

無作為化臨床試験における予後因子の不均衡と

ログランク検定のサイズとの関係

新潟大学医学部附属病院医療情報部	赤澤 宏平
長崎大学環境科学部	中村 剛
九州大学大学院医学研究院医療情報部	絹川 直子
九州大学大学院医学研究院医療情報部	野瀬 善明

はじめに

治療法の優劣を評価して優れたものを医療の現場に還元するためには、治療効果を正確に判定できる臨床試験の実施が必要不可欠である。そこで、臨床試験のデザインや統計的方法をさらに向上させる取り組みがなされてきた。たとえば、ICH Steering Committee は臨床試験実施のためのガイドラインを提唱し、日米欧で薬効評価方法を統一しようとしている¹⁾。また、アジアにおいても、Asia Clinical Conference 1998 が香港で開催され臨床試験の質の向上にむけての統計学的方策が議論された。このように、ガイドラインを作成し質の高い臨床試験を標準的に行えるようにすることは大切であるが、その一方で、これまでに行われてきた臨床試験の問題点と原因を洗い出しその解決方法を探索することも重要である。

生存時間をエンドポイントとする層別ログランク検定の検出力は、層内の症例が不均一な場合に低下することが知られている²⁾⁻⁴⁾。今回の報告では、層内での既知予後因子の不均衡性、すなわち、治療群と対照群とで既知予後因子の分布が等しくないこと(Baseline imbalance)が層別ログランク検定のサイズに与える影響を報告する。

臨床試験における治療効果判定は、一般にサイズが 0.05 となるように棄却域が設定される。ところが、予後因子分布に不均衡が生じると、サイズが 0.05 より大きくなり、治療効果がないにもかかわらず治療効果ありと誤って判定してしまう確率が 5%より大きくなる。臨床試験において妥当な治療効果判定をいかに行えば良いのかを考える際に、サイズの計量的評価は重要である。本研究では、予後指数（比例ハザードモデルの対数ハザード）が $N(\mu, \sigma^2)$ に従う症例を、対照群と治療群に無作為割付した時の群間不均等を検討した。層別ログランク検定のサイズをモンテカルロ法によるシミュレーションにより推定した結果、群間で不均衡が生じると名目上 5%に設定した有意水準が実際にはかなり大きなサイズとなること、また、症例数を増やせば、2 群のハザード比は 1 に近づくが、サイズの増加を防ぐことはできないことがわかった。

・不均等度・不均一度の定義

群間での既知予後因子の不均等ならびに症例の不均一の程度を以下で定義する。生存時間モデルとして比例ハザードモデル $\lambda(t | PI) = \lambda_0(t) \exp(PI)$ を仮定する。ただし、 PI は予後因子情報を集約した予後指数(Prognostic Index)で $N(\mu, \sigma^2)$ に従うとする。 $\lambda(t | PI)$ は PI が与えられた下での時間 t におけるハザード、 $\lambda_0(t)$ はベースラインハザード関数とする。 $N(\mu, \sigma^2)$ から無作為抽出された症例が治療群、対照群の 2 群に M 例ずつ同数割り付けられたとする。このときの治療群、対照群のハザード和を H_T, H_C で表す。一般に、 $H_T \neq H_C$ であるが、その違いの程度を表す量として不均等度指数 $V = (H_T - H_C) / (H_T + H_C)$ を定義する。ハザード比は V を使って $H_C / H_T = (1 - V) / (1 + V)$ と表される。

・不均等度指数 V の分布

予後指数 $PI(i=1, \dots, M)$ が正規分布 $N(\mu, \sigma^2)$ に従うと仮定すると、 V は漸近的に $N(0, Q^2)$ 、ただし $Q = \sqrt{2M / (\exp(\sigma^2) - 1)}$ 、に従う。

実際の胃癌臨床試験において、 V がどの程度の値をとりうるかを Q^{-1} に対する倍率を尺度にとり調べてみた。組織学的治癒切除を受けた進行胃癌患者 1300 例を対象とした。 V を計算するための予後因子は、TNM 分類、ボルマン分類、リンパ管侵襲、占居部位、郭清度である。組織学的漿膜浸潤(ps)とリンパ節転移(n)の有無により層別化割付けを行った。その後で、 V ならびに Q^{-1} を各層ごとに計算し V の Q^{-1} に対する倍率を求めた。その結果、 $n(-)ps(-)$ 、 $n(-)ps(+)$ 、 $n(+)ps(-)$ 、 $n(+)ps(+)$ の 4 層における V の Q^{-1} に対する倍率は、それぞれ、1.38、0.67、1.71、0.467 であった。この不均等度が層別ログランク検定のサイズにどの程度の影響を与えるのかを以下のシミュレーションで検証する。

・層別ログランク検定のサイズの推定

層別ログランク検定のサイズをモンテカルロ法のシミュレーションにより推定する。最初に、各群 200 ずつの予後指数 $PI(i=1, \dots, 400)$ を $N(2, 0.8^2)$ から繰り返し無作為抽出した。その中で不均等度指数 V がおよそ $2Q^{-1}$ となるサンプルを 1 セット選び出した。次に、その 400 例の予後指数サンプルをもとに、指数分布による死亡時間 $-\ln(U_i) / \exp(PI_i)$ (ただし、 U_i は一様乱数、 PI_i は i 番目の症例の予後指数) を 4000 セット生成した。最後に、予後指数サンプルと死亡時間との対データ 4000 セットの各々に対して、有意水準 5% の棄却域をもつログランク検定を行いサイズを推定した。シミュレーションの結果、症例数を増やすとハザード比は 1 に近づく、すなわち、2 群の予後指数分布は症例数の増加とともに均等になるように近づくが、サイズは症例数の大小に関係なくおよそ 15% 以上であることがわかった。さらに別のシミュレーション結果より、不均等度 V が Q^{-1} のおよそ 1.5 以上あるときサイズは 10% 以上になる。従って、前節で述べた胃癌臨床試験の場合、 V の Q^{-1} に対する倍率が 1.38 や 1.71 の層が存在するのでサイズの増加が起こることを示唆している。

．まとめ

症例の無作為化割付は、治療群間での既知の予後因子ならびに未知要因の均等化を図る方策である。たしかに、症例数を増やすと治療群間のハザード比は 1 に近づくので、直感的にはそれらの不均等の影響は除去できるかのように見える。しかしながら、シミュレーションの結果から、大標本であるがゆえにそのわずかな不均等がサイズを増加させることがわかった。層内の不均等は層別ログランク検定のサイズを症例数に関係なく増加させる。

謝辞

臨床試験データを使用させていただきましたがん集学的治療研究財団の井口潔理事長、野本亀久雄副理事長、ならびに、財団関係者の方々に感謝致します。

参考文献

- 1) ICH Steering Committee. ' ICH Harmonised Tripartite Guideline, Statistical Principles for Clinical Trials ' , 1998.
- 2) Schumacher, M., Olszewski, M. And Schmoor, C. ' The impact of heterogeneity on the comparison of survival times ' , *Statistics in Medicine*, 6, 773-784(1987).
- 3) Lagakos, S.W. and Schoenfeld, D.A. ' Properties of proportional-hazards score tests under misspecified regression models ' , *Biometrics*, 40, 1037-1048(1984).
- 4) Akazawa, K., Nakamura, T. And Palesch, Y. ' Power of logrank test and Cox regression model in clinical trials with heterogeneous samples ' , *Statistics in Medicine*, 16, 583-597(1997).