医療統計学を用いた新しい治療効果判定法の開発

新潟大学医学部附属病院医療情報部
赤 澤 宏 平

An Alternative Statistical Method to Evaluate the Treatment Effect in Randomized Clinical Trials

Kohei AKAZAWA

Department of Medical Informatics
Niigata University Medical Hospital

This paper evaluates the loss of power of the simple and stratified logrank tests due to heterogeneity of patients in clinical trials. The results of the paper are based on the analyses of survival data from a large clinical trial which includes more than 6,000 cancer patients. Major findings from the simulation study on power are: (1) for a heterogeneous sample, such as advanced cancer patients, a simple logrank test can yield misleading results and should not be used; (2) the stratified logrank test may suffer some power loss when many prognostic factors need to be considered and the number of patients within stratum is small. To address the problems, due to heterogeneity, the Cox regression method with a piecewise linear hazard model is recommended.

Key words: randomized clinical trial, sruvival time, treatment effect, Cox regression method, piecewise linear hazard model
無作為化臨床試験，生存時間，治療効果，Cox 回帰法，折れ線ハザードモデル

は じ め に

生存時間を評価尺度とする臨床試験において，治療効果判定に用いられる統計学的検定の問題点とその解決方法を述べる．臨床試験の症例は不均一である，即ち，各症例のもつ背景因子は同一ではない．国際的に用いられるログランク検定で治療効果を判定する際に，この症例の不均一性が治療効果の検出を妨げることをコンピュータシミュレーションにより立証する．ログランク検定に代わる新しい治療効果判定方法として折れ線 Cox 回帰法を提唱する．

用 語 の 説 明

1） 生存時間を評価尺度とする臨床試験

生存時間をエンドポイントとする臨床試験の概要を図1にまとめた．まず，適格条件に合致する症例にインフォームドコンセントを行い同意を得る．適格条件とは，たとえば，年齢が40歳から70歳まで，臨床進行期が

Reprint requests to: Kohei AKAZAWA
Department of Medical Informatics
Niigata University Medical Hospital
Niigata City, 951-8520 JAPAN

