Wu, M. C. and Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics* **44**, 175-188.

Xie, F. and Paik, M. C. (1997). Multiple imputation methods for the missing covariates in generalized estimating equation. *Biometrics* **53**, 1538-1546.

付　　録

A　重み付き推定量の漸近分散

欠測メカニズムがMARのもとで，式(8)における回帰パラメータβの重み付き推定を考える．βの解は，以下の重み付き推定方程式を解くことで得られる．

$$U(\beta,\hat{\gamma}) \equiv \sum D_i^T(x_i\,;\beta)V_i^{-1}(\hat{\alpha})w_i(\hat{\gamma})[y_i - g(x_i\,;\beta)] = \sum U_i(\beta,\hat{\gamma}) = 0$$

ただし，$g(x_i\,;\beta) = E(y_{it} \mid x_i)$, $D_i(x_i\,;\beta) = \partial g(x_i\,;\beta)/\partial\beta^T$, $V_i(\hat{\alpha})$はその要素が(9)式から計算される作業共分散行列，$w_i(\hat{\gamma})$は重み行列（対象者が各時点で観察される累積確率の逆数）である．

(10)式のような重みを推定するためのモデルが正しければ，上記の重み付き推定方程式から得られるβの解$\hat{\beta}$は一意に定まり，$U(\beta_{true},\hat{\gamma})$と$(\hat{\beta}-\beta_{true})$は漸近的に平均がゼロで，分散共分散行列がそれぞれ，C, $(\Gamma^{-1})C(\Gamma^{-1})^T$の正規分布に従うことが示されている（Robins et al., 1995）．これらの分散共分散行列の推定値は，$\hat{\Gamma} = \partial U(\hat{\beta},\hat{\gamma})/\partial\hat{\beta}^T$, $\hat{C} = \hat{A} - \hat{B}\hat{\Omega}\hat{B}^T$から求めることができる．ただし，$\hat{A} = \sum U_i(\hat{\beta},\hat{\gamma})U_i(\hat{\beta},\hat{\gamma})^T$, $\hat{B} = \partial U(\hat{\beta},\hat{\gamma})/\partial\hat{\gamma}^T$, $\hat{\Omega}$は重みを推定するための式(10)のモデルに対する尤度関数から計算される観察情報行列の推定値である．

上記のβに対する漸近分散は，$(\hat{\Gamma}^{-1})\hat{A}(\hat{\Gamma}^{-1})^T - (\hat{\Gamma}^{-1})\hat{B}\hat{\Omega}\hat{B}^T(\hat{\Gamma}^{-1})^T$と表現できる．この分散の第1項$(\hat{\Gamma}^{-1})\hat{A}(\hat{\Gamma}^{-1})^T$は，重みの推定誤差を考慮していない通常のロバスト分散（Liang and Zeger, 1986）である．したがって，重みを考慮した分散のほうが通常のロバスト分散よりも少なくとも効率が良いことがわかる．

B　パターン混合モデルにおける周辺平均$\sum_d p^{(d)}\theta_t^{(d)}$の漸近分散

ある時点における（脱落パターンに関して平均した）周辺平均$\hat{\theta}_t = \sum_d p^{(d)}\theta_t^{(d)}$の分散は，$\pi = (p^{(d)},\ldots,p^{(D)})^T$, $\delta = (\theta_t^{(1)},\ldots,\theta_t^{(D)})^T$とすると，$\hat{\theta}_t = \sum_d p^{(d)}\theta_t^{(d)} = \pi^T\delta$なので，デルタ法により，

$$Var(\hat{\theta}_t) = \hat{\pi}^T\,Var(\hat{\delta})\hat{\pi} + \hat{\delta}^T\,Var(\hat{\pi})\hat{\delta}$$

となる．ただし，$Var(\hat{\pi}) = diag(\hat{\pi}) - \hat{\pi}\hat{\pi}^T$である．

一般に，任意の関数$h(\hat{\delta},\hat{\pi})$に対する漸近分散は，$Var[h(\hat{\delta},\hat{\pi})] = \Omega\begin{pmatrix} Var(\hat{\delta}) & 0 \\ 0 & Var(\hat{\pi}) \end{pmatrix}\Omega^T$である．ただし，$\Omega = \begin{pmatrix} \frac{\partial}{\partial\delta}h(\hat{\delta},\hat{\pi}) & \frac{\partial}{\partial\pi}h(\hat{\delta},\hat{\pi}) \\ \frac{\partial}{\partial\delta}h(\hat{\delta},\hat{\pi}) & \frac{\partial}{\partial\pi}h(\hat{\delta},\hat{\pi}) \end{pmatrix}$である．

2) 統計解析の実際

東京大学大学院医学系研究科生物統計学　松山　裕・大庭　幸治
Yutaka Matsuyama　Koji Oba
（助教授）

Summary

医学データの統計解析を行う際に，統計専門家以外のユーザーにとって最も悩ましい問題は，多くの統計手法が存在し，統計ソフトを用いれば簡単にそれらが利用できる状態にある一方で，どの手法を選択すればよいかの指針がはっきりしないことであろう。統計手法の選択を決める要因は複数考えられ，①研究の型（データ収集方法），②変数の数と興味のある関連，③目的および証明したい仮説，④想定できる前提，⑤変数の型・分布・データの質，⑥事前情報の質と量，などさまざまなものがある。本稿では，結果変数（エンドポイント）が生存時間データの場合における統計解析手法（率の算出，カプラン・マイヤ法，ログ・ランク検定，マンテル・ヘンツェル法，Cox 回帰）について紹介する。

Key Words　発生率，生存時間解析，Kaplan-Meier 法，log-rank 検定，交絡調整，Cox 回帰

Surgery Frontier 11(4)：82-89，2004

はじめに

医学研究でよく利用される統計手法を表にまとめた（表 1）。結果変数（エンドポイント）の型の観点から分類したのが表 1 である[1]。本稿では，誌面の都合上，あるイベント発生までの時間を結果変数とする場合の統計解析（生存時間解析と総称される）について述べる。生存時間解析を含むその他の手法については，医学統計の教科書[2)3)]を参照いただきたい。

率の算出

特定の個人が研究開始時に研究対象としている疾病にかかっておらず，将来その疾病にかかる可能性があり，かつ研究者による積極的な追跡を受けている状態を「at risk」とよぶ。この「at risk」の状態にある対象者が新たに疾病にかかった場合，それを疾病発生（incidence）という。コホート研究（前向き研究：prospective study）では「at risk」状態にある対象者の集団（リスク集団）から研究期間中に起きた疾病発生数を数え上げる。集団中の疾病発生の指標としては，割合（proportion）で示す場合と率（rate）で示す場合の 2 通りの考え方がある。割合とは，「分子が分母に含まれる分数」であり，単位をもたず，必ず 0 から 1 の間の値をとる量である。一方，率とは，「単位時間あたりの変化を表す比」であり，単位が存在し，0 から無限大の値をとる量である。

簡単な例で両者の違いを示す（図 1）。A，B，C，D の 4 人からなるリスク集団（population at risk）を 1 年間追跡した結果，A と D に対象疾病の発生がなく，B と C はそれぞれ 3 ヵ月目，9 ヵ月目に疾病を発生したとする。このとき，疾病を発生したのは 4 人の集団の中で B と C の 2 人なので，疾病発生割合（incidence proportion）は，2/4 = 0.5 と計算できる。このように疾病発生割合とは，研究期間中にリスク集団で発生した疾病の数を研究開始時のリスク集団の対象者数で割ったも

表1　医学研究でよく用いられる統計手法

解析の目的	結果変数の型		
	二値データ	連続データ	生存時間データ
分布の記述	頻度集計 分割表	ヒストグラム 平均，標準偏差 散布図，相関	カプラン・マイヤ法
単純な群比較	カイ二乗検定 リスクの推定	t-検定 Wilcoxon 検定	ログ・ランク検定 発生率の推定
層別解析	マンテル・ヘンツェル法 調整リスクの推定	分散分析	マンテル・ヘンツェル法
回帰モデル	ロジスティック回帰 調整リスクの推定	重回帰分析 分散分析	Cox 回帰 調整発生率比の推定

のとなる。研究開始後には新たに対象者を登録しないコホートを「登録に関して閉じたコホート」とよぶが，疾病発生割合はこの閉じたコホートでなければ定義することができない。

一方，疾病発生率(incidence rate)は，観察期間中の疾病発生数をリスク集団の合計観察時間で割ることで得られる。図1の場合，リスク集団の合計観察時間は，1+0.25+0.75+1=3 人年(person-years)なので，「3 人年あたり 2」，あるいは「100 人年あたり 67」と計算できる。このような疾病発生率の算出方法は，人－時間法(person-time method)とよばれ，対象者がいつ疾病を発生したかがわからないと求めることができず，割合と異なり時間の逆数の単位をもつ。疾病発生率は閉じたコホートでも，あるいは研究開始後の対象者の登録，転出などを許す開いたコホートのいずれであっても定義可能であり，コホート研究で特に関心のある疾病発生の指標である。なお，

