第 11 回日本癌治療学会　がん臨床試験協力・参加メディカルスタッフのためのセミナー「ガイドラインの背景と最新のトピック」

MSS-5

臨床試験のための生物統計学

山　中　竹　春
（国立がん研究センター　生物統計部門）

本稿では，がんの臨床試験の研究デザインおよび統計学的側面について知っておくべきことを述べる．

第 I 相試験

第 I 相試験の目的は，試験治療の最適用量（至適用量）を決めることにある．細胞障害性抗がん剤の場合には，毒性と有効性が比例することを仮定し，推奨用量の決定を行ってきた．この目的のために最もよく用いられるデザインが 3 例コホートである．これは，同一用量を 3 例ずつ（そのため、“3 例コホート”と呼ばれる）に投与して，0/3 の毒性発現であれば、用量を増加、最大耐用量を決定する方法である．3 例コホートに統計的な根拠は実はあまりなく、実際、統計“解析”が必要となるわけではないが、この方法でこれまで多くの開発が行われ，また、デザイン上やりやすくわかりやすいからということで，現在でもよく用いられる経験則的な方法である．3 例コホートは細胞障害性抗がん剤のために開発されてきたデザインであるが，分子標的薬の場合にも，（明らかに優れた方法がほかにないため）標準的に用いられているというのが現状である．

3 例コホートはかなり低い用量から開始し，順に増量するため，なかなか最適用量に達しない，最適用量付近でのサンプルサイズが小さい，といった問題が指摘されている．そこで，この点を解消するために提案されている方法が連続再評価法 CRM（continual reassessment method）である．この方法は，1 例ごとに統計的に用量反応曲線を推定し，それに基づいて最適と思われる用量を投与，その結果に基づいて用量反応曲線を推定し次の対象者の用量を定める，を繰り返すベイズ流の方法である．ただし、実際に実施するにあたっては，3 例コホートに比べ，実務上の手順がかなり複雑になる．

第 II 相試験

第 II 相試験の目的は，試験治療の開発を今後も続け，将来的に標準治療を対照とした第 III 相試験を行うに値するかどうかを判断することにある．第 III 相へ進むかどうかの決定は，安全性やほかの薬剤の開発状況などにも依存するので，統計的判断は有効性の足切りの目安として使われることが一般的である．がん領域では、シングルアーム試験やランダム化選択デザインなど通常 100 例以下のデザインが用いられることが多いが，検証的な結果ではなく，いずれにしろ，その先に第 III 相が必要であることを理解すべきである．おもなデザインは，閾値・期待値を用いたシングルアームのデザインである．これは primary endpoint（奏効割合 response rate など）に対して，それ以下なら開発を中止すると考える値を閾値として設定し，真の response rate がある値（期待値）であった場合には，実際に得られるデータによって閾値以下であることが高い確率で統計的に棄却できるようにサンプルサイズと decision　rule を設定するデザインである．閾値以下であるということが棄却されれば統計的には有効性が期待できるということになる．このデザインのオプションとして，試験途中で中間解析を行い、その時点で無効が証明された場合には登録を終了して無効中止を行う 2 段階デザイン（2-stage design）が用いられることも多い．

第 II 相試験の結果を解釈する際に重要なのは、試験に実際にエントリーされた対象者の成績が，対照としている治療の成績と比較可能かどうかをチェックすることである．たとえば，ヒストリカルコントロールとの比較の場合，実際に試験にエントリーされた対象者とヒストリカルコントロールの対象者がどのくらい比較可能かをチェックすることが重要である．適格規準に記載された対象者が必ずしも満遍なく試験にエントリーされるわけではないため，実際に試験に登録された対象者がどのような属性の分布をもっているか，それがヒストリカルなデータとの間で大きく異なっていないかを検討する必要がある．特に time-to-event（全生存期間や無増悪生存期間）をエンドポイントとする場合には，ヒストリカルコントロールのデータがよほど安定していないと比較でき

ない．Time-to-event のデータは，試験に組み入れられた対象者がどういう集団かによって大きく異なる．そのため、比較可能性をより担保するために，ヒストリカルコントロールではなく，第 II 相試験であっても，対照群を設定したランダム化第 II 相試験を計画する場合がある．ランダム化第 II 相試験で注意すべきなのは，治療アームの中に対照群（標準治療群）が含まれていたとしても，サンプルサイズが小さいため、検証的な第 III 相試験の代わりにならず、あくまで第 III 相試験に向けたスクリーニングのためのデザインと考えなければならない、ということである．ときには、ランダム化第 II 相試験において、対照群をおかずに新治療同士をランダム化比較するランダム化選択デザイン（randomized selection design）が用いられることもある．これは，患者を 2 つ以上の新治療にランダムに割り付けし，最も高い奏効割合が得られた治療を選択して第 III 相試験の候補とする、というデザインである． 2 つ以上の試験治療がある時に、開発を継続する優先順位をつけるデザインといえる．このデザインでは，いずれかの治療の成績がより高いと仮定して、そちらの治療法を高い確率で選択できるようにサンプルサイズと decision rule を決定する．このデザインで注意すべきなのは，比較する 2 つの治療法が同じ有効性を有する（たとえば、奏効率が等しい）とき、50-50 の確率で強制的にどちらかを選択してしまうことである．差がない場合でも常にどちらかの治療法が選択されるため、まったく検証的な試験とはいえない。

