Japan Clinical Oncology Group

ポリシー No. 07
タイトル：統計的原則と試験デザイン
適用範囲：データセンター、運営事務局を含むプロトコール作成に関わるすべての研究者

検証的試験の統計的原則と試験デザイン

Statistical principles and study design for confirmatory trial

1 目的 Aim

本ポリシーは、主としてJCOGが行う検証的な臨床試験における統計的原則および試験デザインの指針を定めるものである。ただし、臨床試験の具体的な方法や判断規準については各臨床試験の内容によって異なるため、実際の臨床試験の統計手法および試験デザインについて必ずしも本指針を強制するものではない。また、プロトコールの標準的な統計関連事項の記載についてはプロトコールマニュアルに例を示すこととする。

なお、中間解析については本ポリシーに基本的な考え方を定める他、中間解析審査の方針と手順は JCOG ポリシーNo.21「効果・安全性評価委員会」、No.22「中間解析の方法と審査」にて定める。

2 基本原則 Principles

JCOGの第III相試験の原則は「単純（simple）かつ保守的（conservative）」である。

3 適用範囲 Application

以下に示す、すべてのJCOGの検証的臨床試験を対象とする。

1） ランダム化比較第III相試験
2） 非ランダム化比較第III相試験
3） 単群第III相試験

4 試験デザイン Study design

4.1. 試験の種類の分類および試験の相（phase）の考え Study type and phase

ICH E9「臨床試験のための統計的原則」によると、検証的試験とは、「事前に定められた仮説を評価するための、適切に計画・実施された比較試験」のことである(1)。企業が実施する臨床試験の場合、未だ薬効の存在が証明されていない医薬品等について薬事承認取得のために十分な根拠を得ることが主な目的とされる。一方、JCOGが実施する臨床試験の場合は、通常は既に薬事承認を取得した薬剤や臨床現場で既に治療選択肢の一つとして実施可能な医療技術（外科手技、放射線治療、内視鏡治療等）を用い、新たな標準治療を決め、その臨床試験の結果を日常診療に直接反映することが主な目的であるという違いがある。検証的試験で最も多く用いられるデザインはランダム化比較試験であるが、治療のアウトカムが既に非常に良好な患者（例：5 年生存割合≧95%）を対象にした試験治療に対する評価の場合や、適切なヒストリカルコントロールがある場合などは、適切な比較が可能であれば、ランダム化比較以外のデザインの試験を検証的試験として実施する。

検証的試験以外の臨床試験は基本的に探索的試験と位置づけられる。探索的試験では、次の研究で検証すべき仮説を作る、次の研究で評価に値する有望な治療法かどうかを判断する、検証的試験の結果の解釈の参考とする、などが主な目的とされる。

ICH E8「臨床試験の一般指針」 (2)で述べられているように、開発の相（phase I, II, III）とい

う概念が医薬品以外のモダリティの試験治療を評価するための臨床試験の分類の基礎としてふさわしいとは必ずしも言えない。なぜなら、開発の相と試験の種類は多くの場合で一致するものの、疾患または治療の性質などのさまざまな要因により、一般的な開発順序が必ずしも当てはまらないためである。また医薬品の臨床試験に限ってみても、抗がん薬以外の領域であれば、一般に用量探索を目的とした試験は第II相で行われるが、抗がん薬では用量探索は第I相で行われるなど、疾患領域や医薬品の作用機序・特性により、試験の種類と試験の相は必ずしも一致するわけではない。試験の相という概念は単に開発過程の進捗段階を示すものであり、試験の種類を示すものではないため、本来は臨床試験の分類は開発の相ではなく、試験の種類（臨床薬理試験、探索的試験、検証的試験など）によることが望ましい。しかし、試験の種類に基づく分類よりも試験の相に基づく分類の方が広く認知されており、がん領域では「第 III 相試験」≒「検証的試験」と認識している研究者も多い。そのため、JCOG の臨床試験は、試験の相に基づく分類である「第 III 相試験」を試験の種類に基づく分類である「検証的試験」を表す用語として用いる。

4.2. 試験の目的-優越性試験と非劣性試験 - Purpose of study – superiority vs. non-inferiority

一般に、がんを対象とした第 III 相試験における目的は標準治療と試験治療の治療効果、毒性、薬剤費、QoL、利便性などのバランスにより大きく以下の 3 つに分類される。

① 試験治療は標準治療よりも毒性が強いことや長期間の入院を要するなどのデメリットを有するため、新たな標準治療とするために試験治療が有効性で標準治療よりも優れていることを示すことを目的とした試験

② 試験治療は標準治療よりも毒性が軽いことや治療期間が短いことなどのメリットを有するため、新たな標準治療とするために試験治療が有効性で標準治療よりも一定以上劣っていないことを示すことを目的とした試験

③ 2 つの治療法は毒性や利便性の観点で同等であると考えられるため、標準治療を決めるためにどちらか一方の治療が有効性でもう一方の治療よりも優れていることを示すことを目的とした試験

①・③のような目的の試験を優越性試験（superiority trial）、②のような目的の試験を非劣性試験（non-inferiority trial）という。JCOG では標準治療と試験治療の試験のリスク・ベネフィットバランスに応じて優越性か非劣性かを決定する。試験治療を新たな標準治療とするために非劣性が示されれば十分である状況において、試験治療が有効性で優越性を示す見込みがあるために、優越性試験としてデザインすることは原則として行わない。

②の非劣性試験には、試験治療を標準治療に置き換えることを意図した試験と試験治療を標準治療のオプションのひとつにすることを意図した試験がある。例えば、後者は試験治療を提供する場合に特別な医師の技術や施設の要件が必要な場合や、倫理的問題などから、対照としてプラセボの代わりに実薬を用いて、試験治療の有効性を証明することを目的とした非劣性試験が該当する(3)。

