〈教育報告〉

平成 21 年度専門課程Ⅱ
生物統計分野

特発性肺線維症の急性憎悪の予後因子に関する統計学的検討

勝田知也

Statical Analyses of Prognosis Factors in the Acute Exacerbation of Idiopathic Pulmonary Fibrosis

Tomoya KATSUTA

抄録

目的　特発性肺線維症（IPF）は，予後不良な急性増悪（AE）という病態を引き起こす頻度が高く，有効な治療が乏しい難病である．AE の防止効果があるとされる治療薬が注目されており，AE を起こす予後因子について検討することが必要とされる．本研究では AE と FVC（努力性肺活量）のベースラインからの低下率との関連について統計学的に検討することを目的とした．

方法　IPF 患者 74 名について後ろ向きに情報を収集し，AE を起こすまでの時間を結果変数，FVC のベースラインからの低下率の影響を Cox 比例ハザードモデルで非時間依存性共変量，時間依存性共変量として取り扱い，あわせて個人差指数を利用した解析を行った．低下率と個人差指数に関する感度分析を行った．

結果　FVC の低下率と個人差指数による異常低値は時間依存性共変量とした場合に AE に対する有意な予後因子として検出された．さらに個人差指数を導入することで，個人ごとの FVC の生理的変動範囲からの逸脱をより鋭敏に判別できる可能性が示唆された．

結論　AE に及ぼす予後因子に関して，時間依存性共変量を用いた解析の重要性および個人差指数の臨床的な応用の可能性が示唆された．

キーワード： 特発性肺線維症，急性増悪，Cox 比例ハザードモデル，時間依存共変量，個人差指数

Ⅰ．背景

特発性肺線維症（IPF）は，生存時間の短縮をもたらす急性増悪（AE）という病態を引き起こす頻度が高く，有効な治療が乏しい難病である．AE 防止効果があるとされる治療薬が注目されており，AE を起こす予後因子について検討し AE の予防を可能にすることが必要とされる．本研究では FVC（努力性肺活量）のベースラインからの低下率の影響について統計学的に検討することを目的とした．

Ⅱ．目的

本研究の目的は，AE の予後因子として FVC のベースラインからの低下率を取り上げ，次の 3 点を明らかにすることである．

1）　King[1] らが提唱した FVC のベースラインからの 10% 以上低下が，AE の予後因子として有意な関連を示すかについて，非時間依存性共変量として取り扱い，Cox 比例ハザードモデル（CPHM）により分析する（表の model1).

2）　FVC のベースラインからの 10% 以上低下を時間依存性共変量として取り扱い，CPHM により検討する．また，この際，他の低下率（5%，15%，20%，25%）についても合わせて分析する（10% 以上低下を表の model2）．

3）　FVC の AE 前の生理的変動範囲（変動範囲）を，個人差指数 2）3）を用いて評価し，その逸脱の有無を時間依存性共変量として取り扱い，CPHM により分析する（個人差指数 0.75，個体内分散 0.74 の場合が表の model3）．この個人差指数，個体内分散は安定期の COPD（慢性閉塞性肺疾患）患者から求めたため，その妥当性を検討する目的で，個人差指数と個体内分散に関する感度分析を行う．

指導教官：山岡和枝，丹後俊郎（技術評価部）

Ⅲ．研究デザイン

単一の病院で IPF と診断された患者 74 名をレトロスペクティブに分析した．患者の臨床背景として年齢，診断時の BMI（Body mass index），FVC（経時的データを含む），動脈血酸素分圧，肺胞洗浄液の文画，性別，喫煙歴（診断時までの喫煙の有無），外科的肺生検の有無，MRC（British Medical Research Council によるスコアー；本研究では MRC が 2 以上か否かで 2 区分に分類），および診断から AE までの期間を取り上げた．結果変数は急性増悪の有無とした．

表 model ごとの解析結果

		非時間依存性共変量	時間依存性共変量	
		HR[1] (95% CI)		
		Model1[6]（多変量）	Model2[2]（多変量）	Model3[3]（多変量）
FVC の低下	推定確率[6]	1.00 (0.98-1.03)		
	g1(t|10)[2]		4.86 (1.51-15.7)	
	g2(t|10)[3]			5.33 (1.90-15.0)
年齢[4]		0.95 (0.88-1.03)	0.98 (0.92-1.05)	1.00 (0.94-1.06)
BMI[4]		1.26 (1.02-1.56)	1.2 (1.01-1.43)	1.27 (1.06-1.52)
ベースライン FVC[4]	（ℓ）	0.37 (0.13-1.07)	0.92 (0.31-15.9)	0.87 (0.34-2.18)
性別[5]	男性	10.2 (1.19-87.9)	2.26 (0.32-15.9)	4.13 (0.65-26.4)
MRC2 以上[5]	あり	2.79 (0.96-8.09)	2.38 (0.79-6.54)	2.58 (0.93-7.17)
外科的肺生検[5]	あり	1.23 (0.39-3.90)	0.55 (0.22-1.37)	0.89 (0.32-2.48)
喫煙歴[5]	あり	0.27 (0.07-1.00)	0.83 (0.22-3.16)	0.46 (0.19-2.22)
追跡期間（月）[5]	イベントあり	42.3(25.3) [9.3,111.5]	25.0(22.5) [1.0,104.2]	21.3(17.6) [0.2,82.2]
	イベントなし	45.1(28.6) [7.8,104.2]	35.2(25.6) [5.0,111.5]	27.2(20.2) [0.1,96.7]

