統計数理（1998）
第46巻 第1号 81-95

特集「ヘルスサイエンスと統計科学」
［研究ノート］

検証的比較臨床試験の計画において考慮すべきこと†

——ICH統計ガイドラインの理解のために——

東京理科大学* 吉村 功

（受付 1997年12月1日；改訂 1998年1月8日）

要 旨

本論文では，新薬承認の基準を日欧米の3極間で共通化しようということで作られつつある「統計ガイドライン」の内容のうち，臨床試験の計画の部分について，問題になりそうなことを取り上げて議論している．焦点としている主要な問題の第一は，日本で好まれている実薬対照の採用が，一般には決して合理的なものではないことである．第二は，信頼区間方式での信頼水準と検定方式での有意水準が片側検定を採用したとき混乱しがちなことである．第三は，非劣性試験での受容すべき同等限界は，試験計画書の段階で根拠と共に明確にしておくべきことである．第四は，今回新たに導入された非劣性試験という概念は，対照が確実に薬効を持っているときのみ容認可能なことである．第五は，多施設試験が採用されたり，脱落等に関して解析対象が変更されたりしたときには，偏りが入らないように解析に注意しなければならないことである．

キーワード：統計ガイドライン，国際調和，臨床試験，仮説の検証，試験計画．

1. はじめに

薬は多くの国で承認制になっている．薬として市販を申請された製剤（申請製剤）を承認する条件（承認条件）は国によって違う．一般に共通な承認条件の一つは，申請製剤が検証的比較臨床試験で承認の価値がある（有用）と客観的に認められることである．価値は主として，申請製剤に薬効があることとその安全性が高いことの両面で評価される．ここで臨床試験とは，ヒトを対象とした実験で，その遂行には法律上あるいは倫理上で多くの制限が設けられている（厚生省（1997），World Medical Association（1989），光石（1996）参照）．比較臨床試験は，比較を目的とした臨床試験，つまり治療目標とした疾患の患者をいくつかの群に分けてそれぞれに別の処置を行い，各群間での結果の違いを通して処置の違いを評価する実験である．検証的比較臨床試験とは，申請製剤が有用であるという主張を検証する目的の比較臨床試験である．一般に比較臨床試験で申請製剤の比較基準にされる処置あるいは製剤を対照あるいは対照薬といい，対照として用意された群を対照群という．対照としては，プラセボ（薬剤の形をしていながら有効成分がないもの），市販薬，申請製剤の少量処方（低用量）などが使われる．ここではそれらを，プラセボ対照，実薬対照，低用量対照，と呼ぶことにする．

* 工学部：〒162-8601 東京都新宿区神楽坂1-3．

† 117頁より討論あり．

申請製剤の承認条件が国によって大きく違うと，同じ製剤なのに承認する国と承認しない国とが出てくる．あるいは国ごとに同一目的の試験を別に行うことになる．これは良薬の普及を妨げたり無効薬の市販を許す原因となりやすいし，一般に実験台とされる患者数を大きくする．これは薬の国際的流通を妨げるし，真実を認識する面でも不合理であり，患者の利益にならない．ということで承認条件をなるべく共通化しようという動きが国際的に生じた．これを具体化したものが，ICH（International Conference on Harmonization; 新薬承認基準の調和のための国際会議）である（D'arcy and Harron（1996））．ICH は現在まで4回の全体会議と，非常に多くの作業会議を，日本，欧州，米国，の間で持っている．

調和における一つの課題は，臨床試験，とりわけ検証的比較臨床試験の設計，実施，解析，報告を，統計学的側面から合理化することである．実際，米国，日本，欧州はそれぞれで臨床試験の統計解析についてのガイドライン（以下ではこれらを，それぞれ「米国ガイドライン」：FDA（1988），「日本ガイドライン」：新医薬品統計解析指針検討会（1992），「欧州ガイドライン」：Lewis et al.（1995），と呼ぶ．）を持っているが，その内容にはかなりの違いがある．これらが調和の対象にされるのは当然である．

ICH はその調和を課題とした専門家作業委員会（E 9-EWG：Efficacy に関する第9課題を議論する expert working group）を設け，共通なガイドラインを作製することにした．その委員会が1997年2月の時点で作製したガイドライン案が“Statistical Principles for Clinical Trials”という標題の文書である．（以下ではこれを「ICH ガイドライン」と呼ぶ．）（吉村 他（1997），Sato and Yoshimura（1998）参照．その後多少の修正があり，本稿はその修正版に基づいている．）

標題から読みとれるように，ICH ガイドラインは原則を記述しているだけである．その原則を実現するための具体的な手続きは，「適切な教育を受け，しかもしかるべき経験を持った臨床統計家」にゆだねられている．臨床統計家の責任は重い．しかしこれは臨床統計家がそれぞれ勝手に具体的な手続を定めるということではない．具体的な手続は十分合理的でなければならないし，それがどのような意味を持つかは，公の場で議論され合意されなければならない．議論すべきことや合意すべきことは沢山あるが，本稿では，ICH ガイドラインの “III Trial Design Considerations” で私が気になったことを取り上げ，それについての個人的見解を述べる．私自身が ICH ガイドラインの作成に関与していながら，あえて個人的見解と言うのは，ICH ガイドラインの実効化の際に行われる関係者間の意見調整で，私の見解と異なる見解が得られるかもしれないからである．

2. 試験の型

ICH ガイドラインでは，比較結果をどのような差としてとらえるかで，試験の型を，優越性試験，同等性試験，非劣性試験，用量反応関係試験に分けている．用語から想像できるように，優越性試験は申請製剤がプラセボより優れた薬効を持っていることを確かめる試験である．同等性試験は申請製剤がすでに市販されているある薬剤と同等であることを確かめる試験である．非劣性試験は申請製剤が薬効に関して，治療目標疾患についてすでに市販されている標準的な薬剤と同等またはそれ以上であること，すなわち少なくとも劣ってはいないこと，を確かめる試験である．用量反応関係試験は申請製剤の特性が，用量を変えたときどうなるかを調べたり確かめたりする試験である．比較臨床試験では薬効の検証と同時に安全性の吟味もされるが，その内容は型によって異なるものではないから，型の区分けは薬効の評価法によっている．