4.3. 患者選択規準 Patient selection criteria

試験の適格規準（inclusion criteria）・除外規準（exclusion criteria）の使い分けに関する考え方は大きく 2 つに分かれる。一つは、患者選択規準を表す文言が試験の対象に含めることを意図している場合は適格規準とし、試験の対象から除外することを意図している場合は除外規準とするものである (4)。言い換えれば、「～を満たす」という肯定表現のものを適格規準、「～ではない」という否定表現のものを除外規準とするものである。もう一つは、試験の結果、治療法の有効性が示された場合にその治療を適用することが妥当と判断する対象集団を規定するものを適格規準とし、適格規準で示される対象集団には属するが、治療のリスクが高いために試験に組み入れることが倫理的でないか（倫理的側面）、試験で必要な有効性・安全性の評価に影響を及ぼすと

判断される対象を除外する（科学的側面）条件を規定するものを除外規準とするものである (5)。JCOG では、後者の考え方に従い、患者選択規準を規定する。

倫理的側面としては、例えば、心疾患（心筋梗塞・狭心症）、肝疾患（慢性肝炎）、肺疾患（間質性肺炎・肺線維症）、特定の薬剤服用中など、治療で得られるベネフィットに対してリスクが高いと考えられる患者を除外規準に規定する。科学的側面としては、適格規準をすべて満たすが心筋梗塞の既往を有する患者の場合、将来、試験の結果が得られて新しい標準治療となった治療を、心虚血に注意しながらそうした患者に行うことは正当化される。しかし、そうした患者を試験に組み入れることにより、その患者が心筋梗塞で死亡した場合には、対象疾患であるがんに対する治療効果が薄まり、正しい評価に影響を及ぼし得る。このような懸念がある患者の条件を除外規準に規定する。

また、試験治療の有効性が示された場合に適切な集団に適用できるよう、エンドポイントの評価ができない患者（例えば、奏効割合を primary endpoint とした時に測定可能病変がない患者）や、規定のプロトコール治療の一部が行えないことが予め判っている（髄液注入療法を含む治療レジメンの試験におけるクモ膜下出血の既往など）患者が適切に除外される患者選択規準を設定する。

このように、対象集団の設定すなわち患者選択規準は、「試験の目的」、「エンドポイント」、「治療内容」と密接に関連する。狭すぎる適格規準の試験結果は特定の患者集団にしか適用できないものとなる（一般化可能性が低い）。逆に広すぎる適格規準が原因で治療効果が期待できない患者が多く含まれるようでは、試験結果を適用する患者集団として不適切であり、さらに、治療効果の差が薄まってしまい、真に有効な治療法を見逃してしまうリスクが大きくなるという観点でも問題である。従って、患者選択規準は治療効果が期待できる最大の患者集団を規定することが望ましい。

4.4. ランダム化の有無 Randomization

潜在的な背景因子による交絡を防ぎ、治療群間の比較可能性を担保するためにランダム化は重要な役割を果たす。従って、すべての第 III 相試験はランダム化試験で行うことを第 1 選択とする。

一方、治療のモダリティが明らかに異なる場合は、事実上患者からランダム化の同意取得が困難であるため、ランダム化比較試験が実施不可能な場合も少なくない。そのような場合、一般的なランダム化比較試験の精度を保つために必要なサンプルサイズと同等以上の登録が可能であることを条件に、同時対照の非ランダム化比較試験が選択肢となる。

また、標準治療についての信頼できるヒストリカルコントロールがあり、かつその患者集団の primary endpoint が非常に良好（例えば、5 年生存割合≧90%）な場合において、試験の実施可能性を考慮すると十分な検出力を担保することが困難であること、選択バイアスの影響で本来は有効でない新治療を誤って新しい標準治療に採用してしまうリスクは小さいことから、単群の第 III 相試験が選択肢となる。

4.5. ランダム化の方法と割付調整因子 Randomization method and adjustment factor

4.5.1. ランダム化に用いるシステム Randomization system

JCOG データセンターが支援する臨床試験においては、ランダム化は JCOG データセンターが提供する web システム（JCOG Web Entry System）による中央方式を基本とする。なお、術中ランダム化の場合など、施設によっては web システムによるランダム化が困難な場合は、データセンターへの電話により、研究者に代わりデータセンターのオペレーターが web システムにアクセスする方法も許容する。

4.5.2. ランダム化の方法 Randomization method

一般に、ランダム化の方法としては以下のものがある：①単純ランダム割付（Simple

randomization)、②層別置換ブロック法（Stratified permutated block randomization)、③最小化法（Minimization method)。①単純ランダム割付は、数千例の大規模な試験を除き、重要な予後因子のバランスが群間で偏るリスクが大きいため推奨しない。②層別置換ブロック法では、施設毎に層を作成した場合に多数の層ができることによって却って群間に不均衡が生じることから、後述するように JCOG 臨床試験ではデフォルトとして施設を割付調整因子とすることが標準であり、③最小化法を用いることを基本とする。

最小化法の中でも決定論的な要素を含むアルゴリズムは予見性を高める可能性があるため用いない。代わりに、割付調整因子に基づいて決められた群へ高い確率（2/3 など）で割り付けるアルゴリズム（biased coin method）を用いる。

4.5.3. 割付調整因子 Adjustment factor

割付調整因子による調整は、治療効果の差よりも明らかにエンドポイントに大きな影響を与える因子について、群間で不均衡が生じないようにすることを目的とする。従って、既に確立された予後因子の中から、予後に対する影響が大きいと考えられる因子を優先して選ぶことを推奨する。附随研究や試料解析研究での群間差を減らす目的で検体提供への同意の有無を調整因子に加えることは、予後に大きく影響するリスクが小さいと考えられるならば原則として不可とする。

