

誰がための臨床統計？ わが国で実践された「患者の立場」からの 臨床評価の原則と統計的方法の役割[†]

筑波大学大学院* 椿 広 計
国立公衆衛生院** 藤 田 利 治
元東京大学医学部 佐 藤 倚 男

(受付 1997 年 10 月 23 日；改訂 1998 年 1 月 16 日)

要 旨

わが国の科学的な臨床試験は、1970 年代初頭に医薬品評価に責任のあった臨床家が構築した臨床試験の原則に基づいて推進されてきた。すなわち、臨床試験情報の偏りの防止、使用者指向の原則とその実践、臨床試験の標準化である。こうした臨床試験の原則は多くの臨床家から受け入れられて標準的なものになったが、一方で、その一部の形式のみを借用した形骸化した臨床試験が次第に多くを占めるようになった。こうした中で、日・米・欧医薬品規制ハーモナイゼーション国際会議の下で「国際的に通用する臨床試験」が目指されて、これまでのわが国の臨床試験批判と相俟って、制御された環境下での新薬の有効性に関する「科学的」仮説検証が強調されている。本稿では、従来のわが国臨床評価の原則の意義を明確にし、最近の国際化の動向がわが国の臨床試験にどのような変質をもたらす危険があるかを指摘した。

キーワード：新薬の臨床評価，コントローラー，品質保証システム，有用性評価，多施設臨床試験。

1. はじめに

本稿では、新薬の臨床評価にあたって、必要な質を有した基本情報を提供するには、臨床試験がどのようにあるべきかということと、それに対する統計的方法の寄与について、議論する。

まず、わが国の臨床家が 1970 年代初頭にどのような臨床評価の原則を構築し、その論理構造に合致した臨床統計的方法論を設計したかについて、その歴史を総括する。

第二に、この臨床評価の原則は受け入れられ、事実上標準的なものになった一方で、原則の一部を形式的に借用した形骸化した臨床試験が次第に多くを占めるようになった歴史についても簡単に触れる。

第三に、当初の基本原則に基づく約 20 年の実践の中で、どのような問題が生じ、統計的方法

* 経営システム科学専攻：〒 112-0012 東京都文京区大塚 3-29-1。

** 疫学部：〒 108-8638 東京都港区白金台 4-6-1。

[†] 117 頁より討論あり。

の手直しの必要性が生じたかを紹介する。その対策として、臨床家と統計家の共同作業の産物として制定された「臨床試験の統計解析に関するガイドライン(1992年)」の意義についても述べる。

近年の動向として、日・米・欧医薬品規制ハーモナイゼーション国際会議(以下、ICH: International Conference on Harmonization of Technical Requirements for registration of Pharmaceuticals for Human Use)の下で「国際的に通用する臨床試験」が目指されている。本稿では最後に、この国際化が従来のわが国の臨床評価の原則に対して、どのような変質をもたらす可能性があるかを指摘したい。

なお、議論の前提として強調しておきたい点は、臨床試験のあり方に対する要求が2つの立場、すなわち、新薬の「生産者(製薬企業)の立場」と「使用者(患者)の立場」によって厳然と食い違うことである。また、そのどちらか一方を科学的真実とは、判定できないということである。そして、筆者らは「使用者の立場」からの要求を保証する方法論確立に、特に関心がある。この意味で本稿の展開は、通常の中立的統計科学の方法論とは本質的に異なり、品質マネジメントの方法に基盤を置いている。本論に入る前に、新薬の臨床評価に関する筆者らの基本的認識を明示しておこう。

新薬の臨床試験は、通常新薬の市販を意図した「生産者」がスポンサーとなって計画・実施する。一方、「使用者」の代弁者(臨床家、統計家、規制当局などを含む)が可能なのは、生産者による臨床試験からの情報に基づいて、新薬を市場で使用して良いか否かを評価することのみである。新薬についての情報を自ら作り出せない以上、使用者側には本来、評価に値する情報が臨床試験から得られることを保証する枠組みを生産者側に課す権利がある。さらに、使用者の代弁者は、この使用者の権利に係わる適切な要求を継続的に行う責務がある。

さて、このような社会的枠組みの中で、新薬の臨床評価という行為に統計的方法を適用する意義は何であろうか。筆者らは、市販前の最終段階、すなわち第三相臨床試験など検証的臨床試験における統計的方法の最終目的は、「消費者危険(consumer's risk)」に対する保証だと考えている。ここでいう消費者危険とは、新薬の最終使用者(end user)である患者が有用性の乏しい薬を投与される危険、すなわち、有効性の乏しい薬や有効性に比べて安全性に問題のある薬を投与される危険である。つまり、日常期待できる治療水準を下回る治療を消費者が受ける危険であり、消費者が被る迷惑である。逆に、よく効く薬が日の目を見ないことも社会的な損失ではあるが、このような薬が世に出ないことは消費者の現状を悪化させるものではなく、いわば「生産者危険(producer's risk)」と呼べるものである。

規制当局による新薬の許認可システムの統計的側面として重要なのは、消費者に代わって消費者危険が生じる確率を定められた水準以下にすることである。「臨床評価に関する情報は生産者から提供され、消費者はそれをチェックする程度の防備しかできない」という両者間の情報量の非対称性を補填するために、生産者に課している責務がこの規制当局による新薬の許認可制度ともいえる。この種の制度としては、他分野で著名なものとして、例えばエコラベリング制度を挙げることができる。

本稿では、上記の立場の下で臨床評価に対する一つの視点を提供する。勿論、生産者危険の最小化も重要な問題であるが、これは開発側にとっても大きな関心事であることから、それ自体に合理的な方法論が展開されている。

2. わが国における臨床評価の原則

2.1 科学的な臨床評価の曙

新薬が安全かつ有効であることの科学的な証拠を要求するようになったのは、わが国では、

厚生省による1967年の「医薬品の製造承認等に関する基本方針」が契機といわれている。この中で初めて、臨床試験成績資料は「精密かつ客観的な考察がなされているものであること」の必要性が指摘された。この基本方針は、サリドマイド事件を契機として1962年に改正された米国の食品・医薬品・化粧品法（Kefauver-Harris 修正法）を受けたものであった。

基本方針が出た頃のわが国は、長期間にわたって輸入薬に依存し、評価は権威者の経験や輸出国の研究に依存していた。二重盲検法を用いた臨床試験自体は、1957年以降、抗結核剤、精神安定剤、精神薄弱へのグルタミン酸などで先駆的に実施されてきた。しかし、新薬の承認審査での申請資料のほとんどは使用経験の集積に過ぎなかった。1965年当時の中央薬事審議会新薬調査会では、提出資料の信用性についての激しい議論がたびたび繰り返され、「こういうあいまいな報告はいかにたくさん集められても国として販売承認を与えて良いかどうかは判断できない」「こうやって治験例を漫然とかき集めさせるのではなく、二重盲検法による科学的で納得できる論文を提出させなければ審議できない」といった批判が交わされていた（佐藤（1982））。

1971年7月には、厚生省薬効問題懇談会より「薬効問題懇談会の答申について（以下、薬効懇答申）」が出された。この中で「医薬品評価のあり方」が示され、その方針は現在にまで及んでいる。また、1967年以前と比べて、「現在…（略）…精密かつ客観的な観察が要求され、特殊な医薬品を除き、原則として二重盲検法等の比較試験法を採用した治験成績が重要な資料となっている」との認識も示されている。しかしその一方で、1970年に日本学術会議医薬研連の設置（月間薬事（1970））を経て、1972年に日本学術会議から内閣総理大臣に対して「医薬品の臨床試験評価に関する体制の確立について」の勧告（佐藤（1974））が出されており、「医薬品の臨床試験評価に関し政府は公正なチェック・システムを設けること、被験者に対する救済手段を充実することをはじめ、…（略）…緊急に適切な措置を講ずるよう勧告」されている。科学的な医薬品評価の方法論の普及に対して、それを適正に機能させる社会システムが未整備であり、その体制整備が模索されていた当時の状況が窺える。

こうした時代の中で、薬効懇答申および日本学術会議勧告の精神の実現を目指し、厚生省中央薬事審議会新薬調査会メンバー有志を中心として、1970年8月にコントローラー委員会が活動を開始し、1972年に正式に発足した。その後、委員会は常設の事務局機能を整備すると共に、臨床家のみならず、前臨床試験成績のレビューの専門家、臨床試験倫理など法律の専門家、統計家などを加え、ボランティアな活動を発展させ、今日に至っている。

委員会発足の背景には、新薬調査会に提出された資料に「こう書いてあるから、これでいいのだ」といった書類審査による責任回避の立場に終始することに満足せず、信用できる臨床試験データを自ら作り出し、患者が被る実際の危険・迷惑を最小限に抑えるという強い意思があった。

2.2 臨床評価の原則の確立

コントローラー委員会が推進した臨床評価の原則はわが国の臨床試験に大きな影響を与え、事実上の標準として現在に至るまで機能してきた。椿は、1994年8月東京で行われた the 1st Annual Biostatistics Meeting of Drug Information Association での講演において、品質マネジメント視点からその原則を次の3つにまとめた。すなわち、「情報の偏りの防止」、「使用者指向」、「標準化」である。

以下では、1970年代に確立された臨床試験の原則について、今日的な視点を含めて総括する。

2.2.1 臨床試験情報の偏りの防止

臨床試験の多くは生産者側のスポンサーシップで実施されており、薬効懇答申の中でも二重盲検比較試験の必要性が論じられているように、その公平性の確保は1970年代から大きな問題であった。そこで、盲検下の症例検討や生産者によるモニタリングなどでの作為が働くことを

避けるために、欧米のように治療法の無作為割付けを生産者に任せるのではなく、第三者が割付けを行い、臨床試験の公正性を保証する役割を果たす「コントローラー」というわが国独自の方式を開発し、定着させた。コントローラーは、1970年代以降次のような役割を行ってきた(内藤(1983))。

- ア. 臨床試験計画の使用者の立場からのレビュー
- イ. 薬剤の識別不能性の保証
- ウ. 無作為割付けおよび割付け結果の秘匿保管
- エ. 薬剤サンプルの抜き取り検査(規定含有量や崩壊性の確認)
- オ. 統計解析計画の妥当性のレビュー
- カ. 個々の症例を統計解析に採用することの是非を検討する手続きの公平性をレビュー(ブラインドレビューの保証)
- キ. 症例記録の回収と検査, その統計解析データとしての固定および関係者への公示
- ク. 主要な臨床評価項目に対する検証的統計解析の実施
- ケ. 臨床試験の論文原稿のレビュー
- コ. 論文公表(結果の正否に関わらず出版)

こうした第三者監視・介入による試験情報の公平性確保は、わが国独自のシステムであった。そして、欧米で許認可された薬がプラセボに有意に劣ることを明らかにしたり、実施における生産者の作為や抜け道による歪みの検出やデータ改竄の告発など一定の成果を上げてきた(佐藤(1982), コントローラー委員会(1992), 佐藤(1992))。

今日的に言えば、コントローラー・システムは、臨床試験情報の生成において「外部品質保証(External Quality Assurance)」のためのプロセスを第三者が管理・運営するものであり、第三者適合性評価機関の思想と共通するものでもある。

ここで国際的に認知されている「品質保証」という概念は、ISO 9000-1あるいは、その完全一致規格であるJIS Z 9900-1994によれば、「あるものが品質要求事項を満たすことについての十分な信頼感を供するために、品質システムの中で実施され、必要に応じて実証される全ての計画的かつ体系的活動」を意味し、特に外部品質保証とは顧客に信頼感を与える活動のことである(これに対して内部品質保証とは、経営者に信頼感を与える活動である)。飯塚(1996)は、通常の品質管理活動との取り違えを避けるために、この種の狭義の品質保証活動は「品質と信活動」と考えるのが良いとしている。また、「適合性評価(conformity assessment)」とは、製品またはシステムが技術的規則の要求事項に適合しているか否かの評価を意味している。(適合性評価, 品質システムおよびその監査に関する規格・ガイドについては、例えば日本規格協会(1997)参照)。

欧米の生物統計学や官庁統計学分野でも、近年、データ品質について議論されることが多くなったが、その多くは母集団規定、臨床検査値の品質管理、データのチェックシステムなどを中心としたものである(例えば、Liepins and Uppuluri ed. (1990))。これらも品質保証活動として有用ではあるが、「利害関係者(第一者)」である生産者の品質保証活動を透明化して使用者に信頼感を与える活動という点からは、「誰が」行うかという観点で不足している。データに偏りを生じさせる最大の要因は、モニタリングや症例検討など原データが変更される可能性のある機会に「利害関係者(interested party)」の作為が介入することにある。これを避けるために、わが国では欧米のように治療法の無作為割付けを利害関係者である生産者に任せず、上で列挙したように第三者たる「コントローラー」が割付けおよび割付け結果の保管に責任を持つ制度が定着したのである。欧米ではこの方式を「三重盲検」と呼ぶが、日常的には定着してはいない。なお、これに加えてコントローラー委員会独自の方法として、コントローラーを2名

指名し、両者が独立に所有する「暗号鍵」を同時に参照しない限り、単独のコントローラーでは割付け結果を知りえない「四重盲検」すら実施されてきたことも指摘しておきたい。

1996年に提示されたICH GCP（新医薬品の臨床試験の実施に関する基準，Good Clinical Practice）案において、「品質保証」とそれに対する「品質保証システムの監査」の概念が示され、国際的にも定式化されたことは一応注目に値する。しかしながら、外部品質保証のためのシステム監査は、外部審査機関による「第三者品質システム（マネジメントシステム）監査」が原則である。この種の保証活動が、第一者監査（内部監査）に留まるのであれば、「使用者に対する品質与信活動」としての意味を全くなしてはいない。わが国の医薬品業界以外（鉄鋼、化学、電気など）が指向しているように、第三者品質システム審査機関を適正に立ち上げることによって初めて、一応の品質システム透明化には寄与することとなろう。なお、こうした「第三者監査」については、「チェック・システムをつくり、またこれに必要な制度を設けること」が、既に1972年の日本学術会議の勧告で示されており、その後もコントローラー委員会は、一貫して第三相臨床試験などを監督する第三者組織の設立を提唱してきている。

ここで、「第三者適合性評価機関」の導入に比べて、「第三者マネジメントシステム監査（management system audit）」という手段は、監査報告の公表可能性の問題やパフォーマンス評価の欠如などの点で外部品質保証活動としては、本来劣った方法であることは強調しておきたい。さらに、他国に類を見ない定常的三重盲検確保に代表される、わが国の「適正なコントローラー」による、品質保証システムのコア・パート自体への第三者の実質的介入は、監査・適合性評価などの活動とは次元の異なる優れた保証システムである。1997年7月に国立医薬品食品研究所（旧国立衛生試験所）の中に医薬品医療機器審査センターが設立され、新医薬品の承認申請に関する諸審査を実質的に行うシステムが出発した。しかし、臨床データの品質は臨床試験プロセスの中で作り込まれるべきものであり、現状におけるこの制度の欠陥は、随時、臨床試験プロセスを審査監督しないことであり、審査に耐える臨床情報の品質保証としては不十分であるといえることができる。すなわち、1970年当時の新薬調査会のメンバーが共有していた強いモチベーションと、それに基づくデータ品質保証システム構築の努力が、この制度では欠如している。

さて、これらの活動以外にも特筆したいのは、コントローラー委員会が、臨床試験実施プロセスの体系的なマネジメントを依頼され実施した全ての臨床試験の成績を公表するという原則を実践したことである。この種の例外なき情報公開により、臨床試験情報を使用者の批判に曝す仕組みを確立したのも、わが国独自のシステムであった。このための雑誌として1972年12月に発刊されたのが、「臨床評価（Clinical Evaluation）」誌である。臨床評価誌発刊の辞（佐藤（1972））には、「本誌は薬効判定に止まらず、商品価値のないとされる無効論文や副作用報告、および臨床の他の分野の地味な研究を中心に組み、余力がある場合には対象疾患の認定や症状改善のさいの判定基準の標準化問題、それによる科学的診断治療と予後判定の問題、新薬に限らず、確率的予測の下に行われる治療一般に必然的に内在する倫理問題、さらには多くの統計的検定法が人間を対象とする場合にもつ意味と限界の問題など我々が直面している事柄を考え、かつ記録して行く予定である。特にスポンサーが消極的なために報告されるべくして、埋もれている貴重なデータを可能な限り掲載し、一部の専門家による情報独占の弊害を主体的に減少させる場として行きたい。」と明言され、実際患者の立場に基づく臨床試験の方法論に関する研究論文、総説も多く掲載された。通常の学術雑誌のレフェリー制度は実験・調査の結果だけを書類上審査しているが、「臨床評価」誌のレフェリーは当該臨床試験のコントローラーから選ばれるのが通常であり、論文化された試験の計画・実施プロセス・症例固定・統計解析をも審査している。海外では生産者側の利益となるか、研究者の関心のある成績のみ公表するという「出版バイアス（publication bias）」の問題が、以前から生物統計分野でも深刻な議論の対象

となっている (Pocock (1983)). コントローラー委員会が提唱した臨床評価の原則の一つは、この点でも情報の偏りの防止に寄与してきたといえる。

近年の SBA (Summary Basis of Approval) 制度は、薬事審議会による承認根拠を明示すれば臨床試験報告を論文化する必要がないという国際潮流に乗ったものだが、承認されたデータだけの公表は、既に Selection Bias を強く受けていることを忘れてはならない。

2.2.2 使用者指向の原則とその実践

わが国の臨床評価の最も基本的な原則は、「患者の立場」からの臨床試験の推進であった。患者が主人公であり、医師は患者に対するテクニカル・アシスタントと位置づけ、自分が患者になったときを想定しながら臨床試験の方針が検討された。こうした「使用者指向の原則」は、1980年代後半以降に一般的品質マネジメントの原則として定着したものであり、今日では「顧客満足度 (CS: Customer Satisfaction)」向上を最終目標とした品質改善活動は全世界で実践されている。

臨床評価における使用者指向の原則の実現に関しては、第三相試験によって社会における有用性を保証することが一つの要点であった。薬効懇答申での「医薬品評価の在り方」には、第三相は「第二相までの成績よりみて、それが医薬品としての有用性に期待が持てると判定されたとき、多数の患者について、有効性および安全性を明確にするために、厳密な試験計画に基づき大規模な臨床試験を行う」段階であり、第二相は「第一相の結果よりみて患者についての治験がほぼ安全であろうと判定されたとき、その治験薬を少数の患者について経験ある医師の慎重な監督下に、精密かつ客観的な検討が加えられるような臨床比較試験を行い、有効性の範囲を調べる」段階であることが明記されている。第三相試験は、まさに新薬が認可され、広範な治療現場に投入するに足る有用性を持つかどうかを検証する「実践的試験 (pragmatic trial)」あるいは「技術評価 (technology assessment)」と位置付けられている。すなわち、単に実験室的な制御された条件下で有効性の検証や安全性の保証を得るだけでは不十分なのであり、新しい技術、すなわち新薬を社会に導入した場合の挙動を予め評価すべきなのである。これは、それ以前の開発段階である後期第二相試験までが「科学的」仮説検証 (制御された条件下での新薬の有効性・安全性などの検証) を狙いとした「説明的試験 (explanatory trial)」あるいは「科学的実験 (scientific experiment)」の性格が強いことと対比される (臨床試験の統計解析に関するガイドライン: Q&A (1992))。すなわち、説明的試験 (「科学的」仮説検証) の上に実践的試験 (技術評価) が加わって、初めて新薬の有用性の科学的な根拠が得られるのである。

このため当時の第三相臨床試験に対しては、以下のような実施原則が臨床側の主要な要求となっていた。

- ア. 市販後の実際を予想して、患者対象集団を可能な限り広範囲にする。
- イ. 開業医なども含めた一般的な治療現場を偏りなく反映するために、多数・多種類の施設の参加を求める。
- ウ. 有用性が確立している薬が存在しているのであれば、可能な限り実薬を対照薬とする。
- エ. 治療方針 (投与用量、併用薬) などについても必要以上に束縛しない。薬の切替えが治療現場で予想されるのであれば、相互作用の点検などのため、wash-out など可能な限り避ける。
- オ. 試験に参加した患者の情報は、原則としてすべて解析対象とする。(現在でいうところの ITT (intent-to-treat) の原則)。

「使用者 (患者) の立場」からの評価指標としては、「有用性」が重視された。薬効懇答申の中にも、社会が望む薬として「有用性」の高い薬が挙げられている。「有用性の低い医薬品をそのまま放置することなく、積極的に淘汰を推進するために、再検討を国家規模で実施すること

は、有用性の高い医薬品の育成の面からも不可欠である」という理念は、現在でいえば「品質方針」を明示したものと考えられる。この品質方針を実施する方策として、第三相試験での評価指標として治療現場での治療法（新薬）の「質」の計測（評価）、すなわち「有用性評価」が中心となった。有用性評価とは、仮想的に同一病状の患者に対して治療法を再び行う状況に置かれたとき、患者の視点に立ってどの程度今回の治療法を繰り返すモチベーションが生じているかを評価するのが原則であり、患者自身による評価が基本となっている。臨床医には日常診療において常にこのような評価を行い、治療法を決定する責務があるし、これが診療の基本であることは古今東西を問わない。有用性という総合評価は、治療現場を反映した評価方法である（山本（1991））。さらに、臨床試験の計画段階では予期しなかった事象・作用を評価する必要性、質的に異なる作用を総合的に評価する必要性、臨床試験が実際の治療に比べて短期間であることや中止・脱落が不可避免的に発生するため医師の臨床経験から評価を補う必要性などが、有用性評価が重視された背景にあった。

なお、現在のCS（顧客満足度）調査の目的は商品の繰り返し購買（「次回も当社の製品を使用するか？」）行動の可能性を計測し、それに影響を与える要因を分析することであり、これと構造的に極めて類似している。1970年代以降一貫して有用性評価を実施し続けたことは、品質マネジメントでの現在の現場主義潮流から顧みると、極めて先端的であったともいえる。