今までのガイドラインは，米国ガイドライン，日本ガイドライン，欧州ガイドラインのどれでも，この種の区分けを明示していなかった．その区分けの不十分さがときに混乱や誤解の原

因になっていた．そこでE9-EWGは臨床試験の区分をICHガイドラインに入れることにし，どのような区分けを明示すべきかについて，かなりの時間を割いて議論した．その結果現段階で合意に至っているのが上記の区分である．

3. 優越性試験

申請製剤が確かに薬効（対象患者における疾患の治癒あるいは症状の寛解をもたらすこと）を持っていることは，プラセボ対照に対して明確な薬効差があることで証明できる．あるいは後で述べるように，プラセボ対照に劣らないと思われる場合には，プラセボ対照の代わりに実薬対照や低用量対照に対して薬効差があることで証明できる．いずれにしろICHガイドラインは，“Generally in this guideline superiority trials are assumed, unless it is explicitly stated otherwise.”という文章で，優越性試験が検証的比較臨床試験の最も基本的なものであることを述べている．

優越性試験で確かめたいことは単純なことである．しかしその目的を達成するための実際の計画や解析についてはいくつか心得ておくべき注意点がある．以下にそれを述べる．

3.1 優越性試験での対照群

薬効の発現には偶然性がある．したがって申請製剤の対照に対する明確な薬効差を認めるには，その優越性を仮説検定という形式で確かめるのが自然である．教科書的な意味で典型的なのは，以下のような方針を採用するものである．

まず試験計画の段階で対象疾患についての主評価変数（primary variableあるいはprimary endpoint）を一つ決めておく．対象疾患と想定された患者を一定数集め，ランダムに2群に分け，一方に申請製剤，他方に対照を処方し，試験計画書に定めたやり方で主要評価変数を測定する．試験計画の段階で

$$T=t(X-Y),\ \text{ただし}X,\ Yはそれぞれ申請製剤群とプラセボ群での主要評価変数の平均，関数}t\text{は単調増加関数},$$

という形の検定統計量Tとそれに対する片側有意水準αの棄却限界値cを定める．実際のデータを得たらTの観測値T_0を調べ，それが$T_0>c$を満たしていれば，申請製剤の優越性を認める．すなわち申請製剤には薬効があると判定する．αは第1種の過誤の上限として定める値で，後で述べるように，0.05（5%）あるいは0.025（2.5%）を採用することが多い．

こういう判定法で申請製剤の薬効を認めるのが典型的な優越性試験であるが，対照の選び方については以下のような問題がある．

FDA（Food and Drug Administration: アメリカの医薬品行政を司る官庁）は，対照としてプラセボを用いるのが原則だとする．標準的な市販薬を処方した群あるいは申請製剤の低用量を処方した群を対照群にすることも認めてはいるが，それによる薬効の証明についてはどちらかというと疑問を出す傾向がある．ところが日本の臨床医は一般にプラセボを対照にするのを嫌う．その理由の一つは，治療の中で申請製剤の薬効を吟味するのが臨床試験だという位置づけにある．すなわち，医師は治療をするのが仕事であって，治療をしないのは医師の責任を放棄するものである．プラセボを処方するのは基本的に治療をしないことであるから医師としては行い難い，というわけである．これは臨床試験が基本的に薬効を確かめるための実験である，と認識してもらう以外に克服できない理由である．臨床試験については，それもまた医師の責任を全うする一つの仕事であることを論理的に明確にし，それを日本の臨床医の常識にしてもらうことが必要である．

臨床試験を実験と見なしたとしてもなお日本の臨床医はプラセボを嫌う．一般に臨床医は薬効があるとして市販薬を使っている．申請製剤でも動物実験等で薬効を確認して臨床試験に供している．そういう市販薬あるいは申請製剤の低用量は，悪くてもプラセボより悪いことはあり得ない．それを対照薬にしないでプラセボを対照薬にするのは，患者に意図的に無効な処置をするものとなる．これは倫理的に許されないはずだというわけであろう．

この主張は日本では広く受け入れられているが，日本以外では一般に受け入れられていると言い難い．ICH ガイドラインも FDA も，底流として，実薬対照よりプラセボ対照が原則という視点を固持している．なぜだろうか．

それは，優越性試験で薬効が証明されていない実薬対照は，薬効ではプラセボ以下にならないとしても，有害作用の面でプラセボに劣る可能性があるからである．臨床試験は薬効の確認だけを役割にしているのではなく，有害作用の検討すなわち安全性の吟味もその役割にしているから，有効性が証明されておらず安全性では劣るかもしれない薬剤を対照にするのは，かえって非倫理的になる．日本の主張が説得的でないのはこのためである．

日本の通念には，市販されている薬剤は薬効があるに決まっているという思いこみ，あるいは科学的検証という手続を不要とする経験的な薬信仰，による部分が少なくないのでなかろうか．もし「日常の経験で有効性を感じているのだ」というのであればその経験の科学的証拠を客観化すべきであろう．それなしに欧米的合理主義者を説得するのは無理である．市販薬を対照にするには，それに薬効があることと安全性においてプラセボより劣らないことが，客観的に吟味されていなければならない．ICH ガイドラインで “The appropriateness of placebo control vs. active control should be considered on a trial by trial basis.” とあるのはこのためである．

しかしながら，致命的あるいは短期間で確実に症状が悪化する疾患においては，プラセボが非倫理的であることが十分予想される．その場合は実薬対照もやむを得ないというのが ICH ガイドラインの現時点の到達点である．そしてそれについてのさらに細かい議論は，対照群の構成，利用，特徴吟味ということで，ICH-E 10-EWG で現在も議論が続いている．