1) HR; ハザード比
2) FVC がベースラインより 10% 以上低下の有無の時間依存共変量
3) 個人差指数 0.74，個体内分散 0.54 で変動範囲からの逸脱と判定された有無の時間依存共変量
4) 連続量
5) カテゴリー値
6) ベースラインから 10% 以上低下をイベントとした Kaplan-Meier 推定量

Ⅳ．結果

Model1 では AE に関する因子は MRC スコアー 2 以上と BMI が選ばれた（表第 1 列）．Model2 では 10% 以上低下した場合は AE に関連を認めた（表第 2 列）．そのハザード比は低下率が上昇すると増加傾向にあった．Model3 では個人差指数による変動範囲からの逸脱も同様に有意な関連を認めた（表第 4 列）．さらに個人差指数や分散の値にかかわらず推定されるハザード比はほぼ一定であった．

Ⅴ．考察

非時間依存性共変量として解析した場合，FVC の低下は有意な予後因子として検出できなかった．しかし時間依存性共変量として解析した結果，それは有意な関連が認められ，時間依存性を考慮した解析を行う必要があると考えられた．個人の変動がその集団での変動よりも小さいことに着目し，個人差指数を用いて，個人ごとの変動範囲からの逸脱を検出することが可能であった．その有無を時間依存性共変量として解析したところ，同様に有意な関連が認められた．個人差指数，個体内分散を変化させて感度分析を行ったがハザード比もほぼ一定であることから個人差指数の妥当性が示唆された．

Ⅵ．まとめ

AE と関連する予後因子について，レトロスペクティブに CPHM を用いて検討したところ，FVC のベースラインからの低下率は，時間依存性共変量とした場合に有意な予後因子として検出され，さらに個人差指数を導入することにより，個人ごとの変動範囲の逸脱をより鋭敏に判別する可能性が示唆された．

文献

1） King TE Jr, Safrin S, Starko KM, Brown KK, Noble PW, Raghu G et al Analyses of Efficacy End Points in a Controlled Trial of Interferon- γ 1b for Idiopathic Pulmonary Fibrosis. Chest 2005; 127:171-177
2） 丹後俊郎：臨床検査の個人差指数．臨床病理．28:789-793, 1980
3） 丹後俊郎：医学データ . 共立出版 ; 2002　65-68

〈教育報告〉

平成 21 年度専門課程 Ⅱ
生物統計分野

部分的区間打ち切りデータにおける二標本検定法の比較と評価に関する研究

川口修

Comparison and Assessment of 2-Sample Tests for Partly Interval-Censored Failure Time Data

Osamu KAWAGUCHI

抄録

背景と目的　臨床試験において区間打ち切りデータに対する 2 標本検定を行う場合，初めてイベントが観察された時点（発生区間の右端）をイベント発生時点として補完し，log-rank 検定などを適用する方法が一般的に使用されているが，他に提案されている 2 標本検定の方法との性能比較や，特性の検討などはほとんどされていない．そこで本研究では，一点代入（左端 / 中点 / 右端）による方法と，2 種類の方法（Pan の方法，Zhao らの方法）の性能と性質を比較し，それぞれの特徴について検討を行う．

研究デザインと方法：現実の臨床試験の状況に則した様々なシナリオを想定したシミュレーションを実施し，各検定法の性能と性質を検出力およびサイズを指標として比較，考察した．

結果と考察　最後に観察された時点を右側打ち切りの発生時点とみなす通常の状況の下で，左端代入が最も安定した方法であることが示唆された．また右側打ち切りが多い状況では，特に右端代入法は誤った結論を導く可能性があることが示された．Zhao らの方法を実際の臨床試験に適用するためには，さらなる性能調査が必要であることも示唆された．

キーワード：生存時間解析，区間打ち切りデータ，ログランク検定，生存時間関数

Ⅰ．背景と目的

医学領域では，死亡や治癒までの期間など何らかの関心のあるイベントが発現するまでの時間に関する評価を行うことが多く，その解析手法として生存時間解析が広く応用されている．このようなイベント発現までの時間を評価する臨床研究では，イベント発生時点が正確に観察される場合もあるが，通常は検査や診断によって初めてイベントと判断される場合が多い．その場合，イベント発生時点は，発生「なし」と判定された最後の検査時点から，発生「あり」と判定された最初の検査時点までの間にあるため，イベント発現までの時間が「区間打ち切りデータ」として得られる状況がしばしば存在する（図 1）．

