

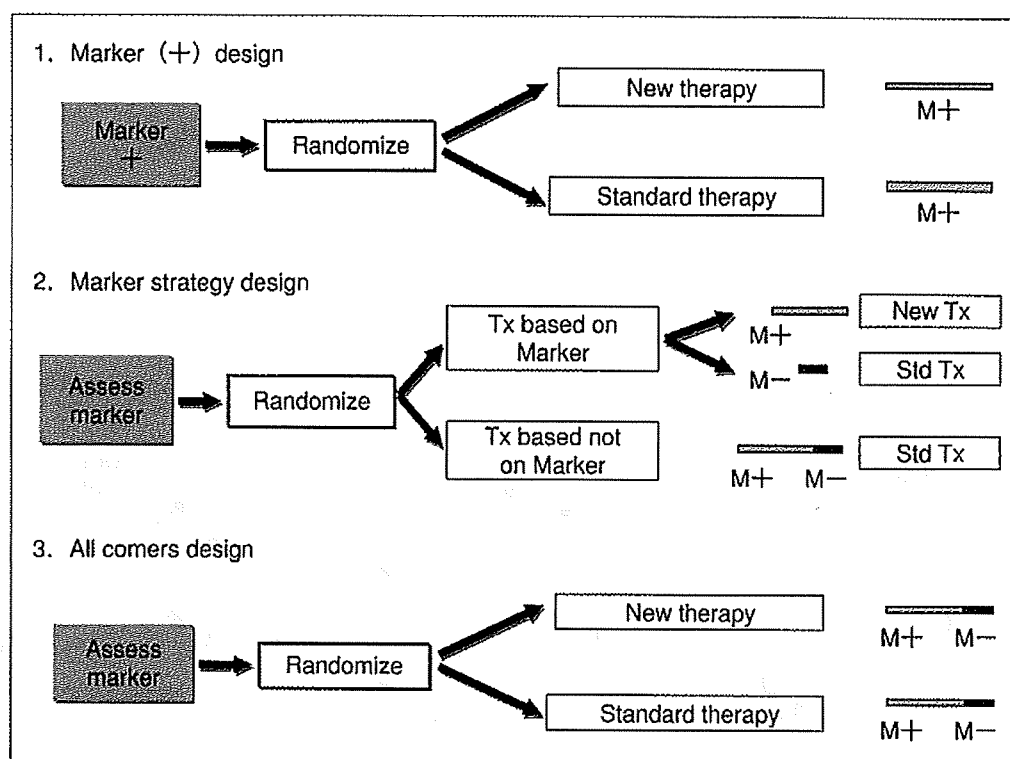


図7 分子標的薬治療開発に有効として提案されている臨床試験デザイン

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つと言える。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効果があると思わ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のが不可能であり、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ず、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られるかもしれない。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られる。

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。

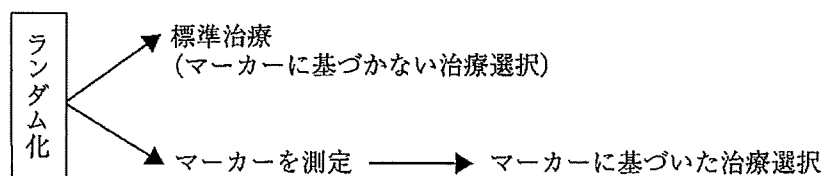
実際の臨床で治療選択に役立つかどうかの検証を行うこと、すなわち予後因子の妥当性研究を行うことは、個別化治療開発のマーカーを開発することでもある。予後因子を探す研究を繰り返したり、同じ予後因子が別の研究でも予後因子として選ばれることを確認しても、不十分である。Simon¹⁾は、予後因子の妥当性検討というコンテキストで個別化治療のデザインを議論している。非常に参考になる考え方なので、ここではその一部をかみ砕いて解説することによって、個別化治療開発のデザインの紹介を行うこととしたい。

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

* 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戦略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。

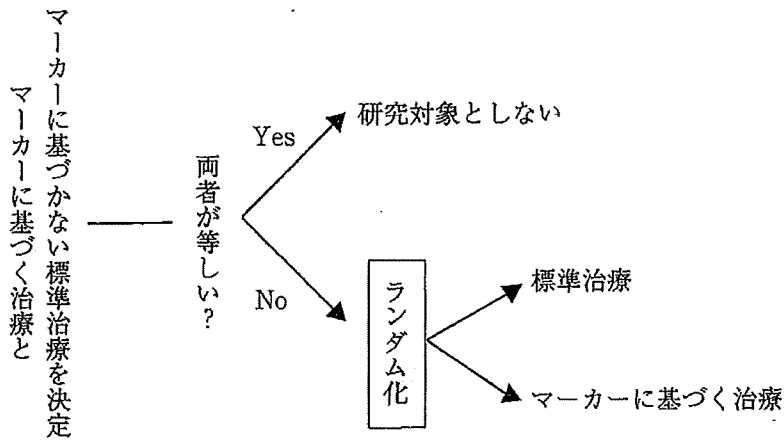
マーカーによる治療選択の評価を新しい臨床試験として実施することを考える。例えば、対象患者をマーカーに基づいて治療選択する群とマーカーに基づかないで治療選択する群のどちらかに割り付けるといった研究である。このデザインでは、マーカーは半分の患者のみで測定される。残念ながらこのデザインは大抵の場合効率が悪く、大きなサンプルサイズを必要とする。なぜなら、多くの患者はどちらの群に割り付けられても同じ治療を受けることになるからである。例えば Oncotype-Dx に基づくリスクスコアによると、node(-)、ER(+) の乳がん患者のおよそ 1/3 は再発リスクが低いと分類される。もし、低リスクに分類さ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)⁹⁾。

