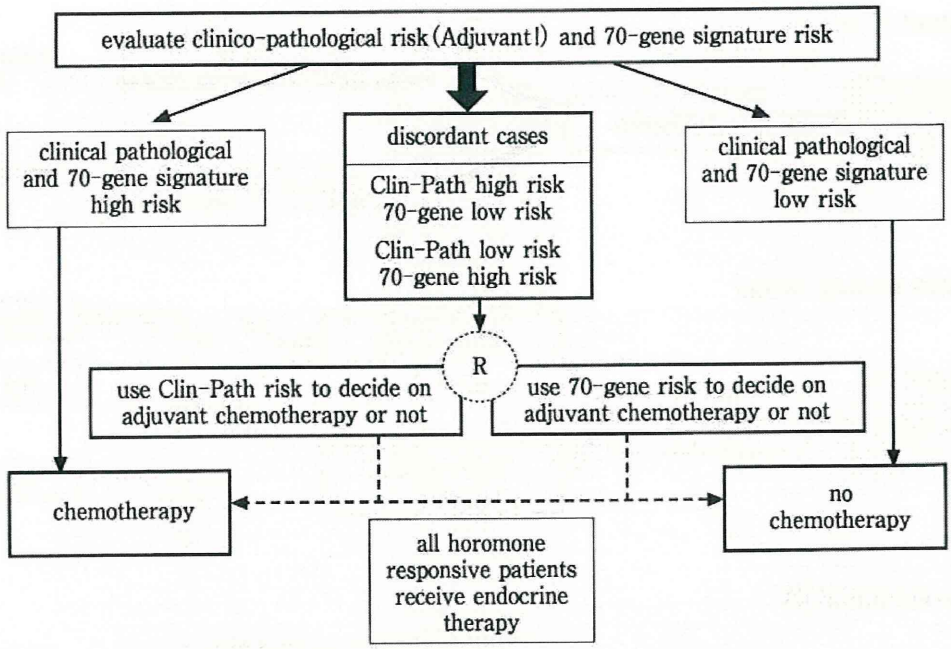


図4 Marker(+) designの例2(EORTC MINDACT Trial)
[Reprinted from Buyse M, et al: Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. J Natl Cancer Inst 98(17): 1183-1192, 2006 by permission of Oxford University Press]

って，まず手術材料を用いて予後を予測し，低リスクと予測された群は経過観察とし，高リスクと予測された群に対しては化学療法と経過観察のランダム化比較試験を行うとするものである．

また，EORTC MINDACT(Microarray In Node (−) Disease may Avoid ChemoTherapy) Trialでは，腋窩リンパ節転移陰性の早期乳癌患者に対し，Adjuvant!というソフトウェアによって推定された臨床病理学的な再発リスクと70の遺伝子発現による再発リスクのスコアを求め，これらに基づいた治療選択の評価を行っている[3]．すなわち，臨床病理学的な方法ならびに遺伝子発現による方法の両方で高リスクと推定されたものには化学療法を行い，両方で低リスクと推定されたものには化学療法を行わず，臨床病理学的なリスクと遺伝子発現によるリスクによる予測が不一致だった場合には，その対象者に対し，化学療法ありなしを比較するランダム化比較試験を行うものである(図4)．

これらの例でもわかることは，marker(+) designとは，①効果があると考えられる対象を選択してその対象に臨床試験を行う(enrichment)，といったものだけでなく，②予後が悪く，更に治療開発が必要な集団を同定し，その集団に追加治療を行うかどうか調べる，③予後が良く，overtreatmentの可能性のある集団を同定し，治療を手控えられるかどうか調べる，④予後が予測できず，治療が必要かどうか不明な集団に必要な治療を調べる，といった場合にも適用可能な臨床試験デザインであると考えることができる．

b. Marker strategy design

二つめのデザインは'marker strategy'といえるもので，まず始めにマーカーを測定し，ランダム化後，一方の群はマーカー結果に基づいて治療を決定し(marker-based)，もう一方の群はマーカーとは関係なく標準治療を行い(not marker-based)，これを比較するというものである．プライマリな解析は，marker-based or notのストラテジーを比較する，つまり試験に参加した対象者全体を比較する．このデザインは，マーカーによって規定されるそれぞれのサブグループに対して異なる治療を行うという

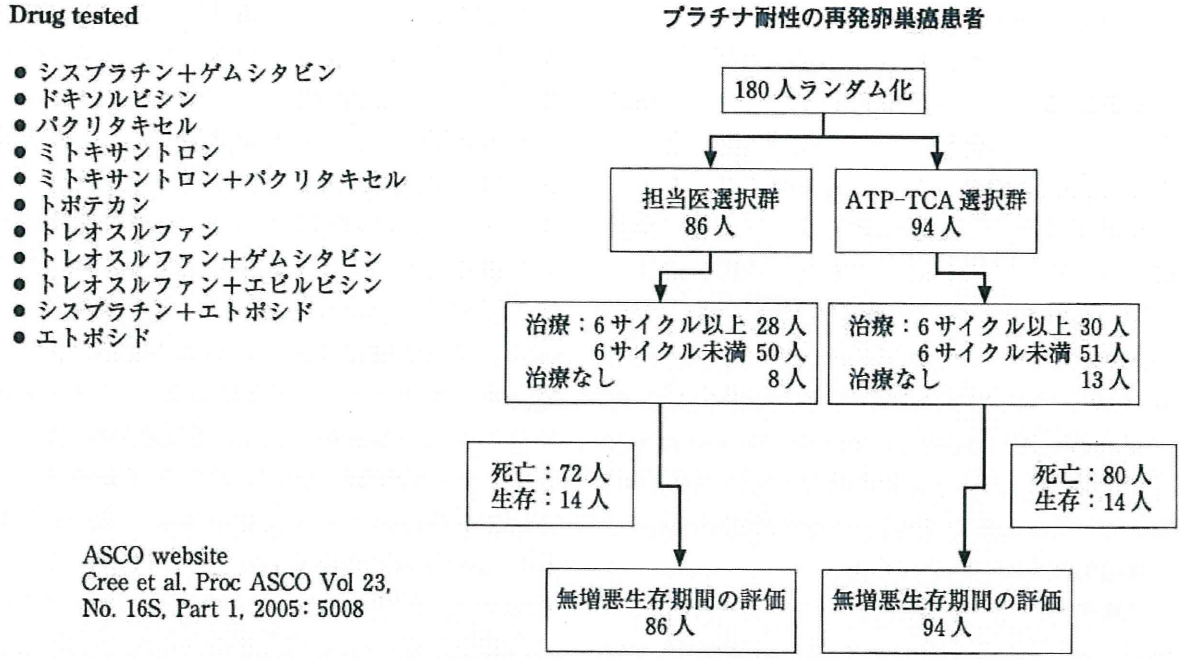


図5 Marker strategy designの例

'個別化治療ストラテジーそのもの'を評価できるデザインと考えられる．ただ，効果のある集団を同定する＝予測因子マーカーを同定する，あるいはproof of principleを調べるための臨床試験デザインではない．なぜなら，マーカー(−)のサブグループには，marker-basedの群でもそうでない群でも同じ治療が行われているため，プライマリな解析で有意に差があった場合でも，マーカー(−)に対する治療効果がわからず，マーカー(+)の群のみで治療効果があるのかどうか(交互作用があるかどうか)調べることが不可能だからである．プライマリな解析(対象者全体の解析)で有意な差がみられなかった場合には，マーカー(+)サブグループでの治療効果も不明である．なぜなら，マーカー(+)サブグループで差があり，マーカー(−)で差がなかった場合，マーカー(+)サブグループの割合が多ければ全体として有意差あり，少なければ有意差なしとなる可能性があり，結果が試験参加者の中のマーカー(+)の割合に依存してしまうからである．

Kurbacherらは，再発卵巣癌患者に対し，ATP-TCA(tumor chemosensitivity assays)の結果に基づいて化学療法を選択したパイロット研究の結果と経験的に化学療法を選択しているヒストリカルなデータを比較し，ATP-TCAに基づいて化学療法を選択した方がヒストリカルなデータに比べ，奏効割合や無増悪生存期間において良好な成績を示すことを報告した[4]．Creeらは，この結果をもとに，プラチナ耐性の再発卵巣癌患者に対し，13種類の化学療法の中からATP-TCAに基づいて化学療法を選択するというストラテジーと担当医が化学療法を選択するというストラテジーを比較するランダム化比較試験を行った[5]．結果として，180人をランダム化して86人が担当医選択群，94人がATP-TCA選択群に分けられ，無増悪生存期間ならびに全生存期間を比較したところ，大きな差はみられないという結果を得た(図5)．この例では，対照群として効果が検証された標準治療がない場合でも，担当医や患者により決定された日常診療に基づく治療に対して，一定のストラテジーによって治療方法を決定する治療ストラテジーどうしを比較することができ，魅力的である．ただし，結果は対照群の治療結果に依存するため，治療方法選択に大きな違いがなければ大きな差が生じず，治療ストラテジーの効果の差を検出することが難しいことが予想される．

c. All comers design

三つめは'all comers'と呼ばれるデザインで，まず始めにマーカーを測定するが，その結果によらずランダム化し，一方には新治療，もう一方には標準治療を行うというデザインである．プライマリな解析としてマーカーによらず全対象者の群間比較を行い，セカンダリな解析としてマーカー(+)/(−)それぞれで治療効果の群間比較を行う．このデザインを用いることができれば，マーカー(+)/(−)を合わせた集団に対する治療効果に加え，マーカーサブグループ間で治療効果が異なるかどうかも調べられるので，マーカーが治療効果の予測因子であるかどうかを調べることができる．