JCOG の第 III 相試験は臨床研究を行う体制が整備された臨床研究中核病院から一般病院まで幅広い施設で行われるため、施設を割付調整因子に含めることを標準としているが、施設のみを割付調整因子とすることは予見性を高めるため推奨しない。最小化法を用いる場合、理論的には割付調整因子の数に制限を用いる必要性は小さい。ただし、割付調整因子の数が多くなることで、割付調整因子の組み合わせで構成した層を用いた層別解析で、特定の層でイベントが生じず、その層の患者が解析から除外される、あるいは他の層と統合する必要が生じ得る可能性がある。また、ある層が非常に少ないと予想されても予想に反することもあり、バランスが崩れていることが結果の解釈に影響を及ぼし得る。従って、割付調整因子の上限を決めることはしないがSWOGの指針を参考として3つ程度を目安にすること､因子数を定める際には、予定登録数に応じて、層あたりの患者数（イベント数）が極端に少なくならないように検討することを推奨する。

4.5.4. ランダム化の比 Randomization ratio

ランダム化は 1:1 割付を基本とする。試験治療の方が標準治療よりも優ることが予想されるために 2:1 割付をすることは、ランダム化の原則である“equipoise”(6)が成立していないと考えるため推奨しない。ただし、試験治療の安全性のデータをより多く収集するためなど、妥当な理由がある場合は許容する。統計学的には、群間でサンプルサイズが等しい時にもっとも検出力が高くなるため、1:1 がもっとも効率的である。

4.6. . エンドポイント Endpoint

4.6.1. 有効性のエンドポイント Efficacy endpoint

何が試験の primary endpoint として相応しいかは対象患者や治療のモダリティにより異なる。原則として、その治療（薬）に治療効果（薬効）があるかどうかを示すために設定するのではなく、治療（薬）が患者にベネフィットがあることを示すことを優先して設定する。つまり、検証的試験では患者のベネフィットを直接測ることのできる真のエンドポイントを用いる。そのため、対象患者によらず、全生存期間（overall survival）が第一選択となる。

しかし、全生存期間以外を primary endpoint として用いることができる場合がある。一般に、再発後は治癒困難と考えられるため、治癒を目指した治療である術後補助療法や早期がん患者を対象とした手術手技の比較試験では、無再発または無病であることが患者の真のエンドポイントであるとみなすことができる。この場合、無再発生存期間（relapse-free survival：

RFS - 再発・死亡がイベント）や無病生存期間（disease-free survival：DFS - 再発・死亡・二次がんがイベント）を primary endpoint として設定することも可能である。また、無再発生存期間や無病生存期間が全生存期間のサロゲートエンドポイント（surrogate endpoint）であることがメタアナリシスなどの統計学的手法により証明されている場合も、無再発生存期間や無病生存期間を primary endpoint とすることは可能である。

一方、切除不能がん患者を対象とした試験の場合、一部の疾患・状態を除いて無増悪生存期間や奏効割合を患者の真のエンドポイントとみなすことは困難である。そのため、治癒が望めない患者を対象とした検証的試験では、妥当な理由がない限り無増悪生存期間や奏効割合を primary endpoint とすることは推奨しない。妥当な理由の例としては、仮に無増悪生存期間等が全生存期間のサロゲートエンドポイントでなかったとしても、増悪を経験せずに生存できることが患者の治療選択に重要な意味を持つ場合や患者の日常生活や社会生活に大きな影響を与える場合などが挙げられる。新規治療による生命予後の延長は重要な目標の一つであるが、それは JCOG 基本規約で掲げている「最善の医療を確立することを目的として研究活動を行う」唯一の在り方ではなく、最善の医療の確立に直接・間接に寄与し得るエンドポイントか否かの観点での検討を形式的に排除するべきではない。

また、当該がん種において、無増悪生存期間や奏効割合などが全生存期間のサロゲートエンドポイントであることがメタアナリシスなどの統計学的手法により証明されている場合、無増悪生存期間や奏効割合を primary endpoint とすることは可能である。ただし、サロゲートエンドポイントの証明のために用いられた試験よりも増悪後の治療の選択肢が増えている場合は増悪後の生存期間が大幅に長い可能性がある。この場合、無増悪生存期間を primary endpoint とすることには慎重になるべきである。

腫瘍縮小効果の判定には、RECIST ガイドライン (7)に基づいた判定を基本とする。血液腫瘍や脳腫瘍など疾患特異的なガイドラインが存在する場合も、RECIST を応用する形で対応可能であることが多いため、CRF などの標準化も鑑み、RECIST ガイドラインに基づいた判定を用いることを推奨する。免疫療法など薬効特異的なガイドラインのうち、RECIST を応用する形で策定されたもの（iRECIST(8)など）が存在する場合、それらを用いることも許容する。

4.6.2. 安全性のエンドポイント Safety endpoint

がんを対象とした臨床試験では Common Terminology Criteria for Adverse Events（CTCAE）に基づいて判定することが一般的であるため、JCOG 臨床試験もこれに従う。ただし、外科領域では術後合併症の規準として Clavien-Dindo 分類 (9)が広く用いられるようになっていることから、これを併用することも許容する。ただし、Clavien-Dindo 分類を用いる場合はオリジナルのものではなく、AE term の共通化、grading の詳細の共通化を行った JCOG 版 (10)を用いることとする。

