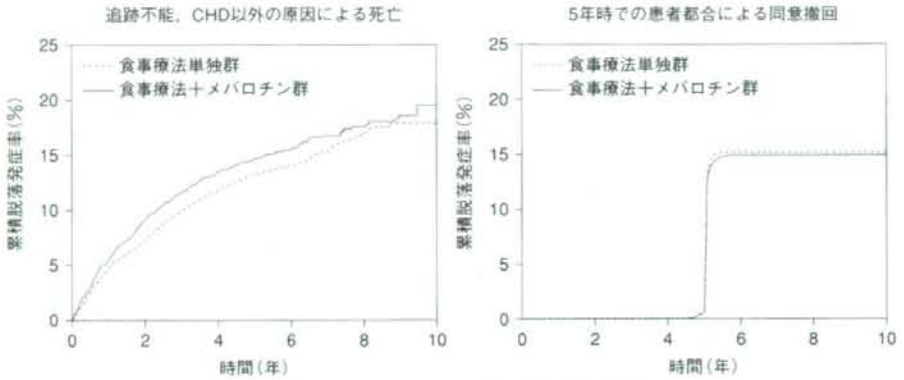

図2　MEGA Studyにおける累積脱落発症曲線

ロール(LDL-C))，時間依存性要因として5因子(TC，TG，HDL-C，LDL-C，治療群)を候補変数とした．LDL-CはFriedewald式を用いているため，TC，TG，HDL-C，LDL-Cのすべての変数を含めると，変数間の相関関係が非常に強くなり，推定が不安定になる問題(多重共線性)が懸念されるが，TG，HDL-Cは対数変換値を用いたため，多重共線性は回避されている．観察確率の推定(説明変数の組み合わせ)として，以下の5つのモデルを考えたが，最終的な結果はいずれのモデルを用いたとしても変化しなかったので，モデル3に基づく結果を**表4**に示す．

モデル1：すべての要因(ベースライン12因子＋時間依存性5因子)
モデル2：モデル1からTC値と時間依存性治療変数を除く
モデル3：モデル2からすべてのベースライン脂質値を除く
モデル4：モデル1からTC値と時間依存性TG，HDL-C，LDL-Cを除く
モデル5：モデル2において有意な因子のみ

表4に示したように，脱落理由ごと，治療群ごとに，脱落に関与する要因は異なっており，別々にモデル化したことは適切であったといえる．ただし，いずれの理由による脱落においても，高血圧や糖尿病，および高脂血症治療歴のない対象者，つまり，ベースライン時において，比較的状態の良い対象者ほど脱落しやすい傾向がみられた．MEGA Studyは一次予防試験であることから，相対的に軽症患者が多く，例えば，通常の臨床試験でみられるような治療の副作用による脱落，状態悪化による脱落といった状況とは異なり，試験に参加し続けることに対する煩わしさやモチベーションの低下が脱落を引き起こした可能性があると思われる．「追跡不能，CHD以外の原因による死亡」のうち，追跡不能の理由のほとんどが同意撤回であり(食事療法単独群：51.8%，メバロチン群：64.3%)，CHD以外の原因による死亡者数は食事療法単独群の方が多かったのに対し，試験途中での同意撤回者数はメバロチン群の方が多かったこと，さらに薬剤投与されていない対象者ほど脱落傾向にあったことなどからもその可能性がうかがえる．また，時間依存性要因としてのTGやLDL-C値が高い対象者ほど脱落傾向にあったことからも，試験期間中に，適切な食事指導に従わず，食事コントロールができなかった対象者ほど脱落しやすい傾向にあった可能性が考えられる．

上述のモデル(**表4**の結果)から，対象者ごとの観察確率の推定値を経時的に求め，その累積確率の逆数を重みとしたIPCW解析(重み付きlog-rank検定，重み付きハザード比推定)を行った結果を**表5**に示す．表中の「標準」と記載した方法は，すべての脱落を情報のない打ち切りとみなした解析のことで，一次評価項目のCHDに関しては**図1**の結果に対応する．表中には二次評価項目の脳卒中に関しても結果を記載している．いずれの評価項目に関しても，IPCW解析によって脱落を補正したとしても治療効果は大きく変化しないことがわかる．また，IPCW解析の結果は観察確率の推定モデルに依存していない．IPCW解析が妥当であるためには，観察確率を正しく推定する必要があるが，**表4**に示したように多くの臨床的に重要な予後因子を考慮しており，仮定からの乖離は大きくないと思われる．したがって，MEGA Studyで観察された5年時の同意撤回

表4　脱落に影響する予後因子の検討(モデル3の結果)

要　因	追跡不能，CHD以外の原因による死亡				5年時での患者都合による同意撤回			
	食事療法単独群		メバロチン群		食事療法単独群		メバロチン群	
	HR	95% CI	HR	95% CI	OR	95% CI	OR	95% CI
ベースライン変数								
年齢(歳)	1.01	0.99, 1.02	1.00	0.99, 1.02	0.99	0.97, 1.01	1.00	0.98, 1.02
女　性	1.08	0.85, 1.37	1.00	0.80, 1.27	0.80	0.55, 1.18	1.28	0.86, 1.89
BMI(kg/m^2)	1.01	0.98, 1.04	0.98	0.95, 1.01	1.10	0.97, 1.16	0.98	0.94, 1.03
現在の喫煙歴	1.13	0.88, 1.44	1.21	0.94, 1.54	1.14	0.75, 1.72	1.23	0.81, 1.85
現在の飲酒歴	1.14	0.91, 1.45	1.03	0.83, 1.28	0.81	0.55, 1.17	1.17	0.80, 1.70
高脂血症治療歴	0.84	0.66, 1.08	0.63	0.48, 0.81	0.87	0.58, 1.29	0.70	0.45, 1.08
高血圧	0.82	0.68, 0.98	0.91	0.77, 1.07	0.79	0.60, 1.05	0.76	0.57, 1.01
糖尿病	1.02	0.82, 1.25	0.72	0.58, 0.90	0.83	0.60, 1.18	1.01	0.73, 1.42
時間依存性変数								
TG (mg/dL)	1.11	0.92, 1.35	1.30	1.08, 1.57	1.56	1.08, 2.25	1.72	1.20, 2.48
HDL-C (mg/dL)	0.82	0.54, 1.26	1.11	0.74, 1.67	3.14	1.54, 6.41	1.76	0.84, 3.69
LDL-C (mg/dL)	1.00	0.99, 1.01	1.01	1.01, 1.01	1.00	1.00, 1.01	1.00	1.00, 1.01

TG，HDL-Cは対数変換値(自然対数)．
HR：ハザード比，OR：オッズ比，CI：信頼区間．

表5　IPCW法により脱落を補正した治療効果

方　法		冠動脈疾患			脳卒中		
		HR	95% CI	p値	HR	95% CI	p値
標　準		0.67	0.49, 0.91	0.010	0.83	0.57, 1.21	0.33
IPCW	モデル1	0.65	0.48, 0.89	0.007	0.81	0.56, 1.18	0.27
	モデル2	0.66	0.48, 0.90	0.008	0.81	0.56, 1.18	0.28
	モデル3	0.66	0.49, 0.90	0.009	0.81	0.56, 1.17	0.26
	モデル4	0.66	0.49, 0.91	0.009	0.81	0.56, 1.18	0.27
	モデル5	0.66	0.48, 0.90	0.008	0.82	0.57, 1.19	0.29

HR：ハザード比，CI：信頼区間．

を含む脱落は，予後因子に依存した大きな選択バイアスを引き起こすものではなく，主解析結果は脱落の影響をほとんど受けていないことが示唆されたといえる．

ノンコンプライアンスを補正した解析

表2に示したように，MEGA Studyにおいて治療に対するコンプライアンスは完全ではない．ここでは，ノンコンプライアンスを「割り付け治療と反対の治療を一度でも受けた」，つまり「食事療法単独群に割り付けられたがメバロチンの投与を受けた，あるいはメバロチン群に割り付けられたが休薬を一度でも経験した」と定義し，その影響を補正する．

そのようなノンコンプライアンス(治療法の交換)を起こした対象者のCHD発症状況(10年間)を**表6**に示す．CHD発症リスク差のITT推定値を計算すると，

$$\frac{66}{3,866}-\frac{101}{3,966}=-0.0084$$

となり，メバロチン割り付け群の方がCHD発症リスクを絶対リスクで0.84%減少させていることがわかる．一方，実際に受けた治療に基づいた群間比較を行うと，

$$\frac{19+46}{844+1,425}-\frac{82+20}{3,122+2,441}=0.01$$

となり，メバロチン治療を一度でも受けた群の方がCHD発症リスクを1%増加させるという結果となる．後者の実際に受けた治療に基づく解析は，治療法の交換がランダムに起きていない限りバイアスのある結果を導くことが知られている．状態が悪くなってきたの

表6　MEGA Studyにおけるノンコンプライアンス状況(10年間)

割り付け群	対象者数	CHD発症数	実際の治療	対象者数	CHD発症数
食事療法 単独群	3,966	101	食事療法 メバロチン[a]	3,122 844	82 19
食事療法＋ メバロチン群	3,866	66	食事療法[b] メバロチン	2,441 1,425	20 46

[a]：メバロチン治療を一度でも受けた患者.
[b]：メバロチン治療を一度でも休薬した患者.

