

第10回データサイエンスラウンドテーブル会議 Oncology領域におけるestimand

PMDA
統数研医療健康データ科学研究センター
JPMA DS部会

担当ファシリテータ
塩田 駿介(PMDA)
野村 尚吾(東京大学)
加藤 凌輔(国立がん研究センター)
久力 洋(中外製薬株式会社)
作井 将(武田薬品工業)

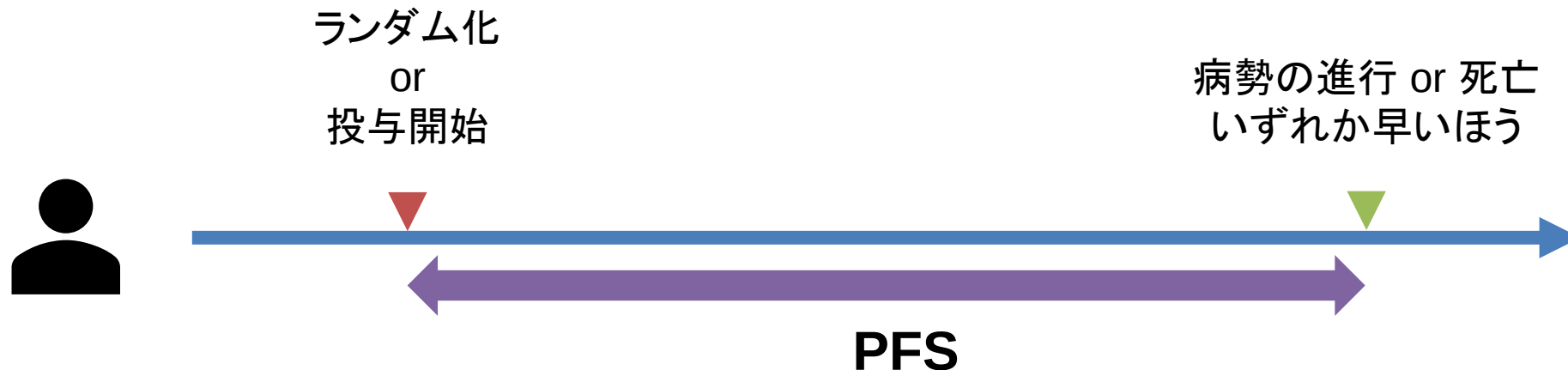
議論ポイント1

無増悪生存期間(PFS)を題材とした中間事象の取扱い

PFS (progression-free survival) とは

- PFS: Time-to-event型のエンドポイント(関心のあるイベントが起きるまでの期間を対象としたエンドポイント)のひとつ。ランダム化(単群試験の場合は投与開始日)から病勢の進行又は死亡のいずれか早い方までの期間を対象とする。
 - 病勢の進行は疾患ごとに定義され、例として固形がんではRECIST v1.1*がよく用いられる。

* Eisenhauer EA、et al.、2009、New Response Evaluation Criteria in Solid Tumours: Revised RECIST guideline (version 1.1)、Eur J Cancer、45(2):228-247.



がん臨床試験のCQと対応するEstimand

- Time-to-event型のエンドポイントは多くのがん臨床試験で中心的な役割を果たす。
- 一方で、Time-to-event型のエンドポイントでは、何をイベントとして扱い、何を打ち切りとするかの議論が先行してきた。
- Time-to-event型のエンドポイントでも関心のあるClinical Question (CQ) と一致するように、治療効果を定義する必要がある。近年になり打ち切りや欠測の理由となる中間事象とその中間事象に対するstrategyの観点で議論されることが増えている。

イベント/打切りの定義：打切表の導入

- FDA guidance (2007) では、Time-to-eventエンドポイントの定義を明確にするために、どのようなイベントをPFSイベントとするか、打切りとするかを事前規定することを推奨した。
 - 本ガイダンス以降、(特に製薬企業では)プロトコルやSAPに打切表を提示することが標準的になり、感度解析として打切りの取扱いを変更した解析が合わせて実施されてきた。
 - 2018年にガイダンスは更新されたが、PFSの計画・解析に関する内容はほとんど変更されていない。

Guidance for Industry Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics

- a. Table A represents a sensitivity analysis that only includes well-documented and verifiable progression events. Other data are censored. In Table A, the progression dates are:
- Based only on radiological assessments verified by an IRC. *Clinical progression* is not considered a progression endpoint.
 - Assigned to the first time when tumor progression was noted.
 - The date of death when the patient is closely followed. However, deaths occurring after two or more missed visits are censored at the last visit.

