日本統計学会誌
第37巻，第1号，2007年9月
37頁～54頁

統計科学と医・薬・生物・健康科学との相互寄与を未来に向けて思量する[1]

吉　村　　功*

An Overview for Future Collaboration of Statistical Science and Medical, Pharmacological, Biological and Health-related Sciences

Isao Yoshimura*

健康科学と統計科学は歴史的に相互の発展に寄与してきた．近年はその様相が多少変化してきて，健康科学から統計科学に提供される研究課題が以前より多種多様多量になっている．治療と薬剤開発のための臨床試験では非定型的な試験計画法，遺伝子解析では超多変量小標本での感受性遺伝子の検出法，創薬ではインシリコ試験の活用法，等々が求められている．計算機技術とマイクロテクノロジー等のハードウエア的新技術の開発が進み，新しい型のデータが提供され，生命への新しい切り口が必要になってきたからであろう．

これらの要求に応えることで，統計科学は，従来からの数理モデル至上主義的ではない，たとえばFDRを指標としたSAM等の新しいデータ解析法を創り出し，内容を深めている．現時点ではこの統計科学の進展が健康科学に寄与をしているが，その趨勢を今後につなげるには，かって統計科学が遺伝学に多くの問題を提起したように，健康科学に新しい課題を提起することが必要であろう．

Historically, the statistical science and health sciences have mutually stimulated their progress by presenting problems to be resolved each other. Recently, however, more problems tend to be presented to the statistical science by health sciences, which is due to the development of information technology and micro-technology. Devising analysis methods for identifying disease associated genes from microarray data is one of the problems. The efforts of statisticians to answer the problems are making much contribution simultaneously to the statistical science and health sciences. For the future progress of both sciences, it is necessary for the statistical science to present important and interesting problems to health sciences.

Key Words and Phrases: 統計科学の未来，統計的原則，適応的臨床試験，遺伝子解析，代替法，インシリコ試験法，エピスタシス

1. はじめに

幼時に病気がちであった上，飢えた経験も持っている筆者は，「飢えと病」をキーワードにして今までの生き方を作ってきた．その立場で研究者としての社会的寄与を考えたとき，筆者が医学，薬学，生物学，健康科学（以下では「健康科学」と総称）を活動の舞台にしてきたのは当然として受け入れてもらえると思う．統計科学は，筆者がたまたま数値扱いを得手にして

[1] 本稿は，2006年度の統計関連学会連合大会（仙台，2006年9月）における日本統計学会75周年記念講演—21世紀の知識創造社会を支える統計科学の現状と展望—での講演を加筆・修正したものである．

* 東京理科大学工学部：〒162-8601 東京都新宿区神楽坂1-3　E-mail: isao@ms.kagu.tus.ac.jp

いたがゆえに，その舞台での小道具として使っているものである．

統計科学の学術雑誌である Biometrika は，1901 年に K. Pearson や W. F. R. Weldon らによって発刊された．発刊の辞に，“memoirs on variation, inheritance, and selection in Animals and Plants, based on the examination of statistically large numbers of species ...” とあるから，健康科学のための雑誌である（Sowan, 1982）．Journal of the Royal Statistical Society の第 1 巻は 1839 年発行であるから，それに 60 年遅れてはいるが，古くから健康科学は統計科学の世界で大きな存在であったと言えよう．

健康科学の統計科学への最も大きな寄与は，研究課題の提出であった．遺伝なる現象の解明，生態の測定・把握，病気の治療法の開発などには，現象の定量的把握と因果関係の究明が必要で，それを効率的に行えるようなデータ取得も必要であった．これが統計科学として，新しい統計解析法の考案と実験計画法・調査法の創出・体系化をもたらしてきている（Fisher, 1935）．

つまり，統計科学は，健康科学から課題を提出され，それに解答をだすことで発展してきた．その結果として統計科学は，医学に対して治療の有効性・安全性検証の方法論を提供して誤った治療法を正し，生物学に対して探索的側面での知見・仮説を提示して検証のための実験・調査の焦点化を行い，品種改良や生物種の発展の歴史の解明に寄与してきた．統計科学と健康科学は相互に寄与しあってきた，というのが両者の歴史である．

この相互寄与の歴史は，ときどきに関連の強さを変えながら現在に至っている．そして近年は，その相互寄与の様相がまた一段と強まっていると感じられる．そこで以下では，例を通してその相互寄与の強さや今後の関係について，感じていることを述べてみる．

2. 健康科学と統計科学の協同の場

健康科学にはいろいろな課題があるが，筆者の流儀で主として目指すところを区分けすると，（T1）生命の延長を阻むもの，（T2）生命の発生を阻むもの，（T3）生きることの質（quality of life; QOL）を低下させるもの，等についての機序の解明とそれへの対策が重要である．すなわち，（T1）に対しては治療と予防の方法を工夫することが対策であり，（T2）に対しては環境汚染の防止が対策であり，（T3）に対しては治療の質の改善が対策である．これらは健康科学と統計科学が協同して解決すべき研究課題である．

これらの課題への挑戦においては，たとえば，（S1）臨床試験，（S2）市販後調査，（S3）遺伝子解析，（S4）動物実験代替法，（S5）インシリコ試験（*in silico* assay）で得られるデータが利用される．

筆者自身あるいは筆者の周辺の研究者には，これらの試験法・実験法・調査法を用いて得られるデータの解析法や取得法について課題を見出し，その研究の中で課題に対する解答を追求している人が少なくない．健康科学の側からどのような課題が出され，統計家の側がそれにどのような試みで答えているかを，そのような研究事例を通して説明してみる．

3. 臨床試験

3.1 国際的統計ガイドラインの確立

様々な薬害・医療災害を経験して，その予防のためには「根拠に基づく治療」（evidence based medicine; EBM）が必要であるという認識が強まったのは 1995 年頃である．根拠として最も主要なものはデータである．データの中でも，ヒトに計画的に投薬・治療を施した結果のデータ，すなわち臨床試験のデータは証拠能力が最も大きい．（注：ヒト，サル，というようにカタカナで書いたときは動物の種を意味している．）ということで EBM の常識化の進行と共に，臨床試験における統計家の役割が重視されるようになり，臨床試験のデータ取得・解析

における統計ガイドラインが整備されてきた（厚生省医薬局審査管理課長, 1992; CPMP, 1995）. 筆者が臨床試験に本格的に関与するようになったのは，そういう機運が盛り上がってきて医学の分野で統計家の需要が増えはじめた 1993 年頃である.

臨床試験は投薬治療だけでなく，治療一般，あるいは食品の健康寄与効果や安全性についても用いられるが，説明を簡単化するため，本節では，被験薬の効果を対照薬に比べて比較するという形式での用語法を用いる. 新薬申請のためのこの種の臨床試験は「治験」と呼ばれ，投与される被験薬と対照薬を合わせて「治験薬」と呼ばれるが，この用語法もそのまま使うことにする. 臨床試験に参加する人を本節では「被験者」と呼ぶことにする. 被験者は然るべき割付法で群に分けられ，各群に一つの治験薬が対応づけられ，その群の被験者にはその群に対応する治験薬が投与される.

1990 年頃の日本の臨床試験では，事後解析・後智恵解析が横行しており，プロトコル違反などでの脱落が 30%を超える試験も稀でなかった. しかしそれは，米国を先頭とした臨床試験の国際的共通化を図る試み，International Conference on Harmonization（ICH）によって，逐次是正されてきた. ここで事後解析・後知恵解析（*post hoc* analysis）とは，層別や評価基準の変更を通して被験薬が有効と見られる条件を探し，その場合には有効であると結論づける解析のことである.

統計科学に関連するところでは，1998 年に「臨床試験には然るべき知識と経験を持つ試験統計家（trial statistician）が関与すべきである」ことを宣言した「臨床試験のための統計的原則（Statistical Principles for Clinical Trials)」というガイドライン（通称，ICH 統計ガイドライン）が世界共通の原則として公認された（Lewis et al., 2001；厚生省医薬安全局審査管理課長, 1998）. 適切な対照の選択についても原理が明確化された（厚生労働省医薬局審査管理課長, 2001）.