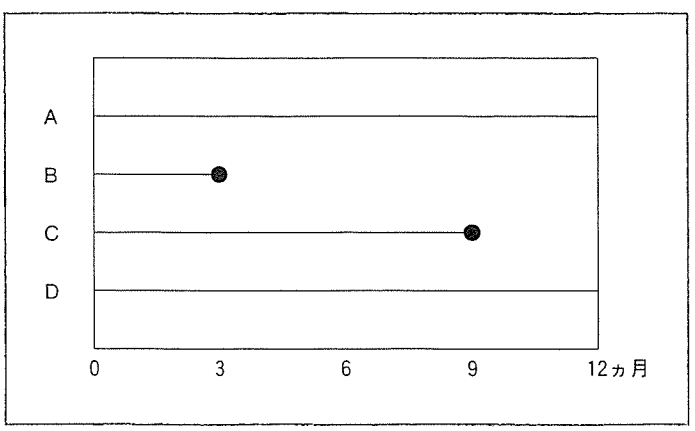

図1　4人のリスク集団の追跡例
●が疾病の発生を表す。

生存時間解析では，疾病発生率のことをハザードとよぶことが多い。

乳がんに対する内分泌療法剤であるタモキシフェンの服用と二次がんの関係を調べるために行われたヒストリカル・コホート研究の際に得られた乳がん再発に関するデータの一部[1]を示す(表2)。全国9施設の乳がん患者全例のうち1982年から1990年までの間に治癒切除手術が実施された一側性の原発女性乳がん患者が対象である。乳がん術後にタモキシフェンを服用したグループの再発率は，1,000 人年あたり 10 (手術時のリンパ節転移なし)，49 (手術時のリンパ節転移あり)であり，タモキシフェン非服用グループでのそれらは，それぞれ 14，52 である。

表2　乳がん術後のタモキシフェン服用と乳がん再発率(リンパ節転移の有無別)

リンパ節転移	タモキシフェン服用	再発数	対象者数	観察人年	再発割合	再発率
なし	あり	96	1334	9713.8	0.072	0.010
	なし	171	1592	12599.3	0.107	0.014
あり	あり	368	1215	7514.1	0.303	0.049
	なし	253	760	4862.2	0.333	0.052

time-to-event データの解析

発生率を疾病発生の指標として用いる場合，対象者が研究で関心のある疾病を発生したかどうかだけでなく，研究参加時点から疾病発生までの時間を問題としなければならない。心筋梗塞などの生活習慣病の発症，あるいは究極的には死亡をエンドポイントとした場合には，エンドポイント発生までの時間を長引かせる治療法は臨床的に意義がある。疾病を発症するまでの時間，死亡するまでの時間，再発するまでの時間といった何らかのイベントが起きるまでの時間(time-to-event)をエンドポイントとした場合のデータ解析は一般に生存時間解析とよばれる[5][6]。

生存時間解析では，イベント発生までの時間をグループ間で比較するこが目的となるが，対象者全員に対してイベント発生までの時間を観察できることはまれである。ほとんどの研究では，研究終了時点で関心のあるイベントをまだ発生していない対象者，追跡途中でイベントの発生状況を確認できなくなる対象者，関心のあるイベントの発生以外の理由で死亡する対象者などが存在する。このような関心のあるイベントの発生までの時間を定義できない対象者は打ち切り(censoring)とよばれる。特に，追跡途中でイベントの発生状況を確認できなくなる対象者のことを追跡不能(loss to follow-up)，関心のあるイベントの発生以外の理由で死亡する対象者のことを競合リスク(competing risk)による打ち切りとよぶ。打ち切りを受けた対象者は，研究参加時点から打ち切りを受けるまで「at risk」の状態にあり，それらのイベント発生時点は「打ち切り時点以降である」という情報をもっているので，解析対象から除くことはできない。

通常の生存時間解析においては打ち切りデータが存在するために，イベント発生までの時間を直接比較することはできない。しかしながら，「単位時間あたりに何例イベントが起きるか」というイベント発生の強さの指標である発生率の計算・比較は可能である。打ち切りを受けた対象者に関しては，打ち切りを受けるまでの時間を観察時間と考え，人-時間法の分母を「打ち切りを含めた合計観察時間」とすればよい。ただし，どのようなタイプの打ち切りでも許されるわけではない。通常の生存時間解析で想定している打ち切りは，関心のあるイベントとは無関係な理由による打ち切りである。例えば，研究終了時点でまだイベントを発生していない場合やイベントとは関係のない理由による転居(追跡不能)，参加拒否，死亡(例えば，偶然の交通事故)などである。逆に，打ち切りが(その後に潜在的に観察される)イベント発生に関して何らかの関係をもっている場合は情報のある打ち切り(informative censoring)とよばれる。例えば，心筋梗塞発生をエンドポイントとしている場合に，コレステロールが高くなってきた対象者ほど打ち切りを受けやすい場合や危険因子がよく似ている疾病(例えば，脳梗塞など)による死亡などである。情報のある打ち切りに対しては，単純に打ち切り例として解析を行うと結果にバイアスが生じることが知られている。最近は，そのようなバイアスを補正する解析方法が提案されつつあるが，汎用的で広く

表3　タモキシフェン服用による乳がん再発率の抑制効果(リンパ節転移の有無別)

リンパ節転移	再発率の比(95％信頼区間)	再発率の差(95％信頼区間)
なし	0.71 (0.57, 0.93)	-0.0036 (-0.007, -0.001)
あり	0.94 (0.80, 1.10)	-0.0031 (-0.011, -0.005)

受け入れられている解析方法は存在しないのが現状である。情報のある打ち切りを否定できない場合には，イベントの定義を複数行い，当該の事象を打ち切りとみなす場合とイベントとみなす場合の2通りの解析を行い，結果の整合性をみるのが現実的な対処法である。

表2のデータに対するタモキシフェン服用の再発抑制効果を示す(表3)。リンパ節転移なしのサブグループでは，タモキシフェン服用グループの方が再発率が低く(再発率を約3/4倍，あるいは10,000人年観察すると再発が36人減少)，その差は統計的に有意である(再発率比の信頼区間が1を含まない，あるいは差の信頼区間が0を含まない)。リンパ節転移ありのサブグループでもタモキシフェンを服用した方が再発率は小さいが，その差は統計的に有意ではない。なお，発生率比，発生率差の95％信頼区間の計算方法は疫学の教科書[7][8]を参照していただきたい。

人-時間法による率の算出・比較は理解しやすく計算も簡単であるが，その方法が妥当であるためには，「発生率が観察期間を通して一定である」という条件が必要である。この条件が成立している場合，人-時間法による発生率の逆数は，「合計観察時間/発生数」なので，イベント発生までの平均的な時間として解釈できる。したがって，「2グループ間で発生率が等しい」ことは「2グループ間の平均的なイベント発生時間が等しい」ことを意味し，打ち切りを含むイベント発生までの時間の比較を行っていることになる。

しかしながら，人-時間法で必要な条件「発生率は観察期間を通して変化しない」は研究開始から1年たっても5年たっても発生率は同じであることを要求しており，かなり厳しい仮定である。この仮定が満たされないと思われる場合には，観察期間を発生率が一定とみなせる期間に区切って比較を行う必要がある。一方，上の条件を緩めて「発生率は時間とともに変化してもかまわない」とすると，発生率は時々刻々と変化するので，もはや人-時間法のような安定した推定値を得ることはできなくなる。しかし，時間によって変化する発生率自体を推定することは困難であっても発生率が累積して観察される発生曲線(累積発生率)

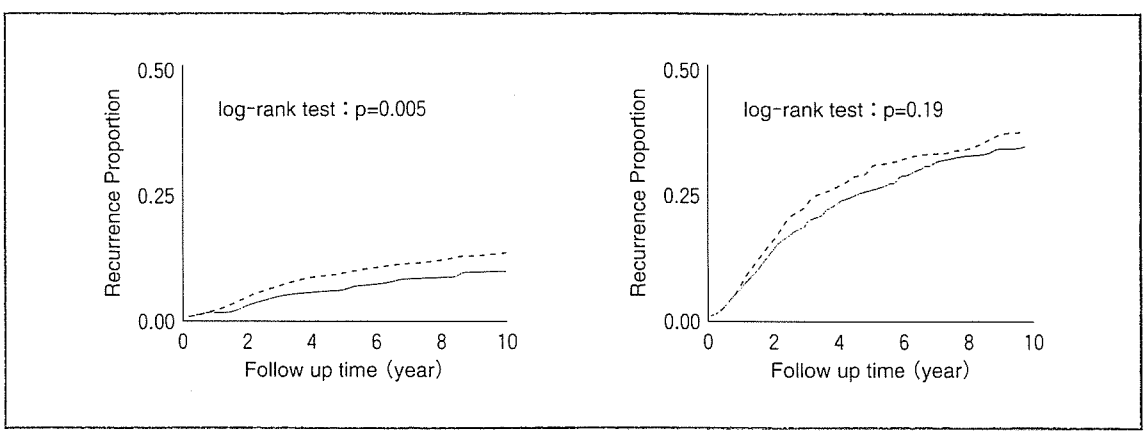

図2　乳がん再発曲線(実線：タモキシフェン服用グループ，点線：非服用グループ)
左図：リンパ節転移なし，右図：リンパ節転移あり

表4　乳がん術後のタモキシフェン服用と乳がん再発率(リンパ節転移の有無と閉経状況で層別)