3.2 優越性試験での有意水準

薬効の証明には検定の考え方が使われる．それゆえ臨床試験の統計解析法を定める際には，検定統計量とともに棄却限界値を定めることが必要になる．棄却限界値は有意水準によって定まる．統計学の理論では有意水準の決定と検定統計量の選択は独立な考慮事項である．すなわち有意水準は第1種の過誤と第2種の過誤の重要性の比較衡量に基づいて決め，検定統計量は帰無仮説と対立仮説の関係で決めるべきである．さらに試験参加者数（標本の大きさ）は必要な検出力の大きさに基づいて決めるのが通常である．

もしこの常識に基づくなら，試験計画者はまず有意水準を，たとえば5%，というように定める．優越性試験では申請製剤の薬効が対照より大きいことを確かめたいのであるから検定は片側検定とする．その上で棄却限界値を定め，検出力を計算し，必要な試験参加者数を定めるのが自然である．ところが臨床試験ではこの統計学の常識が通用していない．ICH ガイドラインでは，

“The choice of type I error rate should be a consideration separate from the use of a one-sided or two-sided procedure”

と言っておきながらさらに，次のように続けている．

“(Historically, the conventional probability of Type I error is set at 5% or less for a

two-sided test, which implies a 2.5% or less one-sided Type I error. It is not the intention of this guideline to modify this standard).”

この部分はこのガイドラインにおいて最も歯切れの悪い部分の一つであろう．確かに伝統的には，日本の優越性試験の計画でも両側5%検定に基づく試験参加者数の設計がなされていて，私がある臨床試験において片側5%検定での設計をしたところ，「それで本当に申請のとき問題が起こらないのですか」と驚かれたことがある．そのとき私は，「少なくとも日本の統計家で新薬審査に関わっている人の中にはそれがおかしいという人はいないと信じています．」と答えた．実際の新薬審査が有意水準5%の片側検定で行われているからである．ただしFDAでは，優越性の検定での有意水準を片側2.5%にしているから，FDAで認可を得ようとする臨床試験ならそれに合わせた例数設計が必要になる．実際に第1種の過誤をいくらにとった判断を下すべきかは，今のところ統計ガイドラインでの「調和事項」になっていないのである．

3.3 優越性試験における信頼区間方式

データの縮約においては，一般に検定の結果が有意かどうかを記述するよりp-値を記述する方が多くの情報保存になる．さらに，信頼区間による記述はそれより多くの情報保存になる（たとえば，吉村（1996）参照）．

信頼区間には片側区間も無いわけではないが，「信頼区間方式が多くの情報保存になる」というときは，両側区間を指すのが統計学の常識である．ICHガイドラインもそれを認めた上で要所要所で信頼区間方式を勧めている．では信頼区間方式で優越性試験の結果を表示することと優越性の確認はどのような関係になっているだろうか．

先に述べた検定統計量に基づく信頼区間は，たとえば次のようにして作られる．主要評価変数の各群での期待値の差を薬効差とする．対象母集団（対象疾患の患者集団全体）での薬効差Δにたいして，

$$\text{帰無仮説 } H_0: \Delta \leq \Delta_0 \quad \text{vs.} \quad \text{対立仮説 } H_1: \Delta > \Delta_0, \text{ 有意水準} \alpha,$$

という仮説検定問題を考える．そのための検定方式を単調増加関数tを含んだ検定統計量$T = t(X - Y - \Delta_0)$に基づいて作り，十分広い範囲に渡ってΔ_0を動かすと，その検定で有意にならないΔ_0の範囲がある値Δ_Lより大きいところという形で得られる．これが信頼区間の下限である．対立仮説の方向を逆にして考えると，同様にして信頼区間の上限Δ_Uが得られる．このとき(Δ_L, Δ_U)で与えられる区間が信頼水準$1-2\alpha$の信頼区間である．

この構成法で信頼水準を95%にとったときの薬効差の信頼区間が完全に正の領域に含まれるならば，すなわち信頼区間の下限が0より大きいならば，これは有意水準2.5%の優越性検定で有意に薬効差が認められたことと同じである．どちらで判定しても結論は変わらない．ではこのとき信頼区間の上限は何を意味するのだろうか．

優越性試験として検証的臨床試験を設計したときの目的についていうと，この場合の信頼区間上限は何の意味もない．しかしこの目的以外にも，治療効果の推定という目的が臨床試験にはあるわけで，それを考えるとこの信頼区間上限は無意味でない．それはたとえて言えば，入学試験に合格したときの試験の点数のようなもので，合否には影響しないが資料として価値あるものとなる．後に他の側面での検討を行うときの情報・資料として，報告と記録保存の価値がある．優越性試験だからといって片側信頼区間でよいとしないのはこのためである．

上に説明した両側信頼区間の作り方は，別の言い方では，「両側検定で棄却されない帰無仮説での薬効差の全体」である．優越性試験でも両側検定を用いるのは，このような信頼区間方式との首尾一貫性の保持のためのようである．日本では，多分問題に応じて検定方式を変えるべ

きだという統計家の感覚によるのであろうが，信頼区間方式で優越性の評価をするときは信頼水準を90%にとるのが普通になっている．これにたいして欧米では，優越性試験に使うときであっても，信頼水準を95%に固定した両側信頼区間を使う方がよいと考えているようである．

これを正当化する議論は，多くの場合，毒性試験や安全性評価を例に取り上げているから，そういうところでの観念的な思いこみが欧米の常識を形作っているのかもしれない．少なくとも私は，信頼区間方式でも検定方式でも第1種の過誤の確率をどのように制御するかは，下したい判定とのかねあいで決めるべきことであって，両側検定なら5%片側検定なら2.5%というつじつま合わせで決めるべきことではないと考えている（たとえば，森川（1994）参照）．

4. 同等性の判定

ICHガイドラインでは，同等性試験と非劣性試験とを同じ項に入れている．両者に，実薬対照を用いることが必須という共通特徴があるからである．申請製剤がプラセボと同等，あるいは劣っていないことを確かめる試験などというものは，実際問題として意味がないのである．しかしながらこの二つの試験は概念的には明確に区分けした方がよい．問題点が大きく異なるからである．このICHガイドラインにその区別が不十分にしか記されていないのは私にとって不満なことである．それはともかくとして本節では，同等性の問題とその判定法について注意すべきことをいくつか述べる．

