表1　検査や診断法の妥当性の評価

		真とする結果 gold standard		尺度
		あり	なし	
検査結果	陽性	真陽性 true positive	偽陽性 false positive	陽性的中度 positive predictive value （真陽性÷検査陽性）
	陰性	偽陰性 false negative	真陰性 true negative	陰性的中度 negative predictive value （真陰性÷検査陰性）
尺度		感度 sensitivity （真陽性÷あり）	特異度 specificity （真陰性÷なし）	

きるが，情報バイアスはランダム化によっては小さくできないため，サンプルサイズが十分大きいランダム化比較試験であるからといって情報バイアスから逃れられるわけではない(少なくない臨床医が，ランダム化しているからバイアスは問題にならないと誤解している)．

また，cisplatinと比べてcarboplatinの嘔気が軽いことは確立された知見であるが，開発段階で第II相試験までの毒性に関する情報が必ずしも第III相試験以降で再現されるとは限らない(第II相試験までの毒性データは真の毒性を過大評価もしくは過小評価したものであるかもしれない)ことから，新薬治験の第III相試験においても，情報バイアスによって誤った意思決定につながる情報が生み出される可能性は常にあるといえる．そして，毒性が必発であり，その情報を患者さんに伝え，患者さんの毒性の情報が担当医に適切にフィードバックされることを通じてリスクを最小化することが一般的となっている「がんの臨床試験」においては，こうした情報バイアスは不可避である．この情報バイアスと比較可能性の問題はHRQOLだけではなく，PROによる毒性評価にもあてはまる．

一方，がん以外の臨床試験や，がん患者を対象とはしていても(治療をマスキングすることによる患者さんのリスクの増大がないとみなせる)制吐剤や鎮痛剤の臨床試験では，こうした情報バイアスはプラセボを用いた二重盲検により回避できるため，PROによる毒性評価やHRQOLによる評価は比較可能性が保たれうる．プラセボ対照を用いたランダム化試験は，車の例でいえば，内装と外装を加工して，試乗する車がJとDのどちらであるかわからないようにした上で「乗り心地」を答えてもらい，「どちらの車の乗り心地がよいか」をもって車の優劣をつけることに対応する．QOL調査否定派の言はとかく誤解されがちなので念のために述べるが，筆者は「すべてのQOL評価(やPRO)」に否定的なのではない．「盲検(マスキング)ができないがん治療の臨床試験」において，治療法の優劣を決定する意思決定に用いる評価指標としてHRQOLやPROを用いることは，誤った意思決定を導きうる点において有害ですらありうることを認識する必要があると考えるのである．

QOL尺度の妥当性評価：いわゆる“validation”

臨床検査や診断法の妥当性は，表1に示すように，通常，なんらかの「真」とする結果(たとえば病理診断でのがんか非がん．これを「gold standard」と呼ぶ)を定め，検査や診断に基づく判定(たとえば内視鏡肉眼診断での良悪性)と真の結果との一致/不一致を評価し，真陽性，偽陽性，真陰性，偽陰性の4つの区分から，感度，特異度，陽性的中度，陰性的中度を求めることで評価がなされる．こうした評価を行うことを「妥当性の検討(または検証)(validation)」と呼び，従来の検査法や診断法に比べて，新しい検査法や診断法の感度や特異度が十分高い，もしくは遜色がなければ，新しい検査法や診断法の「妥当性が検証(validate)」されたとする．

一方，QOL調査票(自記式アンケート用紙)に

ついて行われる「validation」はこれとは異なる方法による．それは「計量心理学(psychometrics)」的手法によるもので，もともと測れない，すなわち「gold standard」を置くことができない，「心」や「知能」等の評価に用いられる方法論である．計量心理学的な“validation”における指標には，大きく分けて「信頼性(reliability)」と「妥当性(validity)」がある．