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戦略を評価するデザイン (2)



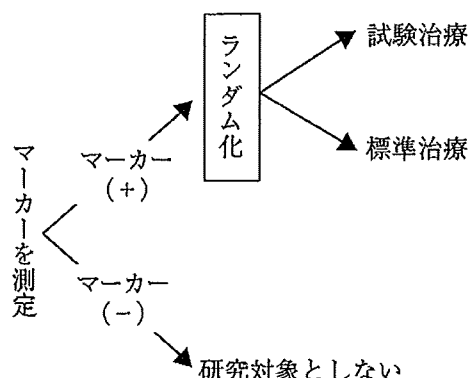
試すことである。レトロスペクティブの研究では測定を1カ所のラボで精度高く行うことができるかもしれないが、多施設共同臨床試験では、多くの施設から腫瘍組織を集めたり、日常診療で使用できるリアルタイムアッセイができるかなども含めた形での評価となる。それゆえに、(新規の、前向き)多施設共同臨床試験はマーカーの外的妥当性を調べるためのゴールドスタンダードとなるのである。

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臨床的ベネフィットが得られるかどうかの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の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

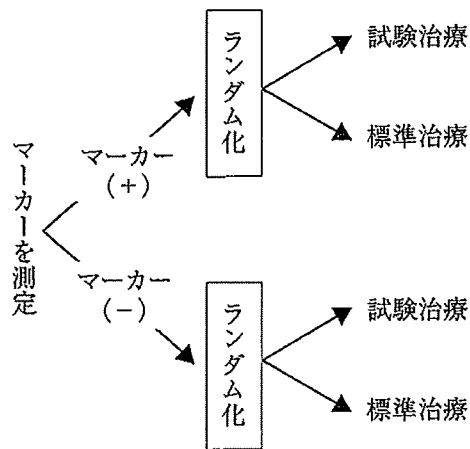
ターゲティングができれば、第Ⅲ相臨床試験が効率良く行えるだけでなく、治療のリスクベネフィット比を大きく向上させることができるので、治療によりベネフィットを受ける患者の割合を大きく上げることができる。

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スとノンレスポンス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³⁰⁾。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nder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 Simon¹⁾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 DJ, 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 signa-

ture 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

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、非小細胞肺癌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逃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

ものであるが、その他にも(b)標準治療では成績が十分でない集団を選別して、その集団に追加治療の有効性を検討する、逆に(c)標準治療のみで成績が十分によく、過剰治療(over treatment)の可能性のある集団を選別して、その集団に治療が手控えられかどうかを調べる、といった場合にも適用できる。ただし、エンリッチメントデザインではマーカー(-)に対するデータは得られないので、マーカーが真の意味で治療効果の予測因子となるかの評価は行えない[マーカー(-)でも新治療が有効である可能性を否定できないため]。先行研究のデータや明確な生物学的根拠などから、マーカー(-)を試験に含めることが合理的ではないと判断できる状況に適したデザインである。

試験開始時にマーカーのバリデーションができていない場合は、マーカー検査の結果にかかわらずに登録を行い、全体効果(overall treatment effect)の検証およびマーカーの検証を同時に実施することが求められる。すなわち、マーカーの妥当性評価も行う解析計画が必要になる。たとえば、以下のようなデザインが考えられる⁹⁾。

2) 交互作用の検定に基づく解析

このデザインでは、治療法(新治療, 標準治療)×サブグループ(マーカー(+), (-))の交互作用の検定をまず実施する。交互作用の検定が有意とならなければ、全例を対象として新治療 vs. 標準治療の検定を行う(全体効果の検証)。交互作用の検定が有意となれば、マーカー検査の結果別に(すなわちサブグループ別に)新治療 vs. 標準治療の検定を行う。これは二元配置分散分析の伝統的なアプローチに従った方法といえる。

3) 有意水準の分割

Simon & Wang⁷⁾によるアプローチでは、全例を対象とした新治療 vs. 標準治療の検定を有意水準 α_1 (たとえば $\alpha_1=0.04$) でまず実施する。この検定が有意となれば、新治療が全体に対して有効であると結論する。有意とならない場合にはマーカー(+)に対して、新治療 vs. 標準治療の検定を有意水準 $\alpha_2=$

$\alpha-\alpha_1$ (たとえば $\alpha_2=0.01$) で行い、この検定が有意となれば、新治療はマーカー(+)に対して有効であると結論する。2回検定を行うことになるが、有意水準の分割によって試験全体の有意水準を $\alpha_1+\alpha_2=\alpha$ ($0.04+0.01=0.05$) 以下に抑えることができる。すなわち、有意水準の分割によって検定の多重性を避けることができる。このデザインでは、新治療の効果を広い対象に対して期待しているが、万一のときに(全体効果が示されなかったときに)検証的に機能するサブセット解析をあらかじめ準備しておくものである。全体における検定とマーカー(+)における検定の相関を考慮したより正確な有意水準も研究されている⁸⁾。

