2. Daniels S, Bhattacharrya S, James A, et al. A genome-wide search for quantitative trait loci underlying asthma. Nature 1996: 383: 247–50.
3. Laitinen T, Daly MJ, Rioux JD, et al. A susceptibility locus for asthma-related traits on chromosome 7 revealed by genome-wide scan in a founder population. Nature Genet 2001: 28: 87–91.
4. Haagerup A, Bjerke T, Schiotz PO, Binderup HG, Dahl R, Kruse TA. Asthma and atopy – a total genome scan for susceptibility genes. Allergy 2002: 57: 680–6.
5. Cookson WO, Young RP, Sandford AJ, et al. Maternal inheritance of atopic IgE responsiveness on chromosome 11q. Lancet 1992: 340: 381–4.
6. Moatt MF, Sharp PA, Faux JA, Young RP, Cookson WO, Hopkin JM. Factors confounding genetic linkage between atopy and chromosome 11q. Clin Exp Allergy 1992: 22: 1046–51.
7. Shirakawa T, Deichmann KA, Izuhara K, Mao I, Adra CN, Hopkin JM. Asthma, atopy and genetic variants of IL-4 and IL-13 signalling. Immunol Today 2000: 21: 60–4.
8. Shirakawa T, Li A, Dubowitz M, et al. Association between atopy and variants of the subunit of high affinity immunoglobulin E receptor. Nature Genet 1994: 7: 125–30.
9. Hook S, Cheng P, Holloway J, et al. Analysis of two IL-4 promoter polymorphisms in a cohort of atopic and asthmatic subjects. Exp Clin Immunogenet 1999: 16: 33–5.
10. Morahan G, Huang D, Wu M, et al. Association of IL12B promoter polymorphism with severity of atopic and non-atopic asthma in children. Lancet 2002: 360: 455–9.
11. Tanaka H, Miyazaki N, Oashi K, et al. IL-18 might reflect disease activity in mild and moderate asthma exacerbation. J Allergy Clin Immunol 2001: 107: 331–6.
12. Arbour NC, Lorenz E, Schutte BC, et al. TLR4 mutations are associated with endotoxin hyporesponsiveness in humans. Nature Genet 2000: 25: 187–91.
13. Van Eerdewegh P, Little RD, Dupuis J, et al. Association of the ADAM33 gene with asthma and bronchial hyperresponsiveness. Nature 2002: 418: 426–30.
14. Romagnani S: Immunologic influences on allergy and the TH1/TH2 balance. J Allergy Clin Immunol 2004: 113: 395–400.
15. Liu AH. Endotoxin exposure in allergy and asthma: reconciling a paradox. J Allergy Clin Immunol 2002: 109: 379–92.
16. Hershey GK, Friedrich MF, Esswein LA, Thomas ML, Chatila TA. The association of atopy with a gain-of-function mutation in the α subunit of the interleukin-4 receptor. N Engl J Med 1997: 337: 1720–5.
17. Heinzmann A, Mao X-Q, Akaiwa M, et al. Genetic variants of IL-13 signalling and human asthma and atopy. Hum Mol Genet 2000: 9: 549–59.
18. Gao P-S, Mao X-Q, Roberts MH, et al. Variants of STAT6 (signal transducer and activator of transcription) in atopic asthma. J Med Genet 2000: 37: 380–2.
19. Hobbs K, Negri J, Klinnert M, Rosenwasser LJ, Borish L. Interleukin-10 and transforming growth factor-β promoter polymorphisms in allergies and asthma. Am J Respir Crit Care Med 1998: 158: 1958–62.
20. Rosenwasser LJ, Borish L. Genetics of atopy and asthma: the rationale behind promoter-based candidate gene studies (IL-4 and IL-10). Am J Respir Crit Care Med 1997: 156: S152–5.
21. Lech-Maranda E, Baseggio L, Bienvenu J, et al. Interleukin-10 gene promoter polymorphisms influence the clinical outcome of diffuse large B-cell lymphoma. Blood 2004: 103: 3529—34.
22. Walley AJ, Cookson WO. Investigation of an interleukin-4 promoter polymorphism for association with asthma and atopy. J Med Genet 1996: 33: 689–92.
23. Noguchi E, Shibasaki M, Arinami T, et al. Association of asthma and the interleukin-4 promoter gene on Japanese. Clin Exp Allergy 1998: 28: 449–53.
24. Risma KA, Wang N, Andrews RP, et al. V75R576 IL-4 receptor alpha is associated with allergic asthma and enhanced IL-4 receptor function. J Immunol 2002: 169: 1604–10.
25. Woitsch B, Carr D, Stachel D, et al. A comprehensive analysis of interleukin-4 receptor polymorphisms and their association with atopy and IgE regulation in childhood. Int Arch Allergy Immunol 2004: 135: 319–24.
26. Nishiyama C, Akizawa Y, Nishiyama M, et al. Polymorphisms in the Fc epsilon RI beta promoter region affecting transcription activity: a possible promoter-dependent mechanism for association between Fc epsilon RI beta and atopy. J Immunol 2004: 173: 6458–64.
27. Hopp RJ, Bewtra AK, Watt GD, Nair NM, Townley RG. Genetic analysis of allergic disease in twins. J Allergy Clin Immunol 1984: 73: 265–70.
28. Duffy DL, Martin NG, Battistutta D, Hopper JL, Mathews JD. Genetics of asthma and hay fever in Australian twins. Am Rev Respir Dis 1990: 142: 1351–8.
29. Nieminen MM, Kaprio J, Koskenvuo M. A population-based study of bronchial asthma in adult twin pairs. Chest 1991: 100: 70–5.
30. Fiorentino DF, Bond MW, Mosmann TR. Two types of mouse T helper cell: Th_2 clones secrete a factor that inhibits cytokine production by Th_1 clones. J Exp Med 1989: 170: 2081–95.
31. Iikura Y, Iwasaki A, Tsubaki T, et al. Study of liver function in infants with atopic dermatitis using the ^{13}C-methacetin breath test. Int Arch Allergy Immunol 1995: 107: 189–93.
32. Qian S, Lu L. Mechanisms of hepatic tolerance. In: Gerchwin ME, Vierling JM, Manns MP, eds. Liver Immunology. Philadelphia: HANLEY &BELFUS, INC, 2003: 115–28.
33. Thomas NS, Wilkinson J, Holgate ST. The candidate region approach to the genetics of asthma and allergy. Am J Respir Crit Care Med 1997: 156: S144–51.
34. Nishiyama C, Akizawa Y, Nishiyama M, et al. Polymorphisms in the FcεR1β promoter region affecting transcription activity: a possible promoter-dependent mechanism for association between FcεR1β and Atopy. J Immunol 2004: 173: 6458–64.