ICH 統計ガイドラインでは，臨床試験で統計科学的に最も重視すべきことは，偏りを最小にして，かつ精度を最大にすることであるとしている. そのための原則として，試験のやり方とデータの解析法を事前に計画書に詳しく書き，それを変更するときにはその妥当性を明確にすること，被験者の割付はランダムであること，主要評価変数はできるだけ一つにしてしかも客観的なものにすること，データ解析計画も盲検解除の前に固定すること，すなわちどの被験者がどの群に割り付けられているかが分からない状態のときに解析法を定めること，等を求めている. 優越性試験，同等性試験，非劣性試験等の概念の明確化も，この ICH 統計ガイドラインで初めてなされたことである.

しかしこの原則は，臨床試験の効率的実施に強すぎる制約を与えているということでそれなりの批判が出され，以下に示すように新しい手法が提案・検討されているのが現状である.

3.2 非ランダム割付法

臨床試験で被験薬の有効性を確認するためには，試験計画書で事前に定められた選択規準を満たし，しかも同様に定められた除外規準に抵触しない被験者をできるだけ偏りなく募集する. このとき，事前に定めた有効サイズ（effect size）に対応する対立仮説を（例えば有意水準 2.5%の）仮説検定で，（例えば検出力 80%で）検出できるように，被験者数を，（例えば各群 120 人というように）確保する. その被験者を確率 1/2 でそれぞれの群に逐次，ランダムに割り付ける.「逐次に」というのは，全被験者を一度にということでなく，同意を得られ次第 1 人 1 人，という意味である. 例えば毎月 10 人のペースで，3 年間かかって約 360 人の被験者を集める，といったことになる.

このように割り付けられた被験者に，試験計画書に指定されたやり方で治験薬を投与する. 投与においては，どの被験者がどちらの群に割り付けれられたかが，被験者にも治療担当医師

にも分からないようにしておき，（例えば有効率といった）評価のための変数を観測し，その群平均を2群間で比較する．比較の結果，被験薬群の値が有意に大きければ被験薬の有効性を認める，というのが治験の最も原則的なやり方である．このようなやり方の試験を，「ランダム化2重盲検並行群間比較試験」（randomized, double-blinded, parallel-group clinical traial）という．このやり方は，群間比較に確率論的原理を適用して仮説検定を行うためのFisher 3原則を応用したものである（Fisher, 1935；佐藤，1995；椿 他，1999）．ここで2重盲検とは．被験者と治療医師の両方共，どちらの治験薬を治療に用いているか知らないことである．

この原理は一見良さそうであるが，これを単純に適用すると，以下に述べるように「偏り無く精度の良い比較」が必ずしも実現できない，という問題が生じる．

治験の被験者はある疾患の患者であるが，その特質は，環境条件と遺伝因子を均質にして飼育した実験動物とは違って，非常に多種多様で個性的である．性，年齢，遺伝的因子，食事の質，肥満度，疾患の重症度等がまちまちである．これに加えて，治療を行う医療施設，医師の違いも無視できない多様性をもたらしている．

これらの中には，治療の結果を大きく左右する因子，すなわち予後因子（prognostic factor）となるものが少なくない．これらの因子の影響を無視してランダム割付で2群比較を行うのは，これらの因子による被験者の反応の違いをすべて「誤差」と見なすことになり，誤差を非常に大きなものとする．

単に誤差を大きくするだけでなく，結果として明らかな偏りが実現してしまうことも稀でない．例えば治療効果が劣る重症患者が30人いたときに，ランダムに割り付けた結果，一方の群に17人，他方の群に13人となったとしよう．当然17人割り付けられた群の方が治療の結果が平均的に悪くなる．真の治療効果とは別の原因が交絡（confound）して真の状態とは異なった偏った結果をもたらすのである．これが分かっていながら，ランダムに割り付けたのだからその差は偏りでなく偶然誤差である，というのは牽強付会である．結果としてこうなった場合には，予後因子の影響を調整して偏りが少なくなるように，推測を行うべきである．ただしその調整には，何らかのモデル想定が必要である．そのモデル次第で結論が異なるという問題が生じる．

Fisher（1935）はこれに対して，明らかに影響する因子を層別因子として扱い，層別ランダム化でその影響が偏りをもたらさないようにすることを提案している．層別ランダム割付である．よく知られるように，層別ランダム割付は群間比較の精度を上げるための優れた工夫である．

しかしながら臨床試験では，層別ランダム化がきわめて利用しにくい．総被験者数がたかだか200〜300人なのに，考慮しなければならない層の数が10を超えて多くなることが稀でないからである．実際，医療施設を層として考慮すると，各層での被験者数が多くても10人程度，少ないときは，4人程度になる．これに重症度という因子を含めて層別ランダム化を行うことは実際上不可能である．

このような状況の下で，層別割付よりは実用的で，ランダム割付よりは確実に予後因子のバランスを図ることができる割付法が提案され，使用されている．最小化法（minimization method）と通称される方法がそれで，たとえばPocock and Simon（1975）の方法がその典型である．

最小化法は，ある種の癌など，被験者数を多く集めるのが困難でしかも影響の大きい予後因子が存在する分野で，積極的に採用される傾向があった．やがてそれが，その必要性がない臨床試験にまで用いられるようになって欧州の規制当局（Committee for Proprietary Medicinal Products; CPMP）は，ICH統計ガイドラインと絡めてそれを強く否定（strictly discourage）す

る文書“Point to Consider”（CPMP, 2004）を発行した．その結果，最小化法の是非についての論議が激しく交わされるようになった（McEntegart, 2003; Buyse, 2004a; Senn, 2004; Buyse, 2004b）．

“Point to Consider”は，その根拠・理由があまり明確にされていない．これが議論を混乱させているのであるが，擁護者はその論議の中で，系統的にバランスを持たせることの利点がほとんど無い（minimum）のに予見可能性が生じることで結果が偏るから用いるべきでない，と主張している．そこで研究として必要なことは，予見可能性がほとんど生じないバランス割付法があるかどうかということと，系統的であってもバランスをとることに利点があるかどうかを明らかにすることである．

前者については，Pocock-Simonの原法などに偏コイン法（biased coin method）を導入することが考えられる（Efron, 1971）．すなわち，アンバランスを決定論的に減らそうとするのではなく，ある確率でバランスを取りやすくする方法の提案である．癌の臨床試験の条件でその方法を提案し，シミュレーション実験を通してその有用性を示したのがHagino et al.（2004）の研究である．

後者については，計量的予後因子のバランス割付を提案したNishi and Takaichi（2003）の割付法に，Endo et al.（2006a, 2006b）が，新しい評価基準と確率化を導入する研究を行い，バランス割付を採用することが予後因子調整に関する薬効評価の頑健性をもたらすことをシミュレーション実験で示している．これは2群だけでなく3群の場合にも適用でき，予後因子に質的因子と計量的因子が混じっているときでも適用できる方法なので，割付法の選択肢を増やしたものと言えよう．

3.3 複数評価変数への対処

検証的臨床試験における主要評価変数は，できるだけ一つに絞ることが勧められている．しかし，現実には複数の評価変数を用いざるを得ない疾患が少なくない．米国の規制当局である食品薬品庁（Food and Drug Administration; FDA）がやむを得ないと見ているものには，アルツハイマー病，慢性閉塞性呼吸器疾患等，20疾患がある（Offen et al., 2007）．

このようなときに問題になるのは被験者数の設計法（sample size design）である．臨床試験の原則では，必要な結論が明確に出せるという条件の下で，可能な限り少ない被験者の臨床試験を行うことが倫理的な観点から原則とされている（佐藤，1995）．より具体的には，被験薬の実際の効果サイズ，あるいは対象疾患で最低限必要な効果サイズに対して，一定の有意水準（ICHガイドラインでは2.5%）での検定で一定の検出力が保持されるような被験者数を計算し，臨床試験からの脱落確率を考慮に入れて被験者数を設定することになる．

