dustry- Patient-Reported Outcome Measures : Use in Medical Product Development to Support Labeling Claims. 2009. Available from : URL : http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.

8) 金谷武洋．日本語に主語はいらない．東京：講談社；2002.

*　　*　　*

腫瘍内科，12(4)：372-381，2013

特集 治療効果の判定基準と臨床試験のendpoint

RECISTとirResponse Criteria 1）総論：Immune Related Response Criteria（irRC）—背景，定義，問題点，JCOGはどう考える？*

江場淳子**
中村健一**
柴田大朗***
福田治彦***

Key Words：irRC, RECIST, response criteria, immunotherapy, comparability

はじめに

2011年米国で抗CTLA-4（cytotoxic T lymphocyte-associated antigen 4）抗体であるipilimumabが承認され，免疫治療に対する注目が高まっている．CTLA-4は，T細胞上に発現する受容体で，T細胞の活性を抑制する．この抑制性に働くCTLA-4を特異的に抑制してT細胞の活性化を維持するのがipilimumabである．2010年Hodiらは第III相試験でipilimumabがplaceboとの比較で進行期悪性黒色腫患者の全生存期間を延長することを報告した[1]．その後，非小細胞肺がんでもcarboplatinとpaclitaxelにipilimumabを上乗せするランダム化第II相試験がLynchらによって行われ，primary endpointのimmune-related progression-free survival（irPFS），secondary endpointのPFS（WHO規準で評価）ともに，ipilimumabを順次併用した群で延長することが示された[2]．2012年には，CTLA-4と同様に免疫抑制性の受容体であるPD-1（programmed death-1）に対する抗PD-1抗体，さらに，がん細胞上に発現しPD-1に結合してT細胞の活性化を抑制するリガンド（PD-L1）に対する抗PD-L1抗体についても，早期の安全性ならびに有効性の報告が行われた[3][4]．これらの結果を受けて，今後，免疫治療がさらに注目を集め，治療開発が展開されることが予想される．

現在，これらの免疫治療薬の開発とともに注目されているのが，上述のLynchらの試験ですでに導入されている新しい効果判定規準のImmune-Related Response Criteria（irRC）である．本稿では，irRCが提唱された背景および定義を述べるとともに，その問題点を論じ，JCOGデータセンター/運営事務局の提案を示す．

irRC提唱の背景

免疫治療薬は，細胞傷害性の抗がん剤とは作用機序が異なる．そのため，腫瘍縮小効果が現れるのに時間がかかること，長期にわたって腫瘍縮小効果がない場合でも生存期間が延長する可能性があること，新病変の出現や一時的な増大のあとに縮小または消失することが知られている[5]．治療開始後の一時的な腫瘍増大は，免疫治療が治療効果を発揮するまでの腫瘍増大，あるいは，一時的な免疫細胞の浸潤や炎症性の変化を反映していると考えられており，それらは病理組織学的にも証明されている[6]．

irRCの提唱者らは，もともとWHO規準や

* Immune Related Response Criteria（irRC）—background, definition, problems, and solutions in JCOG.

** Junko EBA, M.D. & Kenichi NAKAMURA, M.D.：独立行政法人国立がん研究センター多施設臨床試験支援センターJCOG運営事務局〔〒104-0045 東京都中央区築地5-1-1〕；JCOG Operations Office, Multi-institutional Clinical Trial Support Center, National Cancer Center, Tokyo 104-0045, JAPAN

*** Taro SHIBATA, M.Sc. & Haruhiko FUKUDA, M.D.：独立行政法人国立がん研究センター多施設臨床試験支援センターJCOGデータセンター

表1 irRC，WHO規準，RECISTの比較

	irRC	WHO規準	RECIST
測定方法	2方向測定		1方向測定
測定可能病変	≧5 mm×5 mm	規定なし	Ver1.0 ヘリカルCTで長径≧10 mm リンパ節病変に言及なし Ver1.1 腫瘍病変：長径≧10 mm リンパ節病変：短径≧15 mm
測定病変数	ベースライン 各臓器≦5病変 内臓病変≦10病変 皮膚病変≦5病変 新病変出現時に追加 各臓器≦5病変 内臓病変≦10病変 皮膚病変≦5病変	規定なし	Ver1.0 各臓器≦5病変 計≦10病変 Ver1.1 各臓器≦2病変 計≦5病変
腫瘍量	積和		径和
規準値	(ir)CR：すべての病変が消失 (ir)PR：ベースラインに比べて50%以上減少 (ir)SD：いずれにも該当しない (ir)PD：経過中の最小値に比べて25%以上増加		CR：すべての腫瘍病変が消失 PR：ベースラインに比べて30%以上減少 SD：いずれにも該当しない PD：経過中の最小値に比べて20%以上増加
確定を要する判定	irCR, irPR, irPD	CR, PR	CR, PR

RECISTは細胞傷害性薬剤の治療効果判定を目的に開発されてきたため，免疫治療が有効な患者が見逃され，治療効果が過小評価される可能性があると主張している[7]．たとえば，WHO規準やRECISTでは，治療効果が進行(progressive disease；PD)と判定されると，治療無効の判断が下されて治療が中止されるが，免疫治療では，治療早期にWHO規準やRECISTの評価でPDとなっても治療を中止することが適切でない場合があるという主張である．確かに，2011年に米国の食品医薬品局(Food and Drug Administration；FDA)が発表したがん治療用ワクチンの企業向けガイダンス「Guidance for Industry：Clinical Considerations for Therapeutic Cancer Vaccines」でも，開発を行う上で臨床試験のデザインには従来の細胞傷害性薬剤とは異なる配慮が必要であるとしている[8]．

こうした見解に基づき，免疫治療薬特有の治療効果をとらえる効果判定法として，2009年Wolchokらによって irRCが提唱された．

irRCの定義

米国で2004年と2005年に学術界，産業界，規制当局の専門家が集って免疫治療に関するワークショップが開催され，免疫治療薬の抗腫瘍効果について以下の5つのコンセンサスが得られた[9]．

(1)測定可能な治療効果が出現するまでに細胞傷害性薬剤よりも時間がかかる場合がある．
(2)腫瘍縮小効果は，従来の評価規準ではPDと判定される腫瘍増大が生じたあとに現れる場合がある．
(3)PDを確定する前に免疫治療を中止することが適当でない場合がある．
(4)臨床的に明らかに増悪と判断されない場合は，治療継続を許容することが推奨される．
(5)長期間持続する安定(stable disease；SD)は，抗腫瘍効果を意味する場合がある．

これらの特徴を有する免疫治療薬の抗腫瘍効果を系統的かつ適切に評価することを意図して，新しい効果判定規準としてWHO規準に準じたirRCが作成された．

irRCでは，測定可能病変を5 mm×5 mm以上とし，治療開始前のベースラインで各臓器5病変以内，内臓病変は計10病変以内，皮膚病変は計5病変以内の測定可能病変のみを標的病変と

する(表1)．そして，各標的病変の直交する2方向の最長径の積和(SPD：sum of the products of the two large perpendicular diameters)を計算して総腫瘍量(total tumor burden)とし，測定不能病変は総腫瘍量には含めない．治療開始後の評価時点で新病変が出現した場合，測定可能な新病変に限り，各臓器5病変以内，内臓病変は計10病変以内，皮膚病変は計5病変以内を標的病変の総腫瘍量にさらに加える．つまり，経過中の総腫瘍量は以下の式で求められる．

総腫瘍量＝SPD(標的病変)＋SPD(測定可能な新病変)

このように定義した「総腫瘍量」を用いることで，ベースラインで測定した病変が縮小すると同時に新病変が出現している場合や，縮小している病変と増大している病変が同時に混在する場合に，ただちにPDとはならないというロジックになる．

標的病変の効果判定規準は，2方向測定を行うWHO規準と同様で，すべての病変(測定可能病変，測定不能病変，新病変を含む)が消失した場合を完全奏効(immune-related complete response；irCR)，総腫瘍量がベースラインに比べて50％以上減少した場合を部分奏効(immune-related partial response；irPR)，経過中の総腫瘍量の最小値に比べて25％以上増加した場合を進行(immune-related progressive disease；irPD)，いずれの規準にもあてはまらない場合を安定(immune-related stable disease；irSD)と定義する(表1)．

なお，WHO規準，RECISTではCRおよびPRで確定(confirmation)を行うことを必須としている(RECISTでは腫瘍縮小効果がprimary endpointである非ランダム化試験の場合にのみ確定が必須)[10][11]．すなわち，判定された効果が測定誤差でないことを担保するために，最初に規準を満たしてから4週以降に再評価を行い，その規準を満たすことを確認するのである．一方，PDは1回の評価で決まり，測定可能か測定不能かを問わず新病変の出現が認められた時点でPDと判定される．一方，irRCでは，CR, PRに加えて，明らかな臨床的増悪を認めない限りPDでも治療を継続して確定を行うこととしている(表1)．また，新病変が出現した場合でも，測定可能な場合に限り総腫瘍量に含めて評価を行い，PD規準を満たさなければPDとは判定しない．