詳細は別稿に譲るとして簡単に説明すると，「信頼性」は，「得られる値が安定しているかどうか」の指標であり，「再現性(reproducibility)」と「内的整合性(internal consistency)」からなる．平たくいうと，「再現性」は，「同じ」であることが一貫しているかどうかの指標，「内的整合性」は「違う」ことが一貫しているかどうかの指標，である．一方，「妥当性」は，「測定したいことを本当に測定しているかどうか」の指標であり，「内容的妥当性(content validity：内容が適切か？ たとえば痛みを見たいときに痛みを見ているか？)」，「構成概念妥当性(construct validity：他の項目との関係が事前に決めた仮説に沿っているか？)」，「基準関連妥当性(criterion validity：既存の尺度との一致/相関：一致し過ぎてもダメ)」，「応答性/感度(responsiveness/sensitivity：状態の変化に応じて値が変化するか)」からなる[1)5)]．

QOL調査票は，こうした信頼性と妥当性が検討されて計量心理学的に“validate”されたものを用いなければならないとされる．こうした“validation”がされていない調査票を用いるべきでないという見解には筆者も賛同するが，逆に，こうして“validate”されたからといって，「治療法の優劣の意思決定に用いること」が妥当であることまで証明されたわけではないことに注意が必要である．つまり，調査票が“validate”されることは「治療法の優劣の意思決定に用いること」についての「必要条件」であるが「十分条件」ではない．「治療法の優劣の意思決定に用いること」が妥当であることの検証がなされたQOL調査票などというものは世の中に存在しない．

ただし，そもそもOSにしろ，担当医評価による毒性にしろ，「治療法の優劣の意思決定に用いること」が妥当であることの検証がなされたエンドポイント自体，世の中には存在しないため，その点ではQOL調査のみが特別劣っているというわけではない．しかし，「治療法の優劣の意思決定に用いるベネフィット/リスク評価」にOSや毒性を使うことは世界共通の“慣習”であり，それで十分か不十分かの議論はあったとしても，これらを用いることそのものについての異論はないだろう．問題とすべきは，QOL評価を治療法の優劣の意思決定に用いることが万人のコンセンサスになっているわけではないにもかかわらず，ことQOL評価においては「調査票が“validate”されていること」が，あたかも「治療法の優劣の意思決定に用いる妥当性」までが検証されているかのごとき，過大な解釈をされる傾向があることである．「この調査票は“validate”されている」と聞くとき，それがなにを意味するのかを踏まえる必要がある．

患者さんの負担(respondent burden)

里見氏も著書で引用している，2008年にJCOに掲載された論文を紹介する[6)]．これは，北アイルランドの3つの病院の共同で行われたランダム化比較試験であり，PS 0～2の切除不能肺がん患者を対象とし，対照群の患者には通常の日常診療が行われ，介入群の患者にはEORTC QLQ-C30の日記形式の調査票が渡され，質問に対する回答を16週間，毎週記録することとされた．両群の“QOL評価”にはFACT-LとFACT-Gが用いられた．結果は，primary endpointであったTrial Outcome Indexでは有意差はなかったが(介入群で低く)，secondary endpointsのいくつかの指標では介入群で有意にQOLスコアが低かったというものであった．「Discussion」では，「繰り返し健康状態について回答を迫られたり，日記の調査票を自宅に持ち帰ることで，繰り返し自身の病気についてくよくよ考えたり，心配が増したりしたのではないか？」と考察されている．もっともな見方である．QOL調査が患者さんに苦痛を与える実例といえる．

もう15年も前のことになるが，かつてJCOGの試験で使われていたQOL調査票の質問項目の中には，患者さんが病気になったことを責めるかのような表現や，明らかにデリカシーに欠けると思われる表現がみられた．裏面に「こんな質問

に答えさせられてつらかった」，「こんな質問になんの意味があるのかわからない」といったコメントが書かれた調査票もあった(当時筆者がJCOGにおけるQOL調査をいったん廃止した背景にはこうした事情もあった)．患者さんと文書で直接コミュニケーションをとるQOL調査やPRO評価にはこうしたリスクもあることを知る必要がある．一方，physician-assessed QOL measureでは，直接患者さんに「善いこと」もしていないかわりに「悪いこと」もしていない．

QOL肯定派の諸氏の中には，HRQOLやPROの調査を行うことが，当然「患者さんに善いことをしている」かのようにいう人も少なくないが，実はかえって患者さんに精神的苦痛を与えている可能性もあることはほとんど語られない．「言語的コミュニケーション」は「両刃の剣」である．すべてのQOL調査において「respondent burden」が慎重に吟味され，最小化されなければならない．