しかし、現実には試験開始時にマーカーに関する情報がほとんどない、あるいはマーカーとなりうる連続量の測定値は利用できても(例えば遺伝子発現量など)、その閾値[マーカー(+)または(-)に2群化するためのカットオフ値]が不明であるため、マーカー(+)の集団さえ定められない、という事態に直面しうる。こういった状況に対処するため、1つの試験のなかでマーカー(ないしその閾値)を確立し、かつ(そのマーカーが定める)マーカー(+)における治療効果の検証までいっぺんに行ってしまうデザインも提唱されている^{9,10)}。

4) 適応的 signature デザイン

このデザイン(adaptive signature design)では、所定の追跡期間が終了した後、試験前半部分で登録された症例(図1のA)を対象に、適切な統計手法を用いて予測マーカーの構築を行う。そのうえで、主たる解析は以下の2ステップをとる。全例を対象とした新治療 vs. 標準治療の検定を有意水準 α_1 で実施する(図1のA+B+Cで新治療 vs. 標準治療の検定を行う)。この検定が有意となれば、新治療が全体に対して有効であると結論する。有意とならない場合には構築されたマーカーを用いて、試験後半部分で登録された症例(図1のB+C)の中からマーカー(+) (図1のB)の集団を選別し(エンリッチメントし)、この集団で新治療 vs. 標準治療の検定を有意水準 $\alpha_2=\alpha-\alpha_1$ で行う(図1のBで新治療 vs. 標準

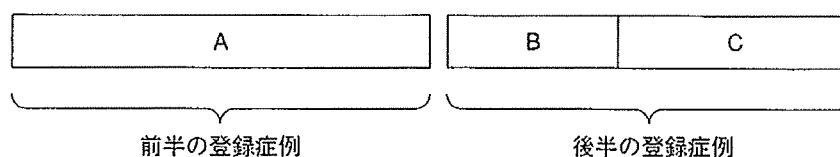


図1 適応的 signature デザイン

治療の検定を行う)。この検定が有意となれば、新治療はマーカー(+)に対して有効であると結論する。

すなわち第1ステップとして全体効果の検証を行い、そこで有効性を証明できなければ、引き続く第2ステップとしてマーカー(+)における治療効果の検証を行う。上記 Simon & Wang によるアプローチと同様の手順であるが、利用可能なマーカーの存在を前提としているかどうかの点でまったく異なる。マーカーの構築も同じ試験内で行ってしまおうというのが適応的 signature デザインの主旨である。もし試験開始前にすでにいくつかのマーカー候補が存在する場合には、そのなかからいずれか1つを統計的に選択して(マーカーを新たに構築するという最初の手順を、候補マーカーのなかから1つ選択するという手順に置き換える)、第2ステップでその選択されたマーカーを用いることもできる。薬剤設計のために標的分子は同定していても実際の患者標本では何をマーカーとして測定すべきかが明確ではなく、たとえば点突然変異や遺伝子増幅、タンパク過剰発現の有無など、マーカー候補が複数存在しているケースは多いので、そうした場合にもこのデザインは有効である。

デザイン上重要なのは、「マーカー構築のためのデータ」と「マーカー(+)での治療効果を検証するためのデータ」を分離している点である(図1でマーカー構築のための集合Aとマーカー(+)での治療効果を検証するための集合Bは別である)。これらを分離しないと、マーカーの予測精度を過大評価してしまうことが知られている。一方、全体効果の検証ではマーカーに依存した検定を行うわけではないので、全例を対象にして解析することに問題はない。

5) 適応的バイオマーカー閾値デザイン

試験開始時にマーカーとなる連続量の測定値は利

用可能だが、閾値が定まっていない場合を想定したデザイン(biomarker-adaptive threshold design)も提案されている。1つの試験内で、第1ステップで全例を対象とした全体効果の検証、第2ステップで最適な閾値を探索し、その閾値から適応的に定めたマーカー(+)に対する検証も計画するというものである。適応的 signature デザインと同様、1つの試験内でマーカーやその閾値を決定し、その妥当性評価まで行ってしまうことの是非はあるが、治療開発の速度を上げる有効な方法となりうるものである。

複数のマーカーを組み合わせる治療戦略を決定し、その戦略自体の妥当性を評価したいような場合には、以下のデザインが有用である。

6) マーカーベース戦略デザイン

このデザイン(marker-based strategy design)は、ランダム化された2群のうち、一方の群はマーカーに基づく治療戦略をとり、もう一方の群は一律に標準治療を行うといった状況を想定している(当然、後者の群がコントロールとなる)¹¹⁾。たとえば、複数マーカーを組み合わせる治療選択する戦略の検証に利用できる。進行期非小細胞肺癌を対象とした MADeIT 試験¹²⁾は、ゲムシタビン、カルボプラチンの感受性にそれぞれ関連のある RRM1, ERCC1 の発現レベル(high, low)を組み合わせ、都合4通りのサブグループ毎にレジメンを変える治療戦略をとる群と、一律に(標準治療として)ゲムシタビン+カルボプラチンを行う群を比較する研究として実施されており、マーカーベース戦略デザインと考えられる(図2)。ただし、このデザインはエンリッチメントデザインと同様、マーカーの妥当性自体を直接示すためのものではない¹³⁾。マーカーが規定するサブグループに対して最適と思われる治療法を定める“治療戦略”そのものを評価しようとしているデザ