表7　全員が食事療法単独治療を受けていた場合に予想される結果

割り付け群	CHD発症	CHD非発症	合計
食事療法単独群	$101-844\,\delta$	$3{,}966-101+844\,\delta$	3,966
食事療法＋メバロチン群	$66-1{,}425\,\delta$	$3{,}866-66+1{,}425\,\delta$	3,866
合　計	$167-(844+1{,}425)\,\delta$	$7{,}832-167+(844+1{,}425)\,\delta$	7,832

でメバロチンの投与を受けた，あるいは状態が良いので休薬したなどのようにランダムでない治療法の交換がMEGA Studyでは起きていると予想されるので，実際に受けた治療法に基づく解析はノンコンプライアンスの補正方法として適切でない．実際，上記で計算したように，メバロチン治療はCHD発症リスクを増加させるという解釈困難な結果であり，因果の逆転，すなわち状態の悪い(CHD発症リスクの高い)対象者が選択的にメバロチン投与を受けた結果と考えるべきである．

上述のようなランダムでない治療法の交換が生じている場合に，その影響を適切に補正し，コンプライアンスが完全に保たれた場合に観察されたであろう治療効果，つまり，全対象者が試験期間を通じて割り付け群どおりの治療を受けた場合の理想的な治療効果を推定するための方法[6,7]が近年いくつか提案されている．g推定法と呼ばれるランダム化に基づく解析[6]と，intensity score法と呼ばれるモデルに基づく解析[7]の2つが主に提案されているが，本稿では前者の方法を紹介する．なお，後者のintensity score法とは，各対象者が当該治療(MEGA Studyではメバロチン治療)を受ける確率(propensity scoreと呼ばれる)を様々な患者背景因子から推定し，それをもとに前述の実際に受けた治療に基づく解析のバイアスを補正する方法である(実際に受けた治療とpropensity scoreの差をintensity scoreと呼ぶ)．

対象者 i に対する潜在的なCHD発症状況を2通り想定する．1つは対象者 i がメバロチン治療を受けた場合の結果(Y_{i1} と表記する)であり，もう1つが同じ対象者 i が食事療法単独治療を受けた場合の結果(Y_{i0} と表記する)である．現実には，対象者 i はメバロチン治療を受けたか受けていないかのどちらかなので，この2つの潜在結果変数のうちの一方のみが観察されることになる．しかし，このような想定のもとでは，本来推定すべき治療効果は，Y_{i1} と Y_{i0} の差として定義できる．つまり，同じ個人に別の治療法を施したときに違いが生じれば，その違いの理由は治療法で説明されることになる．

議論の単純化のために，g推定法によるノンコンプライアンス補正を分割表解析において説明する．ここで，真の治療効果(δ)がすべての対象者に対して共通と仮定し($\delta = Y_{i1} - Y_{i0}$)，すべての対象者が食事療法単独治療を受けた場合の結果を予測してみる．**表6**において，実際に受けた治療がメバロチンの場合には，その治療効果(δ)を差し引くことでその予測は可能である(実際に受けた治療が食事療法であった場合には予測したい状況が実現している)．その結果を**表7**に示す．この分割表は「すべての対象者が食事療法単独治療を受けた場合の結果」なので，CHD発症に関して割り付け群間に全く差がないはずである．**表7**に対するカイ二乗検定統計量を計算すると，

$$Z(\delta)=\frac{3{,}966\times66-3{,}866\times101-(1{,}425\times3{,}966-3{,}866\times844)\,\delta}{\sqrt{\frac{1}{7{,}832}\times3{,}866\times3{,}966\times[167-(1{,}425+844)\,\delta][7{,}832-167+(1{,}425+844)\,\delta]}}$$

となるが，上式の $Z(\delta)=0$ の解が求めるべき真の治療効果の推定値となる．この方程式を計算すると，

表8　g推定法，intensity score法によりノンコンプライアンスを補正した治療効果

方　法	冠動脈疾患				脳卒中			
	5年		10年		5年		10年	
	HR	95% CI	HR	95% CI	HR	95% CI	HR	95% CI
ITT法	0.70	0.50, 0.97	0.67	0.49, 0.91	0.65	0.43, 0.97	0.83	0.57, 1.21
g推定法	0.65	0.30, 0.91	0.64	0.39, 0.83	0.54	0.26, 0.87	0.63	0.33, 1.26
intensity score法	0.68	0.44, 1.05	0.59	0.36, 0.99	0.44	0.25, 0.79	0.51	0.28, 0.95

ITT：intent-to-treat，HR：ハザード比，CI：信頼区間．

表9　治療不遵守例におけるノンコンプライアンス率(/年)

割り付け群	冠動脈疾患イベント		脳卒中イベント	
	あ　り	な　し	あ　り	な　し
食事療法単独群[a]	0.098	0.119	0.026	0.122
食事療法＋メバロチン群[b]	0.137	0.217	0.169	0.216

[a]：メバロチン治療を一度でも受けた患者における服用率(/年)．
[b]：メバロチン治療を一度でも休薬した患者における休薬率(/年)．

$$\delta = \frac{\frac{66}{3,866} - \frac{101}{3,966}}{\frac{1,425}{3,866} - \frac{844}{3,966}} = \frac{-0.0084}{0.1558} = -0.054$$

となり，メバロチン治療を受けた群の方がCHD発症リスクを絶対リスクで5.4%減少させるという結果となる．上式の真の治療効果(δ)の分子はITT解析によるリスク差であり，分母は各割り付け群における割り付け治療を守った割合の差である．したがって，ITT治療効果を割り付け治療を遵守した程度で補正することで，ITT治療効果の過小方向のバイアスを修正していることになる(実際，**表6**におけるITTリスク差は−0.0084である)．

同様の補正は，**図1**のようないわゆる生存時間解析においても可能である[4,6]．また，ノンコンプラインアスは各対象者に対して経時的に測定されるものであり，その状態の時間変化も考慮可能である．観察期間を1年間隔で区切り，1年ごとにノンコンプライアンス(治療法の交換)を集計し，その方法をMEGA Studyに適用した結果を**表8**に示す．CHD(一次評価項目)と脳卒中(二次評価項目)の両方に関して，追跡期間を5年とした場合と10年とした場合の結果を示す．いずれの評価項目に関しても，ノンコンプライアンスを補正することで，より大きなイベント発症抑制効果が得られている．すなわち，全対象者が割り付け群どおりの治療を受け続けた場合の真の治療効果は，ITT治療効果よりも大きいといえる．特に，その補正効果はCHDよりも脳卒中において顕著である．

脳卒中において，補正値とITT推定値の間に大きな乖離がみられた原因を検討するために，治療不遵守例に対する探索的な検討を試みた．一般に，ノンコンプライアンスが増えれば，ITT推定値の真の治療効果からの乖離が大きくなるが，その主な理由としては，1．「食事療法単独群に割り付けられたが，メバロチンを服用したためにイベントが減った」，2．「メバロチン群に割り付けられたが，休薬してしまったためにイベントが増えた」の2つが考えられる．これらの影響度合いがCHDと脳卒中で異なるために，**表8**のような結果が観察されたと推察される．

割り付け治療と反対の治療を一度でも受けた対象者(治療不遵守例)におけるノンコンプライアンス率(/年)と各イベント発症状況を**表9**に示す．まず食事療法単独群について考える．メバロチン服用の影響が，CHDと脳卒中に対して同じように作用していたとすれば，**表9**において，イベント「あり」「なし」のメバロチン服用率は，CHDと脳卒中で同程度の大きさとなっていたはずである．イベント「なし」においては，メバロチン服用率は，いずれも0.12(/年)前後であるのに対し，イベント「あり」においては，CHDの方がメバロチン服用率が大きくなっている．各対象者のメバロチン服用率がイベント「なし」に与える影響をロジスティック回帰で検討したところ，CHDに対してはオッズ比＝14.5，脳卒中に対してはオッズ比＝144となり，メバロチン服用によるイベント抑制効果は脳卒中においてか

なり大きくなっている．

次に，メバロチン群について考えてみる．イベント「なし」においては，メバロチン休薬率はいずれも0.22(/年)であるのに対し，イベント「あり」においては，メバロチン休薬率が脳卒中の方で大きくなっている．各対象者のメバロチン休薬率がイベント「あり」に与える影響は，CHDに対してオッズ比＝1.26，脳卒中に対してオッズ比＝5.71であり，メバロチンを休薬したことによるイベント発症への影響は，CHDに比べて脳卒中の方が大きかったといえる．以上のことから，ノンコンプライアンス(メバロチン服用/休薬)の補正の影響度は脳卒中においてより強く現れることが予想される．つまり，CHDよりも脳卒中において，メバロチンのイベント抑制効果をより強める方向にノンコンプライアンスの影響が補正されており，このことがITT推定値との差が脳卒中においてより顕著にみられた理由と考えられる．