欠測値の問題

QOL評価における「欠測値(missing data)」の問題は広く知られており，DeVitaにも「analytic consideration」として取り上げられている[1]．つまり，QOL評価における欠測値はランダムに発生するのではなく，患者さんの状態が悪くなったがゆえに生じた欠測であることが多く，欠測値を解析から単純に除外してしまうことでQOLの値を過大評価してしまうことになるという問題である．原病の増悪や死亡，治療の毒性のために調査票の記載ができない健康状態になってしまうことにより欠測が生じる．完全にランダムに生じた欠測(missing completely at random)なら解析から除外してもバイアスは生じず，そうした欠測は「uninformative missing」と呼ばれるが，QOL評価における欠測の多くはそうではないため「informative missing」と呼ばれ，バイアスを小さくするための解析上の工夫がなされなければならない．具体的には，欠測値に，最悪値(ゼロ点)を当てはめたり，欠測の前の直近の値を当てはめたり(last observation carry forward approach)，回帰モデルによる予測値を当てはめたりする方法(imputation approach)などがあるが，いずれも「真の値」を用いるわけではないため限界があり，どの処理が最適かに関するコンセンサスはない．DeVitaでは「調査期間を長く取り過ぎない」という，解析以前のデータ取得レベルでの現実的な対応を推奨している．

PROに関するFDAガイダンス

2009年に米国FDA(Food and Drug Administration)が公表したPROに関するガイダンス「Guidance for Industry-Patient-Reported Outcome Measures：Use in Medical Product Development to Support Labeling Claims」[7]が，あたかも「FDAがPROを公認した」証拠かのように言われることがあるが，それは事実に反する．

このガイダンスでFDAは，「QOLやPROの改善を添付文書で謳う医薬品の効能効果(labeling claim)に含めるのであればこのガイダンスに従う必要がある」と述べているにすぎず，「医薬品の薬事承認の審査に(HRQOLを含む)PROの結果を用いるべき」や，「PROを推奨する」といったことは一言も書かれていない．むしろ，マスキングしていない試験におけるPROに対する否定的見解が明記されており(open-label clinical trials are rarely adequate to support labeling claims based on PRO instruments)，このガイダンスによれば，FDAはむしろ「マスキングできないがんの臨床試験におけるPRO」に対しては否定的と考えるのが正しい理解であろう．このことは，DeVitaにも「the U.S. Food and Drug Administration (FDA) is reluctant to consider claims about HRQOL improvement in approval decisions」と言及されている[1]．

「日本語」の問題

先述の「respondent burden」に絡めて「日本語」の問題に触れる．HRQOLやPROの質問文の日本語に不自然さを感じたことのある人は少なくないであろう．これには一応理由がある．

QOL調査票のオリジナル版は多くが英語であり，そのため日本を含め，英語を母国語としない国におけるQOL調査では，当然のことながら使用国の言語に翻訳したQOL調査票が用いられる．しかしオリジナルの英語版作成者にとって

は，せっかく「validate」された調査票を勝手に不適切に翻訳して使われては困るので，QOL調査の質問内容は使用国の言語に順翻訳(translation)されたのち，それを英語に逆翻訳(back translation)し，逆翻訳した英語とオリジナルの質問文を対比して「同じことを尋ねている」ことを確認する，という作業が通常要求される．QOL調査票の質問の日本語の不自然さは，(少なくとも幾分かは)この「translation－back translation」のプロセスに起因する．しかし，「不自然」なだけでは問題とするには当たらないだろうが，なかには「不自然」では済まず，日本語に敏感な患者さんだったら「人を(小)馬鹿にしている」と感じるのでは？ と思われる質問文をみることが少なくない(少なくとも自称「日本語にうるさい」筆者はしばしばそう感じる)．「人を小馬鹿にした」かのような日本語の質問票は，QOL調査に協力してくださる患者さんの精神的苦痛(mental respondent burden)になりうる(イヤな感じを抱きつつ質問に答えさせられることが苦痛であることに異を唱える人はいまい)．