一方、欧米では、「有用性」という治療現場での目的（objective）の設定が回避され、治療法の「有効性」および「安全性」に係わる項目を治療現場で計測ないし評価し、治療現場から離れた生産者および規制当局によって、集団としての有効性および安全性に基づいて統合化した概念として治療法の有用性が判断されるのが通常である。（なお、ここで「有用性」などを括弧付きにしたのは、個々の患者での評価であることを強調するためである）。わが国においても集団としての有効性および安全性には大きな関心はあるが、治療法の各患者での「有効性」および「安全性」に係わる計測・評価は「有用性」の重要な構成要素に過ぎず、集団としての治療法の有用性の判断は個々の患者の「有用性」に基づくという立場をとっている。

品質マネジメントの基本モデルである水野・赤尾（1978）の「品質機能展開（QFD: Quality Function Deployment）」に即して述べれば、「有用性」は治療法の1次要求品質、「有効性」・「安全性」などは1次要求品質を展開した2次要求品質、さらに「有効性」を諸症状の改善度に展開したものは3次要求品質、さらにその裏付けとなる諸症状の重症度や臨床検査値は4次要求品質と考えることができる（QFD的調査については、椿（1994b）を参照）。そして「品質方針」の最終達成は、あくまで1次要求品質である「有用性」に関する事実で判断されるべきものであり、現場と独立したところでの一部の専門家の総合判断が本質とはならないとの立場をとったこととなる。

臨床医による「有用性評価」の論理構造を確固なものにするため、わが国の臨床評価で用いられてきた調査票の多くは、上記の要求品質の階層性を強く意識した項目配置が工夫されてきた。欧米のマーケティング専門家が、企画・設計品質向上のために、QFD的階層を持った調査票を構築し、消費者行動分析を行うようになったのは1990年代であり、これにより共分散構造分析の適用が多くなったのである。それ以前は、外的規準を配置しない調査票から、単純な因子分析などを使って、消費者の実際の満足度（購買行動）に追随するとは限らない尺度を構成してきた。現在においても、学術論文で提唱される「評価尺度」提案では、1次要求品質を代表とする外的規準の主観性、時代・環境影響への反応性から、外的規準の計測が軽んじられる傾向がある。椿（1994b）で指摘したように、適切な主観評定を外的規準として計測することは、評価計測のモデルを診断し、改善を加速化するためにも必要なことである。

2.2.3 標準化

端的に言えば、生産者側にとっては実施した臨床試験の成績が全てである。したがって、生

産者が新薬に有利な計画で、臨床試験を実施しようとするのは自然なことである。また、生産者側の統計家は、その情報をデータ解析の観点からも琢磨して、生産者危険を最低限に抑えた解析を行うのが使命である。

一方、使用者の代弁者にとっての臨床試験情報は、当該の臨床試験のみならず当該領域の過去に行われた臨床試験情報なども参考に評価される。同じ領域において臨床試験ごとに、その情報の集計スタイル、基本的な統計解析の方法、さらには臨床試験の計画や主要な評価項目が毎回変わるようでは、ある臨床試験の情報に対する的確な判断を行うことが困難になる。したがって、使用者の立場からは、臨床試験の計画・実施・結果報告に対するある程度の「標準化」が必要であった。こうした使用者の立場からの標準化が、コントローラー委員会の先導の下で1970年代に進展した。

この標準化活動の力点は、一貫性ととともに、臨床試験の骨格の単純さにあった。すなわち、基本的にこの標準化に沿った臨床試験は小規模の生産者、あるいは場合によっては治療法の有効性を解明したいと考えた臨床家によっても実施可能なものであり、臨床試験への参入障壁を作らないという性格を持っていた。当時の二重盲検比較試験においての主要な標準化方策は、次の通りであった。

- ア. 有用性、有効性、安全性に関して、主治医の主観的総合評価を常時実施する（単純で普遍的な評価の実施）。
- イ. 原則として、割付けられた全症例を対象とした解析結果を先ず示す。
- ウ. 統計解析は、前提条件の緩い単純なノンパラメトリック法を先ず行う。
- エ. 比較を行うための推論では、すべて単純な2標本問題に分解する。

この標準化は群間比較試験およびクロスオーバー試験のそれぞれで実施され、その解説および統計解析の手引きが臨床評価システム解説第1部、第2部としてまとめられた（コントローラー委員会（1975, 1976））。この種の標準化を基に、わが国の臨床試験は1980年代後半まで、約20年間運営されてきた。また、形式的側面においては、現在に至るまでも多くの臨床試験がその影響下にある。

3. 臨床評価の形式化・形骸化の進行

2章で示した「患者の立場」からの臨床試験の推進などの臨床評価の原則は、皆保険制度の下で治療を求めて受診した患者の同意を得て臨床試験を実施せざるをえないわが国の医療環境に適合するものであり、公正性が確保されているという安心感もあって、わが国の臨床医の支持を得て、普及した。しかし、その一方で、これらの臨床評価形式の一部を借用した形骸化した臨床試験も多数実施されるようになった。

1979年の薬事法の改正によって、新薬の臨床試験が「治験」として法的な性格を持つようになった。そして、治験の依頼基準（治験の文書による依頼、被験者の同意の取得、治験中の事故等に対する賠償対策など）や規制当局への治験計画の届出制度などが定められ、行政による事務的形式の整備が伸展した。また、製薬企業によって、薬価制度の下で少ない開発費で安易に利潤が得られる類似新薬（ゾロ新）の開発が横行し、臨床的意義の乏しい治験が多数実施されるようになった。こうした製薬企業主導の治験では、単に規制当局が要求する事務形式を満たせば品質が保証されたことになるため、公平性を保証する第三者監査の役割を適正に果たさない「名義貸しのコントローラー」を作り出した。そして、偏りを度外視した、あるいは偏りを活用したデータ解析が行われることもあった。こうした状況の中では、医学研究の中に占める臨床試験の位置づけは、当然、相変わらず低いままであった。この種の臨床試験は研究費稼

ぎの片手間仕事と認識され、経験の乏しい臨床医による実施も一般化していった。

臨床試験の形式化の流れは、臨床試験の質の改善を促進する動機付けを失わせ、初期になされた創意工夫も次第に衰弱していった。こうした中で起きるべくして起きたのが、1982年に発覚した臨床試験資料の捏造事件、副作用情報隠し事件である。改竄・捏造が日常化していることを示す「メーキング」という業界用語があることが明らかになり、ドラッグ・スキャンダルとして大きな社会問題となった。そして、1960年代後半の「信用できるか否か」という議論が再び持ち上がった。

これに対して、臨床試験の適正化のために、厚生省は1983年専門家会議を設置し、「医薬品の臨床試験の実施に関する基準（案）」（通称、GCP: Good Clinical Practice）を公表し、1989年に正式に通知した。また、規制当局自身によるデータのサンプリング照合や製薬企業・医療機関の査察なども順次実施されるようになり、規制が一層強化された。また、1992年には「新医薬品の臨床評価に関する一般指針」、「臨床試験の統計解析に関するガイドライン」（以下、統計解析ガイドライン）が発効した。さらに、薬効群ごとの臨床評価方法に関するガイドラインも順次整備されて、臨床評価の形式的な標準化はさらに伸展した。

しかし、一層の規制強化や事務形式の整備にも拘わらず、ソリブジン禍、薬害エイズといった大きなドラッグ・スキャンダルが発生し、臨床試験資料の改竄事件も相変わらず跡を絶たないのである。

4. 臨床試験の統計解析に関するガイドライン

さて、今までに記した臨床評価の全般的な流れ、臨床試験の質の見掛けと実質との乖離を背景としながらも、一方には、当初の臨床試験の原則を踏まえた統計的側面の検討の歴史があった。本章では統計解析ガイドライン作成までの経緯について述べる。

4.1 統計解析ガイドライン作成前の統計的側面の流れ

コントローラー委員会方式の臨床試験の統計解析が一般化する中で、臨床試験を実施する過程で統計解析に係わる疑問（今日的にいえば多重性や同等性など）が1970年代にも臨床家から提起されていた。しかしながら、これらの疑問に正面から総合的に取り組む統計家はほとんどいなかった。

臨床統計分野における「統計的方法」を正す運動は、国内外でほぼ1980年前後にスタートした。発端は、Tukey (1977) のサイエンス誌における臨床試験での「多重性 (multiplicity)」問題の提起であり、それに応じて New England Journal of Medicine が投稿論文の統計学的側面に対する生物統計家によるレビューを開始した。海外に投稿した医学論文に対して、統計解析の問題指摘が増加し、臨床試験での統計解析への関心が高まった。こうした中、わが国でも長谷川 他 (1981) は、「医学のあゆみ」誌の「くすり」欄に掲載された臨床試験の論文の統計解析に対する批判的検討を米虫ら統計家との協力下に行い、その結果を報告した。長谷川 他の批判はわが国の臨床試験データ解析の枠組み自体に係わる本質的、具体的批判であり、臨床家も深刻に検討を開始した（清水 (1984)）。長谷川 他の批判は、その後国際的にも検討されたことの大半を含むものであり、先駆的な業績といえる。当時、これらの批判の中で特に重要と考えられたのは、統計的推論の多重性を無視しているという指摘であった。これは有効性、有用性の推論に関していえば、第一種の過誤の増大を意味しており、有効性、有用性の乏しい薬を許容する危険、すなわち「使用者の危険」（消費者危険）を増大させることに繋がっていた。特に、データ固定後に行う数多くの探索的解析、すなわち後知恵解析を生産者側が有利なように活用すれば、消費者危険を一定確率以下に抑えることが困難なことは明らかであった。

当時、厚生科学研究「医薬品の臨床試験における統計的解析等に関する研究」の主任研究者であった佐藤は、臨床家が長年にわたって疑問としてきた臨床試験での多重性などがわが国の臨床試験でどの程度の問題を起しているのか、また、どのようにそれらを克服すべきかについて組織的・実証的な検討を企画した。この検討は、1983年11月の佐藤、広津、椿、藤田による座談会が契機となり開始された（座談会の議事録は佐藤 他（1990）に公表）。臨床試験における統計的解析は既存の統計手法の単なる適用では不十分であり、複雑な臨床判断に対応する仮説構造に適合した統計的方法を検討する必要性が認識された。そして、統計的な検討が進められ、藤田 他（1984）、藤田・椿（1985）、藤田 他（1986）、広津（1986 a）、椿・藤田（1987）の一連の論文となった。また、1984年に東京で行われた第12回IBC(International Biometrics Conference)を機会にコントローラー委員会が主催した衛星シンポジウム“Recent Topics on Clinical Trials”でも、Armitage（1985）、Pocock（1985）らによって多重性問題の総括報告がなされ、臨床試験の統計解析システムの見直し作業が加速した。

臨床試験の多重性問題対処の中で、基本的にプロトコルにおいて事前に明示された「検証的データ解析」と事後的な「探索的データ解析」を明確に切り分けることの重要性が認識された。特に、これは単に統計解析の問題ではないことから、臨床家向けに「検証的臨床試験」と「探索的臨床試験」という臨床試験の性格付けを明示することとなった（藤田 他（1986））。なお、これについても、既に米虫らの指摘が行われていたことは指摘しておきたい。

さて、これらの動きとは独立にわが国の臨床評価の統計側面に重要な影響を与えたのは、1984年に発効した「抗不整脈剤の臨床評価方法に関するガイドライン」において、佐久間らが提示した方法である（これについては、例えば日本公定書協会（1992）を参照されたい）。すなわち、「治験薬は標準薬と同等またはそれ以上」であるという仮説を検証するための統計的方法と例数設計についての提案である。わが国の臨床評価では、統計的検定で標準薬と有意差の無かった新薬は標準薬と同等と解釈されることもあった。そして、安全性などその他の項目で明確なメリットが示されれば、新薬は社会的に有用なものと解釈されていた。

有意差が検出されなくても認可に繋がるのでは、多重性問題をいくら厳密に検討しても消費者危険の保護にならない。試験情報を管理せずノイズの大きなものにすれば有意差が出難くなり、生産者に有利になることを意味しており、臨床試験の内部品質保証のモチベーションさえも阻害するという観点からも許容できないものであった。勿論、これは統計的検定の論理の典型的誤用である。

こうした誤用に歯止めをかける試みとして、既に柴田・開原（1981）の推測確率に基づく方法の提唱があった。さらに、同等性仮説の検定の必要性が、その後、広津（1986 b）、椿・藤田（1987）でも論じられた。新薬が標準薬に対して Δ 以上有効率において劣っていることを帰無仮説として、これを一定の有意水準で棄却することで「臨床的に同等またはそれ以上」ということを検証する方法が提唱された。これは、現在ではNegative Trialに対する推論という考え方で呼ばれて（Pocock（1983））、1997年のICH統計ガイドライン案では「非劣性試験」の中で取り上げられている。

多重性および同等性の検討を踏まえ、わが国の臨床試験の標準的な統計解析システムに対して具体的変革を提案したのが、藤田・椿（1987、1988）である。提案した統計解析システムには上記の概念以外にも、生産者危険の不必要な増大を避けるための背景因子の不均衡の薬効評価への影響調整や背景因子と薬効の交互作用解析などの機能が組み込まれた。これらには、主として生物統計学における離散多変量解析の成果、とりわけ、統計家でなくても解析結果が直観的にも把握可能である統計的方法（Mantel-Haenszel法やBreslow-Day検定など）の重要性についての柳川（1987）の指摘が大きく影響した。

こうした統計的問題の検討が進められていた1986年11月に第1回の新医薬品統計解析指針

検討会が開催されて、統計解析ガイドラインの検討が開始された。この検討会は、臨床の論理を統計的に実現するため、臨床家と統計家の共同作業として進められた。そして、多重性および同等性についての上記の検討成果はガイドラインに組み込まれ、1990年2月に案が公表され、1992年3月に正式に通知された。

4.2 臨床統計解析ガイドラインの骨格

統計解析ガイドラインは、「同等性」と「多重性」のガイドラインなどとも呼ばれる。このガイドラインは、許認可の審査に必要な情報を系統的に列挙している米国FDA (Food and Drug Administration)のガイドラインを参考にして、作成された。しかし、有用でない新薬が市場に流出する危険、すなわち「消費者危険」を統計的方法の適正使用により定められた確率以下に抑えることを保証するという目的意識の下での方策を提示しており、独自の姿勢も見られた。その主な原則は、次のような点である。

- ア. 仮説検証主義（多重性を利用しえる探索的解析による結論付けの排除）
- イ. 同等性推論の厳密な適用
- ウ. 実践的試験の意義の強調
- エ. 不完全症例を可能な限り解析除外しないことの推奨
- オ. 有用性、有効性、安全性に関する3つの主要な解析項目の設定とそれらに対する検証的推論

これらの目的を達成するために、臨床試験の計画段階および実施・解析段階でどのようなことに留意するのが望ましいかについても列挙され、治験実施計画書の重視が特に強調された。この意味で、「統計解析に関するガイドライン」という名前を超えた臨床評価全般に係わるものであった。

さらに、臨床仮説探索のために副次的に実施される探索的解析の奨励、共変量の調整、交互作用解析の必要な状況やその方法についても触れられていた。また、臨床試験の「統計解析責任者」が公式に規定された。

今日から見ると、このガイドラインの狙いや重要な記述がガイドライン本体ではなく、「注」や「Q&A」に記述される一方、Q&Aやマニュアル類に記述するのが適当な具体的な統計的方法の詳細が「注」に含まれているなど、編集上の問題はあった。しかし、使用者側の代弁者としての統計家の要求事項は一通り組み込まれていると思われる。

同等性推論については、1994年2月の中央薬事審議会の「医薬品特別部会の申し合わせ事項」として、審議中の許認可申請の新薬を含めて適用されることになり、同等性推論の結果が否定的であった新薬については厳密に他のメリットの検証が要求されるようになった。しかしながら、本来的には、同等性が検証されれば良しとする新薬の臨床的有用性（当該の臨床試験では検証できないメリットなど）が、計画段階で明示されていなければならないのである（藤田(1993)）。

5. ICH とわが国の臨床評価の変質

GCP (1989年)、一般指針 (1992年)、統計解析ガイドライン (1992年) などにより、倫理性とともに、臨床試験の「科学性」（ここで、科学性に括弧を付けたのは、精密科学性の意味を強調するためである）が強く意識されるようにはなった。多重性の調整、患者背景因子の調整、同等性の検証といったことは「当たり前」のように実施されるようになった。しかしながら、一方で、規制や標準的形式を見掛け上は満たすものの、その臨床試験のデータの正確さ（バイ

アスと精度)が問題になる新薬の許認可申請が相変わらず見られ、わが国の臨床試験の質の一般的な保証はなお大きな問題として残されたままであった。

こうした中で、「優れた医薬品をより早く世界の患者に」を合言葉に、日・米・欧医薬品規制ハーモナイゼーション国際会議(ICH)が1991年から開始された。近年、臨床試験に係わる「有効性分野」も急展開し、わが国の臨床試験のガイドラインが全面的に見直されるなど、大きな転換点にある。

既に総括したように、わが国の臨床評価の原則が有意義なものであることは、十分に理解されたことと思われる。しかし一方、こうした原則を形骸化させた多くの臨床試験が現実的に存在している。こうした臨床試験の問題は、単にわが国の臨床評価の原則を否定し、欧米の原則を形の上で模倣しただけでは、到底、解決できるものではないのである。

ここでは統計解析ガイドラインの副作用やICH「臨床試験のための統計的原則(案)」(ICH E9, 以下, ICH 統計ガイドライン)等の問題点について、わが国の臨床評価の原則を踏まえて、いくつかの指摘を行いたい。

5.1 技術評価の必要性

ICH は、三極の医薬品使用者の代弁者としての規制当局、臨床家のみによる規制システム構築ではなく、欧米日三極の生産者側代表も原案作成に関与している。これまでのわが国のやり方と比べて、オープンに生産者側の「実施可能性(feasibility)」が考慮されている。ICH 活動では「科学的な臨床試験の実施」が目指されているが、これが生産者にとって最も経済的な戦略であることは注意する必要がある。

ICH では、「プラセボに優る」「実対照薬に優る」「用量-反応関係を示す」ことを検証する優越性試験が、「有効性の確立」での最も説得力のある臨床試験として重視されている。こうした有効性仮説検証のための「科学的」臨床試験(ここで、敢えて「科学的」を括弧書きとしたのは、これまでの本稿で用いている科学的臨床試験と区別するためである。すなわち、実験室的な制御された環境下での新薬の有効性・安全性を検証するための精密科学的試験という意味で用いている。以下でも、括弧書きの「科学的」は、括弧の無い科学的と区別して使っているので注意されたい。)では、事前にエンドポイントを絞り込み精密化する方向、および試験参加施設を少なくして1施設当たりの症例数を増加させる方向に向うことになり、実際、ICH 統計ガイドラインでも強調されている。すなわち、制御された条件下での「科学的」仮説検証を狙いとした「説明的試験(explanatory trial)」あるいは「科学的実験(scientific experiment)」の性格が強いものになっている。確かにICH 統計ガイドラインの中には、「検証的試験が実施されるときまでには、被験者を意図した使用者に十分近いものにしておくべき」「目標とした適応症の範囲内でできるだけ組み込み基準と除外基準を緩める」などとされている。しかし、ICH の方向は、基本的には「技術評価」の側面を弱める危険がある。

これまでわが国の臨床評価で重視してきた新薬を社会に導入した場合の挙動を患者の立場から評価する「有用性評価」は、不要になったのであろうか。ICH 統計ガイドラインではこれを「一般化可能性(generalizability)の検討」としているが、こうした検討は認可後の市販後調査で行うという路線も提示されている。

これまでのわが国の後期第二相までの説明的試験については、有効性仮説の検証が不十分であり、技術評価以前の問題があったことは否めない。その意味では制御された環境下での「科学的」仮説検証は必要であり、提唱されてしかるべきである。また、この強調はある意味での臨床試験の質改善の動機付けになろう。

しかしながら、「この段階までで新薬の有用性についての保証が十分であるか」が、問題なのである。極端な言い方をすれば、臨床試験の性格から見て、技術評価としての第三相段階を省

略して、実験室的な環境下での「科学的」仮説検証に重点のある後期第二相までで新薬を許可して良いかということに近い。さらには、ICHの「海外臨床データ受け入れにおける人種差要因(E5)」において、臨床試験の重複を最小限にして薬を迅速に供給するための人種的要因の影響を評価する枠組みが提供されている。主要な臨床試験の成績は海外で行ったものを使用し(わが国での臨床試験を空洞化し)、薬の作用を反映すると考えられる薬理学的エンドポイントを用いた薬力学的試験や対照群を持たない一般臨床試験と呼ばれる使用経験のみで、将来的には許可される方向性が示されている。

1971年の薬効懇答申の方針に示されたように、「科学的」仮説検証のための説明的試験の上に、技術としての治療法を評価する実践的試験がなされることによって、初めて新薬の科学的評価がなしえることを再度認識する必要がある。ICH統計ガイドラインでは実際的には2つ以上の検証的試験を要求しているが、同様の試験の繰り返しではなく、性格の異なる試験の実施も許容されるべきと考えられる。第三相試験として、「科学的」仮説検証を目的とした説明的試験が実施された後に、技術評価としての実践的試験が実施されることが強く望まれる。しかしながら、十分な技術評価を抜きにして新薬を社会に導入する「実験的試み」が不可避な国際的動向というのであれば、百歩も千歩も譲って、薬禍を小さく抑えるためには注意深い市販後監視は必須であり、問題発生に対する迅速な対処を可能にするシステム構築が前提となるべきである。しかし、これまでのわが国の市販後監視システムが不十分なことは、薬害の歴史が教えるところである。

5.2 患者の立場からの有用性評価、総合評価

わが国の臨床試験の質が悪いのは有用性評価と施設当りの症例数が少ないことに原因があり、これらを改善すべきとする議論が多い。そうであろうか。

統計解析ガイドラインで多重性調整の必要性和仮説検証主義が強調された。この副作用として、生産者の一部が新薬の有効性に関する主要な解析項目(エンドポイント)を奇妙に絞り込み始めた。ある疾病の全般改善度をターゲットにせず、薬効の同等性推論をクリアするため、例えば脳循環・代謝改善薬では精神症候のみの改善度をエンドポイントにするといったことも散見される。また、計画段階での目的絞り込みは、ガイドラインの意図とは逆に副次的項目解析の軽視という副作用を生み出した。臨床家は、新薬に対する評価を適切に行うだけの多様な情報を臨床試験から得られなくなる傾向が助長され、ややもすると臨床試験に基づく新薬の有用性判断は統計側の責任という逆転現象、思考の停滞傾向が起き始めた。

