病、膵癌の4つで、いずれもPhase II単アーム試験である。

② 内部スタッフと委託、臨床試験における役割分担

小児がんデータセンターが行っている業務は、臨床試験プロトコール作成支援、症例登録業務、データ収集と蓄積、データの品質管理（モニタリング）、統計解析、試験結果の公表における支援などである。看護師のデータマネージャー2名、薬剤師のデータマネージャー2名、生物学士のシステム担当1名、事務2名（医長秘書兼務）を擁しているが、常勤の統計家はおらず、日本臨床研究支援ユニット（J-CRSU）へ委託する形で、統計等の専門知識を得ている。上記データマネージャーもJ-CRSUで訓練を受けた後は独学で実践を行いつつ学んでいる状況である。

ハード面は、4台のパーソナルコンピュータを使用し、SASをベースに機能する臨床試験管理ソフト「DEMAND」と進捗管理ソフト「KIWIs」を使用している。基本的には紙ベースのCRFを使用しているが、一つの臨床試験では、USBメモリ内に内蔵された電子CRFを使用して、インターネット経由のデータ授受を行うペーパーレスの品質管理を試みている。

現在の受託している4つの研究は、異なる研究費によってスポンサーされており、それぞれが独立に人材を雇用し、データセンターに集約して役割分担するという形を取っている。

上記に加えて、データセンターでは臨床試験オペレーション業務も行っている（一部は予定）。具体的には、施設管理とパフォーマンス管理、広報、印刷、中央病理診断、二次利用のための検体保存、薬事安全管理（＋研究計画管理）、財務管理、などである。これらの業務は米国の小児がんの多施設臨床試験グループであるChildren's Oncology Groupのオペレーション・オフィスを見学したことも参考になっている。

表2．小児がん臨床試験組織に必要なオペレーション機能

1) 研究グループ代表者からなる運営委員会
2) オペレーションセンター
3) 委員会事務局（会議計画、財務、出版、広報等）
4) 施設メンバー管理・パフォーマンス管理
5) 研究計画管理、薬事安全管理
6) リファレンスセンター・組織バンクの充実

③ 抱えている問題点など

上記データセンター業務・オペレーション業務のうち、統計業務や中央病理診断、検体保存などは別の研究支援組織にさらにアウトソーシングを行っている。また、小児がんの臨床試験は、疾患グループ毎に班研究がベースになっているが、いくつかの疾患グループに対しては、データセンター業務自体をアウトソーシングし、グループ間の標準化や橋渡しを小児がんデータセンターが担っているといえるものもある。

これらは、小児がんの疾患特異的な問題から発生している部分がある。例えば、稀少疾患であるため症例集積が困難、長期生存が望める故に観察期間が長期、多種多様、かつ多臓器に発生するため、試験データ管理の標準化が困難、研究組織の未熟性の問題、稀少疾患であるため研究費が僅少、といった理由である。結果として、疾患特異的・班研究ベースでグループ化が進み、独自の活動・非効率性・リソースの分断につながってしまっているともいえる。これを打開するためには、組織化・効率化された全国多施設のネットワークが必須である。これは、データセンターの充実と公的認知とともに、表2に示す各種オペレーション機能の充実によって実現されると考えられる。

5．主に研究者自主研究のサポートを業務とする研究支援会社の紹介

① 行っている研究の種類と内容

「メディカル・リサーチ・サポート」は、臨床研究者が行う自主研究のサポートを目的とした会社である。本稿著者の一人である石川が行

う臨床研究や、全国の大学、研究所に所属する研究者が実施している食品成分や薬による発癌予防試験や、新しい内視鏡や手術機材などの比較試験、適正な内視鏡検査間隔の比較や、癌術後のコホート研究などのデータマネジメントを請け負っている。現在、進行中のものが大小含めて7つ、今年中に開始予定が5つある。新薬開発などの治験に関する業務の受託は行っていない。

② 内部スタッフと委託、臨床試験における役割分担

業務内容は、データの入力や集計などのデータマネジメント業務、受診勧告などの連絡、ランダム割付、プロトコールや配付資料の作成業務、統計解析、各種委員会の設定などである。現在、スタッフの数は事務員が4人、管理栄養士が2人であり、実施する試験に応じて、事務員は遺伝カウンセラーなどの勉強も行っている。

組織に専任の研究者がいないため、試験などの管理については受託している研究に石川が共同研究者として参加し、業務の指導や管理を行っている（石川はこの会社から給料他金銭的サポートを受けていない）。利点は、きわめてスリムな組織のため比較的費用が安くてすむことである。また、石川に加え組織のメンバーがこれまで大腸領域の研究を継続して行ってきたため、これまでの経験を活かしたプロトコール作成や試験実施のサポートができることが大きな利点として挙げられる。また、商業CROと異なり利潤追求を目的としていないため、オペレーション業務を含め、契約にとらわれず臨床研究者サイドに立ったサポートを行っていることも大きな特徴といえる。

③ 抱えている問題点など

もともと、特定の研究者個人が行う臨床研究をサポートすることを主な目的として作られた組織であるため、組織のパフォーマンスがその特定個人に大きく依存してしまう部分がある。例えば、他の研究者から個人的に業務を依頼された場合、契約内容や料金設定が両者の力関係によって決まってしまったり、業務の質が個人の知識によって左右されてしまう可能性があることである。研究者サイドに立った組織であるため、委託する研究者にとって非常に有利な組織であり、依頼が殺到して特定個人に負担がかかりすぎると、それ以上の依頼を受けることができなくなる。しかしながら、将来、専任の研究者を擁する企業として発展し、なおかつ安定経営できるようになると、研究者主導臨床研究をサポートする組織の理想的な形態の一つとなるかもしれない。

6．研究者が設立した臨床研究支援のための専門組織

① 行っている研究の種類と内容

② 内部スタッフと委託、臨床試験における役割分担

NPO日本臨床研究支援ユニット（http://www.crsu.org）は、研究者主導の疫学研究・臨床研究を支援する目的で、本稿著者の一人である大橋が中心となって2001年に設立した。現在は、専任・パートのNPO職員、専任の派遣あるいは出向スタッフ、そして研究者（20%以上のエフォート）からなる混成部隊が3部門に分かれ、表3のような研究の支援を行っている。

NPOは他に教育事業（日本メディカルライター協会・SoCRA、スタッフ数3）と総務部門（2名）を抱え，事業規模は平成16年度で収入ベース約2億5千万である。支援の内容は（臨床試験の場合は）症例登録・データ管理・統計解析が中心であり、派遣モニタリング機能は持たない。研究グループに初期から参加しプロトコール作成にも積極的に関与していることが民間CROとの違いであろう。医師主導治験においては事務局業務を担当し調整役を担っている。現制度下で大きな負担となっている重篤有害事象の報告は別のCRO（J-DSRU）の協力を仰いでいる。

表３．NPO 日本臨床研究支援ユニットによる研究の支援

1) CSPOR(乳癌臨床研究支援事業)
   - 大規模補助療法試験3、ヘルスアウトカム研究7、第II相試験2、国際共同試験2(ATLAS、IBIS II)
   - スタッフ数:総括2(データセンター全体も管掌)、DM7、登録／進捗管理／入力3、解析1
2) 血液ユニット
   - 医師主導治験(造血幹細胞移植時の免疫抑制剤)2、厚労省班研究8
   - スタッフ数:総括1、DM3、登録／進捗管理／入力2
3) その他
   - 循環器大規模コホート研究と高齢者高血圧患者追跡研究(動脈硬化予防研究基金)、前立腺癌PSAスクリーニング研究、腎臓移植全例登録追跡、乳癌発症登録、骨粗鬆症大規模臨床試験(ATOP)1、臨床試験登録割付18、厚労省班研究による臨床試験(糖尿病2、小児腎症3)、膀胱癌診断研究
   - スタッフ数:16(総括1含む)

(http://www.crsu.org)