まとめると，WHO規準やRECISTと大きく異なる点は，①治療開始後に出現した新病変を総腫瘍量に含めること，②治療開始後早期に新病変が出現しても総腫瘍量がPD規準を満たさなければPDとしないこと，③PDの判定に確定を要することの3点である．

Wolchokらは，ipilimumab(10 mg/kg)が投与された切除不能進行期の悪性黒色腫患者227人を対象に，WHO規準とirRCの両方で抗腫瘍効果を評価した[9]．その結果，治療効果が認められた患者では，4つの腫瘍縮小パターンが観察され(A. 治療開始後早期に縮小する，B. 安定している，C. 治療開始後早期の一時的な増大後に縮小する，D. 新病変出現後に縮小する)(図1)，いずれのパターンの予後も良好であることが示された．さらに，治療開始後早期の治療効果判定においてWHO規準でPDと評価された患者のうち，少なくとも約10％の患者がirRCで評価するとirPRあるいはirSDと判定された．また，この患者群の予後はWHO規準でCR, PR, SDと評価される患者群の生存期間に匹敵し，WHO規準でPDと判定される患者群よりも明らかによいことが示された．つまり，irRCで評価を行うとipilimumabが真に有効である患者をさらに10％同定することができるとされた．

また，最近，irRCの2方向測定と1方向測定を比較した報告があり，両者で効果判定の評価は一致し，さらに1方向測定では再現性がより高いことが示されている[12]．

irRCの問題点

irRCは先述したように，細胞傷害性薬剤と異なる作用機序を持つ免疫治療薬の特有の治療効果を評価するために考案された．もともとipilimumabの臨床試験データに基づいて定義された評価規準ではあるが，他の免疫治療薬でも一貫して観察される知見を基礎としているため広く適用できるとirRCの提唱者らは主張している[9]．

しかし，筆者らはirRCの提唱者らの主張につ

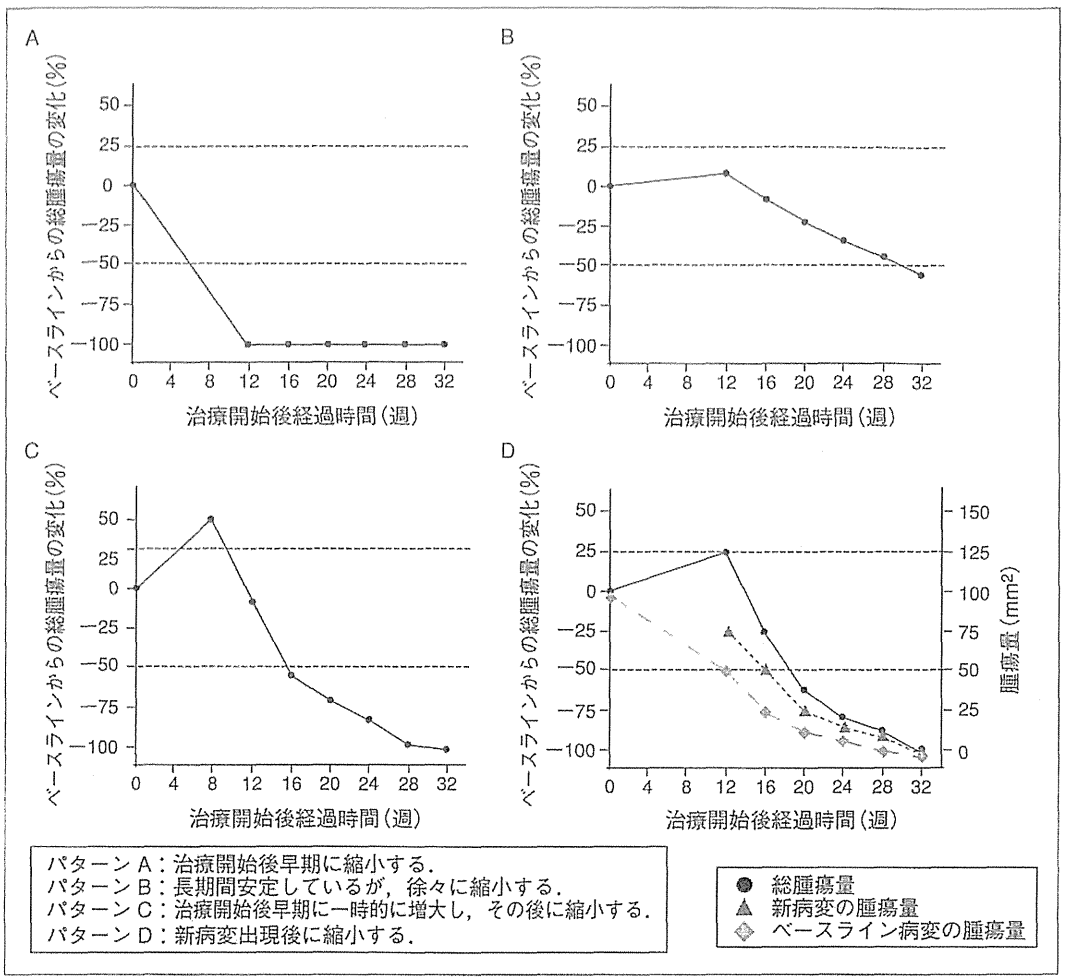

図1　4つの腫瘍縮小パターン

いて，以下の点で疑問がある．

まず1点目は「効果判定規準の意義」そのものに関する疑問である．Wolchokらの原著には，「irRC提唱の背景」でも述べたように，免疫治療の開始後早期にWHO規準やRECISTの評価でPDとなっても，治療を中止することが適切でない場合があると述べられている[9]．しかし，RECIST version 1.1には，「(RECISTは)個々の患者における治療継続の是非についての意思決定に用いられることを意図していない」とある[11]．つまり，「判定がPDとなった時に一律に治療を中止すべき」とはRECISTのどこにも書かれていない．この誤解の背景には，細胞傷害性薬剤において1回のPD判定で治療を中止することが妥当と見なされる状況が多くの場合に当てはまってきたことがあるだろう．そのため，「RECISTに従えば，PD＝治療中止」と誤解され，RECISTが「腫瘍縮小効果の判定規準」であって「治療中止規準」ではないということが正しく理解されていないと思われる．効果判定規準は比較可能性を優先して規定すべきであり，一方，治療中止規準は臨床的な妥当性が優先されるべきであって，いずれにしても両者はきちんと分けて考える必要がある．irRC提唱の必要性として述べられている，腫瘍縮小効果が現れるのに時間がかかること，長期にわたって腫瘍縮小効果がない場合でも生存期間が延長する可能性があること，新病変の出現や一時的な増大のあとに縮小または消失することは，いずれも治療中止規準に関する問題であって，効果判定規準にRECISTを用いない理

由にはならない．要は，irRC提唱者らは「治療中止規準の問題」を「効果判定規準の問題」と取り違えているのである．

2点目は「比較可能性」に関する疑問である．免疫治療が標準治療と見なせるがん種はごく一部であるため，新しい免疫治療薬の臨床試験において比較相手となるのは，ほとんどの場合が従来の細胞傷害性薬剤である．すなわち，複数ある免疫治療の候補の中で最もpromisingなものを選択するという段階はさておき，ある免疫療法について検証的試験を実施するか否かを判断する段階では，従来の治療法との相対関係を考察するための手だてが必要となる．そのような検証的試験の前段階で行われる単群試験でirRCを用いて免疫治療薬を(従来の標準治療よりも有望であるか否か)評価するのであれば，細胞傷害性薬剤をirRCで評価したヒストリカルコントロールと比較する必要があるが，現実にはそのようなデータは存在しない．また，ランダム化比較試験で細胞傷害性薬剤のレジメンと免疫治療薬を含むレジメンを比較する場合にも，先述した治療中止規準と効果判定規準の分離を考慮せずにirRCを細胞傷害性薬剤のみのレジメンにも適用してもよいかという問題が生じる．すなわち，新病変が出現しても，あるいは標的病変や非標的病変がPDとなっても，「PD確定」まで治療を継続しなければirRCの評価はできないが，そのような治療継続は，多くの細胞傷害性薬剤では臨床的に不適切であろう．このように「比較」という観点を入れた際にはirRCを現実に適用できる状況はきわめて限られているといわざるをえない．特に開発が早期から後期へと進むにつれて，比較相手の多くは現時点での標準治療と考えられている細胞傷害性薬剤となるため，irRCの意義はさらに薄れる．このように，irRCは免疫療法を日常診療に導入するためには避けて通れない「免疫治療薬以外の薬剤との比較」という視点に欠けている．