これは、臨床評価の最終目的を矮小化したものである。既に総括したように、わが国の臨床評価の原則における臨床試験の最終目的はあくまで治療法の有用性の評価である。生産者に都合な「エンドポイント」はその代用特性に過ぎないし、本来の評価目的からすれば大きなバイアスが生じている可能性もある。また、エンドポイントの改訂は、新たなエンドポイントの改善が有用性評価の改善に明確な影響を与え、従来のエンドポイントよりも有用性評価との関連性が強く、再現性にも優れることなどが検証された後になされるべきものである。また、生産者の都合で行うのではなく、臨床側の合意(学会の基準、厚生省ガイドラインなど)の上でなされるべきものである。さらに、新たなエンドポイントの採用が妥当とされても、常に当該臨床試験結果の外的妥当性を診断する手段を確保するために、調査票の構造はあまり変化させないか、少なくとも総合評価を含む、エンドポイントより低次の要求品質は、すべて個々の患者に対して評価されるような配慮がなされる必要がある。

一方、臨床評価の原則において推奨された有用性評価が有意義であり続けるためには、いくつかの条件が必要であると考えられる。例えば、当該臨床領域での豊富な臨床経験に基づく総合評価であること、積極的な参加による精度の高い評価であること、真摯な批判的吟味の継続

(妥当性や信頼性の検討, 特に, 真の治療目標の予測性に対する客観的測定などとの比較検討)などを挙げることができよう。しかしながら, 上述のように現実の一般的な臨床試験についてのわが国の歴史はこれらを満たすものではなく, 総合評価の正確さについての疑念を招いたことも否めない。さらには, 観察項目などに基づく機械的なマトリックスでの当てはめを総合評価と呼ぶなど, 専門家の経験に裏打ちされた総合化でない合成変数まで含まれるかに誤解され, 用語の混乱も著しい。

こうした現状を勘案して一步譲るならば, ICH による「科学的」臨床試験の中で検証すべき有効性仮説を具現化したプライマリーエンドポイントとして重視される方向にある客観的測定に対して, 総合評価は患者の立場からその妥当性を確認・補正する役割を少なくとも担わなければならないと考える。客観的測定の多くは代替変数(サロゲートエンドポイント)であり, 欠測などが殆どの試験で不可避である。また, 予め全てを予見して試験計画を立案することは不可能である。そのため, 1次情報を知る臨床医の臨床判断(本来の総合評価)によって, 客観的測定などのプライマリーエンドポイントの意義を確認・補正すべきなのである。また, 上記の有用性評価が有意義であり続けるための努力を精力的に推進して, 客観的測定値と総合評価とを比較する実証的検討を行い, 臨床的意義の良否を具体的に検討することが重要である。なお, ICH 統計ガイドラインにおいて総合評価変数を主要ないし副次的変数として使用する場合に治験実施計画書に記述すべきとされている事柄のいくつかには, 総合評価の本来の意義を阻害するものがある。

5.3 多施設臨床試験

ICH では, 可能な限りレベルの高い医師と医療施設において正確かつ精密な評価を行うのが良いとして, また薬・施設の交互作用の検討によってデータの信頼性を主に点検するために, 「科学的」少施設試験が推奨されている。そして, こうした「科学的」少施設試験の2回以上の繰り返しを要求しており, 「一般化可能性 (generalizability) 検討」のための折衷的方式が提示されている。こうした方式は, 新薬の有効性の「科学的」な仮説検証には好ましいものであるが, 新薬を用いる治療法の技術評価としては不十分なものである。

性質の異なる2つ以上の臨床試験を要求してはいるが, 個々の試験を Fisher 流の科学的実験計画の3原則(繰り返し, 無作為化, 局所管理)の理念のみを拠り所に行う限り, 新薬の有用性の下流再現性を議論するためには効率的でない。なお一般に, 製品機能の下流再現性とは, 実験段階では通常取り上げていない多様な使用条件に対しても, 製品が機能上のトラブルを起こさないことを意味する。

一步譲って, Fisher 流の古典的実験計画の立場をとっても, 結論の普遍性確保のためには, ブロック因子としての施設を積極的に多数導入し, 実験の中に占める完全無作為化試験のウェイトを減らすことが必要だというのが, 局所管理の原則の実際的狙いである。例えば, 奥野・芳賀(1969)は, ストープの芯2タイプの性能比較試験において芯の数が同じ場合に, 芯を取り付けるストープの数を1台にするのが良いか, できるだけ多くするのが良いかについての議論を詳細に行い, 結論の普遍性の立場, そして実際家の立場からは, ストープ数(施設)を増やすのが優れた方法であると結論付けている。

また, 筆者らの主張のように第三相試験における臨床評価を「科学的」実験ではなく, 市場薬効に関する標本調査と見なしても, 複数施設での情報収集は施設内の患者集団というクラスター抽出(通常は有意抽出)を行ったことに過ぎず, クラスター間分散がクラスター内分散に比して十分小さくならない限り, 同一標本サイズにおいてはクラスター数を多くとるほど調査・計測精度の観点からは望ましい。このことは, 調査設計の基本認識である(例えば, Cochran(1961))。

さらに進んで、技術評価のための試験は、製品機能の下流再現性確保という田口の哲学に基づく「品質工学的実験」、すなわち現場を模擬するために十分な「ノイズ因子 (noise factor: 田口の用語では誤差因子)」を意図的に配置する実験計画に基づく方が、より効率的である。欧米の機械・電気産業でも、新技術・新製品の開発評価においては、タグチメソッドを用いることが、Kacker (1985) 以来、日常化している (Ryan ed. (1988))。

わが国の臨床試験は、意図的にノイズ因子を配置するというよりは、欧米から見れば膨大な施設数で試験を行い、医療現場そのもので試験をするかのような単純な構造にはなっている。しかし、現場での治療法の有用性を計測する観点からは、「精密科学的」少施設試験よりブロックを多く導入したこの方式の方が決定的に優れている。何故ならば、「研究者のための統計的方法」では、単に施設を母数的ブロック効果として表現して薬効検定を行っており、これは「併行精度 (repeatability variance)」を基に薬効評価をしていることになる。これに対して、「技術者のための統計的方法」では、多施設臨床試験という治療現場に存在するノイズをむしろノイズ因子 (標示因子) として捉え、これらノイズ因子と薬効 (制御因子) との交互作用を基準に薬効検定を行う (古典的には、Rao (1973) に記述されているように変量的ブロック因子と制御因子の交互作用を分母に制御因子の検定を行う)、あるいは、実際にはこれらの交互作用と誤差を併合した分散で薬効検定を行っており、「再現精度 (reproducibility variance)」を基に検討していることになる。併行精度に基づく「研究者のための統計的方法」よりは再現精度に基づく「技術者のための統計的方法」の方が、直接的に市場での薬効の有用性を評価したことに繋がる (この種の議論の哲学的詳細は竹内(1986)を、統計技術の詳細は、宮川 (1988)、ISO 5725-1 (1994) を参照されたい)。

無論、椿 (1993, 1994 a) で既に指摘したように、ノイズ因子の全てが臨床試験に必要な訳ではない。ノイズには、例えば評価の信頼性の不足などのような適切な標準化により低減すべきものも多い。しかし、少なくとも新薬の使用予想期間には治療の現場に存在するであろうノイズを無視したり、統計解析的に除去するといった哲学は、使用者 (患者) 側の立場からは取り得ないものである。

実際、製品の品質、安全性、環境影響などについて、生産者が全面的に責任を負っている多くの産業界においては、新技術開発に際して製品の本来の機能や安全性の市場再現性を得ようと、様々な市場に存在する諸環境条件を模擬した「品質工学」の実験を行い、成果を上げている。この種の活動の動機付けとなっているのが、やはり市場原理である。使用者の満足を得られないような製品は市場から速やかに淘汰されるという現実である。医薬品業界においても、当面の認可よりも、市場で勝ち残ることの困難さを意識付ける市場メカニズムが構築できれば、自ずと生産者側も「下流再現性」の優れた新薬を投入するためのプロセスを再構築することができよう。

いずれにせよ従来、わが国の第三相での多施設三重盲検臨床試験を「市販前の営業的行為であった」「わが国独自の質の悪い試験」と生産者側が意図的に自嘲し、ISO 5725-1 (1994) 6.3 節 selection of laboratories for the accuracy experiment に見られる「併行分散」, 「施設間再現分散」の推定誤差などに関する統計数理的検討もなく、Ojima et al. (1988) のような実証的検討もせずに、単純に 1 施設 20 例以上の試験を推進することは是認できない。精度管理分野における多施設共同試験、とりわけ ISO 5725 の原案作成を巡り故石川 馨らが、国際商取引のために試験値の許容差設定を行う多施設共同精度実験を計画する原則 (ISO 5725-1) を、20 年かけて米国を除く世界のコンセンサスにした交渉例に倣って、わが国の多施設臨床試験の意義を主張することが必要であろう。

5.4 同等性推論

統計解析ガイドラインの副作用の一つとして、エンドポイントの奇妙な絞り込みについては既に述べた。もう一つの副作用は、有効性の乏しい新薬を症例数を増加させることで、同等性の検証に繋げようという統計的試みの増加である。この点に関して、椿・藤田(1987)では、現行の「10%以上有効率の低い新薬」という帰無仮説の設定の妥当性が標本サイズに依存することを指摘し、最低標本サイズを標準化すれば後は単純な点推定値の大小に基づく決定関数を使うのが良いと提案した。

尤も、統計的推論の技法に頼るよりは、むしろ市場原理でこの種の薬が淘汰されるか、経済原理でこの種の薬の開発意欲を低下させることの方が健全であろう。中央社会保険医療協議会の1995年11月の建議において類似新薬の薬価を抑制する方向が明確にされ、また最近のマスコミの薬価批判の中心にも取り上げられており、経済的インセンティブも世論も擬似新薬の開発意欲を殺ぐ方向に動いている。

ICH 統計ガイドラインでは優越性試験を重視しており、同等性推論にあたる非劣性試験ではプラセボを組み込んだ実薬対照試験の必要性が強調されている。欧米の圧力のためか、わが国の生産者側も擬似新薬開発抑制の方向を決断した(?)とも見られる。

臨床的有用性の改善を見込めない擬似新薬開発の横行は、患者の得難い協力に十分に報いているとは言い難く、その抑制が達成されることを見守る必要がある。また、統計家としては、臨床試験の質を度外視したクロスオーバー法を用いた同等性推論といった奇妙な議論に対して、健全な姿勢を保つ必要がある。

5.5 品質保証

皮肉な言い方をすれば、ICH GCPに基づくわが国の「医薬品の臨床試験の実施に関する省令(新GCP)」(1997年3月)が臨床試験の計画、実施、品質保証システムの構築、外部品質保証に至るまでの全責任を当事者の一方の「生産者のみ」に負わせたことは、興味深い実験といえる。これに関係して、わが国でも臨床試験報告の論文公表義務の廃止、第三者コントローラーからメーカーへのアドバイザーへの役割の転換などが伸展している。原理的に欧米の水準を上回る日本の品質保証システムが、欧米流のシステムとの実証的比較に基づく議論なしに解体される可能性がある。

新GCPに規定されたモニタリング、監査といった生産者(製薬企業側)が行う品質保証と規制当局による査察によって、わが国で臨床評価の公平性は実際に保証されるのであろうか。確かに多くの場合、これらのシステムは臨床試験の質の向上に役立つであろう。しかし、1982年のデータ捏造事件で「メーカー」という業界用語が端的に示しているように、「学問の自由、企業活動の自由の中における主体的な自己規制を国民は信用していない。自由を自ら悪用して自由主義の息の根を止める一部の人間がいる」という佐藤(1982)の警鐘は現状でも生きている。臨床試験に誠実に関わった多くの経験のある者なら、規制を潜り抜け「利益を目指す企業の本務」を遂行しようとする一部が必ずいることを経験しているであろう。職能に伴う倫理綱領(professional code of ethics)に基づく社会システムを、それが未熟な社会に単に移植しただけでは、公正な品質保証の担保は書類上のことだけである。患者、国民に対して信頼感を与える外部品質保証としては、外部審査機関による「第三者監査」、さらには「第三者適合性評価」が良いことは明らかである。わが国の機構は、こうした役割を果たしえるものであろうか。いずれにしても、わが国の「適正なコントローラーの役割」を正當に評価し、「第三者制度」のより充実した確立を目指すことがわが国の臨床評価の歴史から当然必要なのである。

6. おわりに

医薬品開発側に立つ統計家も今般の国際化により、臨床研究の合理性に対する相当の責任を課されることとなった。

一方で、直接の利害関係者ではない統計家は、彼らの行動の監視者として、臨床統計分野に本質的・実質的な貢献を行うことが社会人として期待されていると考える。

既に臨床研究者である山内 (1997) は、「今後新しい制度下で臨床試験が実施されることになるが、従来のわが国の臨床試験が有していた優れた点が失われ、米国流の方法論の問題点だけが表出するような事態に陥ることは避けなければならない。今後、そのための運用面での方策が検討される必要がある。したがって、新しい制度に移行しても、わが国が従来採用してきた方法論と ICH に基づく方法論の間で、実証的に比較検討を進めることが極めて重要であると思われる」という指摘を行い、臨床統計分野での実証的検討が全く不十分であることを訴えている。

情報マネジメント専門家集団としての統計家への社会からの期待と責任は、好むと好まざるにかかわらず非常に大きくなっていることを自覚しなくてはならない。

謝 辞

本論文の執筆を勧めて下さった統計数理研究所佐藤俊哉先生、及び草稿に対して、様々な指摘を下された査読者及び光石忠敬先生(弁護士、臨床評価編集委員)、執筆時に諸資料を整備して頂いたコントローラー委員会事務局の皆様に、深く感謝の意を表します。

参 考 文 献

- Armitage, P. (1985). Two areas of controversy in the design and analysis of clinical trials, 臨床評価, 13, 別冊III, 29-40.
- Cochran, W. G. (1961). *Sampling Techniques*, 3rd ed., Wiley, New York.
- コントローラー委員会 (1975). 臨床評価システム解説第1部, 臨床評価, 5, 99-115.
- コントローラー委員会 (1976). 臨床評価システム解説第2部, 臨床評価, 6, 549-583.
- コントローラー委員会 (1992). 創立20周年にあたって, 臨床評価, 20, 別冊VI, 1-14.
- 藤田利治 (1993). 第3相臨床試験における統計学的諸問題 (その1), 臨床薬理, 24, 315-319.
- 藤田利治, 椿 広計 (1985). 臨床試験の結果の検定方法による相違, 臨床評価, 13, 601-611.
- 藤田利治, 椿 広計 (1987). 最近の臨床統計における問題点, 新しい薬効評価の試み, 臨床評価, 15, 別冊IV, 39-51.
- 藤田利治, 椿 広計 (1988). 薬効評価解析システムの試作, 臨床評価, 16, 3-23.
- 藤田利治, 椿 広計, 佐藤倚男, 栗原雅直, 藤本 聡 (1984). 多数の反応項目およびサブグループに対する統計的検定の繰り返し適用に関する問題点, 臨床評価, 12, 827-835.
- 藤田利治, 椿 広計, 佐藤倚男 (1986). 臨床試験における多重性——多群比較を中心として——, 臨床評価, 14, 477-486.
- 月間薬事編集局 (1970). 今日の薬学, 医薬研連シンポジウム 臨床評価の問題点, 月間薬事, 12, 409-419.
- 長谷川宏明, 大谷靖世, 栗谷典量, 米虫和子, 米虫節夫 (1981). 「医学のあゆみ」誌“くすり欄”の臨床比較試験結果の要点一覧表(1)-(8), 医薬ジャーナル, 17, 439-450, 613-624, 819-836, 963-972, 1145-1159, 1309-1315, 1499-1503, 1651-1655.
- 広津千尋 (1986a). 臨床試験における統計的諸問題(1), 同源性検定を中心として, 臨床評価, 14, 467-475.
- 広津千尋 (1986b). 臨床試験における統計的諸問題(2), 統計的検定の多重性について, 臨床評価, 14, 707-722.
- 飯塚悦功 (1996). ISO 9000 とは何だろうか, 品質月間テキスト 267, 品質月間委員会.
- ISO 5725-1 (1994). *Accuracy (Trueness and Precision) of Measurement Methods and Results—Part 1:*

- General Principles and Definitions*, International Organization for Standardization, Geneva.
- Kacker, R. N. (1985). Off-line quality control, parameter design, and the Taguchi method, *Journal of Quality Technology*, **17**(4), 176-209 (with discussions).
- Liepnins, G. E. and Uppuluri, V. R. R. eds. (1990). *Data Quality Control (Theory and Practice)*, Marcel Dekker, New York.
- 宮川雅巳 (1988). 実験計画法をめぐる諸問題——田口メソッドの意義と問題点, 品質, **18**(3), 12-19.
- 水野滋, 赤尾洋二 編著 (1978). 『品質機能展開——全社の品質管理へのアプローチ』, 日科技連出版, 東京.
- 内藤周幸 (1983). 臨床試験における問題点, 臨床評価, **11**, 253-263.
- 日本規格協会 (1997). 『適合性評価, 品質システム, 環境マネジメントシステム』, JIS ハンドブック 64, 日本規格協会, 東京.
- 日本公定書協会編 (1992). 『臨床評価ガイドライン集』, 薬事日報社, 東京 (注: 臨床試験の統計解析に関するガイドライン Q&A を収録している).
- Ojima, Y., Tsubaki, H. and Fujita, T. (1988). Inter-clinical consistency of drug evaluations in multi-center trials, Biometry, *Clinical Trials and Related Topics* (ed. T. Okuno), 47-57, Excerpta Medica, Amsterdam.
- 奥野忠一, 芳賀敏郎 (1969). 『実験計画法』, 培風館, 東京.
- Pocock, S. J. (1983). *Clinical Trials, A Practical Approach*, Wiley, New York (コントローラー委員会監訳 (1989). 『クリニカルトリアル』, 篠原出版, 東京).
- Pocock, S. J. (1985). Current issues in the design and interpretation of clinical trials, 臨床評価, **13**, 別冊III, 41-58.
- Rao, C. R. (1973). *Linear Statistical Inference and its Applications*, 2nd ed., Wiley, New York (奥野忠一他訳 (1977). 『統計的推測とその応用』, 東京図書, 東京).
- Ryan, N. E. ed. (1988). *Taguchi Methods and QFD: Hows and Whys for Management, A Special Collection of Papers on Today's Quality Issues and New Quality Technologies*, American Supplier Institute Press, Dearborn.
- 佐藤倚男 (1972). 発刊の辞, 臨床評価, **1**, 1-2.
- 佐藤倚男 (1974). 「医薬品の臨床試験評価に関する体制の確立について」の日本学術会議より内閣総理大臣への勧告について, 臨床評価, **2**, 125-128.
- 佐藤倚男 (1982). 臨床評価のあり方について, 医薬品研究, **13**, 997-1005.
- 佐藤倚男 (1992). コントローラー委員会創立の理念と20年の活動, 臨床評価, **20**, 別冊VI, 15-26 (with discussions).
- 佐藤倚男, 広津千尋, 椿 広計 (1990). 臨床評価における多重性の問題, 臨床評価, **18**, 3-18.
- 柴田義貞, 開原成允 (1981). 試験薬が標準薬と同等またはそれ以上の有効率をもつことの判定法, 臨床薬理, **12**, 421-426.
- 清水直容 (1984). 臨床比較試験成績の読み方, 内科懇話会において, 日本医事新報, **3132**, 8-19.
- 竹内 啓 (1986). 問題提起: 田口の実験計画法, その論理と哲学, 品質, **15**(2), 54-62.
- 椿 広計 (1993). 臨床試験における統計の問題点——同等性推論と多施設試験を中心として——, 臨床薬理, **24**, 321-325.
- 椿 広計 (1994a). 臨床統計解析に求められる課題, 『SBA と薬効評価』(日本公定書協会 編), 薬事日報社, 東京.
- 椿 広計 (1994b). アンケート調査の設計, 標準化と品質管理, **47**(9), 95-99.
- 椿 広計, 藤田利治 (1987). わが国の臨床試験における統計的検定の問題点——同等性仮説の検定と多群比較について, 応用統計学, **16**, 55-68.
- Tukey, J. W. (1977). Some thoughts on clinical trials, especially problems of multiplicity, *Science*, **198**, 679-684.
- 薬効問題懇談会 (1972). 薬効問題懇談会答申, 臨床評価, **1**, 3-26.
- 山本皓一 (1991). 臨床試験におけるくすりの有用性の概念とその評価, 臨床評価, **19**, 205-212.
- 山内慶太 (1997). わが国の臨床試験の方法論の再検討——特に多施設試験と有用性の意義について——, 『向精神薬の開発をめぐる』(上島国利 編), 精神医学レビュー25, 92-102, ライフ・サイエンス, 東京.
- 柳川 堯 (1987). 『離散多変量データの解析』, 共立出版, 東京.

Historical Remarks on the Patient Oriented Clinical Evaluation in Japan and Its Statistical Aspects

Hiroe Tsubaki

(Graduate School of Systems Management, Tsukuba University)

Toshiharu Fujita

(Department of Epidemiology, National Institute of Public Health)

Yorio Satoh

(Former professor, University of Tokyo)

In order to get qualified information for a new drug application from the view point of patients, Japanese clinicians had sincerely discussed requirements for clinical evaluation in 1970's and three fundamental principles of clinical evaluation had been established consequently, those are the least information bias principle, the patient oriented evaluation principle and the standardized trial and evaluation principle. Following these historical remarks, the authors point out that the Japanese original methodology has been affected and skewed by formalism in Japanese activities and by globalization through the discussion in the International Conference on Harmonization whose view point is concentrated on the scientific accuracy for verifying a hypothesis on drug efficacy.