医薬統計の原理・原則

吉村　功
東京理科大学

1．動機

学術雑誌に自分の研究結果を掲載してもらおうとする研究者は，論文原稿をその雑誌に投稿して審査を受ける．雑誌の編集責任者は，あるレベル以上と思われる関連分野の専門家にその審査を依頼する．依頼された専門家（査読者）は，(1) 採択が妥当である，(2) 棄却が妥当である，(3) 適切な改訂が行われれば採択が妥当である，という三つの選択肢の一つを選ぶ．

(3) を選んだ場合査読者は，必要な改訂の指示・指摘を査読結果として提出する．編集責任者は論文原稿の著者にこの指示・指摘を伝える．著者は指示・指摘が妥当であれば改訂を行い，妥当でなければ反論を行う．編集責任者はその改訂・反論の妥当性を査読者と共に検討し，最終的な採択・棄却の判断を下す．これが学術雑誌の通常のやり方である．

私は数人の共同研究者と共著で，ある雑誌に「複数の主要評価変数を想定した臨床試験における被験者数設定法」についての論文原稿を投稿した．この投稿に対して，しばらく経ってから，「... 相間の正確な見積りは難しいと考えられます．だとすればよく行われているように相間 0 の前提下に症例数を見積もるのが，一般に安全で賢明なやりかただと考えられます.」という指摘が編集責任者から送られてきた．われわれの研究によると，相関を 0 にすると，そうでない場合に比べて被験者数が多くなる．従ってこの指摘は「被験者は多めに評価するのが安全で賢明」と述べていることになる．

われわれは，この指摘を，被験者数設定の原理・原則に反するものと考え，反論を行った．結果としてそれは受け入れられたが，査読者というものは，それなりのレベルの専門家である．それが原理的に妥当でない判断を出してくるのには，背景に，この考え方を当然とする多くの人がいることを意味している．だとすれば単に個別に反論するのではなく，公の場で一般論として，原理・原則を確認する機会を用意した方が良いのではなかろうか．これが今回，この課題を設定した動機である．原理・原則というものは，それをお互いが納得した上で，実用上の柔軟性をどのように取り入れるかを考えるべきだと思うからである．

2．想定する状況と用語法

短い時間で実りある議論をしたいので，ここでは話題・課題をある程度限定する．用いる用語もそれに合わせて限定することにする．私が近年関係している分野は医薬統計なので，課題を医薬統計に設定する．主たる対象は医薬品であるが，毒性試験等では化粧品や環境汚染物質などにも言及する．

医薬品についての一般原則は，「薬としての価値がある物は可能な限り早く，安い費用で，患者と医師の手元に届け，薬としての価値がない物は患者と医師が使わない状況を作る」ことである．化粧品や環境汚染物質などについての一般原則は，危険と利得のバランスという視点ではなく，「社会的に受け入れられる水準を超える危険性を持つものは規制する」ことである．

統計学ではサンプルサイズと言われているものが，医薬統計では慣用的に「症例数」（ヒトの場合），「動物数」「例数」（動物の場合），「n 数」（インビトロ試験等）と言われている．本稿では，ヒトが対象のときに「被験者数」，動物個体が対象のときに「動物数」，インビトロ試験での繰り返し数の類のものを「サンプルサイズ」と言うことにする．

薬事法では，「薬」という用語をかなり限定したものとして使っている．それをそのまま使うと，開発中の薬の候補物には薬という用語が使えない．それは不便なので，ここでは「被験薬」「被験物」「薬物」という用語を慣例に合わせて使う．

以上の前提の下では，薬効も毒性も，複数の区分での評価変数・反応の比較という形で評価がなされる．この区分を「群」と言うことにする．群がある被験物・薬のある用量と対応している試験を「用量群試験」と言うことにする．

3．臨床試験と非臨床試験

医薬統計では，臨床試験と非臨床試験には大きな状況の違いがある．ヒトと非ヒトとの間の倫理面での重要度の違いと，被験者・物の個体差・個人性の認識の仕方の違いである．

倫理面について言うと，臨床試験も非臨床試験も，実質は実験であるから，実験台として生体に損害・ダメージが生じる可能性が存在する．その損害が致命的な場合，対象がヒトであれば申し訳ないと頭を下げるだけで済まないが，動物であればやむを得ないとして済ますことになる．倫理面の重要性に質的な差があるというのは社会の共通了解であろう．