別刷請求先： 〒951-8520 新潟市旭町通一番町 754
新潟大学医学部附属病院医療情報部 赤 澤 宏 平

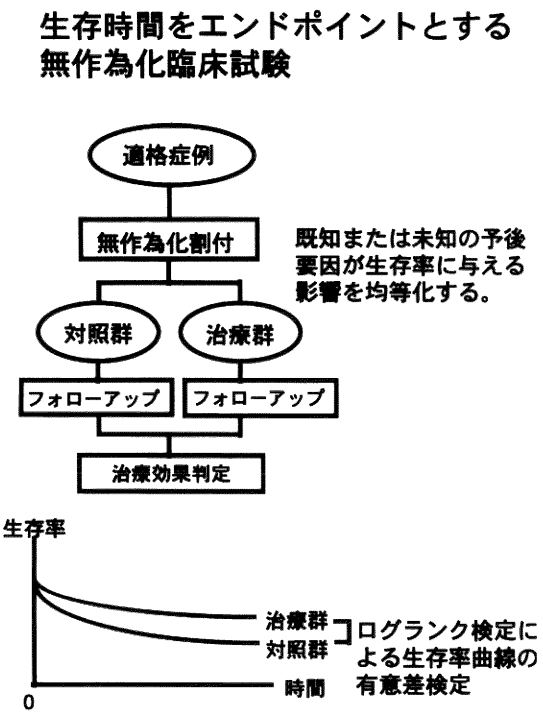

図1　無作為化臨床試験の概要

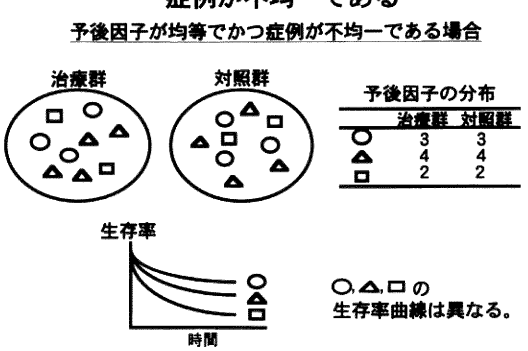

図2　不均一な症例群による臨床試験

ⅡかⅢ，Performance Status がⅡ以下の症例というように，臨床試験に登録できる症例の条件をいう．臨床試験に登録された症例は，一般に，対照群（無治療群，または，従来から行われてきた標準治療群）か治療群（新しく開発された新治療群）の2群にランダムに割り付けられる．プロトコールに定めた治療を実施した後あるいは実施しながら症例のフォローアップが行われる．フォローアップの開始日は，確定診断日や手術日，治療開始日である．通常，数ケ月に1回の定期検査や年一回の生存確認調査をもとにその症例の生死と死亡年月日が採取される．各症例のフォローアップを終えた後，通常ログランク検定か層別ログランク検定で治療効果判定を行い，治療群が対照群に比べて生存率が有意に上昇したかを評価する．

2）症例の不均一性

症例が不均一であるとは，図2に示すように，異なる生存率曲線（生存時間分布）に従う症例が治療群・対照群に混在することをいう．図2では症例の不均一を○，△，□で表した．○，△，□は，たとえば，臨床進行期StageⅡ，Ⅲ，Ⅳを表している．一方，症例が均一であるとは，治療群・対照群のすべての症例が同一の生存率曲線に従うことをいう．動物実験では被験動物は均一である．臨床試験において，症例を均一に近づけようとすると適格条件を厳しくする必要がある．たとえば，年齢50歳，男性，臨床的病期 StageⅡ，手術根治度 B，術前 Performance Status 0または1，・・・という具合である．ところが，このように適格条件を厳しくすると，条件に合致する症例を少数例しか集積できず治療効果判定が行えない．通常，対象疾患の程度を決定づける，あるいは，治療効果に大きな影響を与える数個の因子のみを選んで適格条件とする．従って，適格条件で使われた因子以外の要因にばらつきが生じ生存率曲線も症例ごとに異なるので，登録された症例は不均一である．

3）ハザード（瞬間死亡率）

時間 t の直前までは生存したという条件の下で次の瞬間（時間 t）に死亡する確率のことをハザードと呼ぶ．

4）不均一の程度

年齢，リンパ節転移の程度など生存率に影響を与えうる予後因子を Z で表す．予後因子 Z（たとえば，リンパ節転移の程度は Z＝0，1，2，3）をもつ個体の時間 t におけるハザード $\lambda(t \mid Z)$ が，以下の比例ハザードモデルに従うと仮定する．

$$\lambda(t \mid Z) = \lambda_0(t)\exp(Z\beta)$$

ただし，$\lambda_0(t)$ は未知の正値関数，β は予後因子のハザードへの重みを表す回帰係数である．

このとき，不均一の程度（不均一度）を対数ハザードのレンジ β（Zmax−Zmin）で表す．ただし，Zmax，Zmin はそれぞれ予後因子 Z の最大値，最小値を表す．

5）検出力

本当に存在する治療効果を，有意差検定において治療効果あり（有意差あり）と検出できる確率をいう．サン

プルサイズが小さいと，本当は治療効果があるにもかかわらず有意差検定で治療効果なしという誤った結論を下す可能性がある．本研究では，症例が不均一な場合にも，ログランク検定の検出力は低下することを示した．

胃癌臨床試験における不均一推定

実際の無作為化臨床試験において不均一の程度がどれくらいであるのかを，がん集学的治療研究財団特定研究1のデータを用いて推定した．本臨床試験は，胃癌の肉眼的治癒切除症例に対する免疫賦活剤の延命効果を評価することにある[1]．登録症例は全例 MMC とテガフールによる化学療法を受けた．この共通の化学療法に加え，免疫賦活剤無投与，PSK のみ投与，OK－432 のみ投与，PSK＋OK－432 投与の4群が設定された．1981年4月から1983年8月までに全国で7,637例の胃癌治癒切除症例が登録され5年間の追跡調査が行われた．各群の5年生存率は約65％であり，4群の生存率曲線の有意差を単純ログランク検定で検定した結果，治療群間での有意差は認められなかった（$p=0.50$）[1]．