このデザインは，最初にマーカーを測定しないで，試験終了後(ランダム化後)にマーカーを測定するのと同じではないかと考えられるかもしれない．Fisherらはリンパ節転移陰性，ER(+)の乳癌女性に対し，化学療法の効果を調べる臨床試験を行い，2,306人の乳癌患者をランダムにタモキシフェン群，タモキシフェン+化学療法(MFT)群，タモキシフェン+化学療法(CMFT)群の3群に分け，化学療法群の方が予後が良いことを示した(NSABP B20)[6]．

Paikらは，この試験に参加した対象者の組織と臨床試験の結果を用いて，遺伝子発現に基づくOncotype DXによる再発スコア(RS)の有効性を調べた[7]．NSABP B20の対象者のうち，651人(227人がタモキシフェン群であり，424人が化学療法群)についてOncotype DXによる再発スコアが調べられた．再発スコアにより高リスク，中リスク，低リスクと分けた場合，高リスク群のみ化学療法群とタモキシフェン群に予後の差が認められた(**図6**)．この結果により，化学療法の追加により恩恵を受けるのは高リスク群のみである，言い換えると高リスク群のみ化学療法を追加すべきであると主張している．

この研究はまさに事後的な解析によって，all comers design的な臨床試験を仮想的に行った例と考えられる．もし，事前に再発スコアを測定し，測定した対象者に対しランダム化比較試験を行い，その結果において3つのリスクカテゴリで予後を比較し，高リスク群でのみ化学療法による予後が優っていることが示されれば，リスクスコアと治療法の間に交互作用があり，高リスク群では化学療法を追加することが推奨されると結論づけることができるであろう．しかしながら，この例のように事後的に測定を行った場合には，ランダム化した全例に再発スコアを測定できなかったためにリスクスコアごとの治療効果比較はランダムな比較とはいえず，差があったとしてもRCTによるエビデンスということはできない．また，事後にマーカーを測定した場合には，群ごとのサンプルサイズ設計も十分行われているとはいえず，差がない場合にも検出力不足である可能性がある．ただ，それでも，観察研究のデータより，ランダム化比較試験に基づいた事後的な解析の方が結果の妥当性は高いと考えられるのは明らかであろう．

事後的に行った場合のもう一つの問題点として，事前に特定しないマーカーをいろいろと測定し，それぞれ群間比較してしまうと，統計的多重性，つまりαエラーの調整ができないという点がある．αエラーの調整ができないと，結果を探索的にしか解釈できなくなり，治療効果の差がみられたマーカー・サブグループに対してもう一度検証的な臨床試験を行わなければならなくなってしまう．しかしながら，試験前にマーカーを特定していれば，一度の臨床試験で全体比較とマーカー・サブグループ内での比較の両方を検証的に行う試験デザインを立てることができる．もし，試験前にマーカーを特定できない場合でも，'全体で比較→差がない場合にも臨床試験のデータを用いてマーカーを特定→更にその対象に対してサブグループ比較'を一つの臨床試験の中で検証的に行うようにデザインすることも可能である[8,9]．

all comers design実施上の問題として，治療効果が期待できるようなマーカーサブグループ(マーカー(+)とする)が特定されている場合に，そうでないサブグループ(マーカー(−))も含めてランダム化臨床試験を実施するのが困難であると考えるかもしれない．しかし，マーカー(−)のグループに治療効果がないことが証明さ

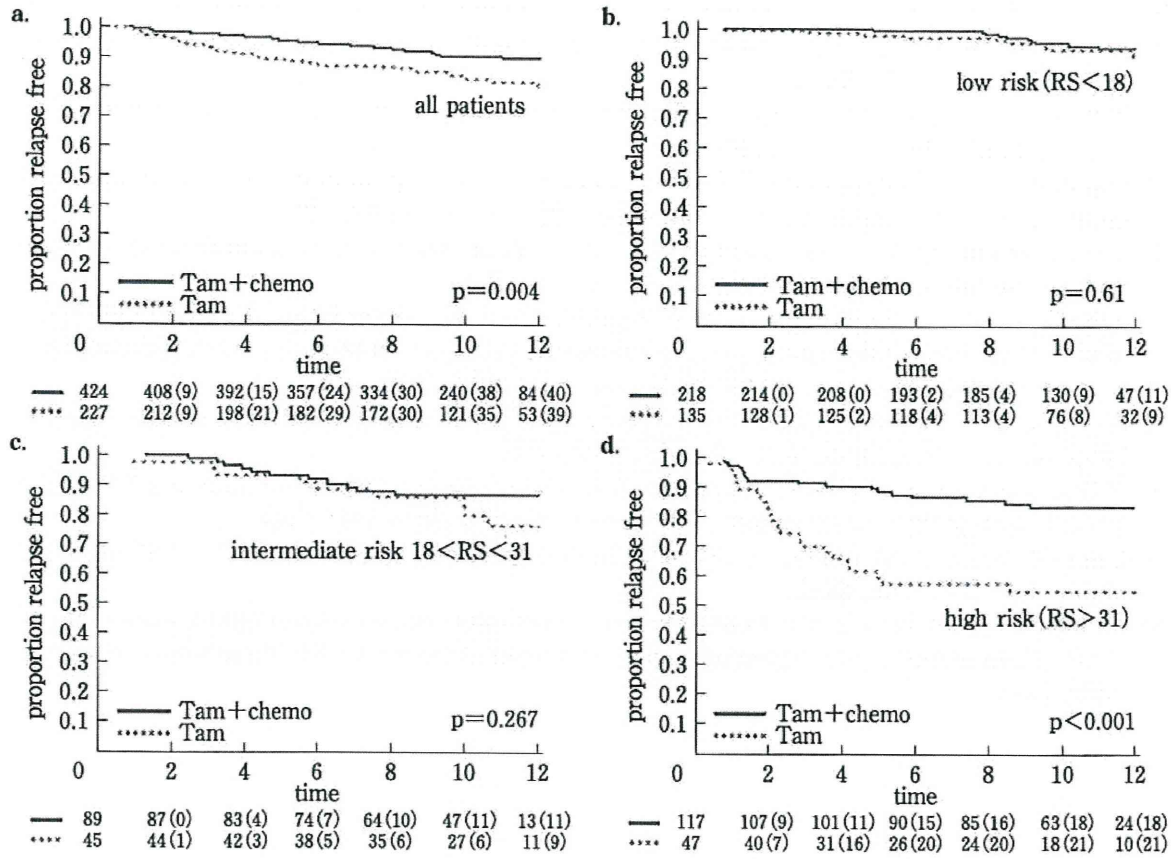


図6 事後的 all comers design の例

(NSABP B20対象者の Oncotype DX による再発リスクスコア別治療成績)

数字は at risk の人数, ( )内の数字は再発数.

[Reprinted from Paik S, et al: Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. J Clin Oncol 24(23): 3726-3734, 2006. with permission. © 2009 American Society of Clinical Oncology. All rights reserved]

れたわけでなければ, マーカー(−)の対象者を含めたランダム化比較試験を行うことは十分 rationale があると考えられる.

おわりに

分子標的薬をはじめとして, がんの個別化治療開発は全世界的に精力的に取り組まれており, いろいろな考え方, いろいろなデザインで研究が行われている. 治療効果を調べたいサブグループを同定し, そのサブグループでの治療効果をランダム化比較試験で検証するというのが基本戦略となると考えられるが, それ以外のサブグループでの治療効果を確認することも重要である. いずれにしろ, 治療効果の検証にはランダム化比較試験が必須であるので, 予測因子マーカーの探索とその検証を効率よく行うために, 質の高い付随研究が実施できるような体制を構築することが重要となると考えられる.

■文　　献

1) Sargent DJ, et al: Clinical trial designs for predictive marker validation in cancer treatment trials. J Clin Oncol **23**: 2020–2027, 2005.
2) Potti A, et al: A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer. N Engl J Med **355**: 570–580, 2006.
3) Buyse M, et al: Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. J Natl Cancer Inst **98**(17): 1183–1192, 2006.
4) Kurbacher CM, et al: Use of an ex vivo ATP luminescence assay to direct chemotherapy for recurrent ovarian cancer. Anticancer Drugs **9**(1): 51–57, 1998.
5) Cree IA, et al: A prospective randomized controlled trial of ATP-based tumor chemosensitivity assay (ATP-TCA) directed chemotherapy versus physician's choice in patients with recurrent platinum-resistant ovarian cancer. Proc ASCO Vol 23, No. 16S, Part I, 5008, 2005.
6) Fisher B, et al: Tamoxifen and chemotherapy for lymph node-negative, estrogen receptor-positive breast cancer. J Natl Cancer Inst **89**: 1673–1682, 1997.
7) Paik S, et al: Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. J Clin Oncol **24**(23): 3726–3734, 2006.
8) Simon R: Roadmap for developing and validating therapeutically relevant genomic classifiers. J Clin Oncol **23**: 7332–7341, 2005.
9) Freidlin B, Simon R: Adaptive signature design: An adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. Clin Cancer Res **11**: 7872–7878, 2005.