吉村論文および椿・藤田・佐藤論文に関する討論

コメント

椿・藤田・佐藤論文への論評

東京理科大学* 吉 村 功

(受付 1998 年 2 月 26 日)

1. はじめに

私は、この論文にたいする私以外の論評者の氏名を知っている。したがってこれらの方々が指摘しそうなことは、省くか簡単に触れるだけにする。また私の論文に対する論評への回答で述べることになりそうなことにも触れない。貴重な紙面であることを考えて、儀礼と敬称を省く。臨床評価は医薬品だけでなく治療法でも行われるが、ここでは「医薬品」と一括する。この議論をしているときの私は、なるべく良い医薬品のみを早く世に出したいと希望している第三者である。患者になる可能性はあるが現在は患者でない。もちろん生産者、医者、薬剤師でもない。

2. 文体について

著者らの物の考え方になじみのある私でも、本当に言いたいことが何であるか分かりにくいところが多い。歴史上の事実について、鵜呑みにできないところも多い。たとえば以下のことが感じられる。

第1は類比(analogy)に頼りすぎていることである。特に、品質管理の分野での概念をそのまま持ち込み、それに依拠しているところが多い。これは二つの面でよくない。一つは、品質管理の専門家さえ解釈の違いが生じ得るものを、既に確立していることとして門外漢に受け入れさせることである。他の一つは、医薬品にはそれなりの特殊性があるにもかかわらず、その特殊性を配慮していないことである。

第2は事実についての主体がしばしばあいまいなことである。たとえば、「委員会発足の背景には…強い意志があった。」 (§ 2.1 末尾) とあるが、委員会発足のときに全委員でそういう申し合わせをしたのだろうか? 「わが国の臨床評価の最も基本的な原則は…であった。」 (§ 2.2.2 最初) とあるが、この原則はどこで確立しており、誰にとって異論のないものだったのだろうか? 本論文でいうところの「臨床評価」は、「コントローラー委員会が目標としてきた臨床評価のやり方」という意味だろうか? 「以下のような実施原則が臨床側の主要な要求になっていた。」 (§ 2.2.2 第4段落) とあるが、臨床側という主体があるのだろうか? 実際には、臨床評価に関係する人が非常に多くいて、それぞれ異なる立場や主張を持っていたと私は考えるのだが。

第3は自己賛美的な態度が目立つことである。たとえば「…といった批判が交わされていた(佐藤(1982))」 (§ 2.1 第2段落) とか、「こうした第三者監視・介入…わが国独自のシステムであった。」 (§ 2.2.1 第2段落) とか、「臨床医による「有用性評価」…工夫されてきた。」 (§ 2.2.2

* 工学部経営工学科: 〒162-8601 東京都新宿区神楽坂1-3.

最後からの第2段落)などである。これらはすべて著者自身の解説文や主張文を根拠にしている。このような議論の仕方が全体に満ちあふれている。このように過去の自分らの主張論文を主たる根拠にするのは、自己の信念の表明としては自然であるが、歴史の記述、評価としては望ましいことでない。

第4は個別の医薬品や疾患をほとんど取りあげていないことである。議論はすべて一般的、観念的、抽象的な評価に終始している。§5.2など、どういう事実を指摘しているのか、さっぱりわからない。こういう議論は、新たな息吹を呼ばない争い、昔の学生運動用語でいうところの「空中戦」をもたらすだけである。

この他に、主題にとって不要と感じられる部分が多いのも主張の明確さを弱めている。

3. 医薬品の特殊性

医薬品は使用する人間が限定されている。原則として、教育を受けて医師や薬剤師の資格を得た人のみが使用できる。その使用が不適だったとき、その責任を医薬品に向けるか、これらの資格を持つ人に向けるかは、事例ごとにていねいに検討しなければならない。本論文にはそのようなきめ細かさが見られない。

医薬品の良さには「患者自身による評価」 (§2.2.2 第3段落) に依拠できないところがある。科学的臨床試験が必要なのはそのためである。患者の意向を尊重するというのは、たてまえとしてはうるわしい。しかしそのうるわしさに酔うことは危険である。また、「臨床医には日常診療においてこのような評価を行い」とあるが、著者らも多施設試験に関連して自ら認めているように、臨床医の質はそれほど理想的なものでない。医薬品の危ない使い方をする臨床医と患者が存在するという現実に基づいた議論をすべきであろう。たとえば稀少医薬品の中には、患者から早期承認の嘆願書が出ているものが少なくない。しかしそれをそのまま受け入れるのは、決して患者の立場に立つことではない。安全性の吟味のために、ときに結論を遅らせることも必要である。

上の二つの特徴が原因で、医薬品では市場原理が他の製品より働きにくい。今の健康保険制度の下での薬価の決め方はそれに輪をかけているようである。それに対する吟味を抜きにして、市場原理を持ち出すのは議論が粗すぎる。

良い薬を早く世に出すことと、悪い薬を世に出さないこととは、矛盾をはらんだ押し合い関係にある。一方を過度に重視することは他方を過度に無視することになる。後者だけを強調することが患者にとって本当に良いというわけではない。両方の均衡点をどこかに設定しておいて、その中で臨床試験とその結果の評価の質をあげることが、われわれの工夫すべきことである。

4. ICH 統計原則について

本論文で、「こうした「科学的」少数施設試験の2回以上の繰り返しを要求しており…」 (§5.3 第1段落) としているのは誤解である。そのような記述はない。

§5.2の議論では、何を攻撃したかったのか、何を錦の御旗として守りたかったのか、読みとれない。攻撃する点については欠点のみを取り上げ、守りたいところについては主観評価の問題点を軽く見るなど、冷静さを欠いている。

有害作用を承知の上で有益部分を利用するのは薬物の宿命である。両側面を、有用かという問いかけで総合評価するのもおかしくない。それを否定する者は無いであろう。しかしだからといって、「有用性」という主観的1次元変数を主要変数として薬剤の検証評価をして良い、というのは信頼区間(仮説検定)方式を用いる下では無茶である。たとえば500人の臨床試験で

2～3人に重大な副作用が生じたといった現象は、検証に用いる変数において非常に劣る例が少数あったというだけのことではない。別の次元で検討すべきことである。

5. おわりに

本論文の3人の著者は良い医薬品を開発することに大きく寄与してきたし、今もそのために真摯な努力をしている方々である。ここでは論文の問題点のみを取り上げたが、それをもって、著者らを誤解することのないように望みたい。

吉村論文へのコメント

筑波大学大学院* 椿 広 計
国立公衆衛生院** 藤 田 利 治

(受付 1998年2月16日)

吉村論文ではICH統計ガイドライン推進の立場から、臨床試験の計画におけるいくつかの論点が提出されている。我々の報告と比べると、吉村論文で主として強調された臨床試験の性格が制御された限定的な医療環境下での「(精密)科学的」仮説検証を狙いとした「説明的試験(explanatory trial)」ないし「科学的実験(scientific experiment)」であることがよく理解できると思われる。実際、我が国の臨床試験の質に対する批判は、説明的試験による薬剤の有効性の「(精密)科学的」検証が不十分であったことに多くが起因している。

これに対して、我々の主張のひとつは、説明的試験による「科学的」有効性検証後の「実践的試験(pragmatic trial)」あるいは「技術評価(technology assessment)」の必要性の強調にあった。すなわち、単に実験室的な制御された条件下で有効性の検証や安全性の保証を得るだけでは新薬承認には不十分であり、これまで臨床家の立場からは当然のことと考えられてきたように、新しい技術、すなわち新薬を社会に導入した場合の挙動を予め評価すべきだということの再確認を要求したのである。この中には、プラセボや標準薬と比べた平均的な有効性の差に関わる検証のみではなく、実際に使用される患者での多様な反応を評価する意図が含まれている。ICHによって、一応複数の検証的臨床試験が要求されることになったが、これが積極的に解釈されて説明的試験と実践的試験とを行う方向に向かえば、両者相俟って新薬の市販前の科学的評価が文字通り確立するといえる。

検証的臨床試験の1つである説明的試験の推進は自ずから臨床試験の質の向上のインセンティブとして働き、その後に実施されるべき実践的試験にとっても望ましいことである。その意味で吉村論文の指摘した論点の多くは意義のあるものと考えている。以下では主に意見を異にする点および吉村論文が事実を誤認している点のみを簡単に挙げるにとどめる。

1. プラセボ対照に関する視点の追加

プラセボを対照とする議論について、2点を指摘しておきたい。

* 経営システム科学専攻：〒112-0012 東京都文京区大塚 3-29-1。

** 疫学部：〒108-8638 東京都港区白金台 4-6-1。

第1は、現在、わが国で実薬対照試験が実施されている薬効群の殆どにおいて、歴史的に早期の段階でプラセボ対照が実施されてきた点が無視されていると思われることである。実際、必要な段階ではプラセボを対照とした優越性試験を実施すべきであるというのが、わが国臨床の多数意見である。現在、わが国で標準薬と呼ばれている薬は、プラセボ対照試験によって薬効が確立したものか、あるいはプラセボに対する薬効が確立した他の薬との比較の中で薬効が保証されたものであり、そうでない実薬対照群をおくことは少なくとも臨床家には許容されない。

第2は、プラセボ対照が必ずしも世界の臨床のコンセンサスではない点である。医学研究の倫理綱領である「ヘルシンキ宣言」について、1996年10月の南アフリカ共和国サマーセットウェストでの第48回世界医師会総会で改訂がなされたが、特にプラセボ対照と関係する部分は次のようになっている。

〈ヘルシンキ宣言からの抜粋〉

II. 専門職としてのケアと結びついた医学研究（臨床研究）

2. 新しい方法を治療に応用する場合には、予想される効果、危険性、及び不快さを、現行の最善の診断法や治療法による利点と比較考慮しなければならない。
3. いかなる医学研究においても、対照群に割り付けられた患者を含めて、現行の最も有効と考えられている診断法や治療法を受けることができるという保証が与えられなければならない。これは、立証された診断法あるいは治療法が存在しない研究段階における非活性プラセボの使用を除外するものではない。
6. 医師は、ヒトを対象とする生物医学的研究を専門的な医療の一部として行なうことができるが、この場合、その目的は新しい医学的知識をうることにありと考える。しかし、このような場合もその研究が患者に対して潜在的な診断的又は治療的価値があることが理由づけられる場合に限られるべきものである。

下線部が、何らかの理由で今回の修正で初めて付け加わった部分であり、プラセボ対照について新たに倫理綱領に盛り込む必要性が生じたことを示している。これを見ても米国FDAを除く世界の臨床のコンセンサスは、確立した治療がある場合には実薬対照であったことが理解できると思われる。

我々自身も説明的試験でのプラセボ対照を否定する立場にはなく、むしろどのような条件が整えば可能であるかを模索すべきと考えている。しかし、わが国で臨床試験に誠実に参加している現実の臨床家の状況は、「医師は治療するのが仕事であって、治療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しないことであるから医師としては行ない難い」といった単純なものではない。我々の知る限り、吉村論文が批判したような上記の姿勢を公言している研究機関は、わが国では、ある私立単科医科大学一校のみに過ぎない。

2. 優越性試験での有意水準が片側5%などの記述は事実誤認

「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、下した判定とのかねあいでは決めるべきこと」は個別の研究では妥当な場合があるが、薬の許認可の基準は研究者の任意の判断によるべきものではない。また、「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」及び「信頼区間方式で優越性を評価するときは信頼水準を90%にとるのが普通になっている」という記述があるが、これは事実誤認である。新薬調査会では、プラセボ対照を含めて優越性推論の場合には、有意水準2.5%の片側検定を2回実施するという基準（つまり、両側5%）で承認審査に当たっている。この点については、他の承認審査の統計担当者にも確認している。優越性試験での有意水準の考え方にFDAとの大差はないと思われる。

3. 非劣性試験についてはさらに統計家による検討が必要

エンドポイントが主観評価のように質的データの場合に、非劣性検証を目的とする臨床試験にノイズをいれることでかえって検証が容易になるというパラドックス（たとえば、佐藤（1994））が紹介されているが、統計学的に非常に重要である。これは、むしろ離散データ解析における基本モデルとして、ポアソン分布や二項分布といった「散らばり母数（dispersion parameter）」を持たない分布を前提とすることの危険性を端的に示している（椿（1996））。データの品質が本質となる分野においては、少なくとも散らばり母数を含む負の二項分布やベータ二項分布などが基本分布として採用されない限り、統計的に正しい結論、すなわち、「ノイズを増やせば、必ず有意性は示しにくくなる」ということは導けないと思われる。質的データの群間変動と群内変動の把握の必要性は、品質管理分野では、赤尾（1970）らによって古くから注意されていることであり、医学統計分野でも統計学者が積極的に貢献すべき研究課題である。

4. 多施設臨床試験に関わる施設間差の本質は医療の質の問題ではない

吉村論文では施設間差を「質の悪い医療」や「信頼性の低い測定・評価」と考えているようであり、その限りにおいて妥当な議論である。確かに現状ではこの問題は大きい。しかし、実際の臨床治療における患者の反応は多様であり、実践的試験としての多施設臨床試験において、施設間差が本質的な薬効のばらつき情報を含んでいる可能性を無視することは、現実的姿勢ではない。

5. 対応のある試験での非劣性の検討は不適当

薬効の存在を検証する説明的試験では、対応のある試験による非劣性の検討はごく限られた特殊な場合であろう。一方、患者の反応の多様性（個人差）を対応を取ることによって除外することは、実践的試験の性格に反するものである。

また、実際に行われている対応のある非劣性試験では、主要評価変数などで全く同じ評価が9割を超えることが殆どの状況である。真に薬効が同じためなのか、あるいは、測定精度の問題や同等であるとの思い込みバイアスによる同一評価であるかは、通常区別できない。測定精度やバイアスを度外視した対応のある非劣性試験は避けるべきである。

6. 日本の臨床試験の伝統はPC解析ではない

我々の報告で述べたように我が国の臨床試験を推進した臨床家によって全ての被験者（患者）を可能な限り統計解析に採用する方針（ITT解析）が当初から採られてきたのであり、新薬の承認審査においてもその方向の指導がなされてきた。確かに1995年の第3回計量生物学会セミナーでの製薬協関係者からの講演の中でPC解析優先のような主旨の報告があり、以降、誤解が広まったと思われるが、こうしたPC解析優先の考え方は製薬企業の臨床統計担当者や開発担当者の認識ないし望んでいる方向なのであろうか。

以上、吉村論文の本筋から外れた細かなものを含めて、若干気になる点のみを指摘した。勿論、むしろ賛意を表する点が多いことを付け加えておく。

参 考 文 献

- 赤尾洋二（1970）. 非復元同時サンプリングによる不良率の平均値と範囲の管理図, 品質管理, **21**, 11月臨時増刊号, 1646-1650.
- 佐藤俊哉（1994）. ランダム化にもとづいた intent-to-treat 解析, 応用統計, **23**, 21-34.
- 椿 広計（1996）. データ品質とマネジメント, 第64回日本統計学会要旨集, 52-53.

有用性と超多施設試験の意義

——椿・藤田・佐藤論文について——

東京大学* 大橋靖雄

(受付 1998年2月24日)

論文(以後略して「椿論文」とよびます)を一読して絶句いたしました。前提に対する認識そしてアプローチにおいて私とは著しく異なる立場からの論文であり、通常の科学雑誌上の討論のように、論点を明らかとしつつ誤解を解き双方の理解を深めるという進め方が取れそうにありません。「困惑」というのが私の実感です。根本的な疑問を投げかけると非難めいた口調になりそうですので、ですます調で述べたいと思います。もちろん私は3人の先生方にはいずれも面識があり(よく存じあげているという方が正確です)、尊敬する方々でもあられますので、私の見方を強要するつもりはありませんし、一方、私も見解を変える気はありません。議論はすれちがいでしょう。

1. 全ての薬効分野で、医師の評価する「有用性」が患者の視点に立ったベネフィット評価に直結するのでしょうか?

糖尿病の患者さんにとっては合併症を予防すること、つまり失明や透析入り、神経障害や血管障害を防ぐことが有効な治療に望まれることでしょう。大規模試験 DCCT や Kumamoto-study により確立されたと考えられる、短期で測定可能な代替エンドポイント HbA1c を主たる評価項目として開発研究が行われています。このような定量的な代替エンドポイントが存在する例には、高脂血症、骨塩量をエンドポイントとする骨粗鬆症(さらに抗圧薬)があります。

治癒がほとんど望めない進行した肺癌・大腸癌など多くの固形癌の患者さんにとっては、経済性を無視すれば生存時間の延長と QOL の向上・維持がベネフィットでしょう。標準治療あるいは first-line として確立させるにはこれらを比較試験において評価すること、標準治療失敗例で他に確立した治療が存在しない場合には、代替エンドポイントとして腫瘍縮小で新薬を認可することが行われています。ただし日本では、生存時間をエンドポイントとした標準治療との比較試験は市販後に行われるようになっていきます。

決定的な治療法がほとんど存在せず、新薬が期待されている多発性硬化症や筋萎縮性側索硬化症などの神経難病の患者さんにとっては、ヒトとしての機能が徐々に損なわれていくことが最大の問題です。世界中でさまざまな治療法が試験されており、その成果を共有するためには、評価者によるぶれを小さくした(すなわち評価者間信頼性を確保した)機能の評価スケールを用いることが、不幸な患者さんにできるだけ早く有効な薬剤を届ける道ではないのでしょうか。現在の臨床試験はもちろんこのように行われています。さらに、介護者の負担の軽減もエンドポイントとなりえますし、軽症患者の場合なら患者自身による QOL も評価すべきでしょう。

私は、新薬開発の臨床試験に、会社側のアドバイザーあるいはコントローラーとして1984年から関与しております。適応症を数えると、癌以外では約50、癌では70以上の開発研究(癌では一つの薬剤を多くの癌種に用いるので数が増えてしまいます)に参加しています。上記の例は私が多く参画している分野ですが、この数年、全般改善度を主たるエンドポイントとし有用度

* 医学系研究科 健康科学・看護学専攻 生物統計：〒113-0033 東京都文京区本郷7-3-1。

を重視した臨床第Ⅲ相試験に参加しておりません。これは分野の偏りかもしれません。私は「有効性」と「安全性」を総合的に評価する「有用性」事体を否定するものでは決してありません。知りたいのは、どのような分野において、医師の評価する「有用性」が消費者たる患者さんのベネフィットの評価になり、これまで日本で行われてきたような100-300例程度の超多施設臨床第Ⅲ相試験が実態のシミュレーションの価値を持ちうるのか、です。

抗精神薬はそうかもしれません。痴呆もそうかもしれません。現に、世界で初めてアルツハイマー病に対してFDAが承認したタクリンの有効性評価は医師による総合評価も用いて行われましたし、この肝毒性の強さを考慮すると、有用性評価を行うべき薬剤であると、私も思います。患者さんの立場に立つ評価の極限は患者さん自身によるQOLの評価でしょうが、これらの領域では不可能でしょう。議論はあるでしょうが、ある程度有効な薬剤の存在するパーキンソン病など、神経内科領域の薬剤や、慢性の喘息薬にも有用性評価が意義を有するものがあると思います。抗菌効果の強さとスペクトルが調べられていることを前提とするなら、抗菌薬の評価も候補かもしれません。

これら候補となる分野を並べてみると共通点が現れてきます。

ある程度有効な薬剤が存在すること、(それとの間接的な比較で)短期で効果を判断でき、それが医師の処方行動に直結すること、です。実薬対象の「同等性」でゼロ新薬が認められてきたことは他の分野と共通ですが、疑惑の、すなわち日本でのみ承認されているような薬剤もこの分野に多いように思います。

くどいようですが、有用性評価自体を私は否定しません。(もちろん、患者さん自身によるQOL評価も多くの場合において有意義と思っています。) 椿論文の前提である、有用性評価が消費者の立場からの評価に直結する薬剤あるいは薬効分野には限定があるのではないかと、言いたいのです。

2. 第Ⅲ相試験で有効性の評価は終わるのでしょうか？

椿先生も藤田先生も長期の市販後臨床試験に参加されたご経験がありますので、長期大規模の市販後臨床試験で真の薬効が初めて検証される分野が多いことはご存知のはずです。臨床試験の社会的認知のレベルと学問上の地位が我が国では低く、薬物治療法研究への貢献がほとんどなかった大きな原因は、国による大規模試験への投資がなかったことであることを、事あるごとに書いています(大橋(1997))。ようやく、高脂血症、骨粗鬆症、糖尿病、脳梗塞予防、肝硬変の分野などで、長期大規模臨床試験が開始され一部結果も公表されるようになってきています。癌の分野では、第Ⅱ相後に承認される新薬は一つの弾丸にすぎず、組み合わせ第Ⅰ/Ⅱ相試験、そして比較試験を通じ標準治療が確立されていきます。このような長期試験と抗癌剤開発の仕事が主となっているためでしょうか、私は小規模・短期の(抗癌剤以外の通常のこれまでの我が国の)第Ⅲ相試験が実践的である、あるいは実践的であるべきとはとても思えません。異論はあるでしょうが、私は、実践的試験とは、標準治療確立をめざし真のエンドポイントを対象として行われる(通常は市販後の大規模)臨床試験への形容であると考えています。従って、通常の第Ⅲ相試験に対しては、実践的であることを根拠とする椿論文の実施原則(2.2.2のアからオ)を認めません。もちろん、「認めない」というのは全ての薬剤・薬効分野に常に当てはめるべき原則としては賛同しないということであり、医師による有用性がそのまま患者のベネフィットに通じ市販後臨床試験の存在を前提としない領域(どのような領域でしょう?)では、意義を認めるのにやぶさかではありませんし、市販後の大規模試験に対しては、一つの原則になりうると評価します。実際、乳癌術後の補助療法として確立しているタモキシフェンの投与期間を検討する目的で国際共同で行われているATLAS研究(Davies 他(1997))は、このような考え方にのっとって行われています。

なお、先日(1998年2月6-7日)私が主催した第10回臨床腫瘍研究会の招待講演のために来日した米国の National Surgery Adjuvant Breast Bowel Project (乳癌、大腸癌の術後補助療法開発研究を行ってきた研究グループ)の Dr. Wickerham と Mr. Cronin は、彼らが行う大規模長期の臨床試験とは実験ではなく state-of-the-art care であると言っていました。私の思う実践的試験とはこのような試験です。