リンパ節転移	閉経状況	タモキシフェン服用	再発数	観察人年	再発率	再発率比(95%信頼区間)
なし	前	あり	48	4992.7	0.0096	0.78
		なし	80	6455.6	0.0124	(0.54, 1.11)
	後	あり	48	4721.1	0.0102	0.69
		なし	91	6143.7	0.0148	(0.48, 0.97)
あり	前	あり	205	4213.6	0.0487	1.03
		なし	130	2758.5	0.0471	(0.83, 1.29)
	後	あり	163	3300.5	0.0494	0.84
		なし	123	2103.7	0.0585	(0.67, 1.07)

を推定することは比較的容易である。イベント発生までの時間の分布やイベント発生率に特に強い仮定を必要としないイベント発生曲線を求める方法がカプラン・マイヤ法(Kaplan-Meier method)である[5][6]。表2のデータに対する乳がん再発曲線を示す(図2)。リンパ節転移のあり・なしにかかわらず,タモキシフェン服用グループの方が観察期間を通して常に再発割合が低いことがわかる。

時間によって発生率が変化してもかまわないが,「2グループ間の発生率は等しい」という仮説を検定する手法がログランク検定(log-rank test)である[5][6]。図2のデータに対しては,リンパ節転移なしのグループでは,ログランク検定のp値は0.005であり,リンパ節転移ありのグループでのそれは0.19である。2グループ間での発生率が等しいかどうかの検定手法としては,ログランク検定のほかにもウィルコクソン検定を拡張した一般化ウィルコクソン検定を用いることもできる。しかし,一般化ウィルコクソン検定は研究開始直後に死亡率が異なると検定結果が有意になりやすい性質がある。このため,早期打ち切り例が多数存在すると,もし打ち切り例をもっと追跡すれば差がなくなるような場合であっても誤って差があるという結果になる可能性があり,注意が必要である。また,カプラン・マイヤ曲線を比較する際にどうしても研究期間の最後の方に目がいってしまいがちであるが,後半部分では「at risk」の状態にある対象者が少なくなっていることが多いので,生存曲線の推定も不安定である。数名のイベント発生で曲線の形状がかなり変化する場合があるので,「研究の後半で差が開いている」という解釈には注意が必要である。そのような場合,生存曲線の横軸に適当な時点ごとの「at risk」の人数を併記して,注意をうながしておくことが重要である。

交絡の調整・制御

表2のデータにおける「リンパ節転移の有無」のような変数は,交絡因子(confounding factor)とよばれる[9]。交絡因子とは,当該のイベントに対するリスク因子で,かつ比較する群間でその人数が異なるような変数である。交絡因子は原因と結果の間の因果関係をゆがめるので,バイアスのない治療効果を求めるためにはその影響を調整する必要がある。交絡の調整は,デザインでも統計解析でも可能である。前者の方法には,対象者の限定,マッチング,ランダム化の3つの方法が存在する[7]。ここでは,後者の統計解析による方法について述べる。

統計解析による交絡調整方法には,層別解析によるものと回帰モデルによるものの2つが存在する[7]。層別解析(stratified analysis)とは,各層内では交絡の影響が無視できるように交絡因子でデータを十分に層別し,層ごとの効果の指標を要約する方法である。例えば,リンパ節転移なしで閉経前のグループでは,全員がリンパ節転移なしで閉経前なので,これらの要因による

交絡は生じない(表4)。この層別したデータに対して,「治療によって疾病発生率は変化しない(疾病発生率の比が1)」という帰無仮説の検定は,マンテル・ヘンツェル検定(Mantel-Haenszel test)で調べることができる[7][8]。マンテル・ヘンツェル検定は,層ごとのカイ二乗検定を総合したような検定である。このデータの場合,カイ二乗値は4.03,p値は0.04であり,「リンパ節転移の有無と閉経状況の影響を調整すると乳がん術後にタモキシフェンを服用することにより再発率が変化する」ことがわかる。

統計学的に「有意であった($p<0.05$)」,「有意でなかった(n.s.)」などのように検定結果(p値)のみを用いて結果の議論をしている医学論文をよくみかける。しかしながら,本来の仮説検定の意義は“decision making”であり,知識の積み重ねが要求される科学研究には仮説検定は不向きである。仮説検定の考え方は重要であるが,単に治療効果が「ある」か「ない」かよりも治療効果の大きさがどれくらいかを求めることの方が,はるかに情報量が多い。このためには,適切な効果の指標(リスク差,リスク比,ハザード比など)を定め,その大きさを推定し,その信頼区間を併記する必要がある。

表4のデータに対して,マンテル・ヘンツェル発生率比を求めると,0.87(95%信頼区間:0.76,0.99)となり,リンパ節転移の有無と閉経状況の影響を調整すると乳がん術後にタモキシフェンを服用することにより再発率が約5/6倍になることがわかる。マンテル・ヘンツェル推定量は,比較的計算が簡便で,全体のサンプル数が多ければ各層内のサンプル数は少なくても妥当な結果を導くなど統計的な性質も優れているので,臨床・疫学研究で最もよく用いられる解析手法である[7][8]。

なお,各層ごとの効果の指標だけを求めて,それらの結果を要約しない方法を層別解析とよんでいる論文をよくみかけるが,それは「サブグループ解析」であって,層別解析ではない。サブグループ解析とは,リンパ節転移なしで閉経前といった特定のサブグループに対する結果を調べる方法で,事前に計画されていないサブグループでの効果の大きさを事後的に推定するためには一般に非常に大きなサンプルサイズを必要とする。

層別解析による交絡の調整方法は,数学的な仮定も少なくわかりやすいという利点をもっているが,多くの交絡因子で層別すると,層内の対象者数が少なくなったり,時には比較する2グループの一方に対象者がいる層がなくなったりということが生じる。そのため,効果の指標の推定値が不安定になる,あるいは効果の指標の推定自体ができなくなるという問題が存在する。回帰モデルに基づく方法とは,リンパ節転移や閉経状況などの交絡因子と乳がん再発との間に,数学的な強い仮定を設ける方法である。例えば,乳がん再発率の対数が術後からの時間に関係した効果,リンパ節転移の効果,閉経状況の効果の和で決まると仮定するモデルが比例ハザードモデル(Cox回帰)とよばれ,医学・疫学で頻繁に用いられるモデルである[5][6]。数式表現すると,ある対象者に対する疾病発生率(ハザードともよばれる),$h(t|x,z)$に対する以下のようなモデルである。ただし,タモキシフェンの服用の有無をx($x=1$ならあり,$x=0$ならなし)で表し,交絡因子(ここでは,リンパ節転移の有無,閉経状況,ステージの3つを考える)をそれぞれ,z_1,z_2,z_3で表すとする。

$$\log h(t|x,z) = \log a(t) + bx + c_1z_1 + c_2z_2 + c_3z_3$$

このモデルにおいて,$x=0$,$z=0$の場合の疾病発生率$a(t)$は,ベースラインの疾病発生率(あるいは,ベースラインハザード)とよばれ,時間とともに変化してもかまわない。しかしながら,このモデルでは,治療変数や交絡因子の効果は時間によらず一定(比例ハザード性とよばれる)と仮定している。この条件のもとで治療や交絡因子の効果を表す回帰パラメータ(b,c_1,c_2,c_3)をデータから求める。回帰パラメータの推定には部分尤度に基づいた複雑な計算が必要であるが,統計解析ソフトを利用すれば容易に実行可能である[6]。得られた結果の解釈は,例えば,$\exp(b)$はリンパ節転移,閉経状況,ステージといった交絡因子の影響を調整したタモキシフェン服用の再発率比(incidence rate ratio)を表している。

タモキシフェンデータに対するCox回帰の結果を示す(表5)。交絡因子の

影響を調整すると，タモキシフェンの服用は再発率を 0.83 倍(95 %信頼区間： 0.73, 0.95)にすることがわかる。この結果は，リンパ節転移と閉経状況のみを調整したマンテル・ヘンツェル発生率比の結果とほぼ同じである。また，交絡因子に関しては，リンパ節転移の有無とステージが統計的に有意な独立したリスク因子となっている(それぞれ，再発率を 3.7 倍，1.6 倍にする)。