腫瘍内科，4(1)：40-47，2009

特集 がんの個別化治療を目指したバイオマーカー

バイオマーカーを用いた臨床試験計画*

山本精一郎**

Key Words : biomarker, clinical trial, study design

治療開発研究計画における バイオマーカーの役割

バイオマーカーを測定するということは，それがなんらかの意味で薬剤，もしくは治療効果と関係がある(と期待する)からであろう．分子標的薬であれば，ターゲットで薬が作用していることを測定するためや，マーカーで規定されるサブグループによって治療効果や予後を予測することがバイオマーカーに期待される2つの大きな役割と考えられる．これらは治療開発の段階では異なる役割を担う．順に考えてみたい．

Phase 1の目的は，用量の設定と毒性の評価である．細胞障害性薬剤の場合，毒性と効果が比例しているという性質を利用し，毒性が許容される中で最大の投与量がもっとも効果があることを利用して用量を設定する．つまり，用量制限毒性(DLT)を用いて最大耐用量(MTD)を探索するが，3例コホートがその典型的なデザインである．これに対し，毒性と効果が比例しない可能性のある分子標的薬の場合，phase 1ではそもそも薬剤が作用しているかどうか調べるために，ターゲットと考えられる部位での作用がみられるようなマーカーを測定することが必要となる．しかし，現実にはいいマーカーがなかったり，作用部位がはっきり特定されなかったりして，マーカーの動きをもって効果を測ることは必ずしも容易ではない．結果として，現実には伝統的な3例コホートデザインを用いることが多いようである．

Phase 2の目的は，有効性によるスクリーニングである．分子標的薬剤では，必ずしも抗腫瘍効果として腫瘍縮小がなくても生存期間の延長が期待できる可能性がある．したがって，腫瘍縮小以外の有効性のマーカーがあればそれを有効性のサロゲートエンドポイントとすることもできるが，サロゲートエンドポイントであるためには有効性の真のエンドポイントである生存期間と相関が高い挙動を示すマーカーである必要があり，そのようなバイオマーカーを薬剤ごと，臓器ごとにphase 2の段階までに開発するのは容易ではない．また，実際には分子標的薬でもある程度の腫瘍縮小効果が期待できるものも多いことから，腫瘍縮小効果をエンドポイントにしたり，生存時間そのものをエンドポイントにしたりすることが一般的である．後者の場合，研究にエントリーした患者背景によって生存期間自体が大きく変わる可能性があるため，標準治療を対象としたランダム化phase 2を行うなど，適切なコントロールの値と比較することが重要である．

Phase 3の目的は真のエンドポイントにおいて

* Clinical trial designs using biomarker.

** Seiichiro YAMAMOTO, Ph.D.: 国立がんセンターがん対策情報センターがん情報・統計部〔〒104-0045 東京都中央区築地5-1-1〕; Cancer Information Services and Surveillance Division, Center for Cancer Control and Information Services, National Cancer Center, Tokyo 104-0045, JAPAN

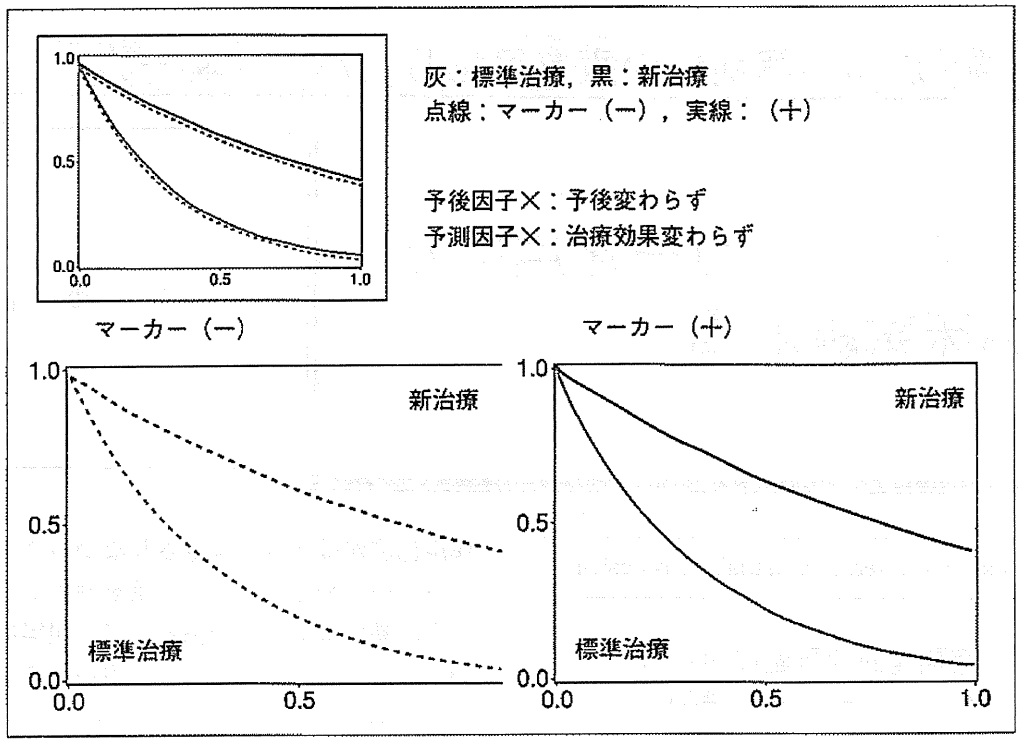


図1　マーカーが予後因子でも予測因子でもない場合

標準治療より優れていることを証明することである．多くの場合，新治療による生存期間が標準治療より長いか，もしくは，生存期間においては新治療と標準治療で同等でも，副作用の軽減などその他の大きなメリットが新治療によりもたらされるかどうか，ということになる．すなわち，ターゲットにおいて薬剤が作用しているかという傍証よりも，実際その薬剤で生存期間の延長が認められるか，ということが目的となり，ターゲットでの作用を調べるようなバイオマーカーをエンドポイントとした臨床試験が組まれることは想像しにくい．このコンテクストでのバイオマーカーとは，以下に述べるようなサブグループを特定するようなマーカーのことである．非臨床の研究段階，もしくはphase 1，phase 2での検討で，薬剤がよく作用するようなサブグループを同定することができれば，その集団を対象に治療開発を行うことが効率的と言えるであろう．

予後因子と予測因子を詳しく考える

ここで，あるサブグループをバイオマーカーで特徴づけることを考える．いわゆる，予後因子マーカーと予測因子マーカーについて詳しく考えてみよう．

予後因子とは，無治療の場合，もしくは標準治療を行った場合に，マーカーで規定されたサブグループ[簡便のためにマーカー(＋)とマーカー(−)と呼ぶが，3群以上となってもよい]によって予後が異なるものを言う．しかし，現実には無治療群のデータはあまりないことから標準治療を行った場合の予後を規定する因子のことを言う．これは必然的に標準治療が変わると予後因子も変わる可能性があることを意味している．

予測因子とは，マーカーで規定されたサブグループによって治療効果が異なる因子を言う．たとえば，標準治療ではマーカー(＋)と(−)で予後に差がないのに，新治療ではマーカー(＋)とマーカー(−)の間で予後に差がある場合，予後の差を生じる要因はマーカーと治療であることから，この因子のことを治療効果の予測因子といい，統計的にはこの状態をマーカーと治療効果に交互作用あり，と言う．これらを生存曲線で表したものが図1〜4である．それぞれマーカー(＋)(−)のサブグループに，標準治療を行った場合，新治療を行った場合の生存曲線を示し