ところが複数評価変数の場合に現実に採用されている設計法は，評価変数ごとに必要被験者数を設計してその大きい方を採用するという方法と，各評価変数を独立なものとして多変量的に被験者数を設計するという方法であった．これらの設計法は，原則的な視点での被験者数とは原理として異なった被験者数を与えることになる．

主要評価変数が複数の場合には，通常，多重性の調整無しにすべての変数で有意差がついたときのみ被験薬の有効性を認めるのが標準である．この原則で被験者数を設計するには，変数が正規分布に従う場合でも，多変量t分布ではなく，多変量正規分布の確率をウィシャート分布で重み付けを行って積分する必要がある．その視点と計算方法が臨床試験家の間では最近まで確立していなかった．

これに対して，Sozu et al.（2006）は，ウィシャート分布に関するモンテカルロ積分が確かな計算結果を与えることを示し，そのための計算プログラムを用意した．モンテカルロ積分が必要な理由は，それぞれの変数における検定統計量がいわゆるt統計量であっても，分散・共

分散がそのまま局外母数となり，その推定量に関する積分が，仮に正規分布が数値積分で行えたとしても，3次元（2変数）あるいは6次元（3変数）というように高次元になるためである．

この被験者数設計法では，効果サイズ以外に相関係数が局外母数となるので，現実にはその評価が設計の際に必要となる．このような局外母数の問題は，次項で述べる適応的試験計画導入の大きな動機となっている．

3.4 適応的試験計画の提案

本当に良い治療・薬剤をできるだけ短期間で開発するには，ICHガイドラインに定めているような「堅い」やり方ではなく，もっと柔軟にやり方を変更して試験を実施した方がいい，という声ががん治療や生殖医療など先端性のある分野で大きくなった．標準的なやり方は時間と手間がかかり，良い治療・薬剤を早く普及させることの障害になっているというわけである．

この種の要求はやがて，適応的試験計画（adaptive design, flexible design）というキャッチフレーズの下で大流行することとなった（Chou and Chang, 2007; Röohmel, 2006）．この3，4年での関連する研究論文数は，他のカテゴリーのものに比べて群を抜いて多い．

適応的試験計画の必要性は，ICHガイドライン制定の過程でも多少議論されていた．そのときのおおよその合意は，検証的臨床試験，つまり治療や薬剤の有効性を明確に確認するための臨床試験では，事前に推定していた局外母数の値の調整というような，比較的限られた場合のみにするべきだということであった．ガイドラインの文章としては，§4.4「必要な被験者数の調整」で次のように述べている．

> （被験者数の計算根拠となる仮定の）確認は，試験計画の詳細が予備的情報もしくは不確実な情報，又はその両方に基づいている場合，特に重要であろう．盲検下のデータを用い中間での確認を行うことにより，それまでの試験全体での，反応の分散，イベントの発生率又は生存状況が予期していた状況と異なることが明らかにされる場合がある．その場合，適切に修正した仮定に基づいて被験者数の再計算を行うこととなるが，その正当性を明らかにし，治験実施計画書の改訂及び総括報告書に記録しなければならない．

これに対して，現在の議論はこのような水準を遙かに超えて，柔軟性を極限まで追求したいというものである．すなわち，試験の途中で盲検を解除してデータを調べ，それに基づいて以下に列挙するような点について，試験計画を変更するというものである．

- 被験者数を計算し直して変更する．
- 計画されていた総被験者数はそのままにして各群への割付の比率を変更する．
- 複数の群のいくつかを途中で停止させて，残りの群のみで試験を継続する．
- 選択基準をたとえば重症者のみにする，というように変更する．
- 当初に設定していた有効性の主要評価変数を，例えば当初の副次評価変数で置きかえる．
- 単純に平均が大きいことの検定を行う優越性試験を，下駄（non-inferiority margin）を履かせた検定を行う非劣性試験に置きかえる．

このような試験法の変更は本来なら，実施場面での必要性と妥当性が先導して，それに問題がないかどうかを統計科学的に吟味するべきものである．ところが現在進行している状態は逆である．このような適応的試験計画を採用すると，第1種の過誤がどの程度不安定になるか，一部のデータが開示された後の条件付き検出力を制御する被験者数計算はどのようにすればよいか，途中で被験者数を変更したときの最終的な有効性評価はどのように行うべきか，ということなどについての数学的モデル上の議論・研究が活発なのである．被験者母集団の変更のような，どのように数学的仮定が変わるのか統計科学的に想定し難いことまで，単純な数学モデ

ルで議論されているのが現状である．

筆者は，このような統計科学独走の研究は，研究としては成立するが，実際の臨床試験の発展に寄与するかどうかについて疑問があると考えている．もっと丁寧にかつ実際的に，臨床試験関係者と統計家が協同で議論・検討・理解・協調・合意を試みるべきではないだろうか．

4. 市販後データの利用

4.1 薬剤の市販後調査の必要性

薬事法上の「薬」は臨床試験で有効性と安全性が確かめられた上で市販が許されたものである．しかし臨床試験は，危険性が経験的に明確でないこともあって非常に限られた条件でしか行うことができない．実際，臨床試験では，(1) 疾患の重症度が選択規準で指定される，(2) 併用薬の使用は原則として禁止される，(3) 合併症を持つ患者は除外される，(4) 事前に別の治療が行われている患者は除外される，(5) 妊娠期の女性と幼少年は除外される，…… というように，選択規準と除外規準が細かく定められている．さらに担当医師は原則として専門医であり，治療施設はある程度以上の技術・施設水準を持っているのが普通である．

倫理的理由によって，被験者数は多くても 1,000 人くらいまでというのが普通である．しかも臨床試験では，有効性の吟味が主眼とされていて，安全性を保証する条件が十分でないことが多い．したがって，現実に使用される多種多様な条件下で多数の患者でのみ発見できる稀な副作用についての情報は，臨床試験で得ることができない．そこで主として安全性を監視するために，市販後調査が必須となる．

市販後調査のデータには，(1) 製薬企業が使用成績を調べる調査，(2) 製薬企業に自発的に送られてくる報告，(3) 行政当局に自発的に送られてくる報告，(4) 臨床試験として計画される市販後試験の結果等いくつかの種類がある．このそれぞれでデータ解析での困難点や要工夫点が異なるので，統計家にはそれぞれに応じた解析法の開発と適用が求められる．

困難点を一般的に列挙すると，(1) 母集団（大きさ，特徴）が不明確で，得られた結果の適用範囲が確かでない，(2) 情報に偏りがあるにも関わらずその程度・大きさが評価し難い，(3) 交絡している予後因子・共変量が無数にあり，そのどれが真に結果に影響しているかが調べにくい，(4) 例えば現在論議の的になっているタミフルと少年の異常行動のように，副作用と疑われるものが見いだされても，その因果関係を確認する手段が乏しい，(5) 報告の精度・正確さが確かめ難い，等がある．

市販後調査担当者はこのような困難の中で，可能な限り早く，知られている副作用の頻度の推定，知られていない稀な副作用の発見に努めている．近年はそのための技法の研究が，データマイニング手法を応用したシグナル検出法という形で進められている（Matsushita et al., 2007）．

4.2 モデルを導入した使用成績調査データの解析例

市販後調査データは，臨床試験データと違って，一般には整然とした条件に従っていないので，解析も単純に頻度分布を比較する程度なのが普通である．しかし，特定の疾患，薬剤の組み合わせのみに注目する使用成績調査では，その特徴をモデル化して有用な情報を集約することが可能な場合もある．その一例が Fukushima et al.（2006）の研究である．以下にその要点を紹介する．

対象薬剤は抗がん剤 TS-1 である．抗がん剤は，がん細胞を攻撃すると同時に正常細胞である白血球や赤血球も破壊するので，副作用として血液毒性が出現する．したがって抗がん剤治療では，投与スケジュールの管理が重要になる．血液毒性が重篤な有害作用を発現させない範囲で，投与と休薬とをクールという単位で繰り返し，腫瘍縮小効果が最大になるように用法・

用量を定めるのが普通である.

そのスケジュールは開発過程での検討に従って，市販承認時に定められるが，その妥当性は市販後の使用成績調査で確認されなければならない．そこでTS-1の投与が行われた患者には定期的な血液検査が行われる．血液毒性は検査でしか調べられないからである.