計測された予後因子は全部で25個あるが，その中から手術時年齢，組織学的ステージ，ボールマン分類の3因子に注目して不均一度を推定した．これら3因子の他に，3因子による2次の交互作用因子を加えた合計6因子からなる以下の比例ハザードモデルを使い，前節の方法で不均一度を推定した．

$$
\begin{aligned}
&\lambda(t \mid \text{年齢，ステージ，ボールマン}) \\
&= \lambda_0(t)\exp(\beta_1 \times \text{年齢} \\
&\qquad + \beta_2 \times \text{ステージ} \\
&\qquad + \beta_3 \times \text{ボールマン} \\
&\qquad + \beta_4 \times \text{年齢とステージの交互作用} \\
&\qquad + \beta_5 \times \text{年齢とボールマンの交互作用} \\
&\qquad + \beta_6 \times \text{ステージとボールマンの交互作用})
\end{aligned}
$$

ただし，$\lambda(t \mid \text{年齢，ステージ，ボールマン})$ は年齢，ステージ，ボールマンが与えられた下での時間 t におけるハザード，$\lambda_0(t)$ は時間 t におけるベースラインハザード，$\beta i\ (i=1, \ldots, 6)$ は回帰係数である．

推定の結果，本臨床試験における不均一の程度は3.75であると推定された．実際の臨床試験から推定された不均一の大きさに基づき，不均一性が治療効果判定の検出力に与える影響を定量的に評価する．

不均一性による治療効果判定の検出力に与える影響

よくデザインされた無作為化臨床試験において，たとえ無作為化が完全で治療群間で予後因子の分布がまったく同一であっても，群内に症例の不均一性が存在する場合には治療効果判定の検出力は低下する．このことは数学的にも証明されている[2]．本稿では，症例の不均一度と治療効果判定の検出力との関係を定量的に評価した結果を示す．前節の結果を使ってモンテカルロシミュレーションを行った．次の条件を仮定する．

（1）治療群と対照群の2群のみ
（2）症例数は1群あたり200例，合計400例
（3）死亡例数は320例
（4）追跡不能例は無い
（5）治療効果は比例ハザードモデルに従って生存率を高めるものとする．その大きさとして，対照群の生存率50％を治療群60％に高めることを仮定する．
（6）検出力は2,000回の繰り返し計算から求める．