また，比較可能性を考えた場合，irRCがRECISTの1方向測定でなくWHO規準の2方向測定を採用したことも問題といえる．2000年にRECISTが公表されてすでに13年がたっており，免疫治療薬が比較相手とすべき細胞傷害性薬剤からなる標準治療のデータの多くは1方向測定のRECISTによるものである．irRC提唱者らが言う先述の(1)～(5)の免疫治療の特徴に起因する課題の解決策として，irRCがRECISTの1方向測定からWHO規準の2方向測定に戻ることにいかなるメリットや論理的必然性があるのか，筆者らがirRCの論文を読んだ限りにおいて納得のできる説明を見出すことができなかった．

3点目が「標準化」に関する疑問である．もともとRECISTが作成されたのは，1990年代までにさまざまな修飾が加えられていたWHO規準を標準化かつ簡略化して，試験間の比較をより適切に行えるようにすることが目的であった．そして，すでに多くの固形がんの臨床試験で広く用いられているRECIST version 1.1で標準化と簡略化がさらに徹底された[11]．現時点でirRCという規準を新たに設けることは，その標準化に逆行することになる．また，irRCが採用するWHO規準に準じた2方向測定であること，標的病変の数がRECIST version 1.1よりも多いこと，新病変が出現すると評価病変数がさらに多くなることは，簡略化という観点からも時代に逆行するといえる．

腫瘍縮小効果判定規準は，世界中で行われる臨床試験で利用可能であることが理想的であり，そのためには標準化された方法で容易に実行できる必要がある．固形がんの腫瘍縮小効果の評価には，現在の標準であり今後も広く用いられることが予想されるRECISTを免疫治療薬の評価にも採用することが望ましい．

では，果たしてRECISTは免疫治療薬の評価において本当に不適切なのであろうか？

irRCの代替案(JCOGデータセンター/運営事務局の提案)

治療効果の現れ方が従来の細胞傷害性薬剤と異なる新規薬剤の治療効果を臨床試験で適切に評価する上で，ありうる解決策は「新たな効果判定規準をつくる」ことのみではない．「効果判定」にのみ眼を向けるから「新たな判定規準」をつくるという発想にしかならないのであって，「臨床試験で正しく評価(比較)する」という本来の目的に立ち返って考えるならば，現実の個々の臨床

試験における，エンドポイントの定義，試験デザイン，意思決定の規準等を工夫することをまず考えるべきであろう．

繰り返すが，免疫治療薬の治療開発を進め日常診療に導入する際には，一連の開発の過程で従来の細胞傷害性薬剤との比較が必須である．そのため，免疫治療薬をirRCを用いて評価するのであれば，逆に，従来の細胞傷害性薬剤がirRCによって適切に評価できることが担保されなければならないが，前項で述べたように，それにはかなりの無理がありそうである．

そうすると，われわれが考えなければならないことは，irRCという複雑な規準を新たに持ち出すことなしに，RECIST準拠の範囲内で(多くの試験との比較可能性を保ったまま)，試験デザインや意思決定の規準を工夫することで免疫治療薬の効果を適切に評価することはできないのか？ ということである．以下，「irRCの問題点」で掲げた3点「効果判定規準の意義」，「比較可能性」，「標準化」に言及しながら，開発の相ごとに，エンドポイント，デザイン，意思決定の規準について考察する．

1．第III相試験

第III相試験では，既存の標準治療である細胞傷害性薬剤との比較が行われる(将来，免疫治療が標準治療となった暁には，以後免疫治療の新旧比較となりうるが)．しかし，腫瘍が縮小する前に増大が起こりうる免疫治療では，治療が無効で腫瘍が増大しているのか(true progression)，治療は有効だが一時的に増大しているのか(pseudo progression)を画像所見と臨床所見のみで正確に判別するのは困難である．2013年American Society of Clinical Oncology(ASCO)年次総会の教育講演「Other Considerations in Immunotherapy Trials：Endpoints, Toxicity Management」でも，PSの変化や症状の出現がtrue progressionとpseudo progressionを見極める一助となるが，現時点では両者を鑑別する決定的な方法が存在しないため，少なくとも非小細胞肺がんの免疫治療の臨床試験では，全生存期間がより信頼できるエンドポイントであると結論されていた[13]．したがって，少なくとも標準治療を決めるための第III相試験，すなわちefficacy(薬効)ではなくeffectiveness(臨床的なベネフィット)を評価しようとする検証的試験では，primary endpointは患者の真のベネフィットを反映する全生存期間(overall survival；OS)とするべきであり，細胞傷害性薬剤でもそうであるように，第III相試験においてirRCによる奏効割合をエンドポイントとする必然性はなく，効果判定規準としてirRCを用いる必然性もない．

2．後期第II相試験

細胞傷害性薬剤の後期第II相試験では，当該がん種において腫瘍縮小効果がOSの延長を反映すると見なされており，かつ適切なヒストリカルコントロールがある場合には，腫瘍縮小効果をprimary endpointとした単群試験が行われることが一般的である．しかし，免疫治療薬では腫瘍縮小が必ずしも予後を反映せず，適切なヒストリカルコントロールも存在しない状況もある．そのような状況は免疫治療薬がはじめて遭遇するものではなく，過去に多くの分子標的薬が同様の問題に直面してきた．その場合の解決法の一つは，他の分子標的薬と同様，primary endpointとしてOSを用いたスクリーニングデザインのランダム化第II相試験である．効果判定規準の問題点の解決を，新たな効果判定規準の創出に依るのではなく，試験デザインに求めるという発想である．

さらに，予後がよい対象でOSをprimary endpointとすることが適切でない場合には，これも他の分子標的薬と同様，無増悪生存期間(progression free survival；PFS)をprimary endpointとしたランダム化第II相試験が解決策の候補となる．ただし，RECIST version 1.1での通常の定義のPDをイベントとするPFSをそのまま用いると，irRC提唱者の言う「治療開始後，腫瘍量が増大したあとに減少することがある」という問題点への解決策とはならない．筆者らはこのような課題に対する解決策を見出す必要があるという点について異論はないが，前述の通りRECISTに代わる効果判定規準を新たにつくる必要はないと考えており，代替案の一つとして「landmark method」を用いたPFSをprimary endpointとしたランダム化第II相試験デザインの利用を提案する．

「Landmark method」は，さまざまな状況下で

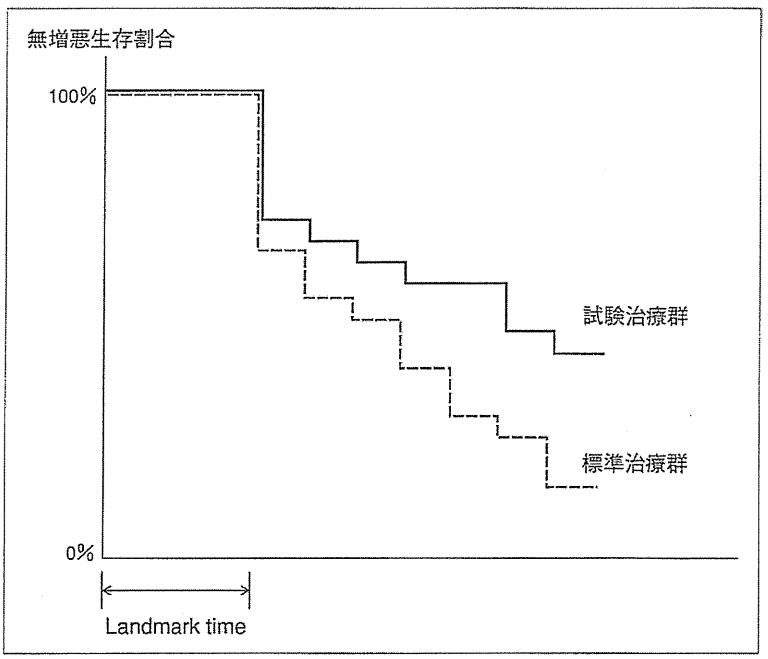

図2　Landmark methodの生存曲線

使われる手法であり目新しいものではないが，免疫治療の臨床試験に用いる場合には，治療開始後の一定期間(「landmark time」と呼ぶ)までに認められた腫瘍の増大あるいは新病変の出現はイベントとしないという形で適用することが考えられる(図2)[14]．試験ごとに決める「landmark time」の時点で，ベースラインと比較して20%以上の径の和の増大(RECISTの1方向測定で評価)を認める場合や，出現した新病変が残存している場合，あるいはlandmark timeまでに死亡した場合にはlandmark timeでPFSのイベントとするという方法である．Landmark timeより前に行われる画像検査や診察での評価では効果判定を行わず，その間，明らかな臨床的増悪が認められた場合のみ，治療を中止する規準を設けることになる．もちろん，この治療中止規準が適切でないと，無効な治療が継続されることになるため，中止規準は慎重に設定する必要がある．たとえば，プロトコール治療中止規準の一つとして，細胞傷害性薬剤の臨床試験と同様に「治療開始後に原病の増悪が認められた場合」を規定し，以下をただし書きとして加える．