個体差について言うと，ヒトの個体差は制御不可能な形で現れ，しかもその一人一人の存在を個別なものとしなければならない．これに対して，非ヒトの個体差はヒトに比べてはるかに小さくて均一的であり，しかも多くの場合ある程度制御可能である．認識においても個体差を無視した集団現象が検討対象になる．

上記２点は，原理・原則論議でも異なった側面をもたらす．今回の例会で両者を臨床の部と非臨床の部に分けたのはそのためである．

4．被験者数の設定

被験者数設定の原則は，「主要評価変数での検証が可能という条件の下で可能な限り人数を少なくすること」につきる．この原則は，検証の精度を上げるには被験者数が多い方が良いが，危険がある臨床試験では危険を冒す人すなわち被験者を少なくする方がよい，という矛盾した要求の妥協として，被験者数を必要最小限にするという原理を立てたことによる．

この原理・原則は現在，総論として多くの人に受け入れられている．それにもかかわらず論議が無くならないのは，必要な精度を定めるための情報が臨床試験を始める時点で不十分であることと，ほとんどの場合，新薬開発に利害のある製薬企業（あるいはそれと密接な関係者）が臨床試験の実施者であるためである．

想定しているモデルの下で検出力を計算しようとしたとき，たとえモデルが妥当だとしても検出力はモデルに含まれる局外母数の値によって変動する．それにもかかわらず局外母数の値を十分な精度で推定することは困難な場合が多い．その結果，曖昧さの残る被験者数の見積もりに対して，どの値を採用するかが臨床試験の実施者に委ねられる．そのときに，実施者の「願望」「先入観」「経験的予断」が「安全で賢い選択」「保守的な選択」を招き，結果としてコストと時間が許される範囲において大きめの被験者数を定める傾向を産むことになる．

臨床試験の実施者がすでに数百億円の開発投資をしていて，しかも，第 III 相試験に踏み切るだけの事前の確信があるときに，なんとしても承認を受けられる状況を作りたいと願うことは，自然であり理解可能である．しかし，それを認めて良いかというと，そうではない．その正当性は，原理として，「被験者」と「将来その薬剤を使用する可能性がある患者」の立場・視点で吟味すべき事柄である．「成功の確率を大きくする」という視点で判断すべきことではない．

このようなときは，得られている情報に基づいて最も確率的にありそうな（most probable）状態を中心にして，脱落率等を考慮して決めることを原則とすべきである．こう言うと，「正義の味方は気楽で良いですね」と嫌みを言われたりするが，でもそれを言うことが倫理委員会（あるいは治験審査会）の責務であり，それを言えない倫理委員会が少なくないことが，倫理委員会の「資格審査，教育，指針」が必要という（例えば私の）主張につながっている．

5．被験者数の再設定

事前情報としての局外母数の曖昧さを凌ぐ一つの工夫は，臨床試験の進行と共に得られる局外母数についての情報を，被験者数の設定に反映させ，適応的に被験者数を変更することである．これは，情報の有効利用という視点で当然の方針である．

しかしこれは，臨床試験の実施者に利害がないときにのみ当然の方針であって，治験のように実施者が利害関係者の場合には，当然の方針でなくなる．最終結果を願望の方向にゆがめようとする操作の余地を大きくするからである．

もちろん多くの関係者はそのような偏りを入れないように努力している．しかし，一緒に仕事をしているチームの中からそういう「公正の原則から外れた発言」があったとき，「正義の味方」としてそれを諫めることは実質上不可能である．

これが「有害事象の隠蔽」といったことであれば，正義の味方面をしなくても，過去の実例を惹いて，「市販されてから大きな副作用が発現したら後の損害が大きい」と言うことで，偏った方針が採用されることを抑制し得る．しかし，中間解析の結果として，例えば，有害事象は対照と同程度で，有効率は最初に想定していた「10%勝っている」という条件でなく，「よくて5%勝っている程度」という結果が得られたとき，「いずれにしろ勝っているのだから，被験者数を増やして優越性を証明できるようにしても倫理的に問題はない」と主張されたならば，反対できる人が居るだろうか．私にはこういう状況で反対を主張することなど，とてもできるものでないと感じている．

このような日本的土壌のもとで「正義の味方」が演じにくいときの倫理的歯止めを用意するとしたら，利害による偏向を抑制できるような「被験者数再設定の機械的規則化」と他者による再吟味が可能である「根拠の透明化」ではないか，と私は考える．

治験の中間解析においては，局外母数の推定値を考慮に入れることはあっても良いが，群間差を被験者数再設定に取り入れることはすべきではない，という私の主張はこのような背景を考慮してのことである．

6．優越性試験と非劣性試験の違い

優越性試験のデータ解析法は，薬として承認されるべき薬物は，プラセボ対照よりもある大きさの差でもって優れているべきだ，という原理によっている．これは異物を生体に入れれば基本的に何らかの危険・副作用をもたらすのが普通だから，その損失を超えた利益がなければならない．それはまた，費用をそれにあてるという面から

も合理的だ，という原理的な見方から来ている．裏を返せば，一般の食品のように，副作用がなくて生体の自然な快復力に寄与するようなものは，薬という特別高価なものとする必要はないし，規制の必要もないということであろう．

この原理から出てくることは，優越性試験の対象となる被験物には，有用性において，対照との差があるレベル（effect level）以上で無ければならないということが，明確な根拠を持って示されなければならない，ということである．優越性試験での決定方式が，仮説検定方式あるいはそれと同等な意味を持つ信頼区間方式で与えられるのは，そのためである．

