標だからというよりは，臨床目的によく整合している場合である．従って，計量的な代用特性が臨床試験の主要評価変数として真に有用であることを主張するためには，それを大規模臨床試験によって確認する必要がある．例えば，拡張期血圧の虚血性心疾患に対する影響として，85～70mmHg に最適値があるとする J-shape 効果の議論は未だに決着していないようである(Cruick, 1987；後藤，守谷，1997)．この点が明確にされないと，降圧効果を主要評価変数とすることの是非を論じることができない．他の例として，骨代謝マーカーと将来の骨折リスクの関係などが挙げられる．

5－2 医師による総合評価の長所と欠点

ICH E9 ガイドラインでは，場合によって医師による総合評価も全般改善度，安全度，有用性の測度として容認し，あるいは奨めてもいる．しかしながら日本では，過去における全般改善度の濫用に対する反動もあって，使用をむしろ差し控える傾向にある．このタイプの変数は客観指標に，患者の状態変化に対する医師の全体的な印象を加味して構成され，必然的に主観的要素を含んでいる．そこで今が，いつ，どこで，どのように総合評価を用いるべきかを真剣に議論すべきと思われる．例えば，ガイドラインにも述べられているように，その総合評価がどうその試験の臨床目的と関連するか，また，種々の計量測定値と医師の印象とをどう結び付けて一つの総合

指標にするかのプロセスをプロトコルで明確にしておく必要がある．しかしながら，これを過度に要求することは，本来，主観評価を場合によって価値あるものと認めていることと矛盾するように思われる．

このような総合指標の長所は次のようにまとめられるだろう．

(i) 数学的な多変量解析による要約統計量とは異なる一つの臨床的総合変数であり，Quality of Life の原型とも見なされる．

(ii) 非線形な臨床効果もある程度評価することができる．

(iii) 患者の初期症状やいろいろな個人情報，そして時間的推移が判断に組み入れられる．

(iv) 臨床評価にはつきものの欠測値があっても，ある程度評価可能である．

この種の変数ではばらつきが増大することは避けられないが，標準化や訓練によってある程度防ぐことができる．実際表 1 の例でも，医師による全般改善度と BPRS のような評価尺度で交互作用の大きさに大した差は見られない．そこで，このような指標をもっと使い易くするためのガイドラインの作成が望まれる．

データに基づいて，総合評価が他の客観的測定値とどの程度相関し，合理性があるかをチェックすることも必要である．その方法はいろいろ考えられるがここでは紙面の都合上割愛する（例えば広津，1995）．

6．その他

その他の問題として，

(1) プラセボ（偽薬）使用の問題，

(2) 海外データ利用（Bridging）における人種差の問題

などがあるが，これらについては当日議論した．

参考文献

Cruickshank, J.：Benefit and potential harm of lowering high blood pressure. Lancet, 1: 581-584, 1987.

後藤　英司，守谷　昭彦：J-曲線仮説と虚血性心疾患発証．循環器 Today, 1: 312-316, 1997.

Hirotsu, C.：An approach to comparing treatments base on repeated measures. Biometrika, 78: 583-594, 1991.

広津千尋：経時測定データの解析のためのモデルとその応用．日本品質管理学会誌，19, No.3, 16-23, 1989.

Hirotsu, C.：An approach to comparing treatments base on repeated measures. Biometrika, 78: 583-594, 1991.

広津千尋：実験データの解析－分散分析を越えて－．共立出版，東京，1992.

広津千尋：比較臨床試験解析－最近の話題から－．臨床評価，23: 489-521, 1995.

広津千尋, 栗原雅直, 清水直容他: 臨床試験のための統計的原則適用に関する座談会. 臨床評価, 27: 13-66, 1999.

石郷岡　純 : プラセボ比較試験の問題点. 臨床精神薬理, 2: 145-153, 1999.

厚生省 : 臨床試験の統計解析に関するガイドライン : 1992.

Marcus, R., Peritz, E. and Gabriel, K. R. : On closed testing procedures with special reference to ordered analysis of variance. Biometrika, 63: 655-660, 1976.

Morikawa, T. and Yoshida, M. : A useful testing strategy in phase III trials: combined test of superiority and test of equivalence. J. Biopharmaceutical Statistics, 5: 297-306, 1995.

丹後俊郎 : 臨床試験における経時的測定データの解析のための混合分布モデル. 応用統計学会誌, 18, N0.3, 143-161, 1989.

なお, 本論では極めて簡略化して概要を述べた. 臨床試験にかかわる統計的方法についてより詳しくは

1. Hirotsu, C. (1993). Beyond analysis of variance techniques : Some applications in Clinical trials. Int. Statist. Review 61, 183-201.

2. 広津千尋 (1993). 臨床試験の統計学的側面. 日本統計学会誌 22, 475-492.

3. Hirotsu, C. (2002). Multiplicity problems in the clinical trial and some statistical approaches. Proceedings of Int. Symposium on Advanced Methods in Statistics, Inst. Statist. Math., 1-26.

を参照されたい. とくに3. は多重比較にかかわる問題を中心に扱っている. また, より philosophical な側面については

広津千尋 (2001). 科学技術としての統計的方法. 臨床精神薬理 4, 763-773,

およびその参考文献を参照されたい.