そもそも，聖書を持つキリスト教文化の英語圏と，聖書に相当する文書を持たない日本文化では「文書」が持つ意味自体が違うし，「書かれたものがすべて」の米国の契約社会文化と，「行間を読む(それを言っちゃぁおしめぇよ)」，「縁起の悪いことは言わない/書かない(言霊思想)」，「甲乙両者が誠意をもって対応(和をもって尊しとなす)」の日本文化とでは「言語的コミュニケーション」の意味やあり方も異なる．自分の意見を明言することに関する態度も違うではないか(自分の意見を言わないと叱られる米国 vs. 自己主張し過ぎると嫌われる日本，first nameで呼び合う米国 vs. 本名は呼ばず職名で呼ぶ日本)．さらに，主語の扱いがまったく異なるという文法の違い[8]等々の彼我の差を考えれば，オリジナルの英語のQOL調査票の質問を，back translationを経てもなおオリジナルの英語と同義とみなされるように，かつ日本語として自然で適切な(失礼でない)質問に翻訳する，という作業には，卓越した日本語能力と語学センスが必要であろう．

難しいことはわかるし，「back translationを前提とするため不自然な日本語になるのはやむをえない」という言い訳を全否定するつもりはないが，QOL調査票やPROが「言語的コミュニケーションツール」である以上，少なくとも，調査に協力してくださる患者さんが読んで不愉快に思う「小馬鹿にしたような表現」になっているか否かを見極める程度の日本語能力は必要であろう．必要な語学力のレベルと文化の違いを考えると，この「translation－back translation」という方法自体にそもそも無理があるのではないかと筆者は考える．オリジナルの英語とback translationした英語を比べるのでなく，たとえば，オリジナルの英語と翻訳した日本語を(英語日本語双方に精通した優秀な翻訳家などの)第3者がチェックするといったような“validation”を英語圏のQOL研究者に飲ませられないものだろうか？ もともときわめて“ソフト”な計量心理学である．ここだけ厳密にしても仕方なかろう．ちなみに，“I love you”を，夏目漱石は「月がキレイですね」と訳し，二葉亭四迷は「わたし，死んでもいいわ」と訳したらしい．「translation－back translation」の枠組みでは漱石も四迷も確実に“失格”である．

おわりに

エンドポイント(endpoint)は「患者さんのベネフィットを測るものさし(criterion by which patient benefit is measured)：by Richard Simon」であり，「ハード」なエンドポイントと「ソフト」なエンドポイントがある．「ハード」とは「誰が見ても同じ・何回見ても同じ」ことを意味し，「ソフト」とは「見る者によって評価が異なる・時々によって違う・他の影響を受けやすい」ことを意味する．HRQOLやPROが「ソフト」なエンドポイントであることは，QOL調査否定派かQOL調査肯定派かに依らず共通認識であろう．そして，研究者個人や研究組織として，「ハード」なエンドポイントを優先する立場と「ソフト」なエンドポイントを優先する立場がありえ，前者は「conservative」なスタンス，後者は「liberal」なスタンスともいえる．

「conservative」なスタンスとは，「本当はよくないものを誤ってよいと判断する偽陽性の誤りを避ける」ことを優先する立場/嗜好/方法であり，確率的なαエラーの最小化を優先すること

に対応する．「liberal」なスタンスとは，「本当はよいものを誤ってよくないと判断する偽陰性の誤りを避ける」ことを優先する立場/嗜好/方法であり，確率的なβエラーの最小化を優先することに対応する．偽陽性の誤りと偽陰性の誤りの両方とも最小化できればよいのだが，個々の試験のデザインや結果の解釈等においては，この「conservative」と「liberal」はトレードオフの関係にあり，残念ながら「非常にconservativeであり，かつ非常にliberalである」ことはできない．

QOL調査肯定派か否定派かの違いは，ソフトなエンドポイントを重視または許容する「liberal」なスタンスか，ハードなエンドポイントを重視する「conservative」なスタンスかの違いであるともいえる．世の中の方法論がすべて「conservative」であれば，本当はよくない治療を誤ってよいとしてしまうリスクは小さくなるが，本当によい新しい治療が日の目をみないことによって医学の進歩は遅くなるかもしれない．逆に，世の中がすべて「liberal」ならば，医学の進歩の効率はよいかもしれないが，本当はよくない(場合によっては害のある)治療を受けるという国民の不幸やリスクは増える．つまり，これはどちらが正しくてどちらが間違っているという問題ではないのである．もちろん「善悪」の問題でもない．