図1は，そのような検査結果で発見された血液毒性の初発の発現頻度分布の例である．このデータの特徴は，検査日に依存して観察される発現頻度が左右されることである．実際には，その検査日以前に毒性が発現しているのであろうが，その日を特定できないから，いわゆる打ち切りデータ（censored data）になっている．このようなデータでは，単純な頻度比較で有用な情報を得ることができない．だからといって，たとえば1週間単位に頻度を集約したのでは，治療開始後のどの時期にどのように副作用が発現しているかということについて，精度の良い情報を得るのが困難である．情報の欠損が大きくなりすぎるからである.

これについてFukushima et al.（2006）は，有害事象の初発までの時間の分布（生存時間分布）についてモデルを想定して，最尤法で適合度を評価し，妥当と思われたモデルを前提にして有害事象発現プロファイルを評価した．ここでの有害事象は，白血球の数，赤血球の数等が，ある基準値以下になる事象のことである.

候補としてのモデルには，多少の試行錯誤の後で，後に示す式で表される，滑り混合ワイブルモデル（slip-mixed Weibull model）と滑り混合対数ロジスティックモデル（slip-mixed log-logistic model）が用いられた．ここで$h(t)$はハザード関数，すなわち時点tで生存している個体が有害事象を発現する瞬間確率，$\lambda_1, \lambda_2, \gamma_1, \gamma_2$は分布のパラメータ，$w$は第1クールと第2クール間の有害事象発現についての違いを表す重みパラメータ，$I(\bullet)$は命題“●”が真のとき1，そうでないとき0という値を取る指示関数である．時間tの単位は日で，第2クールの始まりが42日目であるため，第2クールについては時間を42日ずらしてある．このずらしを入れたことが「滑り」という形容詞を用いた理由である．（詳細は原論文を参照されたい.）

滑り混合ワイブルモデルのハザード関数

$$h(t)=W(t)\gamma_1\lambda_1(\lambda_1 t)^{\gamma_1-1}+(1-W(t))\gamma_2\lambda_2(\lambda_2(t-42))^{\gamma_2-1}I(42<t)$$

ただし，$W(t)$は次式の値である.

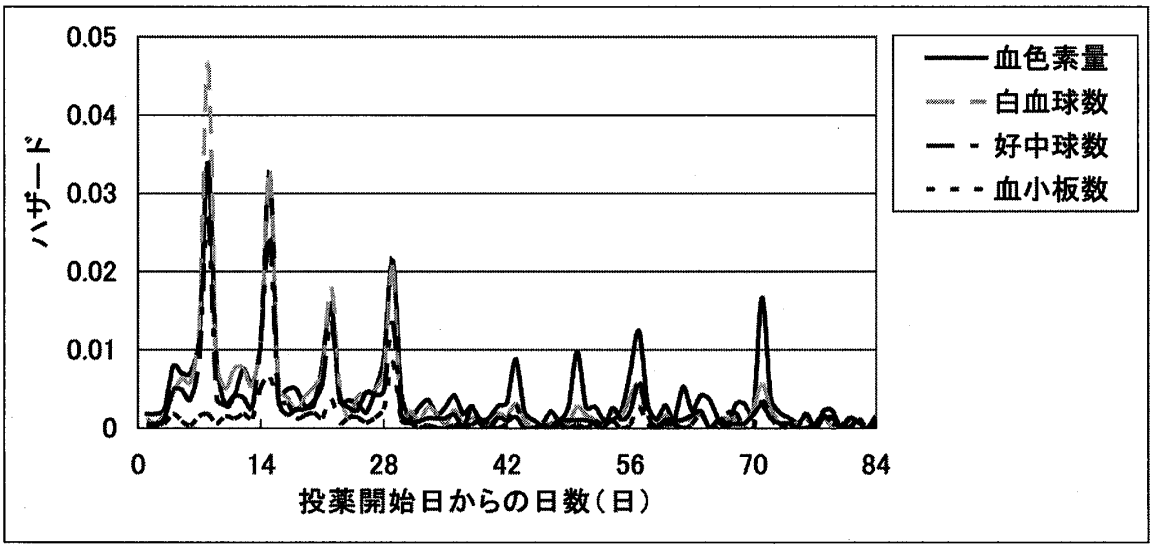

図1 血液検査値に基づく有害事象の出現状況

$$W(t)=\frac{w\exp(-(\lambda_1 t)^{\gamma_1})}{w\exp(-(\lambda_1 t)^{\gamma_1})+(1-w)(1-(1-\exp(-(\lambda_2(t-42))^{\gamma_2}))I(42<t))}$$

滑り混合対数ロジスティックモデルのハザード関数

$$h(t)=W(t)\frac{\gamma_1\lambda_1(\lambda_1 t)^{\gamma_1-1}}{1+(\lambda_1 t)^{\gamma_1}}+(1-W(t))\frac{\gamma_2\lambda_2(\lambda_2(t-42))^{\gamma_2-1}}{1+(\lambda_2(t-42))^{\gamma_2}}I(42<t)$$

ただし，$W(t)$ は次式の値である．

$$W(t)=\frac{w\dfrac{1}{1+(\lambda_1 t)^{\gamma_1}}}{w\dfrac{1}{1+(\lambda_1 t)^{\gamma_1}}+(1-w)\left(1-\dfrac{(\lambda_2(t-42))^{\gamma_2}}{1+(\lambda_2(t-42))^{\gamma_2}}I(42<t)\right)}$$

これらのモデルを実際のデータを当てはめたところ，滑り混合対数ロジスティックモデルの方がデータによく当てはまったので，Fukushima et al.（2006）は，これにもとづいて有害事象発現プロファイルを推定して薬理学及び実際の医療現場に接している担当者に確かめた．その結果この結論は妥当なものであるという評価を得ることができた．統計科学的検討によって有用な知見が得られたと言える．

5. 遺伝子解析の進展が提起する統計科学の問題

5.1 マイクロアレイデータの特徴

ヒトの全塩基配列を同定するというヒトゲノム計画が2000年に一段落したとき，専門家でない人たちの中には，これで個性がすべて計算機で調べられることになる，と誤解した人が少なくなかったであろう．

実際はそうでなかった．30億対の塩基配列が記号として計算機内に蓄えられたからといって，それが直ちに個人の個性の表現と対応づけられるわけではない．極端なアナロジーで言えばそれは，30億ステップのコンピュータプログラムを目の前に出されたからといって，そのプログラムが何をするか分かるわけでないのと同じである．現在のわれわれができるのは，その一部分に注目して，その部分が少し違ったプログラム（塩基配列）を並列的に流して（測定して），それによる結果の違いを見て，そのプログラム部分（対立遺伝子，allele）の役割を推測することである．これがプログラムについてのことであればプログラム解析と呼ばれ，塩基配列についてのことであれば統計的遺伝子解析と呼ばれることになる．

遺伝子解析では，単位をある遺伝子座（locus）における対立遺伝子（allele）の遺伝子型（genotype）の違いを取り上げる場合と，ある単塩基（single nucleotide）の位置（いわば所番地）における単塩基多型（single nucleotide polymorphism; SNP）の違いをとりあげるときがある．いずれにしろ，その型の違いが薬剤への反応の違いになるとき，これを「薬剤感受性がある」と表現し，その違いが疾患のかかりやすさの違いになるとき，これを「疾患感受性がある」と表現する．そういう用語法でいうと，遺伝子解析というのは，薬剤感受性や疾患感受性のある遺伝子やSNPsを突き止め，それがどのような情報伝達経路で，表現型（phenotype）の発現につながるかを検討することとなる．

薬剤（あるいは疾患）感受性遺伝子（あるいはSNP）を調べるのには，網羅的接近法が多く用いられる．たとえば，数千あるいは数万個の遺伝子に対応するDNA（遺伝子から産生される特定の蛋白）を網目状のプレートに捉えてその量を調べる技術が確立された．これがマイ