> ただし，免疫治療群における「原病の増悪」に基づく治療継続の是非は，免疫治療の特性を考慮して決定してよい．すなわち，治療開始後早期に新病変の出現かつ/またはベースラインで測定した腫瘍の増大で総腫瘍量がPDに相当する場合にも，4週後の再評価まで治療を継続してもよい．

すなわち，効果判定規準としてはあくまでRECISTに準じるが，治療継続の是非はirRCの基本的な考えを導入するという案である．そして，landmark timeで登録時と同じ検査法でRECISTに従って標的病変および非標的病変の評価を行い，総合効果がPDである場合には，通常のRECISTのロジックどおり「確定」を要さず「PD」と判定する．

irRC提唱者らの指摘する問題の多くは，RECISTによる効果判定を行い「増悪による治療中止規準」に上記のただし書きを追加するだけで解決できるはずであると筆者らは考えている．irRCがいかに余分な複雑な論理を持ち込んでいるか，理解いただけると思う．

また，すでにこのlandmark methodは，Ribasらの報告でもirRCという新規準作成の代替案として言及されているのだが，landmark timeの期間設定が難しいこと，予後が数か月に限られる患者には適用できないことを「問題」とし，irRCが正しい方向への第一歩であると結論されている[7]．しかし，Wolchokらの進行期悪性黒色腫患者を対象にした試験では，12週時点で治療開始後早期の評価が行われ，その後4週後にirPDの確定を行うように設定されており，これは，治療早期に腫瘍が増大する場合や新病変が生じる場合でも，初回の効果判定時点から4週後までには腫瘍が縮小する傾向があるという知見に基づいている[9]．これを前提にすると，進行期悪性黒色腫患者においては12週＋4週の16週をlandmark timeに設定すればよいはずであり，このようにlandmark timeを設定することと，再評価の時期を決めて確定を行い効果を判定することは，その元となる原理は同じといえる．また，予後が数か月に限られる患者を対象としている場合には，わざわざPFSを使う必要はなくOSで評価すればよいのであって，予後が限られていることは，landmark methodよりirRCがよいという根拠にはならない．

以上より，後期第II相試験においても，irRCを用いるよりも，臨床的妥当性を損わずに細胞傷害性薬剤との比較可能性が保たれるlandmark methodを用いたPFS，もしくはOSをprimary endpointとしたランダム化第II相試験デザインを用いることを筆者らは推奨する．

なお，有効性の指標としてSDを含めるか否か(CR＋PR＋SDを分子とするいわゆる「腫瘍制御率：disease control rate」をエンドポイントとするかどうか)について，Wolchokらの原著では，「腫瘍縮小効果が長期間ない場合」も治療効果が認められた4つのパターンの1つとして論じている．しかし，SDを「効果あり」と扱うことは，進行の遅いがんでは，治療効果がほとんどない場合，あるいはまったくない場合にも薬剤が有効であると判断され得る点が問題となる．すなわち，特にSDまでを効果ありとして評価するのであれば，その比較対照の選択方法が他の場合に比べてより重要となる．当然のことながら比較対照も同じ規準でSDの判定がされていることが最低限必要な条件となるが，そのような条件を満たす外部対照を得ることはほとんどの場合困難である(つまり，このような状況下で適切な比較対照の取りようがないということは，臨床試験の結果に基づき開発を進めるか否かの判断を下すことができないということに等しい)．実質的には，特に単群の試験では，SDが薬剤の治療効果によるものか否かは判別できないため，この問題に対する解決策も，irRCではなく，標準治療を対照群においたスクリーニングランダム化第II相試験デザインであると筆者らは考える．

3．第I相試験/前期第II相試験

さらに早期の段階の試験では，免疫治療が「有望であるかどうか」を見定めることが目的である．試験結果に基づく意思決定は「さらに(後期)第II相試験に進むかどうか」であるため，「(数百例以上を対象とする)第III相試験に進むかどうか」を意思決定する前項の状況よりもさらに探索的であり，より少数例の試験が適切であることから，(抗がん剤の早期開発試験が一般にそうであるように)要求される比較可能性の厳密さは低くなる．

この場合，前項のランダム化第II相試験デザインは，必要以上の被験者を用いる点でオーバースペックであり，単群の試験が行われるべきだが，やはり既存の細胞傷害性薬剤のヒストリカルコントロールとの比較が前提となる．この状況で，前項で示した，landmark timeまでに腫瘍増大あるいは新病変の出現が認められてもただちにPDとは判定しないことにして求める奏効割合(landmark timeでPR以上の効果が得られていた患者の割合)を用いることに不都合はあるだろうか？

標準治療である従来の薬物療法ではこの特殊な奏効割合を評価していないため直接比較することはできないが，少なくとも免疫治療におけるこの特殊な奏効割合が従来の薬物療法の奏効割合より上回っているならば，その薬剤は有望であると判断して(後期)第II相試験に進めるという意思決定が誤っているとは思えない．また，当該免疫治療の利点に「細胞傷害性薬剤よりも毒性が軽い」があるのであれば，上回っていなくて

も奏効割合が標準治療である細胞傷害性薬剤と同等程度であれば第II相試験に進めることも妥当と判断されよう．

この段階の臨床試験においても，筆者らはirRCよりもlandmark methodを用いたRECIST準拠の奏効割合を用いることを推奨する．

おわりに

哲学の分野には「オッカムの剃刀(かみそり)」という考え方がある．「ある事柄を説明するためには，必要以上に多くを仮定するべきでない」や「同様のデータを説明する仮説が2つある場合，より単純な方の仮説を選択せよ」(Wikipedia)と説明される．

臨床的意思決定においては，たとえば，患者の病状を説明できる病態生理について1元論と2元論が同様のもっともらしさ(plausibility)で考えられる場合，(一刻を争う病状である場合は別だが)1元論に基づく治療介入を優先させる，といった応用がなされる．1元論を採用する方が，治療が功を奏さなかった場合に病態生理の読みを修正して次の手を打つロジックがよりシンプルになり，より速やかに正解(＝患者の病状の改善)に到達する確率が高いと考えるのである．

筆者らは，本稿に求められた問い「irRCかRECISTか？」について，irRCが2元論，RECISTの一部修飾が1元論に相当すると考えており，免疫治療の臨床試験において2元論を持ち出さなくても1元論で対応可能と考えている．

標題の最後の問い「JCOGはどう考える？」について筆者らはこう答えよう．

―JCOG試験におけるirRCの使用をJCOGデータセンター/運営事務局は推奨しない―．

補遺：なお，本稿の内容は，独立行政法人国立がん研究センターがん研究開発費23-A-16(主任研究者：福田治彦)に基づく研究成果によるものである．

文　献

1) Hodi FS, O'Day SJ, McDermott DF, et al. Improved survival with ipilimumab in patients with metastatic melanoma. N Engl J Med 2010 ; 363 : 711.
2) Lynch TJ, Bondarenko I, Luft A, et al. Ipilimumab in combination with paclitaxel and carboplatin as first-line treatment in stage IIIB/IV non-small-cell lung cancer : results from a randomized, double-blind, multicenter phase II study. J Clin Oncol 2012 ; 30 : 2046.
3) Topalian SL, Hodi FS, Brahmer JR, et al. Safety, activity, and immune correlates of anti-PD-1 antibody in cancer. N Engl J Med 2012 ; 366 : 2443.
4) Brahmer JR, Tykodi SS, Chow LQ, et al. Safety and activity of anti-PD-L1 antibody in patients with advanced cancer. N Engl J Med 2012 ; 366 : 2455.
5) Hoos A, Eggermont AM, Janetzki S, et al. Improved endpoints for cancer immunotherapy trials. J Natl Cancer Inst 2010 ; 102 : 1388.
6) Hodi FS, Butler M, Oble DA, et al. Immunologic and clinical effects of antibody blockade of cytotoxic T lymphocyte-associated antigen 4 in previously vaccinated cancer patients. Proc Natl Acad Sci U S A 2008 ; 105 : 3005.
7) Ribas A, Chmielowski B, Glaspy JA. Do we need a different set of response assessment criteria for tumor immunotherapy? Clin Cancer Res 2009 ; 15 : 7116.
8) US Food and Drug Administration. Guidance for Industry : Clinical Considerations for Therapeutic Cancer Vaccines. 2011. Available from : URL : http://www.fda.gov/downloads/biologicsbloodvaccines/guidancecomplianceregulatoryinformation/guidances/vaccines/ucm278673.pdf.
9) Wolchok JD, Hoos A, O'Day S, et al. Guidelines for the evaluation of immune therapy activity in solid tumors : immune-related response criteria. Clin Cancer Res 2009 ; 15 : 7412.
10) Miller AB, Hoogstraten B, Staquet M, Winkler A. Reporting result of cancer treatment. Cancer 1981 ; 47 : 207.
11) Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours : revised RECIST guideline (version 1.1). Eur J Cancer 2009 ; 45 : 228.
12) Nishino M, Giobbie-Hurder A, Gargano M, et al. Developing a common language for tumor response to immunotherapy : immune-related response cri-

teria using unidimentional measurements. Clin Cancer Res 2013 June 6 [Epub ahead of print].
13) Chow LQ. Exploring novel immune-related toxicities and endpoints with immune-checkpoint inhibitors in non-small cell lung cancer. Am Soc Clin Oncol Educ Book 2013 ; 2013 : 280.
14) Anderson JR, Cain KC, Gelber RD. Analysis of survival by tumor response. J Clin Oncol 1983 ; 1 : 710.