表４．単施設の臨床試験組織

1) 基本的に多施設の臨床試験組織と同じ
   ・単施設であるため多施設にわたる標準化が不要
2) データマネージメントやオペレーション業務について多科でのリソース共有
   ・オペレーション業務の透明性と公正さ
   ・業務手順の共有と標準化
3) 医師、データマネージャー、CRCなどすべての人材に対する院内教育制度
4) Electric Data Capturingや病棟からのRemote Data Entryの活用
5) 自由度の高い研究費や資金の獲得
6) CRCやDMなどの安定雇用の仕組み
7) CROなど外部組織の利用と育成

③ 抱えている問題点など

運営上の最大の課題は、研究予算に制限がある中での人材の確保と教育であり、民間 CRO との協力関係を模索している。東京大学の ARO 構想の中で、データ管理・プロジェクト管理の教育の場として機能させることが将来の目標である。

## 7．ま と め

単施設であろうと、多施設であろうと、質の高い臨床試験に必要な要素は基本的に同じである。要素によって単施設では実現しやすいものもあれば、しにくいものもあるため、多施設の臨床試験組織を参考にしながら、それぞれの要素をどのように担っていくかを実現することとなる。表４に単施設の臨床試験組織を考える際に必要なポイントをまとめた。

臨床試験にデータマネジメントが必要であるということはもはや多くの研究者の理解するところとなった。データマネジメントを行うために、研究者が自前で行うにせよ委託するにせよ、とにかく人が必要であり、そのためにこそお金が必要であるという認識もできつつある。次には、人や業務を含めてデータマネジメントの質の管理、チェック体制が必要であるということも実際に臨床研究のデータマネジメントに携わったものでは理解できるところである。

しかし、臨床試験に必要なものはデータマネジメントだけではない。様々な第三者的監視を行うための委員会機能などが必要であることも

理解されつつある。しかしこれに加えて、臨床試験推進のためには、オペレーション機能が必須である。いろいろなプロジェクト・マネジメントも含め、試験の管理も必要であるし、ヘルプデスク的な機能、参加施設との対応も必要で、これなくしては臨床試験の円滑な推進はありえない。これらの機能をよく理解し、どれを自分たちで担うのか、どれを委託するのか、委託する場合には具体的に何を委託するのかを明確に特定しなければならない。また、これらの業務を委託する組織と友好な関係を保つとともに、自らそれらを育てていく努力も必要であろう。

## 8．謝　辞

この研究は本稿著者らが属する組織、研究班での活動を通して得たものであり、本稿著者らのオリジナルというより、それらの経験を代表してまとめたものと言える。これまでの研究に協力してくれたすべての方々と、我々にこれらの経験を与えて頂いた多くの研究参加者の方に心から感謝し、この研究結果が今後の臨床研究の発展のために少しでも役立つことを祈ります。本研究は厚生労働省研究費（がん研究助成金、厚生労働科学研究費）の補助を受けた。

# Data management outsourcing for investigator-initiated clinical trials

Seiichiro Yamamoto[1], Haruhiko Fukuda[1], Tetsuya Hamaguchi[1],
Takuji Okusaka[1], Atsushi Makimoto[1], Hideki Ishikawa[2] and Yasuo Ohashi[3,4]
[1]National Cancer Center
[2]Department of Molecular-Targeting Cancer Prevention,
Kyoto Prefectural University of Medicine
[3]University of Tokyo
[4]Japan Clinical Research Support Unit

*Key Words: clinical trial, data management, single institute, contract research organization*

It has been well established that data management is an essential component in conducting clinical trials. To obtain data management of satisfactory quality, a certain amount of funding is needed, regardless of whether investigators have their own datacenter or a contract with a CRO (clinical research organization) for their data management. In addition to data management, an operations office is also necessary for smooth conduct of clinical trials. This office is responsible for the management of trials, institutions, investigators, budgets, and various projects. The principal investigators of a study should therefore familiarize themselves with all the functions necessary for clinical trials and decide who or what organization will be responsible for them.

血液・腫瘍科，54(1)：100-105，2007

Lecture

解　説

# がん治療研究におけるランダム化第II相試験の意義*

山本精一郎**

**Key Words** : molecularly targeted agent, trial design, randomization, phase II trial

## 最近の薬剤開発事情

最近，多くの薬剤が開発されており，その中には多くの分子標的薬が含まれる．これらは従来の細胞傷害性抗がん剤(cytotoxic drug)とは異なる毒性プロファイルをもつ．たとえば，単剤でのactivityが必ずしも高くない，毒性がオーバーラップしないといった特性をもつなど，標準的な化学療法との併用が効果的と考えられるケースがある．こういった薬剤の開発には従来のcytotoxic drug開発とは異なるストラテジーが必要となってきた．

これまでのcytotoxic drugでは「用量－毒性」「用量－効果」が相関関係にあり，結果として毒性と効果が比例していたことから，毒性による最大耐用量を探すことがそのまま効果を最大化する用量を推定することにつながるというロジックに基づいてphase I trialが行われてきた．それに対し，分子標的薬をはじめとするいわゆる“cytostatic drug”では，かならずしも用量と効果が比例しない．そのため，毒性の面からの最大耐用量を推定するプロセスを通して効果の面での最適用量を推定することが難しくなった．その結果，「用量－ベネフィット/リスク比」関係を検討して至適用量(optimal dose)を決めるという，一般薬開発で用いられてきた考え方と同じ方法をとる必要が出てきた．つまり，最大耐用量を探す古典的なphase Iデザインでは，至適用量が決定できなくなり，その後の開発ステージであるphase IIや場合によってはphase IIIで用量設定を行うという必要が生じてきたということである．

また，最近，phase IIIで新規薬剤と標準治療の比較を行ったが，有効性が証明されず失敗に終わったという例がいくつかあった．多くのphase IIIを実施することは通常困難であるため，失敗した場合には，その薬剤の開発が中止されるだけでなく，他の薬剤の開発の遅れにもつながる．新規薬剤がそれほど多くない時代には，phase IIIで検討の対象となる他の選択肢がないためにこれはそれほど大きな問題ではなかったが，ほかに多くの候補薬がある現在では，phase IIIの失敗は治療開発全体の遅れにつながってしまう．そのため，医療に携わるものとしてはできるだけ成功する確率の高い薬剤を選んでphase IIIを行いたい．

ではどうやって有効性のスクリーニングをすればよいのであろうか．単剤で抗腫瘍効果のみられる薬剤であれば伝統的なphase IIデザインで奏効割合(response)などの腫瘍縮小効果をエンドポイントにすればよい．たとえば，CMLやGISTに対するimatinibやNSCLCに対するgefitinibなどがその例であろう．しかし，分子標的薬の中には必ずしも単剤としてのactivityが高くなく，標準治療との併用療法として期待されているものも多い．このような場合，phase IIIへ進めるべき薬剤をどのようにスクリーニングすればよいのであろうか．

このような状況下でphase IIに求められる役割は，これまでの有効性によるスクリーニングとい

* Randomized phase 2 design for molecularly targeted agent.