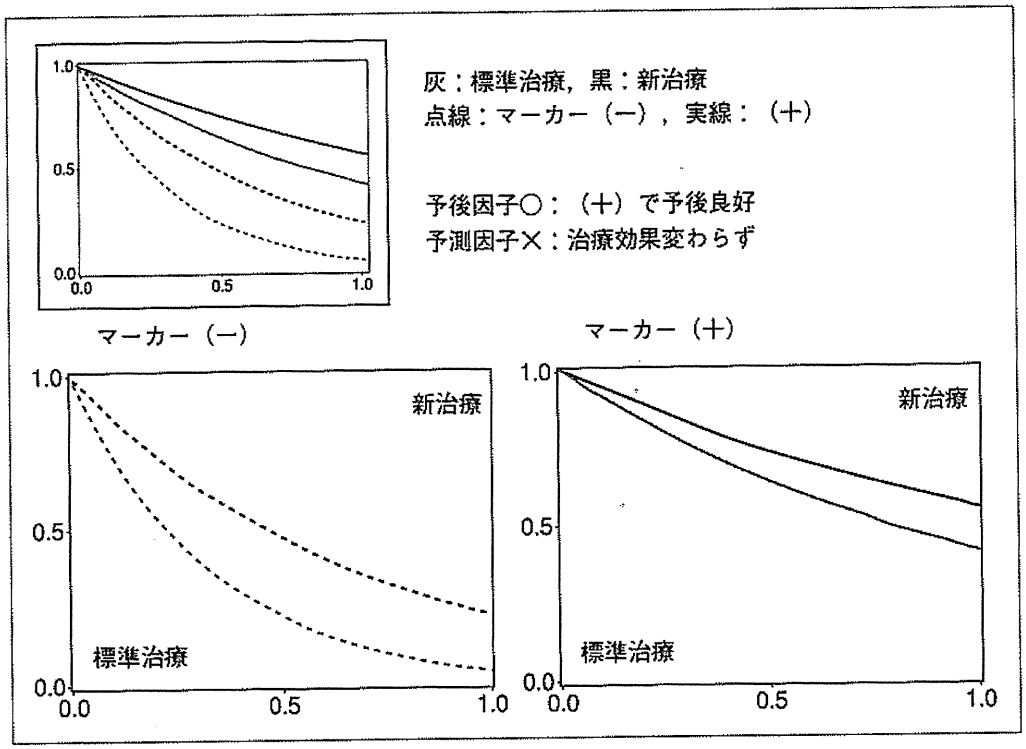


図2　マーカーが予後因子であるが予測因子でない場合

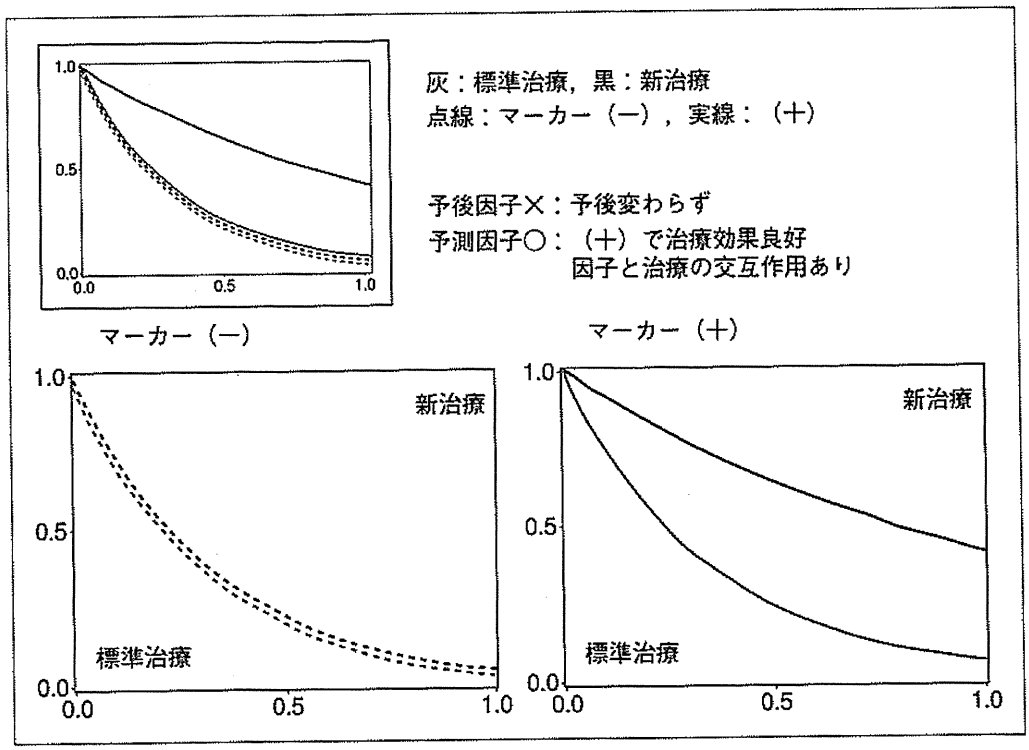


図3　マーカーが予後因子ではないが予測因子である場合

ている．

図1は標準治療群でマーカー(＋)とマーカー(−)の間で予後の差がないことから，マーカーは予後因子ではなく，マーカー(＋)とマーカー(−)の間で治療効果に差がないことからマーカーは予測因子でもない．図2は標準治療群でマーカー(＋)とマーカー(−)の間で予後に差があるためマーカーは予後因子であるが，マーカー(＋)

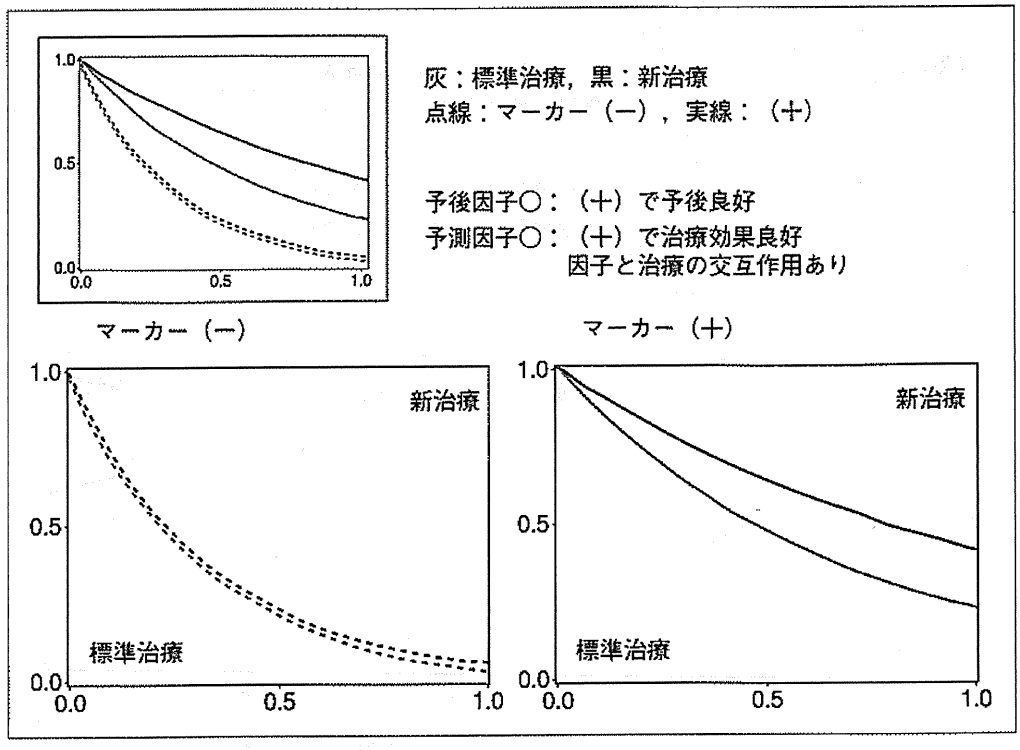


図4　マーカーが予後因子かつ予測因子である場合

とマーカー(−)の間で治療効果に差がないことからマーカーは予測因子ではない．図3は標準治療群でマーカー(＋)とマーカー(−)の間で予後に差がないためマーカーは予後因子ではないが，マーカー(＋)とマーカー(−)の間で治療効果に差があることからマーカーは予測因子である．図4は標準治療群でマーカー(＋)とマーカー(−)の間で予後に差があるためマーカーは予後因子であり，マーカー(＋)とマーカー(−)の間で治療効果に差があることからマーカーは予測因子である．これを別の形で表現したものが図5である．縦軸に5年生存率をとり，新治療と無治療で比較している．無治療を標準治療と考えて，図1〜4がどれに対応するか考えてみてほしい．図1はb，図2はe，図3はc，図4はfにそれぞれ対応するのだが，a，bは治療効果がない場合である．この場合でも予後因子は定義できることに注意してほしい．また，マーカーが予測因子である場合で，マーカー(＋)で標準治療より新治療で大きな治療効果がある場合でも，マーカー(−)でも新治療の方が少し治療効果がある場合，まったくない場合，逆に標準治療の方が効果がある場合に分けられる．最初の場合を量的交互作用，あとの2つの場合を質的交互作用と呼ぶ．量的交互作用の場合は，マーカー(−)の群でも治療効果が期待されるため，新治療を導入すべきということになる．質的交互作用ありの場合には，治療効果に差がなければマーカー(−)の群に新治療を導入するかどうかはほかのエンドポイント(副作用など)の成績によるが，標準治療が上回っている場合にはもちろん新治療を導入すべきでないということになる(図6)．

また，治療効果の差で予測因子を定義する場合，あまり指摘されていないが，効果の指標によって予測因子かどうかが変わってしまうという問題がある．たとえば，あるマーカーが予後因子であるとする．マーカー(＋)のサブグループにおいて標準治療群での5年生存率が60%だったとして，新治療による5%の治療効果の上乗せがあったとするとこれはハザード比に換算すると0.8となる(生存期間が指数分布に従うことを仮定)．ハザード比は生存期間中央値(MST)の比の逆数と覚えておけばよい．これに対し，マーカー(−)のサブグループにおいて標準治療群での5年生存率が40%の場合，新治療による5%