上述の条件を満たすモンテカルロ法のシミュレーションのしくみを図3に示した．プログラム言語は FORTRAN と NUMPAC を用いた．症例の不均一の作り

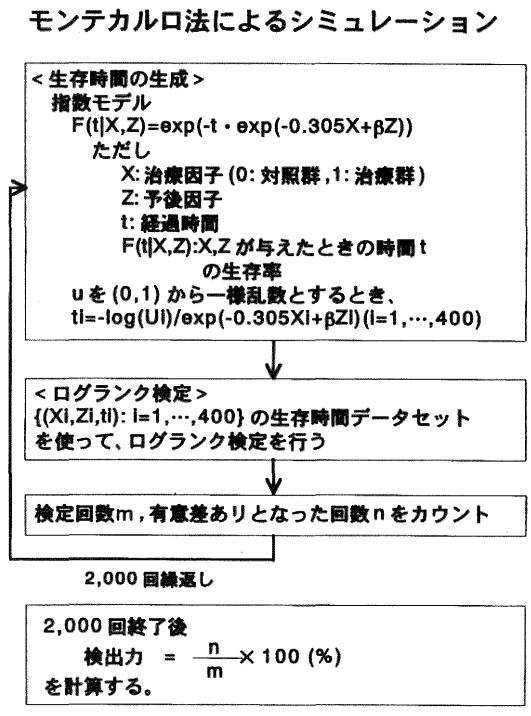

図3　ログランク検定の検出力を推定するためのシミュレーション

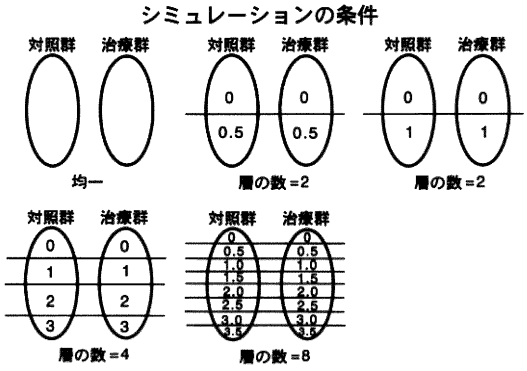

図4　シミュレーションにおける不均一な症例群の作り方

方を**図4**に示した．シミュレーションの結果を**表1**に示す．1列目は層の数（不均一の原因となる予後因子のカテゴリー数），2列目は各層に割り当てられたハザードの大きさ，3列目，4列目は単純ログランク検定と層別ログランク検定の検出力の推定値をパーセントで表している．表中の1行目に示された層の数が1，対数ハザードの大きさが0のときは群内の症例が均一であることを意味する．このときの検出力は77.8％である．即ち，上述の条件の下で同一の無作為化臨床試験を100回繰り返し行ったとき，治療効果ありという結果をおおよそ78回得ることを意味する．即ち，対照群50％の生存率を治療群60％に引き上げる治療効果を，検定で治療効果ありとの結果を検出できる確率が0.78ということである．先に述べた胃癌臨床試験で推定された不均一の程度は3.75であった．この不均一度に対応する検出力は，**表1**の対数ハザードのレンジが3.5（層の数が8の場合）の場合を参照すればよい．このときの単純ログランク検定検出力は33.2％である．従って，同様な臨床試験を繰り返し行いそのたびに単純ログランク検定で治療効果を判定したとすると，3回に1回程度しか治療効果を検出できないことになる．もともと治療効果を検出できない状況下で検定を行っていることを意味する．

これらのシミュレーション結果より次のことがわかる．即ち，症例が不均一な場合には，単純ログランク検定の検出力は著しく低下する．従って，単純ログランク検定をがん臨床試験に用いるべきではない．

症例の不均一性による層別ログランク検定の検出力はほとんど低下していない（**表1**）．理由は層内の症例が均一であるからである．実際の臨床試験では，たとえば，ステージ分類で4層に層別しても，手術時年齢やボールマン分類，腫瘍長径など層内の症例を不均一にする予後因子がまだ多数存在する．層内の不均一は層別ログランク検定の検出力を低下させることが別のシミュレーションで確認されている[2)3)]．

表1　単純ログランク検定および層別ログランク検定の検出力

標本数は各群200ずつで死亡数は320，薬効は対照群50％の生存率を治療群60％に上昇させることを仮定している．症例が均一ならば，検出力は77.8％である．層内の標本は均一を仮定している．

層の数	対数ハザード	検出力	
		単純ログランク検定	層別ログランク検定
1	0	77.8	−
2	0，0.5	76.7	78.0
2	0，1	66.8	78.2
2	0，2	36.5	78.0
4	0，1，2，3	35.3	77.2
8	0，0.5，…，3.5	33.2	76.8
16	0，0.3，…，4.5	21.4	75.2
32	0，0.2，…，6.2	8.1	72.2

折れ線 Cox 回帰法

生存時間をエンドポイントとする無作為化臨床試験の治療効果判定では，多数の予後因子の影響を補正する必要がある[4)−7)]．予後因子の影響を補正する方法として，層別ログランク検定がしばしば用いられる．しかし，多数の予後因子で層別すると層内標本数は減少し，層別ログランク検定の検出力は低下する．少数の予後因子だけで層別すると層内の症例不均一が増大し，この場合もまた層別ログランク検定の検出力は低下する．層内標本数と層内不均一とのバランスを考え層別ログランク検定の検出力の低下が最も小さくなる条件を見いだすことはほとんど不可能である．予後因子の影響を補正する別の方法として，多変量生存時間回帰モデルの適用がある．従来の回帰モデルによる治療効果判定では，モデルの適合性が良くないと治療効果の大きさを正しく推定できなかったり検出力が低下することが知られている．そこで，これまでのモデルよりも予後因子と生存時間との関係を柔軟に関係づけることができ実際のデータに良く適合するモデル構築法を提唱する．