なお，図 1 の例のように正確なイベント発生時間が得られるデータ（被験者 3）と，区間打ち切りデータ（被験者 1, 2）が混在している場合は，特に「部分的区間打ち切りデータ」とよばれる．区間打ち切りを含むデータの群間比較を行う場合には，初めてイベントが観察された時点（発生区間の右端）を発生時点として補完し，正確な時点が得られた場合と同様に log-rank 検定などを行う方法が一般的に用いられている．また一方では，多重代入法を用いた Pan

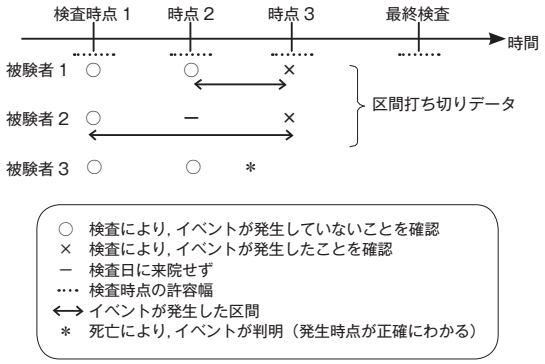

図 1 区間打ち切りを含むデータ

指導教官：飛田英祐，西川正子，丹後俊郎（技術評価部）

の方法[1]や，log-rank 検定の拡張法である Zhao らの方法[2]など，Turnbull[3]が提唱した区間打ち切りデータによる生存時間関数の推定法および 2 標本検定法の考え方にもとづいた，イベント発生区間内の一時点による補完（以降,「一点代入法」と記す）を用いない方法もいくつか提案されている．

しかしながら，これらの検定方法に関する性能比較や，性質の検討はほとんどなされておらず，ただ簡便であるという理由だけで，主に右端を代入する一点代入法が広く使われているのが現状である．そこで本研究では，一点代入法（発生区間の左端，中間時点，右端の時点を代入して，log-rank 検定を行う方法），Pan の方法，および Zhao らの方法の計 5 種類の検定方法の性能，性能を比較するためのシミュレーションを行い，これらの検定法がどのような状況下でどのような特徴を持つか，検定の検出力およびサイズを指標として評価，考察した．

Ⅱ．研究デザインと方法

西川ら[4]が実際の臨床試験を模して作成したシミュレーションデータをもとに，現実の臨床試験において想定される状況に則した様々なシナリオを設定し，計 5 種類の検定方法を比較するためのシミュレーションを行った．その結果と，区間打ち切りデータを生成させる元データ（全イベントおよび右側打ち切り例の打ち切り発生時間が，すべて正確に観測されているデータ．以降,「発生させた真のデータ」と記す）を log-rank 検定を用いて検定した結果を比較し，各検定方法の性能および性質を検討した．

Ⅲ．結果

いずれのシナリオにおいても，左端代入法の結果が，発生させた真のデータにおける結果に近く，最も安定していた．また，中点代入法と Pan の方法の結果はいずれのシナリオにおいてもほぼ同様の結果であった．

最終検査時までに期待される右側打ち切りが多い（50%）シナリオでは，右端代入法の検出力が特に高い結果が得られたが，発生させた真のデータにおける結果と異なる結果を得る割合（以下,「不一致率」と記す）が他の方法よりも高かった．

Zhao らの方法は，いずれのシナリオにおいてもサイズ 0.05 を保っておらず，検出力も他の方法と比較して低い結果であった．

Ⅳ．考察

いずれのシナリオにおいても，左端代入法が最も安定した方法であることが示唆する結果が得られたが，これは，右側打ち切りが発生した際に，最後に観察できた検査時点（打ち切り原因の発生区間の左端）を打ち切り時点として採用していることが原因だと考えられる．左端代入法を用いた場合では，打ち切り時点とイベント発生時点がともに，発生区間の左端にシフトされることとなるため，イベント発生数と観察対象者（at risk 数）のバランスが大きく乱れることがなく，安定した 2 群比較の結果が得られると考えられる．

次に，右側打ち切りが多いシナリオにおいて右端代入法の不一致率が増加する原因としては，観察時期の後半において，生存率が低い群ほど，高い群と比較して生存率を過小評価してしまう傾向が強く，結果として 2 群間の差を過大評価することが考えられた．

以上より，実際の臨床試験において一般的に使用されている，イベント発生時点を発生区間の右端で代入し，打ち切り時点を左端で代入する方法は，代入による見かけ上の差までを検出してしまうことで，2 群比較において誤った結論を導く可能性が示唆されたが，本研究はシミュレーションによる検討であるので，理論的な原因については今後より詳細に検討したい．

