連載 5

統計学から医療を斬る

大橋靖雄
中央大学理工学部
人間総合理工学科 教授／
日本臨床試験学会 代表理事

ハザード比を超えて：生存時間解析の常識を疑う

米国臨床腫瘍学会の７月オンライン版に、現在ダナファーバーがん研究所におられる日本人統計家・宇野一氏（東京理科大学出身）の方法論に関する論文が発表された*。興味深いのは、アメリカの生物統計家のみならず、行政あるいはレギュラトリーサイエンスの立場からわが国の医薬品審査に関わっておられる独立行政法人医薬品医療機器総合機構（ＰＭＤＡ）の宇山佳明氏と元厚生労働省医薬食品局審査管理課の宮田俊男氏が共著者に名を連ねていることである。お二人とも若手のホープである。この論文の主張は、疾患発症・再発・死亡などtime-to-eventのエンドポイントを採用している臨床試験においては定番の「ハザード比による治療効果の相対評価」について、

①解析の前提である比例ハザード性が満足されない場合には、ハザード比による要約はもちろん適切ではない。とくに追跡期間によって結果が変わる可能性も高い。

②追跡期間と症例数が十分でもイベント発生が少ない場合（とくに死亡や重篤有害事象）には、信頼区間の精度不足から、情報不足という不適切な解釈を与える。

③ハザード比に替わる、臨床的に意味がありモデルに依存しない指標の採用をデザイン時点から考えるべきではないか。

という、もっともではあるが、かなり踏みこんだものになっている。本連載第２回で、比例ハザード性が成り立たない例を偶然挙げたばかりであるが、今回は、今後の承認審査や研究動向にも影響を与えかねない本論文の骨子を紹介する。

比例ハザード性の成立はログランク検定の最適性とCox回帰の妥当性の前提

まず理論を簡単におさらいしておく。あるイベントが時刻T=tの直前まで発生せず、次の瞬間に発生する速度がハザードであった。これを時刻tと共変量（治療群や背景因子）xの関数として$\lambda(t,x)$で表したとき、λを時刻による変化パターンと共変量の寄与の部分に、

$$\lambda(t,x)=h(t)\,\psi(x)$$

と分離できるのが比例ハザード性の仮定であり、この下でh(t)を同定せずに共変量の寄与部分を推定するのがCox回帰である。

比例ハザード性の下では、治療群と対照群それぞれの生存関数$S_1(t)$、$S_0(t)$の間には治療群の対照群に対するハザード比をλとして、

$$S_1(t)=(S_0(t))^{\lambda}$$

の関係が成り立つ。一方の曲線が他方の必ず上あるいは下に位置することになる。実は、生存関数の群間差検定であるログランク検定とCox回帰には密接な関係がある（そのためログランク検定をCox-Mantel-Haenszel検定と呼ぶこともある）。

２群間のログランク検定は、イベント発生があった時点iごとに（群×イベントあり・なし）の２×２表を構成し、O_i（治療群での観測イベント数）、E_i（ハザードが等しいという仮定のもとで治療群に期待されるイベント数）から$v=\Sigma_i(O_i-E_i)$とその分散V_i（式省略）の和$V=\Sigma_i V_i$を計算し、χ^2検定統計量