回帰モデルによる交絡の制御方法は，多くの交絡因子を一度に調整できるという利点をもっているが，仮定されたモデルが正しくなければ結果にバイアスを生じる。モデルはあくまでモデルであり，すべてのモデルは誤っていると考えた方がよいので，強い仮定を必要とするモデルを用いた解析に過度に依存することは危険である。実際のデータ解析の場面では，単純な層別解析の結果とモデルによる解析結果の整合性を検討しておくことが重要である。

また，実際のデータ解析において交絡の調整となると，ある要因が交絡因子かどうかを判断する必要が生じる。交絡因子の見極めのためによく用いられる方法は，治療グループと対照グループの間で背景要因や共変量が t 検定やカイ二乗検定で有意であれば交絡因子と判断しようという方法である。多くの論文でこのような方法をみかけるが，この方法は誤りである。なぜなら，統計的仮説検定とは，部分から全体を推論するために必要な演繹的論理であり，帰無仮説が否定されれば「帰無仮説は誤っている」と主張できるが，帰無仮説が否定できないときには「帰無仮説が正しい」とは積極的に主張できないからである。例えば，閉経状況が交絡因子かどうかを判断するために，「タモキシフェン服用グループと非服用グループで閉経前の人の割合が同じである」という帰無仮説を考えて，割合の差の検定(カイ二乗検定と同じ)を行い，5 %水準で有意差があれば「2 グループ間で閉経前の人の割合が異なる」ことはわかっても，5 %水準で有意差がなければ「2 グループ間で閉経前の人の割合が同じ」，つまり「交絡はない」とはいえず，「交絡があるとはいえなかった」という判断になる。交絡は存在するかどうかが問題なのではなく，その程度が問題である。ある要因が交絡因子かどうかを見極める際には，まずその要因が交絡因子の必要条件(交絡因子はリスク因子である，交絡因子は比較する集団間でその分布が異なっている，交絡因子は原因と結果の間の中間変数ではない)を満たしていなければならないが，その見極めには臨床的・疫学的な常識，知識などの事前情報・過去の研究成果がなによりも重要である。

おわりに

臨床研究における生物統計学の貢献は，本質的にバラツキを伴うデータ，およびその曖昧さ・不確実さの中で得られる結果に対する信憑性を確保・評価することである。これらの信憑性を支える柱が，研究計画・データ管理・統計解析の 3 つである。いくら素晴しい統計解析手法を用いたとしても，研究計画とデータ管理がいい加減であれば，そこから得られる結論は弱いことが予想される。したがって，妥当で効率のよい研究デザインはどのようなものであり，プロトコルをどのように開発するか，調査票をどのように設計するかなどの研究計画の問題，モニタリング，データ入力，標準化とデータベース管理をどのように行うか，そしてそれらの品質保証をどうするかなどのデータ管理の問題をまずクリアすることが大事であることも忘れないでいただきたい。

表 5　Cox 回帰の結果

変数	推定値	標準誤差	再発率の比	95 %信頼区間
タモキシフェン (あり v.s. なし)	-0.181	0.068	0.83	0.73, 0.95
リンパ節転移 (あり v.s. なし)	1.300	0.076	3.67	3.16, 4.26
閉経状況 (後 v.s. 前)	0.084	0.067	1.09	0.95, 1.24
ステージ	0.452	0.056	1.57	1.41, 1.76

文　献
1）大橋靖雄：統計処理の落とし穴．Diabetes Journal 24 (4)： 15-20, 1996
2）Altman DG (木船義久，佐久間昭訳)：医学研究における実用統計学．サイエンティスト社，東京，1999
3）Armitage P and Bery G (椿美智子，椿　広計訳)：医学研究のための統計的方法．サイエンティスト社，東京，2001
4）Matsuyama Y, Tominaga T, Nomura Y, et al ： Second cancers after adjuvant tamoxifen therapy for breast cancer in Japan. Annals of Oncology 11 ： 1537-1543, 2000
5）Clark TG, Bradburn MJ, Love SB, Altman DG ： Survival analysis Part I ： basic concepts and first analyses. British Journal of Cancer 89 ： 232-238, 2003
6）大橋靖雄，浜田知久馬：生存時間解析—SASによる生物統計—．東京大学出版会，1995
7）Rothman KJ and Greenland, S ： Modern Epidemiology, 2 nd ed. Lippincott-Raven, PA, 115-134, 1998
8）宮原英夫，丹後俊郎(編)：医学統計学ハンドブック：16章 疫学，442-473，朝倉書店，1995
9）松山　裕：生物統計学の基礎的考え方．Surgery Frontier 10 ： 415-421, 2003

1) コントロール(対照群)の考え方

東京大学大学院医学系研究科生物統計学　松山　裕
Yutaka Matsuyama
(助教授)

Summary

ある治療法の効果を評価するためには，別の治療法との比較が不可欠である。単に，ある治療法を実施する前と実施した後の結果を比較するだけでは，得られる結論は弱く，より適切なコントロール群の結果と比較する必要がある。本稿では，科学的な研究においてある要因に関する効果を調べる際に，なぜコントロールが必要なのか，コントロールにはどのような種類があるのかについて述べる。

Key Words　治療効果，コントロール，比較の可能性，ランダム化，マスク化

Surgery Frontier 11(1)：76-80，2004

はじめに

臨床研究や疫学研究などの医学研究は，実験研究(experimental study)と観察研究(observational study)の2つのタイプに大きく分けることができる。新薬候補物質を標準薬と比較するランダム化研究は，実験研究の代表的なものであり，コホート研究やケース・コントロール研究などの多くの疫学研究は，観察研究の代表的なものである。この2つの研究方法の最も大きな違いは，前者では研究者自身がその研究で最も関心のある要因を人為的に操作するのに対して，後者ではそのような操作が行われない点である。

臨床医学系学会では Evidence Based Medicine (EBM)という言葉が大流行している。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[1]。最も高い証拠を提供するのが実験研究，とくに長期大規模ランダム化試験そして複数の実験研究を統計的に併合するメタ・アナリシスとされている。しかしながら，すべての研究のゴールは，研究目的である仮説について科学的に妥当な証拠を得ることにあり，観察研究から得られる証拠の質は，もし正しく計画された実験が可能であったならば得られたであろう証拠の質と同じであることを期待している[2]。

本稿では，実験研究・観察研究のいずれにおいても科学的に妥当な結論を得るために要求されるいくつかの要件のうち，最も基本的な要件であるコントロール(対照群)の必要性について述べる。

なぜコントロールが必要なのか？

ある特定の個人(以下，Aさん)に関して，「アスピリンを飲むことで頭痛が治るかどうか」について考えてみる。もしAさんがアスピリンを飲んで2時間以内に頭痛が治ったとする。この

事実(観察)だけによって，アスピリンはAさんの頭痛を治すのに効果があったといえるだろうか。

Aさんが何も薬を飲まなくても2時間後に頭痛が治っていたかもしれないので，その答えは否である。すなわち，Aさんに対するアスピリンの治療効果を調べるためには，「アスピリンを飲んだAさんが2時間後に頭痛が治った」という事実はそれ自体だけでは意味をなさず，以下の2つの状況を「同時に」知る必要がある。

状況1：Aさんが2003年12月1日にアスピリンを「飲んだ場合」に2時間以内に頭痛が治るかどうか

状況2：Aさんが2003年12月1日にアスピリンを「飲まない場合」に2時間以内に頭痛が治るかどうか

この2つの状況の結果がわかれば，Aさん個人に対してアスピリンの効果があるかどうかを調べることができる(表1)。状況1でAさんの頭痛が治り，かつ状況2では治らなければ，Aさん個人にとってはアスピリンの効果があると判断できる。しかし，どちらの状況でも頭痛が治れば，Aさんにとってはアスピリンの効果なしということになる。また，どちらの状況でも頭痛が治らなくてもAさんにアスピリンの効果なしということになる。これら2つの状況は，一方が観察されれば他方は絶対に観察することができないので，反事実的(counterfactual)と呼ばれる。

薬剤の評価に限らず科学的な因果推論を行うためには，人為的に操作可能な要因(アスピリン)を個人(Aさん)に対して加えた場合と加えない場合の反事実的な結果を比較しなければならない。この原理が，科学的な研究においてある要因に関する効果を調べるためにコントロールを必要とする根拠となっている。

表1に示した反事実的な結果は，実際に受けた治療によらず概念的には存在するものの同時に観察することはできない。したがって，データからは検証不能な仮定(例えば，アスピリンを飲まなかったこと以外はAさんとまったく同じ他人Bさんが存在するなど)をおかない限り，個人に対する治療効果を調べることはできない。しかしながら，このように「効果」というものを定義することによって，因果関係を調べるための2つの重要な要件を導くことができる。