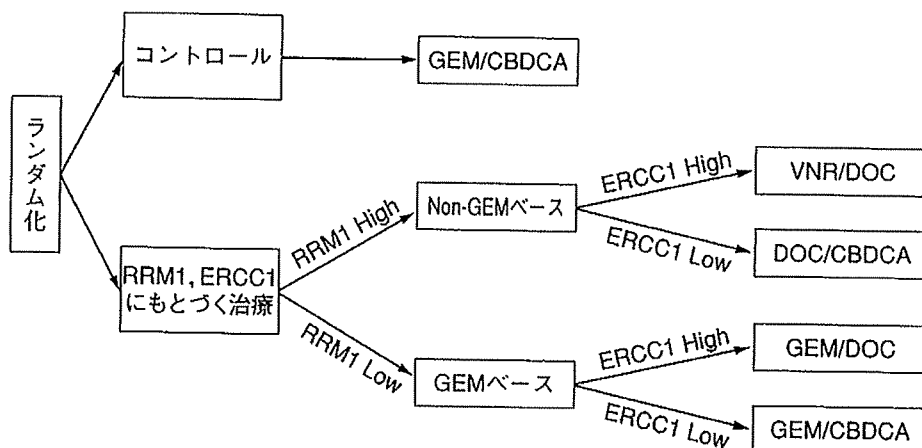


図2 遺伝子発現レベルに基づくマーカーベース戦略デザイン

インである。

2. 第Ⅱ相試験

分子標的薬の開発が進められるなか、いかに第Ⅱ相試験を実施して第Ⅲ相試験の成功へとつなげるかに関して様々な議論がなされている。「第Ⅱ相試験の方法論には多く変更の余地があり、より効率的に薬剤開発を行うためにいかに改良していくか?」という議論であり、バイオマーカーの役割についても予測因子やエンドポイントなど複数の観点から提案がなされている¹⁴⁾。これまでの分子標的薬の第Ⅱ相試験(ただし単剤試験に限る)のレビュー結果に基づき、奏効率をエンドポイントとした非ランダム化試験での実施を推奨する向きもあるが^{15,16)}、個々の試験が抱える状況にも依存するため画一的な議論は難しい。しかし、(是非は別として)従来どおりの単純な設定で行うにしても、今後はバイオマーカーを積極的に取り入れた試験実施が必要となるであろう。第Ⅱ相段階でマーカーを用いた患者の絞り込み(エンリッチメント)まで実施するのは困難であるにしても、候補マーカーを評価して次相で活用できるような情報を得るための試験計画は可能である。たとえば、マーカー(+)と(-)では奏効率が異なることをあらかじめ想定して古典的なサイモンの2段階法(two-stage design)を拡張するデザイン¹⁷⁾や、全体に対する奏効率が基準をクリアできなかった場合に

マーカー(+)に対する奏効率の評価へと移行するタンデムデザイン(tandem design)¹⁸⁾などが提案されている。

ベイズ流デザインの適用への注目も高まっている¹⁹⁾。たとえば、バイオマーカーに関する情報をランダム化の際の割付け確率に逐次的に反映させるアダプティブランダム化を採用したBATTLE試験²⁰⁾などのような事例も出てきている。BATTLE試験は既治療の進行期非小細胞肺癌を対象としたランダム化第Ⅱ相試験で、EGFR変異・K-ras変異・VEGF発現など、複数マーカーの組み合わせによって患者集団を計5つのサブグループA～Eに分割する。サブグループ別に4つの分子標的薬治療の評価を行うことが目的であるが、各サブグループで4つの治療法の効き目は異なることが予想される(たとえば、AグループはEGFR変異を有するのでエルロチニブが効きやすいetc.)。そこで各サブグループにおいて効果がより期待できると思われる治療法により多くの患者が割付けられるように割付け確率を逐次変更してランダム化を行うのである(したがって、登録実施中に割付け確率は変化し、またA～Eごとの割付け確率が異なってもよい)。割付け確率の変更はベイズ流に、すなわち、割付け確率を変更しようとする時点までに得られているアウトカム情報をもとに変更する。この手順で試験を進めてデータを集積し、A～Eごとに4治療法の有効率(BATTLE試験では1時点での無増悪生存率)を推定することが目的である。従来のデザインを改良し

て、より効率的な実施を試みる目的で、今後、ベイズ流デザインの早期フェーズへの適用は増加していくことが予想される。

おわりに

一部の患者サブグループにベネフィットをもたらす治療法の開発には、標的集団を選別するためのバイオマーカー、およびそれを明示的に組み込んだ臨床試験のデザインが必要である。ここではいくつかのデザインの整理を試みたが、少なくとも本稿の執筆段階では上記に述べた方法を含め、標的集団の存在を意識したデザインによる試験の経験はほとんど積まれていない。分子標的薬時代の到来がわれわれに要請する「治療法とバイオマーカーの前向きな同時開発」は新たな試みであり、試験の複雑性が増す、さらなる方法論の研究、それを可能にする試験実施体制の充実、試験実施を通じた経験の蓄積が求められる。