$\chi^2 = v2/V$

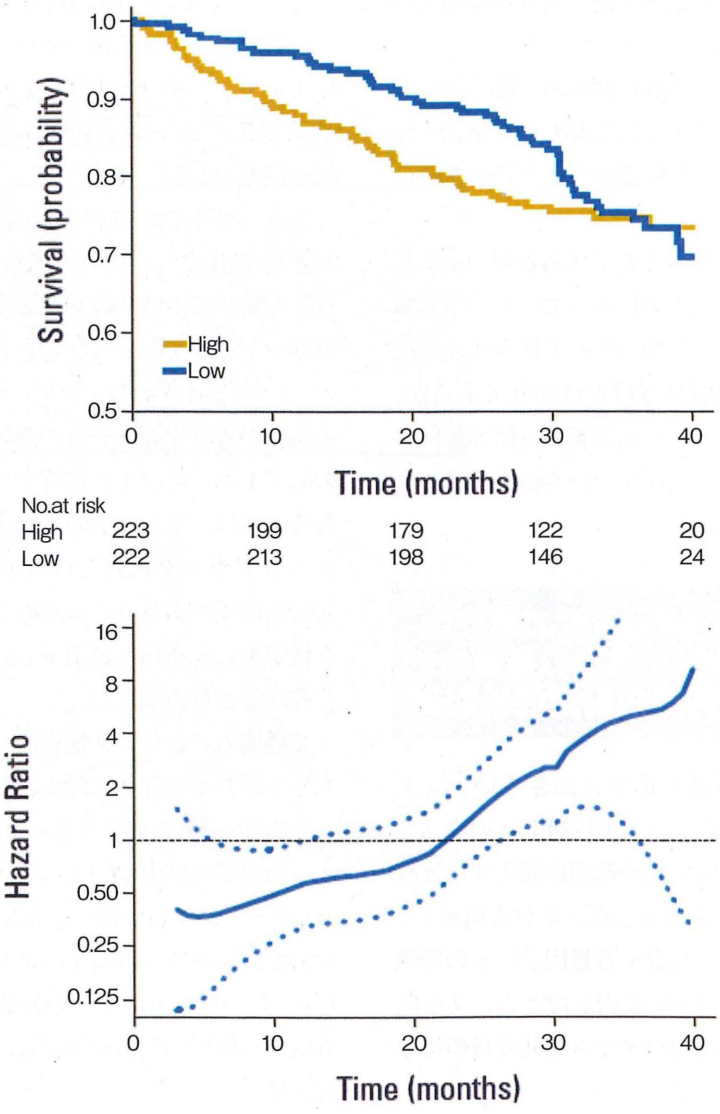
を計算して実行される。当該イベント発生時点直前の両群の追跡者数が治療群：対照群＝n：mなら両群合わせたイベント数をD_iとして$E_i = D_i \times n/(n+m)$である。治療群のハザードが全時点にわたって対照群より小さければ、$O_i - E_i$は負で累積され強い治療効果が示されることになる。

ここでv/VがCox回帰で求められる対数ハザード比の近似となることが証明されている（標準誤差は$V^{1/2}$）。比例ハザード性が成り立つことが、ログランク検定の最適性とCox回帰の妥当性の前提であり、ログランク検定の有意性とハザード比の有意性はほぼ同等となる。

ハザード比要約が不適切な例

原論文では三つの例が紹介されている。最初の例はEastern Cooperative Oncology Group（ECOG）で実施された多発性骨髄腫に対するデキサメタゾンの高用量と低用量の試験であり、カプランマイヤー曲線と対数ハザード比は**図1**のようになる。ハザード比と時刻の間には「質的交互作用」、つまりハザードの大小関係が追跡時間により逆転するほどの質の違いが見られるものの、$O_i - E_i$全体を「足し合わせて計算される」ログランク検定のp値は0.47、ハザード比推定値は0.87（95%CI：0.60～1.27）と有意でない。追跡初期は低用量群が明らかに優れており、1に近いハザード比から全区間にわたって群間差がないと解釈するのは誤りである。

図1　（原論文Fig.1 A,B）ECOGの多発性骨髄腫臨床試験



3番目の例は、J Clin Oncol（JCO）に報告されたステージⅡ/Ⅲ結腸がんの術後補助療法試験であり、標準治療FOLFOXに対して分子標的薬ベバシズマブの上乗せの有無が比較されている。**図2**に見られるように、両群のカプランマイヤー曲線は重なり、ハザード比推定値は0.95（95%CI：0.79～1.13）、ログランク検定p値は0.56である。イベント数が少ないため情報不足という判断がなされるであろうが、後述するようにこれは適切な解釈ではない。