おわりに

MEGA Studyの主解析の結果を補足するための2つの解析(脱落補正とノンコンプライアンス補正)を行った．多くの予後因子の影響を考慮したIPCW法による脱落補正結果は，主解析結果とほとんど同じであり，5年時の同意撤回を含む様々な脱落は，予後に依存した情報のある打ち切りを引き起こすものではないことが示唆された．一方，g推定法(あるいは，intensity score法)によってノンコンプライアンスを補正した結果，CHD，脳卒中のいずれに対してもITT治療効果よりも大きなイベント抑制効果が観察された．一般に，ノンコンプライアンス割合が増えれば，ITT推定値の真の治療効果からの乖離が大きくなるため，全期間(10年間)の解析結果は，5年時までの解析結果に比べてノンコンプライアンスの影響が大きいと思われる．これらのことを考慮すると，メバロチンの薬剤としての薬理学的有効性(efficacy)を議論する場合やサブグループでの治療効果を検討する場合は，5年までのデータを用いた検討が良いと思われる．

文　献

1) Nakamura H, Arakawa K, Itakura H, et al：MEGA Study Group：Primary prevention of cardiovascular disease with pravastatin in Japan (MEGA Study)：a prospective randomised controlled trial. Lancet 2006；**368**：1155-1163.
2) 完成したMEGA Studyのすべて―11年の労苦とその克服の軌跡―．Prog Med 2006；**26**(Suppl 2).
3) Yoshida M, Matsuyama Y, Ohashi Y, for the MEGA Study Group：Estimation of treatment effect adjusting for dependent censoring using the IPCW method：an application to a large primary prevention study for coronary events (MEGA study). Clin Trials 2007；**4**：318-328.
4) Tanaka Y, Matsuyama Y, Ohashi Y, for the MEGA Study Group：Estimation of treatment effect adjusting for treatment changes using the intensity score method：application to a large primary prevention study for coronary events (MEGA study). Stat Med 2008；**27**：1718-1733.
5) Robins JM, Finkelstein DM：Correcting for non compliance and dependent censoring in an AIDS Clinical Trial with inverse probability of censoring weighted (IPCW) log-rank tests. Biometrics 2000；**56**：779-788.
6) Mark SD, Robins JM：A method for the analysis of randomized trials with compliance information：an application to the Multiple Risk Factor Intervention Trial. Control Clin Trials 1993；**14**：79-97.
7) Brumback B, Greenland S, Redman M, et al：The intensity score approach to adjusting for confounding. Biometrics 2003；**59**：274-285.

I. 総　　論

大規模臨床試験の意義

Large scale clinical trials: Gold standard or magnificent waste ?

大橋靖雄

Key words：ランダム化臨床試験，CONSORT，臨床試験登録，evidence based medicine(EBM)，メタアナリシス

1. 我が国における治験・臨床試験の基盤整備

臨床医学の目標は，患者の疾患を正確に診断し適切に治療を行うことを通じ，患者の生命予後とQOLを向上させることにある．しかし，医療提供者側の知識・技術の不完全さ，類型化によってしばしば切り捨てられる患者・疾患の多様性から，診断結果や治療結果には永遠に除去不可能な不確実さが伴う．したがって，診断・治療によってもたらされるベネフィットと被るリスクは，ともに確率変数的性格を帯び，それらのマスとしての評価には，程度の差こそあれ統計的要約が必要となる．一方，ベネフィット・リスクに対する患者の重み付けは患者個々の価値判断を反映して異なるはずのものである．したがって，診断・治療法の選択は，リスク・ベネフィット両者のバランスと資源の制約，統計的要約で切り捨てられる個別事情の斟酌の中で，本来は十分な情報提示と理解，そして自発的意思を前提としたインフォームド・コンセントによるべきであるとされている(これがevidence based medicine: EBMの実践であろう)．しかし，医療提供者と患者の有する情報の不均衡，情報理解の不完全さ，医師・患者双方の意識の問題もあり，これまでの治療上の意思決定は，医師主導のパターナリズムの中で，曖昧な状況下で行われてきたといってよい．

近年の情報公開あるいは患者の権利主張の流れは，このような意思決定プロセスに大きな変革を与えつつある．厚生労働省は，1990年代末から疾患ごとに標準治療をまとめた診療ガイドライン策定を各関連学会に依頼し，数多くのガイドラインやその案が発表されてきた(例えばMINDSデータベース[1]参照のこと)．そして，このガイドライン策定過程で，多くの医療関係者には周知であった次の事態が浮き彫りになった．'我が国には客観的証拠(エビデンス)がない！'．臨床医学系学会では，1990年代半ばからEBMという言葉が大流行したが，ここでエビデンスを提供するのが患者を対象とした臨床研究成果であり，本特集の大規模臨床試験はその中核をなす．

一方，1993年のソリブジン事件を重要な契機として，日本の臨床試験，特に治験の制度・それを取り巻く環境・そして内容は大きな変貌をとげた．制度の面では，現在の医薬品医療機器総合機構(PMDA)設立につながる審査体制の変革と1998年の新Good Clinical Practice(GCP)完全施行に代表される国際ハーモナイゼーション(以下略してICH)の受け入れ，そして2003年から開始された医師主導治験制度の開始が象徴である(**表1**参照)．産業界は，これまでの官・産一体の護送船団方式からICH-E5ガイド

Yasuo Ohashi: Department of Biostatistics, School of Public Health, The University of Tokyo　東京大学医学系研究科 公共健康医学専攻生物統計学

0047-1852/08/¥40/頁/JCLS

表1 最近15年の臨床試験の歴史

1993.9	ソリブジン事件
1996.5	ICH-GCP(E6ガイドライン)国際合意(ステップ4)
1996.6	薬事法改定
1997.3	答申GCP，省令GCP通知
1997.4	医薬品機構誕生
1997.7	医薬品・医療機器審査センター誕生
1998.4	GCP完全実施
1998.8	ICH-E5ガイドライン'海外臨床データ受け入れにおける人種要因差'通知
1998.11	ICH-E9ガイドライン'臨床試験のための統計的原則'通知
1999.2	'適応外使用に係る医療用医薬品の取り扱い'通知
2002.7	薬事法改定
2003.4	治験活性化3ヵ年計画
2003.6	'医師主導の治験の実施の規準'通知
	'臨床研究に関する倫理指針'通知
2004.4	医薬品・医療機器総合機構誕生
2007.4	新たな治験活性化5ヵ年計画
2008.3	'高度医療評価制度'通知

ライン(統計ガイドラインE9を含む臨床試験ガイドラインについてはPMDAホームページ[2]参照)に基づく海外データの受け入れ開始を経て，市場と開発の場を海外に求めるに至り，現象としては，国内治験の空洞化と治験グローバル化，そして相次ぐ企業合併を引き起こした．治験受け入れ側では，大学の(少なくとも一時的な)地盤沈下が発生し，一方で治験を収益源とする開業医やSite Management Organization(SMO)の雨後の竹の子状の設立と淘汰，モニタリングなどを企業から受託するContract Research Organization(CRO)の勃興，そしてClinical Research Coordinator(CRC)が試験成功のキーパーソンとして認知されるに至った．上記のEBMの考え方が普及し，日本からエビデンス発信が希薄であったことの反省とともに，医師主導臨床試験の重要性が認識された．そしていまだに十分ではないものの厚生労働省からの臨床試験に対する研究投資の増加とともに，がんを中心として数多くの医師主導試験実施のための組織が誕生した．

このような流れを促進する目的で，2003年には文部科学省・厚生労働省の'治験活性化3カ年計画'，次いで2007年には'新たな治験活性化5カ年計画'[3]が発表された．いずれも'治験'活性化をうたっているものの，底上げとしての臨床研究の基盤整備を目指した方針であり，これに基づいて治験拠点病院・中核病院が指定され，また大学を中心として橋渡し研究(トランスレーショナルリサーチ)拠点施設も指定された．いずれの計画においても臨床試験を支えるCRCの重要性の指摘，その養成・適正配置の必要性がうたわれ，更に5カ年計画においては，試験研究を計画・実施する医師研究者の養成，臨床試験の計画・解析を担当する生物統計家の活用とその養成，データの信頼性を保証するデータマネジメントの重要性とそれを担当するデータマネージャの養成と配置の必要性がうたわれている．未承認や適応のない医薬品・医療機器を臨床試験の中で保険医療と混合して利用可能とする'高度医療評価制度'も，2008年3月の医政局長通知[4]で動き出したが，ここでもデータの信頼性を保証するモニタリングとデータマネジメント体制が，制度適用の必要条件として挙げられている．製薬会社主導の治験ばかりではない臨床研究全体に対するインフラストラクチャ作りが，ようやく本格化したものととらえることができる．2008年7月に案が公表される(本原稿執筆時点では予定)'臨床研究に関する倫理指針'がその象徴となるであろう．こ

の指針ではすべての臨床試験に補償が要求されるものと予想されるが，その受け皿となる保険商品の開発も予定されている．また2004年に新設されたPMDAについても，本省組織と融合しての本省回帰，あるいは独立行政法人のまま本省組織を取り込む形での拡大の2案が検討されている．

本稿では，まず新GCP以前の我が国の治験の問題点を概観し，診療エビデンスの中核となる大規模臨床試験がなぜ我が国で困難であったかを振り返る．次いでがん領域で起きつつある臨床試験パラダイムシフトについて述べ，臨床試験実施と報告の公正さを保証する試験登録などの最近の動きを紹介する．最後に(大規模)臨床試験結果の解釈上のチェックポイントをまとめる．イベント発生をエンドポイントとした臨床試験結果を解釈するうえで必要となる統計的概念・用語について，参考として末尾に付した．本節から第4節前半までは原著論文としては大橋[5]が初出であり，これに必要な加筆修正を行っている．また，第4節のb.‘疫学研究と臨床試験との接近’と第5節は大橋[6]が初出である．用語集も大橋[6]とほぼ同様であるが読者の便宜のため再掲した．