1. 因果関係を調べる要因は，人為的に操作可能な要因でなければならない
2. 人為的に操作可能な要因の使用以外の条件が全て同一のコントロールが理想のコントロールである

表1　ある個人に対するアスピリンの効果

		飲まない	
		治る	治らない
飲む	治る	なし	効果あり
	治らない	負の効果	なし

なぜランダム化？

個人に対して2つの治療を同時に実施することはできないので，個人に対する治療効果を調べることはできないとしても，個人の集まりである集団に対する治療効果(平均的な治療効果)ぐらいは調べることはできないだろうか。

次の仮想的な臨床研究の例を通して考えてみる。「高脂血症に対する治療法として食事療法のみよりも，食事療法に抗高脂血症薬を加えたほうが心筋梗塞の発症を減少させることができるかどうか」という仮説を調べるために，6人の高脂血症患者さんを対象とした臨床研究を行うことを考えた。試験治療(抗高脂血症薬＋食事療法)グループとコントロール治療(食事療法のみ)グループ(以下，対照治療)の2グループ間で，試験開始から5年後の心筋梗塞発症割合を比較することをこの研究における主な評価項目(エンドポイント)とする。

前節で述べた治療効果の定義に従うと，この仮想的な臨床研究における6人の対象者全体での治療効果を調べるためには，やはり6人全員に試験治療を行った場合の5年間の心筋梗塞発症数と，「同じ」6人に対照治療を行った場合の5年間の心筋梗塞発症数を「同時に」比較しなければならない(図1)。

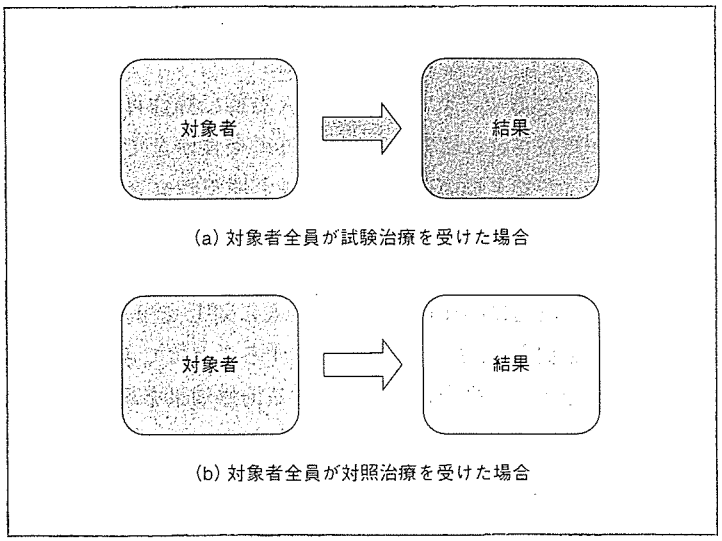

図1　臨床試験で知りたいこと
状況(a)が事実であれば、状況(b)は反事実（逆も同様）

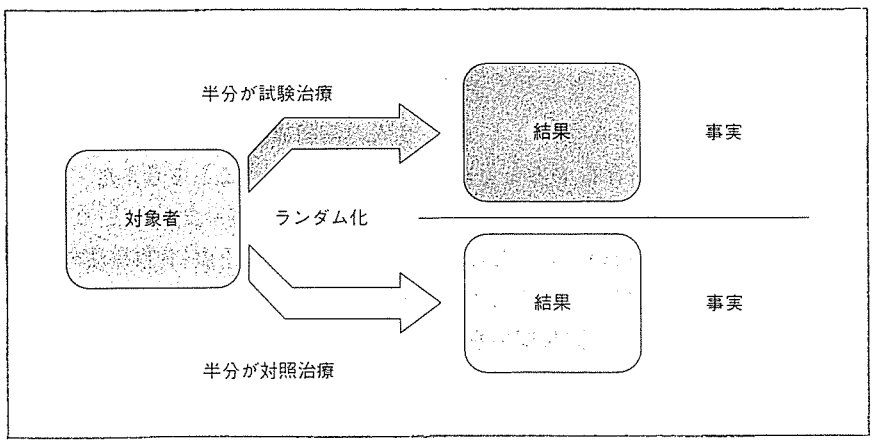

図2　ランダム化の意義

つまり，理想的な研究(理想のコントロールを用いた研究)を行うためには，「その研究で関心のある結果(心筋梗塞の発症)に関わるすべての要因のうち，操作可能なひとつの要因(抗高脂血症薬を与えるかどうか)のみがグループ間で異なる状況」を作り上げる必要がある。この目的を達成するためには，結果に影響をおよぼす可能性のあるすべての要因を制御する必要があるが，残念ながら，本質的にバラツキを伴うヒトを対象とした研究においては，すべての要因を制御できることはまれであり，現状の知識ではわからない未知の要因も存在するのが常態である。

集団に対しても理想のコントロールグループを観察することはできないが，例えば，コインを投げたりサイコロを振ったりという純粋に偶然の要素にのみ基づいて対象者を2つのグループに半数ずつ分けるとどうであろうか(図2)。今の仮想例では，対象者が6人と少ないが，もし対象者数が増えれば，グループ分けに恣意性が入らず，性，年齢，総コレステロール値などの心筋梗塞発症に関わる既知の要因のみならず，生活習慣などの(われわれの現在の知識ではわからない)未知の要因をも含めて，平均的には心筋梗塞発症リスクが均一な2つのグループを作り上げることができる[3]。すなわち，このランダム化という操作で治験治療に割り付けられた半分の患者さんの結果(心筋梗塞の発症割合)は，図1(a)の状況の代用として，また，ランダム化という操作で対照治療に割り付けられた半分の患者さんの心筋梗塞発症割合は，図1(b)の状況の代用として用いることができ，それら現実に観察可能な量を比較することで集団に対する治療効果を調べることができる(図2)。

ランダム化コントロールを採用したとすると，試験治療，対照治療グループそれぞれの対象者に割り付けられた内容を5年間守ってもらい，それぞれの心筋梗塞の発症割合を比較することで，「対象者全員に試験治療を行っ

た場合の5年間の心筋梗塞発症数と，同一の対象者に対照治療を行った場合の5年間の心筋梗塞発症数の比較」を平均的な比較可能性という観点からバイアスなく行うことができる。さらに，この治療法のランダム化という「強い」人為的な操作を行うことにより，治療効果の検定や推定といった統計計算の基礎も与えられる。

コントロールの種類

理想のコントロールグループを作ることは実際にはできないが，試験治療の効果を調べるための現実のコントロールグループの設定方法には，前節で述べたランダム化コントロールも含めて以下の4つの方法がある(表2)[4]。

①ランダム化コントロール
②同時コントロール
③既存コントロール
④コントロールなし

いずれの方法も，研究者自身による人為的な操作により試験治療を行うかどうかを決定することができるが，どの方法が「理想のコントロール」に一番近いであろうか。理想的なコントロールグループは実際には存在しないが，このような研究が実施できたと仮定し，そのような研究を「比較可能性(comparability)がある」と呼ぶことにすると，①から④へ番号が大きくなるにしたがって，この比較可能性(バイアスのひとつ)を保証することが難しくなる(表2)。

①のランダム化コントロールの場合でも，上の仮想例のように対象者数が6人と少なければ，偶然の要素により一方のグループに総コレステロール値の高い対象者が集まってしまうこともあり，比較の妥当性が損なわれる可能性がある。コホート研究やケース・コントロール研究などの疫学研究では②の同時コントロールを採用していることになるが，その場合，データとして測定したリスク要因に関しては比較グループが同一であることを示せるものの，未測定あるいは未知の要因に関しては同一であるかどうかを検討できない。さらに，③のヒストリカル研究では，①や②の場合と異なり，対照群に関してはこれからデータを測定するわけではないので，さらに比較の妥当性が損なわれる。最後に，④のコントロールなしは，「何もしなければ状態も変化しない」というコントロールを採用していることに相当する。物理や化学の分野では，「何もしなければ状態も変化しない」ことが多いであろうから，このタイプのコントロールを用いて因果関係を示してきたが，個人差の大きい医学領域では④のコントロールなしで比較の妥当性を確保することは難しい。

表2　コントロールのレベル[4]

コントロールグループ	設定の仕方	バイアス
ランダム化コントロール	試験治療と対照治療をランダムに割り付ける	測定できない特徴も平均的には同一にできるが，偶然のバラツキによるバイアスは起こりうる
同時コントロール	試験治療を実施したグループと同時期に対照治療を実施したグループ	測定できる特徴は同一であるかどうか確認できるが，測定できない特徴はどうしようもない
既存コントロール	過去に対照治療を実施したグループの既存のデータ	記録に残っていた特徴は同一であるかどうか確認できるが，記録されなかった特徴はどうしようもない
コントロールグループなし	「対照治療で30％が改善する」といった研究者の常識(暗黙のコントロール)	バイアスがないとまったく保証できない