●文 献

- 1) Saijo N, Takeuchi M, Kunitoh H. Reasons for response differences seen in the V15-32, INTEREST and IPASS trials. *Nat Rev Clin Oncol*. 2009; 6: 287-94.
- 2) 柴田大朗. 選択的治療を可能にする方法論 生物統計学の側面から. *Mebio Oncology*. 2008; 5: 4-9.
- 3) Slamon DJ, Leyland-Jones B, Shak S, et al. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *N Engl J Med*. 2001; 344: 783-92.
- 4) Simon R, Maitournam A. Evaluating the efficiency of targeted designs for randomized clinical trials. *Clin Cancer Res*. 2004; 10: 6759-63.
- 5) Maitournam A, Simon R. On the efficiency of targeted clinical trials. *Stat Med*. 2005; 24: 329-39.
- 6) Simon R. The use of genomics in clinical trial design. *Clin Cancer Res*. 2008; 14: 5984-93.
- 7) Simon R, Wang SJ. Use of genomic signatures in therapeutics development in oncology and other diseases. *Pharmacogenomics J*. 2006; 6: 166-73.
- 8) Song Y, Chi GY. A method for testing a prespecified subgroup in clinical trials. *Stat Med*. 2007; 26: 3535-49.
- 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*. 2005; 11: 7872-8.
- 10) Jiang W, Freidlin B, Simon R. Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. *J Natl Cancer Inst*. 2007; 99: 1036-43.
- 11) 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.
- 12) Simon GR, Begum M, Bepler G. Setting the stage for tailored chemotherapy in the management of non-small cell lung cancer. *Future Oncol*. 2008; 4: 51-9.
- 13) Hoering A, Leblanc M, Crowley JJ. Randomized phase III clinical trial designs for targeted agents. *Clin Cancer Res*. 2008; 14: 4358-67.
- 14) McShane LM, Hunsberger S, Adjei AA. Effective incorporation of biomarkers into phase II trials. *Clin Cancer Res*. 2009; 15: 1898-905.
- 15) El-Maraghi RH, Eisenhauer EA. Review of phase II trial designs used in studies of molecular targeted agents: outcomes and predictors of success in phase III. *J Clin Oncol*. 2008; 26: 1346-54.
- 16) Booth CM, Calvert AH, Giaccone G, et al. Design and conduct of phase II studies of targeted anticancer therapy: recommendations from the task force on methodology for the development of innovative cancer therapies (MDICT). *Eur J Cancer*. 2008; 44: 25-9.
- 17) Jones CL, Holmgren E. An adaptive Simon Two-Stage Design for Phase 2 studies of targeted therapies. *Contemp Clin Trials*. 2007; 28: 654-61.
- 18) Pusztai L, Anderson K, Hess KR. Pharmacogenomic predictor discovery in phase II clinical trials for breast cancer. *Clin Cancer Res*. 2007; 13: 6080-6.
- 19) Berry DA. Bayesian clinical trials. *Nat Rev Drug Discov*. 2006; 5: 27-36.
- 20) Kim E, Herbst R, Lee J, et al. Phase II randomized study of biomarker-directed treatment for non-small cell lung cancer (NSCLC): The BATTLE (Biomarker-Integrated Approaches of Targeted Therapy for Lung Cancer Elimination) clinical trial program. *J Clin Oncol*. 2009; 27(15s): abstr 8024.

〈山中竹春 山本精一郎〉



研修医からの質問 Q & A

今回の質問者：文田壮一(近畿大学医学部内科学講座腫瘍内科学部門)
吉田えり(三沢市立三沢病院腫瘍内科)

● 最新の臨床試験結果はどこで仕入れるものなのでしょうか？どのような場合に実際に使用できるエビデンスと言えるのでしょうか？

■ A 最新の臨床試験結果は国際学会や学術雑誌で発表され、学会などでコンセンサスを得ることによって実際に使用されるエビデンスとなります。

最新の臨床試験結果は、国際学会などで発表されます。最も大きなものは米国臨床腫瘍学会(American Society of Clinical Oncology, ASCO)で、最近ではASCO-GI(消化器がん)などその分科会でも新しい研究結果の発表が行われます。血液がんでは米国血液学会(American Society of Hematology, ASH)、乳がんではサンアントニオ乳癌シンポジウム、なども代表的な発表の場です。これらの学会やピアレビュー雑誌に新しい治療が有効であるという信頼できる結果が掲載されると、学会もしくは、さまざまな場で議論が行われ、診療ガイドラインなどでそれらが新しい標準治療としてコンセンサスが得られると実際に使われるようになります。通常、信頼できる結果とはこれまでの標準治療を対照群としたランダム化比較試験の成績です。広く使われるためには、保険でカバーされるようになる(されている)ことも条件となります。

● 回答者：山本精一郎(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)

特集

婦人科がん臨床試験参加に必要な知識

2. がん臨床試験の生物統計学

やまもと せい いち ろう
山本精一郎*

国立がんセンターがん対策情報センター/JCOG データセンター*

要旨

Phase1 臨床試験の目的は、試験治療の最適用量を決めるために毒性によるスクリーニングを行うことである。Phase2 試験の目的である Phase3 へ進むかどうかの決定には、毒性の情報やほかの開発薬剤の状況なども加味するので、統計的判断は有効性の足切りの目安として使うのが一般的である。Phase3 試験の論文を読む際には、primary endpoint に関する結果が重要であり、有意かどうか、どこまでいえるか（何が検証的で何が探索的か）を理解しておくことが必要である。

Key Words 生物統計学, 臨床試験デザイン

はじめに

本稿では、がんの臨床医ががんの臨床試験の論文を読むのに必要な知識、おもに統計学的側面について知っておくべきことを述べる。これらはもちろん、臨床試験に参加する場合にも知っておくべき知識であり、自ら臨床試験を計画する研究者はもっと深く知る必要がある。また、論文を読むには統計だけでは不十分なので研究デザインについて理解することが必要であるため、デザインについても説明する。逆に、統計の概念自体を本稿のみで説明するのは難しいため、参考図書も合わせて学習していただきたい。