クロアレイである．

マイクロアレイでのDNA発現量を測定し，その発現量が正常細胞とがん細胞で有意に違っていればそれをがんに対して感受性を持つ対立遺伝子であると判定することになる．

ここで「有意に違う」というのは，たとえば有意水準5%のt-検定というようなものでの有意差ではない．マイクロアレイで一度に捉える対立遺伝子は数千から数万であり，英語表現では‘1000s or 10000s’という仮説検定を行うわけであるから，多重性によって第1種の過誤確率が極端に上昇する．おまけに，費用，手間，倫理性，プライバシー等の関係で，サンプルサイズが，英語表現では，‘10s or 100s’という程度に少ないので，真の感受性遺伝子を検出する確率はきわめて小さい．実際，筆者が今手許に持っているデータは，各群4～6匹の3群のマウスについての，12,489個の遺伝子のマイクロアレイデータである．このような小さいサンプルサイズで多重性を調整するのは無謀であろう．

5.2 感受性遺伝子同定法の例

超多変量小標本のマイクロアレイデータで感受性遺伝子を検出するのに，伝統的な有意水準指定の検定手法は無力である．統計科学上の工夫が必要である．いろいろ出されている提案の一つを例として紹介しよう．

マイクロアレイ上でN個の遺伝子の発現量が計測されたとしよう．被験薬群と対照薬群それぞれがn匹の動物（たとえばラット）からなっているとすると，各動物について，1枚分のマイクロアレイデータが得られるので，感受性遺伝子を調べることはサンプルサイズnの2標本問題になる．

発現量の差について，たとえばt統計量のような適当な検定統計量を用意すると，一つの遺伝子に一つの帰無仮説と対立仮説が対応することになる．検定ということであればN個の仮説の検定を同時に行うことになる．仮にN=1万個の遺伝子があるとすると，その中のたとえば50個というオーダーの遺伝子が薬剤感受性であろうというのが遺伝子学研究者の口にすることである．その感受性遺伝子の一つ一つが異なる強度の感受性を持っているから，対立仮説はある分布として存在していることになる．これでは検出力という概念が無意味で，対立仮説分布とその存在割合に応じて，棄却限界値を定めるという攻め方が合理的となる．

この視点から，第1種の過誤確率と第2種の過誤確率のバランスで棄却限界値を決めるのでなく，有意とされる遺伝子の中で誤って有意とされるものの割合，すなわち偽検出率（false discovery rate; FDR）を制御しようという考え方が出てきた．たとえばTusher et al.（2001）が提案している“significance analysis of microarrays（SAM）”と呼ばれるものがそれである．

第1種の過誤確率であれば，帰無仮説のみが確率計算に必要で，分布族を適当に想定すれば，たとえばt分布から，有意水準に対応する棄却限界値が計算できる．しかしFDRの場合にはその計算が感受性遺伝子の分布に依存する．それにもかかわらずその分布は不明である．何らかの方法でFDRをデータから推定しなければならない．それに並べ替え確率を用いようというのがTusher et al.（2001）の工夫である．

検定統計量を別のものにしたらどうか，対立仮説の分布に何らかの想定をしたらどうか，ということでSAMの改変版がいろいろと提案されている．Hirakawa et al.（2006）の提案もその一つである．

Hirakawa et al.（2006）は，検定統計量としてt-型統計量を用い，FDRを指定値以下にするような棄却限界値（cut-off value）を混合正規モデルの下で推定することを提案した．これは個々には他の研究で提案されているものであるが，両者を同時に用いたときに良い性能の感受性遺伝子検出法ができるというのが提案の特徴である．t-型統計量というのは，通常，統計量の分母におく標準誤差（SE）の推定量に，ある意味でJames-Stein的に，他の遺伝子から求め

たSEの情報を加味するもので，式で書けば次のようなものである．

$$t=\frac{\bar{x}-\bar{y}}{\sqrt{s_x^2/n+s_y^2/n+a_0}}$$

この式で，x, yはそれぞれ，被験薬群と対照薬群の測定値を意味し，分子は測定値の平均の差，分母の平方根の中は不偏分散の和，最後に加えてあるa_0は，N個の不偏分散の上側10%点である．分母にこのようにa_0を加えるのは，サンプルサイズがわずか$n=4$というオーダーであることから来る不偏分散の不安定さを，他の遺伝子の情報を利用して安定化させるもので，従来の統計科学の枠組みからはなかなか思いつかない工夫であろう．実際，シミュレーション実験で調べると，この工夫はかなり良い性能を解析法に付与している．

6. 動物実験代替法のバリデーション研究

6.1 代替法の3Rs原則

ヒトの健康のためだとしても動物を虐待するのはよくないという社会風潮が，ヨーロッパを起点にして世界に広がった．最初のターゲットは化粧品の安全性の吟味に動物を使うことの禁止であったが，現在では，一般的な生命倫理（bioethics）の課題として3Rs原則が法律化されてきている．2007年8月には，3Rs原則を普及するための第6回国際会議（6th World Congress on Alternatives & Animal Use in the Life Sciences）が日本で開催される．

3Rs原則というのは，

Replacement：動物を使わない試験法の採用

Refinement：動物の苦痛の少ない試験法の採用

Reduction：使用動物数の少ない試験法の採用

であり，毒性試験の分野ではそのための試験法が次々と開発されている．日本ではこれらを一括して「動物実験代替法」あるいは単に「代替法」と呼んでいる．

代替法にはどのようなものがあるかというと，例えば人の皮膚を侵す皮膚腐食性を評価するために「3次元ヒト皮膚モデル」が市販されている．これはヒトの皮膚表面の3層をヒトの細胞を培養増殖させることでインビトロ的な試料として作成したものである．紫外線等の光にさらされることで皮膚に炎症が起こる現象を助長する光皮膚刺激性を評価するのには，例えば酵母菌の増殖が光を当てた場合と当てない場合でどれくらい違うかを調べる試験も日本で開発されている．この他にも，発癌性を調べる細胞形質転換試験等々多くの試験法が開発されつつある．

参考のために言うと，薬理試験，毒性試験，安全性試験の生物的側面は，臨床薬理試験を除けば，ヒト以外の生物あるいは生物由来の試料を用いて行われる．その試験には，インビボ試験（*in vivo* test），インサイチュウ試験（*in situ* test），インビトロ試験（*in vitro* test）の区別があって，費用，時間，労力，精度，を勘案してそのどれを用いるかがおおよそ定まっている．*vivo*は生体，*situ*は存在しているその場という意味である．*vitro*はガラスを意味するラテン語であるが，ここでは試験管を意味している．これらにはそれぞれ独自のデザイン法と解析法の統計的問題があって，例えばOmori et al.（2002），Matsunaga et al.（2002），Soek et al.（2006）などの研究があるが，これについての議論は本論文から割愛する．

話を戻すと，代替法の開発が進む中で統計科学の課題も次々と登場してきている．例えば，開発された試験法が確かにヒトへの安全性を担保するか，予測性能の保証をどのように行えばよいか，というバリデーション研究の方法論の確立である．

一般論として言えば，Organization for Economic Co-operation and Development（OECD）が

バリデーションの規準をガイドライン“OECD series on testing and assessment No. 34: Guidance document on the validation and international acceptance of new or updated test methods for hazard assessment”にまとめている．代替法のバリデーションは世界的にこのガイドラインに沿って行なわれている．このガイドラインが，個々の代替法の妥当性がどの水準に達しているか，を評価するための一里塚（“modular approach”）として出しているのが次の項目である．

（i）Test definition（including purpose, need and scientific basis）；（定義）
（ii）Intra-laboratory repeatability and reproducibility；（施設内再現性）
（iii）Inter-laboratory transferability；（技術易移転性）
（iv）Inter-laboratory reproducibility；（施設間再現性）
（v）Predictive capacity（accuracy）；（予測性能）
（vi）Applicability domain;（適用対象範囲）and,
（vii）Performance standards.（標準性能）

これらの各段階では常に，統計科学的評価が必要である．(i) では，毒性に関する陽性，擬陽性，陰性の判定基準（criteria），あるいは毒性の強さの指標が適切に定義されていることが必要であり，これが統計科学的に妥当なものであることが根拠を持って示されていなければならない．(ii) では，同じ実験施設で実験をすればその代替法が，許容可能な程度のばらつきで同じ結果を与えるかどうかを実験データから評価することが求められる．(iii)～(vii) でも同様な課題が提起されている．