なお，Zhao らの方法は，今回想定したすべてのシナリオにおいてサイズ 0.05 を保てず，検出力も低い状態であった．今回，観察期間の終了の設定有無が Zhao らの方法の性能に影響を与える可能性を示唆する結果も得たれたため，理論的な検討も含めて今後，原因を検証したい．

Ⅴ．まとめ

計 5 種類の検定方法の性能，性質を比較した結果，左端代入法が最も安定した方法であることが示唆され，右側打ち切りが多い場合には，右端代入法は誤った検定結果を導く可能性があることが示された．

また，新しい検定法である Zhao らの方法を実際の臨床試験の解析に使用するためには，さらなる性能，性質の調査が必要である．

文献

1） Pan W. A two-sample test with interval censored data via multiple imputation. Statistics in Medicine 2000; 19(1):1-11

2） Zhao X, Zhao Q, Sun J, Kim JS. Generalized Log-Rank Tests for Partly Interval-Censored Failure Time Data. Biometrical Journal; 50(3):375-385

3） Turnbull, B. W. The empirical distribution with arbitrarily grouped censored and truncated data. Journal of the Royal Statistical Society 1976; Series B 38, 290–295.

4） Nishikawa, M., and Tango, T. Behavior of the Kaplan-Meier Estimator for Deterministic Imputations to Interval-Censored Data and the Turnbull Estimator. Japanese Journal of Biometrics. 2003; 24(2):71-94

〈教育報告〉

平成 21 年度専門課程Ⅱ
生物統計分野

複数の読影者による診断法の比較のための対応のあるカテゴリカルデータの統計的推測

佐伯浩之

Statistical Inference of Matched-Pair Categorical Data for the Comparison of Diagnostic Tests Involving Multiple Readers

Hiroyuki SAEKI

抄録

目的　非劣性試験において，試験実施施設とは独立した複数の読影者から，reference standard の結果が盲検化されたもとで得られた読影結果に基づき，二つの診断法の正診率を比較するための新しい手法を提案する．

方法　試験デザインとして，各患者が二つの診断法を受け，全ての読影者が全画像を評価する状況を想定した．本試験デザインから発生する対応のあるカテゴリカルデータに対して多項分布を仮定し，二つの診断法の正診率の差のエフィシェントスコアに基づく非劣性検定とスコア信頼区間を誘導した．さらに，モンテカルロシミュレーションによる本提案法の妥当性の検討と，眼科学研究のデータに対する本提案法の応用を行なった．

結果・考察　モンテカルロシミュレーションにより，本提案法は既存法に比べて第一種の過誤が名目の有意水準に近いことが確認された．また，検出力についても本提案法は既存法と大きく異ならないことが示された．

キーワード：　対応のあるカテゴリカルデータ，複数の読影者，非劣性，スコア検定

Ⅰ．目的

二つの診断法を比較する臨床試験では，真の状態を正しく診断する確率である正診率（又は感度，特異度）が指標として利用される．この正診率は，診断法以外の情報の混入によるバイアスの発生が考えられることから，試験実施施設とは独立した複数の読影者が，reference standard の結果を盲検化した上で読影する方策が取られている．一方，新規の診断法を従来の診断法と比較して非劣性を示すことを目的とした臨床試験では，Tango[1] により導出された対応のあるカテゴリカルデータの非劣性検定を利用できるが，Tango の検定は複数の読影者による読影結果を総合的に利用することができない．そこで本研究では，複数の読影結果を総合的に利用して非劣性を確認することを目的とした，対応のあるカテゴリカルデータの新しい統計手法の導出を検討した．

Ⅱ．方法

1. 試験デザイン

各患者に実施した診断法 A 及び診断法 B について，k 名 $(1,\cdots,K)$ の読影者による読影結果に基づく正診率を比較する状況を，本研究で検討する統計手法を利用する試験デザインとする．

2. 検定と信頼区間

非劣性の帰無仮説である

$H_0: \delta = P_A - P_B + \Delta \le 0$

に対して，式 (1) のスコア検定を導出した．

$$Z = \frac{\frac{1}{K}\sum_{k=1}^{K} k(\hat{P}_{Xk} - \hat{P}_{Yk}) + \Delta}{\sqrt{\frac{1}{nK^2}\left[K\tilde{q} + \tilde{r} + \tilde{s} - K^2\Delta(1+\Delta)\right]}} \qquad (1)$$