同等性試験には少し性質の異なる二つのものがある．一つは生物学的同等性を確かめる試験であり，他の一つは臨床的同等性を確かめる試験である．

試験をしたい製剤（試験製剤）が，既に市販されている薬剤と同じ薬効物質であるのに，製法などが異なっているため，確かに両者が同じ振る舞いをするかどうか吟味しなければならない場合がある．これをヒト（種としての人間を指すときにはカタカナで表記する習慣がある．）上で確かめるのが生物学的同等性試験である．たとえば市販薬が動物の生体から抽出したもので，試験製剤が遺伝子工学的に製造されたもの，という場合である．生物学的同等性は，血中濃度の時間変化を測定して証明するのが普通であり，本稿が取り上げているICHガイドラインとは別に，生物学的同等性試験のガイドラインがある（鹿庭（1997）参照）ので，ここでは触れないことにする．

外用薬のように，血中濃度で挙動を測ることができない投与形態や薬効物質では，実際の投与で得られる主要評価変数の測定値から両者の差を調べ，試験製剤が実対照薬と同等かどうかを判定する．これが臨床的同等性試験である．

臨床的同等性試験に関しては，それ以前に多くの非臨床試験や物理化学的試験，あるいは過去の臨床試験や治療の経験があるので，それに基づいて主要評価変数を定める．単に薬効というだけでなく，薬理学的反応や副作用に着目して主要評価変数を定めることもある．そうして定めた主要評価変数の，対象母集団における差について，測定値から信頼区間を作る．その信頼区間が試験計画において定められた医学的に容認可能な差の範囲に入っていれば，同等性があると判定する．これがICHガイドラインでの原則である．

生物学的同等性試験の場合には，その容認できる範囲を，「血中濃度の平均値の比を5対4の範囲にする」というように統一的に定めているが，臨床的同等性試験の場合は主要評価変数が多種多様であるから，このような統一的指定ができない．そこでICHガイドラインは次のように指定する．

“An equivalence margin should be specified in the protocol; this margin is the largest difference which can be judged as being clinically acceptable and should be smaller than

differences observed in superiority trials of the active comparator. For the active control equivalence trial, both the upper and the lower equivalence margins are needed, ... The choice of equivalence margins requires clinical justification."

つまり同等とみなす範囲を，しかるべき臨床上の根拠を持って定め，試験計画書に記述しておかなければならない．その範囲は実薬対照の優越性を承認した試験での実薬対照の薬効の範囲（信頼区間）より狭くなくてはいけない，というわけである．

このような範囲の設定は，日本の従来の議論には無かったものであるから，ICH ガイドラインの実効化においては，各分野でその範囲をいくらにするか議論し，合意を得なければならない．大きな問題である．これについての積極的な提案は私にも無いが，統計家がしばしば口にするような，「それは固有の分野の知識に基づいて決めることであって，統計学が決めることではない．だから統計家は関与しなくて良い」，ということではないと思う．実対照薬についてはすでに情報や資料があるのだから，統計家も，固有の分野の医学者や臨床医と共にそれを分析し，薬効発現における統計的・誤差的変動の大きさを評価し，有害作用の程度を参考にし，基準を作るのに寄与すべきだと考えている．どうだろうか．

5. 非劣性試験における対照

5.1 非劣性試験の意義

非劣性という概念は，それを明らかにする必要性がいろいろな機会に指摘されていながら，ICH ガイドラインに採用されるまで表だってはとりあげられなかった．実際，ICH ガイドラインの下敷きになった欧州ガイドラインはこれを区別していないし，日本では，非劣性を確かめることに「同等性検証」というラベルを貼り，非劣性を認めたことを「同等性を立証した」などと言っていた．ところが実際には，非劣性と同等性の間に大きな違いがある．少なくとも今までの日本の承認条件は非劣性での検証を求めており，それが米国が原則としている優越性と違った視点からのものであることは，十分認識しておかなくてはならない．

承認条件において，日本は優越性試験と非劣性試験を原則として採用し，米国は原則として後者を認めずプラセボ対照にたいする優越性試験を採用しようとしている．それには，診療費，薬価決定，行政官庁の責任範囲などについての日米の行政体制の違いが反映しているようである．

日本では，多くの患者が健康保険制度を利用している．そこでは薬価が国によって決められる．その際，薬効が劣っているという理由でもって特定の薬剤の値段を低くすることはない．もし明らかに市販薬より薬効が劣るものを認可したら，それについても患者は通常のルールで定めた薬価でのお金を払うことになる．そこで日本の薬事行政は，原則として「市販薬より劣ったものは薬剤として承認しない．少なくとも劣っていないもののみを承認する」という方針を取る．これにたいして米国では，薬価が自由市場で決められる．多少でも薬効があれば安いという利点を患者が利用できる．「薬効さえあれば薬剤として市場に出して良い」ということになる．非劣性試験を採用するかどうかにはこのような状況の違いが影響しているようである．

5.2 実薬対照の倫理性

すでに述べたように，実薬対照をどちらかというと肯定的にとらえるか，どちらかというと否定的にとらえるかで，日米間にかなりの違いがある．それが最もはっきり現れるのが非劣性試験である．非劣性試験には実薬対照が不可避だからである．

日本では一般に，申請製剤と同種同効の市販薬があるときには，その中の最も信頼性のある

薬剤を実薬対照として非劣性試験を行い，それで非劣性が認められれば申請製剤を承認する．プラセボ対照を使って優越性試験を行うのは倫理的に許されがたいとするのが普通である．

これにたいして欧米は(より正確には欧米からのICHガイドライン検討委員は，というべきかもしれないが)，薬効が証明されていない市販薬を使うのは非倫理的で，特に非劣性試験で使う実薬対照は優越性試験で薬効が確かめられたものに限るべきである．しかもその試験条件は原則として優越性を確かめた試験と同じでなければならない．そうでないときは非劣性試験にプラセボ対照を入れ，非劣性と同時に優越性の試験を行うことが必要だとし，非劣性試験での承認を厳しく制限しようとする．両者の食い違いには，倫理性と対照の選択という二つの問題が密接に絡んでおり，しかも両者を現実問題で分離するのはむずかしい．それが議論をわかりにくくしているのであるが，本節では倫理性に焦点を限って考えてみよう．