* * *

特集　治療効果の判定基準と臨床試験のendpoint

PFS or OS
1）総論：PFSは第III相試験のprimary endpointとなりうるか？―知っておくべき考え方のフレームワーク*

中村健一**
水澤純基***
柴田大朗***
福田治彦***

Key Words：progression-free survival, clinical trial, true endpoint, surrogate endpoint, hybrid design

はじめに

従来，第III相試験のprimary endpointのgold standardは全生存期間(overall survival；OS)であったが，近年各種のがんで有効な治療法の開発が進む中，第III相試験のprimary endpointとして無増悪生存期間(progression-free survival；PFS)が用いられる試験が増えてきた．そもそもOSがgold standardとして受け入れられ続けてきた理由は，それが患者のベネフィットを直接反映する指標であり，かつ，死亡という誰が見ても迷わない事象をイベントとするハードなエンドポイントであったからである．これは万人が認めるところであるが，①OSはプロトコール治療が中止となったあとに行われる後治療の影響を受けるため，OSでは差がつきにくい，②標準治療を行った際のベースとなるOSが延長するに従って，同じOSの上乗せ幅であってもサンプルサイズが飛躍的に増加する，といった理由により，実際にはOSが使いにくいという状況が各種のがんで生じている[1]．しかし，①に対しては後治療によってOSに差がないのであれば，そもそもフロントラインで新たな(そしてしばしば高価な)治療を第一選択として用いる意義はないという反論があり，また，②については臨床的に意味の「ない」差を検出しても仕方ないという反論がある．ではそもそもPFSのイベントは客観的に拾えているのか，という違った角度からの議論もあり，学会などでのPFSをめぐる議論は堂々巡りになって結論が出ないということがよくみられる．

本稿ではこれらの混沌とした状況を整理すべく，PFSをめぐる議論を行う際のフレームワークとすべき基本的な考え方を提示したい．なお，PFSがprimary endpointとなりうるかどうかは試験の相や目的によって異なるが，ここではefficacy(薬効があるかどうか)ではなくeffectiveness(clinical benefitがあるかどうか)の検証を目的とする，つまり，標準治療を決定するための第III相試験を想定して議論を進める．新薬の早期開発で，一定の薬効(efficacy)の存在をまず確かめる必要がある場合などは，異なる考え方がありえるため注意が必要である．

エンドポイントの定義とフレームワーク

米国National Cancer Institute(NCI)の統計家

* Will progression-free survival be appropriate as a primary endpoint in phase III trials?

** Kenichi NAKAMURA, M.D.：独立行政法人国立がん研究センター多施設臨床試験支援センターJCOG運営事務局〔〒104-0045 東京都中央区築地5-1-1〕；JCOG Operations Office, Multi-institutional Clinical Trial Support Center, National Cancer Center, Tokyo 104-0045, JAPAN

*** Junki MIZUSAWA, M.Sc., Taro SHIBATA, M.Sc. & Haruhiko FUKUDA, M.D.：独立行政法人国立がん研究センター多施設臨床試験支援センターJCOGデータセンター

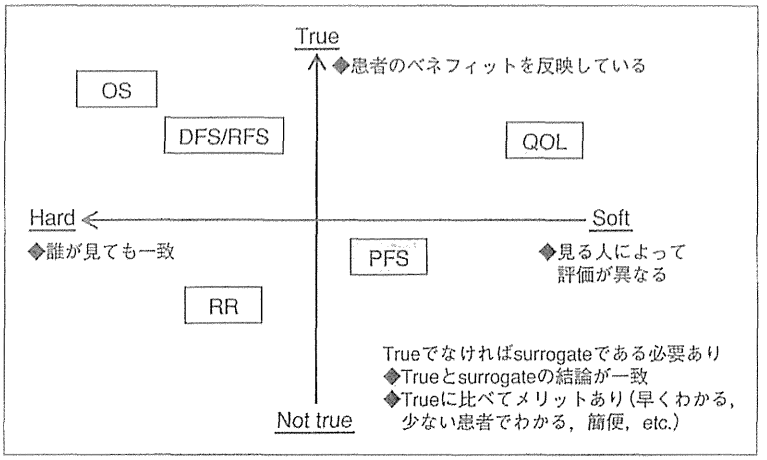

図1　True軸とHard-soft軸

であるRichard Simonは，エンドポイントとは「患者のベネフィットを測るものさし」であると定義している．Primary endpointは主要評価項目，secondary endpointは副次的評価項目と訳されることが多いが，Simonの定義に従えばエンドポイントは単なる評価項目というだけでは不十分で，患者のベネフィットを反映するものであり，かつ，ものさしとして正確に測定できる必要がある．本稿では前者の「患者のベネフィットを反映しているかどうか？」をtruenessと呼び，後者の「きちんと測れているか？」をhardnessと呼ぶ．

1．True軸

Truenessの観点からみた際のエンドポイントには2種類存在する．患者の真のベネフィットを反映するエンドポイントがtrue endpointであるが，患者にとってのベネフィットが明らかではないエンドポイントであれば，true endpointの「代替(surrogate)」として用いられるsurrogate endpointである必要がある．True endpointとしていちばんわかりやすいのはOSで，長生きすることが患者にとってのベネフィットではないと主張する研究者は存在しないだろう．これに対して典型的なsurrogate endpointは奏効割合であり，腫瘍が縮小すること自体は患者のベネフィットとはいえないが，腫瘍が縮小することが治療効果の指標となり，それがしばしば延命につながることから古くからOSの代替(surrogate)として頻用されてきた．奏効割合が主に第II相段階で頻用されてきた理由にはOSに比べて結果を早くに知ることができるというメリットの存在があげられる．つまり良いsurrogate endpointであるためには，①true endpointとよく相関し，②surrogate endpointを用いることでなんらかのメリットが存在することが必須である．ではすべてのエンドポイントはtrue endpointかsurrogate endpointに分類できるかというとそうではなく，世の中にはtrueでもなくsurrogateでもないエンドポイントが多数存在する．True軸で考えた際に優れたエンドポイントであるためには，trueであるか，あるいはtrueでなければsurrogate endpointの要件を満たしているか，いずれかを満たす必要があり，いずれも満たさないエンドポイントはtrue軸の観点から考えると不適切ということになる(図1)．

2．Hard-soft軸

一方，Simonが定義したようにエンドポイントは「ものさし」であるため，ものさしとしての機能が優れている必要がある．優れたものさしとは，誰が見ても同じ結果を導き出せるということであり，このようなエンドポイントをhardなエンドポイントという．たとえば死亡日は誰が見ても違わないし，同じ条件で測定されるCTでの腫瘍径も比較的hardであるといえる．これに対して痛みのスケールなどは我慢強い人かどうかで点数は変わってくるだろうし，鎮痛剤を飲むタイミングによっても点数は変わるだろう．このように評価者の主観的な判断や他の要因により評価が異なりうるエンドポイントのことを

softなエンドポイントという．試験のprimary endpointとして設定するにはhardであることは必須の条件であり，hard-soft軸で考えた際に優れたエンドポイントであるためには，hardの矢印の先端に近づくことが必要不可欠である(図1)．

さまざまなtime-to-event endpoint

これらtrue軸と，hard-soft軸で考えた際に，一般に受け止められている各エンドポイントの位置関係は図1のようになる．

無再発生存期間(relapse-free survival；RFS)と無病生存期間(disease-free survival；DFS)もPFSと名前は似ているが，trueness, hardnessのいずれの観点からもまったく別物と考えるべきである．RFSは再発と死亡がイベント，DFSは再発，二次がん，死亡がイベントであるが，この両者はいずれもいったんは手術などの治療により無病状態となり，その無病状態が続いている期間を測定していることになる．再発をきたした場合，多くはその後の治癒は望めず死亡につながるため，再発は患者にとってtrueな意味を持つ．

それに対してPFSは担がん状態にある患者で，腫瘍ボリュームが一定の割合で大きくなるまでの期間である．腫瘍径の20%増といってもベースの腫瘍径が2 cmの場合と10 cmの場合には自ずとその「20%増」の持つ意味は違うであろうし，腫瘍が大きくなったとしても症状はなく，生存期間が変わらないのであれば，腫瘍径の増大は患者にとってtrueではないといえよう．また，たとえば腹膜播種病変の「増悪」の判定は難しいが，肝臓や肺に出現した新病変の判定は比較的ぶれが少ないだろう．Hardnessの観点からもPFSは，DFSやRFSよりsoftということができる．