2. 新GCP以前

a. Exploratory対confirmatory，explanatory対pragmatic

臨床試験の分類機軸にexploratory(探索的)対confirmatory(検証的)，explanatory(説明的)対pragmatic(実践的)がある．これらの言葉を使えば，新GCP以前の日本の治験の問題はconfirmatoryであるべき統計解析と解釈をexploratoryに行い，本来相反するexplanatoryとpragmaticアプローチを混同し，どちらつかずの臨床試験を行っていたことに帰着する．ちなみに治療エビデンスの中核となる大規模臨床試験は本質的にpragmaticである(べきである)．

exploratoryという言葉は，統計学全般の中ではTukeyの‘革命的’著書Exploratory Data Analysis[7]によっていわば‘提唱’された．Tukeyは数学的に体系化された仮説検定論(Neyman-Pearson理論)に代表されるこれまでの統計的データ解析をconfirmatoryと呼び，

- データ解析は数学的最適理論ではなく，データが語ろうとしていることをいかに抽出するかが本質
- そのためにコンピュータ利用を含む新しい手法が必要
- モデルが真実を反映しているかの診断が重要
- モデルあるいは前提と真実とが乖離していても大きな影響を受けないロバストな手法が有用

という現在の薬効評価の統計学にも通じるデータ解析の基本的哲学を提唱した．ちなみに，アメリカにおける医薬品の認可当局FDAは‘ロバスト’という言葉を極めて重視する．

臨床試験の分野では，confirmatoryとexploratoryの区別は，(臨床試験に携わる統計家にとって周知であったが)1998年のICH-E9統計ガイドラインにより強く認識されるに至った．もちろん一つの臨床試験にはいずれの側面も存在し，統計解析計画書では検証的部分と探索的部分を明確に分けることが要求されている．より後期の試験，例えば第III相試験では検証的側面が主になるものの，有効なサブグループの探索(治療効果と背景要因の交互作用の検討)など探索的解析も適正使用のために重要な情報を提供する．

同ガイドラインにより，多重性(multiplicity)，すなわち検定の繰り返しによって第一種の過誤を上昇させる行為はきつく戒められることとなり，これは現在の我が国の医薬品審査にも厳密に適用されている．多重エンドポイント・多重時点・多重サブグループ・多群の場合の2群比較の繰り返し・複数の検定方法がこの例である．割付け開示後の追加解析も広義にはこれに含まれる．この分類機軸に対する理解は，研究者(スポンサー)側・審査側にも十分浸透していると思われる．

なお，臨床試験におけるconfirmationの統計的・数学的な基盤は，ランダム化で生成される確率空間から計算されるp値である．すべての応用分野を通じp値の数学的根拠が明確な場は

2つしか存在しない．無作為抽出(サンプリング)あるいはランダム化がなされている場面である．前者が実質不可能な臨床試験においては，結論の一般化可能性を当面は犠牲としてもconfirmationのためにランダム化を行うのである．したがって，確率空間あるいは結論をゆがめる(バイアスを生む，あるいは第一種の過誤を増大させる)解析上の取り扱いは極力避けることになる．そのために採用されるのがintent(ion)-to-treat(以下略してITT: 用語集参照)の考え方である．

b. これまでの反省

上記の分類機軸を用いれば新GCP施行前の，我が国の臨床試験の方法論上の問題点は以下のようになる(詳しくは大橋[6])：

0. confirmationという概念の理解が申請者側・審査側とも不十分であった．
1. 1施設あたりの症例数が少ない．
2. 実薬を対照とした同等性試験で最終的な検証が行われる．
3. 主治医による主観的総合評価が主たる(プライマリ)エンドポイントである．

0については既に述べた．全体で有意性が証明できない場合にサブグループの結果から承認に至った例もあったように思われるが，(2に関連して)有意でない場合には同等として承認されたという最大の問題がある．2群比較を例にとれば，'統計的に有意でない'理由として次の2つが考えられる：

・2群間に本当に差がない．
・差を有意に検出できるだけの情報量が存在しない．

1群1例の臨床試験を行えば，2群間の差が有意となることは絶対にありえない．またいいかげんな臨床試験を行えば(極端には調査票をいいかげんに医師に書いてもらえば)，大数の法則から大規模臨床試験の結果は2群間で同等になる．後者が'比較試験の感度(assay sensitivity)'の問題である．プラセボや極めて有効な薬剤の効果を検出できない臨床試験系が'感度の低い'試験系であり，そこからは同等の結果が出やすくなる．我が国でかつて抗痴呆薬として処方されていたホパテン酸カルシウム(ホパテ)が1989年に劇薬指定され，実質上市場から撤退した．これを受け，ホパテを実薬対照として同等性試験により認可された複数の薬剤のプラセボ対照試験が実施された．その結果，1剤を除きプラセボに対する優越性が証明できず市場撤退するという惨憺たる結果となった．以前のホパテ対照試験の試験感度が低かったことが最大の理由であろう．

上記1についていえば，稀少疾患については1施設の登録例数が少ないことは当然である．このような場合でもランダム化が正しく行われれば結論の妥当性は保証され，かつ施設が多いことから試験はpragmaticで一般化可能性も高くなる，という議論がありうる．しかしこれは日本の多施設少数患者試験には全く当てはまらない．1施設内で同様の適格条件を有した試験が並行に行われることから患者選択のバイアスが入るからである．実薬対照および主治医評価は，一見試験をpragmaticな立場から行う方法であるかのように思われる．しかしこれまでの日本の臨床試験の第III相試験の組み立ては，適格条件設定・評価・解析とも短期・小中規模なexplanatoryな試験のそれである．この中に，感度と信頼性が保証されないまま2，3を取り込んだことに問題があったと考える．

一方，(QOLを考慮した)生存や疾患発症・再発をエンドポイントとした大規模試験をpragmaticに行うインフラストラクチャは我が国ではいまだに不十分である[7]．製薬会社主導で(現時点での我が国の)GCP準拠で大規模なpragmatic試験を行うことはコスト上極めて難しい．そこまで厳しい質を求める必要もない．また併用も含んだ治療法選択といった目標からは，1社スポンサーによる臨床試験には無理がありconflict of interestsの観点からも望ましくない．一方，研究者主導でこれを行うには，研究費獲得・研究者のインセンティブ・施設の体制・データセンターおよびデータマネージャなど支援組織と人材の点でまだまだ問題が多い．市販後の研究者主導研究の体制強化とともに，研究の中立性を保ちつつ財政上の支援を可能と

するシステムが必要であると考える．前節で紹介した高度医療評価制度がこの打開策の一つとなる可能性はあるものの，'確立'までにはまだ時間がかかりそうである．

3. なぜ疫学・臨床試験研究が日本で育たなかったか？

a. 臨床試験に何が必要か？

初めにも述べたように，臨床医学系学会ではEBMという言葉が1990年代に大流行した．これまでの医療科学に対する新しいパラダイムであるという熱狂から，反感・誤解に至るまで様々なレベルの理解と態度が存在したが，現在では医療科学の重要な視点として一定の評価を得たものと位置づけられる．EBMとは，目の前の患者の問題点を一定の手順で定型化し，主に文献検索と抽出された文献の批判的吟味により過去の'証拠・根拠'を点検し，そこから有効な情報を引き出し，目の前の患者に対して実践することである，とされている．そして既に述べたように，最も強力なエビデンスを提供するのが臨床試験(特に長期大規模試験)そして複数の臨床試験を統計的に併合するメタアナリシスである．

b. 我が国の臨床医学は，このエビデンス作りにどれだけ貢献してきたか？

循環器系疾患の薬物療法に関して大きな貢献を果たした288の大規模臨床試験研究をコンパクトにまとめた'循環器メガトライアルThe State of the Heart'(エルゼビア)の編者，松崎益徳山口大学教授は，引用した日本からの論文0という事態に対し，'本邦から報告されたものが皆無であるのは実に残念であり'と述べておられる．これは実は循環器に限らない実態であり，乳癌術後補助療法のメタアナリシスEarly Breast Cancer Trialists' Collaborative Group(EBCTCG)の133試験中5試験，症例数75,000例中6,239例(1992年当初)という貢献は極めて珍しい例である[10]．

c. 何がそうさせたか？

'臨床研究を行ううえで我が国にないもの'という問いかけに対し，西條長宏氏(当時は国立がんセンター中央病院，現在は同センター東病院)はこう答えられている(1999年The 1st US-Japan Workshop on Clinical Trialsの講演から)．'Everything!'．項目を挙げれば具体的には以下のようになろう．

(1)医師研究者を支援するコーディネータ(CRC/リサーチナース)

(2)プロトコルを書くことのできる医師研究者

(3)生物統計学者(試験統計家)あるいは生物統計学の教育

(4)効率的なデータマネージメント・システムと中立的なデータセンター

(5)申請書類や研究論文を執筆するメディカルライター

(6)監査などによるデータの品質保証体制

(7)医師研究者主導型の臨床研究に対する法的規制，特に薬剤提供のシステム

(8)国民に対する臨床試験の意義の理解と患者参加に対するインセンティブ

(9)臨床試験に参加する医師研究者，支援スタッフへのインセンティブ(臨床試験研究の学問的地位)