4.6.3. Patient Reported Outcome（PRO）のエンドポイント PRO endpoint

Patient Reported Outcome（PRO）は、がんの臨床試験においても年々重要性が高まりつつある。

PRO 評価のための調査票は日本語で記載されており、その妥当性が示されたものを用いることを原則とする。例えば、がん臨床試験では健康状態を測定するための尺度として欧州 EORTC（European Organisation for Research and Treatment of Cancer）で開発された EORTC-QLQ-C30 や米国で開発された FACT-G などがよく用いられる。また、効用値を測定するための尺度として EQ-5D を用いることもある。CTCAE に対応した PRO 調査票として、PRO-CTCAE を用いることも可能である。これらは代表的な調査票であるが、これ以外にも疾患領域に応じて適切な調査票を用いることを許容する。

PRO データの解析では、欠測データの取り扱いが重要な問題となる。近年の ICH E9(R1)の

動向も抑えつつ、プロトコール作成時に欠測データをできるだけ予防するような仕組みを取り入れることを推奨する。また解析については、閾値を設定して改善割合を求めるなどの二値に落とし込む方法を推奨する。二値を用いる場合、missing data（欠損値）については保守的な方向に補完することを基本とする。ただし、がん臨床試験における PRO データの報告法の標準化を目的とした working group(11)が立ち上がっていることも踏まえ、今後、より良い解析法が確定した時点でその手法に従うこととし、ポリシーを変更する。

現時点の PRO のエンドポイントに関する詳細は JCOG ポリシー「30. QOL 調査」に記載されている。

4.6.4. 医療経済に関するエンドポイント Economic endpoint

欧州と異なり、これまでは医療技術の保険収載や保険償還価格の判断材料に費用対効果（cost-effectiveness）は重視されてこなかった。しかし、2016 年 4 月より、診療報酬改定における医薬品・医療機器の評価について、費用対効果評価の試行的導入がなされた(12)。今後は、特に保険収載を目的とした臨床試験において費用対効果を評価することが必須になると考えられる。

JCOG では、必要な場合には「中央社会保険医療協議会における費用対効果評価の分析ガイドライン」(13)を参考にして医療経済評価を行う。一つの選択肢として、各群の期待費用と期待効果から増分費用効果比（Incremental cost-effectiveness ratio：ICER）を算出することを検討する。

4.7. サンプルサイズ Sample size

臨床試験の primary endpoint の評価のために必要なサンプルサイズは、アウトカムの型や特殊な解析手法を用いる場合で差違はあるものの、一般に治療効果の差（Δ：デルタ）、有意水準（α）、検出力（1-β：power）、アウトカムのバラツキの 4 つのパラメータで規定される。

治療効果の差（Δ）の決め方は、①試験治療と標準治療の治療効果、毒性、薬剤費、QoL、利便性などのバランスに基づき、優らなければならない治療効果の大きさで決める方法、②先行研究のデータに基づいて見込まれる治療効果の大きさで決める方法がある。①、②両方の観点から計画をすることが必要であり、②のみで決めると臨床的に意味のない試験治療を残してしまう可能性があるため推奨しない。例えば、試験治療が標準治療と比較してかなり毒性の強いレジメンである場合やモダリティが追加される場合は、試験治療がわずかにしか上回っていない場合は、それが新たな標準治療にはならないため、Δは大きく設定されるべきである。一方、標準治療と比較して試験治療の毒性の上乗せが小さければ、Δは小さく設定されるべきである。このような検討は、臨床的に意味の無い差を統計学的に検出してしまう試験計画を避ける上で重要である。この観点でΔを決めた後、先行研究のデータを踏まえた考察により、試験治療の効果/ベネフィットが本当にそのΔを見込めるかどうか、そのΔを設定した場合に実施可能性があるかどうかなどを検討する。この検討無しには臨床試験の実施可能性が評価出来ないため、①、②両方の観点から事前の検討を行うことが必要である。

非劣性試験の場合、①に関する考え方の他に、③現在の標準治療が過去の標準治療に対して示した差の一定以上の効果を担保するように非劣性マージン（非劣性試験のΔに対応）を決める方法がある。非劣性マージンが大きすぎると、非劣性が示されたとしても、現在の標準治療に劣る過去の標準治療と同程度の効果しかない治療法が新たな標準治療になる可能性があるため、①と同様に②の視点も考慮すべきである。

①の考え方の場合、どのくらいの大きさであれば良いという客観的な規準は存在しない。そのため、グループの研究者が議論を重ね、研究者間でコンセンサスを得る必要がある。場合によっては、患者団体などに意見を求めることも考慮してもよいかもしれない。

JCOG 臨床試験においては、多くの場合、毒性や侵襲の異なる治療法どうしの比較であるため、

検定の推論は多くの場合、両側仮説ではなく片側仮説を選択することとなる。言い替えると、試験治療が標準治療に優越するという片側仮説を採用するのは、標準治療に対して試験治療が劣る可能性が低いためという理由ではなく、標準治療よりも毒性が強く、侵襲の大きな試験治療の臨床的価値を否定するために標準治療に統計学的有意に劣るという結果自体が必要無いためとの理由による（詳細は 5.2.4 節参照）。

統計学的な精度として、有意水準（α）両側 5%に相当するのは片側 2.5%である。第 III 相検証的試験について、ICH のガイドライン「E9：臨床試験のための統計的原則」には「原則として片側仮説を検証する場合は 2.5%、両側仮説の場合は 5%とする」とあり、少なくとも治験（医薬品等の薬事承認を目的とする臨床試験）における片側検定の有意水準の国際標準は 2.5%である。そのため、JCOG では片側検定を行う場合には有意水準 2.5%、両側検定を行う場合には 5%を推奨する。