Phase1 臨床試験論文を読むのに必要な統計

Phase1 臨床試験の目的は、試験治療の最適用量を決めることである。細胞障害性抗がん剤の場合には、毒性と有効性が比例することを利用し、毒性によるスクリーニングを行うことによって、推奨用量を決定してきた（図1）。この目的のために最もよく用いられるデザインは3例コホートである。これは、3例ずつ同一用量を投与し、毒性で評価した最大耐用量となるまで用量を増加する方法である（図2）。3例コホートデザインは、それを用いて実際に臨床試験を計画し、実施する際にも、また、3例コホートを用いた臨床試験論文を読むためにも統計的な知識は特に必要ない。3例コホートは、2/6の対象者に用量制限毒性（Dose Limiting Toxic-

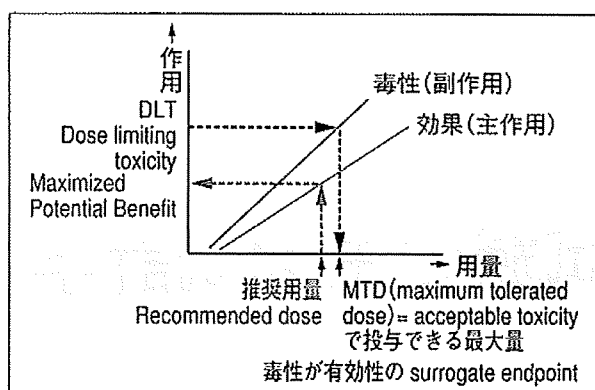


図1 細胞障害性抗がん剤 p1 デザインの rationale
"More is better" with acceptable toxicity

ity: DLT) が発現すると中止するようなデザインとなっているため、抗がん剤としては DLT が 33% あたりが限界ラインと想定しているのかもしれないが、統計的に考察すると、実際には 33% よりも早めに中止となることが多いと考えられる。3 例コホートに統計的な根拠はあまりないが、この方法でこれまで開発を行ってきた、特に問題が生じず、やりやすくわかりやすいからということで、現在でもよく用いられる経験則的な方法であるといえる。3 例コホートは細胞障害性抗がん剤のために開発されてきたデザインであるが、分子標的薬のように、必ずしも毒性と有効性が比例するとは限らない場合にも、(明らかに優れた方法がほかにないため) 標準的に用いられているようである。

3 例コホートはかなり低い用量から開始し、順に増量するため、なかなか最適用量に達しない、最適用量付近でのサンプルサイズが小さい、といった問題が指摘されている。そこで、この点を解消するために提案されている方法が CRM (continuous reassessment method) である。この方法は、1 例ごとに統計的に用量反応曲線を推定し、それに基づいて最適と思われる用量を投与、その結果に基づいて用量反応曲線を推定し直し、次の対象者の用量を定める、を繰り返す方法である。ベイズ流統計学の手法を用いているが、3 例コホートに比べ最適用量付近が

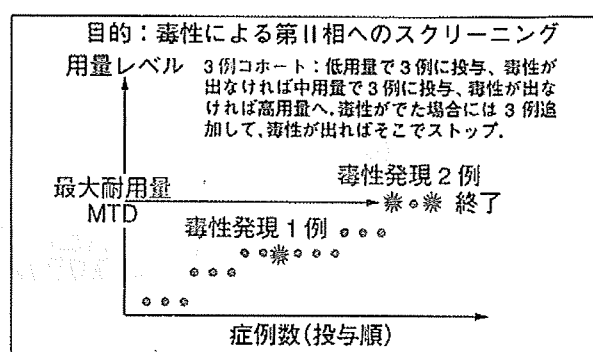


図2 第I相臨床試験の典型的なデザイン
[By courtesy of Dr. Ishizuka]

選択されやすい優れた方法であることを知っていればよい。実際に実施するには、3 例コホートに比べ、手順がかなり複雑になることもあって、それほど頻繁には用いられていない。

Phase2 臨床試験論文を読むのに必要な統計

Phase2 臨床試験の目的は、試験治療の開発を続け、標準治療を対照とした第 III 相試験 Phase3 を行うかどうかの判断を行うことである。Phase3 へ進むかどうかの決定には、毒性の情報やほかの開発薬剤の状況なども加味するので、統計的判断は有効性の足切りの目安として使うのが一般的である。シングルアーム試験やランダム化選択デザインなど通常 100 例以下のデザインが用いられることが多いが、検証的な結果ではなく、いずれにしろ、その先に phase3 試験が必要であることを理解すべきである。

おもなデザインは、閾値・期待値を用いたシングルアームのデザインである。これは、primary endpoint (奏効割合 response rate など) に対して、それ以下なら開発を中止すると考える値を閾値として設定し、真の response rate がある値 (期待値) であった場合には、実際に得られるデータによって閾値以下であることが高い確率で統計的に棄却できるようにサンプルサイ