6.2 バリデーション研究における実験計画

ある試験法がある毒性の代替法として提案されると，その施設間再現性を確認するという問題が生じる．試験すべき被験化合物は典型的なものでも100種以上あり，研究をする労力・施設・機材・時間などはボランティアに依存しているから，バリデーション研究のために行う実験は必要最小限に限ることが求められる．

そのため，いろいろな制約下で最適な研究計画を作ることが「統計家の仕事」として統計科学の分野に持ち込まれる．これは100年前に農場試験のやり方について統計家が提起された課題と似ている．違うのは制約が多種多様で，釣り合い不完備ブロック配置（balanced incomplete block design; BIB）という類の単純なやり方が役に立たないことである．

施設間再現性を調べるのであるから，多くの施設に実験を依頼する必要があるが，各施設それぞれの事情があり，すべてに同じ数の被験化合物，繰り返し，時期を求めることはできない．結果に影響する共変量の制御も難しい．技術力にも差がある．毒性の極端に強い被験化合物と全く毒性のない被験化合物では再現性についてのばらつきが違う．毒性が中程度の被験化合物ではばらつきが大きくなる傾向がある．それらをすべて考慮に入れて，施設への被験化合物の割付を行い，かつ得られたデータを解析して施設間再現性を評価しなければならない．これは結構難しい仕事である．

被験化合物の主効果や実験施設の主効果といったものが母数因子と想定できて，しかも加法模型が仮定できるならば，既に確立している実験計画法が応用できる．実際には施設を変量とせざるを得ないから，混合模型で加法性と等分散性が成り立たない場合に最適な実験計画を求めるという問題に帰着する．現在のところ，これについての方法論は確立していない．

そこで考えられることは，現実上の制約を満たす試験計画を全列挙し，適当な最適化規準に関して指標値を計算し，指標値が比較的良好な割付をバリデーション研究に適用することである．たとえば高沼 他（2006）はこのやり方で皮膚感作性試験代替法のひとつであるLLNA-DA

法の実験計画を検討している．このような問題に対する方法論の確立は今後の課題であろう．

7. インシリコ試験の利用

7.1 インシリコ試験とは

従来はインビボ試験とインビトロ試験がほとんどであった毒性試験，薬理試験の分野に，最近インシリコ試験（*in silico*）が割り込んで来て，活用されるようになっている．*silico* は元素の珪素を意味するシリコンのことで，計算機の主要構成材の半導体がシリコンを素材としている関係で計算機を意味するのに使われている．古い語源を追えばラテン語であろうが特にラテン語として使っているわけではない．

インシリコ試験は，何万とある開発候補化合物から目的に適合している化合物をスクリーニングする際に，計算機内で構造の関係を評価し，有望な物質とそうでない物質を区分けしようとする技術である．これに試験というラベルを貼るのは，有望性の評価の際にいろいろな条件設定をして評価値を求め，それを試験における測定値とみなし，その測定値を総合比較してスクリーニングを実行するからである．

インシリコ試験法の本体をなす道具は，被験化合物の構造や属性を計算機に入力したときに評価値を出力するソフトウエアである．すでに多くのインシリコ試験法が市販されているが，その開発者は主として計算機技術者と化学者である．統計家はあまり関与していないようである．出力をどのように活用するかというマニュアルに統計科学的センスが見られないからである．

入力として用いられる変数は，たとえば受容体の立体構造，被験化合物の立体構造，大きさ，酸・塩基の化学的特性など，かなり多いのが普通である．それらを用いてソフトウエアが両者の結合しやすさを評価値とするとき，使用者（インシリコ試験の実験家）がオプションとして指定すべきパラメータが，シミュレーション回数，許容距離限界等沢山ある．それをどうしたらよいかというのが統計家に問われる問題である．

7.2 インシリコ試験データの解析法の開発例

Kakumoto et al.（2004）は Dock というインシリコ試験法を用いている実験家から依頼を受けて，最適なオプションパラメータの定め方の研究を行った．

研究の当初では，最適なパラメータの同定という形で問題を設定していたが，どのようにパラメータを設定しても，インビトロ試験あるいはインビボ試験の結果を予測する性能があまり良くないという結果を得て，これがこの試験法の限界であろうと考えた．しかしあるとき，特定の最適設定を求めるのではなく，複数のパラメータで実験値を得てそれを多変量的に利用するというアイデアを思いつき，変数選択を行って最適予測式を構成してみたところ，これが非常によい性能を持つことが分かった．

考えてみれば当然で，受容体も物質としては分子量が非常に大きい有機化合物で，被験化合物もそうである．結合のしやすさも1次元的ではあり得ないから，被験化合物の性質に応じて，異なる結合関係で他と同じ強さの結合（活性）を持つことがあり得る．したがってスクリーニングでは，できるだけ多くの側面を評価し得る多変量解析的な接近法が有効なのである．これはコロンブスの卵であって，分かってみて初めて言えることである．

Kakumoto et al.（2004）は，一つの試験法での出力を多変量的に利用することを考えたが，同様な利用法は，複数の試験法の結果を利用するときにも考えられる．実際 Hayashi et al.（2005）は三つのインシリコ試験法をバッテリーとして，ある決定樹（decision tree）にしたがって用いると遺伝毒性の評価に有効であることを報告している．インシリコ試験法のデータは多変量的に利用するのが良いようである．

8. 過去を回顧して未来を思量する

8.1 健康科学は統計科学にどのような寄与をしているか

統計科学は空想の世界に存在するのではない．現実世界からデータに関する問題の提起を受け，それに答えようとすることで発展するものである．100年前のPearsonやFisherが研究していたことがまさにそういう性質の問題であった．上に挙げた事例はいずれも同じような役割で統計科学の前に登場したと筆者は考えている．考えるべきことは，その問題に解答を出す営為が学としての統計科学の内容を豊かにしたかということである．

そういう意味で一般化すると，確率モデルでの定式化を超えた問題，母集団的なものが曖昧な状況，超多変量小標本のデータ，統計科学的帰無仮説での過誤や検出力では評価できない誤りの評価，データの多変量的利用，といったところで多くの新しい試みがなされている．しかもそれが健康科学に十分な寄与を与えていると筆者には感じられる．健康科学が統計科学に新しい課題を提出し，それに答えようとすることで統計科学に新しい展開がもたらされつつあるのではないだろうか．

しかしここで次の自問が筆者の中に芽生えてくる．

自問：現在の時点では統計科学が健康科学からの刺激を受けて発展しており，健康科学にある程度の寄与をしている．しかしこれが未来の両者の同時発展，相互寄与に結びつき得るのだろうか？ そうだとすると，その実現においてはどのような留意点があるだろうか．

8.2 遺伝学と統計科学

自問に対して自答を試みてみよう．それには過去の歴史を回顧するのが良いと思われる．

他の分野で見ると，過去に無数というべき論文が流行的に出されたにもかかわらず最終的には学界から消滅したものが稀でない．たとえば，1980年代にフィーバーが起きた「常温核融合」や，大型プロジェクトの「電磁流体発電」「ごみの完全自動処理」がそれである．

他方で「information technology」や「micro-technology」は，流行がそのまま新しい技術革命をもたらし，社会のありよう全体に巨大な影響をもたらしている．生殖技術はこれらに裏付けられて革命的な人工操作を可能とし，結果として親子関係や兄弟姉妹の関係が定義困難という状況まで創造してしまっている．

どういう要因が短期的流行と永続的発展を分ける，あるいは左右するのであろうか．これは，技術史，技術論の主題の一つであり，気楽に答えられるようなことではないが，統計科学と健康科学の関係については，参考になりそうな面白い文献がある．Cordell (2002) である．

遺伝学にはエピスタシス (epistasis) という概念がある．ある遺伝子によって異なった座にある遺伝子の発現が抑止され，上位の遺伝子の発現が優先される現象である．発現型が現われる遺伝子を上位の (epistatic)，抑止されるものを下位の (hypostatic) 遺伝子という．これについてCordell (2002) は次のように書いている．

> The term 'epistatic' was first used in 1909 by Bateson (1909) to describe a masking effect whereby a variant or allele at one locus prevents the variant at another locus from manifesting its effect. This was seen as an extension of the concept of dominance for alleles within the same allelomorphic pair, i. e., at a single locus.
> (エピスタティックという用語がはじめて使われたのは，1909年にBatesonがある遺伝子座での変異が他の遺伝子座の変異の発現を妨げるという遮蔽効果を記したときである．これは

一つの同じ遺伝子内の対立遺伝子対の間の優性劣性の概念の拡張と見なされた.)