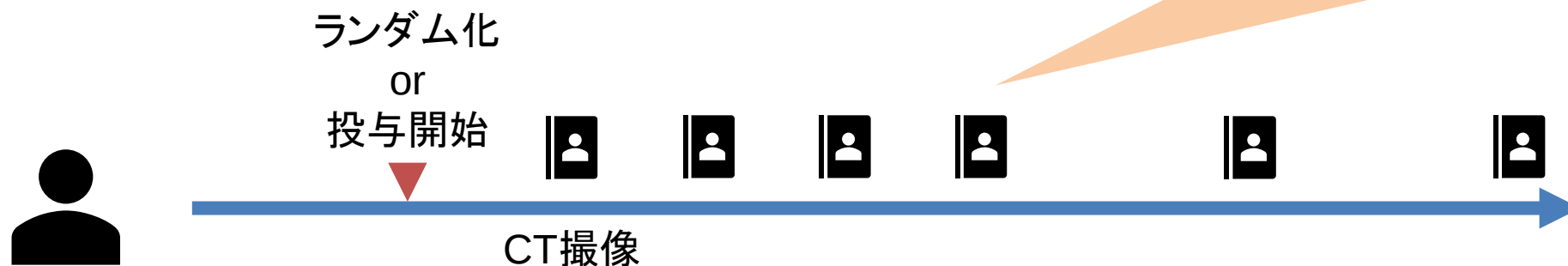
Table A. PFS 1 (includes documented progression only)

Situation	Date of Progression or Censoring	Outcome
No baseline tumor assessments	Randomization	Censored
Progression documented between scheduled visits	Earliest of: • Date of radiological assessment showing new lesion (if progression is based on new lesion); or • Date of last radiological assessment of measured lesions (if progression is based on increase in sum of measured lesions)	Progressed
No progression	Date of last radiological assessment of measured lesions	Censored
Treatment discontinuation for undocumented progression	Date of last radiological assessment of measured lesions	Censored
Treatment discontinuation for toxicity or other reason	Date of last radiological assessment of measured lesions	Censored
New anticancer treatment started	Date of last radiological assessment of measured lesions	Censored
Death before first PD assessment	Date of death	Progressed
Death between adequate assessment visits	Date of death	Progressed
Death or progression after more than one missed visit	Date of last radiological assessment of measured lesions	Censored

PFSに関する一般的な留意点

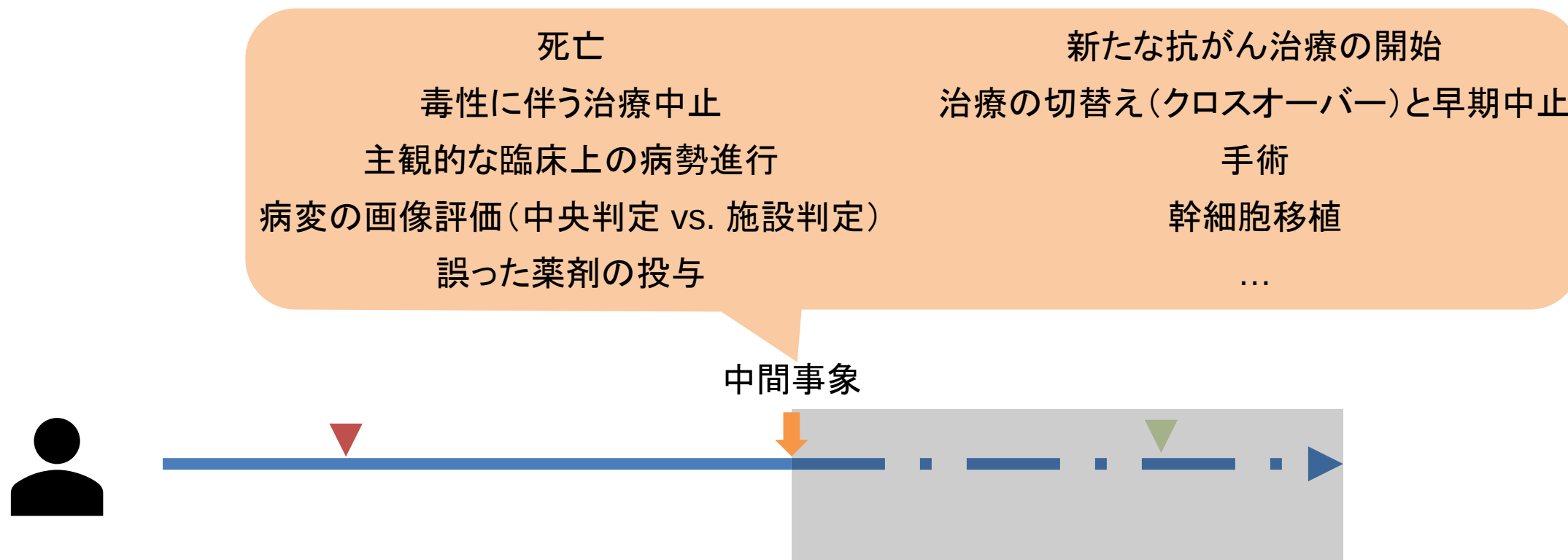
- 病勢の進行を確認するためには、「病変の観察 = CT(疾患によってはPET-CT)撮像」を実施し、進行しているかどうかを定期的に調べる必要がある。このため、一般に以下のような点に留意する必要がある。
 - 群間比較する場合は、撮像(病変の観察)の間隔を群間で統一する。
 - 「追加の(スケジュール外での)病変の観察」や「病変の観察をスキップしてしまう」ことを避ける必要がある。
 - 割り付けられた治療が評価者に分からないように盲検化する。
 - 読影者間の評価の違いを考慮し、主治医による評価だけでなく、独立した読影者(中央評価機関)による評価の実施も検討する。