(10)研究資金，特に公的研究資金の供給と評価

保険制度とのすり合わせなどの臨床研究に対する国家の態度，医療機関の体制，そして国民の意識というように，臨床試験のインフラストラクチャにかかわる問題は根が深い．国際化の中で抜本的な変革が必要とされており，ようやく我が国でも，先にあげた活性化計画のように文部科学省・厚生労働省・経済産業省一体となった治験・臨床試験振興策が具体化されようとしている．このようにようやく芽生えつつある機運をどう成長させるかが，職種・専門を超え臨床試験にかかわるものすべてに対して与えられた(行政用語なら)喫緊の課題である．

d. 必ずしも理解されていないランダム化

臨床試験にかかわる生物統計学の最大の貢献は，疑いなくランダム化の導入である．しかし'ランダム化'が正しく理解されているかは，医療関係者に対してさえ，更に現時点でさえはなはだ疑問である．またランダム化が医師研究者になかなか理解されなかったことが，(大規模)

臨床試験実施の最大の要因であったことも間違いない．ここでランダム化について若干補足しておこう．

臨床試験におけるランダム化とは，治療法選択において恣意的な医師の選択あるいは患者希望を避け，複数の治療介入のいずれかに確率的な要素を伴って患者を割り付ける操作である．これによって，アウトカムに影響する患者背景や病態が治療群間で偏ることから生ずるバイアスを減らし，あるいはバイアスを平均的に除去し，正確な治療法間の比較がなされるとされる．しかし，その数学的な意味が十分理解されているとは限らず，またその普及も必ずしも順調ではなかった．このことは，医療関係者の中でいまだにランダム抽出との区別がしばしばなされていないこと，臨床試験の教科書でさえも統計的推測の基盤としてランダム抽出を説明していることが多いことに見て取れる．Senn[11]の教科書（Glossary）の強烈な批判からも，上記の混同・不徹底が広く存在することがうかがえよう（以下引用）：ランダム抽出は・・・多くの統計理論において理論上重要な概念であるが，臨床試験においてはほとんど現実味をもたない．私見であるが，不用で誤解を招く概念．

ランダム化が初めて提唱されたのは，1930年代にイギリスのRA Fisherが，新種や肥料を評価する農事試験を対象として創案した‘実験計画法’においてである．この実験計画法は1947年に英国Medical Research Councilによって開始された結核患者に対するストレプトマイシンの評価に採用され，その成功を通じ，臨床試験においても方法論としての有効性が確立されたとされている．しかし，1980年代からランダム化試験の重要性を訴える我々統計家に対して，日本の臨床家から投げかけられた言葉は以下のようであった．

80年代　‘がんでランダム化試験ができるか’
90年代　‘外科でランダム化試験ができるか’

今日の我が国では，胃癌における脾摘の有無，大腸癌における開腹・内視鏡手術の比較など，がんの分野では外科手術に至るまでランダム化試験が普及しエビデンス発信に貢献を果たしている．隔世の感がある．現在の臨床家の疑問は

00年代　‘しばしば小規模で患者を限定したランダム化試験の結果が，あるいは海外のランダム化試験の成績が，目の前の患者に適用できるのか’

という一般化可能性の疑問に進化をとげている．実は欧米でもランダム化試験の定着はそれほど古いものではない．N Engl J Medに錚々たる統計家が連名で発表した1976年の特別レポート（Byarら[12]）において，（歴史対照の活用やCox回帰などの背景因子調整の統計解析手法が現れても）ランダム化は最も信頼できる評価法として存続する，という宣言がなされているほどである．

ランダム化の驚くべき特長は，現在の用語を用いれば反事実的（counter−factual）想定の下で‘実際は治療を行わない治療効果も，（個々の患者ではなく）全体の平均ならばバイアスなく正確に推定できる’という点である．すなわちある患者に治療Aを行ったときの仮想的反応をX_A，治療Bを行ったときの仮想的反応をX_Bとする．1人の患者にはいずれかの治療しか行わなければ$X_A - X_B$をすべての患者に対して観測することは不可能である．しかしランダム化を行えば，$X_A - X_B$の集団平均はバイアスなく推定できるのである．‘バイアスがない’とは，全体をA群とB群に分ける組み合わせパターンによって平均値の群間差はばらつく（これを並べ替え分布と呼ぶ）が，すべてのパターンにわたってこの分布の平均値をとれば真値に一致する，という意味である（より具体的説明については大橋・荒川[13]などを参照されたい）．患者数を大きくすれば大数の法則によりバラツキは相対的に小さくなり，ほとんど確実に真値に一致することとなる．この事実から，ランダム化を重視したITT解析が検証的解析の中心となるのである．ランダム化に対するこのような理解と，ランダム化によって初めて検定のp値が計算可能になる，観察研究においてはp値の正当性は保証されない，という数学的構造は医療関係者にはほとんど浸透していないように思われる．

表2 ランダム化臨床試験結果の一般化可能性を高めるために

◆患者背景の解析
　結果の差異の説明
◆サブグループ(部分集団)の解析
　部分集団における効果の差，交互作用の検討
◆再試験
　FDAの方針
◆メタアナリシス
　複数の独立な研究結果の統計的併合
◆緩い選択条件
　large-scale-randomized-evidence
◆ITT(Intention-to-treat)解析

4. 臨床試験のパラダイムシフト

a. 個別化治療に向けて

統計家の立場からすれば，他の応用統計の分野と異なる臨床医学統計の究極の課題，あるいは醍醐味は，一般化と個別化の'行きつ戻りつ'であろう．population pharmacokineticsに代表されるpharmacokineticsの統計解析の分野はこのよい例である．

次の挑戦課題は薬物の作用点・受容体以降の個体差に対してである．

治療法評価のための標準手順であるランダム化とは，患者の個体差は認めたうえで公平な(valid)判定の場を人工的に構築する実験技法であり，これにより集団としての，あるいは平均としての治療法の評価が可能となる．この結果を個別の患者にどう適用するかは医療のアートの部分と見なされてきたが，これを問題解決技法として体系化しようとしたのがEBMであろう．ランダム化試験の対象は，特に同意という選別プロセスを経るがゆえに限定された特殊な患者集団であり，この結果がどれだけ素直に(躊躇なく)患者一般あるいは目の前の患者に適用されるかは，統計学においては一般化可能性(外的妥当性，generalizability)の問題と呼ばれてきた．**表2**に一般化可能性を検討する，あるいはこれを高める方法をまとめた．メタアナリシスあるいはRichard Petoに代表されるイギリスの統計家がsimple and large scale evidenceと呼ぶ適格条件をゆるくした大規模試験は，平均的な効果に対する結論の頑健性を高めようとする方向であり，個別化治療とは，逆にサブグループ解析などにより治療の予測因子を摘出し，これにより治療方法は個別化するものの全体として治療効果を上げる，あるいは無駄な副作用を避けようとする方向である．

以下に，臨床試験の個別化医療へのパラダイムシフトを実感する体験を紹介したい．先にもあげた乳癌術後補助療法メタアナリシスEBCTCGの会議の顛末である．

EBCTCGは，Richard Petoの主導のもと，世界中の研究者から乳癌補助療法のランダム化試験の個票データを集め，これをメタアナリシスの手法により解析し，1992年のLancetの論文以来，1995年，2000年と5年ごとにデータ更新・解析・出版を行い，乳癌術後補助療法の重要エビデンスを提供してきた．ところが2005年の会議の直前(2005年7月5日)に，オックスフォードから全参加予定者にショッキングなメールが送られた．EBCTCGが存亡の危機に陥っているというのである．2000年の解析結果がようやく2005年に発表されるという出版遅れが生じ，また，データ提供を行ってきた研究グループの要求に沿った解析に，オックスフォードの統計グループがなかなか応じようとしてこなかった．そこで不満が爆発した．世界を代表する研究グループが，連名で'このままならデータ提供を行わない'という最後通牒を突きつけたのである．彼らはこれまでのEBCTCGの功績は認めるものの，以下のような主張を行った：

・メタアナリシスを必要とした昔と異なり，いまや試験は大型化し，単独で答えを出せるようになってきている．
・そして集団全体に対する最適な治療というより，(分子マーカーによって定まる)生物学的に意味のあるサブグループに対して最適な治療法を決定する時代にきている．
・trastuzumab(ハーセプチン)や，2005年のASCO(アメリカ臨床腫瘍学会)で進行・再発乳癌に延命効果が証明されたbevacizumab(アバスチン)など，今や有効な治療法が続々