筆者は，「QOL調査は善である」といった主張には反論するが，だからといって「QOL調査は悪だからやるべきでない」というつもりもない．要は肯定・否定は立場/嗜好/方法の違い，平たくいえば「好み」の違いであるということである．そして，好みの問題なのだから，QOL調査はそれを好きな人が，賛同してくれる人を語らって(患者さんにも迷惑をかけないように)やればよいと考える．その際，「やるべきものだから」や「de facto standardだから」や「FDAも認めているから」などの(おそらくは正しくない)根拠を基に，それがキライな人たちにまで押しつけないで欲しいと思っているだけである．

筆者らはQOL“調査”がキライなだけであって，「患者さんのQOLが大事」という信条はQOL肯定派の諸氏と変わりはないと思っている．筆者らは筆者らが正しいと思うやり方で「患者さんのQOLをよくする治療」の治療開発・臨床試験に携わっており，個々の臨床試験において，どうやったら患者さんのリスクをもっと最小化できるのか・・等を日々考えている．QOL調査よりも優先すべき未解決の課題が「がんの臨床試験」にはまだまだあると筆者らは思っているだけである．

最後の最後にお願いで稿を終えたい．どうかQOL調査否定派の私たちのことは放っておいていただきたい．そして「QOL調査」に非協力的な私たちのことを悪人呼ばわりしないで欲しい(そういう「空気」も醸成しようとしないで欲しい)．QOL調査をやるなら，QOL調査が好きな人たちだけでやって欲しい．そうすれば私たちも，他人が好きでやっていることにケチをつけるようなオトナげないことはせずにすむのだから・・・．

補遺：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) Earle CC, Schrag D. Health Services Research and Economics of Cancer Care. In : DeVita VT, Lawrence TS, Rosenberg SA, editors. Cancer- Principles & Practice of Oncology. 9th edition. Philadelphia : Lippincott Williams & Wilkins ; 2011. p. 352.
2) Piantadosi. CLINICAL TRIALS- A Methodologic Perspective. Second Edition. Hoboken : John Wiley & Sons, Inc. ; 2005. p. 207.
3) Green S, Smith A, Benedetti J, Crowley J. Clinical Trials in Oncology. Third Edition. Boca Raton : CRC Press ; 2012. p. 45.
4) Walters SJ. Quality of Life Outcomes in Clinical Trials and Health-Care Evaluation. Chichester : John Wiley & Sons, Ltd. ; 2009. p. 2.
5) Walters SJ. Quality of Life Outcomes in Clinical Trials and Health-Care Evaluation. Chichester : John Wiley & Sons, Ltd. ; 2009. pp. 31-49.
6) Mills ME, Murray LJ, Johnston BT, et al. Does a patient-held quality-of-life diary benefit patients with inoperable lung cancer?. J Clin Oncol 2008 ; 27 : 70.
7) U.S. Department of Health and Human Services-Food and Drug Administration. Guidance for In-