しかしながら、どれだけ気を付けていても、毒性プロファイルの違いにより割り付けられた治療が推測できてしまったり、スケジュール外の追加での病変観察や病変観察が実施されないVisitの発生など、評価間隔が群間で不均衡となることがある。



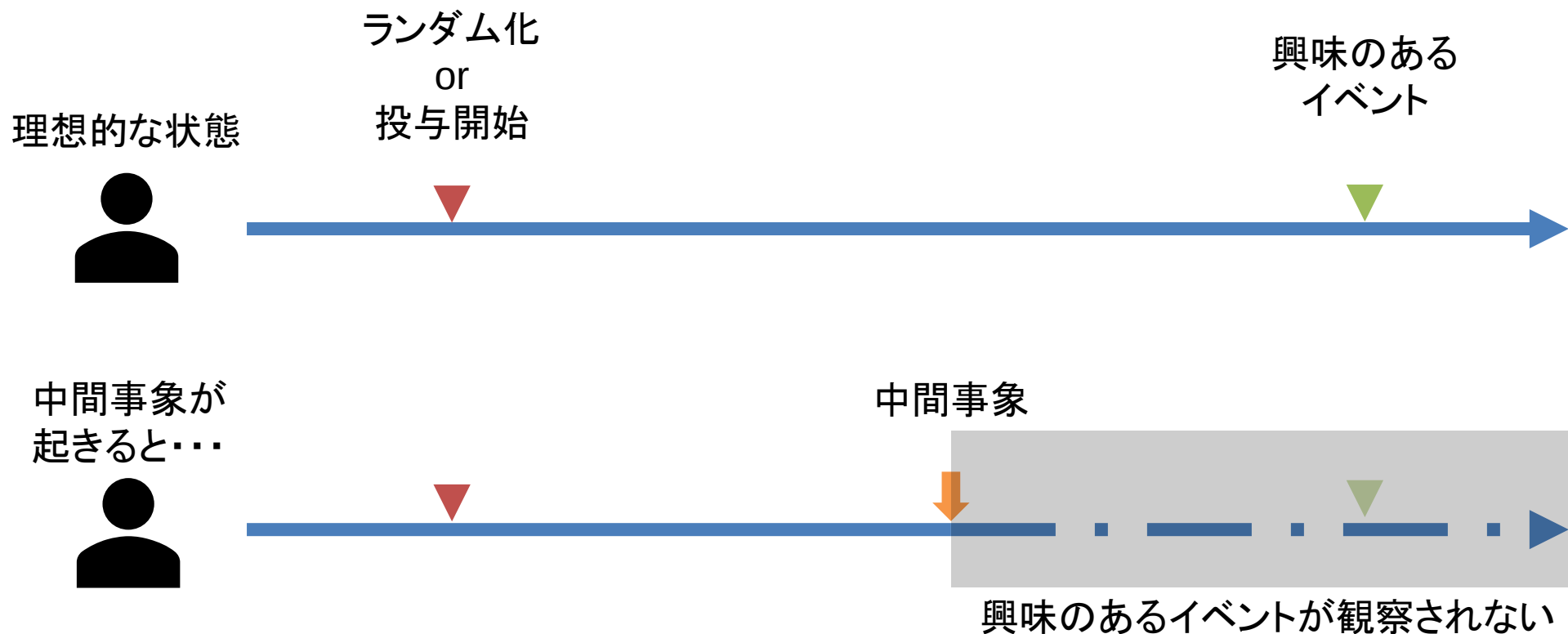
PFSを扱う際の間件事象の例

- 関心のあるCQによっては、以下のような事象が中間事象となる可能性がある。



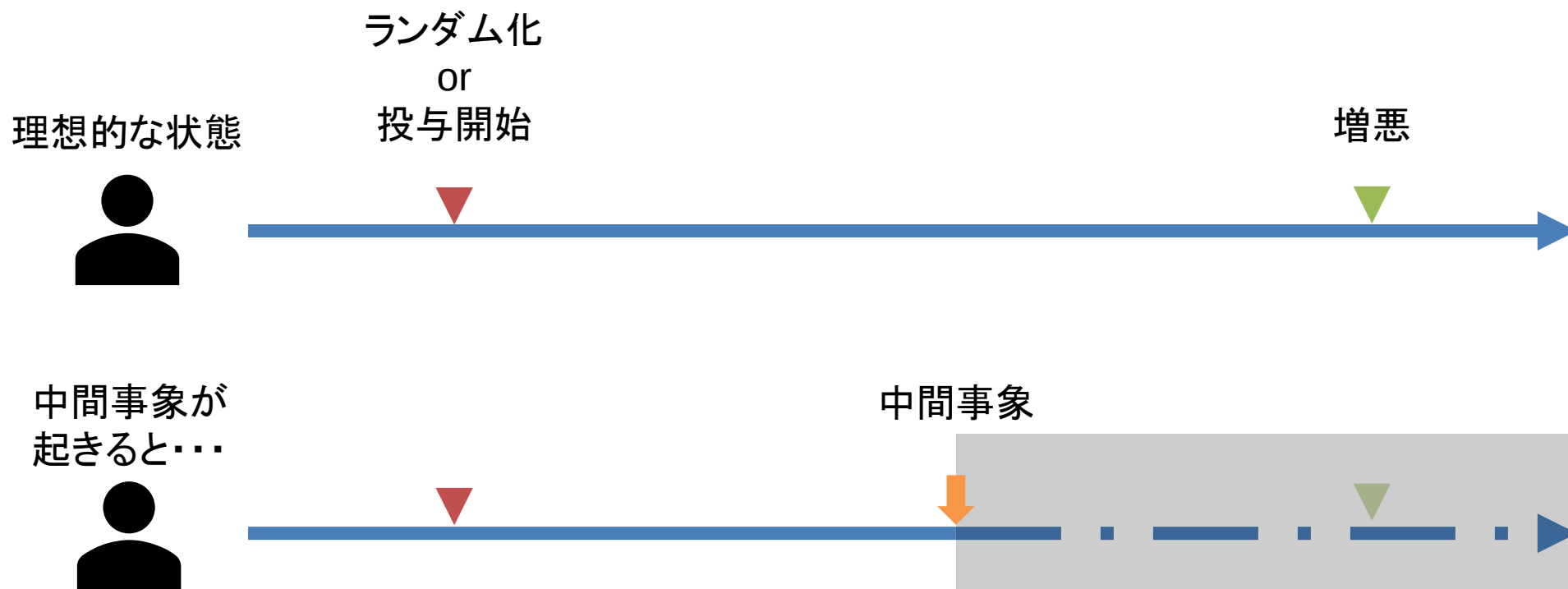
がん臨床試験によくある状況

- 目的とするイベントの前に中間事象が起きる可能性がある。
 - 一部の被験者では中間事象発現後も追跡が可能だが、一部の被験者では興味のあるイベントの観察が出来なくなることがある。



PFSの場合

- 病勢の進行を観察する前に中間事象が起きた場合、中間事象発現後も画像評価が(多くの場合)可能だが、feasibilityや手間を考慮して、中間事象発現後の画像評価を行わない試験が存在する。



中間事象に対するstrategyとその影響

- がん臨床試験において、中間事象の発生時又は中間事象を原因とした追跡不能が発生する可能性がある。その結果生じた打切りはnon-informativeではない可能性が高い。
 - 例：病勢の進行を確認する前の後治療の開始は、医師が患者の状態から総合的に判断している可能性がある
- 中間事象に対するstrategyを検討する際には、各strategyが試験デザインに与える影響を考慮することが有用である。

中間事象に対するstrategy	試験デザインへの影響
Treatment-policy strategy	治療 の定義
Hypothetical strategy	治療 の定義
Composite strategy	変数 の定義
While-on-treatment strategy	変数 の定義
Principal stratum strategy	集団 の定義

関心のあるCQにより中間事象への対応は変わる

- プロトコルで設定した治療ラインで試験治療を実施した場合の、当該治療後のすべての治療ラインを含めた治療効果を見たい(= ITT estimand、effectivenessに興味)



後治療を開始した時点以降も興味の対象 → 中間事象以降も評価を継続

- 試験治療の間に起きた増悪リスクの軽減を、限られた研究期間で評価したい(= 当該治療ラインでの薬剤の効果を見て意思決定したい、efficacyに興味)

後治療の開始までが興味の対象 → 後治療開始で打ち切り、病変の観察も中止



データを収集する範囲は何を見たいかに依存