表1はエピスタシスの一つの表現例である．もし，上位の遺伝子Gが存在しなければ，あるいはそれがg/gという型であれば，下位の遺伝子Bによって髪の色が白か黒になる．ところが相互作用をもつ上位の遺伝子の型がg/GあるいはG/Gのときは白と黒が覆われて，灰色が優先する，というような現象が起こる．これがエピスタシスである．

この概念が何故提起されたかというと，それはメンデルの法則が成り立つという仮説の下で観察データを集めていたら，うまくいく例とうまくいかない例があったことであろう．エピスタシスという仕組みを仮定すれば，統計的に観察された後者の現象がうまく説明できることになる．もしそうだとすれば，これは100年前に統計科学が遺伝学に対して提起した仮説であったに違いない．

Cordellはこれについて，次のように述べている．推測通りと考えて良いであろう．

The situation has been confused further by the fact that in quantitative genetics, following a paper by Fisher in 1918, the term 'epistatic' has been generally used in yet another different sense from its original usage. In Fisher's 1918 definition, epistasis refers to a deviation from additivity in the effect of alleles at different loci with respect to their contribution to a quantitative phenotype.

DNA配列が同定され，遺伝子座が具体的に指定できるようになった現在，統計科学と遺伝子学の共同作業として，具体的にどの遺伝子座の間でエピスタシスが起こっているかが遺伝子—遺伝子相互作用，ということで追求されている．そしてそこでは，先に述べた関連性評価における多重性が，統計的検出力を乱すという困難をもたらし，統計科学はこの困難に，たとえば“Multifactor dimensionality reduction”というようなソフトウエアで挑戦している（Moore, 2005; Ritchie, 2005）.

100年近くの間をおいて今，昔の仮説の追求が再開されているのであるが，この間に起こったことは遺伝子解読という革命的技術の発展である．つまり，統計科学と健康科学（に限るわけではないが）の一方の進展は他方に大きな課題を提起し，それがある時間をかけて解決されると，次にはそれが別の課題を返してくるという関係があることが，研究を単なる流行で終わらせない要因ではなかろうか.

かって筆者は，「統計学は良薬を創るのに役立つだろうか」（吉村，1995）という問題提起を行ったが，それは他の分野に影響を与えられるほどでなければ，相互発展に寄与することはないという視点に基づいたものである．これは現在の多く行われている健康科学関連の統計科学の研究の将来を決めるものではなかろうか.

8.3　未来の相互寄与・発展を願って

Röhmel（2006）が適応的試験計画について示唆的で興味深い次の文章を書いている.

表1　Batesonが定義したエピスタシス現象を，相互作用を持つ2つの遺伝子座での遺伝子型（b/B及びg/G）の表現型（例えば髪の色の白，黒，灰色）で示した例

遺伝子Bの表現型	遺伝子Gの表現型		
	g/g	g/G	G/G
b/b	白	灰色	灰色
b/B	黒	灰色	灰色
B/B	黒	灰色	灰色

> If popularity of a scientific concept were measured by the number of recently published manuscripts, one could only agree that "adaptive design" must be a very popular method. … the theoretical progress is far ahead of practical applications.
> (近年に公刊された論文の数で概念の普及ぶりを評価するなら，適応的試験計画は抜群に普及した概念ということになる．しかし実際は，理論が駆け足で走っていて実用がはるかに引き離されてしまっている.)

ここでいうところの論文は，統計科学の雑誌に載ったものであって健康科学の雑誌に載ったものではない．統計的に検証するのはほぼ不可能であるが，今のところ健康科学の分野で適応的試験計画という概念があまり受け入れられていないのは確かである．これは単に理論が実用を引き離しているというだけでなく，統計科学が健康科学に寄与する内容を提供していないことを意味しているのではないだろうか.

第4節に例示したFukushima et al. (2006) の研究結果は，当初，取り上げている薬剤の開発・営業担当者から，強い異議を受けた．結論が実感と合わないということである．研究の過程でFukushimaは，その異議が薬剤投与初期でのモデルとデータ間の不適合であることをつきとめ，当初用いていたワイブルモデルに加えて対数ロジスティックモデルを候補に入れ，どちらがデータにより適合するかを検討した．結果として，後者の方が適合するという統計的結論が得られ，その結果が異議を唱えた担当者からも受け入れられ，社内での検討会で実際の営業現場でも活用できるということになった.

この例は，統計科学での研究結果を健康科学の分野に提示し，健康科学の分野からの疑問を受け，共同研究を進めることで，統計科学の方法論としての提案を健康科学の成果にできたものと言える.

同様なことは，Sato et al. (2004; 2006), Suganami et al. (2007) においても言える．実際，遺伝子解析のSato et al.の研究は，検出したSNPsが医学的な意味で，確かに疾患感受性SNPsであるかどうか，医学的に吟味する必要があるということで，医学者から発表の保留を求められた．しかしその後，医学的にも妥当性があるということでその分野の研究論文が投稿される段階に来ている.

これらの例を通して私が感じることは，健康科学は次々と統計科学的課題を提出し解答を統計家に要求してくる．それに対して統計家は自らの方法論を駆使して解答を試みる．その結果を健康科学側に提示してその受容可能性を問う．もし，受容可能性が低いならば，健康科学の発展に寄与する形で解答の改善を試み相手に受け入れられる内容にまで結果を昇華させる.

もし，その受け入れが，たとえば後知恵解析の否定のように，正しい原則であるにもかかわらず健康科学側が受け入れを渋っているのであれば，説得性のある説明を通してその普及を図る．このような過程を通して，統計科学への信頼感を健康科学に持たせられるなら将来の相互寄与は確かなものとなるであろう．統計科学の方もそれを通して新しい方法論を作らざるを得なくなり，ひいてはそれが統計科学の発展になるのであろうということである.

飛躍していることを承知で言うならば，統計科学と健康科学の過去の協調・共同研究の歴史は，学のアイデンティティの確立にこだわらず，健康科学にとって何が求められているかをその立場で追求することが，結果として統計科学自体の確立，発展につながったものと感じられる．現在，統計科学の中に健康科学が提起した課題が大きな位置を持っているのは過去のそういう営為の結果であり，それを続けていくことが未来の相互発展をもたらすと思量するのであるがどうだろうか.

謝　辞

本稿をまとめるに当たっては山本拓前会長の適切な問題提起と，国友直人元理事長の寛容なる激励を受けた．匿名の査読者のコメントも本稿を改善するのに役立っている．必ずしも表には出ていないが，例示した各論文の著者からは，その内容に関する詳細な資料の提供を受けた．これらの方々に心から謝意を表したい．

学術誌の慣例にしたがって文中では敬称を省略した．お許し頂きたい．

参 考 文 献

Bateson, W. (1909). *Mendel's Principle of Heredity, Cambridge University Press*, Cambridge. (注：筆者は未入手であるが，Cordell (2002) に引用されているので，読者の参考のためリストに入れておく．)

Buyse, M. and McEntegart, D. (2004a). The CPMP's position discouraging dynamic allocation techniques is unfair, *Applied Clinical Trials*, Letters to Editor, May 1, 20004.

Buyse, M. and McEntegart, D. (2004b). More nonSENNse about balance in clinical trials, *Applied Clinical Trials*, Letters to Editor, July 1, 2004.