これに対して非劣性試験の基本は，実対照薬と同等であれば，両者に未知の一長一短の関係がある可能性があり，一方が効かない患者に対して他方が効く可能性が否定できない．仮に，全く同じとしても，複数の選択があれば競合することでコストを下げる圧力が生じる．したがって，そのような薬物は市販したときに，実対照薬の単一存在より，良い影響を社会に与えるという原理によっている．

問題は，データに基づいて概念としての非劣性を判定する十分満足できる手法がないことである．いわゆる同等性検証であれば，対照薬との差について許容できる幅を誤差的変動に応じて設定するという手法が考えられるが，それは同等であることが十分期待できる場合で，しかも誤差的変動がどの程度であるか対照薬の過去のデータ（historical data）から十分推定できる場合のことである．実際の実対照薬では，そこまで信頼できるデータはないのが普通である．

そこで考えられるのがハンディキャップ方式である．すなわち，ハンディキャップ仮説を帰無仮説とし，同等仮説を対立仮説として検出力がある値（例えば80%）となる仮説検定方式を判定のルールとして用いる，という方法論が登場した．例えば，「臨床試験のための統計的原則」はそれを採用している．

ハンディキャップ方式の一つの問題点は，信頼区間方式で非劣性を承認する非劣性マージンの値の意味と決め方である．非劣性マージンは，判定手法における判定限界値を定めるための作業用の値に過ぎない，と私は考えている．ところが非劣性マージンを導入した統計家は，これを，この程度までは劣っていても医学的には同等として許容できる限界幅，というように説明したがった．すなわち，医学的に定め得る定数パラメータとした．

しかし医学の世界で，医学的に同等と認めて良い劣り方など，あるはずはない．実際，医者の感覚であれば，ほんのわずかでも有用性が優れているのであれば，優れている方を使うべきである，という原則が存在している．そうであれば，ここまでは劣っていても良い，という限界を定めることは困難である．いきおい，それは統計的に

決めて欲しい，というように下駄が統計家に預けられることになる．

統計家の中にはそのような無理な要求に応えるのがスペシャリストとしての統計家の職能である，ということで客観的・形式的な定め方を提案する者がいる．その一つが，プラセボとの中間を非劣性マージンの値にしよう，という提案である．

私はこのような決め方に原則として反対である．それは非劣性という概念に反する結果をもたらすからである．すなわち，実対照薬と被験薬が同等で，プラセボ効果と被験薬・実対照薬の薬効に大きな差があるとき，この非劣性マージンは非常に大きなものとなり，実対照薬よりはるかに劣った被験薬が「非劣性」と判定されることになるからである．

では，非劣性マージンはどのような原理で設定されるべきであろうか．それは非劣性マージンが，対照薬と同等な被験薬が採択される確率がある値（例えば80%）以上になるという条件の下で，それほどかなり劣っている被験薬が誤って採択されることが無いという原則で判定ルールが作られることである．

それを作るには，被験薬の対照薬に対する優越サイズのいろいろな値での採択確率曲線（OC 曲線）に基づくべきだと私は考えている．それは誤差的変動の大きさを示す局外母数（対照薬の有効率，誤差分散，イベント発生率など）の関数となるから，それを事前情報から推理して OC 曲線を描き，それと医学的影響とを勘案して決めるべき性質のものである．

7．優越性と非劣性の切り替え (Switching)

適応的（柔軟）デザインの議論の中で，私にとって最も理解しがたいものが，この意味でのスイッチングである．私はこれを正当化しようとしている人が，非劣性試験という概念を理解していないのだと感じている．優越性試験は被験薬の対照薬に対するある程度以上の優越性を検証する試験であるのに対し，非劣性試験は被験薬が実対照薬に劣っていないことを検証する試験である．

これは，被験薬がどういう意味で社会提供に値するか，という開発の位置づけの問題である．そういう位置づけは対象疾患の，致命性，症状の急変度，現存する競合薬の数と性能（副作用の程度），主要評価変数以外の利点の存在，等によって定められるものであって，試験結果で変わるものではないと考えているからである．

このスイッチングの肯定者は，被験薬に期待した性能が無かったときに，開発費の元を取るために無理矢理その被験薬を承認させたいがために操作を試みているとし

か思えない.

その逆に，非劣性試験を優越性試験にスイッチするというのは，市場での宣伝のためであろうか？そんなことは優越性試験と非劣性試験という概念区分とは無縁のことであって，単に臨床試験の成績をきちんと学術論文として公表して，その優れた性能を世間に知らせるだけで十分である．優越性試験にスイッチするという問題ではないと思うのだがどうだろうか.

8．3群試験のデータ解析

典型的な3群試験が二つある.

一つは被験薬の2用量を対照と比べる3群試験であり，他の一つはプラセボと実薬を対照として被験薬の有効性を検証する3群試験である．前者は，市販においてどのような用法・用量が適切か検証するのが目的であるのに対し，後者は，実対照薬のプラセボに対する優越性が顕著でないが故に，非劣性マージンを設定しての非劣性判定が信頼するわけにいかない，という視点からの試験デザインの採用である.

前者での問題は，被験薬の優越性が確認されるのは当然として，確認された後でどのように臨床用法・用量を決定するかである．私は，原理として，最大対比法で有効性を検証して差し支えないが，臨床用量の決定については，有害事象の出現頻度と有効性の比較で定めるべきものと考えている．この型の試験計画は，現在より多く採用されて良いと感じられる.