プラセボとマスク化

前述の仮想的な臨床研究の例における，ある特定の個人が試験治療(抗高脂血症薬＋食事療法)を受けた場合と対照治療を受けた場合(食事療法のみ)の比較では，厳密にいうと，抗高脂血症薬が食事療法に比べて心筋梗塞発症に効果があるかどうかを調べているのではなく，「食事療法に加えた抗高脂血症薬の服用」が心筋梗塞の発症に効果があるかどうかを調べていることになる。すなわち，「抗高脂血症薬を服用すること」という心理的効果を

含めた比較になっている。抗高脂血症薬がこの心理的効果以上の効果があることを証明するためには，「ある個人が食事療法に加えて抗高脂血症薬を服用した場合」と「その個人が食事療法に加えて高脂血症に対する有効成分を含まない薬もどきを服用した場合」を比較しなければならない。この不活性な物質だけで作った「薬もどき」のことがプラセボと呼ばれる。

対照治療としてプラセボを使用したプラセボ対照比較試験は，対照治療グループの対象者に医療上のメリットがなく，倫理的に問題となることがある。前述の仮想的な臨床研究の例では，対象者全員に食事療法を実施しているので，食事療法だけでもコレステロールが十分コントロール可能な患者さんだけを対象とするように試験の選択基準が設定されていれば，プラセボの使用は倫理的にも許されるであろう。このように標準的治療に上乗せの効果を期待する場合，プラセボの使用に倫理的な問題は生じないかもしれないが，その他の状況でプラセボの使用が倫理的かどうかは，無治療という選択が許されるかどうかに依存する[5]。

無治療が許されない場合には，プラセボを用いることは非倫理的であり，「証明された治療法」を対照治療として用いた研究を行わなければならない[6]。この場合は実薬コントロールと呼ばれる。

前述の仮想的な臨床研究の例において，対照治療を「食事療法＋プラセボ」とすることによって，薬の服用という心理的効果を排除でき，比較可能性が保たれたが，対象者に抗高脂血症薬を服用しているかプラセボを服用しているかがわかってしまうと，試験治療グループに再び心理的効果が入ってしまう。このため，対象者にはどちらを服用しているのかわからないようにする工夫がなされる。これがマスク化(Masking)である。また，プラセボを用いることによって，対象者だけでなく医師に対してもどちらを服用しているかをマスク化することもある(二重マスク化)。対象者が抗高脂血症薬かプラセボのどちらを服用しているのかを医師が知ってしまうと，その対象者に対する医師の評価，管理にバイアスが入る恐れがある。医師に対してもマスク化することにより，これらのバイアスを避けることができる。ただし，治療の種類によっては，マスク化が不可能，あるいはすべきではない状況も存在する。例えば，投与直後の嘔気のようなその治療に特有の副作用が存在する場合には，マスク化が実際には有効ではなかったり，血小板減少が予想される薬剤を投与する場合には，血液検査，血小板輸血の準備のためにどちらの治療群かを前もって知っておく必要があったりする。これらの場合であっても，患者さんにだけはマスク化を行う，最終的に評価を行う医師にはマスク化を行う，などマスク化のレベルを変えて，可能な限りマスク化することを心がけてほしい。

おわりに

コントロールの選択の仕方としては，表2の①から順に実施可能性を考え，どうしても実施できない場合に次のレベルのコントロールを選択すべきである。単に実施が容易だという理由だけから③や④を選択することは避けるべきで，より妥当なコントロールである①や②を採用できない理由，③や④のコントロールでも比較の妥当性が十分確保できるかどうかを示しておく必要がある。また，たとえ①のランダム化コントロールを採用したとしても，その研究の質が非常に低ければ(例えば，治療のコンプライアンスが悪い，脱落例が多いなど)，得られる結果は妥当なものにはなり得ず，②の同時コントロールを採用した良質な疫学研究の方が得られる結果の信憑性が高いことも忘れてはならない。

文　献

1) Evidence Based Medicine Working Group : Evidence-based medicine. JAMA 268 : 2420-2425, 1992
2) Rothman KJ, Greenland S : Modern Epidemiology, 2nd ed. Lippincott-Raven, PA, pp. 67-78, 1998
3) 佐藤俊哉：治療のランダム割り付けと治療効果の検定．医学のあゆみ 173 : 779-784, 1995
4) 佐藤俊哉：Pコントロール．これからの臨床試験：椿広計，藤田利治，佐藤俊哉編．朝倉書店，東京，pp. 21-33, 1999
5) Altman DG : Practical Statistics for Medical Research. Chapman & Hall, London, pp. 440-476, 1991
6) Rothman KJ, Michels KB : The continuing unethical use of placebo controls. N Engl J M 331 : 394-398, 1994

Clinical Bioinformatics

医学研究における統計学入門

1) 医学研究のデザイン概論 1 —ランダム化比較試験—

東京大学大学院医学系研究科生物統計学　松山　裕
Yutaka Matsuyama
(助教授)

Summary

臨床研究や疫学研究を含む科学的な研究は，実験研究(experimental study)と観察研究(observational study)の2つのタイプに大きく分けることができる。この2つの研究方法の最も大きな違いは，前者では研究者自身がその研究で最も関心のある要因を人為的に操作するのに対して，後者ではそのような操作が行われない点である。例えば，実験研究では，抗高脂血症薬を飲ませるかどうかは研究者が決定するが，観察研究ではそのような介入はなされない。本稿では，前者の実験研究，特にランダム化比較試験をデザインする際のいくつかの重要な概念について説明する。ここで，科学的な実験研究とは，「研究者自身が行う人為的な操作によってどのような結果が研究対象に生じるかを適切なコントロールグループが存在する状況で調べるための一連の観察」と定義される。これには，実験動物を対象とした非臨床試験，患者さんを対象とした臨床試験(clinical trials)，主として健康な一般住民を対象とした予防研究(prevention studies)などが含まれる。

Surgery Frontier 12(2)：80-84，2005

Key Words　プロトコル，ランダム化，マスク化，コントロール，内部妥当性，一般化可能性

プロトコルの作成

実験研究，あるいは観察研究に限らずすべての科学的な研究において，質の高い研究を実施するためには綿密な研究計画を立てることが何よりもまず重要である。具体的には，実際に研究を開始する前に，研究の具体的な目的，方法などを文書化したプロトコル(研究計画書)を作成する必要がある。

基本的にはプロトコルと研究実施マニュアル(プロトコルには記載できない詳細な実施手順)をみれば，行われた研究を再現できるようにすべきであり，プロトコルには最低限，表1に示す内容が記載されていなければならない[1]。プロトコル，実施マニュアル，調査票の作成などは，臨床研究者，疫学者，生物統計家，医療倫理の専門家，データマネジャーなどの複数の専門家がチームとして行わなければならない作業であり，その内容は，簡潔で，その研究に携わるすべての者にとってわかりやすいものでなくてはならない。プロトコルにおける「研究計画のまとめ」などは，対象者への説明の手助けとなるように，専門的な用語はできる限り避けて平易な表現で書く必要がある。

特に，臨床試験はヒトに対する実験であり，ヘルシンキ宣言の精神に則って実施される必要がある。そのため，厳密なプロトコルの作成とその遵守，プロトコルに定められた選択基準を満たす患者さんに，試験の内容，試験治療の利益と不利益，試験に参加しなくても不利益を被らないこと，いつでも試験を止められることなどを説明し，患者さんが十分理解したうえで試験の参加に同意すること(インフォームド・コンセント)が義務づけられている[2]。例えば，がん臨床試験においても病名の告知，つまり「がんであること」を知らせることは最低限必要である。担当医がこの患者さんにはがんであることを知らせないほうがいい，と判断することは尊重されるべきだが，「がんであることを知らせないほうが

いい」と判断した患者さんは臨床試験の対象としてはならない。

すべての研究，特に，実験研究においては，作成されたプロトコルは，研究開始前(および，研究終了まで定期的)に，研究(あるいは，倫理)審査委員会(Institutional Review Board：IRB)により審査され，承認されなければならない。また，研究途中で中間解析を予定している場合などは，様々な情報(集積されたデータ，生物学的知識，先行研究からの情報など)に基づいてその研究を途中で中止する，あるいは継続するなどの判断を下す第三者機関である独立データモニタリング委員会(Independent Data and Monitoring Committee：IDMC)を設置することも必要である。綿密に計画され十分に科学的な研究を実施しなければならないことはいうまでもないが，それと同時にその研究が倫理的にも妥当であることを保証することが重要である。

表1　プロトコルに最低限記載すべき事項

項目	内容
研究計画のまとめ	研究全体の簡潔なまとめ
研究の目的	背景・理論的根拠・研究仮説・エンドポイント
研究の実施期間・場所	いつ，どこで，いつまで
研究対象集団の設定	適格条件・除外条件
インフォームド・コンセント	説明と同意の具体的方法
研究の実施方法	研究の型・割り付け方法・用量変更・併用治療・症例の取り扱い・測定項目・測定時期
統計解析を行う項目・方法	エンドポイントの設定と具体的な解析方法
研究に必要なサンプルサイズ	必要な対象者数算出の根拠
研究の組織・責任の所在・連絡先	研究組織一覧・連絡先