プロトコル検討時の議論

後治療を開始した後も病変の観察を継続すべきでしょうか？

後治療開始後も病変観察を継続するのは難しいのですが



- 評価したい対象と実施可能性の観点でデータ収集範囲を検討する
 - Efficacy(試験治療で治療中の効果)を評価したい場合: 中間事象発現(例: 後治療開始)後は興味の対象でない可能性がある。
 - Effectiveness(治療方針の効果)を評価したい場合: 中間事象発現(例: 後治療開始)後もプロトコルで規定したスケジュールで病変の観察を継続する必要がある。
- データ収集範囲により、評価可能なestimandが制限される可能性も
 - Effectivenessを評価したくても、病変の観察を中間事象後に継続することが難しいこともある。→どこまで妥協可能かを事前に検討しておく必要がある。

打切表への反省

- FDA guidanceでは、例えばPFSを対象とする解析で特定の間事象に対して一律で打ち切りとするような記載が含まれていた。
- また、複数の打切表を事前規定して感度分析する場合にも、当該感度分析がこういった仮定からの逸脱（あるいはこういったestimand）を検討しているかを明示することは要求されていなかった。



- 関心のあるCQを評価する上で間事象を打ち切り（又はイベント）と扱うことが解釈上の問題を引き起こすか、またどのような仮定からの逸脱があったかを確認するための感度解析を実施するかを検討することが重要である。

a. Table A represents a sensitivity analysis that only includes well-documented and verifiable progression events. Other data are censored. In Table A, the progression dates are:

- Based only on radiological assessments verified by an IRC. *Clinical progression* is not considered a progression endpoint.
- Assigned to the first time when tumor progression was noted.
- The date of death when the patient is closely followed. However, deaths occurring after two or more missed visits are censored at the last visit.