有効性が証明されている薬剤があるときにプラセボを処方するのが非倫理的であることは，かなり多くの人が認めている．症状の変化が緩慢であったり，致命的でなかったりするなら，優越性について確実な証拠を得るためにプラセボを使っても非倫理的と言うに当たらない，と主張する人は少数派である．

しかし有効性の証明の条件を狭くしようとするか広くしようとするかには，明らかに人による意見の違いがある．優越性を証明したときの年齢・症状などの患者選択条件，目標とした疾患・症状の範囲，用いた主要評価項目などについてどの程度枠を広げて，「市販薬を使わないのは非倫理的」と言い得るかは微妙である．

逆に，優越性試験では有効性が証明されていないが，かなり多くの臨床医が薬効の存在と副作用の不在を（主観的，経験的に）認めて使っている市販薬があるときに，プラセボ対照を使うのが非倫理的かどうかは逆の意味で，意見が分かれる微妙な問題である．こういう例に直面すると私は，臨床医に「経験・主観」なるものの内容の吟味を十分にしてもらい，私のような無知な人間にもそうと感じられる程度にその説明をしてほしいと感じる．少なくとも同じ分野の複数の臨床医の間で合意が得られないようならば，プラセボを使うべきだと思う．どうだろうか．

5.3 実薬対照の妥当性の検証

非劣性試験では，実薬対照として市販薬が採用される．日本の新薬承認審査では，その対照薬が本当に市販薬の中で薬効が最も大きいものであったかどうかを吟味する．その対照薬が優越性試験でプラセボより有意に薬効が高かったかどうかを吟味することは（少なくとも今までは）比較的稀である．

ところがICHガイドラインはこの点をかなり神経質に問題としている．まず，実薬対照は優越性試験で薬効が認められているものでなければならない．これは当然である．問題は実薬対照が認められた過去の臨床試験の試験条件と，現に行おうとしている臨床試験の試験条件の食い違いである．これがどの程度まで許容できるかである．

たとえば昔の胃炎の治療薬での主要評価変数は，たいてい，自覚症状の改善度に基づくものであった．しかし現在では内視鏡での結果に基づく変数の方がはるかに精度の高い，内容の明らかな判定となる．このような技術の変化を無視して，対照薬の有効性が認められた試験条件に固執したのでは，明らかに薬効のある実薬対照を使わずにプラセボを使うことになる．これは現在の市販薬より劣った申請製剤を承認する結果をもたらし，望ましくない．そこでICHガイドラインでは次の注意を記述に入れることにした．

“the new trial should have the same important design features as the previously conducted superiority trials in which the active comparator clearly demonstrated clinically relevant

efficacy, taking into account advances in medical or statistical practice relevant to the new trial."

この最後の1節で，医学あるいは統計学の進歩によって変更するのが当然となった条件については，適切な配慮をすべきであると注意している．

6. 非劣性試験における判定

非劣性試験で判定を下す手続は，どのような視点で構成しかつ理解すべきだろうか．これについては今までにいろいろな議論があり，まだ決着がついていない．それは仮説検定という推測形式の特異性がモデルと判定手続の間にある種の乖離をもたらし，それを正当化するのにいろいろな考え方が持ち込まれているためである．統計的データ解析におけるモデルと判定手続の乖離については，別の機会に一般論をすることにして，ここでは非劣性試験を主にした注意点を述べる．

6.1 平行群試験での判定手続

日本ガイドラインは，平行群試験，つまり試験参加者をランダムに2群に分けて，それぞれに申請製剤と実薬対照を処方する試験の場合について，「たとえば」という形で次の示唆を行っている．

「新医薬品の承認審査に当たっては，原則として治験薬（本稿での申請製剤のこと）は当該効能を有する既承認の薬剤よりも優れているか，または同等と評価された場合にのみ承認されるべきである．比較試験において治験薬が対照薬と有意差がなければ直ちに同等であるとするのは統計学的に問題がある．本指針では臨床的な実用性を加味した観点から，許容範囲を超えては劣っていないことを積極的に証明する方法を使用するよう提案している．」

「臨床的同等性（本稿での非劣性のこと）を立証するには，たとえば，有効率の場合には次のような統計的方法が考えられる．ただし，以下のΔは，治験薬と実対照薬（本稿の実薬対照）の臨床的に許容できると考えられる有効率の差を予め設定しておいたものであり，疾患，薬効群により具体的な数値は異なるものの，例えば10%が一つの目安となる．

a. 治験薬と実対照薬の有効率の差の信頼区間（90%信頼区間）が$-\Delta$より小さい領域を含まない場合には，臨床的に同等と評価する．
b. 「治験薬の有効率が実対照薬の有効率に比してΔ以上劣る」という帰無仮説に対する片側仮説検定（有意水準5%）を用い，これが棄却されたときに臨床的に同等と判断する．」

この判定手続の提案は，それ以前にかなり広く普及していた「有意差がなければ同等と見なすべき」という「仮説検定の結果への不適切な理解」をとがめるためになされた．そしてこの手続は，その意図した役割をある程度果たした．しかし一方でこれは，このガイドラインの作成者が意図しなかった解釈を生じさせた．それはこの手続の立脚点が，申請製剤の有効率が実薬対照より低くてもその低さが10%までなら，申請製剤を実薬対照と同等とするという解釈である．

一般にΔなどと俗称されるこの種の「差」には二つの解釈があり得る，と私は考えている．一つは，文字通り「医学的に容認可能な差」であり，他の一つは，実質的に「医学的に容認可能な差」のものが薬剤として承認されるような手続を構成するための操作的な差である．前者であればその差の設定に統計家が関与することはほとんどできないが，後者であればその差の設定に統計家が寄与することができる．いやむしろそれに積極的に寄与しなければならない．広

津（1986）や椿（1994）の議論を読む限りでは，上に述べたΔ=10%という値は後者の視点に立脚したものとするのが妥当である．少なくとも私はそれが正しいと信じている．これが「意図しなかった解釈」と私が述べる理由である（この議論は同等性試験についても生じる）．