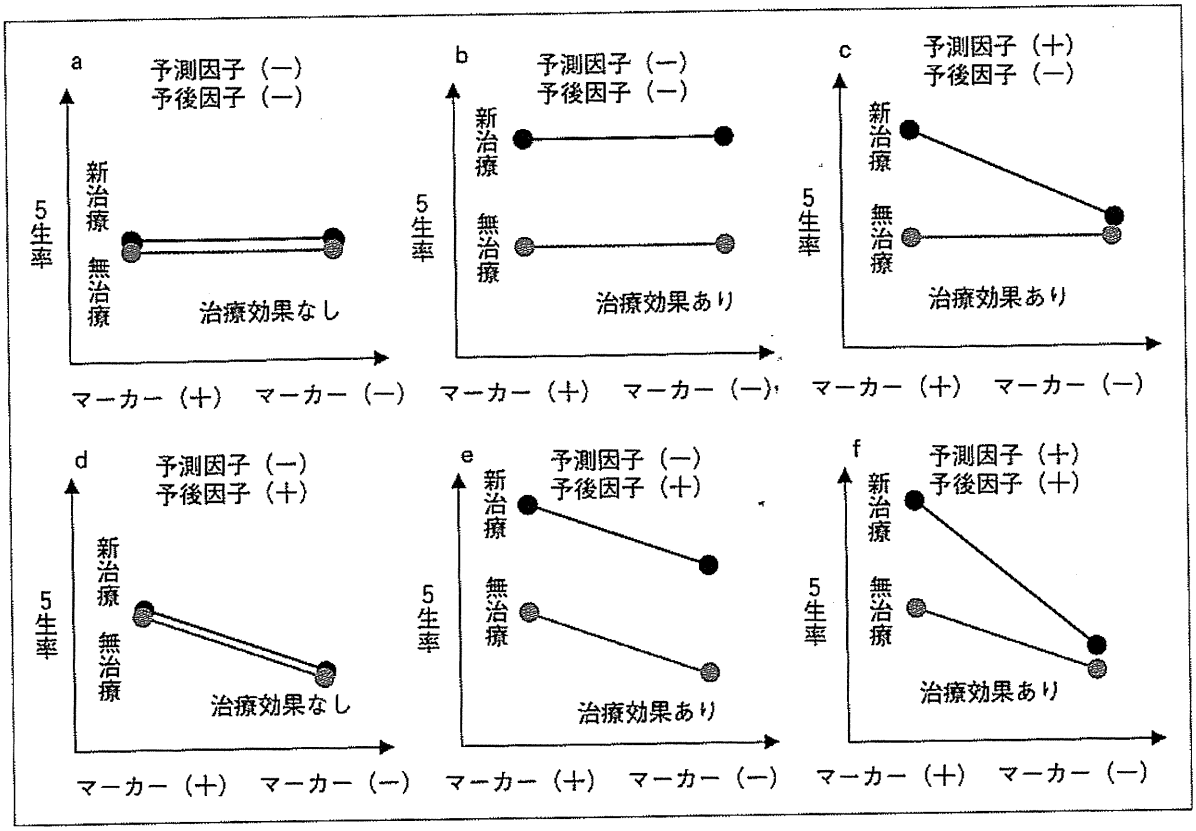


図5　予測因子と予後因子

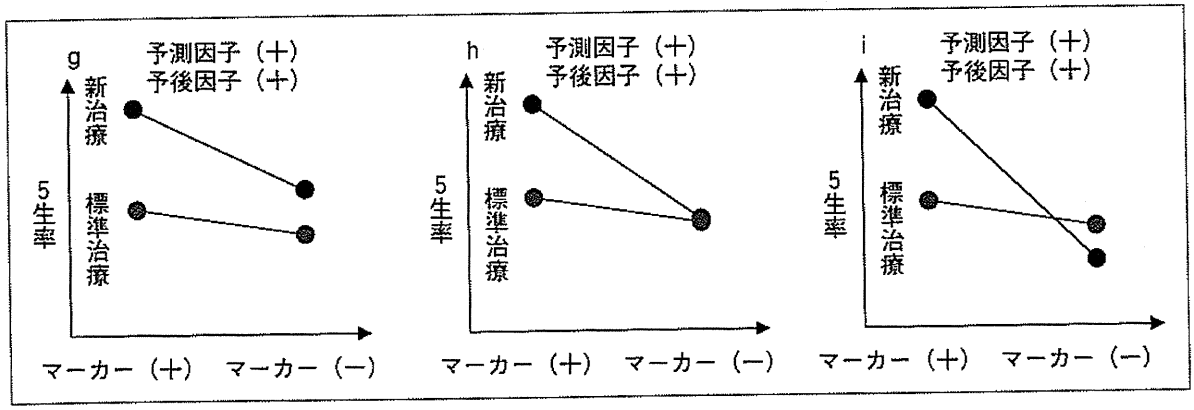


図6　予測因子と治療効果の3つの質的交互作用

の治療効果の上乗せがあったとするとこれはハザード比に換算すると0.7となる．つまり，5年生存率でみるとどちらのサブグループでみても5％の上乗せであるが，ハザード比で考えると治療効果はサブグループ間で異なってしまう．同じ例で，治療効果のハザード比が0.8で一定の場合，マーカー(−)サブグループでの5年生存率の差は6％となる．この場合，治療効果を比で定義するとマーカーは予測因子ではないが，5年生存率の差で定義するとマーカーは予測因子ということになってしまう．どちらが正しいのであろうか．このように統計的交互作用のありなしは用いる効果指標(5年生存率やハザード比など)によって異なってしまうため，疫学では交互作用のことを効果の修飾(effect modification)と呼ぶ．生物学的な交互作用は差の指標における交互作用に対応するという報告もあるが，5年生存率の差は3年生存率の差にも対応しないことを考えると，比の効果指標(通常はハザード比)で交互作用，すなわち予測因子かどうかの判断

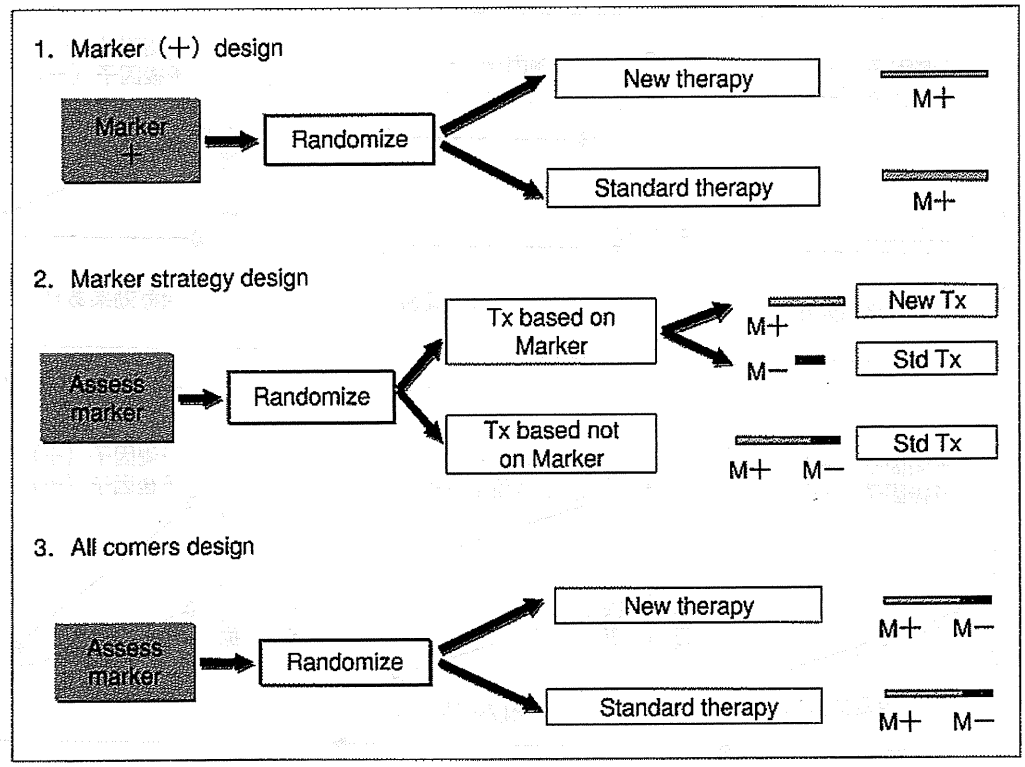


図7 分子標的薬治療開発に有効として提案されている臨床試験デザイン

をすることが便宜的と言えるかもしれない．このように，量的交互作用は必ずしも生物学的な交互作用と関連していない場合があることにも気をつける必要がある．これに対し，質的交互作用は必ず生物学的な治療反応性の違いに対応していると考えてよい．たとえば，標準治療は特定のサブグループに効果があるわけではなく，まんべんなく効果があり，新治療である分子標的薬がある特定のサブグループに効果がある場合，アドオンでなく分子標的治療単剤を用いると，特定のサブグループ以外に効果がないので，むしろ標準治療よりも成績が悪くなる可能性がある(図6-i)．

予測因子探索のための臨床試験デザイン

1．マーカー(+)デザイン

予測因子を探すという考えに基づく個別化治療開発に対して，大きく分けて3種類の臨床試験デザインが提案されている(図7)[1]．まず1つめは「マーカー(+)デザイン〔marker(+) design〕」と言えるもので，マーカーを測定し，(+)の人だけをランダム化して新治療と標準治療の比較を行うものである．いわゆるエンリッチメントデザインもこのデザインの一種と言える．作用機序から考えて分子標的薬の治療効果が予想されるサブグループ〔マーカー(+)〕の集団に臨床試験を行うようなデザインである．この場合，想定しているのは図3の場合，すなわちマーカー(−)の集団には治療効果がない，もしくはあっても小さいであろうことを想定している．しかし，このデザインでは，マーカー(+)の集団で治療効果が観察されても，マーカー(−)に対しては情報が得られないので，予測因子かどうかはわからないし，もしマーカー(−)のサブグループで治療効果がある場合には，その対象の患者さんに有効な治療を提供できないことになる．それを解消するために，最初の比較試験では効果の望めそうなマーカー(+)の対象者で治療法の評価を行い，治療効果が証明されればマーカー(−)の集団で開発を行うというストラテジーも考えられるかもしれない．逆に，後で述べるオールカマーズデザインのように最初はマーカーによらずにランダム化を行い，途中でどこかのサブグループに効果が認められそうになればその集団をエンリッチして試験を続けるという方法も提案されている．しかし，途中でエンリッチ