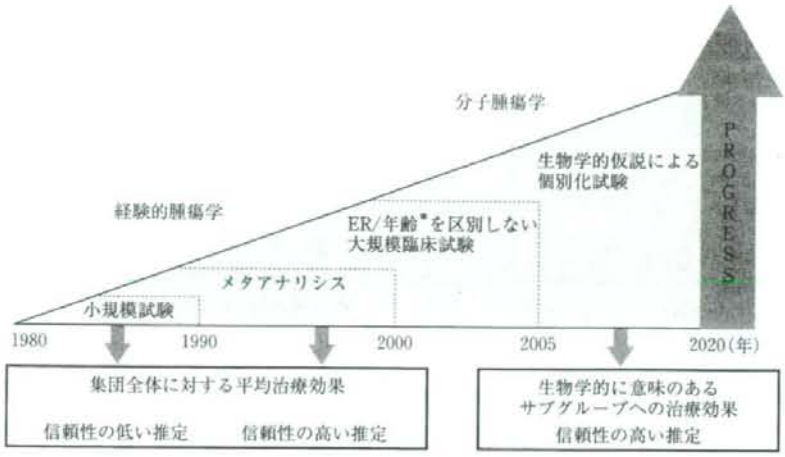

図1 大規模臨床試験・メタアナリシスのパラダイムシフト

登場している．

・これまではオックスフォードに頼ってきたが，自分達でメタアナリシスを行い重要な解答が得られるようになってきた．

一騒ぎの後，データ提供者（会議参加者）へのアンケートにも基づき，臨床家・統計家からなる国際的Steering Committeeが結成され，オックスフォード統計家グループと協力・連携しながらメタアナリシスが進められることとなった．2005年の会議ではデータ提供の遅れにより不十分な会議となってしまったものの，2006年6月のSteering Committeeを経てデータは再び着実に収集されるようになり，2006年9月11日から体制を一新しての会議が無事開催されることになった．具体的には，Petoをはじめとするオックスフォード統計家が一方的にプレゼンテーションするこれまでの形式に変わり，臨床家が解析結果を受けて解釈のコメントを述べる形となった．

図1に，上記の議論の背景となった乳癌補助療法臨床試験の歴史的展開とパラダイムシフトを挙げた．1980–90年代には統計的な詰めが甘く症例数の小さな試験も多数行われ，集団全体に対する平均的な治療効果を精度高く推定するためメタアナリシスが必要とされた．1990年代から2000年代には，統計的にきちんと設計され症例数は多いものの年齢やER（ホルモン感受性）などの重要な患者背景に制約を置かない大規模試験が数多く行われた．そしてtrastuzumabなど分子標的薬の登場，DNAマイクロアレイに代表される分子生物学の知見の活用の時代となり，生物学的に意味のあるサブグループを標的とした臨床試験が設計される個別化治療（評価）の時代となった．例えばあらかじめ標的集団を想定して，標的集団で強い効果が証明されるか，全体である程度の治療効果が証明できればよし，とするデザインや，マーカー閾値設定を検証的に行う第III相試験のアイデアも議論されるに至っている[14]．このような標的を限定した臨床試験結果も将来はメタアナリシスの対象となり，更に細かくサブグループ解析されることにより，新たな仮説を生み出すことになろう．

b. 疫学研究と臨床試験との接近

臨床試験においては，内的妥当性が確保され，正しく実施されるかぎり治療効果に対する正し

い，すなわちバイアスを含まない推測が可能となる．しかしその外的妥当性(一般化可能性)は必ずしも保証されず，EBMの最終ステップである'目の前にいる患者'に適用できるかについては，治療者のある程度の恣意的判断を免れえない．経済評価の実施も臨床試験結果単独からでは不可能であり，対象患者の全体像(特に発症率と有病率に代表される疾患負荷の見積もり)に対する疫学データが不可欠である．薬剤の製造販売後の安全性に関しても様々なデザインによる薬剤疫学的研究が臨床試験結果を補完する．

一方疫学研究においては，最新の解析技術を駆使したとしても完全な正確性の確保は不可能であり，当然，薬剤の治療効果に関しては，疫学研究成果のみで十分なエビデンスとしてはなりえない．更に，倫理的見地から薬剤の有無をランダム化する長期大規模試験がどのような臨床的課題に対しても実施可能とは限らず，実臨床でのアウトカム評価(経済性や患者QOL・満足度)を疫学調査で実施することが必要とされる場面も多い．冠動脈疾患(CHD)あるいは脳卒中の最大の危険因子が高血圧から糖尿病に移行しつつあるように，疾病像と危険因子の相対的重みは時代とともに変遷し，これをとらえるには継続的な疫学研究が必須である．臨床試験を効率化するために，患者の生命予後・QOLを高い精度で予測する代替エンドポイント(マーカー)が重要であり，この確立のために疫学研究は重要なエビデンスを提示する．

このように，疫学研究と臨床試験研究の協調が特に循環器分野で強調されるに至り，臨床試験結果をまとめたwebサイトに加え疫学研究成果のサイト[15,16]も開設されるに至っている．これは健全な方向であり，今後の予防・治療ガイドラインの充実を目指して，公的，企業資金を問わず中立な研究実施のための研究費と研究基盤整備が疫学研究にも及んでいくことを期待したい．

5. 臨床試験結果発表の標準化と臨床試験の登録

a. CONSORT

Consolidated Standards of Reporting Trials (CONSORT)は，臨床疫学者・統計学者・主要臨床医学雑誌の編集者のグループInternational Committee of Medical Journal Editors(ICMJE)によって作られたランダム化臨床試験報告のための国際標準であり，1996年の発表，2001年の改訂版発表を通じ，現在では150を超える主要臨床雑誌で採用に至っている．その活動を伝えるwebサイト[17]を通じて，例題を含むより詳細な解説文書も入手可能である．CONSORTは，対象患者の内訳を示すフロー図と22項目のチェックリストからなるが，後者に対しては日本語版PDFも入手可能である[18]．CONSORT自体は，2群並行群ランダム化試験の執筆・レビュー・評価報告の標準形式であるが，この主な考え方は，他のデザイン，例えば多群試験やクロスオーバー試験，ランダム化を行わない単群の研究，観察研究にも応用可能である．更にこの標準化の普及活動は他の分野にも及んでおり，現時点で，QUOROM(ランダム化試験のメタアナリシス)，MOOSE(観察研究のメタアナリシス)，STARD(診断技術)，STROBE(疫学研究)の4種の標準報告形式が提案されている．いずれも同じwebサイトを通じて参照可能である．

CONSORTが作られた背景は我が国と同様である．すなわち，不適切な報告は偏りのある結論の普及により非倫理的な医療につながりかねないにもかかわらず，また永年の専門家の警鐘・教育にもかかわらず，臨床試験の報告の質がなかなか上がらない，という状況の存在である．成立に至る歴史的な経緯は省略するが，この活動は目覚ましい成果を達成しつつある．多くの主導的学術団体の支援とリストが困難なほどの雑誌の採用を経て，デファクト標準化がなされつつあるといって過言でない．またこの普及の成果を示すエビデンスも集積されつつある．

報告様式の標準化，しかも高い質の要求は，試験デザインに高い質を要求することと同義で

ある．製薬会社の行う治験報告形式の国際標準化はICH-E3ガイドライン[7]によってなされたが，これが研究計画書（プロトコル）のガイドラインとして読まれている実態と同様である．Garbage in，Garbage out（ゴミを入れてもやはりゴミ）の原則はここでもあてはまる．

臨床試験のデザインは，多数の専門家による共同作業である．当該分野の臨床実態とニーズ，そして試験治療の医科学的側面に通じた臨床科学者（一人ですべてをカバーすることができるとは限らない），方法論専門家としての生物統計家，臨床薬理的検討を行う試験ではその専門家，製薬会社主導の治験の場合には開発と薬事の担当者，データマネージャ，文章としてプロトコル（試験計画書）をまとめる専門家であるメディカルライター，時には試験現場に通じたコーディネータ，等々が関係するプレイヤーである．プロトコルの科学的側面の主担当は臨床科学者と生物統計家であり，この担当部門のチェックリストを簡潔にまとめたものがCONSORTである．本標準形式の普及が，臨床家の臨床試験方法論に対する認識向上・統計家との対話の円滑化を通じ，試験の質向上に貢献することをこれからも期待したい．

b. JAMAの宣言と臨床試験の登録

2008年の1月初め，臨床試験結果をJAMAに投稿する際の規定について，興味深い指針が提示された[19]．公正に臨床試験結果が発表されるための条件と考えてよい：

(1)ICMJEが公認するサイトに臨床試験が事前に登録されていること（我が国では国立大学病院医療情報ネットワークUMINが公認サイトとして掲載されている）．

(2)論文はICMJEが策定した臨床試験報告の標準様式CONSORTに準拠して書かれていること．

(3)製薬会社がスポンサーあるいは資金提供者の場合には，中立な立場にあるアカデミアの統計家によって解析がなされていること．

(1)および(2)はこれまでのJAMAの指針の確認であり，特に目新しいことではない．驚いたのは(3)である．統計解析によるバイアスが米国でそれほど深刻にとらえられていることに統計家として驚いた次第である．当然この場合のバイアスとは，自社の製品にとって解釈が有利となるように論文の表現を誘導することである．具体的には，たくさんの結果（主たるエンドポイント，副次的エンドポイント，それらのサブグループ解析）から有利な結果を選んで提示する，逆に不利な結果を隠す，あるいは事後的な解析を（いろいろ探索的に行って有利なものを）追加して提示することになる．さてこの宣言は，日本の大学における生物（臨床）統計家のポスト増につながるであろうか．

最後に臨床試験登録制度について簡単に紹介しておこう．

臨床試験の資金提供と学術的独立性の問題は，アメリカにおける資金提供が公的なもの（主にNIH）から私的（つまり製薬会社）なものに移行するにつれ問題視されるようになった．ありていにいえば，自社製品にネガティブな結果の出版を製薬会社が妨害するという事態[20]である．これに対しICMJEは，著作権には説明責任と独立性が含まれるとして，資金提供者あるいはスポンサーが単独でデータをコントロールしたり出版を許可しないような状況で行われた研究は拒否するとの声明[21]を発表し，conflict of interestの開示を行うことも併せて要求することとなった．しかしこれでも，ネガティブな結果は公表されにくい'出版バイアス'の問題は解決されなかった．ICMJEの最後の手段がJAMAの条件(1)にある'臨床試験の事前登録制'であった．