治療効果判定に用いられる Cox の比例ハザードモデ

ルは以下の式で表される．

$$\lambda(t \mid X, Z) = \lambda_0(t)\exp(\alpha X + f(\beta, Z))$$

ただし，$\lambda_0(t)$ は未知の正値関数，X は治療群を識別する因子，たとえば，X＝0 が標準治療群，X＝1 が新治療群，$f(\beta, Z)$ は，年齢，性別，臨床進行期などを成分に持つ予後因子ベクトル Z とそのパラメータベクトル β で作られる関数，α は治療群を識別する因子の回帰係数である．また，$\lambda(t \mid X, Z)$ は X，Z の情報が与えられた下でのハザード（瞬間死亡率）である．

通常の治療効果判定では，$f(\beta, Z)$ は Z の一次関数かまたは2次関数が用いられる．即ち，Z が1変数のとき，$f(\beta, Z) = \beta Z$ または $\beta_1 Z + \beta_2 Z^2$（いずれも定数項は $\lambda_0(t)$ に吸収される）である．ところが，実データでの予後因子 Z と対応する対数ハザード $f(\beta, Z)$ の関係は，必ずしも直線や2次関数で近似できるとは限らない．たとえば，臨床進行期ⅠとⅡとの間では死亡率がそれほど変らないが，ⅡとⅢの間ではⅡに比べⅢが極端に高く，ⅢとⅣの間ではまた死亡率に差はないという場合，対数ハザードは S 字型関数となり，1次関数や2次関数では近似できない．このように，データに不適合なモデルを仮定して治療効果判定を行っても治療効果を正しく検出できない．モデルの適合性を良くしようとして3次，4次などパラメータを多く含む関数を用意してモデルを構築すれば適合度は高まるが安定した結果が得られない．つまり，必要最小限のパラメータを用いてデータに適合するモデルを構築し治療効果判定の検出力を低下させない検定手法が望まれる．そこで，折れ線ハザードモデルによる Cox 回帰法を提唱する．

説明を簡略化するために，Z は4つのカテゴリー 0，1，2，3 をもつとする．このときの折れ線ハザードモデルは，上述の Cox の比例ハザードモデルにおいて $f(\beta, Z)$ を以下のように定義する．

$$f(\beta, Z) = \beta_0 Z + \beta_1 \langle Z-1 \rangle + \beta_2 \langle Z-2 \rangle$$

ただし，$\langle Z-1 \rangle$，$\langle Z-2 \rangle$ は，それぞれ，1，2 を折曲点とする折れ線関数であり，

$$\langle Z-i \rangle = \{(Z-i) + |Z-i|\} \times 0.5 \quad (i = 0, 1, 2)$$

である．折れ線ハザードモデルによるハザード近似のしくみを**図5**に示した．**図5**の左のグラフは，実データにおける対数ハザードの変化であるが，この変化パターンを右の2つの折れ線関数を使って近似する．このような近似方法により，**図6**に示すいろいろなハザード変化を柔軟に近似できる．

折れ線ハザードモデルによる Cox 回帰法を用いることの長所をまとめると以下のようになる．

（7）多数の予後因子の影響を補正した上で治療効果判定を行える．即ち，既知要因に関する不均一性の影響を除去できる．

（8）すべての予後因子について，（カテゴリーの数－1）個の変数を作りモデル化すると，変数の数が多くなりすぎ検出力の低下につながる．そこで，回帰分析の逐次変数増減法を使って，折れ曲がり方が有意な変数のみをモデルに残し他の変数はモデルから除外する．**図6**の左上にある直線の場合には，$\beta_0 Z$ のみがモデルに残り他の2つの折れ線関数は除外される．また，**図6**の右下の Borrmann 分類では，0と1で折れ曲がる折れ線関数のみをモデルに残し他は除外される．