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

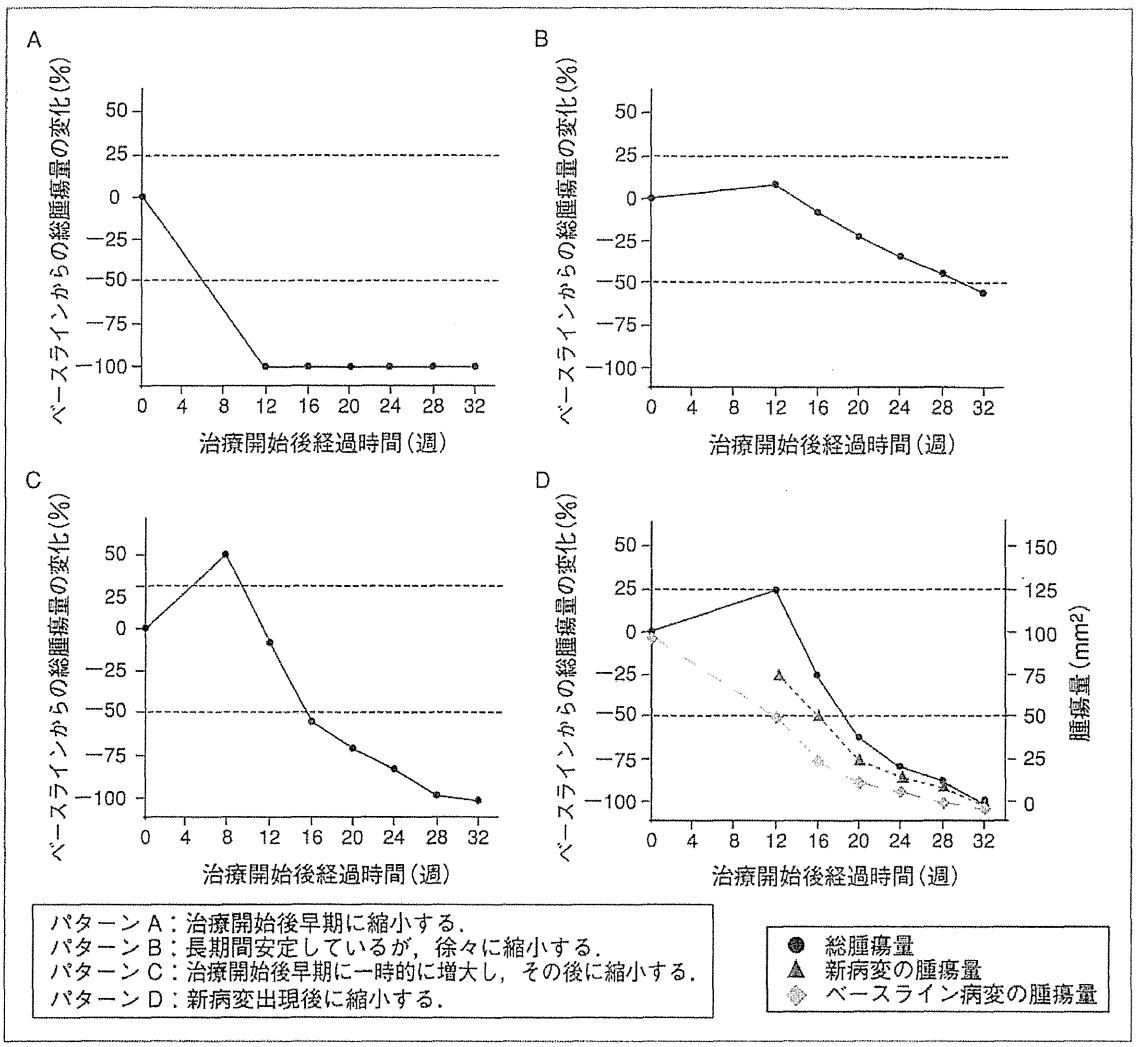

図1　4つの腫瘍縮小パターン

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

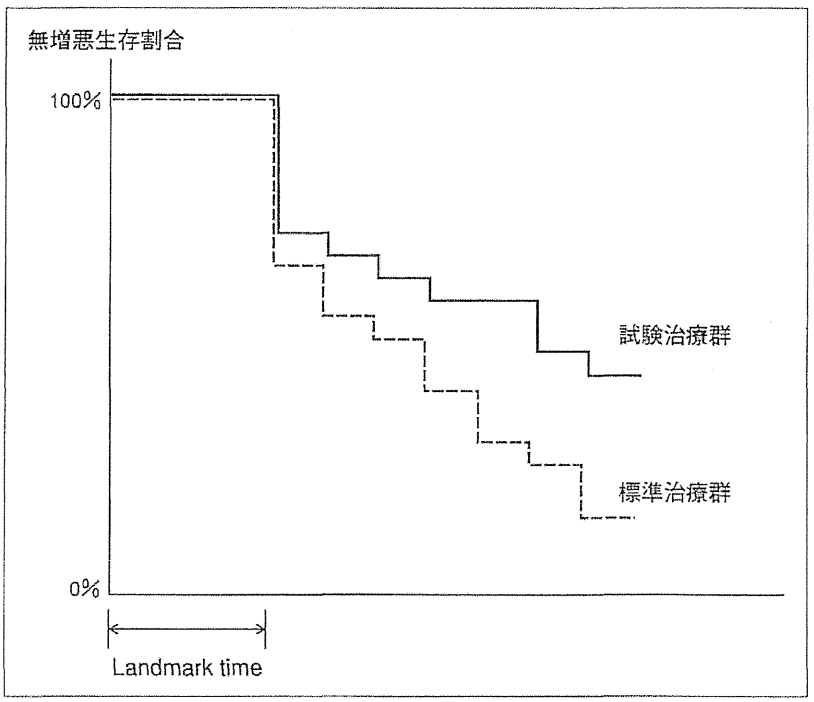

図2　Landmark methodの生存曲線