しかしながら、がん領域では多くの被験者数が確保できず、有意水準 2.5%を採用することで臨床試験の規模が実施可能ではないものになってしまうことが少なからずある。また、治療開発は科学的な研究業績のためだけに行われるのではなく、臨床現場での意思決定に必要な情報を得るために行われるものであるため、精度の高い臨床試験が組めない状況でも次善の策として精度を落とした臨床試験を統計学的に弱点があることを理解して行わなければならない場合も多々ある。そのため、実施可能性の観点並びに試験結果を解釈するコミュニティのコンセンサスが得られる場合には有意水準片側 5%や、場合によってはより緩く設定することを許容する（例：希少がんの第 III 相試験での有意水準片側 10%）。

しかし、がん種によっては国際的コミュニティのコンセンサスで、結果の公表の際に片側 2.5%と設定していないことを理由に、学会や雑誌に検証的な結果として受け入れられない可能性がある。そのようながん種を対象とする臨床試験において片側 2.5%よりも緩い有意水準を設定する場合は、結果の公表時のこうしたリスクを十分に認識した上で、グループが自らの責任において片側 5%（あるいはより緩い水準）を選択する。

検出力（1-β）は 80%以上を基本とする。ただし、検出力 80%は真に差がある場合 5 回に 1 回は差がないと判断してしまうことを意味しているため、必ずしも十分な精度とはいえない。欧米では 90%以上を基本としている臨床研究グループも存在するため、実施可能性を考慮しつつ 90%にすることを検討するべきである。一方、対象集団が細分化されて患者集積が困難になりつつある現状を鑑み、試験開始時には 80%未満（例：70%）で開始し、登録状況をみて検出力を上方修正することも許容する。その場合、予めプロトコールにその可能性を記載しておくことが望ましい。

Time-to-event 型のデータを曲線全体で比較（ログランク検定、Cox 回帰による）する場合のサンプルサイズ計算は、まず冒頭の 4 つのパラメータのうちΔ（ハザード比）、有意水準α、検出力（1-β）で解析時に必要なイベント数を算出する。その後、必要なイベント数を観察するためのサンプルサイズをアウトカムのバラツキ（対象集団の予後）、登録期間、追跡期間を考慮して算出する。主たる解析を実施する時期は、必要なイベント数が観察された時点で行う場合と、臨床的に必要な追跡期間を経た後に実施する場合がある。JCOG では、死亡や増悪・再発といったイベントを収集する CRF は用いておらず、追跡調査用紙にてこれらのイベントを半年に一度収集することを標準としているため、後者の時期による規定を基本とする。

5 統計解析の原則 Principles for statistical analysis

5.1. 解析対象集団 Study population

JCOG の第 III 相試験では、優越性試験、非劣性試験ともに、ランダム化の有無によらず有効性の primary endpoint については全登録例を主たる解析の解析対象集団とする。

優越性の検証を目的としたランダム化第 III 相試験の場合、比較可能性および α エラーを制御

することを考慮すると、全登録例を主たる解析の対象とすることが望ましい。実際、「ランダム化が行われた全被験者を主要な解析に含めるべき」という主張である ITT（intent(ion)-to-treat）の原則 (1)と一致する。また、全適格例を主たる解析対象としている SWOG に対して、Alliance など SWOG 以外の cooperative group では全登録例を主たる解析の対象としている。このため、JCOG 臨床試験も全登録例を主たる解析の解析対象集団とする。

非劣性の検証を目的としたランダム化第 III 相試験の場合、疾患または治療の性質により、プロトコール治療をより遵守している集団（ここでは、Per protocol set：PPS とする）を主たる解析対象とすることも多い。がん領域の場合、割り付けられたプロトコール治療を一部でも行った対象を PPS と定義することが多く、JCOG でもこの考え方を踏襲する。PPS を主たる解析対象とする理由は、全登録例を用いると、プロトコール不遵守の患者が多くなるほど群間の治療効果の差が小さくなり、結果的に非劣性が証明しやすくなる（保守的ではない）ためである。一方、全登録例から多くの患者が除外され、全登録例と PPS の人数が大きく異なる場合、PPS は比較可能性が損なわれる可能性があるため、α エラーが増加し得る。すなわち、全登録例または PPS のどちらを主たる解析対象とした場合も、解析対象集団によって得られる結果が異なる場合は試験治療の効果を慎重に解釈すべきである。プロトコール治療の性質によって PPS の定義を適切に定めた上で、JCOG では PPS に対応する全治療例を主たる解析対象とするものの、CONSORT に未治療例や不適格例の内訳を提示した上で、全登録例や全適格例（全登録例から登録後の情報に基づき不適格が判明した事後不適格例を除外した対象集団）を対象とした補足的解析も行う。

単群第 III 相試験の場合、比較対照のデータが必ずしも全登録例のデータでない場合もあるため、全適格例を主たる解析の対象とすることも許容し、試験毎に解析対象集団を定義する。

5.2. Primary endpoint に対する主たる解析手法 Analysis for primary endpoint

5.2.1. ランダム化比較第 III 相試験の場合 Randomized phase III trial

Primary endpoint に対する主たる解析はランダム化に基づいた手法を標準とする。すなわち、time-to event 型のエンドポイントであれば、優越性試験では施設以外の割付調整因子を層とした層別 log-rank 検定を主たる解析手法とし、非劣性試験であれば施設以外の割付調整因子を層とした層別 Cox 回帰を主たる解析手法とする。いずれの場合も、治療効果の推定は施設以外の割付調整因子を層とした層別 Cox 回帰を用いてハザード比を算出し、その信頼区間は中間解析の多重性を考慮した信頼係数で構成する。