指導教官：　丹後俊郎（技術評価部）

ただし $\tilde{q}$, $\tilde{r}$, $\tilde{s}$ は

$$\tilde{q}=\left(\sum_{k=1}^{K} k\tilde{P}_{Yk}-\sum_{k=1}^{K-1} k\tilde{P}_{Xk}\right)$$

$$\tilde{r}=\sum_{k=1}^{K-1} k^2\tilde{P}_{Xk}$$

$$\tilde{s}=\sum_{k=1}^{K} k^2\tilde{P}_{Yk}$$

である．ここで，式 (1) の P_{Xk} は診断法 B に比べて診断法 A で正しい診断であった読影者の人数が k 名多い確率，P_{Yk} は診断法 A に比べて診断法 B で正しい診断であった読影者の人数が k 名多い確率，さらに P_{Xk} 及び P_{Yk} は帰無仮説のもとでの最尤推定量であり，対数尤度関数から制約付き準ニュートン法を利用して求められる．

スコア検定に対応する 100(1-α)% 信頼区間は，

$$\frac{\frac{1}{K}\sum_{k=1}^{K} k(\hat{P}_{Xk}-\hat{P}_{Yk})-\lambda}{\sqrt{\frac{1}{nK^2}\left[K\tilde{q}+\tilde{r}+\tilde{s}+K^2\lambda(1-\lambda)\right]}}=\pm Z_{\alpha/2} \qquad (2)$$

における λ の解として求められ，右辺の正の符号が下側信頼区間，負の符号が上側信頼区間に対応する．この信頼区間の計算には secant method が利用できる．

3. シミュレーション

読影者が 2 名の状況を想定し，本研究の提案法と Durkalski ら[2] の Wald 型検定統計量における第一種の過誤及び検出力の調査を，モンテカルロシミュレーションにより実施した．シミュレーションにおいて，第一種の過誤の検討では $\pi_A-\pi_B+\Delta=0$，$\Delta=0.1$，検出力の検討については $\pi_A-\pi_B+\Delta=0.1$，$\Delta=0.1$ とし，いずれの検討も有意水準上側 2.5% の条件で繰り返し数 10,000 回とした．

Ⅲ．結果

1. シミュレーション

1）第一種の過誤の検討

有意水準上側 2.5% のもと，本研究の提案法は概ね名目の水準を保っていたが（1.0 – 3.1%），Durkalski らの検定は名目の水準を保てない事例が多く認められた（2.2 – 3.9%）．

2）検出力の検討

本研究の提案法と Durkalski らの検定は概ね同程度の検出力を示した．

3）実データへの応用例

840 名の患者の両目における地図状萎縮の有無について，同一の 2 名の読影者が評価を行なったデータ[3] を用いて，読影者 A の陽性率の読影者 B に対する非劣性を検討した．非劣性マージン $\Delta=0.01$ としたときの，本研究の提案法による結果は P=0.328（Z=0.446），読影者 A と読影者 B の陽性率の差の 95% 信頼区間（lower, upper）は（−0.0166，−0.0009）であった．

Ⅳ．考察

第一種の過誤は，本研究の提案法が Durkalski らの検定方法に比べて名目の有意水準上側 2.5% を概ね保つことが確認された．また，検出力については，本研究の提案法と Durkalski 検定は概ね同程度の検出力を示した．

Ⅴ．まとめ

本研究では，二つの診断法の比較で非劣性を検証する臨床試験において，試験実施施設とは独立した複数の読影者から得られた読影結果を総合的に利用するスコア型の検定及び信頼区間を誘導し，本研究で誘導した検定が優れた性質を示すことをシミュレーションにより確認した．

参考文献

1） Tango T. Equivalence test and confidence interval for the difference in proportions for the paired-sample design. Stat Med. 1998;17:891-908.

2） Durkalski VL, Palesch YY, Lipsitz SR, Rust PF. Analysis of clustered matched-pair data for a non-inferiority study design. Stat Med. 2003;22:279-90.

3） Oden NL. Estimating kappa from binocular data. Stat Med. 1991;10:1303-11.

〈教育報告〉

平成 21 年度専門課程Ⅱ
生物統計分野

癌領域の第Ⅱ相臨床試験の多段階デザインにおける早期無効中止基準のカットオフ値の検討

中川智文

Evaluation of the Cut-Off Points of the Early Stopping Rules for Futility of the Multi-Stage Design of Phase II Clinical Trials in Oncology

Tomofumi NAKAGAWA

抄録

背景および目的　抗癌剤の第Ⅱ相臨床試験は，通常１群無対照で実施し，対象とする癌腫に対して治験薬を組み入れた新しい治療と既存の標準的治療との比較を行う第Ⅲ相試験等のさらなる評価を行うべきかについて判断するために実施される．一方，Wathen *et. al.*（2008）は，２値または time-to-event データに対し予後因子として２つ以上のサブグループを有した１群無対照の第Ⅱ相臨床試験において，薬剤・サブグループ交互作用が存在する場合を考慮したある種のモデルベースのベイズデザインを報告した．本研究では，最近の癌領域における新規治療の開発状況を考慮し，効果予測因子としてバイオマーカーを用いた分子標的療法の開発を想定する．ある種のバイオマーカーの発現の有無により２つのサブグループを有し，主要エンドポイントを奏効率とした癌領域の第Ⅱ相臨床試験において，Wathen により報告された方法の早期無効中止の基準値（カットオフ値）を検討することとした．