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

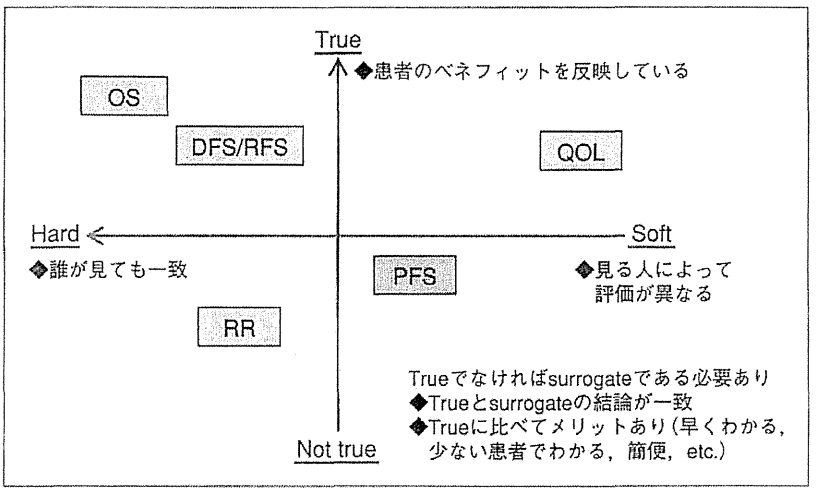

図1　True軸とHard-soft軸

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

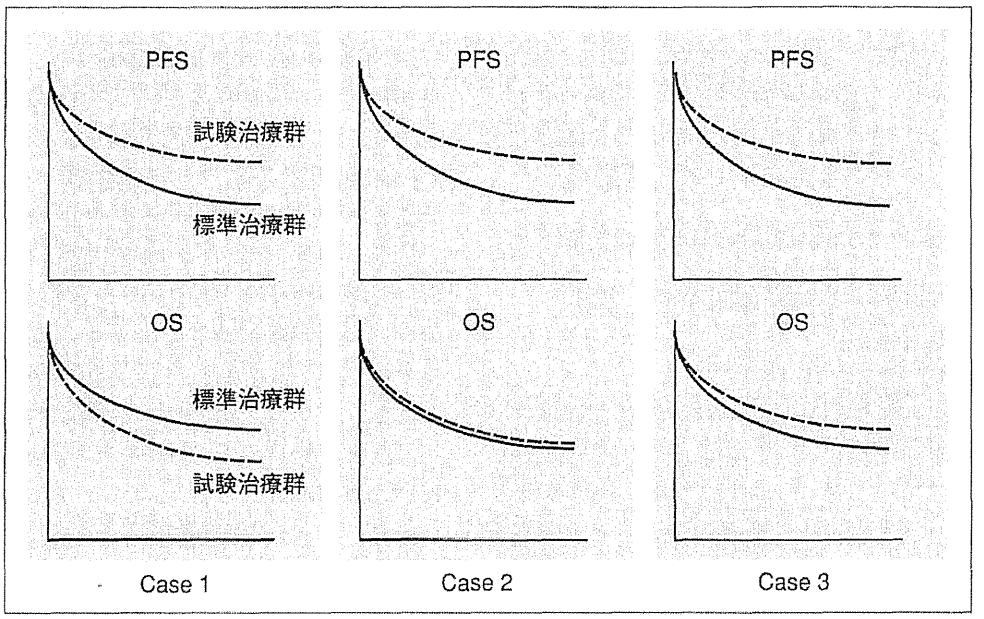

図2　どの状況で試験治療を取るか？

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