** Seiichiro YAMAMOTO, Ph.D.: 国立がんセンターがん対策情報センター/JCOGデータセンター統計部門〔〒104-0045 東京都中央区築地5-1-1〕; CIS, National Cancer Center, Tokyo 104-0045, JAPAN

う目的に加え，いくつかの用量や候補薬剤，combinationの中から，phase III で成功する確率の高い薬剤や治療法を選択するということである．この場合にrandomized phase II が有効なデザインと考えられる．なぜなら，phase II で標準治療のconcurrent controlと比較できれば，ヒストリカルコントロール(historical control)を用いることによる選択バイアスを防ぐことができるため，有効性に関してsingle armの試験より高い確信が得られるからである[1)2)]．ランダム化は，新薬や分子標的薬との組み合わせで目安にすべきヒストリカルコントロールのデータがあまりない場合やresponseがエンドポイントとして適切でなく，time to progressionの方がエンドポイントとしてよい場合にはさらに有効である．なぜなら，time to progressionやsurvivalなどは，ヒストリカルコントロールと同じ選択規準で試験に組み入れられる対象者を選択しても，各被験者がもともともつ予後によって大きく影響を受ける，すなわち，治療以外の要因によってtime to eventに大きな違いが生じてしまう可能性がある．また，抗腫瘍効果は腫瘍が縮小したことを直接観察できるのに対しtime to progressionやsurvivalの延長は直接観察することができない．このようなtime to eventの欠点は，標準治療とランダム化することによって克服することができる．このように，標準治療を含めたランダム化比較をphase II で行うことは大きな利点をもつが，2群を直接比較するとサンプルサイズが倍以上になってしまう．では，どのようにすればよいのであろうか．以下，Rubinsteinらの論文[3)]をもとに，randomized phase II designについてこれまでに提案されている例をあげ，その特徴を議論してみたい．

## これまでに提案された randomized phase II design

### 1．Randomized phase II selection design

1985年にNCIのSimonら[4)]によって提案されたものが有名である．患者を2つ以上の試験治療にランダムに割り付け，もっとも高いresponseが得られた治療が次の研究に進むというデザインである(エンドポイントはほかのものでもよい)．2つの試験治療があるときに優先順位をつけるデザインであって，もっともよい群が高い確率で選択されることを保証するようにサンプルサイズを決定する．たとえば，真のresponseがレジメンAで35%，レジメンBで20%のとき，90%の確率でレジメンAを選択するためには29名が必要，といった具合である．この場合の問題点は，「どちらも同じ場合には50%の確率でAを選択!!!」することであり，その意味でαエラーが50%ともいえるかもしれない．このデザインはdoseやscheduleを選ぶためにも用いることができる．前提として，毒性の程度がほぼ同じことを仮定しているが，毒性が異なる場合にもその分効果に下駄を履かせたdecision ruleを作ることも可能である．しかし，異なる治療薬の選択の場合には本当にこれでよいのかという不安がある．というのも，2群の治療効果がまったく同じでも，50%の確率でいずれかの群を選んでしまうためである．また，観察されたresponseのみで選択するため，とくに事前にどちらがよいかの根拠がまったくない場合だけ使えるともいえる．さらに，試験の対象となる薬剤が異なる製薬企業によって開発されているものであった場合，製薬企業がよいといえばこのデザインによる優先順位づけはできるかもしれないが，大きなコストをかけて開発が行われていることを考えると，このような単純な判断基準が受け入れられるとは考え難い．

Randomized phase II selection designの別の方法として，患者を2つ以上の試験治療に割り付け，それぞれヒストリカルコントロールと比較するというものもある．群ごとにサンプルサイズ計算をするので，群の数だけphase II studyをやるのと同じサンプルサイズが必要であるが，効果と毒性の両方から判断してよい方を次の段階へ進めるという判断が下せる．これは，効果と毒性のプロファイルがまったく違う場合に有効と考えられるが，ヒストリカルコントロールとの比較はあまり信用できないため，選んだ薬剤がphase III で既存治療に勝る確率が高いという確証にはつながらない．

### 2．Reference arm phase II design

Hersonら[5)]は，試験治療群と標準治療群にランダム化するが，サンプルサイズの増大を防ぐため，直接比較はしないというデザインを提案し

た．標準治療群はヒストリカルコントロールの対象者と試験の対象者が同じ集団かどうか(比較可能かどうか)をチェックするためだけに用い，新治療群はヒストリカルコントロールと比較するというものである．標準治療群の成績がヒストリカルコントロールの成績と変わらなかった場合は，試験治療群とヒストリカルコントロールを比較することに問題ないが，標準治療群が予想より低かったり高かったりした場合はどのように比較するかが問題となる．その場合にはもう一度研究を行えばよいとオリジナルの著者らは言っているが，これは現実的ではないし，もう1回やったからといって比較可能な成績が出るという保証はない．このように，ヒストリカルコントロールと同じかのチェックが失敗すると試験結果が解釈しづらくなるといえる．

3．Phase II/III design

標準治療と新治療でランダム化phase IIをやり，その結果に基づいてそのままphase IIIに進むデザインをrandomized phase II/III designと呼ぶ．いろいろなバリエーションがありえるが，共通していえることとして，同じ対象者を使えるのでサンプルサイズの節約になる一方，phase IIIに進まなかった場合にはphase IIIの準備が無駄になる，といった特徴がある．ここでは，提案されているデザインをいくつか具体的に紹介する．

Schaidら[6]は，標準治療Aとそれと比較したい新治療B，さらにはphase II試験をしたい新治療C1，C2，C3...でランダム化比較試験を行い，Cに失敗したらA or Bにさらにランダム化して組み込むというデザインを提案した．このデザインの特徴は，無治療の患者をphase IIに組み込み，失敗したらmore established therapyであるA or Bへ治療を変更するというものである．しかし，無治療の患者をphase IIに入れることに対する倫理的な問題や，前治療としてのCの影響をどう取り扱うかといった問題があり，あまり現実的でないと思われる．

また，Storerら[7]は，後継phase IIIと同じランダム化をphase II部分で行うというデザインを提案している．オリジナルの点としては，新治療をヒストリカルコントロールと比較し，よければphase IIIへ，よくなければphase IIとして終了するというものである．Phase IIの段階ではランダム化した標準治療群との比較を行わないので，ヒストリカルコントロールと比較することの問題点は解消されていない．Phase IIIが終わって，ふたを開けてみたら負けていた，ということも十分ありえる．

Ellenbergら[8]は，phase II部分はresponseで比較，少しでも勝っていたらphase IIIへというデザインを提案した．これは，誤ってphase IIで止めてしまう確率が十分低くなるようにして，phase IIIでの検出力低下を防ぐことを意図したデザインである．Phase IIIのサンプルサイズの1/3位，通常のsingle arm phase IIの2倍くらいのサンプルサイズを提案している．このデザインではresponseがprimary endpointである生存期間のよいサロゲートエンドポイントである場合でも，帰無仮説が正しい，すなわち両群に差がない場合にも50％の確率でphase IIIへ進んでしまうので，スクリーニング目的としてはあまり効率がよいとはいえない．

Schaid[9]，Scherら[10]は，複数の新治療群があっても行うことができるrandomized phase II/III designを提案した．Phase II部分も標準治療群と真のエンドポイントである全生存期間(overall survival；OS)で比較し，ある規準を超えていればその群のみphase IIIへ進むというもので，phase II部分は全体の半分程度のサンプルサイズという提案を行っている．検定の多重性の問題が生じるためαエラーを増大させないためには，なんらかの統計的方法を組み込む必要があるだろう．

私が所属するJCOG(Japan Clinical Oncology Group)でも，実際のいくつかの試験でphase II/III designを採用している．1種類のデザインを用いているのではなく，試験の内容によりさまざまなデザインを採用している．たとえば，phase IIからphase IIIへ進むための判断規準として，

- 新治療群の安全性だけを確認してそれが規準を満たしていればphase IIIへ進む．
- 標準治療群に対し有意に劣っていないことのみ証明できればphase IIIへ進む．
- 両群別々にヒストリカルコントロールと比べて劣っていないかを検証してphase IIIへ進む．

表1 PFSをエンドポイントとした場合のランダム化スクリーニング試験に必要な総イベント数

| | 検出したいハザード比 Δ（化療 vs. 化療＋分子標的薬） | | | |
|---|---|---|---|---|
| | Δ=1.3 | Δ=1.4 | Δ=1.5 | Δ=1.75 |
| (α, β)＝(10%, 10%) | 382 | 232 | 160 | 84 |
| (α, β)＝(10%, 20%) | 262 | 159 | 110 | 58 |
| (α, β)＝(20%, 20%) | 165 | 100 | 69 | 36 |