TruenessからみたPFS

エンドポイントを考える際に必要なフレームワークとして，truenessとhardnessという2つの観点を述べたが，このtruenessから考えた際には，PFSがprimary endpointとして用いることのできる条件は2つしかない．すなわち，1. PFSがOSのsurrogate endpointとなることが示されている，2. PFS自体が患者にとってのclinical benefitを持つ，のいずれかである．

1. PFSがOSのsurrogate endpointである

PFSがOSのsurrogate endpointであることが確立されている場合には，PFS自体がclinical benefitを持つことを示す必要はない．あくまで，PFSがpositiveであれば，OSもpositiveとなることが前提であり，clinical benefitを持つOSの延長を正しく予測できるという点にPFSを用いる意味がある．各種のがんのメタアナリシスでOSに対するPFSのsurrogacyを検証しようとする試みがなされているが，現在のところ一定のsurrogacyが示されているのは大腸がん，頭頸部がんのみであり．反対に胃がん，乳がんなどではsurrogacyがないとされている．

ただ，たとえば大腸がんであってもsurrogacyが示されたメタアナリシスで使われている試験は有効な薬剤が少なかった1980年代から90年代にかけてのものであり，ほとんどがPFSは6か月，OSは12か月程度の試験である[2]．現在は有効な薬剤が増えるにつれて一次治療で増悪となったあとの生存期間(post-progression survival；PPS)が延長する傾向にあるが，PPSが延長すると一般にPFSのOSに対するsurrogacyは示しにくくなる[1]．有効な薬剤が増えつつある現状を考えると，今後大腸がんを含めた多くのがん種でPFSのOSに対するsurrogacyを示しにくくなると予想される．事実，大腸がんでは最近の分子標的薬を含む試験も含めた大規模なメタアナリシスで，PFSがOSのsurrogate endpointとはならなかったという発表がなされている[3]．

一方，RFSやDFSがOSのsurrogate endpointであることが示されているがん種では，多くの場合OS全体に占めるRFSの期間が長いことから，再発後の生存期間が延長したとしても影響は比較的限定的であり，現在OSに対するsurrogacyが示されている胃がんや大腸がんでは，当面このsurrogacyを維持し続けると予想される．

2. PFS自体がclinical benefitを持つ

PFSをprimary endpointとできるもう一つの条件は，PFS自体がtrue endpointであること，つまりなんらかのclinical benefitを持つことである．PFSではないが，頭頸部がんにおける喉頭温存生存期間はOS以外のtrue endpointの好例である．喉頭摘出することになれば，発声，味覚といっ

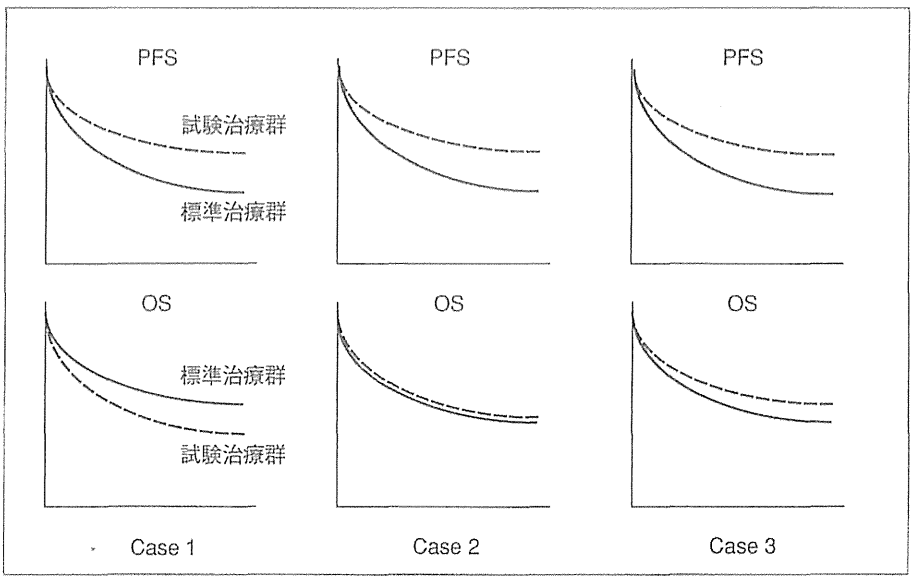

図2　どの状況で試験治療を取るか？

た点に大きな障害が生じることから，仮にOSに差がまったくなかったとしても喉頭温存生存期間が延長するということは患者のベネフィットをダイレクトに反映すると考えられる．

では一般的なPFSの場合はどうであろうか？この点についてFlemingらは，「4～6週間増悪までの期間が延びれば，患者は少しの間，心の平安が得られるかもしれないが，そうしたベネフィットは毒性によるquality of life（QOL）低下や，コスト，投薬に伴う不便さにより相殺されてしまう」，「同じ20％の径和増大といっても，小さな病変が20％増大した時と，大きな病変が20％増大した時では意味が異なる」と述べている[4]．Clinical benefitを持つかどうかあやしい状況でPFSが多少延長したとしても，PFSに比べてPPSがどんどん長くなってきている現状では，PFSのわずかな延長が持つ意味はさらに薄れているといえよう．

一方，PFSがtruenessを持つ場合，PFSのみで標準治療を決めてしまってよいだろうか？PFSにおいて試験治療が標準治療を上回ったとしても，OSで下回っていれば多くの場合には試験治療が選択されることはないだろう．つまり，試験のdecision ruleを決定するにあたっては，OSのtruenessの大きさとPFSのそれとを相対的に比較し，それぞれのtruenessの大きさに応じたdecision ruleを決める必要がある．図2に示したように，どの状況でその試験を“positive”と結論づけるか，ということは，このPFSのtruenessの「強さ」を考察することにほかならない．試験治療のOSが下回ってもPFSが上回っていれば試験治療を取るという状況（Case 1）はほとんどないと考えられるが，先に例としてあげた喉頭温存生存期間などはOSがぴったり重なっていたとしても試験治療を取る状況はありえる（Case 2）．多くのPFSでは，一定のtruenessを持っていたとしても，そのtruenessは価値観に左右される程度のものであり，OSで一定以上上回らないことには（Case 3），試験レジメンを新たな標準治療として受け入れがたいという状況が多いと思われる．この点についてはあとで再び考察する．

以上のような状況から，標準治療を決定する第III相試験において，PFSをprimary endpointとして用いるための必須条件は，1．PFSがOSのsurrogate endpointとなることが示されている，2．PFS自体が患者にとってのclinical benefitを持つ，のいずれかを満たし，かつ，hardであること，となる．このような観点から，NCIの統計家であるKornらも，「clinical benefitを直接的に測る中間的なエンドポイントが存在しないがん種

では，OSをprimary endpointとすべき」と結論している[5]．このrecommendationに従えば，OSに対するsurrogacyが示されておらず，PFSがtrueともいえないがん種では，当面OSをprimary endpointとして使い続けるしかないという身も蓋もない結論になる．

現実的な解決策はあるか？

しかし，現実問題として，OSが20か月以上になるようながん種では，OSをprimary endpointとしてランダム化比較試験を組むことは，登録数の観点から困難である場合も多い．たとえばOSの3か月の延長が「臨床的に意味のある差」である場合，6か月に対して3か月の上乗せを検証するには212例でよかったが，これが12か月に対する3か月の上乗せだと838例が必要になり，18か月だと2,092例，24か月だと4,194例が必要となる(話を単純化するために片側 $\alpha=0.025$，$\beta=0.2$，登録3年，追跡1年に固定)．理論的にOSが望ましいというのは簡単であるが，ではどのようにしてこのような難しい状況のもとで治療開発を進めていくのか，ということを臨床試験の実務家としては考える必要がある．

PPSが延長するに従って増え続けるサンプルサイズに対する対処法としては，次にあげる3通りの方策が考えられる．

1．大規模なサンプルサイズに対応できる体制を構築する

最初の選択肢は最もストレートな解決策であるが，単一の臨床試験グループでは困難なサンプルサイズの試験を，共同試験という形を取ることで実施可能にするというアイデアである．近年，国際共同試験や国内共同試験の必要性が叫ばれているが，共同試験の重要な動機の一つは，OSをprimary endpointとした時に増え続けるサンプルサイズである．ほかにも共同試験の動機には，個別化治療が進むに従ってマーカー別に細分化される試験の対象に対して迅速に患者登録を進めるといったことや，(治験の場合は)同時治療開発によるコストダウンや，結果を一般化できる地域を広げたいといった要因も存在するが，いずれにしても単にお友達を増やそうというnaïveな感情が共同試験の動機ではないことを理解しておく必要はあるだろう．一方で，国際共同試験では国別にアウトカムの差があることが知られておりデメリットも存在する(単に人種間差ということだけではなく，適格性の診断精度の違いや，保険制度の違いによる後治療の差など多くの要因が原因となりうる)．これらのデメリットが存在することから，比較的文化や習慣が似た地域のグループと共同試験を行えるような体制作りが理想的で，それらのグループとお互いの信頼関係を醸成するとともに実施体制の標準化を進めるべきであろう．