治療法の割り付け
—なぜランダム化？—

新薬や新治療法のヒトに対する効果を調べるための標準的実験方法としてランダム化臨床試験が現在広く用いられている。なぜ治療法をランダム化する必要があるのかを次の仮想的な臨床試験の例を通して考えてみる。

「高脂血症に対する治療法として食事療法のみよりも，食事療法に抗高脂血症薬を加えたほうが心筋梗塞の発症を減少させることができる」という仮説を調べるために，6人の高脂血症患者さんを対象とした臨床試験を行うことを考えた。試験治療(抗高脂血症薬+食事療法)グループとコントロール治療(食事療法のみ)グループ(以下，対照治療)の2グループ間で，試験開始から5年後の心筋梗塞発症割合を比較することをこの研究における主な評価項目(エンドポイント)とする。

6人の対象者におけるある特定の個人に関する試験治療の効果を調べるためには，以下のことを「同時に」知る必要がある。

A)その個人が試験治療を受けた場合，5年以内に心筋梗塞を発症するかどうか

B)その個人が対照治療を受けた場合，5年以内に心筋梗塞を発症するかどうか

この2つの状況を知ることができれば，その個人にとって，試験治療の効果があるかどうかを証明することができる。すなわち，状況Aでは心筋梗塞を発症せず，かつ状況Bで発症すれば，その個人にとって試験治療はプラスの効果があると判断できる。逆に，状況Aで心筋梗塞を発症し，かつ状況Bでは発症しなければ，その個人にとって試験治療はマイナスの効果があると判断できる。また，どちらの状況でも心筋梗塞を発症しない，あるいは発症する場合には，その個人にとって試験治療は効果なしということになる。

このように考えると，上の仮想的な臨床試験の例における6人の対象者全体での治療効果(平均的な治療効果)を調べるためには，6人全員に試験治療を行った場合の5年間の心筋梗塞発症数と，「同じ」6人に対照治療を行った場合の5年間の心筋梗塞発症数を「同時に」比較しなければならないことになる。しかしながら，これら2つの状況は，一方が観察されれば他方を観察することは不可能である。

つまり，理想的な実験的研究を行うためには，「その研究で関心のある結果(心筋梗塞の発症)に関わるすべての要因のうち，操作可能な1つの要因(抗高脂血症薬を与えるかどうか)

のみがグループ間で異なる状況」を作り上げる必要がある。この目的を達成するためには，結果に影響をおよぼす可能性のあるすべての要因を制御する必要があるが，残念ながら，本質的にばらつきを伴うヒトを対象とした研究においては，すべての要因が既知であることはまれであり，人為的に操作した要因のみが異なる理想的なコントロールグループを作ることは不可能である。

理想的なコントロールグループを作ることは実際にはできないが，試験治療の効果を比較するための現実的なコントロールグループの設定方法には，4つの方法がある(表2)[3]。

表2　コントロールのレベル

番号	コントロールグループ	設定の仕方	バイアス
①	ランダム化コントロール	試験治療と対照治療をランダムに3人ずつに割り付ける	測定できない特徴も平均的には同一にできるが，偶然のばらつきによるバイアスは起こりうる
②	同時コントロール	試験治療を実施したグループと同時期に対照治療を実施した別のグループ6人	測定できる特徴は同一であるかどうか確認できるが，測定できない特徴はどうしようもない
③	既存コントロール	過去に対照治療を実施したグループの別の既存データ6人	記録に残っていた特徴は同一であるかどうか確認できるが，記録されなかった特徴はどうしようもない
④	コントロールなし	「対照治療で30％が改善する」といった研究者の常識（暗黙のコントロール）	バイアスがないとはまったく保証できない

（文献3より引用）

いずれの方法も，研究者自身による人為的な操作により試験治療を行うかどうかを決定することができるが，どの方法が「科学的な実験的研究の定義」に一番近いであろうか？

前述の「同じ6人に対照治療を行った場合」という理想的なコントロールグループは実際には存在しないが，このような研究が実施できたと仮定し，そのような研究を「比較可能性(comparability)がある」とよぶことにすると，①から④へ番号が大きくなるにしたがって，この比較可能性を保証することが難しくなる。①のランダム化臨床試験の場合でさえも，上の仮想例のように対象者数が6人と少なければ，偶然の要素により一方のグループに総コレステロール値の高い対象者が集まってしまうこともあり，比較の妥当性が損なわれる可能性がある。このような2グループ間の治療以外の要因が比較可能でないことは交絡(confounding)とよばれ，その研究での内部妥当性が損なわれる(バイアスが存在する)1つの原因となる。

しかしながら，このランダム化コントロールは，他の②から④のコントロールとは違い，(例えば，コインを投げたりサイコロを振ったりという)純粋に偶然の要素に基づいて対象者を2つのグループに分けているために，対象者数が増えれば，性，年齢，総コレステロール値などの心筋梗塞発症に関わる既知の要因のみならず，生活習慣などの(われわれの現在の知識では分からない)未知の要因をも含めて平均的には心筋梗塞発症リスクが均一な2つの(理想的に近い)グループを作り上げることができるという利点をもっている[4]。言い換えれば，④のコントロールが存在しない状況，③の過去に記録された情報のみに基づいた既存コントロール，あるいは，②の未知の要因を考慮できない同時コントロールと違い，ランダム化コントロールを採用し，対象者を「平均的には心筋梗塞発症リスクが均一」という比較可能な2グループに分けることにより，理想的な実験的研究の実施に近づくことができる。

ランダム化コントロールを採用したとすると，試験治療，対照治療グループそれぞれの対象者に割り付けられた内容を5年間守ってもらい，心筋梗塞の発症割合を比較することで，「対象者全員に試験治療を行った場合の5年間の心筋梗塞発症数と，同一の対象者に対照治療を行った場合の5年間の心筋梗塞発症数の比較」を平均的な

比較可能性という観点からバイアスなく行うことができる。さらに，この治療法のランダム化という「強い」人為的な操作を行うことにより，治療効果の検定や推定といった統計計算の基礎も与えられる。

プラセボとマスク化

前述の仮想的な臨床試験の例における，ある特定の個人が試験治療(抗高脂血症薬+食事療法)を受けた場合と対照治療を受けた場合(食事療法のみ)の比較では，厳密にいうと，抗高脂血症薬が食事療法に比べて心筋梗塞発症に効果があるかどうかを調べているのではなく，「食事療法に加えた抗高脂血症薬の服用」が心筋梗塞の発症に効果があるかどうかを調べていることになる。すなわち，「抗高脂血症薬を服用すること」という心理的効果を含めた比較になっている。抗高脂血症薬がこの心理的効果以上の効果があることを証明するためには，「ある個人が食事療法に加えて抗高脂血症薬を服用した場合」と「その個人が食事療法に加えて高脂血症に対する有効成分を含まない薬もどきを服用した場合」を比較しなければならない。この不活性な物質だけで作った「薬もどき」はプラセボとよばれる。

対照治療としてプラセボを使用したプラセボ対照比較試験は，対照治療グループの対象者に医療上のメリットがなく，倫理的に問題となることがある。前述の仮想的な臨床試験の例では，対象者全員に食事療法を実施しているので，食事療法だけでもコレステロールが十分コントロール可能な患者さんだけを対象とするように試験の選択基準が設定されていれば，プラセボの使用は倫理的にも許されるであろう。このように標準的治療に上乗せの効果を期待する場合，プラセボの使用に倫理的な問題は生じないかもしれないが，その他の状況でプラセボの使用が倫理的かどうかは，無治療という選択が許されるかどうかに依存する[5]。

無治療が許されない場合には，プラセボを用いることは非倫理的であり，「証明された治療法」を対照治療として用いた研究を行わなければならない。この場合は実薬対照とよばれる。

前述の仮想的な臨床試験の例において，対照治療を「食事療法+プラセボ」とすることによって，薬の服用という心理的効果を排除でき，比較可能性が保たれたが，対象者に抗高脂血症薬を服用しているかプラセボを服用しているかがわかってしまうと，試験治療グループに再び心理的効果が入ってしまう。このため，対象者にはどちらを服用しているのかわからないようにする工夫がなされる。これがマスク化(Masking)である。また，プラセボを用いることによって，対象者だけでなく医師に対してもどちらを服用しているかをマスク化することもある(二重マスク化)。対象者が抗高脂血症薬かプラセボのどちらを服用しているのかを医師が知ってしまうと，その対象者に対する医師の評価，管理にバイアスが入る恐れがある。医師に対してもマスク化することにより，これらのバイアスを避けることができる。ただし，治療の種類によっては，マスク化が不可能，あるいはすべきではない状況も存在する。例えば，投与直後の嘔気のようなその治療に特有の副作用が存在する場合には，マスク化が実際には有効ではなかったり，血小板減少が予想される薬剤を投与する場合には，血液検査，血小板輸血の準備のためにどちらの治療群かを前もって知っておく必要があったりする。これらの場合であっても，患者さんにだけはマスク化を行う，最終的に評価を行う医師にはマスク化を行う，などマスク化のレベルを変えて，可能な限りマスク化することを心がけることが大切である。