（文献[3]より改変）

表2 レスポンスをエンドポイントとした場合のランダム化スクリーニング試験に必要な全体の患者数

| | 検出したいレスポンスの差（化療 vs. 化療＋分子標的薬） | | | |
|---|---|---|---|---|
| | 20% vs. 35% | 40% vs. 55% | 20% vs. 40% | 40% vs. 60% |
| (α, β)＝(10%, 10%) | 256 | 316 | 156 | 182 |
| (α, β)＝(10%, 20%) | 184 | 224 | 112 | 132 |
| (α, β)＝(20%, 20%) | 126 | 150 | 78 | 90 |

（文献[3]より改変）

といったようにphase IIIでのαエラーを消費しない仕組みを組み込んでいる．いますぐにでもphase IIIを行いたいが，その前に有効性や安全性の再確認をしておく必要があるといった場合に用いることが多い．そのため，phase II部分で検証的な解析を行うことには重きを置いておらず，phase II部分でαエラーの増大が生じないよう注意を払っているといえる．これらは統計的にかなりの応用問題であり，統計家が関与しても厳密には解決しつくしていない部分が残るが，サンプルサイズの節約とphase IIIを計画する時間の節約がcriticalな場合には，採用する余地があると考えられる．

Phase II/III designをまとめると，メリットとして，①phase II部分のデータをphase IIIでも使える，②phase II終了後phase IIIをすみやかに開始できる，③phase II部分のエンドポイントにはflexibilityがある，④標準治療群をconcurrent controlを使うのでphase II段階での比較の妥当性が上がる，などがあげられる．デメリットとして，①スクリーニングツールとしての性能はよくない，②通常のphase IIに比べてphase II部分のサンプルサイズが大きい，③phase IIIを実施するためのインフラと準備が必要，④phase III部分の検出力を保つためにはphase II部分のα(＝controlよりたいしてよくないのに間違ってphase IIIに進んでしまう確率)が大きくなる，⑤phase II部分で早く判断できるエンドポイントがなければphase IIの結果が出るまで中断しなければならない，といった点があげられる．したがって，phase II/III designはeffective screeningというより，aggressive interim monitoringをもつphase IIIと考えた方がよいであろう．

4. Randomized phase II screening design

最近提案されたものにランダム化スクリーニングデザイン(randomized phase II screening design)[3]と呼ばれるものがある．これは，標準治療と新治療を直接に，しかし決定的ではない形(nondefinitive)で比較するものであり，標準治療vs. 標準治療への分子標的治療のadd onなどといった比較を念頭において提案されたデザインである．これは，ヒストリカルコントロールとの比較による選択バイアスの影響を避けるために標準治療群とランダム化し，かつ，試験の規模は2群で50～100名くらいのサンプルサイズにしたいという欲求を満たすものである．そのためには，治療効果の差に対してαエラー，βエラーを注意深く設定することによってサンプルサイズ設計を行うことが必要となる．また，決定的な結論を導くことのできるphase III試験ができなくなることがないよう，OSとは異なる，より早期のエンドポイントを使うなどといった工夫が必要である．たとえばPFSやresponse, %PFSなどといったものである．ちなみに，ある時点で全員のprogressionを測定できる場合には，その時点での%PFSの方が観察間隔の影響を受けやすいPFSに比べて観察によるバイアスが少ないといえる．

PFSをエンドポイントとした場合のランダム化スクリーニング試験に必要な患者数を表1，responseをエンドポイントとした場合のランダム化スクリーニング試験に必要な患者数を表2に示す．α>20%，β>20%は薦められないので，それ以下で考えても，両群合わせて100名程度あ

るいはそれ以下で設計できる組み合わせがある．つまり，治療効果の差に対して，α，βを注意深く設定することによって100以下のサンプルサイズ設計が行えるケースがある．たとえば，治療効果をハザード比で表した表1より，Δ＞1.4であればα，βがともに20％の場合必要イベント数は100以下となる．αが大きすぎるとスクリーニング能力がない，つまりphase IIを実施せずにphase Iから一足飛びにphase IIIを実施することと同じになってしまい，逆にβが大きすぎると有効な薬剤の開発を誤って中止してしまうことになる．過度に楽観的な治療効果の差Δを考えることもβを大きくすることと同じである．これらから，スクリーニングにはHR＝1.5（レスポンスでは差が20％）に対してα＝β＝0.2程度が一つの目安といえるだろう．

ランダム化スクリーニング試験の他の特徴として，同じ薬剤を用いて実施されるスクリーニング試験の数が多くなるとfalse positiveの確率が高くなることがあげられる．たとえば，同じ薬剤を用いて異なる3つの対象に試験を実施したとすると，α＝β＝0.2の場合には，その薬剤が3つの対象すべてに治療効果がある場合に1つも治療効果があることを見出せない確率が5％以下となる一方，3つの対象すべてに治療効果がない場合に少なくとも1つの対象で効果ありと判断してしまう確率は約50％となってしまう．

ランダム化スクリーニングデザインでの重要な点は，得られた結果がdefinitive phase IIIから得られたものとは違うことをしっかり意識するということである．とくにOSをエンドポイントにしているときは，研究者がsubstantial evidenceと思ってしまうことが問題である．そのため，PFS, response rate, %PFSといった，できるだけOSと異なるエンドポイントを用いることも必要であろう．このようなデザインで実施された試験結果をdefinitive phase IIIと考えてはならないのは，何度もランダム化スクリーニング試験を実施する状況では検定の多重性の問題と同様の状況が生じ，α＝0.05が判断の基準として大きすぎるためである．Phase IIIをやるにはかなりの証拠が必要だが，phase IIはそうでもない．したがって，通常のphase II-phase III sequenceと同じような証拠を与えるためには，phase IIIの中間解析と同じように考えて，たとえばα＜0.005などといった極端な結果が得られた場合にのみsubstantial evidenceと考え，そのあとのphase IIIを行う必要がないといった判断の方針を採用するのがよいかもしれない．また，サンプルサイズが小さいことを考えると，有意な結果が得られたとはいっても大きな治療群間差が大きな信頼区間を伴っているはずであり，より精確な推定値，すなわちより精確な治療効果の見積もりを得るためにも，さらなる研究が必要といえる．

具体的に考えてみよう．ランダム化スクリーニングデザインが使えるシナリオとして，

(1)ある製薬企業は，多くの臓器で標準化療に対してadd onできるような分子標的薬Xを開発中であるが，薬剤Xは単剤ではinactiveである．しかし，薬剤X＋標準化療をヒストリカルコントロールと比較するsingle arm phase IIはあまり信頼できない．このような場合には，複数のスクリーニング試験を行うというストラテジーが提案できるだろう．それぞれの臓器で実施する一つ一つの試験に対して比較的大きなβエラーを設定しても，どれかの試験では差を検出できる可能性が高まるからである．そのほかのシナリオとして，

(2)研究者はある特定の疾患の標準化療にadd onできるような分子標的薬$Y_1, Y_2,...$に優先順位をつけなければいけないが，single arm phase IIは毒性やfeasibility程度のデータしか与えないし，薬剤の開発状況の違いや製薬企業のreluctanceによって，複数の分子標的薬のrandomized selection designはやり難い．そこで，複数のスクリーニング試験を行うというストラテジーが提案できるかもしれない．この場合も，それぞれの試験に対する比較的大きなβエラーが許容できるからである．

## まとめ

本稿ではさまざまなrandomized phase II designについて議論してきた．最後にphase IIでどのデザインを選ぶかを簡単にまとめてみたい．

①Responseで評価が可能な場合はconventional

phase II design.

②複数の試験治療に優先順位をつけたい場合randomized selection phase II design.

③次にphase III に進むことがほぼ確実な場合はphase II/III design.

④ヒストリカルコントロールとの比較が行いにくい場合はrandomized screening design.