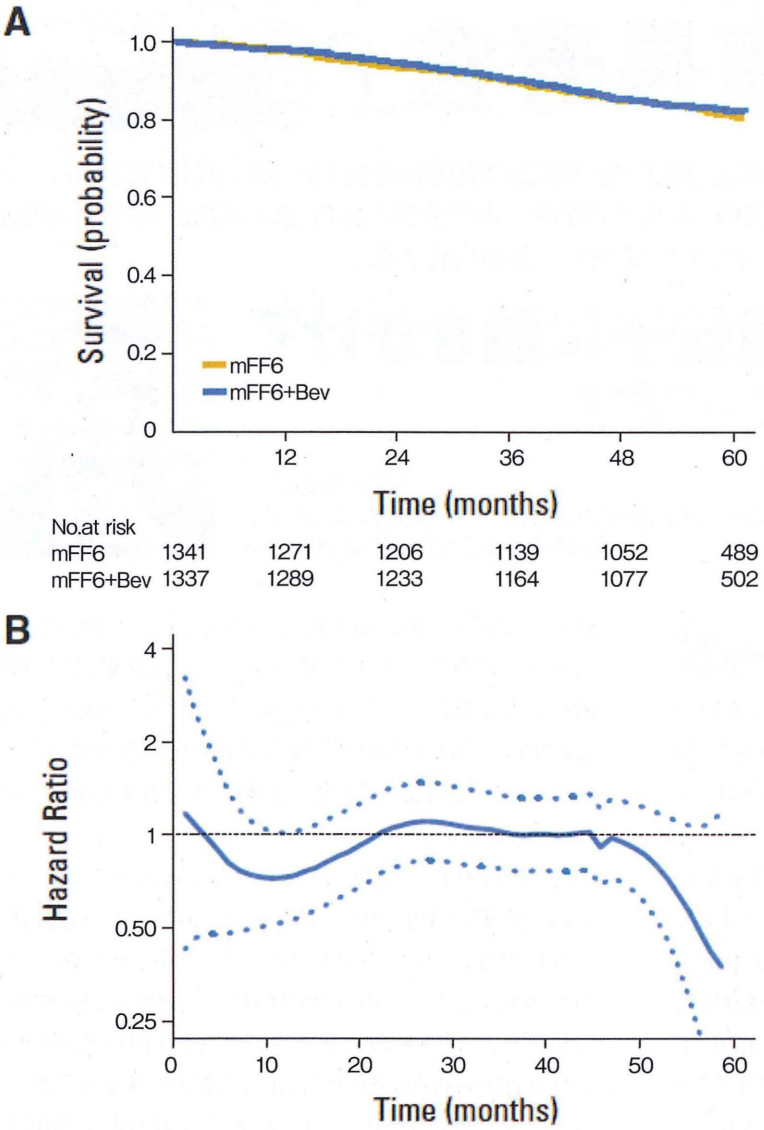
患者の利便性やコストを考慮しハザード比に替わる指標の使用も

ハザード比に替わる、モデルによらず、かつ臨床的に解釈が容易な指標として、著者らは、

①ある時点での生存関数の差あるいは比

②メディアンなど生存関数のパーセ

図2（原論文Fig.3 A,B）結腸がん術後に対するベバシズマブの上乗せ試験



ント点の差あるいは比

③ある時点までを追跡上限とした制限つきの生存時間平均値

を挙げている。①はふつうに使われている指標の見直しであり、②ではメディアンに替わるパーセント点の提唱が目新しい。③はこれまで取り上げられていない指標である。

多発性骨髄腫の例（図1）では40カ月では①の差はほぼ0（比1）であるが24カ月では比は1.13（95% CI:1.03～1.23）と低用量群で有意に高い。②の10%、20%点を用いると低用量群はそれぞれ20.3カ月と30.8カ月、高用量群ではそれぞれ9.5カ月、22.1カ月で10%点には比で有意差がある。

③はカプランマイヤー曲線の上限時点までの曲線下面積として求められる。多発性骨髄腫の例で40カ月を上限とすると、低用量群での制限つき平均値は35.4カ月、高用量群では33.3カ月であり、40カ月から引いた「失われた生存時間」はそれぞれ4.4カ月、6.7カ月でその比0.68（95% CI:0.47～0.98）は有意である。

追跡の実質的な終了時点にもなりうる40カ月という時点設定は解釈にとって本質的であり、試験デザイン時点で臨床的意義と実施可能性の双方から選択されねばならないが、この結果はハザード比による解釈とは異なり、低用量群の方が患者にとって便益は高いという解釈につながるであろう。

大腸がんの例（図2）で（しばしば追跡期間に設定される）60カ月を上限として制限つき平均値を計算すると、ベバシズマブ上乗せ群で55.2カ月、なし群で54.9カ月、差は0.3カ月（95% CI：－0.7～1.2）となる。平均値の絶対値からすれば差はわずかである。ハザード比による情報不足の判断と異なり、この指標に関する限りは両群が臨床的にほぼ同等であることが強く示されるという結果である。

筆者らは、「臨床的に、生物学的に、経験的に」強い正当化がなされる場合にはハザード比のようなモデルに依存する指標は有益であることを認めているものの、そうでなければ（つまり例外を除けば）モデルに依存しない指標の使用を強く勧めている。

2016年から試行予定の保険収載における医薬品の経済性評価においても、質を調整した平均生存年（QALY）が用いられる可能性が高い。何が患者のベネフィットかを、コストを考慮しつつ考える時期に来ているのであろう。そのための評価指標再検討の必要性をこの論文は訴えている。 MA

* Uno Het al: Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis: JCO 2014 June 30 オンライン版

訂正 「統計学から医療を斬る」7月号75ページ中段19行目に「ただしHR－」とあるのは「ただしHR＋」の誤りでした。訂正します。