Table A. PFS 1 (includes documented progression only)

Situation	Date of Progression or Censoring	Outcome
No baseline tumor assessments	Randomization	Censored
Progression documented between scheduled visits	Earliest of: • Date of radiological assessment showing new lesion (if progression is based on new lesion); or • Date of last radiological assessment of measured lesions (if progression is based on increase in sum of measured lesions)	Progressed
No progression	Date of last radiological assessment of measured lesions	Censored
Treatment discontinuation for undocumented progression	Date of last radiological assessment of measured lesions	Censored
Treatment discontinuation for toxicity or other reason	Date of last radiological assessment of measured lesions	Censored
New anticancer treatment started	Date of last radiological assessment of measured lesions	Censored
Death before first PD assessment	Date of death	Progressed
Death between adequate assessment visits	Date of death	Progressed
Death or progression after more than one missed visit	Date of last radiological assessment of measured lesions	Censored

以降の事例では、取り扱う論点

具体的な関心のあるCQ、estimandを踏まえ、以下のような論点を検討していく

1. どのような中間事象が起きうるか？ 中間事象への対処法は？
 - 想定される中間事象に対してどのような strategy を適用するか。
 - ある中間事象が起きた時点でイベント/打切りとする場合、それはどのような estimand が興味の対象であるためか。
2. データ収集をどのように行う(特にエンドポイントに関するデータや中間事象に関連するデータをどのように収集する)？
 - 2で議論したデータ収集方針で、1で議論したestimandを評価可能か。
3. 主たる解析はどのような解析になるか。
 - 打切りにつながる中間事象がある場合、当該事象をnoninformativeと扱うことができるか。
 - どのような感度解析を実施すれば、デザイン時の仮定が検討可能か。
 - どのような補足的解析が考えられるか。

事例検討

事例①

中央判定によるPFSを主要評価項目とする固形がん試験。
様々な試験の制約が存在する。

事例②

医師判定によるEFSを主要評価項目とする血液がん試験。
「治療不成功」の取扱いをめぐる論点について考える。

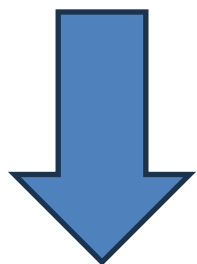
事例①

ある固形がん患者に対する薬剤Xの有効性・安全性を検討する 第Ⅲ相二重盲検プラセボ対照無作為化比較試験	
想定されるCQ① (今回はこちら)	後治療やクロスオーバーを含めた治療の経過全体を通じて、 疾患進行(PD)又は死亡を遅らせる薬剤Xの効果を評価したい
想定されるCQ②	後治療やクロスオーバーが実施される前までの、 PD又は死亡を遅らせる薬剤Xの効果を評価したい
主要評価項目	中央判定によるPFS
主要評価項目の定義	無作為化日から、中央判定によるPD 又は死亡のいずれか早い方までの期間

医師判定と中央判定：同じPDでも役割が違う

医師判定

- 担当医が各施設で実施
- 治療中止、後治療、クロスオーバーなどの臨床判断に用いられる



画像を収集

中央判定(本試験の主要評価項目)