第 II 相試験の結果によって，第 III 相試験へ進むかどうか，どのレジメンを選択するかの決定にあたっては，毒性の情報や他の開発薬剤の状況などを考慮する必要があり，統計的判断は有効性の足切りの目安として使われることが一般的である．サンプルサイズの目安として，シングルアームだと数十名であることが多いが，患者数が少ない疾患などで，次に第 III 相を実施することが現実的に不可能で，やむを得ず第 II 相試験に検証的な意味をもたせたい場合は 100 名以上とすることもある．ランダム化比較している場合は，100 名程度の場合が多く，ランダム化選択デザインだと 100 名以下となることが一般的である。

第 III 相試験

第 III 相試験においては、研究仮説は何か，primary endpoint は何か，デザインはどうか，解析が予定通りに行われているか，などが評価のポイントである．研究仮説として，比較する治療は何と何か，どちらの治療が標準としているかを把握する．適切でない治療を標準治療としてしまっている試験は結果の解釈が困難である．次に，primary endpoint として患者ベネフィットの直接の指標である全生存期間（overall survival：OS）を用いているか、あるいは progression-free survival（PFS）、disease-free survival（DFS）、time-to-treatment failure（TTF）を用いているのかに注意する。PFS、DFS、TTF を用いている場合は、それらがその疾患領域で確立されているエンドポイントであるかを知ることが重要である．そうでなければ，必ずしも検証的な研究結果と捉えられない可能性がある。デザインは優越性試験か非劣性試験か．新しい治療が toxic new であれば優越性，less toxic new であれば非劣性デザインを使う必要がある．非劣性試験の場合，臨床的に意味のある secondary endpoint が定義され，それが標準治療に勝っていることが必要になる．解析が予定通りに行われているかどうかも大きなポイントである．予定通りのサンプルサイズやイベント数を達成しているか、などがポイントである．最近の論文では，CONSORT 声明に従った研究の流れ図が記載されている論文が一般的であるが、これは途中で脱落した人がどのくらいいるか，治療のコンタミネーションがどのくらい起こっているか，Intention-to-Treat（ITT）解析をしているか，などを容易に把握することができ，研究の質を評価するのに有用である．

結果の解釈においては、まず背景因子の比較が示されることが一般的である。ここでは，重要な予後因子に関して，群間に大きなアンバランスがないかどうかを確認する．もし大きなアンバランスがあれば，結果の解釈の際に注意すべきであるし，統計モデルでアンバランスな要因を調整した解析結果なども参考にすることになる．背景因子ごとに p 値が示されていることも多いが，検定の多重性の問題が生じるので、p 値自体にあまり神経質になる必要はない．アンバランスがある場合でも，最も重要な結果はランダム化に基づいた解析であり，統計モデルで調整した解析はあくまで確認のためである。

最も重要なのは primary endpoint の結果である．前述の通り，primary endpoint としては全生存期間 OS が最も望ましく，PFS が primary endpoint である場合でも，OS は大きな意味をもつ．用いられる統計手法は，Kaplan-Meier curve による生存曲線の比較が一般的であり，打ち切りがどのくらいあるかがデータの成熟度の目安となる．生存曲線の比較には，Log-rank 検定を用いることが多い．Log-rank 検定の結果が有意であるかどうかとともに，2 群間の効果の差（ハザード比や生存期間中央値 MST の差）についても着目する必要がある．ハザード比の算出には Cox 回帰が用いられる．

優越性試験の場合，試験がしっかり計画，実施，解析された場合には，primary endpoint で有意な結果が得られた場合，新治療が勝っていると判断する．その際，効果の大きさについても考慮することが重要である．

primary　endpoint で有意な結果が得られなかった場合には，対象とした集団全体では新治療が勝っているとはいえない．この場合にも，探索的にサブグループ解析をして，次につながる仮説を立てることが重要である．非劣性試験の場合には，結果の解釈に特に注意が必要である．非劣性試験とは，試験治療の primary　endpoint（たとえば OS）が，標準治療と比べて「許容できる差」以内にあるかどうかを検証するデザインである．許容できる差以上に劣っている，という仮説を棄却することによって，許容できる差以内であることを検証する．有害事象が少ない等，ほかにメリットがあるため，OS は同等であればよいと考えるというデザインである。Primary endpoint で有意な差，かつ secondary　endpoint で新治療の有効性が示されている場合，新治療が優れている，あるいは標準治療のオプションになりうると判断することになる .非劣性試験においては、非劣性の許容域が広すぎないこと（5 年生存率で 10％など）に注意する必要がある。また，primary　endpoint で有意な結果が得られなかった場合には，新治療は新しい標準治療，もしくはそのオプションとはなれないと考えるべきである．一般に、非劣性試験では、試験がしっかり計画，実施，解析されていないと考えられるような場合は注意が必要である．なぜなら，非劣性試験は，治療がきっちり行わないと有意になりやすい（非劣性が証明されやすい）．これは，治療の不遵守があると、2 群間の差が薄まる、すなわち、（本当は試験治療の方が劣っていて差があるのに）同じ成績に近づくためである。

試験の primary な結果に加えてよく議論されるのは，サブグループ解析の結果である．サブグループ解析をして色々な検討をすることは非常に重要なことであるが，その解釈には慎重になるべきである．2007 年に出された New England Journal of Medicine のサブグループ解析の報告に関するガイドライン（Wang R. et al, NEJM 2007; 357; pp2189-2194）が参考になる．これによれば，サブグループの結果を abstract に報告してよいのは，それらが primary endpoint に対する結果であり，かつ事前に規定したサブグループ解析であることが前提とされている．