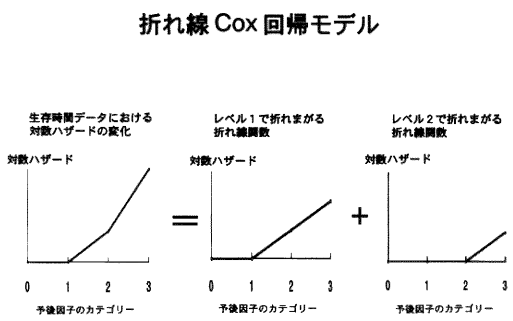

図5　折れ線ハザードモデルによる対数ハザード近似のしくみ

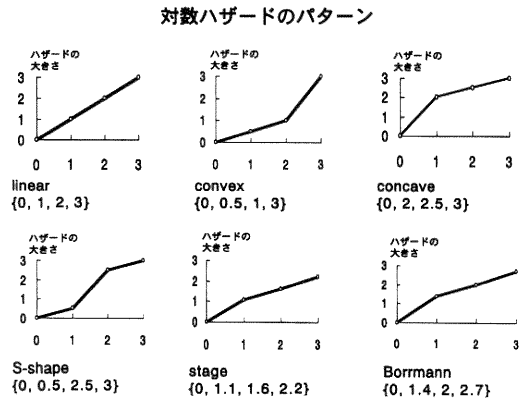

図6　実際の臨床試験データで生じうる対数ハザードパターン

表2　胃癌臨床試験データ解析に用いた因子

解析に用いた因子	カテゴリー	折れ線 Cox 回帰法で使った変数
1. ステージ	0，1，2，3	STG, STG 1, STG 2
2. ボルマン分類	0，1，2，3	BORR, BORR 1, BORR 2
3. 手術時年齢	0，1	AGE
4. 浸潤増殖様式	0，1，2	INF, INF 1
5. リンパ管侵襲	0，1，2，3	ly, ly 1, ly 2
6. 静脈侵襲	0，1，2，3	v, v 1, v 2
7. 腫瘍長径	0，1，2，3，4	LNG, LNG 1, LNG 2, LNG 3
8. 郭清度	0，1，2	R, R 1
9. ステージⅠ，Ⅱ症例への PSK 投与	0，1	LPSK
10. ステージⅢ，Ⅳ症例への PSK 投与	0，1	HPSK
11. ステージⅠ，Ⅱ症例への OK－432 投与	0，1	LOK
12. ステージⅢ，Ⅳ症例への OK－432 投与	0，1	HOK

（例）STG 1＝｛STG－1｝＝（(STG－1）＋｜STG－1｜）/2

折れ線ハザードモデルの治療効果判定への応用

不均一の程度を推定する際に述べたがん集学的治療研究財団の特定研究1胃癌臨床試験データに折れ線ハザードモデルを適用する．OK－432，PSK，OK－432＋PSK の治療効果を判定するために，従来の治療効果判定方法である単純ログランク検定と層別ログランク検定を行った．層別ログランク検定は，臨床進行期，ボールマン分類，手術時年齢で層別された．4群比較の結果，単純ログランク検定では $p=0.50$，層別ログランク検定では $p=0.11$ でありいずれも有意な治療効果を認めないということになる．

表2に挙げる因子を候補因子とし前節で述べた方法でそれぞれのカテゴリーに対して折れ線関数を作り，それらをすべて治療効果判定の候補因子とする Cox 回帰法を行った．その結果，臨床進行期，ボールマン分類，手術時年齢，リンパ管侵襲，静脈侵襲，腫瘍長径などと共に，組織学的臨床進行期Ⅲ，Ⅳ期の群に対する PSK 投与の有意な延命効果が検出された．

このように，生存時間を評価尺度とする臨床試験の治療効果判定において，検出力の低下を防ぎ多数の予後因子の影響を補正する方法として，折れ線ハザードモデルによる Cox 回帰法は有用であると考える．