- バイアス低減と標準化を目的として、独立した評価者が盲検下で実施

ポイント：医師判定PD後に中央判定ではnon-PDとなった場合、その後の中央判定データが欠けると、中央判定PFSの解釈が難しくなる

試験の制約（見直し可）

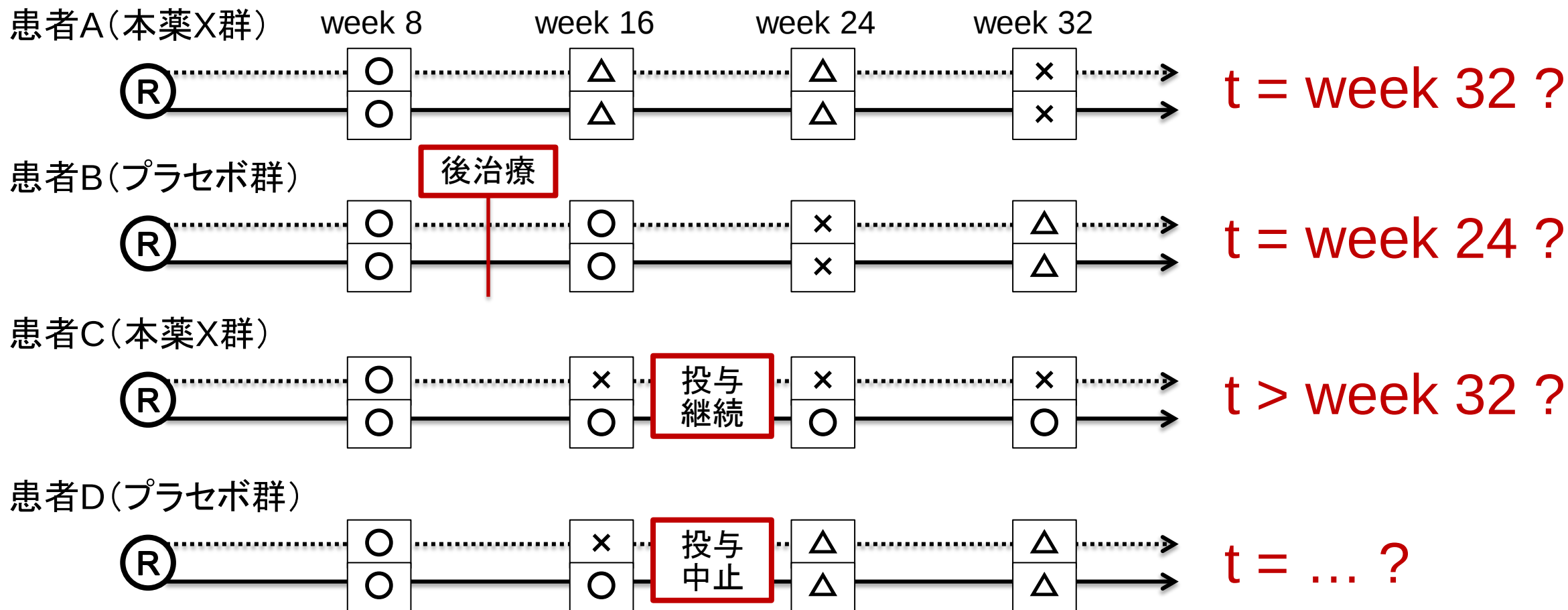
【共通】

- 腫瘍評価は8週ごとに実施
- 医師が必要と判断すれば、イベント観察前でも後治療を実施可
- 医師判定PD後、当該患者の盲検を解除する

【医師判定PD後の取扱い（群別）】

	薬剤X群	プラセボ群
治験薬の投与	臨床的ベネフィットありと判断されれば継続可	中止／薬剤Xへクロスオーバー可
画像評価	治験薬の投与が継続される場合は中央判定も継続	医師判定も実施しない（実務的な負荷やクロスオーバーのため）

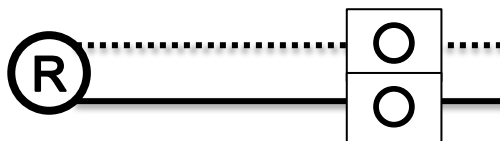
PFS評価における患者経過の例



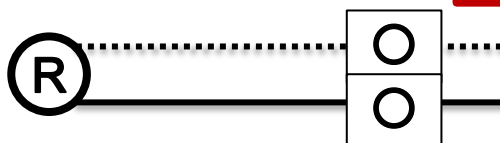
点線: 医師判定 実線: 中央判定 ○: non-PD Δ: 欠測又は評価未実施 ×: PD

PFS評価における患者経過の例

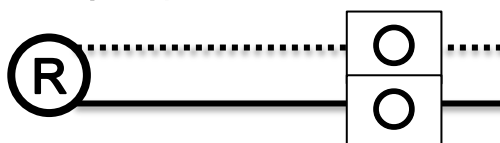
患者A(本薬X群) week 8



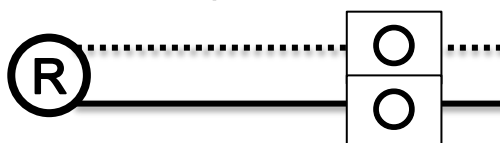
患者B(プラセボ群)



患者C(本薬X群)



患者D(プラセボ群)



この図で考えること(例)

- 各患者について、主解析ではどこをイベント又は打ち切りとするか。
- その取扱いは、どのCQ / estimandに対応しているか。
- 特に、医師判定PD後に中央判定non-PDとなり、その後の中央判定データが欠ける患者では、最終中央判定日で打ち切ることが妥当か。
- その打ち切りをnon-informativeとみなせるか。

点線: 医師判定 実

事例①に対する論点

番号	論点
論点①-1	医師判定PD、クロスオーバー、後治療開始、中央判定未実施等を中間事象としてどのように扱うか。 その扱いは、中央判定PFSを主要評価項目とするCQ/estimandと整合しているか。
論点①-2	スライド18に示した試験制約の下で、そのestimandはどこまで評価可能か。 どのデータ収集は妥協でき、どのデータ収集は妥協できないか。
論点①-3	主解析では、どの時点をイベント／打ち切りとするか。 中央判定データが欠測する場合、その打ち切りをnon-informativeとみなせるか。
論点①-4	医師判定PFS、中央判定PFS、打ち切り規則を変更した解析、医師判定PD後に中央判定non-PDとなった症例の内訳など、どのような感度分析・補足的解析が有用か。