臨床試験事前登録は，2004年9月にNEJM，Lancet，JAMA，Annals of Internal Medicineなど（後にBritish Medical Journalが追随）の主要一般臨床医学雑誌が行った（我々にすれば青天の霹靂の）宣言に始まった．すなわち，2005年7月1日以降に症例登録が開始される臨床試験については，一般国民が無償で検索できる非営利の団体が運営するサイトに当該臨床試験が登録されていないかぎり，上記の雑誌は投稿を受け付けないことが宣言されたのである．

このいわば非常事態に対するため，我が国で

表3 試験結果を正しく理解するためのチェックポイントは

◆エビデンスとしては，CONSORTを参考に・・・
　ランダム化は適切か？
　盲検化されているか，オープン試験ならその影響は？
　症例数設定根拠は妥当か？
　最初の仮説は？　探索的解析が強調されていないか？
　　非劣性試験での閾値設定は妥当か？　多重性の調整は？
　　実薬対照ではその種類と用量選択は妥当か？
　　多重エンドポイントの問題は？
　部分集団解析の解釈は適切か？
　　特定部分集団の結果が強調されていないか？
　　効果は一様か，あるいは予測因子(交互作用)は存在しないか？
　追跡は十分か，意味のある脱落(informative censoring)の影響は？
　p値のみではなく，効果の大きさの臨床的解釈は？
　　そもそも効果を一つのパラメータに縮約できるか？
◆安全性情報は？
　しばしば情報が少ない，グレーディングと標準化は？
◆デザイン論文との整合性は？　(もちろん登録は前提)
◆conflict of interest
　資金提供者の影響は？　対照薬と用量選択は適切か？
　解析は誰が？
◆データマネジメントと品質保証は？
◆EBMの課題'目の前の患者に使えるか？'の判断のためには，更に・・・

UMINが登録を開始し，その後，種々の政治的な調整から，製薬会社主導の治験に関してはJAPIC，医師主導治験に関しては日本医師会の治験支援センターが登録を開始した．その後，これらを統一して検索できる一般国民向けのポータルサイトが保健医療科学院に開設された．UMINはICMJEに公認され，今回のJAMAの指針においても日本のサイトとしてUMINが紹介されることとなった．なお臨床試験登録制度についての最初の著書[22]には，著者を含むUMIN関係者による日本の状況に関する1章が設けられている．

6. 臨床試験結果の解釈

軽から中程度の高脂血症患者に対するプラバスタチンの予防試験MEGA Study[23]は，アメリカ心臓病学会AHAでの発表以来，我が国で実施された臨床試験の解釈としては未曾有の多くの議論を巻き起こしてきた．すなわち

・MEGA Studyは2重盲検を採用していないので，バイアスが入っているのではないか．2重盲検の海外のエビデンスに比べ質が落ちるのではないか？
・MEGA StudyではITTの原則に従うといっておきながら，割り付け後に除外を行っている．恣意的に結果を薬物群に有利なように解釈しているのではないか？
・MEGA Studyのエンドポイントのうち統計的に水準に達していないものがある(全期間での脳卒中，女性の結果など)．これらの点では試験は成功していないのではないか．その治療意義は確立していないのではないか？
・MEGA Studyの結果のNNTは119である．治療効率は良くないのではないか？

これらはいずれも臨床試験方法論とその結果の一般化(実臨床上の位置づけ)に関する疑問である．繰り返しになるが，一つの臨床試験の結果がこれほど熱心に議論されたことは日本の臨床研究史上初めてであったと思われる．このような議論を通じ，臨床試験方法論と臨床試験自体の意義に対する理解が深まったのではないかと考える．MEGA Studyあるいは大規模臨床試験の最大の貢献は，あるいはここにあるのかもしれない．

表4 バイアスの要因とMEGA Studyでの対処

重要なバイアス	MEGA Studyでは
・患者選択 解釈と一般化可能性	患者背景は一般診療を反映
・割り付け(ランダム化)	中央登録と厳密な適用
・治療過程	試験に関する考え方の違い*
・試験継続・中止の判断	基本的には群間で同様
・評価(情報バイアス) 盲検化で対応	評価は盲検化(PROBE)
	血管再建は議論の要ありか?
・交絡 ランダム化と一部解析で対応	厳密なランダム化,補助的解析
・データ管理	独立なデータセンター
・統計解析	事前の解析計画書策定

* explanatoryかpragmaticか

上記の問題についての著者なりの回答は,既に大橋[6]に詳しく与えてある.ここでは**表3**に,臨床試験(特にpragmaticな大規模臨床試験)を解釈するうえでのチェックポイントを示しておこう.

a. 方法論のチェックポイント

方法論上のチェックはCONSORTに従えばよい.大橋[6]にも述べているように,コンプライアンスと併用治療の決定を含めた治療系全体を評価するためには,2重盲検試験ではなくオープンの試験を行うことにも意味がある(用量変更や中止・他剤への変更も含めた治療戦略の実践的試験なら,選択可能な治療オプションが限定される2重盲検より望ましいこともありうる).オープンの,いわゆるPROBE(Prospective Randomized Open-label Blinded Endpoint)試験の最大の危惧はエンドポイント評価に対するバイアスである.心血管系試験の場合には,TIAと狭心症の診断,入院,血行再建術施行をエンドポイントとすることがしばしば問題となる.進行癌臨床試験の場合には,進行までの時間(progression free survival)が画像撮影の時期で決定され,撮影時期の決定が医師の判断に左右されることが問題となる.近年ではpatient reported outcomeと呼ばれる(QOLなど)患者自身による評価もこのバイアスを免れない.その他のバイアス要因とその(主にPROBE試験での)対処法を**表4**にまとめた.

次の大きなチェックポイントは,ランダム化が適切になされているか,探索的な解析結果を'検証結果'として提示していないか,である.'ランダム化'といいながらカルテ番号を用いるなど,割り付け結果が容易に予見可能な割り付けを行う擬似割り付けが,以前にはかなり存在した.また治療法指示書を封筒に密封し,適格患者の同意後に開封し治療法を割り付ける封筒法は,治療に直接かかわらない第三者が開封するかぎり適切なランダム化法でありうるものの,治療を担当する医師が開封する場合は危険な方法である.開封結果次第で早期脱落の決定がなされたり,封筒開封以後に患者選択がなされる可能性さえありうるからである.がん領域では,既に1990年までには封筒法の危険性は周知であり中央登録と割り付け,そして患者背景が均等になるように動的割り付け(用語集参照)を行うことが一般的になりつつあった.しかし他領域では,著者の知るかぎりでさえ,このような封筒法の採用により割り付けが均等にならず患者背景に偏りが生じた例が1990年以降に存在する.

探索的解析結果を過度に強調する典型が2次エンドポイント結果やサブグループ解析結果の強調である.これらは多重性(用語集参照)の'悪用'例であり,がん領域では(ストイックなほどに)医師研究者から批判されるものの,循環器領域ではこれまでは比較的に鷹揚に解釈されてきた傾向がある.Wangら[24]はこれまでのNEJM誌でのサブグループ解析結果提示をレビューし,問題点と今後の対応について提言を行っている.サブグループ解析はあくまで結論の

一般化可能性を検討するための手段(**表2**参照)であり，特定のサブグループでの検証を行おうとする場合には，4節で述べたように，事前の計画への組み込みが必須となる．

試験開始時点と解析時点とでエンドポイントが変更される例はまれながらありうる．アスピリンの心筋梗塞予防効果を検証するPhysicians' Health Studyがそうであった．予想以上に心筋梗塞死亡が少なく，発症をプライマリエンドポイントに変更せざるをえなかったのである．このようなプライマリエンドポイントや主たる解析手法，検証すべき仮説の変更は，正当な理由により盲検下で行われるなら許容範囲である．しかし，解析の結果そのような変更がなされたとしたら言語道断である．試験治療を対照治療と比較して劣らないことを検証する非劣性試験においては，非劣性の検証の後に優越性(有意に優れること)を検証することは第一種の過誤を損なわないため許容される．しかし，優越性が検証できないときに非劣性に言及することは，事前に非劣性の判断根拠を提示しないかぎり許されない．

データの信頼性保証はconflict of interestの観点も加え重要である．具体的には，どのデータセンターがデータマネジメントを行ったか記載のない論文はまず信用に値しない．研究資金提供者が製薬会社であることが必ずしもデータの信憑性を損なうわけでは決してない(米国においてさえ，公的研究費のみで必要な大規模研究をすべて実施することは現時点では不可能である)．資金提供者と，研究計画・実施に責任を有するスポンサー(医師主導治験においては医師，財団やNPOなどが支援する研究においてはそれらの機関)を明示し，データへのアクセス・出版への関与が資金提供者には不可能な体制を作ることで，公正に臨床試験を計画・実施することは可能である．我が国にはありがちな過度なストイズムは研究の，そして国民福祉の阻害要因でさえある．