3. Fisher vs Neyman-Pearson 論争を思い出しました

椿論文あるいは椿発言の根底に、たくさんの類似薬が開発されていく中で、全体としての (in-the-long-run) 消費者のベネフィットを確保する、という考え方があります。ベネフィットの評価の問題は既に議論しましたので、ここでは、たくさんの薬の開発と、in-the-long-run での評価という前提について意見を申し上げたいと思います。私は前者は認めず、後者には違和感を感じます。これまでの我が国のゾロ新開発は異常な事態であり、これが続くことは、規制当局の現在の考え方からも、製薬会社の最近の態度からもほぼありえません。後者はやや微妙です。確かに統計手法の性質は in-the-long-run で評価されますが、個々の薬剤の評価は「科学として」それぞれ固有の特徴を有しており、平均してリスクを押さえることに違和感を感じます。僭越ですが、おそらく Fisher が感じた違和感ではないでしょうか。同じ研究室(東京大学計数工学科奥野研究室)出身ながら椿先生と私の間にこのような考え方の差があることは興味深い現象です。椿先生は田口玄一先生の品質工学に学生時代から心酔し(私も田口先生の思想を高く評価します。しかし残念ながら講演や著書からは、学生時代にはこれが理解できませんでした)ISOの標準化の仕事にも大きく貢献されておられます。一方私は1984年に病院に転出した後は、本質的な仕事は臨床家との共同作業がほとんど全部です。この環境の違いでしょうか。

インパクトのある画期的な薬剤が、理論を育て、ときには新たな需要を作り、薬効評価のあり方さえ変えてしまうことはよくあると思います。最近では、高脂血症に対する HMG-CoA 還元酵素阻害剤(プラバスタチンなど)、骨粗鬆症に対する 2 燐酸塩(アレンドロネイトなど)、膀胱癌に対するゲムシタビンなどが、私の間接・直接に経験した例です。最新の HMG-CoA 還元酵素阻害剤の中には、第II相試験の最低用量で 100% 著効(15%以上の総コレステロールの減少)を達成するものが現れるに至り、従来の判定基準が完全に陳腐化したことを象徴しました。ちなみに、日本では高脂血症の有効性評価にコレステロール値の減少をカテゴリー化して評価していると FDA の Dr. Dubey に言ったところ、「信じられない」とあきれていました。

有効な 2 燐酸塩の出現により骨粗鬆症および骨量減少症の臨床評価ガイドラインは国際的に大揺れとなり、日本のガイドラインも決着にいたっておりません。ゲムシタビンの例では、clinical-benefit-response という患者さんの QOL の視点にたった新しい評価基準による承認申請を FDA が認めました。

このような状況が例外というより普通になるのが、今後の新薬開発ではなかろうかと感じています。このような場面における統計家の役割は、現実の(しばしば誰も経験していない)問題に研究プロジェクトチームの一員として柔軟に対処し解決することであると思うのです。薬効評価はかくあるべし、という理念を応用するのではなく、現実の薬剤あるいは薬効分野の仕事の経験を通じ、共通する考え方が整理・昇華・抽象されて「臨床試験方法論とくに統計的側面」が形成されるという感じがします。少なくとも私はそう行動してきたつもりです。

整理・昇華・抽象する過程で、現実の問題の持つ存在感と複雑さ(いやらしさ)、みずみずしさが失われます。ICH 統計ガイドラインに限らず、臨床試験統計方法論のガイドラインや教科書的講演がしばしば無味乾燥かつ曖昧となるのは、ある程度いたしかたの無いことです。(余談ですが、これに輪をかけたのが昨今盛んなデータ管理に関する議論です。)これからの試験統計家に期待されることは、ガイドラインを覚え応用するのではなく、ガイドラインが作られてき

た背景・問題を理解し、新たな問題に対処できる力を養うことであると思います。そのためには、どうしてもある程度の臨床の知識が必要になると思います。私が癌の分野の研究に比較的積極的に関与しているのは、この分野の臨床の知識がある程度身についたからと思うからです。アメリカの臨床腫瘍学会に出席しはじめてから7年ほどになりますが、研究の背景や意義を臨床の文脈の中である程度理解できることが、計画・解析の大きな助けとなることを実感しています。この教育をこれからの試験統計家やデータ・マネージャにどう行うかは、頭の痛い問題です。

話がずれてきましたが、個々の開発研究の個性を強調したいということ、個々の問題に柔軟に対処しない限り、一般・抽象としての統計的方法論の発展（統計家の成長）も無い、というのが私の言いたいことです。

4. コントローラ委員会の功績は十分に評価したいと思います

我が国の臨床試験黎明期におけるコントローラ委員会の貢献には疑いがありません。コントローラの存在が我が国の臨床試験の公平性確保に果たした役割も大きなものであり、将来にわたっても、何らかの形で第三者が公平性確保に貢献するシステムは存続の意義があると評価します。内藤先生が挙げられたコントローラの役割（椿論文 2.2）のうちアからカまで、そしてケは、一人ないし二人の人間で実行可能であり、コントローラと呼ばれようが統計解析アドバイザーと呼ばれようが、第三者が参与することで、（まだしばしば）未熟な製薬会社スタッフを支援し同時に公平性の確保も図れると思います。キとク、すなわちデータ収集・点検と統計解析の実施は個人ではもちろん対応不可能であり、逆に品質保証もおぼつきません。やはり、現在の流れである専門部署の設置と SOP による業務および責任分掌の明確化、社内監査に代表される内部品質保証に頼らざるを得ないと思います。統計解析の検証がときどき問題となりますが、第三者による独立な解析による検証も行われるようになると思います。もちろん外部の統計専門家に依頼することでも、CRO に委託することでもよいと思います。審査センターが解析の機能を充実させていくことを考慮すれば、公的な第三者機関などの存在は不要でしょう。現在の我が国の社会では、下手な権威化と人事の停滞、非効率が起きそうだからです。

椿論文では職能倫理が未成熟であること、そしてより根本においては、企業性悪説に立ち（生産者と消費者が相反する利益を求めるとする図式です）第三者制度を価値付けています。あまりに後ろ向きではないでしょうか。私は日本の製薬会社で生き残れるところは3-5社がよいところ、と常々言っております。社会に貢献することを標榜すると同時に実行し、消費者から尊敬される会社でない限り（つまり両者の利益が同一の方向を向かない限り）、現在の荒波は耐え切れなと感じています。今年度の日経企業評価 PRISM で、公開性の評価が高いことから（収益では他にも優れた製薬会社は存在したもの）武田薬品が全業種の中でトップに評価されたことは象徴的です。3-5社くらいになれば、職能倫理に富む優秀な人材も確保できるのではないのでしょうか。

5. なぜ超多施設試験がまずいのか

実は、オーガナイザーから臨床試験の結果の「一般化」と施設数の問題についてまとめるようお薦めを受けましたが、椿論文に対する意見の形でここにまとめた方が主張が明確になりそうです。

超多施設試験、極端には1施設1例であっても無作為化が適切になされれば（事前の）偏りは存在せず、この意味で試験は妥当です。FDAの統計家がこのような試験には偏りがあると言っていた、と佐久間昭先生がいぶかっておられましたが、このような概念が彼らの頭の中には存在しないため「うろたえた」というのが本当のところではないのでしょうか。S. Senn も、1996

年の英国レスター大学のシンポジウムで上記のような試験の妥当性には問題ないと発言していました。

私が超多施設少数例試験を問題とするのは、表面の統計的妥当性ではなく、これが大きな歪みの結果であるからなのです。つまり、これまでの保険制度と薬価制度のもとで安定した利益が期待されるゾロ新の開発に企業が走る。見識と「力」・経験を備えた数少ない臨床家（グループ）に総括が集中する。多くは温厚で人格者でもあられるこれらの先生方は、不公平のないよう多くの仕事を引き受ける。当然1施設内で、同じ適格条件の患者さんを対象とした臨床試験が同時進行する（本来なら倫理委員会で問題とすべきことです）。似たような薬が発売される中で、施設で採用されるための条件として「臨床試験から参加」したことが課せられ、状況はさらに悪化する。安全な前例にのっとるため、プロトコル開発とは前の成功した薬剤のそれを写すこととなり、当然のこと、統計的思考の入る余地はなくなる。したがって企業内の統計家の地位は上がらない。

少しでも臨床試験の実務に触れば、症例選択が成績の良否を決定する要であることが理解できます。多少私に経験のある進行乳癌を例にとれば、骨転移のみを評価部位とする患者さんを多数登録すれば腫瘍縮小の奏効率はあがらず、小さなリンパ節転移と皮膚病巣を対象とすれば奏効率は上昇します。10%くらいから50%くらいまでは操作可能でしょう。これほど極端ではなくとも、同時に複数の同じ登録条件の臨床試験が進行した場合の「試験選択バイアス」は、試験結果に大きな偏りをもたらし一般化を困難とするのではないのでしょうか。癌の臨床試験のグループの中には、一般化可能性を高めるため、全ての適格患者に同意を求めることを原則としているところもあるくらいです。適格患者に対する同意取得状況と登録状況のログをとることも多くのガイドラインや教科書で薦められています。

椿論文では、実践性を重んじることから施設数を増やし、治療効果と施設との交互作用を含めた誤差により検定することにより結論の頑健性を向上させる、という論旨をとっています。実践性に対しての疑問は既に述べました。後者はどうでしょう。まず、ありなしの2値応答の場合にこの議論が成立するのでしょうか。次に、Fisherが「3原則」中の繰り返しの原則について述べたとき、その意義を、平均化による安定化と同時にばらつきの大きさの推定に置いたことを思い出します。施設差そして治療効果との交互作用を推定可能とし、得られた治療効果の臨床的意義をこれらの推定値との相対評価で行うような試験、椿論文によれば「科学的」試験が治療法の評価と進歩の貢献には必要ではないのでしょうか。少なくともこれを省略することはできないと思います。また施設の質を評価するためにも、ある程度の症例数が施設あたりに必要です。科学の話ではありませんが、モニタの手間（とくに照合）と費用（旅費）を考慮しても、施設あたりの症例数がある程度確保の方が経済的です。私は超多施設少数例試験の意義を完全には否定しません。使用経験を臨床家に与えるという意味で、上記の「科学的」試験と併存させることも可能かもしれません。

結論の一般化可能性向上のために、ノイズと薬効の交互作用をもとに検定するという田口流の考えを推し進めると何が起きるのでしょうか。検出力確保のために必要サンプルサイズを計算するという思想は田口メソッドには希薄です。とくにありなしの2値反応の場合、施設差を考慮したサンプルサイズの補償の方法を私は知りません。2項分布に基づく単純な式によるサンプルサイズのもとでは検出力は減少し、一方で第一種の過誤は設計通りに保たれますので、正の「予測値」、すなわちプラセボあるいは対照薬と比較して有意に優れている（あるいは有意に劣っていない）と判断された薬剤が真に優れている（劣っていない）確率は減少します。皮肉にも、これは消費者危険を上昇させることに通ずるのではないのでしょうか。

今述べたことは思考実験です。既に、開発研究の個性を私は尊重したいと述べましたように、たくさんの薬剤の候補が存在し、許認可という一種のゲーム（抜き取り検査でいえば取り引き）

の中で統計手法適用の結果が in-the-long-run でどうなるか、という問題の立てかたを私は好みません。ここでは椿論文の土俵に上がり観念的な議論を行ったにすぎません。そもそも田口メソッドは、既に存在する製品の特性をどう評価するか、あるいは性能を頑健化させる（パラメータ設計といいます）ための工学方法論であり、たくさんの候補物質から有用な物質を生産者・消費者双方のリスクを押さえたもとで選択するという、椿論文の土俵になじむ方法論なのでしょうか。

本号の吉村功先生の論文に、特別な教育を受けていない消費者でも問題なく使える工業製品と、資格化された専門家である医師が使用する薬剤とを同列に論ずることの危険が指摘されています(6.2)。私も同感です。親しくさせていただいている臨床腫瘍学の専門家がこの議論を聞いたならば、烈火のごとく怒り出すと想像します。椿論文にある「参入障壁を作らない」(2.2.3)とした方針は、歴史を振り返ってみて果たして正しかったのでしょうか。評価を医師のみに委ねること、広く参加できるよう構造を単純化することにより、臨床試験を行うことは普及したものの、臨床試験を行う基盤が未整備のまま放置されてしまったのではないのでしょうか。

6. おわりに

臨床試験コーディネータの必要性がようやく声高に叫ばれ、試験に参加する施設と医師の要件、さらには試験統計家の資格が議論されています。現在進行中の我が国の臨床試験システムの変革は、他の社会的変革と切り離せない「構造的」な大変革であると感じます。歴史に学び、多様な視点から議論を行い、公開し、社会からの批判を仰ぐことが必要に思います。この意味で本稿が少しでも貢献できれば幸いです。

参 考 文 献

- Davies, C., 野村雍夫, 大橋靖雄 (1997). 乳癌術後補助療法における Tamoxifen の投与期間, 癌と化学療法, 24(10), 1203-1209.
大橋靖雄 (1997). 薬物評価への統計的アプローチ——国際ハーモナイゼーションを受けて——, 呼吸, 16(6), 897-901.

吉村論文および椿・藤田・佐藤論文に対するコメント

国立医薬品食品衛生研究所* 山口拓洋・安藤友紀・秋田倫秀

(受付 1998 年 2 月 13 日)

1. はじめに

ICH 統計ガイドラインは今後の医薬品の開発に大きな影響を与えるものと考えられるだけに、医薬品の研究・開発・承認等に関わる者、特に生物統計学にたずさわっている者にとって重大な関心事であろう。

* 医薬品医療機器審査センター：〒105-8409 東京都港区虎ノ門3-8-21 虎ノ門33森ビル10F.

今回は、吉村論文、椿・藤田・佐藤論文で産官学の学の立場から ICH 統計ガイドラインに対する意見が述べられており、オーガナイザーから「産官学それぞれの意見を交換することにより議論が充実し、ICH 統計ガイドラインに対する理解と認識が深まるのではないか」との趣旨の申し出があった。本稿をまとめるにあたっては、医薬品医療機器審査センター（以下、審査センター）の統計担当審査官の間での意見交換を行っているが、ここでの議論は必ずしも実際の承認審査の場に反映されるものではない。また、審査センターを代表するものでないことも申し添えておきたい。

2. 両論文の構成と議論の進め方

吉村論文では、ICH 統計ガイドラインの「III. Trial Design Considerations」について5つの問題点を挙げて個人的見解を述べている。一方で椿・藤田・佐藤論文では、日本での臨床評価の歴史を振り返り、ICH が我が国の臨床試験の原則にもたらす危険性について特に「患者の立場」から指摘を行っている。視点は異なるものの、日本の臨床試験を考える上で両論文とも大変に意義深いものであり、臨床統計家の果たすべき役割の重要性を切に感じた次第である。以下では、少なくとも一方の論文で議論されていた内容のうち、有用性、同等性試験（ICH 統計ガイドラインでは非劣性試験）、多施設共同試験の3点について、両論文を引用しつつ意見を述べたい。なお、文章中での「ICH ガイドライン」は「ICH E9 Guideline: Statistical Principles for Clinical Trials (1998)」, 「現行ガイドライン」は「臨床試験の統計解析に関するガイドライン（厚生省, 1992）」を指すものとする。

3. 有用性

椿・藤田・佐藤は、

“有用性評価とは、仮想的に同一症状の患者治療法を再び行う状況に置かれたとき、患者の視点に立ってどの程度今回の治療法を繰り返すモチベーションが生じているかを評価するのが原則であり、…(後略)…”

と述べている。「有用性」が有効性と安全性に加え、患者特性や病態などに対する担当医の評価を加味したその治療法の総合的な評価とすれば、新しい治療法の臨床的意義そのものを評価した指標であるといえるだろう。

しかし一方で、有用性評価は benefit と risk に関する総合的なバランスの評価を担当医がそれぞれ個別の価値判断により行うために客観性と再現性に乏しい評価である、というこれまでの数多くの指摘は否めない。有用性の評価尺度例えば有用度が、臨床試験で用いられる評価尺度として要求される妥当性・信頼性（再現性）・感度などを十分満たすものであると説明することは極めて難しいのではないだろうか。

これらを考慮すると、有用性を臨床試験の主要評価項目として用いるべきではないと我々は考えている。ICH ガイドラインでは、

“If global usefulness is specified as primary, it is important to consider specific efficacy and safety outcomes separately as additional primary variables.”

となっており、有用性とは別に有効性と安全性の評価を考慮する必要性を唱えている。

ただし、椿・藤田・佐藤の、

“…1次情報を知る臨床医の臨床判断（本来の総合評価）によって、客観的測定などのプライマリーエンドポイントの意義を確認・補正すべきなのである。”

という意見は参考にすべきと考えている。ここでの主張のように評価項目の一つに有用性を加えプライマリーエンドポイントの意義について確認することは、一つのアプローチとして否定するものではない。

4. 同等性試験

ここでは混乱を避けるため、従来の「同等性」という用語をICHガイドラインでの「非劣性」と同義で用い、1) 同等性の立証方法と許容できるマージン、2) 同等性試験における解析対象集団、の2点について意見を述べる。

4.1 同等性の立証方法と許容できるマージン

現行ガイドラインで同等性を積極的に証明する方法が提示され、かつての「有意差無し＝同等」という検定の誤用はなくなりつつあるが、様々な問題点も指摘されている。臨床的に許容できるマージン(Δ)の設定に関する議論もその一つである。ここでは特に吉村の、

“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きを構成するための操作的な差である。…、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。いやむしろ積極的に寄与しなければならない。… $\Delta=10\%$ 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。”

に対して意見を述べる。

臨床試験に参加している臨床医の大部分は、同等性検証の（特に統計的な）手続きの詳細は理解していないと思われる。ただし、吉村が、

“帰無仮説を形式的・機械的に理解して、10%も劣る薬を受け入れることだとするのは、現実の理解としての的を射ていないと思う。”

と述べているように、10%も劣る薬を受け入れることを認めているわけではないと思われる。実際には、同等以上を示すためには試験薬が対照薬より若干優れていない限りサンプルサイズが非常に大きくなる。そのために10%も劣る薬を受け入れるようなことが現実起きる可能性は非常に小さいのだが、臨床医がこの事実を理解しているかどうかはわからない。ゆえに、吉村の、

“私は、その合理性を認めて臨床医もこの手続きを受け入れているのだと考えている。”

という意見には多少疑問がある。すなわち、吉村の言う“実際の意味のずれ（二つの解釈のずれ）”は臨床医には無意識のうちにしか認識されていないのではないだろうか。それが正しいとすれば、今後適切なマージンを設定する必要性が生じた場合に、多分に臨床的な問題である、の一言では片づけられないと考える。多分に統計的な問題である、では言い過ぎかもしれないが、“実際の意味のずれ”が生じる限り、むしろ統計家が臨床家を先導しつつ適切なマージンを設定する必要があるだろう。

4.2 同等性試験における解析対象集団

ICH ガイドラインでは、

“In superiority trials the full analysis set is used in the primary analysis (apart from exceptional circumstances) because it tends to avoid over-optimistic estimates of efficacy resulting from a per protocol analysis, since the non-compliers included in the full analysis set will generally diminish the estimated treatment effect.”

とあり、優越性試験においては full analysis set が主要な解析であると位置づけているが、同等性試験においては、

“However, in an equivalence or non-inferiority trial used of the full analysis set is generally not conservative and its role should be considered very carefully.”

と記述しているように、同等性試験における PC 解析と ITT 解析との関係は優越性試験におけるそれと異なるので注意が必要である。吉村は、

“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。”

と述べている。同等性試験の場合、どちらが主解析か実施計画書に記載しておく必要があり、都合のいい結果を採用しないよう心がけたい。また、結論が異なればその理由を十分に考察する必要があることも強調したい。

5. 多施設共同試験

分野により程度の違いがあるものの、日本の第Ⅲ相臨床試験は多数の施設から成り立っており、その賛否に関して現在まで様々な議論が繰り返されている。それらの意見には、第Ⅲ相試験を実践的試験 (pragmatic trial) と科学的実験 (scientific experiment) とのどちらに位置付けるかの違いが反映されていると思われる。ICH ガイドライン「3.2 Multicentre Trials」は基本的には後者の立場で書かれており、椿・藤田・佐藤は前者の立場から ICH ガイドラインを、

“基本的には、「技術評価」の側面を弱める危険がある。”

と指摘している。

我々は、1施設あたりの症例数を増やし施設数を少数に限定して試験を行うことの意義は、試験の質の向上が期待できると共に施設毎に質の評価の検討が可能であること、施設と治療効果との交互作用を調べることで治療効果の一般化可能性が検討しやすくなること、施設間差が生じていた場合にその原因を特定することができれば広い意味で医療の質の向上に役立つこと、にあると考えている。したがって、吉村の、

“…誤差の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。”

という意見が正道であると思う。ただし一方で、多数参加者少数施設試験を実施する場合、施設をどう定義するか、1施設あたりの症例数は何例必要か、などについての検討が必要と思われる。

最後に、多施設共同治験の統計解析で用いるモデルに関して意見を述べる。ICH ガイドラインでは「3.2 Multicentre Trials」の最後に、

“Up to this point the discussion of multicentre trials has been based on the use of fixed effect models. Mixed models may also be used to explore the heterogeneity of the treatment effect. These models consider centre and treatment-by-centre effects to be random, are especially relevant when the number of sites is large.”