その確信はもちろん，上に引用した論文だけから来るものではない．それは臨床医が一般に有効率が10%近くも低い薬剤を同等とすることはない，と考えるところからも来ている．たとえば有効率が70%と80%の薬剤，つまり10人の患者の7人が直る薬と8人が直る薬があったとき，「10人で7人直る薬と10人で8人直る薬とは同じなんですよ．ですからあなたには7人しか直らない方のお薬を上げても良いですね．」と言える医者がいるだろうか．私はそんな医者はいないと信じているし，有効率が10%も低い申請製剤は他の利点がない限り承認すべきでないと信じている．この意見に反対する臨床医がいるだろうか．確信はさらに，広津(1986)，椿(1994)，遠藤 他（1996）が説明しているように，この手続は有効率が同じかそれ以上であることを判定する際の，誘導効果を入れた大変合理的な手続であることからも来ている．私は，その合理性を認めて臨床医もこの手続を受け入れているのだと考えている．帰無仮説を形式的・機械的に解釈して，10%も劣る薬を受け入れることだとするのは，現実の理解として的を射ていないと思う．

たとえば，理論家からよく出る批判は，「上記論文などの主張は試験参加者数を各群100程度にしかできないという前提から来ている．そんなことはない．各群数百例の臨床試験はいくらでもある．」というものである．そういう人は実際に検出力を計算してみるのがよい．たとえば，1群100例を1群400例にすると，標準偏差は半分になる．そうしても有効率が5%程度低い申請製剤が承認される確率（検出力）は50%程度にしかならない．そういう薬剤を確実に非劣性試験で承認させるためには，各群800例あるいはそれ以上にしなければならない．そんな自信のない製剤にそんな大変な臨床試験を計画することがあるだろうか．

椿（1994）が述べているように，10%という値は「試験の点数における下駄」というニュアンスで決められている．その下駄は臨床医の経験と直感に依拠して決めたもののようである．その経験と直感は，現実にこの方式を適用すると臨床試験で有効率が3%も劣ったものは承認に至らないという事実に裏付けられていると思うのだがどうだろう．もしそうなら，臨床医は，実際のデータで申請製剤の有効率が対照の有効率より大きいときのみ承認がされることを見ていることになる．データで10%も劣っているものを認めているのではないと言ってよい．

この問題は，統計学におけるモデルと実際のデータ解析における手続の間に，乖離があることを意味している．大学などで仮説検定の講義をするときの公式的解釈をそのまま適用すると，その理解は実際的意味（physical meaning）とずれるのである（このずれについての統計家としての意見は別の機会に論じるつもりである）．

品質管理の分野で抜き取り検査の設計に詳しい技術者はこの種の問題に，「Δ=10%以上劣る申請製剤をどうしても不承認，Δ=0%以上は劣らない申請製剤を承認，ということで第1種の過誤と第2種の過誤の限界をたとえば5%，10%と定めて試験参加者数を定め，それ以上は試験参加者数を増やしてはいけない，ということにすればよいのではないか」という提案をしたりする．この発想を活かすなら，たとえば「有効率が実薬対照より3%以上劣る申請製剤をどうしても不承認，有効率が実薬対照より6%以上優れている申請製剤を承認，ということで第1種の過誤と第2種の過誤の限界をたとえば5%，10%と定めて試験参加者数を定め，それ以上は試験参加者を増やさない」ということにすればよい．もちろんここで入れた，3%，6%という値はただの数値例で，実際には申請薬剤の他の利点・欠点を配慮して決めるべきものである．この場合，中間の質の製剤は偶然不承認になったり承認になったりするが，それは申請者のリスク判断によるから，手続としては合理性を保っている．非劣性試験の計画としてこの提案は考慮に値すると私は考えているのだが，読者の皆さんはどうだろうか．

6.2 多施設試験における非劣性の確認

上に示した平行群試験だと，前節の最後で述べた「モデルと手続での解釈の乖離」は気にならない程度である．ところがときにこの乖離にかなりの注意をしなければならないことがある．多施設試験の場合や対応のある試験の場合である．偶然的変動がかなり大きくなったり，極端に小さくなったりするからである（広津（1995），佐藤（1994 b）を参照）．

現在の日本では，主要評価変数として全般改善度での有効率を採用することがきわめて多い．歴史的経緯もさることながら，その最も大きな動機は，申請側の保守的な態度，すなわち全般改善度ならば従来の経験でどのようなデータを出せば承認が得られるかが分かっているからのようである．主要評価変数として全般改善度を用いるときは，判定手続も2項分布を前提にした単純なものを使うので，判断に影響する統計量はそれぞれの群の平均だけである．施設間差として問題にされるばらつきの指標は無視される．したがって，全般改善度を主要評価変数としているときには，非常に多くの施設を参加させることで現実に生じているばらつきをデータに反映させることができ，非劣性の証明における一般化可能性（generalizability）が確保できるという議論は成立しない．

問題になるのは，なんらかの計量値変数を主要評価変数にした場合に，非常に多施設の参加があり，したがって各施設からの試験参加者数が少数になる，という場合である．この場合には，患者の背景が違う，医療の質が違う，臨床試験の質管理が悪くなる，などの原因で誤差的変動が大きくなる（佐藤（1994 b）参照）．

ここでモデルと判定手続の乖離が問題となる．2値データの場合のように誤差的変動の大きさを配慮せずに，医学的に容認可能な差の範囲を定めると，多分それはかなり小さい幅になる．その幅の下で少数参加者多数施設という臨床試験を行い，原因を問わずに誤差的変動をひとまとめにすると，申請製剤が承認される可能性はかなり小さくなる．ずさんな臨床試験をすると承認がされにくいという内容になる．ある意味で望ましいことである．