5.2.2. 非ランダム化比較第 III 相試験の場合 Non-randomized controlled phase III trial

Primary endpoint に対する主たる解析はプロトコールで規定した因子で調整した手法を標準とする。すなわち、time-to event 型のエンドポイントであれば、事前に規定した因子を層または共変量とした Cox 回帰か、傾向スコアを用いた IPTW（Inverse Probability of Treatment Weighting）などによる Cox 回帰を主たる解析手法とする。治療効果の推定は事前に規定した因子を用いて調整した Cox 回帰を用いてハザード比を算出し、その信頼区間は必要に応じて多重性を考慮した信頼係数で構成する。

5.2.3. 単群第 III 相試験の場合 Single arm phase III trial/single arm confirmatory trial

Primary endpoint が time-to event 型の割合（例：5 年生存割合）の場合、主たる解析は Kaplan-Meier 法で割合の点推定値を算出し、その信頼区間は Greenwood の公式で求めた標準誤差を用いて算出する。Primary endpoint がカテゴリカルデータの割合（例：奏効割合）の場合、二項分布に基づく正確な検定と Clopper-Pearson 法に基づく正確な信頼区間を用いる。

5.2.4. 片側と両側 One-sided vs. two-sided

優越性試験と非劣性試験のいずれにおいても、以下に示す両側検定が妥当な状況に該当しない限り primary endpoint の解析の検定は片側検定を JCOG 標準とする。

優越性試験を行うのは、試験治療が標準治療に比して毒性が強い等のデメリットを有している場合（toxic new）であり、試験治療はそのデメリット（毒性）に見合うメリット（有効性）を有することが示されて初めて標準治療に優っていると言える。Primary endpoint の解析では、試験治療群が標準治療群に有効性で優っているか否かを統計学的仮説検定を用いて判断する。すなわち、検定が統計学的に有意であった時に「試験治療が標準治療よりもよい治療である」と結論し、有意でなかった時には「引き続き標準治療がよりよい治療である」と結論する。一方、試験治療群が有効性において標準治療群に劣っている時は、統計学的に有意か否かによらず「引き続き標準治療がよりよい治療である」という結論は変わらない。つまり、患者に対する「引き続き標準治療を第一選択として推奨する」という意思決定は統計学的に有意か否かによって変わらず、検定に基づいて臨床的意思決定を行うわけではないことから、primary endpoint の解析では片側仮説を評価していることになる。以上より、優越性試験における primary endpoint の解析の検定は片側検定を標準とする。

同様に、非劣性試験においても、有効性以外のメリット（毒性が軽い等）を有する試験治療群（less toxic new）が許容下限（非劣性マージン）を統計学的に有意に上回っているか否かによって「試験治療が標準治療よりもよい治療である」と結論するか、「引き続き標準治療がよりよい治療である」と結論するのかが決定される。試験治療群が許容下限を統計学的に有意に下回っているか否かによって結論は変わらず、同様に検定に基づいて臨床的意思決定を行うわけではないことから、非劣性試験においても primary endpoint の解析は片側検定にて行う。

以上の考え方は、一部の臨床試験で行われる、先行研究のデータから試験治療が優る可能性があるから優越性試験を行う、それが厳しそうであるから非劣性試験を行うという考えに基づくデザインの選択とは全く異なる。がん以外の領域においても、非劣性試験を行うためには試験治療に対して既存治療に優る有効性以外の毒性の軽減等のメリットが求められることが多い。しかし、がん領域の試験治療は、毒性の軽減等のメリットが無いだけでなく更にデメリットがあるというケースであると捉えれば、何故片側仮説の優越性試験のデザインを選択しなければならないのかが容易に理解できる。

両側検定が妥当な状況としては、試験治療のデメリットが標準治療と同等と考えられる場合（equitoxic new）や、デメリットが同等と考えられる 2 つの標準治療がある場合（standard A vs. standard B）で、既に日常診療として受け入れられている 2 つの治療のうち有効性で有意に上回った方を「よりよい治療である」と結論づけるような状況が考えられる。臨床的意思決定は、治療 A が有意に優った時は「治療 A を第一選択として推奨する」、治療 B が有意に優った時は「治療 B を第一選択として推奨する」、両者に有意差がなかった時には「いずれかを第一選択として推奨する根拠が無いため治療 A と治療 B のどちらでもよい」となる。なお、統計学的有意差がないことは 2 つの治療が同等であることを意味しないが、既に日常診療として受け入れられている 2 つの治療のいずれか一方を積極的に優先させる根拠が無いという意味で「どちらでもよい」と判断するのであって、有意差なしを持って同等と判断しているわけではない。

ただし、デメリットの面で同等な 2 つの標準治療候補がある時に、敢えてその両者に優劣を付けるための大規模な第 III 相試験を行う価値があるという状況はあり得る。

5.3. Secondary endpoint に対する主たる解析手法 Analysis for secondary endpoint

Secondary endpoint に対する主たる解析は、厳密に多重性の調整を要せず、primary endpoint の主たる解析結果を補足する位置づけであることから、ランダム化に基づく手法にこだわらない。すなわち、層別しない手法を標準とする。Time-to event 型のエンドポイントであれば、優越性試験では層別しない log-rank 検定を主たる解析手法とし、非劣性試験であれば層別しない Cox 回帰を主たる解析手法とする。いずれの場合も、治療効果の推定は層別しない Cox 回