とくに新規分子標的薬を標準化療にadd onする場合には，definitive phase III を実施可能でかつ研究者が治療効果とαエラーとβエラーをバランスと分別をもって設定できるのなら，ランダム化スクリーニングデザインはよいデザインといえるだろう．

もちろん，デザインが先にあるわけではなく，治療開発全体のストラテジーの中で適切なデザインを選ぶということが基本である．最近randomized phase II が多く議論されているということは，数少ないが一発逆転狙いのcytotoxic drugの開発という時代から，単剤での治療効果は必ずしも高くないが，分子標的薬という数多くの候補薬剤の中から有効なものを選んで開発していくという時代へと世の中が変わっていることを意味している．これは製薬会社や臨床研究者のみならず，臨床試験に携わる統計家にとっても大きなチャレンジといえる．ぜひ協力してよい治療を開発していきたいと思う．

## 文　献


1) European Organisation for Research and Treatment of Cancer. Phase II trials in the EORTC. The Protocol Review Committee, the Data Center, the Research and Treatment Division, and the New Drug Development Office. Eur J Cancer 1997 ; 33 : 1361.

2) Van Glabbeke M, Steward W, Armand JP. Non-randomised phase II trials of drug combinations : Often meaningless, sometimes misleading－Are there alternative strategies? Eur J Cancer 2000 ; 38 : 635.

3) Rubinstein LV, Korn EL, Freidlin B, et al. Design issues of randomized phase II trials and a proposal for phase II screening trials. J Clin Oncol 2005 ; 23 : 7199.

4) Simon R, Wittes RE, Ellenberg SS. Randomized phase II clinical trials. Cancer Treat Rep 1985 ; 69 : 1375.

5) Herson J, Carter SK. Calibrated phase II clinical trials in oncology. Stat Med 1986 ; 5 : 441.

6) Schaid DJ, Ingle JN, Wieand S, et al. A design for phase II testing of anticancer agents within a phase III clinical trial. Control Clin Trials 1988 ; 9 : 107.

7) Storer BE. A sequential phase II/III trial for binary outcomes. Stat Med 1990 ; 9 : 229.

8) Ellenberg SS, Eisenberger MA. An efficient design for phase III studies of combination chemotherapies. Cancer Treat Rep 1985 ; 69 : 1147.

9) Schaid DJ, Wieand S, Therneau TM. Optimal two-stage screening designs for survival comparisons. Biometrika 1990 ; 77 : 507.

10) Scher HI, Heller G. Picking the winners in a sea of plenty. Clin Cancer Res 2002 ; 8 : 400.


*　　*　　*

# 2 サブグループに対する治療開発のための臨床試験デザイン

山本精一郎（国立がんセンターがん予防・検診研究センター情報研究部JCOGデータセンター統計部門）

**Point**

- 個別化治療とは、何らかのマーカーに基づいてサブグループを定義し、そのグループに対して効果のある治療を行うことである。個別化治療であっても、それが有効であるかどうかは臨床試験による検証が必要となる。
- PhaseⅢ Trialでall comers design、

(a)ある薬剤が対象者全体の全生存期間を延長するような標準治療となり得るか

だけでなく

(b)試験薬投与前に腫瘍組織で調べた薬剤に対する反応(マーカー)によって治療効果が予測できるかを主要な目的として研究デザインすることも可能である。

- マーカーが高いサブグループほど治療効果が大きい場合(linear仮説)や、マーカーが高いサブグループにだけ治療効果がある場合(threshold仮説)、どのサブグループに新治療を導入するかは、そのサブグループに対する治療効果がどの程度あるかということに依存する。

分子標的薬などの開発が進むにつれ、個別化治療という言葉をよく耳にするようになった。個別化治療とは、何らかのマーカーに基づいてサブグループを定義し、そのグループに対して効果のある治療を行うことである。個別化治療であっても、それが有効であるかどうかは臨床試験による検証が必要となる。個別化治療の臨床試験デザインとして、大きく分けて図1の3つが提案されている[1]。このうちのall comers designについて、本稿では深く掘り下げてみたい。

All comers designは、まずはじめにマーカーを調べるが、マーカー結果によらず、ランダム化

 0910-0474/06/￥400/論文/JCLS

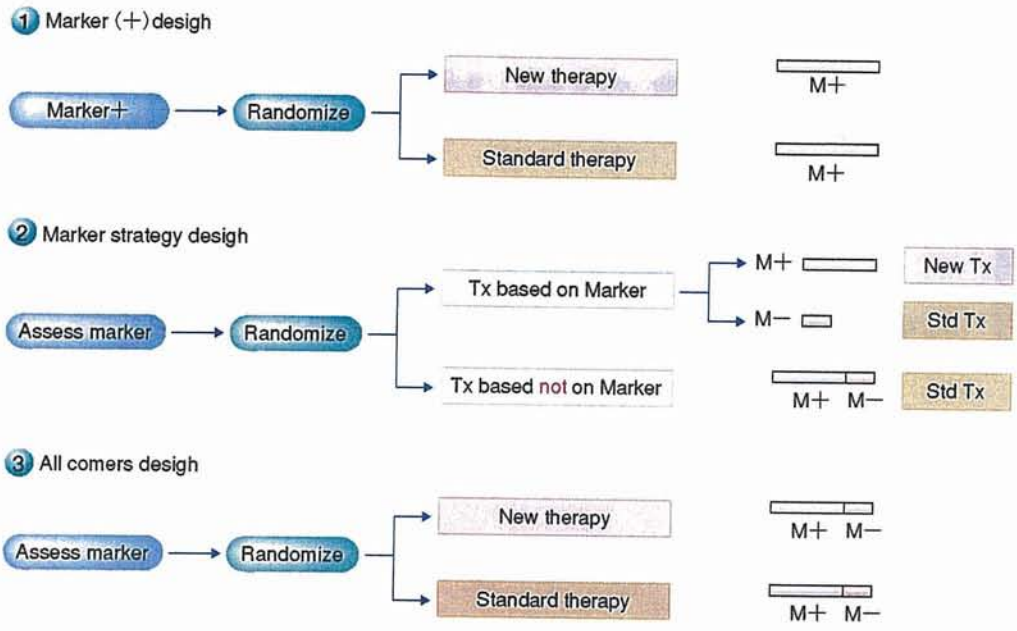



**図1●個別化治療のための臨床試験デザイン**

して一方には新治療、もう一方には標準治療を行うものである。マーカー結果によらず両群を比較することによって、対象集団全体に対し新治療と標準治療の治療効果のプライマリな比較ができるだけでなく、マーカー(+)群、(−)群のそれぞれで新治療の治療効果を調べることができる。つまり、マーカー結果と治療の間に交互作用があるかどうか、言い換えるとマーカーが治療効果の予測因子かどうかを調べることができる。全体で治療効果をみた後に単にサブグループアナリシスをするのではなく、さらに進めて、検証的なサブグループアナリシスを組み込むにはどうすればよいか、感受性仮説などのproof of conceptをどのように行っていけばよいか、について統計的な試みがいくつか提案されているので、本稿ではそれらの方法を議論することによって、個別化治療の臨床試験デザインについて理解を深めていただきたいと思う。

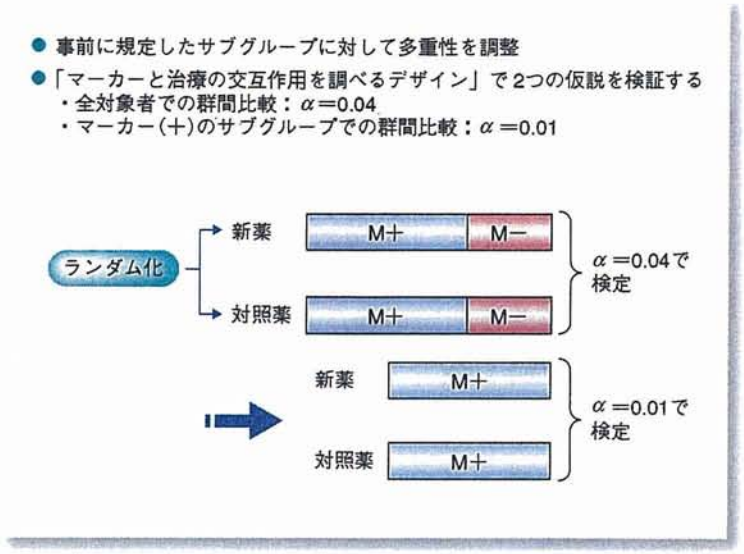
Simon[2]は、事前に規定したマーカー結果によって定義されるサブグループでの検定に対して、aエラー[*1]のコントロール、つまり検定の多重性[*2]を調整するようなデザインを組めばいいの