した場合にはその集団で効果がありそうなことが予想されてしまい，効果があることが予想されているのにランダム化するという倫理的な問題も生じるため，この方法を用いるにはさらなる工夫が必要となるであろう．

マーカー(+)デザインとして，①効果があると期待される集団をエンリッチするものだけでなく，②予後が悪いためさらに治療開発が必要と考えられる集団をマーカー(+)として抽出し，その集団に追加治療を行うかどうか調べるもの，③予後がよく，overtreatmentの可能性のある集団をマーカー(+)として抽出し，治療を手控えられるかどうか調べるもの，④予後が予測できず，治療が必要かどうか不明な集団をマーカー(+)として抽出し，治療が必要かどうか調べるもの，といったものも考えられ，実際にこの考え方に沿って多くの臨床試験が行われている．

2．マーカーストラテジーデザイン

2つめのデザインは「マーカーストラテジー(marker strategy)」と言えるもので，まず始めにマーカーを測定し，ランダム化後，一方の群は「マーカー結果に基づいて」治療を決定し，もう一方の群は「マーカーとは関係なく標準治療を行い」，これを比較するというものである．つまり，プライマリな仮説として，試験に参加した対象者全体に対し，マーカーに基づいて治療選択をするかしないかというストラテジーを比較する．このデザインは，マーカーによって規定されるそれぞれのサブグループに対して異なる治療を行うという「個別化治療ストラテジー」を評価するデザインである．ただ，逆に言うと，効果のある集団を同定する(予測因子マーカーを同定する)，すなわちproof of principleを検証するための臨床試験デザインではない．なぜなら，マーカー(−)のサブグループには，マーカー結果に基づいても基づかなくても同じ治療が行われているため，プライマリな解析で有意に差があった場合でも，マーカー(−)に対する治療効果がわからず，マーカー(+)の群のみで治療効果があるのかどうか調べることが不可能であり，交互作用の検討ができないからである．さらに，対象者全体に対するプライマリな解析で有意な差がみられなかった場合には，マーカー(+)サブグループでの治療効果さえも不明となってしまう．なぜなら，マーカー(+)サブグループで差があり，マーカー(−)で差がなかった場合，マーカー(+)サブグループの割合が多ければよいが，少なければ全対象者でみた場合に有意差なしとなる可能性があり，結果が試験参加者の中のマーカー(+)の割合に依存してしまうからである．

3．オールカマーズデザイン

3つめは「オールカマーズ(all comers)」と呼ばれるデザインで，まず始めにマーカーを測定するが，その結果によらずランダム化し，一方には新治療，もう一方には標準治療を行うというデザインである．代表的な解析法として，まずマーカーによらず全対象者の群間比較を行い，差があればマーカー結果によらず新治療が効果ありとし，差がなければマーカー(+)/(−)それぞれで治療効果の群間比較を行い，どこかのサブグループで差があった場合にそのサブグループで新治療の効果ありとするものである．このデザインを用いることができれば，すべてのサブグループで治療効果があるかどうか調べることができ，さらにサブグループ間で治療効果が異なるかどうかも調べられるので，マーカーが治療効果の予測因子であるかどうかを調べることができる．

このデザインは，最初にマーカーを測定しないで，試験終了後(ランダム化後)にマーカーを測定するのと同じではないかと考えられるかもしれない．しかしながら，事後的に測定を行った場合には，必ずしもランダム化した全例の測定ができるとは限らず，その場合にはランダム化比較ではなくなってしまうので，結果をrandomized controlled trial(RCT)によるエビデンスとして解釈することはできない．また，事後的に行った場合のもう一つの問題点として，事前に特定しないマーカーをいろいろと測定し，それぞれ群間比較することができるので，いわゆる統計的多重性の問題が生じ，αエラーを一定の水準以下(通常は5％)に抑えることができないと言う．αエラーの調整ができないと，結果を探索的にしか解釈できなくなり，治療効果の差がみられたマーカー・サブグループに対してもう一

度検証的な臨床試験を行わないと，プラクティスとして行えないということになってしまう．しかしながら，試験前にマーカーを測定していれば，1度の臨床試験で全体比較とマーカー・サブグループ内での比較の両方を検証的に行う試験デザインを計画することができる．もし，試験前にマーカーを特定できない場合でも，測定をランダム化の前に行っておけば，1つの臨床試験の中で治療効果のあるサブグループを特定し，さらに検証を行うというデザインを立てることも可能である[2)3)]．

オールカマーズデザイン実施上の問題として，治療効果が期待できるようなマーカーサブグループ〔マーカー(+)とする〕が特定されている場合に，そうでないサブグループ〔マーカー(−)〕も含めてランダム化臨床試験を実施するのが困難であると考えるかもしれない．しかし，マーカー(−)のグループに治療効果がないことが証明されたわけでなければ，マーカー(−)の対象者を含めたランダム化比較試験を行うことは十分rationaleがあると考えられる．むしろ，マーカー(−)の対象者にも効果がないかはしっかり調べるべき課題である．

まとめ

マーカーを用いた臨床試験計画はまだ始まったばかりで結果まで到達している具体的な例もまだまだ多くない．今後，積極的にいろいろな研究計画を用い，効率的な治療開発につなげていけるよう経験を蓄積していくことが重要であると考える．

文　献

1) Sargent DJ, Conley BA, Allegra C, et al. Clinical Trial Designs for Predictive Marker Validation in Cancer Treatment Trials. J Clin Oncol 2005；23：2020-7.
2) Simon R. Roadmap for Developing and Validating Therapeutically Relevant Genomic Classifiers. J Clin Oncol 2005；23：7332-41.
3) Freidlin B, Simon R. Adaptive Signature Design：An Adaptive Clinical Trial Design for Generating and Prospectively Testing A Gene Expression Signature for Sensitive Patients. Clin Cancer Res 2005；11：7872-8.

＊　　＊　　＊

トピックス

個別化治療開発の臨床試験デザイン

山本 精一郎*

個別化治療開発とマーカー

がんの治療法は，予後や治療への反応性で定義されたサブグループに対して開発される．あるサブグループにおいて標準とされている治療法に対し，臨床試験において比較が行われ，新しい治療が従来の標準治療より上回る成績を示した場合，それが新しい標準治療として取って代わる．治療法から考えると，ある治療法の開発は，その治療がベネフィットを受けるサブグループの同定と言うことができる．ここでは便宜的にそのサブグループを定義するものをマーカーと呼び，その+/−でサブグループが定義されるものとする．マーカーは組織型や腫瘍の大きさ，遺伝子発現といった腫瘍側因子である場合もあるし，年齢や肝機能，PS といったホスト側の因子である場合もある．マーカーのうち，予後に関係するものを予後因子，治療効果に関係するものを（治療効果の）予測因子と区別して呼ぶこともある．マーカーで規定されたサブグループが受けるベネフィットとして最も考えやすいのが治療効果である．しかし治療効果が同じ場合には，副作用が少ない，治療の負担が少ない，なども患者のベネフィットと考えることができるであろう．すなわち治療法の決定を行う場合，目の前の患者の持つさまざまな要因から患者がどのサブグループに属するか決定し，そのサブグループで最適とされている治療を提供することになる．サブグループを決定するのに関係がある要因の１つは患者の予後であるため，予後を予測する試みが数多くなされてきた．しかし，ただ予後を予測するだけでは治療法選択には結びつかない．また，予後因子は治療効果の予測因子の候補でもある．予後因子としてであれ，予測因子としてであれ，予後因子を治療選択に用いるためには実際のプラクティスで役に立つことを実証することが必須である．この作業がないため，予後因子研究の結果はこれまであまり日常臨床に用いられてこなかった．

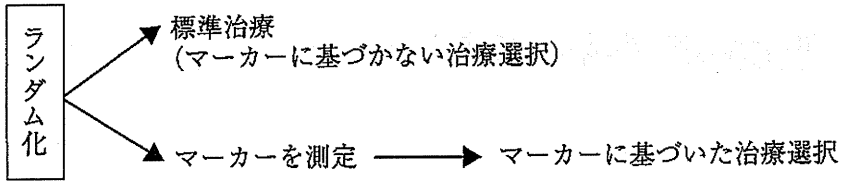
実際の臨床で治療選択に役立つかどうかの検証を行うこと，すなわち予後因子の妥当性研究を行うことは，個別化治療開発のマーカーを開発することでもある．予後因子を探す研究を繰り返したり，同じ予後因子が別の研究でも予後因子として選ばれることを確認しても，不十分である．Simon[1]は，予後因子の妥当性検討というコンテクストで個別化治療のデザインを議論している．非常に参考になる考え方なので，ここではその一部をかみ砕いて解説することによって，個別化治療開発のデザインの紹介を行うこととしたい．