とあり、特に施設数が多い場合には、施設間差（ベースライン、交互作用）を変量効果とみなした混合効果モデルでの解析も示唆されている。このアプローチは椿・藤田・佐藤の要求を多少なりとも満たすものと考えられる。ただし、混合効果モデルを用いる場合には固定効果モデルの場合以上にモデルの適用について詳細な検討を行う必要があり、また、得られた結果の解釈には十分に注意を払うべきである。施設数が多い場合でも、必ずしも治験参加施設が日本の医療施設からの無作為標本とは見なすことができないので、結果が一般化できるかどうかについては検討が必要であると考ええる。

6. おわりに

非常に簡単にはあるが、吉村論文および椿・藤田・佐藤論文に関して意見を述べさせていただいた。もちろん、我々に誤解があれば遠慮なく指摘していただきたいし、忌憚のないご意見が伺えれば幸いである。

これからの新薬開発における課題

旭化成工業* 佐々木 秀 雄

三菱化学** 酒 井 弘 憲

(受付 1998 年 2 月 18 日)

吉村、椿らの指摘を受けて、製薬企業における統計解析担当者としての立場から、いくつかコメントをしたい。主に、ICH での議論に基づいた「原則」をどのように有効に活かし具体的実現に結びつけるか、という視点から述べることとする。

新薬開発の形骸化について

「標準化が進展した一方で、形式のみを借用した形骸化の弊害もある」との椿らの指摘は、製薬企業で新薬開発に携わる我々にとって、大いに反省すべき点であろう。吉村の「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」という指摘にも同様の視点が含まれると考え

* 医薬開発センター：〒 101-8481 東京都千代田区美土代町 9-1。

** 臨床開発部：〒 140-0002 東京都品川区東品川 2-2-24。

られる。過去においては、ガイドラインが施行されると、あたかもそれが“ゴールデンスターダート”であるかのように取り扱われ、本来必要な多くの議論や検討が省略され、単なるガイドラインのトレース、形式的手順の遂行に腐心することも少なくなかった。形式主義により、ガイドラインの精神が反映されないばかりか、逆に非科学的な臨床開発／試験計画に陥いることがあったかもしれない。

E9(臨床試験のための統計的原則)を初めとしたICHの多くのガイドラインは、原則のみの提示を基本としており、このような形式主義に陥る危険性を軽減させていることは重要なことと捉えるべきであろう。ガイドラインが基本的な考え方を提示するに留め、医薬品開発の具体的な手順に言及していないことは、今後は疾患、薬剤その他の状況に応じた適切な新薬開発が求められている、と解釈すべきであろう。

製薬企業の統計解析者として臨床試験に係わる我々にも、勿論求められることが多い。単なる解析家としての参加ではなく、統計的視点に立った計画立案への積極的な関与が、最も重要な使命として捉えられなければならない。目的と仮説の設定、試験デザインの構成、評価指標の妥当性、解析手法の選択、試験規模の設定など、計画の根幹において何が合理的で科学的かという視点に立ち、役割を果たさねばならない。まずは、自らを律する機会としたい。

「原則」の実現について

今後は、「原則」から一歩進んだ、具体的実現を目指した議論の重要性を強調したい。ICHで多くの原則が合意されたことは、疑いなく意義深い。一方で、合意された内容は基本的に「原則」であることも事実である。国際間で合意すべき事項については、原則、総論、概念レベルとならざるを得ないことは十分に理解できる。そこで、今後さらに重要な事は、原則に基づいた適切な実施であるのではないだろうか。原則の議論(総論的議論)から、各論への展開の時期に来ているのではないだろうか。E9を初めとするICHガイドラインの合意は、これからの新薬開発における具体的議論のための布石が打たれたに過ぎないと我々は理解したい。

その意味で吉村の、「(原則を実現するための)具体的手続きは十分合理的でなくてはならないし、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。」という指摘は、極めて有益な提言と受け止めたい。

今後、新薬開発や承認審査に携わる統計家は、疾患や薬剤に特有の問題にまで踏み込んだ、具体的問題の解決に今まで以上に参加する機会が増える事になるであろう。議論の切り口は既に幾つか提案されている。例えば一つは、プラセボを対照とする試験は、如何なる疾患の如何なる目的の臨床試験で実施すべきであるか(する必要がないか)、実施可能であるか(実施は不可能であるか)、という議論であろう。そこでは、当該領域の既存の情報が吟味され、実施可能な臨床試験から如何なる情報が得られ、それが承認審査にどれだけ役立つか、といった議論が必要であろう。当然、臨床研究家と共に統計家の参加が望まれよう。また、別の例としては、実薬対照同等性試験や非劣性試験における「許容出来る差」をどの程度に設定するかという課題がある。さらに有意水準の議論となると優越性試験においては、両側5%、片側2.5%がガイドラインで推奨されており、これまで用量反応試験で用いてきた片側5%の有意水準の見直しが必要となるかもしれない。製薬企業にとってもcost-performanceを考えると、従来のように漫然と既存薬とそれほど違いのないme-too drugを開発する意欲もなくなるであろう。

対照群の選定についての議論はICH-E10(対照群選定のためのガイドライン)で進行中であるが、「許容出来る差」の大きさについては、疾患、薬剤その他の固有の事項が個別に考慮されるべきであろう。「許容出来る差」は、吉村が指摘するように「承認手続きを構成する操作」として受け止めることもできよう。しかしながら一方で、実薬対照の試験結果から試験薬が対照

に劣らないだけでなく、プラセボに上回って有効であることを間接的に示しうるのか、という議論も避けることはできないだろう。これらの精密な議論にも、疾患、薬剤その他の固有の事項が考慮されるべきであると考ええる。

以上はほんのいくつかの例に過ぎないが、「原則」に基づいた適切な実施についての議論が多方面で継続される必要がある。

コントローラー、第三者による評価について

コントローラーは日本独自の制度であり、歴史的に日本の新薬開発に果たしてきた役割は、極めて大きいものであったことは疑いない。しかし一方で、多岐に渡る業務が「コントローラー」の名の下に特定の個人に対して過度に集中してしまったことも否めない事実である。

従来のコントローラーの役割は、以下に大別できよう。即ち、「臨床評価の専門家」、「第三者としての監視者」、「盲検性の維持者」、「統計的アドバイザー」等である。臨床試験がより科学的に厳密となり、個々の業務に高い専門性が求められる今日では、このような種々の役割を一手に担う事にも限界があるかもしれない。今後は、当該試験が置かれた状況によっては、各々の役割を適切な専門家が分担することも考慮の価値があらう（勿論兼務も含めて）。

さて、これらの役割のうち、「第三者による監視者」の重要性が、椿らによって指摘された。この役割は確かに重要であり、「コントローラー」が評価される一つの所以であろう。ところで、新GCPでは、治験のスポンサーである製薬企業側の責任が強調されている。ここで、「製薬企業の責任」をはき違えないことが肝要である。この「責任」は自己完結的な企業責任だけを意味するのではなく、広く社会的責任を担っていると考えるべきであらう。その意味で、「第三者による監視」のシステムを、従来にも増して積極的に導入することも、企業としての「責任」の取り方の一つといえるのではないだろうか。実現のためには各種の方法が考えられる。従来のコントローラー的役割を持つ人を確保することも一つであらうし、さらに「独立データモニタリング委員会」などの機能を積極的に活用することも考えられよう。また、社会が広く係わる機会を与えるためにも、情報公開の場に自らを置くことも別の方法かもしれない。

情報公開による新薬開発から承認までの過程の透明化は、セレクションバイアスを否定できないという批判を受けても、なお高い価値があらう。情報が公開され、広く批判の矢面に立つことは、当事者が社会的に鍛えられることを意味する。また、情報が公開されることは、社会全体が責任の一部を担うことでもあり、新薬開発を取り巻く全ての環境が成熟することにもつながる。その意味で、製薬企業だけでなく新薬開発／承認に関与した全ての当事者が、常に社会的監視を受け、ある種の緊張感を保持し続けることは望ましいことであらう。

有用性について

ICHの議論では、「有用性」を患者毎に評価することは、肯定的には取り扱われていない。これは、欧米では、「有効性」と「安全性」の独立した評価によって、新薬の評価がなされてきたことと、主観的な評価を排除する傾向に起因していると思われる。E9においても、「有用性」という概念の問題点として、「比較される治療が有益な効果と有害な効果のまったく異なったプロファイルを持っている場合でも同等だと断言してしまう場合がある」と指摘されている。しかしながら、このような問題点があったとしても、椿らの指摘するように「有用性」は患者の視点に立った評価としての有益性が期待できるし、事実、E9でも、「医薬品の使用を決定するために利益と危険を秤にかけなければならない臨床医の意志決定過程を反映している」とも述べられている。患者の立場にしろ、医師の立場にしろ、利益と危険を総合的に評価する事は、新薬

の評価に有益な情報となる場合もあるだろう。

さて、この視点に基づく評価指標の開発がなされなかった訳ではない。その一つの取り組みが、QOLの評価であろう(事実、FDAによって、抗癌剤がQOLを主たる評価指標として承認された事例も報告されている)。しかしながら、日本で従来用いられてきた「有用性」は、「患者の視点」や「臨床医の意志決定過程」を真に反映したものとなっていたのであろうか。従来の「有用性(度)」の多くは、単に「全般改善度」と「概括安全度」の組み合わせで、半ば機械的に評価されてきたのではないだろうか。有用性についても「有用性(度)」という概念的過ぎる名称と評価尺度は、改善の余地があるのではないだろうか。実際、有効性に対して「有効性」という名称の評価変数は存在しないし、安全性に対して「安全性」という評価変数も殆ど使用されない。従って、医師の視点に立つならば、「再度この治療を当該患者に使用したいか」、患者の視点に立てば「再度この治療を受けたいか」という評価であることが明瞭となるような、評価変数としての適切な名称と具体的評価尺度の開発こそが望まれるのではないだろうか。したがって、「有用性」という概念の価値を議論することに加え、今後は利益と危険を(さらには治療における操作上の問題点を含め)総合勘案する指標の開発を進める必要があると思われる。

多施設共同試験について

若干の誤解があるようだが、ICHで合意されたガイドラインにおいては、科学的な少施設試験が推奨されている訳ではなく、また、二回以上の検証的試験が要求されてもいない。E9においても、多施設共同試験は「得られた結果を後に一般化するためのよりよい根拠を与える」とされている。

さて、施設当たりの症例数については、本邦でも長く議論され続けているものの、未だ結論が明確になっていない問題である。おそらく主たる論点は次のような事項であろう。

- 1) 対象母集団に対する代表性(治験の対象集団は実際の治療対象集団を反映しているか)
- 2) 施設の統計的な捉え方(施設をノイズと捉えるか、母数と捉えるか)
- 3) 施設-薬剤間交互作用の検討の必要性

施設当たりの症例数の問題は、これらの問題と共に議論されているが、これらの事項は相互に関係しているため、明確に区別されずに混同されて議論されることが多かったのではないだろうか。

1)については、施設数を増やすことが、実際の臨床治療の現場をより反映することになる、という論調が多い。しかし、単に治験参加施設数を増やすこと(逆にいうと施設当たりの症例数を減らすこと)でその目的が達成されるのだろうか。当該治験薬が、市販後には主に開業医で使用されることが予想されるならば、大病院の施設数を多くすることでは、実際の臨床治療に近い状態には近づかない。対象母集団に対する代表性については、施設数だけでなく、施設の医療実態についての議論も必要であろう。

2)の問題については、樁らの指摘のように、薬効差の検出に際して施設をノイズと捉えることでより頑健な結論を得ようとするのか、施設を要因と捉えてモデルに組み込み薬効差の検出力の向上を目指すのか、いずれの立場に立つべきかという問題であろう。

さて、施設-薬剤間交互作用の検討の意義について論じてみたい。施設-薬剤間交互作用は、施設による評価のばらつきを原因とする場合と、施設と交絡した患者特性や治療方針の違いを原因とする場合があらう。評価の均一性の問題は、解決のためには事前の評価者訓練や、良い評価方法の確立など、いくつかのアプローチがありうる。また、患者特性や治療方針の違いについては、それを要因とした分析をすることで、交互作用として捉えられることもある。しか

しながら、患者特性や治療方針の違いは、事前に十分には想定できないかもしれないし、そもそも解析評価すべき項目としてデータが入手されていないかもしれない。このような場合に、施設-薬剤間交互作用の検討をすることによって、施設と交絡した要因が発見される可能性もあるし、未発見の要因の存在が提起されるかもしれない。この意味において、施設-薬剤間交互作用の検討の価値を捉えてはどうだろうか。

例えば、抗生剤の臨床試験を考えよう。抗生剤の過去の治験においては、評価方法は確立され標準化が進んでいたとしよう。また、施設-薬剤間交互作用も過去において存在が報告されていなかったとしよう。このような場合でも、なお施設-薬剤間交互作用の検討の意義はあろう。なぜなら、特定の施設で、一方の治験薬に対する未知の耐性菌が高頻度に出現しているかもしれない、それが施設-薬剤間交互作用として捉えられるかもしれないからである。つまり、治験薬が、既存薬とは異なる新しい薬理作用を有している時には、一般的な患者特性や治療方法を要因とした分析に加えて、施設-薬剤間交互作用を検討する価値があるのではないだろうか。当該領域では、施設間の評価のばらつきが小さいことが知られていたとしても、施設-薬剤間交互作用の検討には、意義があるのではないかと考える。但し、少施設の試験で施設-薬剤間交互作用が見いだされなかったからといって、交互作用が存在しないとは結論できないし、逆に施設当たりの症例数が極めて少ない場合は、そもそも交互作用が検出できないことは当然である。したがって、積極的に交互作用の検討をする場合の症例数は、上記の1)、2)の課題とは別に論議が必要であろう。また、交互作用は開発経過の中のどの段階でチェックしておかなければならないか、についても議論の必要があろう。

さて、施設の症例数の問題は、上記の問題についての議論に加えて、以下の各条件を全く無視して画一的に定めることは、適切ではないだろう。即ち

- ・全体の試験規模
- ・期待できる群間差
- ・参加可能な施設数
- ・薬効評価上問題となる施設-薬剤の交互作用の大きさ
- ・全体の開発計画での他の試験からの情報の量

等の考慮が必要であろう。

従って、実際の臨床試験の計画に際する施設数や施設当たりの症例数の決定においては、最初に述べた三つの課題に対して如何なる態度を取るかを明らかにし、試験の種々の条件を考慮して議論する必要があるのではないかと考える。

最 後 に

吉村、椿らによって、統計的課題を切り口として、実に多くの問題が提起されている。それらの指摘を含めて、今後の新薬開発において解決しなければならない事項には、企業側の努力だけでは達成できないことも多い。

統計的テーマを切り口として、統計関連の学会や製薬企業の会合等で、臨床試験に関する議論がなされる機会も決して少なくない。その取り組みに加えて、今後は、医学、薬学、生物統計学を初めとした多方面の学識者、さらには規制当局側の参加も得た、広い公開の場での議論が望まれるのではないだろうか。また、医療の当事者である看護従事者や患者、さらには弁護士、倫理の専門家の参加も期待されるところである。その意味で、今回このような紙上討論の場が設けられたことは大変意義深い。これを先駆けとして、広く議論される場が多方面に設けられることを期待したい。また、その議論では、原則論の確認に留まらず、疾患や医薬品に固

有の問題が事例と共に議論される事を期待したい。原則の正しい理解と実行には、種々の具体的課題についての検討が有益と信じる。新薬開発に関係する多くの人が同じテーブルにつき、具体的な課題について議論する機会を持つことで、より適切な新薬開発の方法論が生み出されることを望みたい。

回 答

拙著「検証的比較臨床試験の計画において 考慮すべきこと」の補足

東京理科大学* 吉 村 功

(受付 1998年3月23日)

1. 検定の有意水準をめぐる ICH での論議

私は、頭記論文を、1997年7月段階の ICH 案文に基づいて書いた。ところがこの案文はその後で変更になった。最終的に確定したもの、つまり案文ではなく完成した文書としての「臨床試験のための統計原則」“Statistical Principles for Clinical Trials”では、私が問題にした文章が無くなっている。そこでここでは最初にその変更の経緯と理由を述べよう。

第4回の ICH 全体会議 (ICH4) は1997年7月に Brussels で行われた。Step 3 段階、つまり案文を日米欧3極の域内に公表して意見を聴取し、出された意見を取捨し最終案文を作るという段階、に達していた E9-EWG は、ここで最終案文作成の作業を行った。作業はかなり厳しかったが、EWG は会議の末日に一応の合意文書を作成した。しかしやはり時間に追われた仕事だったため、細部に不適當なところが残った。そこで委員各自は、気がついた間違いを帰国後すぐに責任者の O'Neill (米) 氏に送ることにした。同氏がこれらを入れて最終的な文章を作ることになった。この最終文章に問題がなければ、3極の責任者が署名をして Step 3 が終わることになる。

O'Neill 氏の案文が Lewis (欧) 氏を経由して日本に送られてきたのは、1998年1月10日過ぎであった。そこでは頭記論文の第3.2節で引用した文章とその前の一文、

“For superiority trials, the alternative hypothesis is inherently one-sided. The choice of type I error should be a consideration separate from the use of one-sided or two-sided procedure (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.)”

が削除されていた。代わりに、“V. DATA ANALYSIS CONSIDERATIONS, 5.5 Estimation, Confidence Intervals and Hypothesis Testing”の節に、

* 工学部：〒162-8601 東京都新宿区神楽坂1-3。

“It is important to clarify whether one- or two-sided tests of statistical significance will be used, and in particular to justify prospectively the use of one-sided test. … The issue of one-sided or two-sided approaches to inference is controversial and a diversity of views can be found in the statistical literature. The approach of setting type I errors for one-sided tests at half the conventional type I error used in two-sided tests is acceptable in regulatory settings. This promotes consistency with the two-sided confidence intervals which are generally appropriate for estimating the possible size of the difference between two treatments.”

という文章が挿入されていた。他の部分にも多少の変更があったが、それらは一般に明らかな改良だったので、大きな問題はこれだけだった。

ここでの変更は、「優越性試験は本質的に片側検定だ」と言うのをやめ、「片側か両側かは簡単には言えないことだ」と言い、「片側有意水準を両側有意水準の半分にするという慣例を認可当局は受け入れる」としたことである。第5.5節は解析の一般論を述べている節であるから、この内容は優越性試験であるか非劣性試験であるかに拘わらず適用されることになる。以前の案文で優越性試験の項にあったものをこのような一般原則にするのは、明らかに内容自体の変更であり、Brusselsでの合意から外れたものである。

日本の代表委員である私と佐藤俊哉氏は、厚生省の医薬品医療機器審査センターの統計担当者及び製薬協のE9-EWG委員の意見も聞いた上で、これに関する次の意見を欧米の委員に送った。

“We prefer to leave the original text agreed in Brussels. The current change has the different meaning from the agreed :

1. A sentence that “For superiority trials, the alternative hypothesis is inherently one-sided.” has gone. This sentence was supported by the most of EWG members, and it was this sentence which required the following explanations.
2. By moving to Section 5.5, this argument is now beyond the superiority trials. The MHW’s convention is as follows; the two-sided 5% level in superiority trials and the one-sided 5% level in non-inferiority trials. Under the current change, the MHW will have to change the convention on non-inferiority trials, and it will make great confusion.

ここでMHWというのは厚生省 (Ministry of Health and Welfare) のことである。この意見を出すときに特に配慮したのは、現行の日本の統計解析指針が、「たとえば」「目安としては」と断りながらも、非劣性試験で有意水準5%の片側検定を示しており、現実それで審査が進められていることである。もし有意水準を片側2.5%にしようというのなら、臨床医や行政担当者を含めて日本での合意を得なければならない。“in regulatory settings” というからには、統計担当委員だけでこのような方針を出すことはできない。

迅速な反応が欧米の委員から戻ってきた。要約すると、「欧米では優越性試験も非劣性試験も片側2.5%の有意水準で進めたい。しかしだからといって、日本の方針に制約を与えるということではない。それに配慮して、「片側か両側かは論議のつきないこと」という文章も入れたし、元の案文にあった5%という具体的な数値や、あちこちにあった、 α , $1-2\alpha$ という類の表現も除いた。だからこの文章で日本の行政当局が困ることは無いはずだ。」「欧州では今後、両側信頼区間で結果を評価することにしたい。そのときに有意水準を両側で設定しておく、統計解析が首尾一貫して混乱がなくなる。それをはっきりさせたいのでこの新しい文章そのままにしてほしい。それで日本が困ることはないように、解釈に自由度を持たせる表現にしてある。問

題になっている有意水準の選択に関しては日本人が納得できるもっとよい表現があるなら、それを出して欲しい。可能な限り取り入れる努力をする。」というものである。

Fax で文書の往復が繰り返された後で、私と佐藤俊哉氏は、欧州委員の希望にある程度の共感と理解を示して、“acceptable”を“preferable”に直すことに合意した。往復文書には日本が独自の方針を出すことに問題がないと明記されているので、国際的な誤解が生じたときには、それを公にすることで対処できると判断したのである。

このようなやりとりの上で最終文章は3極のE9-EWG委員によって署名され、Step 4、つまり政府間の承認の段階に移された。2月5日にはこれも署名に至り、「臨床試験のための統計的原則」は、国際的に通用する文書となった。次はStep 5、つまり国内措置である。国内措置としては、行政当局つまり厚生省が通達文書を作り、方針を全国に広めることになる。このとき発行される文章は、英文の単純な翻訳ではなく、合意英文文書の内容を入れた上で国内の各法規との整合性を持たせたものとなる予定である。おそらく、現行の統計解析指針は廃棄されるであろう。これはE9-EWG委員である私と佐藤俊哉氏の権限外である。委員としてのわれわれの仕事は終わりとなった。

2. 検定の有意水準についての日本での方針

頭記論文の第3.2節で、私は記述のミスをした。私が設計したのは実薬対照を用いた非劣性試験であって、プラセボ対照の優越性試験ではなかった。また日本の慣例では、プラセボ対照の優越性試験での評価を、片側2.5%の検定で行っているのも現実のようである。訂正をさせていただきたい。

「現実のようである」とあいまいに言ったのは、その根拠がどこにもないためである。現実には、新薬の申請者がそうしているのに対して、審査を担当する調査会委員が異議を唱えていないだけなのである。もし、申請者の誰かが片側5%の検定で申請を行い、それを調査会が駄目だと言って争いになったとき、何が根拠になって決着が付くかは全く不明である。

この件について、佐久間昭氏は「1967年頃に配合剤で有意水準を5%にしようという話が出たとき、そこでは片側か両側かなどということは全く意識になかった。もちろん、そういうことを取り決めたり、文章にししたりしたことは私の知る限り一度もない。」と述べている。また、広津千尋氏は「私が調査会の委員であったときは、優越性試験だから片側2.5%にするということは、まったく意識していなかった。だから統計ガイドラインで同等性試験（非劣性試験のこと）の手法を考えたときには片側5%で当然と考えていた。」と述べている。厚生省の審査センターの統計担当者、意見を聞いた製薬協の人たち、私より後に各調査会の統計担当委員になった人たち、これらの人の中で、新薬審査でのプラセボ対照優越性試験では片側2.5%の検定（信頼水準95%の信頼区間）を用いることが決まっていると教えられたり、あるいはそういう内容の記述文書を見たりした人は一人もいなかった。