Chou, S. C. and Chang, M. (2007). *Adaptive Design Methods in Clinical Trials, Chapman & Hall.*

Cordell, H. J. (2002). Epistatis: what it means, what it doesn't mean, and statistical methods to detect it in humans, *Human Molecular Genetics*, **11**, 2463–2468.

CPMP Working Party on Efficacy of Medicinal Products (1995). Biostatistical methodology in clinical trials in application for marketing authorisations for medicinal products, *Statistics in Medicine*, **14**, 1659–1682.

CPMP (2004). Points to consider on adjustment for baseline covariates, *Statistics in Medicine*, **23**, 701–709.

Efron, B. (1971). Forcing a sequential experiment to be balanced, *Biometrika*, **58**, 403–417.

Endo, A., F. Nagatani, Hamada, C. and Yoshimura, I. (2006a). Minimization method for balancing continuous prognostic variables between treatment and control groups using Kullback-Leibler divergence, *Contemporary Clinical Trials*, **27**, 420–431.

Endo, A., Hamada, C. and Yoshimura, I. (2006b). An allocation method for balancing continuous prognostic variables among treatment groups using the Kullback-Leibler information, *Japanese Journal of Biometrics*, **27**, 1–16.

Fisher, R. A. (1935). *The Design of Experiments*, Oliver & Boyd.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance, *Transactions of the Royal Society of Edinburgh*, **52**, 399–433. (注：現在筆者の手元にないが，Cordell (2002) に引用されているので，読者の参考のためリストに入れておく．)

Fukushima, A., Kashiwagi, W., Sano, M., Hamada, C. and Yoshimura, I. (2006). Estimating a hazard function for each of four items of adverse event induced by the anti-cancer drug TS-1 –Application of slip-mixed log-logistic model for interval censored data–, *Japanese Journal of Pharmacoepidemiology*, **11**, 9–21.

Hagino, A., Hamada, C., Yoshimura, I., Ohashi, Y., Sakamoto, J. and Nakazato, H. (2004). Statistical comparison of random allocation methods in cancer clinical trials, *Controlled Clinical Trials*, **25/6**, 572–584.

Hayashi, M., Kamata, E., Hirose, A., Takahashi, M. Morita, T. and Ema, M. (2005). In silico assessment of chemical mutagenesis in comparison with results of salmonella microsome assay on 909 chemicals, *Genetic toxicology and Environmental Mutagenesis*, 129–135.

Hirakawa, A., Sato, Y., Sozu, T., Hamada, C. and Yoshimura, I. (2006). Estimating the false discovery rate using a normal mixture distribution with microarray data, poster in the *XXIIIrd International Biometric Conference*, July 16–21, 2006, Montreal, Canada.

Kakumoto, K., Yamanaka, S. Hamada, C. and Yoshimura, I. (2004). A statistical analysis of an effective method to conduct *in silico* screening for active compounds, *Chem-Bio Informatics Journal*, **4**, 121–132.

厚生省医薬安全局審査管理課長 (1998)．「「臨床試験のための統計的原則」について」, http://www.nihs.go.jp/dig/ich/eindex.html, last access 2007年1月.

厚生省薬務局新医薬品課長 (1992)．「「臨床試験の統計解析に関するガイドライン」について」, 薬新薬第20号.

厚生労働省医薬局審査管理課長 (2001)．「「臨床試験における対照群の選択とそれに関連する諸問題」について」, http://www.nihs.go.jp/dig/ich/eindex.html, last access 2007年1月.

Lewis, J., Röhmel, J., Huitfeldt, B., Yoshimura, I., Sato, T., Uwoi, T., Uesaka, H., O'Neill, R., Ellenberg, S., Louv, B. and Ruberg, S. (1999). Statistical principles for clinical trials, *Statistics in Medicine*, **18**, 1905–1942.

Matsunaga, N., Kanno, J. and Yoshimura, I. (2002). A statistical method for judging synergism: Application to an

endocrine disruptor animal experiment, *Environmetrics*, **14**, 213–222.

Matsushita, Y., Kuroda, Y., Niwa, S., Sonehara, S., Hamada, C. and Yoshimura, I. (2007). Criteria revision and performance comparison of three methods of signal detection applied to the spontaneous reporting database of a pharmaceutical manufacturer, *Drug Safety*, (in press).

McEntegart, D. J. (2003). The pursuit of balance using stratified and dynamic randomization techniques: An overviews, *Drug Information Journal*, **37**, 293–308.

Moore, J. H. (2005). A global view of epistasis, *Nature Genetics*, **37**, 13–14.

Nishi, T. and Takaichi, A. (2003). An extended minimization method to assure similar means of continuous prognostic variables between treatment groups, *Japanese Journal of Biometrics*, **24**, 43–55.

Offen, W., Chuang-Stein, C., Dmitrienko, A., Littman, G., Maca, J., Meyerson, L., Murhead, R., Stryszak, P., Boddy, A., Chen, K., Copley-Merriman, K., Dere, W., Givens, S., Hall, D., Henry, D., Jackson, J. D., Krishen, A., Liu, T., Ryder, S., Sankoh, A. J., Wang, J. and Yeh, C. H. (2007). Multiple co-primary endpoints: Medical and statistical solutions; A report from the multiple endpoints expert team of the pharmaceutical research and manufacturers of America, *Drug Information Journal*, **41**, 31–46.

Omori, T., Honma, M., Hayashi, M., Honda, Y. and Yoshimura, I. (2002). A new statistical method for evaluation of L5178Y tk+/- mammalian cell mutation data using microwell method, *Mutation Research*, **517**, 199–208.

Pocock, S. J. and Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial, *Biometrics*, **31**, 103–115.

Ritchie, M. D. (2005). A global view of epistasis, *Neurosurg Focus*, **19**, October, 1–4.

Röhmel, J. (2006). Adaptive designs: Expectations are high, *Biomedical Journal*, **48**, 491–492.

佐藤俊哉 (1995).「ヘルスサイエンスのための統計科学：サンプルサイズの設計」,『医学のあゆみ』, **173/13**, 1041–1046.

Sato, Y., Suganami, H., Hamada, C., Yoshimura, I., Yoshida, T. and Yoshimura, K. (2004). Designing a multistage SNP-based genome screen for common diseases, *Journal of Human Genetics*, **49**, 669–676.

Sato, Y., Suganami, H., Hamada, C., Yoshimura, I., Sakamoto, H., Yoshida, T. and Yoshimura, K. (2006). The confidence interval of allelic odds ratios under the Hardy-Weinberg disequilibrium, *Journal of Human Genetics*, **51**, 772–780.

Senn, S. (2004). Unbalanced claims for balance, *Applied Clinical Trials*, Letters to Editor, July 1, 20004.

Seok, K. J., Wanibuchi, H., Morimura, K., Totsuka, Y., Wakabayashi, K., Yoshimura, I. and Fukushima, S. (2006). Existence of a no effect level for MelQx hepatocarcinognicity on a background of thioacetamide-induced liver damage in rats, *Cancer Science*, **37**, 453–458.

Sowan, B. (1982). Biometrika, *Encyclopedia of Statistical Sciences*, **1**, 248–251.

Sozu, T., Kanou, T., Hamada, C. and Yoshimura, I. (2006). Power and sample size calculations in clinical trials with multiple primary variables, *Japanese Journal of Biometrics*, **27**, 83–96.

Suganami, H., Kano, K., Kuwayama, Y., Hamada, C. and Yoshimura, I. (2007). Comparison of methods for parameter estimation in a circular linear mixed effect model incorporating the diurnal variation for evaluating the treatment effects of glaucoma therapy, *Japanese Journal of Biometrics*, **28**, 1–17.

高沼正幸, 寒水孝司, 大森崇, 浜田知久馬, 吉村功 (2006).「動物実験代替法バリデーション研究における被験物質割付の最適性に関する検討」, 日本動物実験代替法学会第20回大会要旨集, 119–120.

椿広計, 藤田利治, 佐藤俊哉 (1999).『これからの臨床試験』, 朝倉書店.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response, *Proceedings of National Academy of Sciences*, **98**, 5116–5121.

吉村功 (1995).「統計学は良薬を創るのに役立つだろうか」,『数理科学』, **389**, 43–49.