*1　αエラーとβエラー
統計的検定において、帰無仮説が正しいときに誤ってこれを棄却してしまう確率をαエラーとよぶ。αエラーの水準は5%とすることが慣習的である。また、対立仮説を誤って採択できない確率をβエラーとよぶ。1−βエラーが検出力となる。

*2　検定の多重性
1つの研究のなかで検定を何度も繰り返すとαエラーが大きくなることをいう。検定の多重性を調整して試験全体のαエラーを名義水準通りに抑えることを多重性の調整という。臨床試験における検定の多重性とその調整の例として、中間解析と最終解析における複数回の検定、複数のエンドポイントの検定、などが挙げられる。

**図2●Simonの提案**

(Simon R, JCO 2005. より引用)




ではないかという提案を行った(図2)。これはマーカーと治療の交互作用を調べるデザインで、2つの仮説を検証する。つまり、はじめに全対象者をランダマイズして、2群間比較の検定を行い、さらにマーカー(+)だけで群間比較の検定を行う。ポイントは、全対象者について行う検定は$\alpha=0.04$でやっておいて、そのあとマーカー(+)のサブグループだけで検定するときには$\alpha=0.01$で行うというものである。この論文のなかでは、厳密な統計的な議論はしておらず、簡単な提案のみ行っている。

これをさらに進めたものとして、Freidlin & Simon[3]は"Adaptive Signature Design"という方法を提案している(図3)。この方法は、マーカーを試験開始前に特定できない場合に特に有効である。この方法は次の3ステップからなる。全員をランダマイズし、新治療群と対照群に分けて治療を行うことは同様である。

a) まず、割付を行った全対象者に対して$\alpha=\alpha_1$ ($<0.05$)で2群間比較をする。全体の対象者で差があるかをみる部分である。

b) 試験前半(1st stage)で登録された対象者のデータに対し、さまざまなマーカーを用いて治療効果の異なるサブグループを探索的に探し出す(training set)。次に後半の対象者(2nd stage)のなかから同一のサブグループを抽出する(test set)。すなわち後半の対象患者のなかでこの薬剤に対して"感受性の高い"と考えられる対象者のサブグループを同定する。

c) サブグループを探索した1st stageのデータと独立である2nd stageのサブグループを用いて治療効果が異なるかどうかの検証を行う。すなわち、2nd stageのサブグループに対し、$\alpha=\alpha_2$で治療効果の群間比較の検定を行う(ただし、$\alpha_1+\alpha_2=0.05$となるように$\alpha_1$、$\alpha_2$を定める)。

彼らは全対象者に対する検定を行う際の$\alpha_1$を全体の$\alpha$の80%、つまり$\alpha=0.05$であれば$\alpha_1=0.04$とし、探し出した"感受性の高い"と思われるサブグループに対する検定を行う際の$\alpha_2=0.01$とすることを勧めている。この方法を用いると、試験全体の$\alpha$エラーは0.05で抑えることができる。$\alpha_1=0.04$にすることによって、全体の対象者

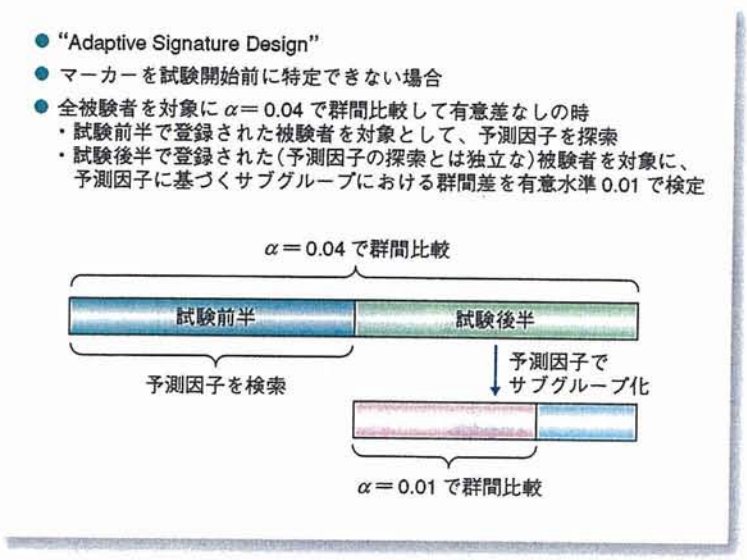



**図3 Freidlin & Simonの提案**
(Freidlin B & Simon R, Clin Cancer Res 2005. より引用)

に対する検出力は少し落ちるが、2nd stageの"感受性の高い"サブグループに対する検定を加えることによって、試験全体の検出力は増加することが期待される。"感受性が高い"(＝十分な治療効果が期待できる)ことが本当であれば、そのサブグループでの群間差は全対象者での群間差よりも大きくなることが期待されるため、$\alpha_2=0.01$でもそれなりの検出力をもつであろうということである。また、$\alpha_1=0.04$で行う全対象者に対する検定の検出力を$\alpha=0.05$で検定する際の検出力と同様となるようサンプルサイズを多少増加させることもできる。その場合には"感受性の高い"サブグループに対する検定を加えた2つの検定を行うことによって、通常のデザインよりも有効な薬剤を誤って捨てる確率は少なくなることが期待できる。新規薬剤が多くの対象者に効果がある場合には、通常のデザインと同じ検出力をもち、マーカーで規定される少数のサブグループにしか効果がない場合にもそのサブグループでの感受性が高く治療効果が大きければ群間差を検出でき、試験全体としての検出力の増加が見込めるためである。

National Cancer InstituteのこのSimonらのグループは、microarrayの解析やpharmacogenomicsの解析に積極的に取り組んでおり、そのなかでどのようにして予測因子を探索するかということに多くの経験をもっている。1st stageの対象者のデータを用いてさまざまなマーカーによって規定される感受性の高いサブグループを探索的に探す、というアイデアはその中から出てきたものだと思われる。感受性の高いサブグループが見つかるかどうか、そのサブグループの割合がどうかといった結果に依存する部分や、1st stageと2nd stageのサンプルサイズ比をどう設定するか、など試験計画のデザイン時に最適化できない部分も多いが、試験開始前に感受性の高いサブグループが十分同定できない場合には有効な方法であると考えられる。

次に筆者が所属する日本臨床腫瘍研究グループ(Japan Clinical Oncology Group；JCOG)で現在計画されている研究デザインを紹介する。これは進行癌に対する化学療法の効果を調べる

表1●薬剤感受性に関する仮説

| 治療と感受性の関係を3レベルで考える | |
|---|---|
| 一様な効果(交互作用なし)uniform effect<br>・薬剤の治療効果は感受性によらない |  |
| 閾値モデル threshold model<br>・薬剤の効果はマーカー値がある値を超えると発現する |  |
| 直線モデル linear model<br>・薬剤の効果はマーカー値と直線的に関連する |  |
| **ポイント** | |
| 治療とマーカーの交互作用として感受性仮説を検討 | |
| 想定した感受性仮説が正しければサンプルサイズは減少する | |
| **感受性仮説の証明と臨床的意義は別** | |
| 帰無仮説は、治療効果なしと交互作用なしの両方が定義できる | |

phase Ⅲ trialを計画する際に検討されたデザインである。研究の主要な目的は、

a) ある薬剤が対象者全体の全生存期間を延長するような標準治療となり得るか。

であるが、

b) 試験薬投与前に腫瘍組織で調べた薬剤に対する反応(マーカー)によって治療効果が予測できるか。

ということも大きな関心である。もちろん、"腫瘍組織で調べた薬剤に対する反応"自体は単なるマーカーに過ぎず、生存期間延長といった治療効果と必ずしも一致するわけではない。bの仮説、すなわち"マーカーの値によって、治療効果が異なる"ことを本稿では"感受性仮説"とよぶことにする。