方法　Wathen の方法を用いて，新薬が有効である事後確率のカットオフ値を偽陽性および偽陰性による全損失より検討した．

結果および考察　薬剤・サブグループ交互作用が存在する場合，偽陽性率に重点をおくと新薬が有効である事後確率のカットオフ値は 0.150-0.175 が妥当であることが示された．

キーワード： 早期無効中止基準のカットオフ値，癌領域の第Ⅱ相臨床試験，ベイズデザイン，薬剤･サブグループ交互作用，シミュレーション

Ⅰ．目的

1980 年代までは有望な新規抗癌剤を見つけることは困難で，抗癌剤開発においては，本来有効な薬剤を見落としてしまう危険性（第２種の過誤）を小さくすることに関心が払われてきた．多段階デザインにおいても「積極的に無効が示せない場合は，有効と判断する」という考えの下，新薬が有効である事後確率のカットオフ値が小さく（早期無効中止になりにくいように）設定されてきた．しかし，現在では効果予測因子としてバイオマーカーを用いた分子標的薬（ある癌腫に対応するバイオマーカーの（過剰）発現の有無により治療への反応性が良い・悪いを予測可能）等，次から次へと新規抗癌剤が開発されており，むしろ本来無効な薬剤を有効と判断してしまう危険性（第１種の過誤）が問題となっている．現在の抗癌剤開発の状況および本来の中間解析の目的の１つとして「無効な薬剤は早期に試験を中止すべきである」という倫理面を考慮すると，新薬が有効である事後確率のカットオフ値を現行よりも大きく設定すべきではないかと考えられる．一方，Wathen *et. al.*[1]（2008）は，多段階デザインにおいて，薬剤・サブグループ交互作用を考慮した回帰モデルを基にベイズ理論による事後確率を用いサブグループごとに早期無効中止の有無を判定する方法を報告した．以上より，本研究では現在の抗癌剤開発の状況を考慮し，ある癌腫に対応するバイオマーカーの（過剰）発現の有無により２つのサブグループを有した抗癌剤の第Ⅱ相臨床試験において，薬剤およびバイオマーカーによるサブグループ交互作用が存在する場合を想定し，Wathen（2008）が報告した解析手法についてシミュレーションを行い，偽陽性および偽陰性の発生による全損失より新薬が有効である事後確率のカットオフ値を検討することとした．

指導教官：　西川正子，丹後俊郎（技術評価部）

Ⅱ．研究デザインと方法

Wathen (2008) と同様に，薬剤・サブグループ交互作用を考慮した回帰モデルを基にベイズ理論を用いサブグループごとに早期無効中止の有無を判定する．*n* 症例分のデータが得られた時点で，サブグループごとの新薬が有効である事後確率（期待奏効率が閾値奏効率＋期待する改善率を上回る事後確率）が，カットオフ値未満であった場合，そのサブグループに対する新規治療は無効であると判定し，早期無効中止とする．本研究におけるシミュレーションの奏効率の真値は，以下のシナリオのとおり設定し，カットオフ値は，0.010-0.200 の範囲で検討した．

薬剤・サブグループ交互作用有り

シナリオ 1:　0.60（予後良好：有効）
　　0.25（予後不良：無効）
シナリオ 2:　0.45（予後良好：無効）
　　0.40（予後不良：有効）

Ⅲ．結果

全損失を以下の式で表しカットオフ値の最適値を決定した．

全損失＝（C_1 ×偽陽性率）＋（C_2 ×偽陰性率）

但し，C_1 は偽陽性による平均損失を，C_2 は偽陰性による平均損失を示す．なお，本研究では偽陽性率に重点をおいており，頻度論を用いた臨床試験では，第 1 種の過誤を 0.05 に，第 2 種の過誤を 0.10 に設定していることが多いことを考慮し，C_1 および C_2 の比を（1.2, 1）から（2, 1）とし，結果を表 1 に示した．

表 1　C_2 に対する C_1 の重みを変化させた場合の全損失およびカットオフ値の最適値

C_1	C_2	全損失	最適値
シナリオ1			
1.2	1.0	0.589	0.075
1.4	1.0	0.621	0.150
1.6	1.0	0.636	0.150
1.8	1.0	0.651	0.150
2.0	1.0	0.666	0.150
シナリオ2			
1.2	1.0	0.741	0.100
1.4	1.0	0.784	0.175
1.6	1.0	0.818	0.175
1.8	1.0	0.851	0.175
2.0	1.0	0.885	0.175

表 1 より，偽陽性率および偽陰性率の重みを（1.2, 1）から（2, 1）に変化させた場合，シナリオ 1 は，（1.2, 1）以外でカットオフ値は 0.150，シナリオ 2 は，（1.2, 1）以外でカットオフ値は 0.175 において全損失が最小となり，また安定していた．