2．デザインの工夫によるサンプルサイズの調整

(1)試験の精度を落とす

試験対象となる集団の予後は変えることができないため，通常の優越性デザインを取る限り，サンプルサイズを減らすために緩められるパラメーターは，α(有意水準)，β(本当は有効な治療を捨ててしまう確率．1-βが検出力)，Δ(臨床的に意味のある差)の3つしかない．Δはこれぐらい差があれば，試験治療の毒性などのデメリットを考慮に入れても試験治療を新しい標準と見なしうるといった臨床的な感覚を反映した数字であるため，このΔを緩めてしまうと臨床的に意味のある差でもnegativeになってしまう．そのため臨床的な意思決定と試験結果を一致させる観点からはΔを緩めることはおすすめできない．

βを緩めるということは検出力を落とすということであり，本当は試験治療が良い時に，試験結果としてはnegativeとなってしまう確率が増えてしまうことを意味する．もちろんβも小さい(検出力が高い)ことに越したことはないのだが，他の2つのパラメーターの優先度が高いため，一定の効果が持つ標準治療が存在する状況では先にβが犠牲にされるケースが多い．

αを緩めるということは，本当はよくない治療法をpositiveと結論してしまう危険性を高めることであり，患者に効果のない治療法を還元することを避けるという観点ではαを大きくするべきではない．ただし，試験全体のα(有意水準)は両側5%もしくは片側2.5%が世界標準であるが，対象疾患の患者数が限られている場合には片側5%や，かなりの希少がんである場合には片

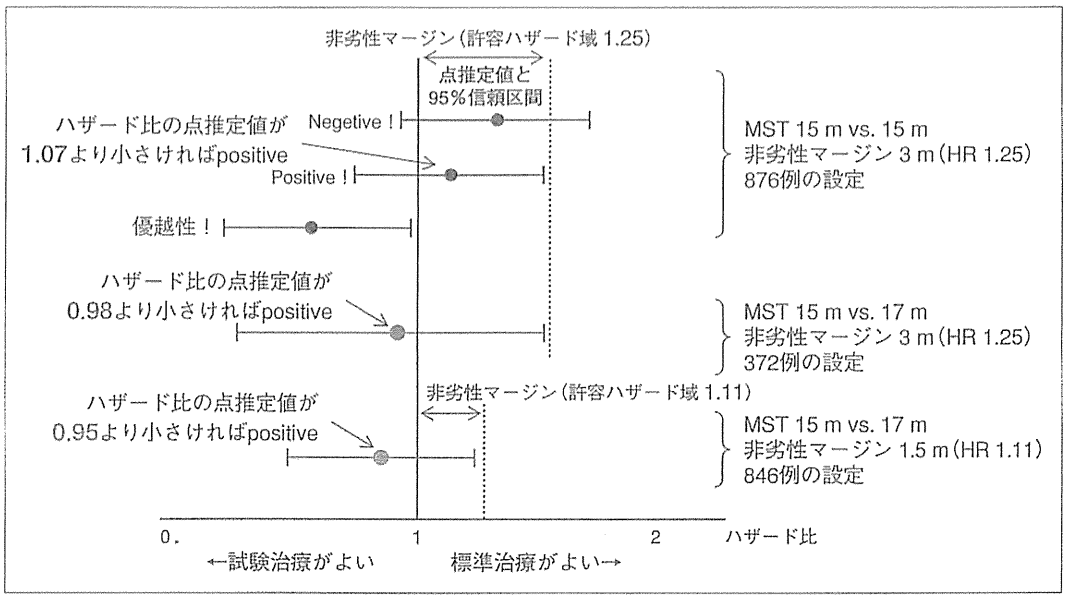

図3　"Hybrid design"と判断境界(boundary)

側10%も許容されるかもしれない．以上のような観点から，一般論として，サンプルサイズが実現不可能なほど多い場合には，まずβ，次にをαを緩めることが奨められ，Δを緩めることは最後まで避けるべきである．

(2)"Hybrid design"

ある特殊な状況では，NCIのFreidlinらが"hybrid design"と呼んでいる非劣性試験の設定を行うことにより実現可能なサンプルサイズ設定を行える場合がある[6](なお，世の中にはさまざまな"hybrid design"が存在し，このデザインだけが一意に"hybrid design"と呼ばれているわけではないことには注意が必要である．また，これは特殊なデザインではなく，単に一般的な非劣性試験の標準治療群と試験治療群で見込まれる予後の設定を変えただけにすぎない)．

非劣性試験では特に事情がない限りは標準治療と試験治療で見込まれる予後は同じに設定される．たとえば標準治療群と試験治療群で見込まれる生存期間中央値(median survival time；MST)がどちらも15か月で，非劣性マージンが3か月といった具合である．この時片側有意水準2.5%，検出力80%，登録3年，追跡1年とするとサンプルサイズは876例となる．これに対して"hybrid design"では少し試験治療の予後が上回るという仮定を置く．たとえば，標準治療群で見込まれるMSTが15か月，試験治療群で見込まれるMSTが17か月，非劣性マージンが3か月といった具合である．先ほどと同じ有意水準，検出力などを用いた場合にこの設定だとサンプルサイズは372例にまで減る．

ただし，サンプルサイズが減った分，得られた結果(ハザード比)の信頼区間の幅は広くなり，positiveになるために越えるべきハードル(判断境界＝boundaryという)は高くなる(図3)．たとえばMSTを両群で15か月と見込んだ場合のboundaryは1.07であり，実際に得られたハザード比の点推定値が1.07より小さければP値が2.5%を下回ることになる．つまり試験治療群の生存曲線が少しだけ劣っていてもpositiveになりうる．これに対して標準治療群のMSTを15か月，試験治療群のMSTを17か月(他のパラメータは同じに設定)とした場合のboundaryは0.98となる．つまり試験治療群の生存曲線がわずかではあるが上回っていないとpositiveにはならない．

ちなみに標準治療群のMSTを15か月，試験治療群のMSTを17か月として，非劣性マージンを3か月から1.5か月に狭めた場合，その分サンプルサイズは大きくなり846例と当初の設定とほぼ同じ程度となる．この場合のboundaryは0.95とな

り，試験治療群が少し勝たなければpositiveとならない設定となる．試験治療が標準治療と比較してなんらかのメリットは持つものの，それがさほど大きなメリットではなく，あまり大きな非劣性マージンが許容されない場合などに，このように「少し勝たなければpositiveにならない」という設定にすることはありえる．

この“hybrid design”は設定次第でサンプルサイズを劇的に減らすことができるデザインの方法であるが，適用にあたって２つの注意点がある．まず１点目は，“hybrid design”はあくまで非劣性試験の亜型であり，標準治療に比べて試験治療にOS以外のなんらかのメリットがなければならないという点である．“Hybrid design”を取るにしても生存曲線がほぼ重なっていた際に“positive”という結論になることは変わりないので，OS以外に毒性が軽い，投与が簡便，といったなんらかのメリットがないことには試験治療を選択することが正当化されない(ただし，先ほどの例のように非劣性マージンがかなり小さい場合には，示すべきメリットも相対的に小さくてよい)．世の中には標準治療と試験治療の毒性プロファイルが異なり，試験治療がtoxic newともless toxicともいえないような，equitoxicとでもいうべき状況が存在するが，このような場合にも試験治療に少なくともなんらかのメリットが存在する場合には，この“hybrid design”が適用できる可能性がある．

２点目の注意点は，試験治療が少し上回るという仮定を置く以上，過去のデータなどから試験治療が少し勝つ見込みがなければならないという点である．“Hybrid design”ではサンプルサイズを減らした分，試験がpositiveとなるために越えるべきハードルが高くなるので，本当に勝つ見込みがなければpositiveとなる可能性がその分低くなる．つまり，単なるunderpowerの試験を行っているだけ，ということになりかねない．“Hybrid design”は設定次第でサンプルサイズを劇的に減らすことができるデザインであるが，その適用にあたっては上記の２点がクリアできているか慎重に検討してから適用すべきである．

3．PFSとOSのどちらをもdecision ruleに組み込む

３つ目のオプションとしては，PFSに多少なりともclinical benefitがあることが示される場合に，PFSでサンプルサイズ設定を行い，さらにOSによるdecision ruleを組み込むという方法が考えられる．PFSでサンプルサイズ設定を行うことによって登録数は現実的な範囲内におさめ，試験のdecision ruleとしてはPFSが有意で，かつOSで試験治療が標準治療を一定以上上回った場合に“positive”と結論づけるというデザインである．