例となった実際のJCOG試験では、主要な目的であるa) を優先させるために、通常の2群間比較を数百例の規模で行い、副次的な目的として(多重性の調整は行わず)、b) の仮説を調べることとした。ここでは思考実験として、bの"感受性仮説"を主要な目的として検証することを考えてみる。治療への"感受性"との関連を期待するマーカーの値が連続的であり、3段階に分けて治療効果との関係を調べることとする(**表1**)。3段階に分けると、さまざまな感受性仮説が検証しやすいからである。感受性仮説の例を以下に挙げる。

①linear仮説:マーカー値が「低」「中」「高」となるにつれ治療効果が大きくなる。

②threshold仮説1:マーカーが「中」以上であれば治療効果が発揮される。

③threshold仮説2:マーカー「高」であれば治療効果が発揮される。

感受性がない場合にも以下の2つの仮説が考えられる。

④uniform仮説:マーカー値が「低」「中」「高」に対し、治療効果が一定。

⑤全体の帰無仮説:どのサブグループに対して治療効果がない。

通常、臨床試験で行われる検定は、"⑤全体の帰無仮説"に対して"④uniform仮説"を対立仮説として行われる。繰り返しになるが、統計的に言うと感受性仮説は治療とマーカーの生存期間に対する交互作用のことを指す。つまり、治療効果があっても感受性がないことを示すuniform仮説は、統計的な交互作用がないことを示す。

| もし感受性仮説が証明されたら、新しい標準治療の決定につながるのか？ |
|---|
| 一様な効果（交互作用なし）<br>・新治療がすべての患者さんに対し標準治療となる |
| 閾値モデル or 直線モデル<br>・それぞれのマーカーサブグループに新治療を導入するかはそれぞれのサブグループでの治療効果の大きさしだい |
| 少なくとも「治療効果はマーカー値によって異なる」という仮説は証明された<br>・標準治療とするためには、次のステップとして新たなRCTが必要 |

表2●感受性仮説に基づいて標準治療を決定

uniform仮説でなく、治療と感受性の交互作用を示す（上記のような）感受性仮説を対立仮説とすると、感受性有り無しについての検証を行うことができる。帰無仮説は棄却したい仮説により、④とすることも⑤とすることもできる。

感受性仮説と臨床的意義について考えてみよう。例えば⑤を帰無仮説として、上に挙げたそれぞれの仮説が証明された場合、次の標準治療はどうなるのだろうか（**表2**）。uniform仮説が証明された場合は、もちろん新治療がすべての患者さんに対して標準治療になる。次にlinear仮説が証明された（マーカーが高いサブグループほど治療効果が大きい）場合や、threshold仮説が証明された（マーカーが高いサブグループにだけ治療効果がある）場合、どのサブグループに新治療を導入するかは、そのサブグループに対する治療効果がどの程度あるかということに依存する。つまり、感受性仮説が証明されただけでは、あるサブグループに対して十分に臨床的に意義のある治療効果があることを証明したことにはならない。標準治療となるためには、感受性仮説のみならず、それぞれのサブグループに対して臨床意義clinical significanceがあるかどうかを検証することが必要となる。

今回のJCOG試験では、感受性仮説の検証を副次的な目的とした。副次的な目的であっても治療効果が腫瘍に対する薬剤の反応によって異なるというproof of conceptができれば、次のステップとしてそれを取り込んだ治療開発を行うことができる。仮説を立てるために利用するデータとそれを検証するためのデータは独立であるべきという科学の一般論からも、感受性のあるサブグループを探索するステップと臨床的意義があるかどうかを検証するステップは分けて考えるべきである。具体的には、以下のようなデザインが考えられる。

①：phaseⅢで副次的にproof of conceptを行い、次のphaseⅢで検証する。
②：前述のSimonたちのAdaptive signature designのように、1つの試験を感受性のあるサブグループを探索する部分と、それを検証する部分に分ける。
③：標準治療をアームに含んだrandomized phase Ⅱ design（例えばランダム化スクリーニングデザイン[4]）で感受性の探索やproof of conceptを行い、次のphaseⅢで検証する。
④：randomized phaseⅡ/phaseⅢを用いて、③のrandomized phaseⅡとphaseⅢをシームレスに行う。

本稿では、SimonやSimon & Freidlinの提案や、われわれのアイデアを紹介した。さらに進んだ方法として、linear、threshold、uniformといった複数の対立仮説から最もデータにフィットした対立仮説を選択するという統計的方法も考え

られる。あるいは、uniformとlinearを組み合わせることによって、より広い対立仮説に対して高い検出力を保つ方法も考えられる。統計的にはいろいろな工夫が可能であるが、臨床的な意義や行いたいproof of conceptを十分に反映させるようなものにしなければいけない。意義のある仮説とそれを効率よく検証できるよう試験をデザインするためには、今までにも増して臨床家と統計家のコラボが必要と言えるだろう。

## 謝辞

本稿の内容は筆者が所属するJCOGデータセンターのメンバーやJCOGの研究者との議論のなかで整理され、考え出されたものであり、その方々との共作といえる。本研究はがん研究助成金指定研究（17指-5）の補助を受けた。

## 文　献

1) Sargent DJ, 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.
4) Rubinstein LV, Korn EL, et al. Design Issues of Randomized Phase II Trials and a Proposal for Phase II Screening Trials. J Clin Oncol 2005; 23: 7199-206.

# Effect of the introduction of minimum lesion size on interobserver reproducibility using RECIST guidelines in non-small cell lung cancer patients

Hirokazu Watanabe,[1,6] Hideo Kunitoh,[2] Seiichiro Yamamoto,[5] Shin Kawasaki,[2] Akira Inoue,[2] Katsuyuki Hotta,[2] Kazu Shiomi,[3] Masahiko Kusumoto,[4] Kazuro Sugimura[1] and Nagahiro Saijo[2]

[1]Department of Radiology, Kobe University Graduate School of Medicine, 7-5-2 Kusunoki-cho, Chuo-ku, Kobe 650-0017; [2]Department of Medical Oncology, [3]Thoracic Surgery, and [4]Radiology, National Cancer Center Hospital; and [5]Research Center for Cancer Prevention and Screening, Statistics and Cancer Control, National Cancer Center, 5-1-1 Tsukiji, Chuo-ku, Tokyo 104-0045, Japan

(Received September 17, 2005/Revised November 24, 2005/Accepted November 28, 2005/Online publication February 16, 2006)

**We evaluated interobserver reproducibility for the response evaluation criteria in solid tumors (RECIST) guidelines and the influence of minimum lesion size (MLS) on reproducibility. The 110 consecutive patients with non-small cell lung cancer were treated with platinum-based chemotherapy. Five observers measured target lesions according to both the World Health Organization (WHO) criteria and RECIST. The percentage changes for unidimensional measurements (UD; RECIST type) and bidimensional measurements (BD; WHO type) were calculated for each patient. Interobserver reproducibility among five observers, that is 10 pairs, was expressed as the Spearman's correlation coefficient for the percentage changes, the proportion of agreement and the kappa statistics for response categories. The same analysis was carried out using MLS. BD was more reproducible than UD (Spearman rank correlation coefficient, 0.84 vs 0.81; proportion of agreement, 84.4% vs 82.5%; kappa value, 0.69 vs 0.61). When MLS was applied to UD, eligible cases decreased by 6.4% and the number of target lesions by 44.6%, whereas interobserver reproducibility for UD improved (Spearman rank correlation coefficient, 0.81–0.84; proportion of agreement, 82.5–84.2%; kappa value, 0.61–0.65). The introduction of MLS to UD could also improve intercriteria reproducibility between WHO and RECIST. It is important to apply the MLS when using RECIST for the comparable interobserver reproducibility attained with WHO. (*Cancer Sci* 2006; 97: 214–218)**

Tumor response to chemotherapy was previously evaluated using the WHO criteria, which stipulate bidimensional measurement (BD; WHO type) of lesions.[1] With these standardized criteria for evaluating tumor response, valid and reproducible results could be obtained by all investigators. However, a number of modifications to the WHO criteria have been developed by different institutions, which made it difficult to compare response rates for screening new anticancer agents across different investigators. This has led to the introduction of a new system, the RECIST guidelines,[2] which have been widely accepted as the new standard.