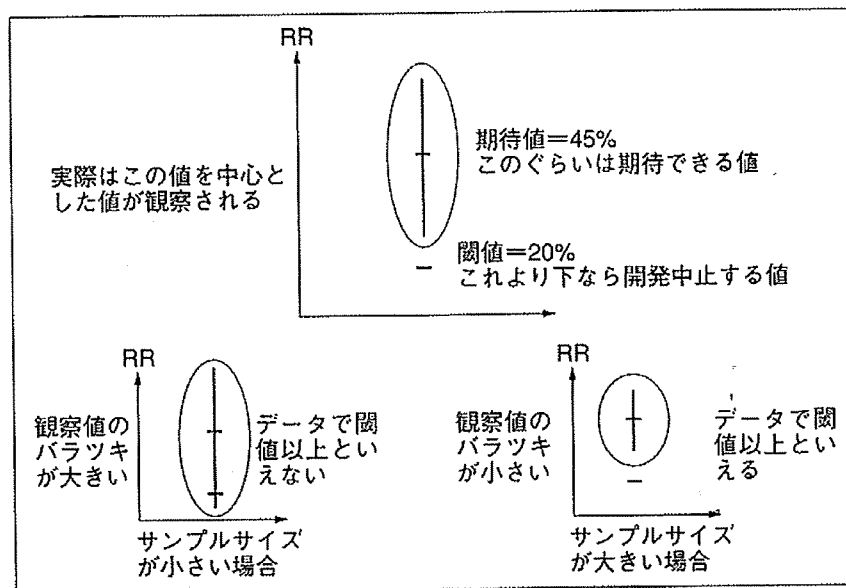


図3 閾値、期待値を用いたサンプルサイズ設計

ズと decision rule を設定するデザインである (図3)。閾値以下であるということが棄却されれば統計的には有効性が期待できるということになる。このデザインのオプションとして、試験の最後まで待たずに、途中ですでに無効が証明された場合には登録を終了する 2 stage design が用いられることも多い。

phase2 臨床試験の論文を読む場合に必要なのは、試験に実際にエントリーされた対象者の成績が、対照としている治療の成績と比較可能かどうかをチェックすることである。たとえば、ヒストリカルコントロールとの比較の場合、実際に試験にエントリーされた対象者とヒストリカルコントロールの対象者がどのぐらい比較可能かをチェックすることが重要である。適格規準に記載された対象者が必ずしも満遍なく試験にエントリーされるわけではないため、実際に試験に登録された対象者がどのような属性の分布をもっているか、それがヒストリカルなデータとの間で大きく異なっていないかを検討する必要がある。特に time-to-event (生存や無増悪生存期間) をエンドポイントとする場合には、ヒストリカルコントロールのデータがよほど

stable でないと比較できない。time-to-event のデータは、試験に組み入れられた対象者がどういふ集団かによって大きく異なるからである。より比較可能性を担保するために、ヒストリカルコントロールでなく、Phase2 試験であっても、ランダム化した対照群を設定する場合がある。この場合に注意すべきなのは、ランダム化した対照群があったとしても、サンプルサイズが小さいので検証的でなく、あくまで Phase3 へのスクリーニングと考えることである。

対照群を置くためではなく、新治療同士をランダム化比較するランダム化選択デザイン (randomized selection design) が用いられることもある。これは、患者を2つ以上の試験治療にランダムに割り付け、最も高い奏効割合が得られた治療を選択し、Phase3 試験の候補とするものである。つまり、2つ以上の試験治療がある時に優先順位をつけるデザインということができる。このデザインでは、非常に promising な治療法があった場合、その治療法を高い確率で選択できるようにサンプルサイズと decision rule を決定するものである。図4の例のように、真の response が Drug A で 35%、Drug B で 20%

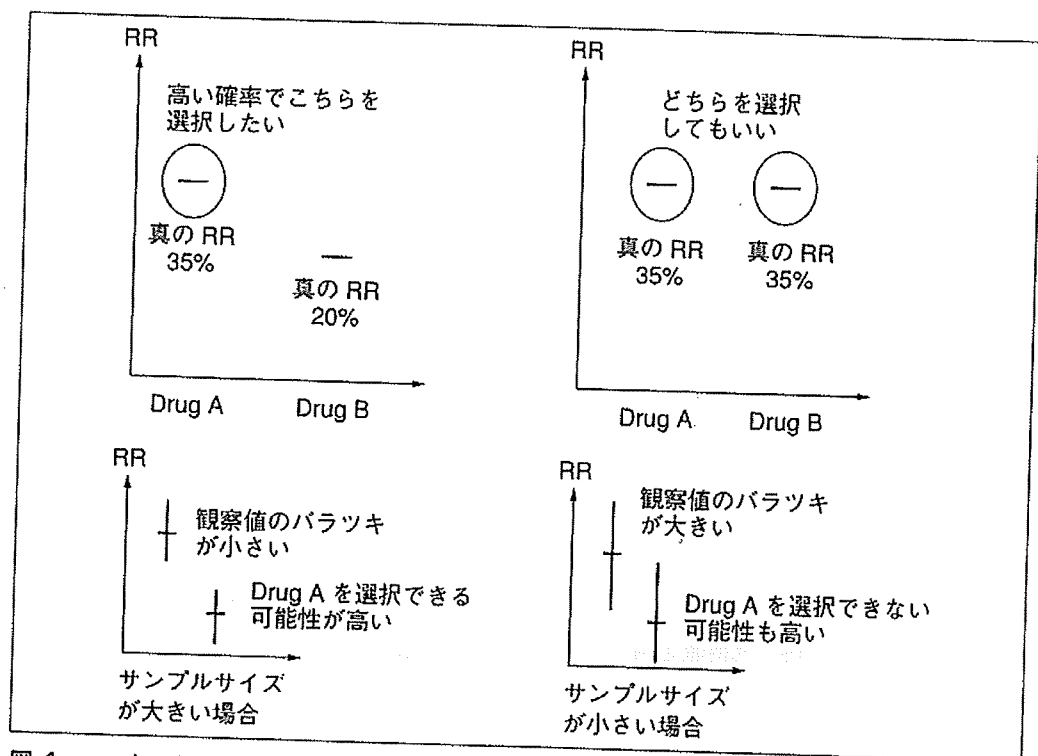


図4 randomized Phase2 デザインの考え方

である場合には、Drug A を 90% の確率で選択できるためには 29 例が必要となる。このデザインで注意すべきことは、どちらも同じ有効性をもつ治療法の場合には、50% の確率で強制的にどちらかを選択してしまうことである。差がない場合にどちらかに優劣をつけることが間違いとするならば、 α エラーが 100% といってもいいだろう。すなわち、まったく検証的試験とはいえないデザインであり、有効性に大きな差がない場合には、有効性の高い治療法を正しく選べない確率も高くなってしまふ。