マーカーの妥当性研究デザイン

日常臨床もしくは治療開発の現場において，マーカーはさまざまな目的に用いられるため，妥当性研究もそれぞれの目的に合うように行わなければならない．ここでは，治療選択の助けとなるようなクリニカルベネフィットがあるかどうかを検証するための妥当性研究デザインに焦点を当てる．例えば Oncotype-Dx のリスクスコアは，node(−)，ER(+) の初発の乳がん患者が原発巣の術後にタモキシフェンの投与を受けた場合の予後を予測するために開発された[2]．ここでの妥当性研究とは，このリスクス

* 国立がんセンターがん対策情報センター
がん情報・統計部 室長

図1　マーカーに基づく治療ストラテジーを評価するデザイン（1）



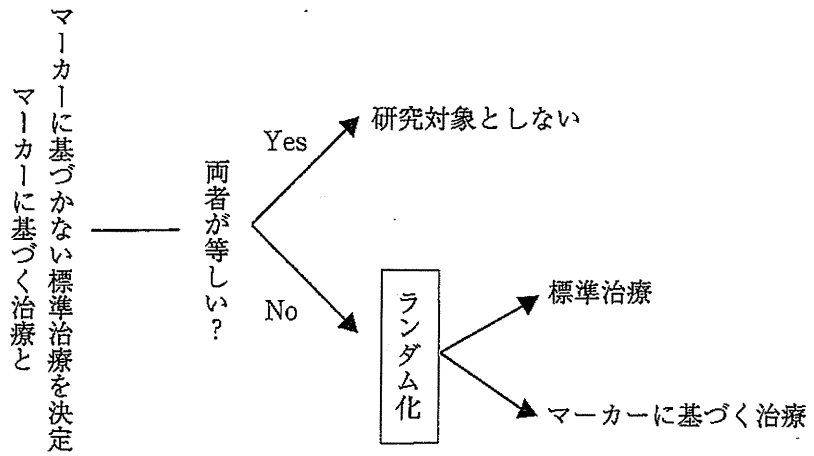
コアを用いることが臨床的に役立つかどうかということである．マーカーの生物学的な機能解析をし，腫瘍との関係を解明することは望ましいが，必ずしも解明できていなくても治療との関連性という点から妥当性は評価できる．

マーカーによる治療選択の評価を新しい臨床試験として実施することを考える．例えば，対象患者をマーカーに基づいて治療選択する群とマーカーに基づかないで治療選択する群のどちらかに割り付けるといった研究である．このデザインでは，マーカーは半分の患者のみで測定される．残念ながらこのデザインは大抵の場合効率が悪く，大きなサンプルサイズを必要とする．なぜなら，多くの患者はどちらの群に割り付けられても同じ治療を受けることになるからである．例えば Oncotype-Dx に基づくリスクスコアによると，node(−)，ER(+) の乳がん患者のおよそ 1/3 は再発リスクが低いと分類される．もし，低リスクに分類されるサブセットで化学療法を手控えることができるかどうかという治療戦略（ストラテジー）を評価したいと思った場合，node(−)，ER(+) の患者をすべてランダム化することは，非常に非効率的である．もし，全員をランダム化してマーカーに基づく治療群の患者のみにアッセイを行った場合，ランダム化された2つのグループは全体として比較されることになるが，2/3 の患者はどちらの群であっても同じ治療を受けることになる（図1）[3]．

それよりも，患者全員に先にアッセイを行い，低リスクに分類された患者だけをタモキシフェン群とタモキシフェン＋化学療法群にランダム化するデザインのほうがずっと効率的である．これは，比較対照となる治療戦略（ストラテジー）が1種類の治療だけでない場合でも同じである．例えば，ステージ1の ER(+) 乳がん患者全体に対して遺伝子発現などのマーカーに基づいた治療選択を評価する場合，マーカーに基づく治療と，マーカーには基づかず診療ガイドラインに基づく治療をランダム化して比較することは，非常に非効率的である．それよりも，適格患者全員の遺伝子発現を調べ，遺伝子発現に基づく治療選択と診療ガイドラインに基づく治療選択の2つのストラテジーが一致しなかったもののみランダム化するほうがずっと効率的である（図2）．このデザインでは，ランダム化された2つの群の治療効果はそれぞれの群で同じ治療を受けたものによって薄められることはない．ただし，2つのストラテジーによって治療方針が異なる患者をスクリーニングするには，多くの患者が必要となるかもしれない．この場合，マーカーに基づく治療選択とは遺伝子発現によって低リスクと予想される患者に対しては化学療法を手控えるというものであり，ガイドラインに基づく治療とは Her2 陽性患者にハーセプチンを投与したり，腫瘍径の小さな患者には化学療法を行わないといったものである．いずれにしても，図2のデザインで評価するためには，それぞれの適格患者に対するマーカーに基づく治療とガイドラインによる治療が何かをランダム化の前に決定できることが必要であり，それをもとに2つのストラテジーによって治療が異なる患者のみをランダム化することになる．

第Ⅲ相臨床試験の目的の1つは，より広い日常臨床の現場で用いることができるかどうかを

図2　マーカーに基づく治療ストラテジーを評価するデザイン（2）

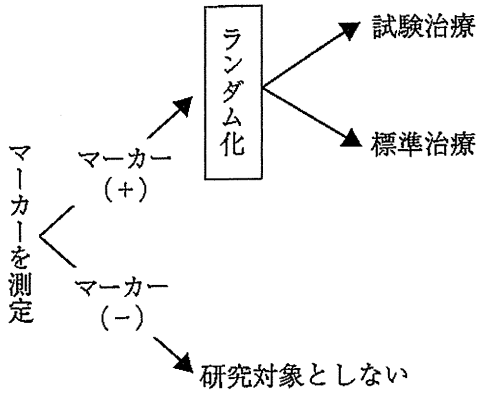


試すことである．レトロスペクティブの研究では測定を1ヵ所のラボで精度高く行うことができるかもしれないが，多施設共同臨床試験では，多くの施設から腫瘍組織を集めたり，日常診療で使用できるリアルタイムアッセイができるかなども含めた形での評価となる．それゆえに，(新規の，前向きの）多施設共同臨床試験はマーカーの外的妥当性を調べるためのゴールドスタンダードとなるのである．

マーカーの妥当性の検証を新しい臨床試験に基づいて行おうとした場合，それが低リスクの患者に対してのものであったとすると，検証するまでに長い追跡が必要となってしまう．そのような状況では，もしすでに行われた多施設共同臨床試験のデータがあり，その中の多くの患者に対して保存された腫瘍サンプルがあるのならば，それを用いて前向きに計画された妥当性研究を行うことが有用であろう．このような状況で妥当性研究を行うためには，新しい患者を登録する場合と少なくとも同様以上に，詳しく，厳密に，前向きに計画しなければならない．

node(−)，ER(+) の乳がん患者に対するPaik ら[3] の Oncotype-Dx に関する研究は，保存された腫瘍サンプルを用い，注意深く前向きに計画された妥当性研究の例と言えるかもしれない．彼らの研究では，低リスクと分類された患者だけをランダム化したほうがずっと効率的ではあるものの，それでも多くの患者が必要であるという考えに基づいている．なぜならこの研究は，観察された治療効果に差がないことをもってクリニカルプラクティスを変えようという意味において治療の同等性を検証する研究と言えるため，小さな差を見逃がさないために多くのサンプルサイズが必要となるからである．さらにこの場合，期待される再発率が非常に低いため，両群の差を見つけるにはより多くの患者が必要となる．しかし，もしマーカーによって予測されるほど再発率が低いのであれば，化学療法によるベネフィットは（あっても）非常に小さなものとなるであろう．そう考えると，別の研究デザインとして，低リスクと分類された患者にタモキシフェンのみで治療を行うシングルアームの研究を行うことが考えられる．長期に観察してもこれらの患者に非常に低い再発率しか見られなければ，マーカーはクリニカルベネフィットを提供することが検証されたと考えて良いであろう．なぜならそのマーカーによって，タモキシフェンのみで十分な効果が得られ，化学療法を行うことによって生じる毒性や不便さ，出費などを経験せずに済む患者を同定することができるからである．これこそが Paik らの研究において用いられたアプローチである．予後因子とされている遺伝子は最初マイクロアレイ研究に基づいて同定されたものであったが，本研究を行うに当たって，多施設共同臨床試験グループである National Surgical Adjuvant

図3 エンリッチメントデザイン



Breast and Bowel Project (NSABP) の研究から保存されたサンプルを用いて，ホルマリン固定した生検組織を用いて実施できるような異なるアッセイを用いたマーカーが開発された．このように事前に特定されたマーカーを用い，タモキシフェンのみを全身療法として受けたNSABP B-14 に参加した 668 人の患者（臨床試験に参加した対象者の一部）のサンプルに対してアッセイを行った．そのうち 51% が低リスクに分類され，その 10 年再発率は 6.8% (95%CI 4.0～9.6) であった．中間リスクと高リスクと分類された患者の再発率はずっと高いものであった（それぞれ 14.3% と 30.5%）．