慣例とは、一体何なのだろう。「FDAではそうしている」、「国際的にはそれが通常である」、ということで申請者がそうしていることの追認だろうか。

これについて佐藤俊哉氏は私に、一つの文献“Steven Piantadosi (1997). *Clinical Trials: A Methodologic Perspective*, John Wiley & Sons”を教えてくれた。この著者の考えは信頼区間間方式を原則とすることで、片側か両側かという見方をやめようということである。これは前節で述べた欧米の委員の考え方と共通である。これはこれで一つの方針である。現実の決定規則、判断方式としてはこれで問題はない。必要なことは、日本の規制当局の規準として優越性試験と非劣性試験で両側信頼区間の信頼水準を同じにするかどうか、それを90%にするか、95%にするかだけである。

非劣性試験で信頼水準を 95%にすると、これは現在の規準と変わり、申請者にとっては規準が厳しくなる。しかしそれは今までの規準が甘すぎたことを意味しているのかもしれない。幸いなことに、前節で述べた「臨床試験のための統計原則」では非劣性の同等範囲 (equivalence margin) について、たとえば有効率で 10% というような一律規準を示していない。とすれば信頼水準を 95%にとるという条件の下で、同等範囲を新たに決めることにすれば、従来の規準の一方的な厳格化という問題は生じなくなる。これは大至急議論しなければならないことである。

3. 椿・藤田両氏の論評にたいする見解

プラセボ対照がときに非倫理的であることは、E9-EWG の議論で日本側委員全員が強調したことである。それは欧米の委員も認めている。その議論はすべて E10-EWG に伝えられている。1997 年 3 月に成田で行われた E10-EWG では、疾患を具体的に挙げて、どの場合がどうかという議論を行った。その中で、日本の臨床医の側にかかなりの実薬対照志向の強いことが認められた、と私は考えている。

私が担当している第 1 調査会では、たとえば肝炎に対する治療が問題にされる。そこでは、有効性が証明されていないにもかかわらず、無処置が悪いにきまっているという理由で、双(二重)盲検をしないことが頻繁に行われている。また、試験の経過がよくないときに有効性が証明されていない治療を併用する例が非常に多い。椿・藤田両氏の論評の第 2 節で言っていることは、空論としか受け取れない。建前でどう宣言しているかではなく、現実がどうなっているかを事例に則して議論する必要がある。

有意水準のことは前節までに述べた。

非劣性試験については特に見解を述べる必要がないと思う。

多施設試験については、現在日本で行われている多施設試験が、大橋靖雄氏が指摘しているように、椿・藤田両氏が理想化して語るようなものになっていないことを注意したい。もし、200 という程度に多くの施設を全国から確率的に選び、患者もまた確率的に選び、解析も変量模型的に行うなら、一般化可能性についてのよい情報が得られる。それが本当にできるならした方がよいだろう。「臨床試験のための統計原則」はそういうことにも配慮していろいろな注意をしている。問題はそういうことが実際にはできないのが普通で、現実に行われている多施設試験は、およそそういうものではないことである。

対応のある試験についての記述は、私の主張に対して何を追加したかったのか理解できなかった。一般的に非劣性試験を対応のある試験として計画することは避けた方がよいというのは、私がいろんな機会に述べていることである。実際、有害事象が出たときにそれがどちらの原因によるかは一般に判断する方法がないし、有効性においても、存在する擾乱を過小評価しがちだからである。しかしこれは対応のある試験を絶対にしてはいけない、ということではない。対照薬の実績や薬理の詳細が確実に分かっている場合なら、適切な解析法をとることでそれを認めてよい場合があるように考えられる。私自身がそういう試験を設計したことはないが。

PC と ITT については、計量生物セミナーを含めて多くの機会に製薬企業の担当者と意見交換をしている。私の記述で特に問題はないと考えている。

4. 山口氏らの論評にたいする見解

引用されている有用性変数についての記述には、その前に、“Therefore it is not advisable to use a global usefulness variable as a primary variable.” という文章がついている。この記述の方が重要であろう。

臨床医が論理の詳細を分かっているとは私も考えていない。私が言いたかったことは、今の“ $\Delta=10\%$ ”という規準が、多くの場合、得られた単純な有効率において被験薬の方が優っているという規準に帰着する。それを臨床医が肯定しているように感じられるということである。

謝 辞

他の方々からの論評にも、もっと詳しい意見を述べたいところがないわけではないが、紙数の制約もあるので割愛させていただくことにして、拙稿にたいして、丁寧な目を通していただき、論評をしていただいた諸氏に心より感謝の意を表したい。

「誰がための臨床統計？ 我が国で実践された 「患者の立場」からの臨床評価の原則と統計的方法の役割」 へのコメントに対する意見

筑波大学大学院* 椿 広 計
国立公衆衛生院** 藤 田 利 治

(受付 1998 年 3 月 31 日)

1. ICH ガイドラインでの「検証的試験の繰り返し」は本当に誤解か?!

事の重大さに鑑み、何はさて置き、「複数の検証的臨床試験」の要求は誤解というコメントについて、反論する。全く失礼とは思われるが、本稿では敬称は省かせていただくことにする。「ICH E9 Guideline: Statistical Principles for Clinical Trials (1998)」を「ICH ガイドライン」、それを検討した「ICH E9 Expert Working Group」をEWGと呼ぶ。

吉村は、椿と藤田も委員を務める我が国のICH ガイドライン検討会の座長であり、日米欧三極交渉のEWGで大きな責任を担われてきた方である。また、佐々木らも製薬企業のメンバーとしてこの活動を熟知しているはずである。こうした立場の方から、「2回以上の検証的試験は要求されていない」と簡単に片づけたコメントは我々にとって大きな驚きであり、この点を明確にすることは急務と考える。大橋及び厚生省の山口からは、この点は誤解であるとの指摘はなされていない。

科学的検証的試験の強調と複数回の検証的試験の要望は、今後の我が国の臨床試験に大きな影響を与えるICH ガイドラインの大きなハイライトと認識している。この点に関して我々の論文が「誤解」を犯しているとの指摘であるので、ICH ガイドラインの英文解釈に紙面を費やさざるを得ない。本来は、最終合意案ステップ4文書で議論すべきであるが、吉村らが現在翻訳中であり、また関連する部分に大きな変化はないことから、一般公開されているステップ2文書とその翻訳に基づいて反論する。

確かに、吉村らが公表した日本語訳(吉村 他(1997))にはそのような記述が一切ない。し

* 経営システム科学専攻：〒112-0012 東京都文京区大塚3-29-1。

** 疫学部：〒108-8638 東京都港区白金台4-6-1。

かし、これは confirmatory trial (検証的試験) の単数形と複数形を訳し分けていないからであり、そのために最も重要な要望事項が隠れてしまっている。ICH ガイドライン原文において「検証的試験の繰り返し」の必要性に触れた文章は、「検証的試験」の意義を記述した 2.1.2 節にある。

2.1.2 Confirmatory Trial の第 1 パラグラフ、第 1 文と第 2 文を原文と吉村他訳で比較する。“A confirmatory trial is a controlled trial in which a hypothesis is stated in advance and evaluated. As a rule, confirmatory trials are necessary to provide firm evidence of efficacy and safety”となっているが、吉村らの訳は「検証的試験とは、事前に定められた仮説を評価するための比較試験である。一般に、検証的試験は有効性または安全性の確固たる証拠を提示するために必要である。」であり、第 1 文の単数形主語と第 2 文の複数形主語のニュアンスの違いを無視している。As a rule は normally よりは弱い、generally の公的な表現であり、この文章は複数の検証的試験の必要性を説いた文章と見なすこともできる。特に、2.1.2 節第 1 パラグラフで necessary to という要求事項を規定しているのは、この第 2 文だけである。残りの文は検証的試験の機能の定義であるから、この第 2 文こそ規格作成者の執筆意識の中では、第 1 パラグラフのハイライトと位置づけられる。

当然、この複数形は単に抽象概念を示したものと反論もありそうだが、更に、第 2 パラグラフでは “Confirmatory trials are intended to provide firm evidence in support of claims and …略… A justification of the design of each trial and of all other statistical aspects such as planned analysis, should be set out in the protocol. Each trial should address only a limited number of questions.” と続く。ここでも、each trial を「個々の臨床試験」と訳しているが、第 2 パラグラフにおける要求事項が存在する should を含む文に何故包括的な総称としての every でなく、一つ一つのといった意味の each が使われたかは、第 1 パラグラフ要求事項である一つの薬剤に対して検証的試験を複数回行う「原則」を受け、その中の「個々」を意味するからだと筆者達は解釈している。もっとも、我々も第 2 パラグラフまでは、極めて「玉虫色」の解釈が許容されることは認める。

しかし、第 3 パラグラフで事態は明瞭になる。吉村他訳では、2.1.2 節最後の文の翻訳として「場合によっては、一つの検証的試験からの証拠の重みだけで十分であることもありうる。」という、唐突な文章が出現し読者を混乱させるのである。吉村らの翻訳では、規格作成者が 2.1.2 節の締めくくりの 1 文として、“In some circumstances the weight of evidence from a single confirmatory trial may be sufficient.” を敢えて配置した必然性が、全く伝達されていない。この一文は、逆に「複数回の検証的試験などの要求は記述していない」という論拠とも安易には解釈されそうだが、第 3 パラグラフ自体の要求事項を記述している 2 番目の文には “The confirmatory trials should therefore be sufficient to answer each key clinical question relevant to the efficacy or safety claim clearly and definitively.” と記述されており、複数形 (confirmatory trials) を主語とした should が、使用されているのである。この第 3 パラグラフ第 2 文が、単数検証試験 (a single confirmatory trial) を主語とする最終文の may より強い要望を主張していること、言い換えれば、最後の文の「許容 (may)」を例外許容とする意図を含むことは明らかである。つまり、製薬企業がこの種の「許容」を規制当局に認めさせるためには、単一試験でも複数回の試験と同様の証拠力を持つことを立証する必要があるといえる。吉村らは、この第 3 パラグラフで唯一「要求事項」を記述した第 2 文を「したがって検証的試験は、有効性または安全性の主張に関連した個々の主要な臨床的問題に、はっきりと最終的に答えているべきである」と不適切に訳したため、最後の許容事項の文が唐突になったのである。ここで、The が confirmatory trials についてのも、第 1 及び第 2 パラグラフで定義したある新薬のための複数回繰り返し試験を指すからである。正しくは「したがって、複数の検証試験に

よって、有効性または安全性の主張に関連した個々の主要な臨床の問題に十分に答えられるようにすることが望ましい。」と訳すべきである(なお、後述する規則にしたがって、should は「望ましい」と訳す)。注意深く読めば、検証的試験について記述した ICH ガイドライン 2.1.2 節全体の規格化意図が複数回の検証的試験の必要性の記述であると理解できるはずであるが、日本語訳では全く複数形を無視したため、単に検証的試験の意義を示す教科書的記述になっている。

国際標準化活動では、どの文章が「概念定義」であり、どこが「要求・要望事項」かの交渉が中心となり、とりわけ国際規格で用いる助動詞 (shall, should, can, may) の選択と配置、名詞に単数形を用いるか複数形を用いるかを巡って審議が難航することが多いのである。翻訳上の問題に関する一般的注意については、厚生省の ICH 統計ガイドライン検討会で椿が再三指摘したところである。通常、ガイドライン原案作成は、非政府標準化機関による任意規格であっても、WTO/TBT 協定 (Agreement on Technical Barriers to Trade: 旧ガットスタンダードコード) を遵守する必要がある。この点は、横浜での ICH で FDA の立場を明らかにする講演で触れられていたのが印象的であった。新薬申請の場合、基本的には許認可は各国の裁量であるが、本文中でも触れた外部品質保証の対象となる「適合性評価 (conformity assessment)」については、TBT 協定の枠内と考えている。FDA のこの点に関する公式見解はそう単純なものではないが、TBT 協定を引用した FDA の Nightingale (1995) を参照されるとよからう。TBT 協定の枠組みの中では、国際指針で使用可能な助動詞及びその意味も ISO/IEC directive Part 3 で規定され、各国政府が責任を持ってそれをどのように翻訳するかについての directive を作成している。我が国では、昨年改訂された日本工業規格 JIS Z 8301「規格票の様式」の中に国際規格・ガイドの翻訳指針が示されており、これは各産業界に公開説明されている。また、ISO の翻訳に基づく日本工業規格原案は、この種の指針に基づく「規格調整委員会」のレビューを受ける。規則や英語の厳密訳に配慮しない訳案は、大抵大幅修正要求が出される。この種の国際ガイドラインの翻訳作業は日本にとって神経を使う困難なものであり、日本工業規格原案でさえもレビュー前には誤訳が頻発しており、今回の ICH ガイドラインのようなことも例外とはいえない。なお、我々の論文にも確かに書きすぎはある。WTO の枠組みでは、強制力のある「要求事項」、すなわち shall 要件を規定できるのは、国際標準化機構の専門委員会だけであり、すべての ICH ガイドには、本来「要求事項」は存在しないのである。should や necessary to は、「要望事項」と呼ぶのが正しい。勿論、ICH ガイドラインでの最も強い要求であっても規則上は「要望事項」と呼べるに過ぎない、ということである。

さて、英文解釈を離れ本質的な議論に入りたい。先ず、ICH ガイドラインには「『科学的』少施設試験を繰り返せ」という記述はないという反論もあり得よう。確かに、その通りである。しかし、ICH ガイドラインでは「科学的」試験による「科学的」有効性検証が強調されており、山口らのコメントもこの点を認めている。また、少施設試験を強調したのは、審査側で影響力のある方が、我が国の新薬申請のための検証的試験を 1 施設 20 例程度の試験に限ると公言される向きがあるからである。「こうした「科学的」少施設試験の 2 回以上の繰り返しを要求」という我々の記述を、全く誤解だとする認識は我々には理解できない。

そもそも少数施設で検証的試験を繰り返すことは FDA の一貫した要求事項であり、これを国際間でどう整合させるかは臨床試験の方法論に興味のある者には長年の大きな関心事であった。ICH ガイドラインの製薬会社側の日本代表もこの点を熟知しているベテランであったはずである。また実際、EWG で検証的試験の複数回実施が検討された跡が明確に残っている。ステップ 2 に上がる前の段階での 1996.10.14 付けの原案における前述した 2.1.2 節の最後の「In some circumstances」で始まる単一の検証的試験を許容しているかの文章は、元々、複数回の臨床試験の繰り返しを主題とする次のようなものであった。「If the main hypothesis being tested has not been firmly established prior to the confirmatory stage of investigation, it may

be necessary to do more than one confirmatory trial to demonstrate robustness and repeatability.”筆者達は、実際にこの交渉に参加された方から、『この項目を necessary to』とする FDA の主張に対して、『全ての場合に複数回の検証的試験の実施を要求されては困る』との EU 側の意見 (may を付加したい) により、妥協の産物として、特殊な状況に限っては不完全ながら許容する場合もあるという趣旨に弱められて落ち着いた。しかし、FDA はほとんど認めないであろう」と聞いている。ただし、日本の交渉責任者である吉村が否定する以上、欧米の解釈と日本の解釈との併存を政治的にねらった「玉虫色のガイドライン案」に収束したということなのであろうか。「1つの『科学的』少施設試験のみでは危険」という、米国の方法論の枠組みの中ではあるが、経験に裏打ちされた FDA 側の複数検証的試験の強い要求に対して、日本では単一試験を許容とするローカル・ルールが国際的に認められたとすれば、それは製薬会社の経済的で容易な新薬の国内開発を推進する意向を踏まえてのことか、あるいは、外国からの新薬の安易な導入を図るためなのだろうか。吉村らのコメントを好意的に解釈すれば、各国で1回ずつコアとなる検証的試験を行えば、三極合わせて複数回の試験が可能という国際共同開発を意識したのかもしれない。

いずれにせよ、ICH ガイドラインのコアパートである 2.1.2 節の複数形を単数形に変更する「日本独自解釈」を安易に主張すべきではない。検証的試験の複数回実施については、EWG の日本メンバー内で確認するとともに、米欧のメンバーとの認識に差がないか、是非確認されたい。なお、検証的試験の複数回実施は、共通認識だと勝手に思い込み、この点を具体的に指摘せず曖昧な翻訳を放置したことは、椿、藤田も ICH 統計ガイドライン検討会メンバーである以上、共同責任といえる。ICH に関わる日本の新ガイドラインの公式発行に当たっては、このような問題が再発しないように監視を強める必要性を痛感した。

2. 立場や観点の違いからくる意見の違いについて

さて、以上に比べれば、他のコメントについて意見を述べることは、その多くが立場の違いによるものであり気が楽である。

今回の論文はオーガナイザーから椿、藤田に「日本の臨床評価について書きたいだけ書いて良い、全く別の立場で別の人が書くから」という依頼により執筆を引き受けた。吉村論文への我々のコメントでも述べたように、臨床試験形骸化の歴史の中で説明的試験（「科学的」実験）による薬の有効性の「科学的」検証が不十分であったことは大きな問題であり、ICH ガイドラインの方向に我々も基本的には賛成なのである。新薬承認の証拠として「科学的」試験が省略できるとは毛頭思っていないし、その必要性を痛感している。この点は論文で「第三相試験として、『科学的』仮説検証を目的とした説明的試験が実施された後に、技術評価としての実践的試験が実施されることが強く望まれる。」と指摘した通りである。今回の論文では、「科学的」試験推進の議論は、吉村及び予定されていた大橋の両論文の使命と割り切って、ICH ガイドラインの発効によってむしろ弱まる可能性が懸念される実践的試験を含む、我が国臨床家によって組織的・具体的に推進された臨床試験原則の再確認に重点を置くことにしたのである。そして、「情報の偏りの防止」、「使用者指向」、「標準化」の3原則に特徴を総括し、その意義と我が国臨床試験の歴史をこの視点から述べた。

コメントの中で我々のバックグラウンドにも若干触れられているが、我々は偏りの防止をはかり公平性確保を第三者として監視するコントローラーの責務を 1970 年代後半から身近にしており、この責務を支援する立場として臨床試験との関わりを持った。また、中央薬事審議会新薬調査会において、椿が 1995 年まで 5 年間、藤田は 1986 年以来現在に至るまで 12 年間に渡って、新薬の承認審査に当たってきた。この点では「なるべく良い医薬品のみを早く世に出

したいと希望している第三者」ではなく、公平性を担保する役割を担う第三者の観点、及び責任のある審査を行ないメーカー（第一者）と対峙する第二者ともいえる観点に立脚している。新薬開発の臨床試験に会社側のパートナーとして参画し、大きな貢献を果たしつつある大橋と観点到違があるのは、むしろ当然といえる。「論文を一読して絶句しました。前提に対する認識そしてアプローチにおいて私とは著しく異なる立場」、「困惑」という感想は、我々にはよく理解できるところである。おそらく吉村も大橋と類似の感じを持ち、そのために理解しようとする努力が損なわれたのであろう。

本論文執筆に当たって1970年代以前の歴史的状況については、椿・藤田では把握できない部分が多いことから、共著者に佐藤を加えて当時の歴史的経緯を確認した。今回は、薬効懇答申という基本原則確立以降に限って述べたが、こうした「臨床試験の原則」は、大橋のいうところの「臨床試験黎明期」に当たる1970年代半ばまでに広く普及したものである。コントローラー委員会が、第三者の立場から管理した臨床試験は、磁気化した再解析可能なデータとして、製薬会社とは独立に保管しているものだけでも800件程度にのぼる。また、1974年から1992年までに管理したプラセボ対照試験は、125試験である。確かに最も多い時期であっても、新薬承認申請の半数程度に過ぎなかったと思われるが、主要な第三相臨床試験の多くに関わっており、新薬承認申請での割合以上に影響は大きかったものといえる。また、直接関与していない他の多くの試験でもその形式的模倣がなされており、この点は、当時の臨床試験報告を実際に見比べれば確認できよう。委員会と立場を異にする専門家でも（例えば、本文中に参照した長谷川他論文）、我が国臨床試験の独特のスタイル確立に果たした役割は、そのスタイルに対する賛否はあるものの異論はないはずである。実際、立場の異なる大橋もコメントの中で、その貢献を認めているところである。吉村は「鵜呑みにできない」とのことであるが、鵜呑みにする必要はなく、具体的に資料に当たり確認していただきたいのである。ちなみに国際的にも、適正な新薬審査を目指して、臨床家自身が行政を巻き込んで推進した系統的・組織的・長期的な臨床試験の品質保証への取り組みを我々は知らない。この点についても、例えばFDAとの1982年以来的の交流や討論会資料、例えばいささか社交辞令的で気が引けるが、Temple (1992) の「高水準の医薬品開発を目指した日本独自の存在、『コントローラー委員会』」という発言などでも確認できる。

ICH GCP (Good Clinical Practice) 以前は、日本の臨床試験の管理は臨床家なしでは考えられないというのが、我が国では常識となっていた。ところが、共著者の佐藤が佐久間昭氏と共に、臨床の立場から新薬調査会に参加した1960年代には、中央薬事審議会は、薬学研究者が多くを占め、会長は製薬会社のトップが務めていた。このような状況の中で、佐藤は新薬臨床試験における二重盲検実施を指示事項として要求し、メーカーの拒絶と中央薬事審議会の冷淡な扱いを受けながらも、これを実施させた。さらにこの問題に関連して、佐藤が国会参考人質疑において二重盲検が科学的臨床試験に必要であることを主張し、更に日本学術会議勧告を取りまとめるなどの活動を行った。こうしたことも一因になって、70年代前半に日本独自の臨床家主導の新薬評価体制が築かれたのである。これらの歴史的な事実は、我々の参照した文献のみならず、広く一般報道でも取り上げられたことである。ただし、今回の論文に対するコメントを執筆する期間は極めて短く、吉村が具体的検討をする余裕がなかったであろう事情は斟酌する。

大橋のコメントの中で「企業性悪説に立ち第三者制度を価値付けています。あまりに後ろ向きではないでしょうか。」は、立場の違いを明確に示している。

臨床試験では不完全例が発生して解析から除外されることがある。これに関連して承認審査の中で少なからず経験することとして、解析採用例では有効性に関わるエンドポイントで有意差が見られないにも拘わらず、解析除外例のみの解析では大きな差が見られることがある。し