Ⅳ．考察

薬剤・サブグループ交互作用は有りとした際のカットオフ値の最適値は 0.150-0.175 と考えられる．想定とは異なり薬剤・サブグループ交互作用は無かった場合，両サブグループともに有効では偽陰性率が若干高くなるが，両サブグループともに無効では真陰性率に大差は認められなかった．また，平均症例数および平均評価回数においても大差は認められなかった．

Ⅴ．まとめ

本研究では，Wathen (2008) が報告した解析手法について，新薬が有効である事後確率のカットオフ値の最適値を，偽陽性および偽陰性による全損失より検討した．その結果，本研究のシミュレーション条件下では，薬剤・サブグループ交互作用が存在する場合，最適値は，0.150-0.175 と考えられた．したがって，当初考えていたカットオフ値を現行よりも大きく設定すべきではないかという点については，偽陽性の観点からは現行の値よりも多少大きくする方が良いと考えられた．

文献

1)　Wathen JK, Thall PF, Cook JD, Estey EH. Accounting for patient heterogeneity in phase II clinical trials. Statistics in Medicine 2008; 27:2802-2815

〈教育報告〉

平成 21 年度専門課程Ⅱ
生物統計分野

非臨床薬理試験の用量－反応試験データ解析に関する研究

本田主税

A Study of Dose-Response Data Analyses of Non-Clinical Drug Examinations

Chikara HONDA

抄録

医薬品開発の非臨床薬理試験において汎用される用量－反応性データ解析モデルは，個体差を適切に取り扱えない問題点がある．近年，個体差を適切に考慮できる非線形混合効果モデル解析が用いられているが，設定例数が少ない非臨床薬理試験に応用される事例は殆どなかった．最近，シミュレーションによる非臨床薬理試験データ解析に適した非線形混合効果モデルの解析モデル探索の報告がなされたが，真値の設定が実データに基づいていないために，薬力学的パラメーターが相互に独立という仮定を置いている点に問題があった. 従って, 本研究では薬理実験で得られた実データをもとにシミュレーションの真値の設定を行い，薬力学的パラメーター相互の相関を考慮して安定した推定が可能な条件をシミュレーションにより検討した．解析モデルに変量効果相互の相関を考慮した場合，収束率は低かったが，無相関を仮定した場合，多くの解析モデルで収束率は高かった．母集団平均の推定の良さは変量効果を仮定した薬力学的パラメーターの数や変動係数の大きさに必ずしも依存せず，分散の大きい薬力学的パラメーターを中心に複数の変量効果を仮定することで良い推定が行えることが示された.

キーワード： 非臨床薬理試験，用量－反応関係，非線形混合効果モデル，母集団パラメーター推定，推定精度

Ⅰ．目的

最近，シミュレーションによる非臨床薬理試験の用量－反応データ解析に適した非線形混合効果モデル（以降，NONMEM）の解析モデル探索の報告（以降，先行研究）がなされた[1)]．先行研究では，シミュレーションの真値及び解析モデルに仮定する変量効果が相互に独立という仮定を置いているが，実際は独立ではないと考えられた．従って，薬力学的パラメーター相互の相関を加味した NONMEM の検討を行い，薬力学的パラメーターの推定の良さを向上させつつ，安定した推定が可能な解析モデルをシミュレーションにより検討した.

Ⅱ．研究デザインと方法

想定する薬理実験は，健常人から採取した血液サンプルを用いたサイトカインの抑制作用を検証する *in vitro* 実験系とした．筆者所属の研究所から入手した薬理実験の実データからシミュレーションの真値を三つ設定し，乱数により各々のシミュレーションデータを作成した．モデルは式（1）に示す薬理実験データ解析に汎用される four-parameter logistic model を仮定した[1,2)]. 薬力学的パラメーターの分布として，四次元正規分布を想定した（式（2））. 解析モデルとして，いずれかの薬力学的パラメーター間に相関があることを仮定したモデル（以降，相関ありの解析モデル）11 通り，薬力学的パラメーター間は全て無相関であることを仮定したモデル（以降，相関なしの解析モデル）15 通り，即ち全通りの解析モデルでの検討を行った.

$$Y_{ij}= \beta 3_j+\frac{\beta 4_j-\beta 3_i}{1+10^{\beta 2_i(\beta 1_j-d_i)}}+U_{ij}$$

$$Y0_j= \beta 4_j+U0_j \qquad \cdots (1)$$

$Y0_j$　個体 j の阻害剤未添加 d0 でのサイトカイン産生量
$U_{ij} \sim N(0, \sigma^2)$ U_{ij} は残差で，平均 0, 分散 σ^2 の正規分布に従うことを仮定する.
$\beta 1$：$\log D_{50}$（以降，D_{50}）
$\beta 2$：曲線の傾き（slope）
$\beta 3$：反応の最小（min）
$\beta 4$：反応の最大（max）を表すパラメーター