以下のスライドに含まれる議論結果は当日の4グループの結果を要約したものとなります。
これらの内容はDSRT参加者の経験に基づく意見及び見解であり、各所属組織の公式のものではありません。

論点①-1：

医師判定PD、クロスオーバー、後治療開始、中央判定未実施等を中間事象としてどのように扱うか。その扱いは、中央判定PFSを主要評価項目とするCQ/estimandと整合しているか。

- ・ 中間事象の候補とその取扱い

医師判定PD、クロスオーバー、後治療開始についてはTreatment policy strategyとして扱い、画像判定の未実施（特に医師判定PD後の画像評価の未実施による中央判定の欠測や、連続した2回以上の評価欠測等）についてはHypothetical strategyとして扱う案が示された。また、後治療開始やクロスオーバー開始、治療中止、アドヒアランス不良、毒性による中止、有害事象、盲検解除、同意撤回なども中間事象の候補として挙げられた。

- ・ CQ/Estimandとの整合性

中央判定PFSを主要評価項目とするCQとの整合性については、理想的には整合しているとの意見があった一方で、実際の臨床判断には医師判定PDが用いられることから、治療方針を評価するのであれば医師判定PDを重視する考え方も示された。さらに、主要評価項目が中央判定PFSである以上、医師判定PD後であっても画像評価を継続すべきではないかとの意見や、治療群によって評価継続の取扱いが異なること自体がCQと整合するのか再検討が必要との指摘もあった。

論点①-2：

スライド18に示した試験制約の下で、そのestimandはどこまで評価可能か。
どのデータ収集は妥協でき、どのデータ収集は妥協できないか。

- ・現在の試験制約下でEstimandはどこまで評価可能か

本事例で想定するCQが「クロスオーバーや後治療を含めた治療方針全体の効果（effectiveness）」の評価であるならば、後治療開始後やクロスオーバー後であっても、PDが確認されるまで追跡を継続し、可能な限りデータを収集する必要があるとの意見が多く挙げられた。特に、Treatment policy strategyを採用する場合には、医師判定PD後を含めて継続的に情報を収集することが理想であり、データが収集されなければ当初想定したestimandを十分に評価できなくなるとの認識が共有された。

- ・データ収集で妥協できる点/できない点

一方で、実務上は追跡継続に伴う負荷や予算上の制約が大きく、データ収集頻度を減らす、医師判定PD後は評価間隔を広げる、一部症例を対象に収集するなどの工夫も考えられるとの意見があった。ただし、補足的解析であってもデータがなければ検討できないため、両群で可能な限り情報を収集することが重要とされた。

また、医師判定PD後の盲検解除や中央画像判定の未実施を、何らかの時点（例えば最終撮像日）で打ち切りとして扱う場合には、情報のある打ち切りにつながる可能性があり、撮像が実施されない期間にPDが発生した場合にはPFSを過大評価する方向のバイアスが生じ得ることが指摘された。さらに、二重盲検試験であっても有害事象から治療群が推測される可能性があるため、中央判定を主要評価項目とする意義は理解できる一方、医師判定との乖離によって評価が困難になるのであれば、医師判定の活用も検討の余地があるとの意見があった。

論点①-3：

主解析では、どの時点イベント／打ち切りとするか。

中央判定データが欠測する場合、その打ち切りをnon-informativeとみなせるか。

- ・ 主解析では、どこをイベント／打ち切りとするか

主解析では医師判定PD、クロスオーバー、後治療開始が発生した場合でも、可能な限り追跡を継続し、最終的にPDまたは死亡が確認された時点イベントとして扱うことが望ましいとの意見があった。一方、画像判定の未実施については適切な評価時点で打ち切りとする案が示された。また、PFS解析の主たる手法としてはKM曲線やCox回帰が一般的であるとの考えが示された。

- ・ 中央判定データが欠ける場合、その打ち切りをnon-informativeとみなせるか。

中央判定データの欠測については、non-informativeな打ち切りとみなすことは容易ではなく、欠測理由を可能な限り収集することが重要との認識が共有された。特に、同意撤回であっても背景理由を把握することが望ましいとの意見があった。また、クロスオーバーや後治療開始は中間事象である一方、中央判定未実施や評価スキップは欠測として整理すべきではないかとの議論があった。評価が連続して実施されない場合には、その期間中にPDが発生していても観察されず、PFSを過大評価する方向のバイアスが生じる可能性も指摘された。

さらに、effectivenessを評価したいのか、efficacyを評価したいのかをCQの段階で明確化することが重要であり、その目的に応じて追跡継続や打ち切りの取扱いを決定すべきとの意見があった。安易に打ち切りとするのではなく、後治療後やクロスオーバー後も含めて可能な限り評価を継続し、欠測理由を把握した上で結果を解釈することが重要との認識が共有された。