後者は，大きな問題を孕んでいる．実対照薬がプラセボと類似の薬効しか無いということであれば，原則としてプラセボ対照を用いた2群比較の優越性試験を採用すべきである．そうでなく，実薬対照のプラセボに対する優越性が過去に認められているにもかかわらず，外国の行政当局の方針に配慮して3群比較試験を採用せざるを得ない，ということであればプラセボ対照の被験者数比率を小さくして3群試験を行うべきであろう．この場合，実験感度（assay sensitivity）を確かめる意味でプラセボと実対照薬の比較をする必要はないのでなかろうか．被験薬のプラセボに対する優越性と実対照薬との非劣性が検証できれば十分なのではなかろうか.

9．中間解析の停止規則

治験のように，利害関係者が臨床試験実施者であってしかも検証性が強く求められる臨床試験の場合，現在提案されている各種の適応的計画が採用できる状況はかなり少ないと思われる.

私が一定の条件下で採用の妥当性を感じるのは，逐次群計画 (group sequential design) である．

計画継続の際の情報秘匿の問題を別にすると，逐次群計画実施上の一つの問題は，主要評価変数のデータが有効停止の条件を満たしているにもかかわらず，（例えば副作用・有害事象を表す）副次的変数の情報が不足しているために継続が望ましい，という判断が当初の試験計画では想定されていない事後的判断として出てきた場合，どうするかである．この判断は，あらゆる中間解析は当初の試験計画で事前に計画されていなければならない，という原理に抵触するからである．

原理的に言えば，本当に事前に想定されなかったことは，その時点の判断で柔軟に対応することになるが，一度そういう経験をした後では，中間解析で副次評価変数の取り入れ方を事前計画で考慮に入れておくべきであろう．「超法規的処置」は初めての経験においてのみ，採用可能なのであって，それが採用された後は法規の中にすぐ取り入れておくべきである．

付言するなら，そういう場合に副次評価変数を被験薬群でのみ観察を続けるのでは情報不足となるというのであれば，そのような状況の妥当性は一般論として議論をしておき，ある程度の合意に達しておくべきである．

１０．モニタリング委員会

臨床試験に対しては，治験審査委員会，独立データモニタリング委員会，倫理審査委員会等，位置づけと名前が微妙に異なる監視組織が設置されている．ここでは仮にモニタリング委員会として一括して考える．医薬統計という側面で考えるならば，独立データモニタリング委員会を焦点にすべきかもしれないが．

被験者数設定の節で述べたように，モニタリングの主要な役割は，被験者の危険の監視と目的にふさわしい情報獲得の可能性の確認である．前者はどちらかというと自明であるが，後者に関しては役割のガイドライン化が必要であるにも関わらず，それを明確に認識している人はきわめて少ない．少なくとも現在進行している数の臨床試験のすべてにおいてこれらの委員会が十分機能を果しているかというと，そうは思えない．

モニタリング委員会が留意すべきこととしては，局外母数の推定値の妥当性，予想脱落率との乖離，（時間分布，背景分布に関する）登録の均一性の吟味，有害事象の発現率などがあるが，現実にはこれが，独立ではないモニタリング委員会で吟味・検討されている．こういう状況は望ましいことでないと考えられるが，これについての

議論はここから割愛する.

１１．非検証的臨床試験における仮説検定

検証的臨床試験では，表現が信頼区間であったり p 値表現であったり，という違いがあっても，判断を下す原理は仮説検定である．すなわち「被験薬は有用でない」という仮説が否定されなければ，有用性が認められないのが「検証」の内容である．

これにたいして非検証的臨床試験，例えば第 I，II 相臨床試験での結論の出し方では仮説検定の視点を採用する必要がない．実際，O'Quigley ら (Biometrics, 1990) の提案も，若菜らの提案 (Statistics in Medicine, 2006) も，非仮説検定の臨床用量探索法である．それにもかかわらず，仮説検定という視点で，デー解析法を認識しようとする傾向が存在している．

たとえば，がんの 2 段階試験計画では，4 パラメータ $(\pi_0, \pi_1, \alpha, \beta)$ で与えられた条件を満たすように被験者数と判定基準が定められる．(π_0, π_1) は，それぞれ，「これ以下であれば開発を進めるべきでない」「これ以上であれば開発を進めるべきである」という判断で設定された奏効率であり，慣例的に「閾値奏効率」「期待奏効率」と呼ばれているパラメータで，実質的内容は「失望奏効率，undesirable level」と「有望奏効率，desirable level」である．(α, β) は，奏効率が π_0 の被験薬が開発継続とされる確率を α 以下にし，奏効率が π_1 の被験薬が開発継続とされる確率を $1-\beta$ 以上にする，という目標のために定められるパラメータである．

n 人の被験者の内でこの被験薬で有効な結果が得られた被験者数を Y とし，Y が二項分布 $B(n, \pi)$ に従うとすると，この条件を満たす判定方式は，帰無仮説 $H_0: \pi = \pi_0$，対立仮説 $H_1: \pi = \pi_1$ の仮説検定で，名目有意水準をα，目標検出力を $1-\beta$ とする仮説検定方式と同じになる．だから，「2 段階試験計画は仮説検定である」と関係者が信じても，一般には違和感が無いようである．

実際はどうだろうか？今，Fisher 先生や Wald 先生が存命でこの言い方を耳にしたら，私の想像では，かなり強く異議を唱えると感じられる．Fisher 先生は「これは仮説検定ではない」と言い，Wald 先生は，「これは二者択一の決定問題・決定方式である」と言うに違いない．そして両先生とも，帰無仮説，対立仮説という表現は避けるべきだと主張するであろう．数学的問題としては同じであっても，抜き取り検査と同様，これは制約条件のある決定問題であって，Fisher 先生が原理的・概念的に提起した仮説検定問題ではないからである．