おわりに

がん領域では，2000年頃から国際的学会において我が国発の臨床試験結果の発信が行われてきた．そしてようやく，これまで我が国では実現困難とされた循環器系大規模予防臨床試験の成果も2006年から着々と発信されつつある．これまで我が国では全く実施されてこなかったがん検診のランダム化臨床試験も，乳癌超音波検査を対象として開始された(J-START[25])．一部の循環器系の試験には前節で指摘したような問題点も見いだされるものの，我が国が臨床試験のブラックホールの状態からは脱しつつあることは事実である．繰り返しになるが，この機運を国民の理解を更に得る形で高めることが関係者すべてにとっての課題である．

[付録：重要な統計用語の解説]

1．バイアスと誤差的バラツキ

完璧な実験系あるいは測定系によって得られる真値からの系統的な誤差をバイアスといい，'不完全な'系なるがゆえに存在する，特に理由が同定できない(あるいは敢えてしない)中心がゼロとなるようなバラツキを誤差的バラツキという．両者の境界は必ずしも絶対的なものではなく，我々の知識の限界による相対的なものであるといってよい．例えばサイコロを振って出る目は，サイコロの初期位置・投げ出す初速・床の弾性係数などの情報が完全に得られるなら予測可能であるが，これらの情報は不完全であり予測方法も極めて複雑である．そこで系統的な偏りがないサイコロを系統的な偏りがないように投げ出すことによって，目の数はすべて等確率1/6で誤差的(ランダム)に出現するとみなすのである．

バイアスのない系を'正確accurate'あるいは'妥当valid'といい，誤差的バラツキの小さな系を'精密precise'という．比較臨床試験の目的は，これらの言葉を用いれば

- いかに精密に介入効果(の差)を推定するか(clarity)
- いかにバイアスなく介入効果(の差)を推定するか(comparability)
- いかに結果の一般化可能性を高めるか(generalizability)

とまとめることができる．

臨床試験において精密さを高める最大の戦略は被験者数を増やすことであり（介入効果差の推定値の分散は被験者数に反比例する），バイアスを除くための技術がランダム化そして盲検化である．

2．ランダム化

臨床試験において，新しい実験的介入の効果を，標準的介入（あるいはプラセボによる介入や無介入）を対照として正しく評価するために，被験者を両群に医師あるいは被験者の意思によらず‘確率的に’割り付けること．これにより，両群の背景因子の分布は平均的には等しくなり，（未知のものも含めて）背景因子の不均等により結果が偏る‘交絡 confounding’を平均的には避けることが可能となる．また，事前に定めた方法により被験者を確率的に割り付けることにより，発生する割り付けパターンの種類とその発生頻度は予測可能なものとなる．これに基づいてp値の計算が可能となる．3群以上の場合への拡張はもちろん可能である．実験群と対照群の割り付け割合は，効率最大の観点から通常は1：1とされるが，実験治療に関する経験を増やしたい場合などに，2：1などの非均等割り付けが用いられることがある．

被験者が登録されるごとに同一確率で割り付ける‘単純な割り付け’は，少数例では被験者数の不均等が生ずるため実際にはほとんど用いられず，研究参加施設内である一定数からなる被験者数のブロック（群数の整数倍の大きさ）を複数用意し，それぞれのブロック内で割り付けを行う置換ブロック法が，盲検を伴う治験の標準的方法である．予後因子の影響が大きく，かつ盲検化がしばしば困難なため中央登録で割り付けがなされるがんの臨床試験においては，予後因子の分布を登録ごとに計算し分布の偏りが小さくなるように割り付ける‘動的割り付け法’が標準的方法である．単純な割り付け，動的割り付けいずれにおいても，まず被験者を重要な背景因子で層別し，それから割り付けを行う‘層別割り付け’を行うこともある．

3．盲検化

ランダム化によって割り付けられた介入特に薬剤の種類を，被験者や実際に治療を行う医療従事者，および治療効果の評価者が知りえないようにすることで，それぞれの思い込みが引き起こすバイアスを減らす技術のこと．マスキングともいう．

被験者のみを盲検化するのが単盲検，被験者と（しばしば評価者を兼ねる）治療者とを盲検化するのが2重盲検法であるが，実際にはデータを取り扱う担当者などにも盲検化が保持されるので，2重盲検法は3重以上の盲検となっている．手術と放射線治療あるいは毒性の強い抗がん剤治療の臨床試験のように，技術的あるいは安全性と倫理の観点から被験者・治療者への盲検化が困難な場合，あるいは市販されている薬剤を用い実際の治療環境の中で併用法も含めて治療指針全体を試験する実践的試験の場合には，結果評価の際に群（や施設名）を盲検化するPROBE（Prospective Randomized Open-label Blinded-Endpoint）法が用いられることが多い．

4．プライマリーエンドポイント

検証的な臨床試験において，検証の対象となる（通常は有効性の）エンドポイントで，通常は1つに設定される．必要症例数はこの検証をめざして設定される．その他のエンドポイントはセカンダリエンドポイントと呼ばれる．

5．ITT

intention-to-treatあるいはintent-to-treatの略で，割り付けた被験者は（たとえ早期中止やコンプライアンス不良でも）その割り付け群として解析する方針．検証的な第III相試験では，ITTを解析方針として採用するのがふつうである．ランダム化によって達成された比較可能性を，被験者除外によって崩さないためである．実はITTには分野によって若干の解釈の幅があり，2重盲検ランダム化を採用する治験において，より具体的にITTを規定したのがFAS（Full Analysis Set）の概念である．そこでは，割り付けられたものの投薬が行われなかった被験者，割り付け以前の情報によって不適格が確認できる被験者（ただしすべての被験者に対して

同じように適格性チェックがなされることを前提とする)，試験開始後の情報が全くない被験者は解析除外してもよいこととなっている．

6．検定の多重性

介入結果に差がないという仮説が正しいもとで'統計的に有意'という偽陽の結果が得られる確率は，採用した有意水準に一致し，通常は0.05(片側0.025)である．これを第I種あるいはα過誤と呼ぶ．一方，仮説が正しくなく実験的介入が優れている場合に，誤って'有意としない'偽陰の結果が得られる確率が第II種あるいはβ過誤である．このα過誤は，仮説検定をただ一つ行う場合に保証される誤り確率であり，複数の仮説を同時に設定し，そのいずれかが成立しているときに'有意差あり'と推論する場合には，誤った判断を行う'全体の'確率は上昇してしまう．エンドポイントを多数設定しそれぞれ検定したり，重症度など被験者背景によってサブグループ(部分集団)をたくさん設定しそれぞれに対して検定を行ったり，長期臨床試験や継時データを扱う臨床試験の場合に複数の時点で検定を繰り返したり，複数の検定手法を同時に適用することなどが，このα過誤を上昇させる'多重性 multiplicity'の例である．ICH-E9において多重性は厳しく指弾されており，検証的な臨床試験においては，多重性を考慮したうえでα過誤を伴う推測方式を採用しなければならない．

具体的には

- 多エンドポイント：プライマリエンドポイントとして一つを選択する，あるいは複数エンドポイントを一つに合成する，あるいは検定の順序を決める．
- サブグループ：サブグループの解析結果を検証とみなさない，あるいはあらかじめ興味のある部分集団を設定し，部分集団と全体と2つの検定を行うことを考慮して有意水準を設定する．
- 長期臨床試験において中間解析を行う場合には，全体として有意水準を保つよう，中間での評価を厳しくする．
- 継時データ：主たる評価時点を設定する，あるいはAUC(Area Under the Curve：曲線下面積)など要約統計量で比較を行う，あるいは継時データ分散分析の方法により全時点に対する平均的な効果を評価する．
- 複数解析手法：事前に策定する解析計画書において，主たる解析手法を指定するなどの方策がとられている．

7．率・割合・比

率 rate，割合 proportion，そして比 ratio は日本語の俗語ではすべて比率と表現されごっちゃに用いられるが，疫学あるいは統計学では厳密に区別される基本概念である．

異なる物理量を割り算して得られる指標が率であり，例えば体重/(身長)^{2}で与えられるBMIや自動車の時速は率の例である．単位は当然，(分子の単位/分母の単位)であり0から無限大の値をとる．割合はある特徴をもった個体が全体のどれだけになるかの指標であり，単位は無次元，0から100％で表されることが多い．比は同じ次元をもった物理量の比であり，単位は無次元，0から無限大の値をとる．生存時間解析あるいは疾病発生の疫学においては，ある疾患(あるいは死亡)の起きる時間あたりの速度である発症率(死亡率)は率，ある時点での累積発症率(累積死亡率)は割合，発症率(死亡率)の比であるハザード比あるいは累積発症率(累積死亡率)の比であるリスク比はいずれも比にあたる．

8．ハザード比

これまでに発生していない疾患がある短い時間間隔中に起こる速度を'罹患率 incidence rate あるいは morbidity'という．同様に死亡がある短い時間間隔中に起こる速度を'死亡率 death rate あるいは mortality'という．いずれもこれまで発生していない事象が発生する速度であることから，これらをまとめて数学的にはハザードといい，これらを2群間で比を取ったものがハザード比である．当然ハザード比が1なら群間に差はなく，治療群の無治療対照群に対する疾患発生ハザード比が1を下回れば，予防効果がみられたことになる．

数学的には説明はやや厄介であるが，死亡ハザードλが時間に対して一定であれば，時間7