繰り返すが、Phase2 臨床試験の結果によって、Phase3 へ進むかどうか（閾値・期待値）、どのレジメンを選択するか（選択デザイン）の決定するには、毒性の情報や他の開発薬剤の状況なども考慮する必要がある。統計的判断は有効性の足切りの目安として使うことが一般的である。サンプルサイズの目安として、シングルアームだとだいたい数 10 名 (<50) であることが多く、患者数が少ない疾患などで、次に

phase3 を実施することが現実的に不可能で、やむを得ず Phase2 試験に検証的な意味をもたせたい場合には 100 名以上とすることもある。ランダム化比較していると場合には、だいたい 100 名程度であり、ランダム化選択デザインだと 100 名以下であろう。いずれにしろ、検証的な結果ではないことを理解することが大切である。

Phase3 臨床試験論文を読むのに必要な統計

Phase3 の論文を読む際には、まず abstract で研究の概略と質を把握することが効率的である。研究仮説は何か、primary endpoint は何か、デザインはどうか、試験が予定通り完遂しているか、などが評価のポイントである。研究仮説として、比較する治療は何と何か、どちらの治療が標準としているかを把握する。適切でない治療を標準治療としてしまっている試験は結果の解釈が困難である。次に、primary endpoint として最も確実である全生存期間 (overall sur-

vival: OS), すなわちあらゆる死因による死亡までの期間, を用いているか(死因別死亡はあまり使われない)を調べる, progression-free survival (PFS), disease-free survival (DFS) や time-to-treatment failure (TTF) を用いている場合はそれらがその疾患領域で確立されているエンドポイントであるか知ることが重要である。そうでなければ, 必ずしも検証的な研究結果としてとらえてはいけないことになる。デザインは優越性試験か非劣性試験か, 新しい治療が toxic new なら優越性, less toxic new なら非劣性デザインを使う必要がある。非劣性試験の場合, 臨床的に意味のある secondary endpoint が定義されていて, それが標準治療に勝っていることが必要となる。試験が完遂しているか大きなポイントである。予定通りのサンプルサイズで終了しているか, 途中終了の場合, 理由は何か, また, 追跡は十分かなどもチェックするポイントである。abstract にこれらの記載がない場合は, 本文の中でチェックすることになる。

introduction では, 研究を行うに至った背景, 特に対象と標準治療の決め方, なぜその新治療が有効であると考えられるかの説明の妥当性を見る。それらを踏まえ, どのようなデザインが適当かを自分で考えてみることはよい勉強になるであろう。

methods では, どのような試験グループが行っているか, eligibility の記載は正確か, 統計記述は十分かなどをチェックする。たとえば, sample size 設計の根拠がきちんと書かれているかといったところである。優越性試験であれば, 期待する治療効果が大きすぎないかをチェックすべきであるが, これは研究の結果, 統計的な差が認められなかった場合に検出力不足のせいかどうかを検討するためである。非劣性試験であれば, 設定している許容域が広すぎるかどうかをチェックすべきで, 広すぎる場合

には, 必要以上に統計的に有意になりやすいことになる。中間解析のやり方を含め, 解析方法がきちんと書かれているかも重要なポイントである。また, データマネジメントについて記述があるかなども研究の質を推測する手掛かりになる。最近の論文では, CONSORT 声明に従った研究の流れ図が記載されている論文も多いが, これは途中で脱落した人がどのくらいいるか, 治療のコンタミネーションがどのくらい起きているか, Intention-to-Treat (ITT) 解析をしているか, などを容易に把握することができ, 研究の質を評価するのに有用である。自分でこの流れ図を書いてみるのもよい勉強になる。

results では, まず背景因子の比較が示されることが一般的である。ここでは, 重要な予後因子に関して, 群間に大きなアンバランスがないかどうかを確認する。もし大きなアンバランスがあれば, 結果の解釈の際に注意すべきであるし, 統計モデルでアンバランスな要因を調整した解析結果なども参考にすることになる。背景因子ごとに p 値が示されていることも多いが, 割り付け調整因子になっていない場合には, 20に1つは統計的に有意にアンバランスになってしまうので, p 値自体にあまり神経質になる必要はない。ランダム化で期待しているのは, 「同じ治療をすることした場合の予後が等しくなる」ことであって, 個々の背景因子のバランスではない。アンバランスがある場合でも, 最も重要な結果はランダム化に基づいた解析であり, 統計モデルで調整した解析はあくまで確認のための結果であると考えべきである。

背景因子の比較の次には, 有害事象・毒性や治療のコンプライアンス, 有効性の secondary endpoint の結果等が順に示されることが一般的である。もともと, 有害事象に関してサンプルサイズ設計を行っているわけではないため, 有害事象が両群で統計的に有意に異なるかどうかの p 値は参考程度にすべきであって, 有意かど