帰を用いてハザード比を算出し、その信頼区間は記述的な位置づけとして 95%信頼係数で構成する。

2 値データのエンドポイントであれば、Fisher の直接確率計算法を主たる解析手法とする。

5.4. 感度分析と補足的解析 Sensitivity analysis and supplementary analysis

ICH E9(R1) (14)に感度分析と補足的解析の定義付けがなされたため、JCOG もそれに従う。また、重要な概念として、試験の目的により提示される関心のある科学的疑問に対応する推定の対象を示す ”estimand” という概念が導入された。Estimand の要素には、対象集団、関心のある変数（または評価項目）、関心のある科学的疑問を反映するために中間事象をどのように考慮するかという説明、および集団レベルでの変数の要約が含まれる。感度分析とは「主とする推定量の、モデル化における仮定からのずれとデータの限界に対する推測の安定性を調べるために実施される、同じ estimand を対象として、異なる仮定を用いた一連の解析」と定義される。がんを対象とした臨床試験であれば、比例ハザード性が成立していない場合に、Cox 比例ハザードモデルではなく、区分指数分布などのパラメトリックなモデルを適用する場合や、層別 Cox 回帰を主たる解析とした試験で層別しない Cox 回帰を行う場合などが考えられる。Time-to event 型のエンドポイントの解析において、Cox 回帰で曲線全体を比較するのではなく、生存曲線下面積（restricted mean survival time：RMST）や特定の時点の割合（5 年生存割合）などで要約した解析は、estimand を構成する要素のうち集団レベルでの変数の要約（population-level summary for variable）が異なるという意味で異なる estimand を対象とするため感度分析とはならない。また、主たる解析を全登録例を対象として実施した場合、全適格例や全治療例などを対象とした解析もモデル化における仮定は同じであること、estimand を構成する要素のうち対象集団（population）が異なるという意味で異なる estimand を対象とすることから感度分析ではない。

補足的解析とは、「主とする解析及び感度分析に加えて、治療効果の理解に追加の考察を与えるための解析の総称。この用語は感度分析よりも広い種類の解析を指す」と定義される。つまり、上述の特定の時点の割合などで要約した解析や、全適格例や全治療例を対象とした解析は補足的解析に該当する。

主たる解析結果で得られた結論が、異なる解析対象集団や異なる解析手法・モデルの仮定によっても同じ結論となるのかどうかを確かめることは、試験の結果を正しく解釈するうえで極めて重要である。具体的にどのような感度分析および補足的解析を実施するかは、試験ごとにプロトコールや統計解析計画書に定める。

5.5. サブグループ解析 Subgroup analysis

サブグループ解析は研究結果の解釈および新たな研究仮説の探索に有用である一方で、多くの問題点が指摘されている。例えば、適切な多重性の調整を行わない限り、α エラーが生じる可能性が高くなってしまう。また、サブグループ間での治療群間に交互作用があるかどうかの検定は、通常のランダム化比較試験のサンプルサイズでは十分な検出力を保持していない。従って、JCOG のランダム化第 III 相試験では事前に α の調整を行ったサブグループ解析を除いて全てのサブグループ解析は探索的な位置づけとする。

また、実施するサブグループ解析は、主たる解析前にプロトコールか統計解析計画書に記載する。

5.6. 競合リスクに対する解析 Competing risk analysis

高齢者を対象とした臨床試験や早期がんを対象とした臨床試験では、原疾患や治療とは関係なく死亡する患者が不可避的に一定数含まれる。競合リスク（competing risk）とは、興味のあるイベント（ここでは原疾患の再発や増悪、または原疾患や治療が原因の死亡）が観察される前に生じる、興味のあるイベントの観察を不可能にするイベント（ここでは他病死）のことを言う。

他にも、例えば手術手技の比較試験で局所再発に興味がある場合、局所再発の前に遠隔転移が生じると局所再発は報告されにくくなり、CRF で初回の再発しか収集しない場合は局所再発が報告されない。このような場合、遠隔転移は局所再発の競合リスクとなる。

競合リスクを伴うデータに対し、競合リスクを打ち切りとした Kaplan-Meier 法はバイアスが生じることが知られているため、用いることは推奨しない。ログランク検定や Cox 回帰も基本的には推奨しない。従って、競合リスクを伴うデータに対して解析を実施する際は、競合リスクを考慮した解析手法を用いることとする。すなわち、cumulative incidence 法や、Gray の検定 (15)、Fine & Gray のモデル (16)などを用いることを推奨する。

5.7. 中間解析 Interim analysis

中間解析審査の方針と手順は JCOG ポリシーNo.21「効果・安全性評価委員会」、No.22「中間解析の方法と審査」に定める。以下の記載は、ランダム化比較試験における中間解析の方針とする

5.7.1. 時期 Timing

中間解析の実施時期は、イベント数（情報時間：information time）に応じて決める方法と登録数に応じて決める方法があるが、JCOG では後者を標準とする。一般に、十分なイベント数が集まった時点で早期有効中止を念頭においた中間解析を規定することが広く行われている。それは将来の患者に対する配慮という意味では妥当であるが、臨床試験の参加患者への配慮という意味では不十分であり、統計学的には検出力が足りなくても、中間解析を行わなければならないことはある。

JCOG では統計的な観点よりも倫理的な観点を重視し、第 III 相試験では計 2 回の中間解析を実施することを標準とする。1 回目の中間解析は、患者登録を継続するかどうかを判断することを目的とし、（有効中止もあり得るが）主として無効中止を念頭に置き、原則として予定登録数の半数の登録が得られた時点以降に問い合わせを行う最初の定期モニタリングのデータを用いて行う。2 回目の中間解析は、無効中止だけではなく有効中止も念頭に置き、一定数のイベント数が生じた状態で行う。そのため、登録が終了した後、すべての登録患者のプロトコール治療が終了する時期、あるいは患者の治療前データのクリーニングが終了する時期を目処に、データセンターと研究事務局で相談した上で適切と思われる時期の定期モニタリングに合わせて行う。