しかしこれにたいして，多数参加者少数施設の臨床試験を行ったらどうだろう．一般には，誤差的変動が小さくなり，医療の質がよくなり，薬効がある申請製剤が比較的高い確率で承認されることになる．もちろん，薬効のない申請製剤は間違って承認される可能性が低くなる．問題は，質の悪い医療が現実に存在するのに，それが反映されずに申請製剤が承認されることをどう評価するかである．患者の立場に立てば，そういう申請製剤は承認された方が良いだろうか，悪いだろうか．

患者にとって望ましいことは，申請製剤が自分にとって薬効を持っているなら承認され，薬効を持たなければ承認されないことである．そのときに，「自分の主治医の医療の質が悪いときには承認されない方がよい」，という考えが成り立つだろうか．私はそうは思わない．本当に悪い医療の質があるなら，それを発見して教育その他の施策で改善することを試みるべきである．別の言い方をするとこれは，誤差的変動の内容がなるべくつかめるような試験計画を立て，試験製剤の薬効はそれ自体として評価でき，同時に医療の質，患者背景と薬効との交互作用についての情報も得られるようにすることである．この方が正道でなかろうか．

多施設試験については，長所短所を巡っての激しい議論がある．その議論においては，特別な教育を受けていない消費者でも問題なく使えることが商品の必要条件であるといったときの製品と，必要な教育を受けて免許制度で資格を制限されている医者しか使えない製品とを，同じように見るのは避けた方がよいと私は考える．その区別をはっきり認識した上で，患者背景，医療の質，試験の質管理等のためにどんな注意をするか議論した方が建設的であろう．

6.3 対応のある試験での非劣性の判定

対応のある試験というのは，患者のそれぞれに両腕，両眼などに別の処置をしてその差で比

較結果を評価するものである．形式的には，2×2クロスオーバー試験と同じになるが，順序関係による残存効果が無くて，代わりに相互作用があるかもしれないという違いがある．クロスオーバー試験のときに残存効果が生じないよう十分長い冷却期間を設けるのと同様に，相互作用が生じないように十分間隔をあけ，境界の状況で相互作用を吟味することが必要になる．

対応がある試験では，前項の多施設試験の場合と逆の問題が生じる．クロスオーバーの場合と同様，個体差の影響が取り除かれる関係で，偶然的な変動が非常に小さくなる．特に主要評価変数が2値の場合にそれが顕著となる．

平行群試験と同様に，たとえば有効率で10%も劣ることを許容すると，わずか数十例の試験で本当に有効率が10%も劣ったものが承認されることになる．すでに述べたように，この場合の10%の差は医学的に容認できない大きなものでないかと考えられる．前項に述べた私の解釈では，“医学的に許容できる差”がしばしば操作的に定められているから，平行群試験での値を機械的に適用することはせず，平行群試験と異なった原理で“医学的に許容できる差”を定めることが必要ではないかと考える．

なお，現実にこのような状況があったときには，偶然変動が小さいことを前提にして，試験参加者数を小さくする傾向がある．これはある種の安定性の欠如を伴いやすいことが日本製薬工業会からのICHガイドライン検討委員から指摘されている．ICHガイドラインでは特に注意をしてはいないが，これも試験の遂行においては考慮すべきことであろう．

6.4 非劣性試験におけるPCとITT

試験を理想的に計画しても現実の試験参加者は生身の人間である．いろいろな原因・理由によって試験の計画が守れないことになる．そのような人を解析対象のデータに含めるか否か，含めるとするとどのように扱うか，これについていろいろな方針が存在する．一つの極は，試験計画に忠実に行動した参加者，いわゆる適合例（protocol-compatible subjects），のみを評価対象にするというものであり，他の極は試験でランダム割り付けに含めた（intent-to-treat subjects）全員を評価対象にするというものである．一般に，前者に基づく解析をPC解析，後者に基づく解析をITT解析という．ただし，ICHガイドラインでは，この用語法（PC analysis, ITT analysis）を用いていない．より細かい区別が必要という視点からである．

日本では，最初PC解析が良いと考えられていた．この方が科学的に正確に差を測定できると考えたからである．ところが欧米では，むしろITT解析の方を重要視すべきだという見解が強かった．試験計画違反や，試験からの脱落，それ自体が薬効や安全性についての大きな情報を含んでいるからであろう．日本でも両者の相違についていろいろな機会に議論がなされた（計量生物学会（1995）参照）．そういう議論において，しばしば出される誤解は，ITT解析の方が申請薬剤の有効性を認め難いという意味で保守的であるというものである．一般論は別の機会にするが，同等性や非劣性試験の場合には，ICHガイドラインがこれについて次の注意を与えている．

“The equivalence (or non-inferiority) trial is not conservative in nature, so that many flaws in the design or conduct of the trial will tend to bias the results towards a conclusion of equivalence. For these reasons, the design features of such trials need special attention and their conduct needs special care. For example, it is especially important to minimize the incidence of violations of the entry criteria, non-compliance, withdrawals, losses of follow-up, missing data and other deviations from the protocol, and also to minimize their impact on the subsequent analyses.”

一般に，試験計画からの外れがあるということは，比較される処置の違いを分からなくする．

それなのにそれを解析に入れると，参加者数が多いということで見かけ上は精度の良い試験をしたことになる．そうすると同等性あるいは非劣性試験の結果を，承認の方向に偏らせることになる．だからといってそういうものを除外すると，重要な情報をある方向で除外することになる．要するに偏りという意味では試験の質を上げて脱落例を減らすことに努力するのが正道であって，どちらかが良いというものではない．ITT 解析の方が保守的であるというのは，同等性や非劣性を評価する場合には誤解なのである（たとえば，佐藤（1994 a）参照）．

7. 用量反応性試験

用量反応性試験は，申請製剤あるいは実薬対照にいくつかの用量を設定し，それに対応する複数の群の結果を比較して，用量反応関係について判定を下すものである．薬効や副作用に関する薬剤間の相対的位置関係の推定など，探索的目的をいくつか相乗りさせた試験であることが多く，細かく分けると型が非常に多くなる．

これについての記述が ICH ガイドラインでごくわずかなのは，検証的比較試験と言いにくい側面が多いからである．実際，この臨床試験での解析法としては信頼区間や図表示が求められている．もちろん仮説検定も使う場面があるが，そのときは評価変数が用量に関して単調であることや直線的に変化することといった条件を取り入れた手法を使う必要があろう．