臨床研究の内部妥当性と一般化可能性

ランダム化臨床試験において最も一般的な研究デザインである群間比較試験のデザインを図示すると図1のようになる[6]。

ランダム化比較臨床試験，特に標準治療の構築と普及を目指す第Ⅲ相臨床試験計画の目標は，精確性(clarity)，比較可能性(comparability)，一般化可能性(generalizability)の3つにまとめることができる[7]。精確性の要求とは，主たる評価項目の計測について，そのランダムなばらつきを小さくし，研究の精度を向上させることである。このためには，より多くの対象者を観察することが最も有効である。比較可能性の要求とは，バイアスを減らすことに対応し，研究の内部妥当性(internal validity)を向上させることである。このための最も強力な手段が前

述の治療法のランダム化である。

しかしながら，通常の臨床研究においては，研究成果の適用を考えている目標集団と実際に研究が行われる標本集団の間には数の上でも質の上でも大きな隔たりが生じる可能性が存在する。したがって，臨床研究においてはランダムサンプリングなどは存在せず，厳密にいえば，ランダム化臨床試験から得られた結論の適用範囲は当該の標本集団のみであり，この結論が目標集団まで外挿できるかどうかが一般化可能性(外部妥当性)の議論である。ランダム化臨床試験からの結論がより広い患者集団に外挿できるかどうかを検討するために，治療効果に関するサブグループ解析や治療と施設との交互作用解析などが試みられる。しかしながら，ある研究からの結論の一般化を確保するためには，「得られた結果を抽象化し，特定の研究結果からその研究を実施した時期や地域を越えた普遍的な仮説を産み出す」というプロセスの積み重ねが重要であり，これは決して統計的な問題ではない。臨床試験における結論の一般化のためには，統計的な議論よりもデザインの段階において患者さんの選択を避け，「試験の適格基準を満たす患者さんにはすべて参加をお願いする」という原則を守ることが重要である。

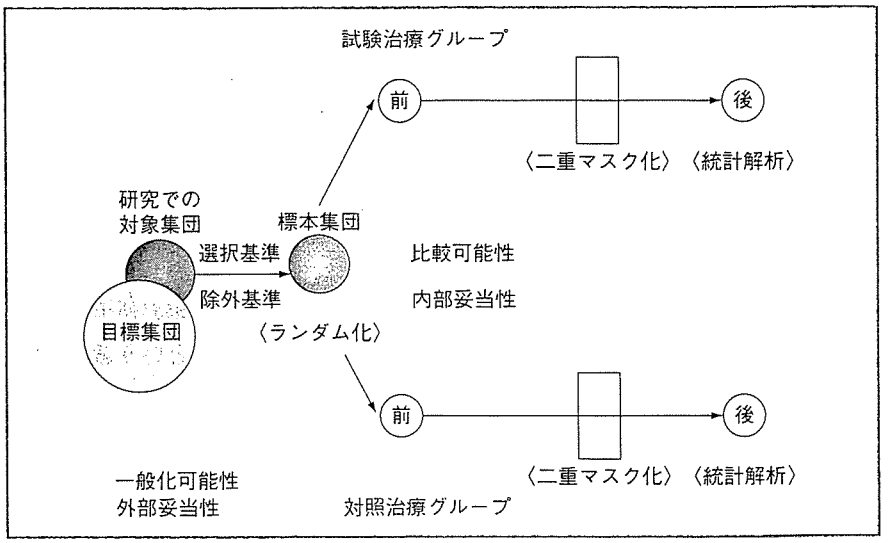

図1　群間比較試験における内部妥当性と一般化可能性

(文献6より引用)

おわりに

臨床試験は新しい治療法や予防法の適切な開発と普及，科学的な医療の進歩のためには不可欠である。そのためには，臨床試験にかかわるすべての研究者が，ヒトに対する実験という観点から，ヘルシンキ宣言の精神に則った上で，科学的に妥当な臨床試験の原理と方法論を十分に理解し，実践していくことが不可欠である。

文　献

1) 佐藤俊哉：調査計画書の作成―調査を始める前の自己チェック―．保健の科学 37 : 72-76, 1995
2) 厚生大臣：医薬品の臨床試験の実施の基準に関する省令：平成9年3月27日厚生省令第28号．1997. (http://www.nihs.go.jp/dig/ich/ichindex.html)
3) 佐藤俊哉：Pコントロール．椿　広計，藤田利治，佐藤俊哉(編)：これからの臨床試験．21-33，朝倉書店，1999
4) 佐藤俊哉：治療のランダム割り付けと治療効果の検定．医学のあゆみ 173 : 779-784, 1995
5) Rothman KJ, Michels KB : The continuing unethical use of placebo controls. N Engl J M 331 : 394-398, 1994
6) 藤田利治：臨床試験とは．椿　広計，藤田利治，佐藤俊哉(編)：これからの臨床試験．1-19，朝倉書店，1999
7) 大橋靖雄：Biostatistician からの本邦臨床試験への提言．癌と化学療法 20 (16) : 2483-2495, 1993

2) 医学研究のデザイン概論2 —疫学研究—

東京大学大学院医学系研究科生物統計学　松山　裕
Yutaka Matsuyama
(助教授)

Key Words　コホート研究，ケース・コントロール研究，交絡，バイアス，コホート内ケース・コントロール研究

Summary

疫学研究では倫理的な理由(健康に有害な影響があると考えている要因を対象者にランダムに割り付けるわけにはいかない)から観察研究を実施することが多い。疫学研究のための代表的な研究デザインがコホート研究(cohort study)とケース・コントロール研究(case-control study)である。例えば，喫煙による心筋梗塞発生リスクを調べる場合，コホート研究では対象者を要因の有無(喫煙者と非喫煙者)で分類し，性，年齢などの背景因子を調べ，数年かけて心筋梗塞の新規発生(incidence)を比較する。一方，ケース・コントロール研究では，結果の有無(心筋梗塞を発生したケースと発生していないコントロール)別に過去にさかのぼって喫煙状況と背景因子を調べる。本稿では，「何らかの要因への曝露」，あるいは「今行われている治療」についていわば受け身の状態で実施する観察研究のいくつかのデザインについて解説する。

Surgery Frontier 12(2) : 85-90, 2005

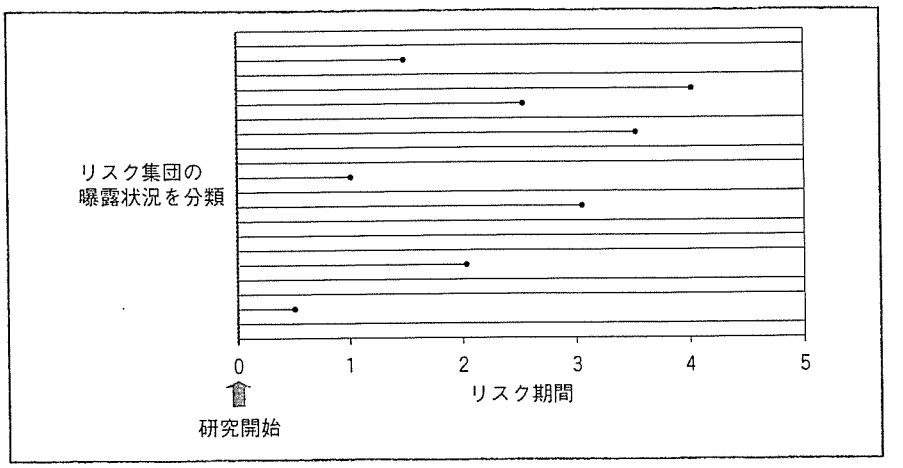

図1　コホート研究の概要
図中の1本の線は対象者の追跡を示し，●はイベントの発生を意味する

コホート研究

コホート研究はランダム化比較試験に代表される実験研究とよく似た設定であり，因果の順序と研究実施の時間的順序が同じであり，わかりやすい研究デザインである。例えば，喫煙による心筋梗塞発生リスクをコホート研究で調べることを考える。その研究デザインの概要は，「研究開始時に心筋梗塞を発生していない対象者を喫煙状況で分類し，性，年齢，運動・食事状況などの心筋梗塞発生に対するリスク因子を調べ，数年間の追跡を行い，心筋梗塞の新規発生状況を調べる」というものである。コホート研究は，その研究で一番調べたい要因(曝露とよばれる)の有無で対象者(特定のリスク集