喉頭温存生存期間のように誰の目にも明らかなclinical benefitを持つ場合には，OSが重なっていたとしても試験治療を選択するであろうが，ほとんどの場合PFSにclinical benefitがあるといっても，OSのbenefitに比べると相対的に小さいことがほとんどである．そのような状況では，試験治療の真のOSが，標準治療の真のOSを下回っている際に，間違って試験治療を取ってしまう確率を一定以下に抑えるために，OSが一定以上上回る(例：OSのハザード比の点推定値が0.85を下回る＝検定は行わない)というdecision ruleを設ける，というのがこのデザインの基本的なアイデアである．

OSがどれほど上回ればpositiveと結論するかは，喉頭温存生存期間のような明らかなclinical benefitを有するエンドポイントである場合にはこのハードルは低くて済むであろうし，PFSのclinical benefitがそこまで明確でない場合には，このハードルは相対的に高くなる．ただ，少なくともPFSがなんらかのclinical benefitを持つ(trueである)ことが前提であり，そうでなければtrueでもなくsurrogateでもないPFSをdecision ruleに組み込むことが正当化されない．また，OSのdecision ruleについては，αとβを大幅に緩めた検定を行っていることと同じであるため，この精度が著しく低くないことは確認する必要がある．

Japan Clinical Oncology Group(JCOG)では，まだこのデザインの実例はないが，1．や2．のオプションが適用できない場合に考慮に値するデザインであると考えている．ただし，いまだ広く受け入れられているデザインではないため，設定に際しては生物統計家への相談が必須で

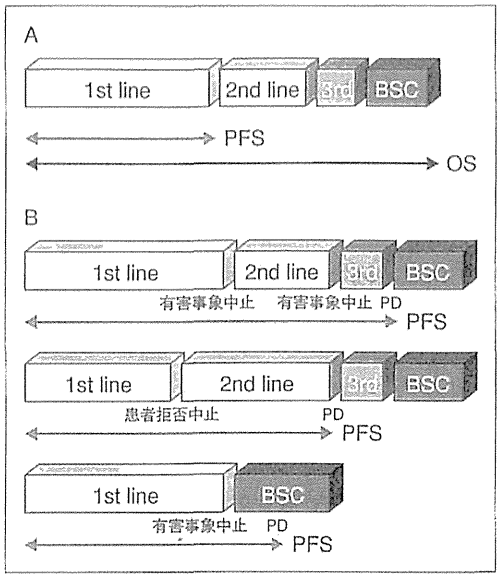

図4　PFSは一次治療の有効性の指標か？

ある．

なお，企業治験などではPFSとOSにたとえば2.5%ずつ有意水準を分割するようなデザインが取られることもあるが，かえってサンプルサイズは大きくなり，PFSがpositiveとなった段階で(OSの結果が出るのは通常その数年後となる)，Food and Drug Administration(FDA)のaccelerated approvalを取りに行くといった企業戦略に基づいたデザインであることから，研究者主導臨床試験で適用できる状況は少ないと考えられる．また，PFSの帰無仮説のハードルをあげる方法も提唱されているが(通常PFSの信頼区間上限が1を下回ればpositiveとなるが，0.9や0.8を下回るべきというdecision rule)[4]，これもPFSがOSに対して一定のsurrogacyを持つことが前提となっており，本稿では割愛する．

HardnessからみたPFS

これまで，truenessの観点から考えた際には，PFSがprimary endpointとして用いることのできる条件は，①PFSがOSのsurrogate endpointとなることが示されている，②PFS自体が患者にとってのclinical benefitを持つ，の2通りしかないということと，状況によって取りうるデザイン上の工夫について述べてきた．ただ，PFSがtrueであれ，surrogateであれ，hardであることは必須条件である．PFS評価にはさまざまなバイアスが入りうることは，多くの論文で述べられておりここでは詳述しないが，代表的なものだけでもassessment bias(検査間隔が群間で異なったり，増悪の評価自体が客観的でない場合に生じるバイアス)，attrition bias(主に試験治療群で評価の脱落が多くなるバイアス)[7]，術前療法の試験で生じるlead time bias(手術日のタイミングが群間で異なることで生じるバイアス)[8]といったバイアスがあげられる．注意が必要なのは，ほとんどの場合これらのバイアスは試験治療群に有利に働くという事実であり，PFSをprimary endpointに設定する際には，これらのバイアスを最小化する努力が払われるべきである．

Hardnessの観点に関連して，PFSをめぐる議論であまり意識されていない事実を一つ指摘しておきたい．よくPFSは一次治療の有効性の指標であるという説明がされ，図4-Aのようなスライドが示されることがあるが，これは誤解を招くシェーマである．特に一次治療レジメンの毒性が強い場合などは，プロトコール治療中止理由の半分以上が増悪中止以外の毒性中止や患者拒否中止であることがある．図4-Bに示すように，この場合のPFSのイベント日は一次治療が増悪中止となるまでの期間ではなく，多くのケースで二次治療や三次治療の増悪中止までの期間である(試験によっては，毒性中止や患者拒否中止の時点で「打ち切り」としている場合もある．ただし，打ち切りとすると治療効果の推定にバイアスを生じさせる可能性がある)．純粋な一次治療の治療中止までの期間を評価したいのであれば治療成功期間(time to treatment failure；TTF)を評価すべきであり，TTFとPFSを混同したような議論が行われることがあるため注意が必要である．

また，hardnessという観点では，実は臨床試験グループや試験によってPFSの定義がまちまちであるという事実にも目を向ける必要がある．JCOGにおけるPFSの標準的な定義は画像によるPDと臨床的増悪(symptomatic deterioration)をイベントとし，最終無増悪生存確認日で打ち切りにするというもので，治療中止日や後治療開

始日では打ち切りにしていない．ただ，試験によっては臨床的増悪をイベントにしない定義や(この場合にはしばしばattrition biasが生じる)，後治療開始日で打ち切りにするといった定義，打ち切り日を最終無増悪生存確認日ではなくて最終生存確認日にするといった定義が取られる場合もある．意図的かどうかはわからないが，PFSの結果を提示する際に生存曲線に打ち切りのヒゲを表示せずに発表を行っている場合もあり，ひとくちにPFSといっても定義も質もさまざまであることは念頭に置いた上で結果の解釈を行うべきである．

おわりに

本稿でこれまで行った議論は，“effectiveness”の評価，つまり，患者にとって第一選択として推奨すべき標準治療を確立するランダム化第III相試験を前提としている．新薬に薬効(efficacy)が存在するかどうかを探索する試験では薬効をよりシャープに評価しうるPFSやTTFをprimary endpointとすべき状況はありうるだろうし，研究者主導試験であってものちに第III相試験が控えている第II相試験などのスクリーニング段階ではPFSをもっと積極的に用いるべき状況もありうると考えられる．「PFSはprimary endpointとして使えるかどうか？」という議論は十把一絡げにできるものではなく，がん種や試験の相，試験の目的によって変わりうる．ただし，標準治療を確立する第III相試験では本稿で述べたようなtruenessやhardnessを十分考察の上で，PFSをprimary endpointとすべきかどうかを決定すべきである．PFSをめぐる議論は往々にして議論のポイントが散漫となり，結論が出ないままに終わることが多いが，本稿で示したフレームワークやデザインの工夫がそれらの議論の一助となれば幸いである．

補遺：なお，本稿の内容は，独立行政法人国立がん研究センターがん研究開発費23-A-16(主任研究者：福田治彦)に基づく研究成果によるものである．

文　献

1) Broglio KR, Berry DA. Detecting an overall survival benefit that is derived from progression-free survival. J Natl Cancer Inst 2009 ; 101 : 1642.
2) Buyse M, Burzykowski T, Carroll K, et al. Progression-free survival is a surrogate for survival in advanced colorectal cancer. J Clin Oncol 2007 ; 25 : 5218.
3) Shi Q, De Gramont A, Buyse ME, et al. Individual patient data (IPD) analysis of progression-free survival (PFS) versus overall survival (OS) as an endpoint for metastatic colorectal cancer (mCRC) in modern trials : Findings from the 16,700 patients (pts) ARCAD database [abstract]. J Clin Oncol 2013 ; 31 Suppl : 3533.
4) Fleming TR, Rothmann MD, Lu HL. Issues in using progression-free survival when evaluating oncology products. J Clin Oncol 2009 ; 27 : 2874.
5) Korn EL, Freidlin B, Abrams JS. Overall survival as the outcome for randomized clinical trials with effective subsequent therapies. J Clin Oncol 2011 ; 29 : 2439.
6) Freidlin B, Korn EL, George SL, Gray R. Randomized clinical trial design for assessing noninferiority when superiority is expected. J Clin Oncol 2007 ; 25 : 5019.
7) Freidlin B, Korn EL, Hunsberger S, et al. Proposal for the use of progression-free survival in unblinded randomized trials. J Clin Oncol 2007 ; 25 : 2122.
8) Nakamura K, Shibata T, Takashima A, et al. Evaluation of three definitions of progression-free survival in preoperative cancer therapy (JCOG0801-A). Jpn J Clin Oncol 2012 ; 42 : 896.

*　　*　　*