試験治療に対する遺伝子発現マーカーの開発

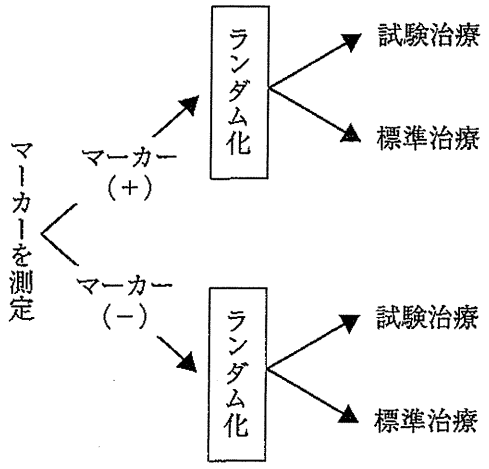
既存の治療を改善するためにマーカーを導入しようとする場合，そのマーカーを用いたほうが用いない場合よりクリニカルベネフィットが得られるかどうかが妥当性検証のポイントとなる．マーカーは，新治療が効果を発揮するような集団を同定するために開発される．これは，作用機序が明らかな分子標的薬がその好例であろう．この場合，マーカー自身の分類能ではなく，マーカー(+) の集団で治療効果があるかどうかが評価の関心事となる．このような場合，必要なサンプルサイズはそのような患者集団を同定できる感度と特異度に応じて大きく変化する．もし感度と特異度の高いマーカーで集団のターゲティングができれば，第Ⅲ相臨床試験が効率良く行えるだけでなく，治療のリスクベネフィット比を大きく向上させることができるので，治療によりベネフィットを受ける患者の割合を大きく上げることができる．

Simon and Maitournam は，図3のようなデザインを用いることを考えた[4]．非臨床もしくは第Ⅰ/Ⅱ相試験の段階で，新治療に高い確率で反応する患者を同定できるマーカーを開発することができたとする．そのマーカーを使って第Ⅲ相試験に入る患者を選んだ場合（いわゆるエンリッチメントデザインの1つ），同じ検出力を得るために必要なイベント数は大きく減少する[5]．

このような第Ⅲ相試験を行えるようなマーカーを開発するためには，より大きな第Ⅱ相試験が必要となるかもしれない．必要な第Ⅱ相試験の大きさは，開発される薬剤のタイプに依存する．例えば作用機序が十分分かっているような薬剤の場合には，ゲノムワイドのスクリーニングは不要かもしれない．しかし，多くの分子標的薬では患者を選択する適切なアッセイは分かっておらず，第Ⅱ相試験でのレスポンダーとノンレスポンダーの発現プロファイルを比較することが最も良いアプローチである場合もあるだろう．そのような場合には，図3の第Ⅲ相試験の登録対象としない患者を第Ⅱ相試験で開発したマーカーを用いて選択するのは十分ではないかもしれない．適格患者をすべて登録可能とするという伝統的なデザインを用い，事前に決めた解析計画に沿ってマーカーの検証をするほうが良い場合もある．

図4は，marker by treatment interaction design を示している[3)6)]．このデザインでは，マーカー(+)/(−) 両方の患者を試験治療と標準治療に割り付ける．解析計画は，2つのマーカーグループ別々に治療効果の差を評価するか，両方のグループで治療効果が等しいことを検証するかのどちらかとなる．このデザインを新治療の開発に用いるなら，後者，すなわち交互作

図4　マーカーと治療の交互作用を評価するデザイン



用を先に検討するのが適切な解析計画となるであろう．もし交互作用が事前に決めた水準で有意でないなら，新治療は標準治療と（マーカーの値によらず）全体で比べることになる．交互作用が有意であれば，マーカーで定義されたサブセットの中で新治療と標準治療を比べることになる．このような研究デザインや交互作用の検討を行う際の適切な有意水準などについては，さらなる研究が必要であろう．

Freidlin と Simon は，図4のデザインに対して別の解析計画を提案した[7]．彼らは，対象者全体に対する帰無仮説の有意水準を 0.04 とすることを提案している．通常の α エラー 0.05 のうち残りの 0.01 を，マーカーが（+），すなわちマーカーがレスポンダーであると予測するサブセットでの新治療の評価のためにとっておくのである．つまり交互作用を先に検討せず，まず全体に対して帰無仮説の検定を行う．全体の帰無仮説が棄却されれば全体に対して効果ありと結論し，マーカーは不要となる．全体の帰無仮説が 0.04 で棄却されなければ，1つだけサブセット解析を行う．マーカーによって最も新治療に反応すると予測されたサブセットで，新治療を標準治療と比較するのである．帰無仮説が棄却されれば，そのマーカーが規定するサブセットでは新治療が効果ありということになる．この解析方法は，良いマーカーを開発すれば広い適応がとれなくなってしまうことを心配する製薬会社に対しても，マーカーを開発するインセンティブを与えるであろう．この方法は，全体の治療効果がマーカー（+）のサブセットでの治療効果のせいではないことを保証するものではないが，マーカー（−）の患者に対する治療効果を評価できるデータを提供できるデザインと言える．

ま と め

本稿では Simon[1] に従って，個別化治療開発をマーカー開発という観点から検討した．今後，個別化治療開発の中で，マーカー開発とその妥当性研究という考え方はより一般的になってくると思われる．ここで言うマーカーとは，ある実験の測定値だけでなく，治療選択に関するサブグループを特定する要因のすべてを指す．今後はこれまで以上に診断と治療選択が直結することが予想され，治療法選択という観点からの診断が評価されることになるであろう．

文　献

1) Simon R: Development and validation of biomarker classifiers for treatment selection. J Stat Plan Inference 138: 308–320, 2008.
2) Paik S, et al: A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 351: 2817–2826, 2004.
3) Sargent D J, et al: Clinical trial designs for predictive marker validation in cancer treatment trials. J Clin Oncol 23: 2020–2027, 2005.
4) Simon R, et al: Evaluating the efficiency of targeted designs for randomized clinical trials. Clin Cancer Res 10: 6759–6763, 2004.
5) http://linus.nci.nih.gov/brb/samplesize/
6) Pusztai L, et al: Clinical trial design for microarray predictive marker discovery and assessment. Ann Oncol 15: 1731–1737, 2004.
7) Freidlin B, et al: Adaptive signature design: an

adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. Clin Cancer Res 11: 7872–7878, 2005.

Clinical Trial Designs for Development of Individualized Medicine

Seiichiro Yamamoto
Cancer Information Services and Surveillance Division, Center for Cancer Control and Information Services, National Cancer Center

4 バイオマーカーを用いた臨床試験デザイン

はじめに

細胞傷害性の抗悪性腫瘍薬の多くは，細胞分裂過程における普遍的な現象に作用し，細胞増殖能の違いに基づいて腫瘍細胞への選択性を示してきた．これに対して，近年開発の進んでいる分子標的薬は理論上，腫瘍細胞に特異的に作用すると考えられている．すでに多くのがん種において分子標的薬の使用は標準的な治療戦略として組み込まれており，今後も分子標的薬の開発は加速度的に発展していくものと期待される．

細胞傷害性の抗悪性腫瘍薬に対する従来の臨床試験では，その薬剤で治療される集団を大きく捉え，比較的広い適格基準を設定してきた．一方，同じがん種でも遺伝子レベルでは多様な差異があることが一般的であるし，設計段階から「標的ありき」で「個別化医療」を目指した薬剤ならば，その薬剤の効果が高い投与対象を選別して治療開発が進められることが本来的には望ましいと考えられる．対象を特定せずに試験を行えば，非小細胞肺がんにおけるゲフィチニブの試験のように[1]，効果のない患者も多く含んでしまい，特定のサブグループへの有効性が見過ごされてしまう可能性が高くなる．

ある治療法が特定の患者サブグループに有効であることを示すためには，そのようなサブグループを選別できる治療効果予測因子の存在，統計学的には(治療法×サブグループの)交互作用を検出できるようなバイオマーカーの存在が必要である(以下，バイオマーカーとはこの意味で用いる．予測因子と交互作用の説明は文献2に詳しい)．しかしながら，試験開始時はバイオマーカーの信頼性が不十分な場合も多く，適切な投与対象を意識しながら試験を計画実施することは新たな課題をもたらしている．以下ではバイオマーカーを組み込んだ臨床試験デザインについて，現在の知見を述べたい．

1．第Ⅲ相試験

本節ではバイオマーカー(以下，マーカー)を利用するランダム化第Ⅲ相試験のデザインについて概説する．

1）エンリッチメントデザイン

すでに特定のマーカーが利用可能な場合には，新治療に反応すると予測されるサブグループ〔マーカー(+)〕のみを対象に，新治療 vs. 標準治療の比較試験を行うことが可能である．いわゆるエンリッチメントデザイン(enrichment design)とよばれるものであり，乳がんに対するトラスツズマブ治療の開発に効果的に利用された[3]．トラスツズマブの試験では免疫組織化学的に HER2 陽性であることを登録の適格条件としている．エンリッチメントデザインと通常のデザイン(マーカーによる選別なしで登録)の効率性を比較した場合，マーカー(+)の全体に占める割合が少なく，新治療がマーカー(−)にあまり効果がないようなケースならば，エンリッチメントデザインのほうがサンプルサイズを大きく節約できて効率的である[4,5]．

エンリッチメントデザインは典型的には(a)効果が特にあると考えられる集団を選別して試験を行う