謝　辞

本講演の座長の労をおとりいただきました新潟大学医学部検査診断学岡田正彦教授に心よりお礼申し上げます．

参　考　文　献

1) **がん集学的治療研究財団**: 特定研究1　研究報告書（井口　潔編）: 1992.
2) **Akazawa, K., Nakamura, T. and Palesch, Y.**: Power of logrank test and Coxregression model in clinical trials with heterogeneous samples. Stat. Med. **16**: 583～597, 1997.
3) **Akazawa, K., Nakamura, T. and Nose, Y.**: A statistical model that takes into account patient heterogeneity in decision making. Medinfo. **9 PT 1**: 525～528, 1998.
4) **Simon, R.**: Importance of prognostic factors in cancer clinical trials. Cancer Treat. Rep. **68**: 185～192, 1991.
5) **Sather, H.**: The use of prognostic factors in clinical trials. Cancer **58**: 461～467, 1986.
6) **Kamura, T., Tsukamoto, N., Tsuruchi, N., Saito, T., Matsuyama, T., Akazawa, K., Nakano, H.**: Multivariate analysis of the histopathologic prognostic factors of cervical cancer in patients undergoing radical hysterectomy. Cancer **69**: 181～186, 1992.
7) **Ishida, T., Kaneko, S., Akazawa, K., Tateishi, M., Sugio, K., Sugimachi, K.**: Proliferating cell nuclear antigen expression and argyrophilic nucleolar organizer regions as factors influencing prognosis of surgically treated lung cancer patients. Cancer

Res. **53**: 5000～5003, 1993.

8) **Tanabe, G., Sakamoto, M., Akazawa, K., Kurita, K., Hamanoue, M., Ueno, S., Kobayashi, Y., Mitue, S., Ogura, Y., Yoshidome, N., Baba, M., Aikou, T.**: Intraoperative risk factors associated with hepatic resection. Br. J. Surg. **82**: 1262～1265, 1995.

質疑応答

若井先生 臨床論文で Cox 回帰法などの多変量解析を行うとき，症例数として何例必要か？

赤澤 Cox 回帰法を行うときの必要症例数については多くの研究がなされている．一般に，統計理論に基づく必要症例数算出のための公式を用いる方法とデータの特質を考慮に入れたシミュレーションを行う方法とがある．前者では公式導出のためさまざまな仮定がなされており，実際の臨床試験データの必要症例数算出には適さない場合もある．私は，後者のシミュレーション法を推奨している．シミュレーションを行うための簡易なソフトウエアも用意されている．

白井先生 現在，統計解析を StatView で行っているが問題はないか？2×2表の直接確率法を使ったところ，別のソフトウエアの解析結果と異なっていた．もっと高性能のソフトウエアを使用する必要があるか？

赤澤 Statview で処理可能な統計処理は，Statview で処理して構わないと思う．直接確率法の検定結果の食い違いについては，詳細な検討が必要である．直接確率法は計算量も多く統計ソフトウエアでもサンプル数などに制限をかけているものが多い．もし，心配であれば直接確率法を専門に扱ったソフトウエアがあるのでそれを使うことをお奨めする．

岡田先生 折れ線ハザードモデルにおいて，各予後因子について（カテゴリー数－1）個の変数を作っていくわけだからパラメータが膨大になる．逐次変数法の適用だけで妥当なモデルは本当に得られるといってよいか？

赤澤 逐次変数法による折れ線ハザード関数の取捨選択が妥当かどうかは，別途シミュレーションを行った．シミュレーションの結果から，**図6**で示された程度の凸関数，凹関数，S 字曲線などに対しては，パラメータの推定値とその標準誤差，治療効果判定の検出力に関して妥当な結果を得ている．

岡田先生 医学研究者が統計解析を行う際の注意点は？

赤澤 最近では，対話型の使いやすい統計解析ソフトウエアが市販され研究室でも先生方が解析されるケースが増えてきた．統計解析結果が独り歩きしないように，統計手法を使う際の条件をよく調べてから使用すべきである．統計解析した結果を論文にまとめて投稿したら，米国の査読者から統計解析方法に問題ありとの指摘を受けた，という相談を年に数十件承る．医療統計学者が数千人いる欧米では，統計解析に関するコンサルテーションや依頼は日常茶飯事に行われ，査読者からクレームが付けられた場合には責任を持ってそのクレームの処理にあたる．データ解析を行う前，できればデータ収集する前に実験計画やデータ収集方法，解析の手順などを打ち合わせさせていただければ先生方の損害も少なくなると考える．