8. おわりに

検証的比較臨床試験の区分けを糸口にして，関連する問題のいくつかを論じたが，これらはまったくの私見である．たまたま私が厚生省で新薬申請の資料の調査と評価に関係しているからといって，この私見を直接資料の評価に反映させるわけではない．かって東京都公害局長であった故田尻宗昭氏が常々口にしていたように，私も，行政上の仕事をする者は法律（ここではガイドライン）に依拠した行動をとるべきであると考えている．したがって本稿での見解をそのまま新薬審査での判断に直結させることはない．現実問題については多くの方と議論をし，意見調整をした後で，それにしたがった判断を下すことになる．本稿の内容は全く私見であることを再度強調したい．

ICH ガイドラインのよりよき理解を目指したつもりであるが，結果としてはかえって議論を呼ぶものになったかもしれない．願わくば感情的でない批判をいただいて議論の掘り下げを試みたい．そういう意図を汲んでいただければ幸いである．

本稿の内容については，統計数理研究所の佐藤俊哉助教授との意見交換が大いに役立った．また，中央薬事審議会新医薬品第一調査会の委員の方々から常々いただいている医学面でのご教示も役に立っている．査読者からは，誤解の生じやすい表現についてのご注意をいただいた．ここに記して謝意を表したい．

参 考 文 献

遠藤 輝，吉村 功，森川敏彦，柳川 堯 (1996)．臨床試験における対応のあるデータでの有効率の「同等性検証」の一方式，応用統計学，**24**, 59-73.

D'arcy, P. F. and Harron, D. W. G. (ed.) (1996). *Proceedings of the Third International Conference on Harmonization — Yokohama 1995*, The Queen's University of Belfast, U. K.

FDA (1988). Guideline for the format and content of the clinical and statistical sections of new drug applications (1991,『申請書類の書き方に関する FDA ガイドライン——臨床と統計——』(翻訳：日本製薬工業協会臨床評価部会，臨床試験における統計学的諸問題検討分科会)，ライフサイエン

ス出版，東京)．
広津千尋 (1986)． 臨床試験における統計的諸問題(1)——同等性検定を中心として——，臨床評価， **14**, 467-475.
広津千尋 (1995)．比較臨床試験解析——最近の話題から——，臨床評価，**23**，489-521.
鹿庭なほ子(1997)．生物学的同等性試験について——新しいガイドラインに向けて，第67回医薬安全性研究会定例会資料，20-31.
計量生物学会 (1995)．臨床試験におけるプロトコール逸脱例の諸問題，第3回計量生物セミナー予稿集，1995.10.28-29，富士教育研修所，東京．
厚生省 (1997)． 医薬品の臨床試験の実施の基準に関する省令，臨床評価，**25**, Suppl. XII, 17-31.
Lewis, J. A., Jones, D. R. and Rohmel, J. (1995). Biostatistical methodology in clinical trials—A European guideline, *Statistics in Medicine*, **14**, 1655-1657.
光石忠敬 (1996)．統計解析法に関する諸問題，『臨床試験』(内藤周幸 編)，123-152，薬事日報社，東京．
森川敏彦 (1994)．同等性問題再考，計量生物学，**15** (2), 111-140.
佐藤俊哉 (1994a)．ランダム化にもとづいた intent-to-treat 解析，応用統計学，**23**, 21-34.
佐藤俊哉 (1994b)．椿口演にたいする討論，計量生物学，**15** (2), 161-163.
Sato, T. and Yoshimura, I. (1998). Expectations for an international harmonized guideline, *Drug Information Journal*, **32** (to appear).
新医薬品統計解析指針検討会(1992)．臨床試験の統計解析に関するガイドライン，『新薬臨床評価ガイドライン1995』(日本公定書協会 編)，68-103，薬事日報社，東京．
椿 広計 (1994)．同等性推論の様々な不都合について，計量生物学，**15** (2), 141-160.
World Medical Association (1989). The "Declaration of Helsinki"，『新薬臨床評価ガイドライン1995』(日本公定書協会 編) 588-593，薬事日報社，東京．
吉村 功 (1996)．統計解析法に関する諸問題，『臨床試験』(内藤周幸 編)，94-122，薬事日報社，東京．
吉村 功，魚井 徹，佐藤俊哉，上坂浩之 (1997)．ICH E 9 ステップ2ガイドライン——臨床試験のための統計的原則，薬理と治療，**25**, Suppl. 4, 1-94.

Proceedings of the Institute of Statistical Mathematics Vol. 46, No. 1, 81-95 (1998)

Comments on Design Considerations in Controlled Clinical Trials for Confirmatory Purposes ——For Better Understanding of "Statistical Principles for Clinical Trials"——

Isao Yoshimura

(Faculty of Engineering, Science University of Tokyo)

This paper addresses some design issues related to comparative controlled clinical trials for the purpose of new drug approval. It is intended to provide a complementary interpretation of the section III of the "Statistical Principles for Clinical Trials", which is now in the state of final draft of the International Conference on Harmonization. First, it notes that Japanese general tendency which prefers to using active controls in the intention of showing the superiority of investigational product is not always considered as rational. Secondly, it explains the relation between the confidence level of interval estimation and the significance level of hypothesis testing because it often becomes a source of dispute. Tertiary, it pointed out that the equivalence margin should be clearly described in the protocol of non-inferiority trial with the rationale of the margins, paying a special attention on the margins observed in previous superiority trials. Fourthly, it asserts that the notion of non-inferiority trial is introduced in the ICH guideline because the use of an well established active control is ethical compared to a placebo control and is consistent with the medical insurance system of Japan, although the confirmation of the efficacy of the active comparator should be strict in the sense that the condition of contemporary clinical trial must be similar to that of the previous superiority trial. Finally, it explains what kind of effects should be taken into consideration when a multi-center design is adopted or the analysis set is changed regarding dropouts and missing observations.

Key words: Active comparator, clinical trial, equivalence margin, multi-center trial, non-inferiority.