私はこの問題を仮説検定として捉えるべきではないと考えるのだがどうだろうか．

【非臨床試験】

１２．定型的な毒性試験

一口に非臨床試験と言っても，定型的な毒性試験（安全性試験），薬効薬理試験，非定形的な毒性試験，創薬におけるスクリーニングなどでそれぞれ異なった側面がある．

医薬安全性研究会が発足した約 30 年前には，毒性試験のデータ解析が大きな問題であった．毒性試験のガイドラインが整備され，good laboratory practice (GLP) が義務化されるという流れの中で，毒性試験が新薬申請においてある程度の重要度を持つようになった．ところが，それに用意する資料において，「無毒性量をどのように決めたらよいか」「多重比較を使う必要があるか」「多重比較を使うとしたらどの手法をどのように使うべきか」「多種・多様なデータ・変数が登場するがそれをどのような手法で解析すべきか」といったことについて，標準的なやり方が確立されていなかった．それを確立するのが最も大きな問題であった．たとえば，「毒性・薬効データの統計解析，サイエンティスト社，1987」，「毒性試験データの統計解析，地人書館，1992」，「統計的多重比較法の基礎，サイエティスト社，1997」等が出版されたのは，そのような状況の反映である．

約 30 年たった現在，ガイドラインに収載された毒性試験データの統計解析については，論議がほぼ収まっている．実験やデータが定型的であれば，データ解析に用いる手法も定型化できる．原理・原則の理論面では問題が残るとしても，現実面ではそれが大きな問題にならないように統計手法が整備されたと言って良い．したがって毒性試験でのデータ解析上の問題は，例えば環境汚染物質の毒性評価・吟味といった，非定形的なものに現れていると感じられる．

１３．安全性確認における仮説検定手続きの意義

実験家には，毒性試験のデータ解析とは仮説検定を適用することであると思いこんでいる人が少なくない．統計家の視点で見ると，この思いこみは誤解であると感じられる．

医薬品開発のための毒性試験の目的は，その安全性がどの程度であるかを確かめることであって，毒性があることの検出ではない．そして，仮説検定という推測の形式は，毒性が存在することは検証するが，存在しないことを検証することは原理的にできない判定手順である．したがって，検定で安全性を確認することはごく限られた場

合以外，原理として無理である．ごく限られた場合というのは，特定のメカニズムがあって，それについての指標変数パラメータ，たとえば実質安全量（virtually sage dose, VSD）が指定できて，安全性が定義できる場合である．医薬品開発過程での毒性試験はこれに該当していない．

しかし，仮説検定で用いるところの，たとえば奇形発生率といった検定統計量に，Dunnett 検定の棄却限界値を適用したり，Williams の手順を適用したりして p 値を求め，それに基づいて毒性を評価することにはある程度の合理性がある．一般毒性試験であれば，使用動物数に標準的な値があって，それを前提にすれば，その毒性判定手順の性能が評価でき，どの程度の毒性が検出できるか分かるからである．

その場合，一般の検定で標準的に使われる 5%，2.5%，1%という有意水準を無批判に用いるのは，望ましくないと私は考える．たとえば「無毒性量を求める」という問題に検定手順を適用して，有意差が生じる前の最大用量を無毒性量にするといったとき，動物数と毒性の重篤性のバランスで，有意水準を変更することがあって差し支えない，と私は考える．ただし，一般毒性の評価では，標準的な有意水準を考慮して標準的な動物数が定められている側面も否定できないので，一般毒性のようなガイドライン化されているものの有意水準は，個別に検討する必要・根拠が無いと感じられる．

１４．インビトロ毒性試験に仮説検定を適用するときの有意水準

仮説検定法をインビトロ毒性試験での毒性評価に使うことについては，検定統計量を評価の指標にするという面において，不合理でない．実際，Dunnett 検定，傾向検定が個々の試験で利用されている．

しかし，エームズ試験，小核試験，マウスリンフォーマ試験（MLA）などの遺伝毒性試験，３次元ヒト皮膚モデルを用いた皮膚刺激性試験，細胞転換試験などの発ガン性試験，h-CLAT などの皮膚光感作性試験，といったインビトロ試験では，実験家の間に，仮説検定法を用いることについての抵抗がかなりある．その理由の一つは，通常の 5%あるいは 1%有意水準の棄却限界値で有意な場合を毒性ありとする方針では，実験家の感覚にあう判定，すなわち，陽性（毒性あり），陰性（毒性無し）がインビボ試験の結果と一致するという判定が出てこないからである．実際，エームズ試験では２倍法，MLA では３倍法，小核試験では有意水準を調整しない２段階手順が用いられていて，Dunnett 検定の対案となっている．

インビトロ試験とインビボ試験の結果の乖離には，メカニズムの違いが大きいことが第一であるが，私が気になることは，それ以上に，統計的仮説検定における第１種の過誤の抑制の原理が，実データに対する妥当な棄却限界値の設定と乖離しているこ

とである．実際，私がMLAのデータで，実験家の判定と統計的手順の結果を対比したところでは，公称有意水準を 0.3%という位に小さくしたときはじめて，実験家の評価とほぼ一致する判定（一致率 90%）が得られている．ということは，「統計モデルの上で用量反応性がない」という仮説が，インビボ試験における毒性の有無に対応していないことを意味している．

このことは，実験家から「インビトロ試験は感度が良すぎる」という表現で言及されたりしているが，その原因としては，ポアソン分布や二項分布といったモデルでの分散より実際の評価変数の分散の方が大きいという過分散性（over dispersion phenomenon）と，生体にある防御機構がある種の閾値を備えていることが考えられる．