In order to standardize the methodology for evaluating tumor response, RECIST simplified the response evaluation through the use of unidimensional measurements (UD; RECIST type) instead of the BD used by the WHO criteria. Furthermore, the MLS allowable for measurement at baseline study was defined as being no less than double the slice thickness on CT or MRI.

The validity and intercriteria reproducibility between the new RECIST guidelines and the previous WHO criteria have been investigated.[2–7] However, to the best of our knowledge, no analysis of the influence of MLS on interobserver reproducibility, specified for measurability in tumor response evaluation according to the RECIST guidelines, has been published in the literature.

The purpose of the present study was therefore to evaluate interobserver reproducibility in tumor response evaluation using RECIST, intercriteria reproducibility between WHO and RECIST, and whether this reproducibility is affected by the application of MLS.

## Materials and Methods

### Patient population

This is a retrospective study of the radiological findings for patients who underwent chemotherapy for advanced NSCLC. The subjects were patients treated during clinical trials at the Medical Oncology Division of the National Cancer Center Hospital in Tokyo, Japan, between January 1999 and January 2001. All clinical trials were conducted in accordance with the Helsinki Declaration and the protocols were approved by the institutional review board. Written informed consent was obtained from each patient for the treatment protocols, which included the secondary use of treatment-associated documents. Patients were staged according to the Union Internationale Contra le Cancer TNM classification of malignant tumors.[8] The 110 eligible patients included those histologically or cytologically diagnosed with NSCLC. Patients were required to undergo CT scans periodically for evaluating tumor response prior to and once after

[6]To whom correspondence should be addressed. E-mail: h-watanabe@kcc.zaq.ne.jp
Abbreviations: BD, bidimensional measurements; CR: complete response; CT, computed tomography; MLS, minimum lesion size; MRI, magnetic resonance imaging; NSCLC, non-small cell lung cancer; PD, progressive disease; PR, partial response; RECIST, response evaluation criteria in solid tumors; SD, stable disease; TNM, tumor node metastases; UD, unidimensional measurements; WHO, World Health Organization.



doi: 10.1111/j.1349-7006.2006.00157.x

treatment, to have at least one bidimensionally measurable lesion, and to be treated with chemotherapy in clinical trials.

Patients treated in clinical practice were considered to be unsuitable and excluded from this study as tumor response evaluation in the clinical practice of oncology is not always carried out according to predefined criteria, but rather is made by subjective medical judgment based on clinical and laboratory data. In addition, tumor response evaluation is not always carried out by CT examination, and the intervals between tumor evaluations can be irregular.

### Image analysis

Almost all images were acquired with a TCT-900S Superhelix (Toshiba Medical, Tokyo, Japan), with the remainder scanned on an X-Vigor helical CT scanner (Toshiba Medical). Helical CT was carried out with fixed scanning parameters, including a table speed of 15 mm/s, a pitch ratio of 1:1.5 per rotation time 1 s, and the same contrast agent for both baseline and follow-up evaluations. Image reconstruction was carried out at intervals of 10 mm.

On chest CT obtained during baseline examination before the initiation of chemotherapy, target lesions up to a maximum of five lesions per patient with longest and perpendicular diameters that could be measured accurately were selected by one diagnostic radiologist. In addition, one follow-up chest CT examination, indicating tumors with the greatest response to chemotherapy, was selected retrospectively. Target lesions included primary lung lesion, pulmonary metastases and lymph nodes.

For the target lesions, the two parameters consisting of the longest diameter and the diameter perpendicular to it were measured with electronic calipers on digitized images. Five observers of different backgrounds, blinded to patient profiles, reviewed all patients independently and no attempt was made to arrive at a consensus. These observers included one diagnostic radiologist, one thoracic physician, two medical oncologists and one thoracic surgeon.

### Tumor response evaluation

The sum of the longest diameters for all target lesions was calculated for pretreatment and post-treatment UD. Similarly, the sum of the products of the longest diameters and their perpendicular diameters for all target lesions was calculated for pretreatment and post-treatment BD. If there were two or more lesions, the sum of all target lesions was calculated. The baseline sum was used as the reference from which objective tumor response could be calculated. The percentage changes were calculated as post-treatment value divided by pretreatment value for both UD and BD.

Percentage changes were then classified using the current RECIST guidelines and the previous WHO criteria tumor response classification system. Tumor response was categorized into CR, PR, SD and PD based on both RECIST guidelines and WHO criteria. The RECIST PR was defined as a 50% decrease in the percentage changes for UD, and the WHO PR was defined as a 30% decrease in the percentage changes for BD. The RECIST PD was defined as a 20% increase in the sum of the longest diameters, and the WHO PD was defined as a 25% increase in the sum of the products of the two diameters of all lesions or in the product of the diameters of one lesion. For the present study, no minimum interval was required for the confirmation of either CR or PR.

### Analysis of intercriteria reproducibility

To examine intercriteria reproducibility, the mean and ranges of differences in the response rate between UD and BD were calculated. We then estimated those between UD-MLS and BD. Interobserver differences among the five observers yielded 20 pair comparisons. Intraobserver differences of the same observer yielded five pair comparisons.

### Analysis of interobserver reproducibility

First, to examine the interobserver reproducibility of the percentage changes according to the two different dimensional measurements, we estimated the Spearman's correlation coefficient of the percentage changes among the five observers, calculated for each pair observed (five observers yielded 10 pair comparisons).

Second, to examine the interobserver reproducibility for two tumor response criteria, we estimated the proportion of agreement to the categories of CR, PR, SD and PD for both UD and BD among the five observers (10 pair comparisons). We then calculated the kappa statistics, a measure of agreement in which agreement is taken into consideration by chance, to assess interobserver reproducibility for tumor response categories.[9]

Third, we examined the influence of MLS on the number of eligible cases and target lesions. The same analyses on interobserver reproducibility were conducted applying the MLS. MLS was introduced into the RECIST guidelines, which specify a minimum lesion size of less than double the slice thickness on images. The slice thickness was 10 mm in the present study, so the MLS was set at no less than 20 mm at baseline evaluation before treatment. Cases that only had tumors smaller than the MLS were excluded from the present study. We defined the RECIST guidelines as the evaluation by UD for measurable cases and the WHO criteria as the evaluation by BD for all cases.

SAS version 8.02 (SAS Institute, Cary, NC, USA) was used for all analyses.

## Results

### Patient population

The characteristics of the 110 patients were as follows: male/female = 80/30, median age = 59 years (range 36–72 years), stage IIIB/IV = 33/77. Chemotherapy regimens are listed in Table 1. A total of 220 CT images were reviewed, comprising 110 CT images each from the baseline study (pretreatment) and from the follow-up (post-treatment) study.

### Tumor response evaluation between UD and BD

The tumor response evaluation was categorized into CR, PR, SD and PD without MLS. The response rate results are shown in Table 2. None of the patients were rated CR. The use of UD resulted in response categories by observers A, B, C, D and E of 35, 28, 26, 34 and 36 PR, 73, 79, 81, 73 and 71 SD, and 2, 3, 3, 3 and 3 PD, respectively. The response rate ranged from 23.6 to 32.7%. For BD, the corresponding response categories were 37, 30, 33, 36 and 36 PR, 67, 73,