指導教官：西川正子，高橋邦彦，丹後俊郎（技術評価部）

$$(\beta 1\ \ \beta 2\ \ \beta 3\ \ \beta 4)^T \sim N(\beta, \Sigma) \qquad \cdots (2)$$

β は平均ベクトル，Σ は分散共分散行列（略記）

STS 法による母集団パラメーター推定は非線形最小二乗法による個体毎のパラメーター推定を行い，算出した個々の推定値の平均及び分散を算出し，母集団パラメーター推定値とした．NONMEM のパラメーター推定はラプラス近似法により行い，最適化計算の初期値として STS 法により算出した各パラメーター推定値を用いた．設定例数および用量ポイント数はそれぞれ 10 例，11 ポイントとし，シミュレーション回数は 1000 回とした．

Ⅲ．結果

相関ありの解析モデルで三つ以上の薬力学的パラメーターに変量効果を仮定した場合の収束率は実用困難な程低く，パラメーター過剰と考えられた．二つの薬力学的パラメーターに変量効果を仮定した場合，設定した真値の分散が大きいほど収束率が高く，相関が大きいほど収束率が低くなる傾向が認められたが，95％を越える収束率が得られた解析モデルはなかった．一方，相関なしの解析モデルの場合，多くの解析モデルで 95％を超える収束が得られた．図 1 に真値 1 を用いたシミュレーションの $\hat{\beta}1$（D_{50}）の中央値の偏り及びその信頼幅を示した．解析モデル M123, M13, M12, M1 を用いた推定値は偏りが大きかったが，それ以外の解析モデルを用いた推定値の偏りの小ささは，必ずしも変量効果を仮定した薬力学的パラメーター数や種類に依存しなかった. このことは真値2及び3を用いたシミュレーションでも同様であった.

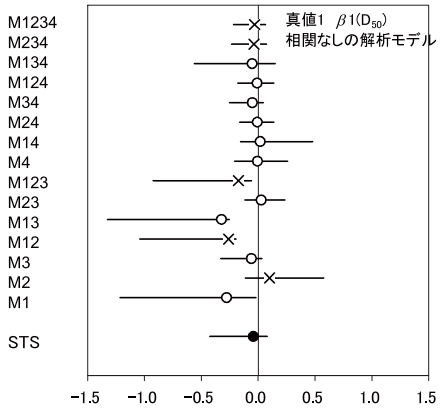

図 1　推定値のパーセンタイルと真値との差（相関なしの解析モデル）

M1234 は全ての薬力学的パラメーターに変量効果を仮定したモデル名，数字は薬力学的パラメーター名と対応している．記号×○●は $\hat{\beta}1$（(D50）の推定値の中央値の真値からの偏り，ひげの長さは 2.5 及び 97.5％点の真値からの偏りを示す．×は収束率 95％未満の解析モデル，○は収束率 95％以上の解析モデル，●は STS 法.

次に，収束率が 95％以上であった解析モデルに限定し，母集団平均の推定の良さについて Bias（偏り）と RMSE（平均二乗誤差）により評価を行なった．その結果，一貫して Bias や RMSE が小さい解析モデルはなく，真値と解析モデルの組合せにより異なった．さらに，各々のシミュレーションデータセット毎に AIC を算出し，AIC が最も小さい（良い）解析モデルの個数をカウントした．その結果，最もカウントの多い解析モデルは設定した真値により異なったが，いずれも $\beta 4$（max）を含む二つ以上に変量効果を仮定した解析モデルであった.

Ⅳ．考察

薬力学的パラメーターの真値に相関を仮定した場合，NONMEM の最良の解析モデルは薬力学的パラメーターの相関，分散の大きさにより変わることが示された．分散が大きい薬力学的パラメーターへの変量効果の仮定は必要であるものの，先行研究が推奨する $\beta 4$（max）に限定して変量効果を仮定する解析モデルが最良というわけではなかった.

Ⅴ．まとめ

薬理実験の実データを基に薬力学的パラメーターの真値を設定し，実際をより反映した条件下にて，薬力学的パラメーター相互の相関を考慮したシミュレーションを行った．その結果，NONMEM による母集団平均の推定の良さは変量効果を仮定した薬力学的パラメーターの数や変動係数の大きさに必ずしも依存しなかった．分散の大きいパラメーターを中心に複数の変量効果を仮定し，AIC を用いて解析モデルを選択することで，良い推定が行えることが示唆された.

文献

1）Yamada M, Hamada C, Yoshimura I. Influence of random effect incorporated in the analysis of pharmacological data based on four-parameter logistic model. Jpn J Biometrics 2009; 30:17-34.
2）Finney DJ. Radio immunoassay. Biometrics 1977; 32:721-40.