１５．決定問題としてのインビトロ毒性試験のデータ解析

インビトロ毒性試験は，その妥当性検証（validation）が無い限り毒性試験法としての地位を得ることができない．妥当性検証では，たとえば“OECD Guideline No. 43”に規定されているような，検証条件を満たすことが必要である．その一つは，その試験法が被験物質の陽性・擬陽性・陰性を判定するための規準（criteria）を定めておくことである．ここで擬陽性は，陽性と断定することは困難であるが陽性にかなり近いと判定することができる存在である．

判定規準は，たとえば「感作性指標（sensitization index, SI）の（用量を変えて得られる）最大値が 2 以上であること」というように単純なものもあれば，「最大反応値が陰性対照での反応値の 2 倍以上で，連続する 3 用量における増加反応が示されている」というような複合的なものもある．いずれにしろ，観測値の値が存在する空間（標本空間）の各点に対して，陽性，擬陽性，陰性を対応させる決定方式のことである．

この決定方式は，適当な統計量の値に陽性，擬陽性，陰性の判定を対応付けることで作られることが多い．擬陽性と陽性を「毒性あり」と見なすと，この決定方式は見かけ上，検定方式と同じ形になっている．そのためであろう．これを本当に検定方式だと思っている人が意外に多い．すなわち，かなり多くの人が，偽陽性率を 5%に制御するのが統計的判定方式・検定だと思いこんでいるのである．この思いこみがあると，別の規準，たとえば偽検出率（false discovery rate, FDR）を制御する方式が提案されたときに，それを検定の枠組みで捉え，検出力を計算しようとしたりする．これは無理な話である．

インビトロ毒性試験では，被験物質の数だけ，毒性の仮説があり，その仮説も単に

無毒性という帰無仮説（null hypothesis）だけでなく，強い毒性物質，中くらいの毒性物質，弱い毒性物質，という状態・仮説が存在し，それぞれの被験物質がどのような状態・仮説に属しているか判定することが求められる．これに応じる決定方式の性能を仮説検定の枠組みで作ることはできない．性能評価の指標を作るとしたら，それぞれの強さの陽性・陰性物質の誤判定率を個別に評価した上で，標準化死亡比のような意味での，標準化した正判定率，あるいは一致率といったものを用いるべきであると私は考える．

１６．薬理試験で得たい情報

薬理試験でデータから得ようとしている情報は，モデル探索・確認・検証であって，「有意差の存在の確認」ではない．それにもかかわらず，薬理試験の実験家には，有意差の判定のみにこだわる人が多い．これは効率的な情報利用の原則に反する物の見方，データ解析のやり方である．

たとえば用量群実験で溶媒群に対する反応の差を評価する状況を考える．このとき，溶媒群と各用量群を比較して有意差を調べている例が薬理試験の論文には非常に多い．ところが薬理試験では，サンプルサイズが 10 というオーダーなので，高用量では有意であるが低用量では有意でないという結果が出やすい．そのときにそれをそのまま論文に記述して，後はデータと無関係に考察を書いていたりする．これは原理的に賢いやり方でない．

このような反応は一般に用量依存的なのが普通だから，高用量で有意差があるときは中用量でもそれなりに反応の増加があり，しばしば直線的用量依存性が認められる．そうであれば，統計的に有意差が無いことが用量依存性反応を否定することにはならず，むしろ，反応の直線的増加を前提にした解析を行うのが合理的ということになる．この方針をとれば，勾配で薬物の反応強度が定量的に評価できるし，他の類似薬物との定量的・統計的比較も可能になる．

よく知られているように，有意水準を定めた仮説検定における有意差の有無だけより，p 値の方が多くの情報を残している．さらにパラメータの信頼区間は，p 値より多くの情報を残している．したがって，データ解析においては，可能な限り信頼区間を求めるべきである．

信頼区間は，知りたい情報が未知母数・パラメータという形式になっているときしか用意できない．したがって，信頼区間方式を採用するには，対象となっている薬理試験のデータに統計的モデルを想定しなければならない．そしてそれには，薬理学的メカニズムに対するある程度以上の事前情報と薬理学的洞察が必要である．つまり，

薬理試験のデータ解析では，原則として，適切なモデル化がキーになる．実験家がそういうモデルを統計家に提示できたときに初めて，有効なデータ解析ができる．

上の議論では，直線性を前提にして信頼区間を作ることを説明したが，ときにはそうでない場合もある．しかし薬理試験での用量水準は，たかだか５水準である．５水準で複雑な用量反応関係を想定するのは危険である．可能な限り薬理学的モデルを想定し，それができないときはある程度単純な統計モデルを利用するのがよいと私は考える．

１７．モデル化の後のパラメータ推定法の選択

薬理試験の分野には，古典的に数学的・統計的モデルが確立している場合がある．薬物動態学におけるコンパートメントモデル，受容体理論における反応速度モデル，酵素阻害理論における Michaelis-Menten モデルなどである．

これらのモデルには，それぞれ，キーになる未知パラメータが存在する．データ解析では，仮説検定ではなく，これらの未知パラメータの区間推定を目指すべきである．

ところが，2006 年の東京理科大の修士論文で明らかになったことであるが，薬理試験の分野では，統計学的に性能の劣るパラメータ推定法が，長い伝統・慣習の下で依然として使われている．これらは，結果として候補化合物のスクリーニングの効率を落とすから，改善がなされなければならない．仮説検定をやめて信頼区間方式を採用するときでも，できる限り効率の良い推定法を使うのは，データ解析の原則である．