かも、ほとんど全ての事例が、新薬に有利な方向であることを「都合のよい偶然」の集積として、性善説から説明できるのだろうか。ある新薬の異なる適応症に対する複数のプラセボ対照試験では、ケースカードの訂正が多数発生していた。新薬とプラセボでは逆方向の訂正がなされており、それが「功」を奏して、見かけ上、プラセボに有意に優る成績となっていた。一つの新薬の承認申請資料でこうした事態が重なることは、「好都合な偶然」であろうか。大橋は現在の臨床試験に纏わる歪みの指摘において「温厚で人格者であられる」臨床家と表現しているが、藤田も皮肉を込めて「人格者」という言葉をよく使う。症例固定のための症例検討会に出席もせず、「症例の扱いが妥当である」などと保証できる「人格者コントローラーの第三者監査」を誰が信用できるのか。最近では公正性に大きな疑問のあるものについては、藤田は治験総括医師（新 GCP での治験調整医師）、コントローラー、さらには世話人会のメンバーなどからの文書を要求することにしている。こうした事例を固有名詞で指摘することは憚られるが、決して現在でも少ない訳ではない。そのいくつかを既に匿名で類型化して報告している（藤田（1995））。現状の書類審査の限界の中で、申請書類を額面通りに受け取ることが審査側の責任を果たすこととは考えられず、我々の論文で述べた姿勢で対応せざるを得ないのである。我が国に比して職能倫理が成熟していると考えられる米国においても、FDA が新薬の承認審査を企業性善説に基づいて行なっているとは思われない。

また、審査側は一定の審査基準を意識しながら、領域別の新薬開発を「集団」として認識せざるを得ないことが必要であることは論文で触れた通りである。勿論、大橋の指摘通り me-too-drug の氾濫があろうとなかろうと、大橋が「違和感」を感じたという long run property の確保、すなわち、審査システムの統計的挙動自体も、本質的な関心事である。勿論、類似新薬の承認数を制限したり、経済的に採算が取れないような更に抜本的な対策を講じれば、審査側の負担は大幅に軽減される。この方向は、ICH と同じく、資金力・開発力のある製薬企業の生き残り選別を意味し、大橋の望む 3-5 社への寡占化を誘導することになるかもしれない。

今後の重要な論点は、承認審査側と申請側とが、従来のように一定の対立的緊張関係を保持しつつ互いの職分を尽くすか、あるいは、協調して両者に共通の規範を作るかにある。我々は後者の活動には懐疑的であるが、吉村が ICH の活動を中心に良心的に後者の方向を模索していることは評価している。いずれにしても、異なる立場があること自体は健全である。

3. 公正性確保について

我々は、「アドバイザー」と正当な「コントローラー」は全く違うと心得ている。

大橋は、「個人」としてのコントローラーではデータ収集・点検と統計解析の実施は不可能であり、品質保証もおぼつかないことを告白している。佐々木らも、「多岐に渡る業務が「コントローラー」の名の下に特定の個人に対して過度に集中してしまったことも否めない事実」とし、「種々の役割を一手に担うことにも限界があるかもしれない」と認めている。

我々が引用した内藤論文は、厚生省の「新医薬品の評価に関する一般指針」（平成 4 年 6 月 29 日）に発展し、その中で「コントローラーの役割は、その試験が安全に、偏りなく公正に行われ、その成績がそのまま公表されるように治験総括医師、治験担当者及び治験依頼者から独立した立場から試験を管理し、それらを保証すること」と規定されている。臨床試験の公正性の確保・保証は、「個人」ではほとんど不可能なのであり、一方、コントローラーという我が国独自の制度は「組織」による適正な運営を要する多量の責務をコントローラーに科してきたのである。

しかし、実際には、個人では不可能な責務と承知しながら、製薬企業は制御可能な「個人」を選択し、また保証できない責務を安受けする「人格者」コントローラーが輩出した。これら

が適正な第三者監視の進展を阻害し、良貨を駆逐した「貢献」は誠に大きいといわざるを得ない。

今後について、大橋が「社内監査に代表される内部品質監査に頼らざるを得ない」と消極的なコメントであったのに対して、佐々木らは「製薬企業の責任」について「自己完結的な企業責任だけを意味するのではなく、広く社会的責任を担っていると考えるべきだろう。その意味で、『第三者による監視』のシステムを、従来にも増して積極的に導入することも、企業としての『責任』の取り方の一つ」と述べている。これは外部品質保証に関する良識ある見解と評価する。しかし、コントローラーにしろ独立データモニタリング委員会にしろ、それを支援して品質保証を全うする第三者「組織」の裏打ちがない場合、「適正なコントローラー」制度を破壊した轍を踏むことになる。また、いま一つの方策は、やはり佐々木らが指摘するごとく情報公開である。社会的な合意を形成しつつ、個人データを含む公開可能な情報を提示・利用可能にしていくことは、公正性の確保とともに科学的医療の進展に寄与すると思われる。

なお、念のために断っておくが、第三者組織は決してコントローラー委員会のみの専売特許ではなく、他の「組織」であっても一向に差し支えない。この点も、前述した日本学術会議から内閣総理大臣への「医薬品の臨床試験評価に関する体制の確立について」（1972年）の勧告や佐藤倚男をはじめとするコントローラー委員会メンバーの言動が示す歴史的事実である。

4. 有用性あるいは総合評価などについて

有用性評価に関して、次のように論点を整理する。

- ① 有効性と安全性などを総括した上位概念（有用性）を、患者ごとに計測することが必要か？
- ② 有用性評価はプライマリー・エンドポイントか？
- ③ 有用性評価に限らず、担当医師の総合評価は必要か？
- ④ 担当医師の総合評価はプライマリー・エンドポイントか？

上記の論点についての我々の見解は、有用性評価がこれまで「事実上のエンドポイント」に使われてきたこと及びその信頼性に問題がある場合があった歴史を認識した上で、①、③はYes、②、④はNoである。

山口ら及び佐々木らのコメントと我々の見解との差は大きなものではない。佐々木らの中で指摘している「組合せでの半ば機械的な評価」は、我々の意図してきた有用性評価ではない。吉村はやはり論文の意図を「読み取れない」としながら、我々の見解を「『有用性』という主観的1次元変数を主要変数として薬剤の検証評価をして良い」と曲解しているが、これも無茶である。また、我々の論文で示した臨床試験の歴史の中では、有用性評価やその他の総合評価のみが一人歩きしたことはなく、臨床検査や評価尺度の成績も常に提示され、照合可能であったことも無視してはならない。なお、他の一部の雑誌において臨床試験成績の不十分な情報提示しかなされない例は、確かに存在することは指摘しておきたい。

大橋のコメントは具体的で貴重な点も多いが、やはりすれ違いも多い。大橋は医師の評価する「有用性」と患者の「ベネフィットの評価」を違う概念と捉えているようであるが、我々は「仮想的に同一病状の患者に対して治療法を再び行う状況に置かれたとき、患者の視点に立つてどの程度今回の治療法を繰り返すモチベーションが生じているかを評価するのが有用性評価」としており、異なるものとは捉えていない。また、歴史的に有用性評価が重視されてきた背景として、「臨床試験の計画段階では予期しなかった事象・作用を評価する必要性、質的に異なる作用を総合的に評価する必要性、臨床試験が実際の治療に比べて短期間であることや中止・脱

落が不可避免的に発生するため医師の臨床経験から評価を補う必要性」などを指摘したが、この点も噛み合っていない。次に、「すれ違い」であるが、大橋のコメントに対する意見を簡単に述べる。

大橋は、多発性硬化症や筋萎縮性側索硬化症などでの信頼性の高い機能評価スケールの必要性をコメントしている。それはその通りであるが、椿（1995）が日本医学会第100回シンポジウムで指摘したように、患者の利益確保に関わる臨床的な総合評価とのトレーサビリティ確保が偏りの混入に対する保護（外的妥当性の確保）のために、更に必要と考えられる。

有効性 efficacy の確固たる根拠は、市販前にほとんど確立しているべきであり、更に実際の医療現場でその有効性に大きな変動はないか、安全性に大きな問題はないかを点検する技術評価が必要と考える。無論、不適切な使用や質の悪い医療は論外であり、それを社会的に許容しない態度は当然である。実践的試験が質の悪い医療を追認するものと考えているなら、大きな誤解である。

限られた条件下での有効性 efficacy の確固たる根拠のみでの市販許可は、それに続く実際の医療現場を借りての実験を意味し、患者に大きな不利益をもたらす可能性がある。市販前の臨床試験の規模で確認可能な大きさの問題は、市販前に評価すべきであり、問題を先送りして適切な使用を行なう医療現場の臨床家に責任を転嫁すべきでない。無論、市販前にどこまで確認すべきかは領域によって異なる。

Surrogate endpoint を有効性評価の指標として承認する薬効群については true endpoint と関連する効果 effectiveness の検証は市販後臨床試験に俟つところであり、指摘の通りである。市販前での有用性評価及びプライマリー・エンドポイントとされた客観的測定値などの臨床的意義は effectiveness との関連から本来評価されるべきであるが、我が国には大規模市販後臨床試験の成功例はないことから、真の臨床的意義についての有用性評価と客観的測定値との比較についての情報は不十分である。なお、大橋が「実践的試験は true endpoint を対象とした市販後臨床試験の形容」とするのは興味深いが、実践的試験 (pragmatic trial) を説明的試験 (explanatory trial) と対照させた Schwartz and Lellouch (1967) 及び、我が国でこの考えが普及のきっかけとなった1987年の生存科学研究所シンポジウムにおけるハーバード大学 Zelen の講演及びそれを採用した日本の統計解析ガイドラインの定義とは、全く異質のものであることは一応指摘しておく。定説では、浜島 (1993) にも記述されているように、本論で科学的試験と呼んだ説明的試験とは、「治療効果が端的に解明できるようにデザインをくむべきだと言う立場から行われる臨床試験」、本論で技術評価とした実践的試験とは「臨床の場でどのように判断をすればよいかという点に注目して研究をデザインした臨床試験」のことである。

5. 多施設共同試験について

日本の超多施設試験は「統計的」に問題であるという理由で批判あるいは自己批判されることが多く、臨床家や実務家の一部が施設数の少ない試験が「誤差変動を精確に評価できる正しい実験」と誤解する傾向がある。これに対して、大橋が「超多施設試験、極端には1施設1例であっても無作為化が適切になされれば偏りは存在せず、この意味で試験は妥当」とコメントをしたのは、統計的には全く正当なものとする。これは、これまで少施設試験に誘導する政治手段に「統計学」が誤って流用されていることが多かったのとは、全く異なる指摘である。また、「臨床試験を取り巻く環境の歪みが超多施設少数例試験を問題のあるものになっている」という趣旨の指摘の多くも肯けるところである。ただし、一つの施設で臨床試験が併行進捗したときの患者選択の問題は、無作為化「比較」試験では検出力を落とす可能性はあっても、審査側の最大の関心であるバイアスの問題にはならない。また、検出力（生産者危険）の問題は当

然存在するが、層別解析を行なうか、大橋のコメントにある方法か、類似領域試験の第三者集中管理などによってある程度回復する。いずれにせよ、生産者側が主張したい仮説の検証に不利な状況は、生産者とそのパートナーにそれを解消する強い動機を自ずから与えるものである。

さて、多施設試験の考察を統計的行なうときには、評価の施設間差 (reproducibility variance) の推定と施設・薬剤間交互作用の検討との両面がある。また、交互作用検出の意義として、特定の外れ施設摘出(施設毎の評価)、施設に関わる臨床評価構造の差違の示唆、ないし治療予後因子の探索が挙げられる。これらの点に関わる佐々木らのコメントは、大変有用である。

ちなみに、同一サンプルサイズならば施設数が多い方が良い試験という立場は、論文中でも紹介した統計的方法の適用に関する国際規格 ISO 5725 (第1部から第6部までで構成される) で主張されているが、これは既に翻訳され日本工業規格 JIS 8402 となることが内定している。誤差の施設間分散成分の推定精度向上のためには、施設数が多い程良いのである。また、ICH でも分析法バリデーションを検討するグループは ISO 5725 を規格文書として参照している。しかし、臨床評価を計測行為とは考えずに適用範囲外と見なすせいか、臨床分野ではその理念も含めてほとんど参照されていない。この国際規格には、誤差分散推定量の精度確保から、7, 8 施設以下の共同試験の非推奨 (should not) があり、それ以外にも山口らが危惧した施設がランダムサンプリングできないことの配慮 (物理計測・化学分析の共同試験でも施設をランダムに抽出可能なことは殆どない)、外れ施設の検出手順、施設の参加資格、品質管理などさまざまな項目についての原則が規定されている。

一方、交互作用解析については、同一標本サイズの下では、ISO 5725 のように施設数が多い共同試験がよいなどと、簡単に割り切ることはできない。論文で紹介した Ojima 他論文が、脳循環代謝改善薬の領域において「超」多施設共同試験から複数回検出した交互作用を臨床に必要な情報と考えるか否かが評価の分かれ道のような場合である。すなわち、臨床評価の施設間分散自体が群間で差があることが予想されるような場合である。このような本質的臨床評価構造や治療予後因子に関わる交互作用情報は、施設間分散成分の推定精度が低い少施設集団での試験では検出される可能性がむしろ低いし、優秀な施設の一枚岩の試験ではほとんど検出の可能性がない。

同じ規模の標本サイズを前提とした場合、共同試験についての施設数に関する我々の見解を単純化していえば、科学的仮説は、施設内再現性を基準に検証可能なのだから、「科学的」試験においては施設数が少ないことが許容される。一方、実践的試験においては薬効評価の施設間再現性改善のために、多施設の参加が望ましいというものである。

さて、本筋ではないが、生物統計学のパイオニア P. Armitage が、日本の統計学の代表的業績 (田口、赤池、甘利の業績をその独創性から評価した) と評価した田口の業績に関係する批評には不正確な点がある。タグチメソッドは、「存在する製品」の研究手段ではない。1970 年代には新製品開発 (目的機能の研究) のためのツール、1980 年代以降は新技術開発 (基本機能の研究) のためのツールと位置づけられている。二値データの実験誤差を二項分布の標本誤差で評価する無意味さについても、田口はタグチメソッド確立以前から臨床試験データ解析の枠組みの中で指摘していた (増山 他 (1974))。一方、数理統計学的にこの種のパラメータ設計を行なう手順は、G.E.P. Box の学派が一般化 SN 比を初めとして既に実現しているし、Nelder らが一般線形模型の枠組みでの推論法を提唱している。ただし、これらの研究の大半は、Technometrics 誌上で統計学者が、タグチメソッドの統計的チューニングのために行なったものである。大橋の批判にあるように、田口自身は検出力設計に限らず統計的推測の実用性については常に疑問視している (田口・矢野 (1996))。「繰り返しの原則」を否定し、代替的に「誤差因子の導入による直積実験」を推奨するのが、タグチメソッドの基本原則の一つだからである。したがって、我々も大橋が危惧するように工業分野の品質工学的手法自体を単純に移植すれば

良いと考えている訳ではない。例えば、椿（1995）が日本医学会で提言したのは、医学、理工学境界領域の検討課題としての次のような議論である：「工業実験では市場以上の極端な条件での加速試験も行われるが、臨床試験では患者を対象とする以上、日常診療で生じ得る状況を考慮にいった計画を立てる以上の追求は危険である。このように臨床試験での技術評価は工業実験と異なり種々の制約のため系統的な方法論が未成熟であり、国際的議論が必要と考えられる。」

6. 品質保証について

吉村は品質管理（正確には品質マネジメント）との類推を叱責するコメントを出しているが、そもそも ICH の統計解析ガイドライン制定の背景にある臨床試験実施に関する指針、ICH GCP が ISO 9000 シリーズの品質マネジメントシステムに準拠した記述を持つことを指摘しておきたい。現在、マネジメントシステム分野の専門家において国際合意されているマネジメントシステムには、通産省工業技術院管理システム規格課の矢野友三郎氏によると、「品質マネジメントシステム」、「環境マネジメントシステム」、「労働安全衛生マネジメントシステム」、「個人情報マネジメントシステム」、「リスクマネジメントシステム」があり、またマネジメントシステムのコアパートとその外部品質保証については compatibility があるものと認識されている。既に最初の 2 つは ISO 9000, 14000 シリーズとして国際規格化されていることは、一般新聞紙上でも取り上げられている。GCP については、論文で紹介したように既に品質マネジメント視点が導入されており、各国で実際に活動が開始されている。DIA (Drug Information Association) の生物統計学年会（東京）のプログラム委員会において、一昨年 EWG のメンバーでもある FDA の O'Neil が「品質保証からの影響」というシンポジウムテーマ案を提出したのもその表れである。また、FDA 上層部が本論で取り上げた「適合性評価」の概念を ICH において意識していることは、Nightingale (1995) から明らかである。我が国の製薬業界も、日本科学技術連盟川村数増氏（JAB 登録品質マネジメントシステム主任審査員）のアドバイスを基に GCP における ISO 9000 流の品質マネジメントの研究を組織的に実施し、研究報告書を論文として次々に公表している（原 他 (1996 a, 1996 b), 石田 他 (1997)）。また、石田 他 (1997) では、「ICH GCP を理解するためには、これまでの日本の GCP にはあまり馴染みがなかった欧米における品質保証の考え方を理解することが不可欠であり、そのためには ISO 9000 を理解することは有益であることから、…」と指摘している。また、「既報と同様に、ISO 9000 でいうところの最終製品を治験の総括報告書、供給者を治験依頼者（メーカー）、顧客を規制当局と設定した上で検討を行った。」との記述もあり、ISO 9000 シリーズについて理解があれば当然のことだが、我々と全く同一のフレーム設定をしている。このように、品質管理（マネジメント）の専門家ですら本質的意見対立を生む理念を、我々が単なる類推で GCP 分野に提起したとは思っていない。むしろ、これまで GCP の国際交渉を品質マネジメントの専門家にマネジメントさせなかった、我が国の現状の方が異常なのである。

また、ISO 9000 シリーズの原案作成を行なっている ISO 第 176 専門委員会第 3 小委員会では、品質マネジメントのツールに対するガイドを構築する作業を開始している。不幸にも、この活動に我が国は専門委員を派遣していないが、この中では、統計的方法のみならず、論文で触れた水野、赤尾らの QFD も要素として取り上げられている。論文で提起した GCP の個々の材料のどの部分に具体的に品質マネジメント観点で合意がないと決め付けているのだろうか。「要求品質」という用語にしても、上記製薬業界の研究レポートにすら登場しているのである。

我々が専門家間で一致を見ないと考えているのは、むしろ GCP における品質システムをどのように構築するか具体論、実際に GCP にどのように ISO 9000 流のマネジメントを適用す

るかである。これは上記の製薬企業のレポートにも指摘されている。椿は、1996年にJAB(財団法人日本適合性認定協会)のメンバーとICH GCPについて議論する機会を持ったが、JABの担当者もGCPを一読して、具体的な品質システムに関する記述がないことを指摘し、このままではシステム監査ができないことを心配されていた。コントロール委員会自体が行なってきた「試験管理システム」自体は、製薬会社の依頼を受けるか否かから始まり、論文投稿に至るまで、例えば本論で参照した佐藤(1992)にフローチャートの形で公開しているので、新たな品質システム構築の一つの参考にして頂きたいと思う。

なお、製薬業界のこの問題に関する研究報告については、第一に品質方針の矮小化、次に、総括報告書をプロダクトとする代わりに臨床試験データ自身をプロダクトとする方がより精緻な品質システムを作れるのではないかなど気になる点もあるのだが、第一印象の域を脱せず、顧客側としての独自の詳細な批判検討を完了していないので、今回はこれ以上の意見は述べない。

最後に、このような貴重な論争の機会を与えて下さった統計数理研究所の佐藤俊哉先生に謝意を表します。

著者の一人の椿は、吉村先生、大橋先生の研究室の後輩です。討論の中でも述べましたが、立場が違う中で互いの最善を尽くそうとするから論争になるのであって、両先生についてはその人間性、能力及び臨床分野及び応用統計分野での真摯な活動と業績について最大の敬意を有しています。「確率論とか理論統計学をやる人は沢山いるわけだが、統計学者の中で、実際の研究に手を汚してもいい、医学と協力してもいいという人は少ない。日本では、特に統計学といった方向が必要なのである。もっと統計学を勉強した人で応用に行ってくれるための職業を用意しなければならない。そういう人に是非頑張ってもらいたい。」と呼びかけたのは、残念ながらロンドン大学のPocock教授(1992, 臨床評価, 20, Suppl. IV, p. 113.)ですが、まさにそれを実践しているのが両先生や今回誌上討論に参加して下さいの方々だと思います。私たちは、不遜に聞こえるかもしれませんが、むしろ今回の討論を観戦された我が国の多くの統計関係者の皆さん方に対して、御自身の統計学について、折りに触れて「誰がための」という自問をして頂きたいと感じています。

参 考 文 献

- 藤田利治(1995).『我が国の薬効評価を考えるII, 薬効評価(日本公定書協会 編)』, 29-54, 薬事日報社, 東京.
- 浜島信之(1993).『無作為割付臨床試験』, 癌と化学療法社, 東京.
- 原信次 他(1996 a). ISO 9000のGCPへの適用, 薬理と治療, 24, 511-520.
- 原信次 他(1996 b). ISO 10011に沿った治験の品質監査, 薬理と治療, 24, 1211-1223.
- 石田信幸 他(1997). ISO 9000のGCPへの適用(第3報) 治験依頼者の組織・体制のあり方, 薬理と治療, 25, 2585-2596.
- 増山元三郎, 奥野忠一, 田口玄一, 竹内啓, 広津千尋(1974).『実験計画法とその発展と最近の話題』, 東京大学出版会, 東京.
- Nightingale, S. L. (1995). Challenges in international harmonization, *Drug Information Journal*, 29, 1-9.
- Schwartz, D. and Lellouch, J. (1967). Explanatory and pragmatic attitudes in therapeutical trials, *Journal of Chronic Diseases*, 20, 637-648.
- 田口玄一, 矢野 宏(1996).『技術開発のマネジメント——技術革新を促進するタグチメソッド』, 日本規格協会, 東京.
- Temple, R. (1992). 医薬品承認——アメリカの考え方, 臨床評価, 20, Suppl. IV, 85-93.
- 椿 広計(1995). 新薬評価における統計的方法のあり方, 第100回日本医学会シンポジウム記録集「新薬と評価」, 61-66.
- 吉村 功, 魚井 徹, 佐藤俊哉, 上坂浩之(1997). ICH E9ステップ2ガイドライン 臨床試験のための

統計的原則 ICH E9 Step 2 Guideline: Statistical Principle for Clinical Trials, 薬理と治療, **25**,
Suppl. 4, S869-S962.