試験ごとの中間解析の実施時期はプロトコールに規定する。例えば、患者登録中に実施する無効中止を念頭に置いた第 1 回中間解析は、両群の毒性の差やプロファイルを検討し、有効性の観点での無効中止を検討する必要性が高くなければ中間解析を延期する場合もあり得る旨を規定する。

5.7.2. 多重性の調整 Adjustment for multiplicity

中間解析と主たる解析の多重性の調整は、Lan & DeMets の α 消費関数を用いて調整し、群間の primary endpoint の差について統計学的有意性を調べる。α 消費関数として、O’Brien & Fleming タイプを用いることを標準とする (17)。試験によっては、SWOG の方法 (4)などを考慮する。

5.7.3. 判断規準 Decision criteria

優越性試験の場合、主たる解析により、標準治療に対する試験治療の primary endpoint での優越性が証明された場合、原則として試験を中止する（有効中止）。一方、試験治療群の primary endpoint が標準治療群のそれを下回っている場合には、検定による判断を行わず、総合的に試験中止の是非を検討することとする（無効中止）。どちらにも該当しない場合は、試験を継続する。

非劣性試験の場合、登録中の中間解析と登録終了後の中間解析で有効中止に関する判断規準

が異なる。登録中は両群の安全性のデータが揃っておらず、試験治療が実際に標準治療よりも毒性が軽いかどうかなどの判断がつかない場合がある。従って、登録中の中間解析では、標準治療に対する試験治療の primary endpoint での優越性が証明された場合にのみ、原則として試験を中止する（有効中止）。一方、登録終了後の中間解析では、安全性のデータも揃っていると考えられるため、標準治療に対する試験治療の primary endpoint での非劣性が証明された場合、原則として試験を中止する（有効中止）。

中間解析の実施時期によらず、標準治療群に対する試験治療群の primary endpoint のハザード比の点推定値がハザード比における非劣性マージンを超えて上回った場合（試験治療群の許容範囲を超えて悪い場合）には、試験を中止する（無効中止）。試験治療群の primary endpoint が標準治療群のそれを下回っている場合には、検定による判断を行わず、総合的に試験中止の是非を検討することとする。どちらにも該当しない場合は、試験を継続する。

参考文献 Reference

1. ICH. STATISTICAL PRINCIPLES FOR CLINICAL TRIALS E9 1999 [2020/3/9]. Available from: https://database.ich.org/sites/default/files/E9_Guideline.pdf.
2. ICH. GENERAL CONSIDERATIONS FOR CLINICAL TRIALS E8 1997 [2020/3/9]. Available from: https://database.ich.org/sites/default/files/E8_Guideline.pdf.
3. ICH. CHOICE OF CONTROL GROUP AND RELATED ISSUES IN CLINICAL TRIALS E10 2000 [2020/3/9]. Available from: https://database.ich.org/sites/default/files/E10_Guideline.pdf.
4. Green S, Benedetti J, Smith A, Crowley J. Clinical Trials in Oncology. 3rd Edition ed. Boca Raton: CRC Press; 2012.
5. Pocock SJ. Clinical Trials: A Practical Approach. 1 edition ed: Wiley; 1984.
6. Freedman B. Equipoise and the ethics of clinical research. The New England journal of medicine. 1987;317(3):141-5.
7. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). European journal of cancer. 2009;45(2):228-47.
8. Seymour L, Bogaerts J, Perrone A, Ford R, Schwartz LH, Mandrekar S, et al. iRECIST: guidelines for response criteria for use in trials testing immunotherapeutics. Lancet Oncol. 2017;18(3):e143-e52.
9. Dindo D, Demartines N, Clavien PA. Classification of surgical complications: a new proposal with evaluation in a cohort of 6336 patients and results of a survey. Annals of surgery. 2004;240(2):205-13.
10. Katayama H, Kurokawa Y, Nakamura K, Ito H, Kanemitsu Y, Masuda N, et al. Extended Clavien-Dindo classification of surgical complications: Japan Clinical Oncology Group postoperative complications criteria. Surg Today. 2016;46(6):668-85.
11. Bottomley A, Pe M, Sloan J, Basch E, Bonnetain F, Calvert M, et al. Analysing data from patient-reported outcome and quality of life endpoints for cancer clinical trials: a start in setting international standards. Lancet Oncol. 2016;17(11):e510-e4.
12. 厚生労働省. 平成 30 年 2 月 7 日医政発 0207 第 10 号／保発 0207 第 5 号「医薬品及び医療機器の費用対効果評価に関する取扱いについて」. 2018.
13. 福田敬. 中央社会保険医療協議会における費用対効果評価の分析ガイドライン 2019 [2020/3/9]. 第 2 版:[Available from: https://c2h.niph.go.jp/tools/guideline/guideline_ja.pdf.
14. ICH. ADDENDUM ON ESTIMANDS AND SENSITIVITY ANALYSIS IN CLINICAL TRIALS TO THE GUIDELINE ON STATISTICAL PRINCIPLES FOR CLINICAL TRIALS E9(R1) 2019 [2020/3/9]. Available from: https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf.
15. Gray RJ. A Class of K-Sample Tests for Comparing the Cumulative Incidence of a Competing Risk. The Annals of Statistics. 1988;16(3):1141-54.
16. Fine JP, Gray RJ. A Proportional Hazards Model for the Subdistribution of a Competing Risk. Journal of the American Statistical Association. 1999;94(446):496-509.
17. Lan KKG, DeMets DL. Discrete Sequential Boundaries for Clinical Trials. Biometrika. 1983;70(3):659-63.