【おわりに】

日本では，産・官・学の間の人の移動が極めて少なく，あったとしてもほとんどが一方通行である．その結果，異なる立場に立って物を考えることを疎かにしがちである．それに歯止めをかけるのが物の見方についての原理・原則ではないだろうか．倫理を一つの指針としてガイドラインの中に明記するのはそのために重要なことと感じられる．それは，例えばソリブジン事件に象徴されるような，データの意味することを抑えてある方向に結論を向けさせることに歯止めをかけるのでは無かろうか．データモニタリング委員会，治験審査委員会，倫理委員会といったものが，建前通りの姿では機能していない現実の下では，ときに，素朴な原理・原則論を論議することが必要だというのが，今回このテーマを設定した理由である．

この試みは，果たして有効だっただろうか．議論をしたいところである．

Title page

Type: Original Article

Title: Interval estimation of the 50% effective time in small sample assay data

Short running title: Interval estimation of the 50% effective time

Author:

Takashi Sozu[1] (寒水 孝司), Ayako Shiraishi[2] (白石 亜矢子), Yohei Hyodo[3] (兵頭 洋平), Chikuma Hamada[3] (浜田 知久馬), and Isao Yoshimura[3] (吉村 功)

[1] The Center for Advanced Medical Engineering and Informatics, Osaka University, 2-2 Yamadaoka, Suita-city, Osaka 565-0871, Japan

[2] Management and Biostatistics Department, Janssen Pharmaceutical K.K., 3-5-2 Nishi-kanda, Chiyoda-ku, Tokyo 101-0065, Japan

[3] Department of Management Science, Faculty of Engineering, Tokyo University of Science, 1-3 Kagurazaka, Shinjuku-ku, Tokyo 162-8601, Japan

Correspondence to:

寒水 孝司 (そうず・たかし)

大阪大学臨床医工学融合研究教育センター

〒565-0871 大阪府吹田市山田丘 2-2

大阪大学大学院医学系研究科 J6 内科系臨床医学専攻 情報統合医学講座 医学統計学

TEL: 06-6879-3597 FAX: 06-6879-3598 E-mail: sozu@medstat.med.osaka-u.ac.jp

Abstract

The time score for 50% cell viability (50% effective time; ET50) is used as the index of skin irritation evaluated by a three-dimensional human skin model, such as TESTSKIN^{TM} and Vitrolife-Skin^{TM}. ET50 is conventionally estimated by linear interpolation of measurements at two time points, which yields cell viabilities above and below 50%. This simple method is problematic in that biased estimates are occasionally obtained and confidence intervals cannot be appropriately constructed.

We compared four estimation methods including a logistic regression method, a log-time regression method, a linear regression method and a newly proposed two-stage method through a Monte-Carlo simulation study in small sample sizes due to the experimental restrictions. The logistic regression method provides almost unbiased estimates, although the confidence interval for ET50 is occasionally not obtained. The log-time regression method and the linear regression method provide positive biased estimates, although the confidence interval for ET50 is obtained in any case. The two-stage method is reasonable, in which the log-time regression method is adopted only if the logistic regression method cannot construct a confidence interval for ET50.

Key words: skin irritation, human skin model, effective time 50 (ET50), interval estimation, two-stage method.

Introduction

Several three-dimensional human skin models (3D skin models) such as TESTSKIN^{TM} and Vitrolife-Skin^{TM} have been developed for evaluating the skin irritation of test substances (Botham, 2004, Morikawa et al., 2002, Morikawa et al., 2005, Morota et al., 1998). The time score for 50% cell viability (50% effective time; ET50) is used as the index of skin irritation (Kojima, 2005a) and conventionally estimated by linear interpolation of measurements at two time points (Morikawa et al., 2002, Sonoda et al., 2002), which yields cell viabilities above and below 50%. This simple method is problematic in that biased estimates are occasionally obtained as a result of linear approximation of convex time-response curves and confidence intervals cannot be appropriately constructed.

Due to the high cost of models and restrictions on the time schedule of experimenters, the scale of experiments using 3D skin models is small and the time points for measurement of cell viability are usually set at 4 time periods within 16 hours after the start of treatment and, if necessary, at 24 hours after the start. Actually, in the validation study of Vitrolife-Skin^{TM} (Kojima et al., 2005b, Kojima et al., 2005c), viabilities were measured at only 4 or 5 time points with two repetitions at each time point in each experiment. The method of data analysis in the experiment under such restricted conditions should be carefully examined because the lack of information tends to hinder obtaining precise estimates of ET50 or constructing confidence intervals for ET50.

We, therefore, compared four methods of point and interval estimation including a logistic

regression method, a log-time regression method, a linear regression method and a two-stage method, which is newly proposed in this article, the results of which are explained in the succeeding sections.

Methods

In this section, we explain the conventional method and the four methods for estimating ET50.

The conventional method for estimating ET50

In the evaluation of the skin irritation using the 3D skin models, the absorbance of the extracts at each time point is generally measured as the cell viability of a test substance as well as a negative control and a blank. The cell viability of 3D skin models at time point i is estimated by

$$\hat{p}_i = \frac{\bar{y}_{i\cdot} - \bar{y}_{B\cdot}}{\bar{y}_{C\cdot} - \bar{y}_{B\cdot}} \times 100(\%), \tag{1}$$

where $\bar{y}_{i\cdot}$, $\bar{y}_{C\cdot}$ and $\bar{y}_{B\cdot}$ are the mean value of the absorbance of a test substance at time i, a negative control and a blank, respectively.

The skin irritation of the test substance is assessed by ET50 defined as the time score for 50% cell viability. ET50 is roughly estimated from a straight line between two time points above and below 50% cell viability on the time-response curve (Morikawa et al., 2002, Sonoda et al., 2002).

A logistic regression method