論点①-4：

医師判定PFS、中央判定PFS、打ち切り規則を変更した解析、医師判定PD後に中央判定non-PDとなった症例の内訳など、どのような感度分析・補足的解析が有用か。

主解析結果の頑健性を確認するため、医師判定PFSの収集・解析や、医師判定PDをイベントとして扱った感度解析が有用との意見があった。また、医師判定PDと中央判定PDの整合率や、医師判定PD後に中央判定non-PDとなった症例の内訳を確認することで、判定間の差異や結果への影響を把握できるのではないかとの指摘があった。

また打ち切りとなった理由を詳細に収集し、その内訳を補足的に提示することも重要との認識が共有された。打ち切りや欠測の発生状況を把握することで、主解析における仮定の妥当性を検討しやすくなると考えられた。

さらに、感度分析をどこまでSAPに事前規定するかは重要な論点であり、実務上はテンプレート等を用いてあらかじめ一定範囲の感度分析を計画することが多いとの意見があった。一方で、情報のある打ち切りによる影響を補正するならば、打ち切りメカニズムを回帰モデル等で要約し、打ち切りなく観測を続けている集団のデータを各時点でIPCWするなどが考えられるものの、解析手法が複雑化することや、事前規定が難しいことが指摘された。

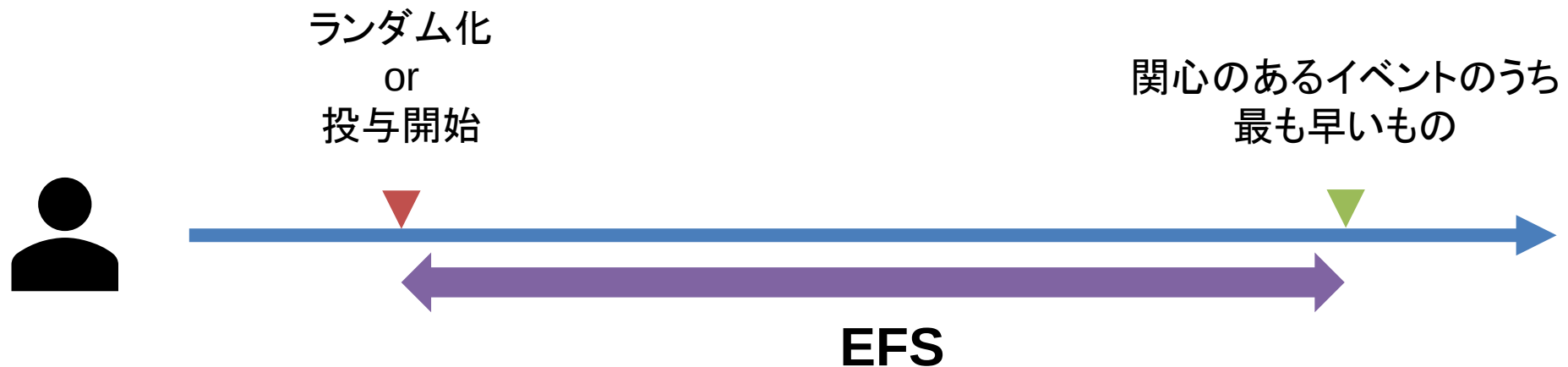
事例②

ある血液がん患者に対する薬剤Xの有効性・安全性を検討する 第Ⅲ相二重盲検プラセボ対照無作為化比較試験	
想定されるCQ①	寛解後の再発又は死亡を遅らせる効果を評価する
想定されるCQ② (今回はこちら)	上記CQ①に加え、寛解に至らないこと自体も 治療効果の欠如として評価する
主要評価項目	医師判定によるEFS
主要評価項目の定義	無作為化日から、 a.寛解後の再発 b.死亡 c.治療不成功(後述) のいずれかが最初に観察されるまでの期間

* 事例②は[イボシデニブ審査報告書](#)を題材に、グループディスカッションでの議論に合わせて改変した。

EFS (event-free survival) とは

- EFS: ランダム化 (単群試験の場合は投与開始日) から関心のあるイベントまでの期間
 - 関心のあるイベントは試験ごとに定義される。
 - イベントの例: 病勢の進行、手術、局所再発、遠隔再発、死亡 など



EFSの評価について

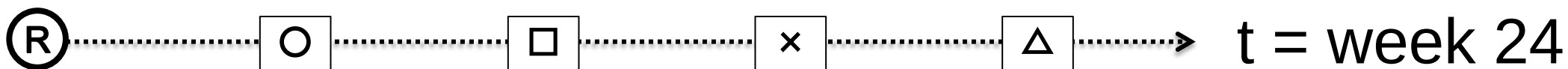
- 事前規定された評価基準に従い、8週ごとに寛解の有無及び再発の有無を評価する
- 評価には骨髓検査、末梢血検査等を用いる
- EFSイベントは、寛解後の再発、死亡、治療不成功と定義する
- 24週までに一度も寛解が認められなかった場合を治療不成功と定義する
- 本試験では、治療不成功を投与開始1日目、すなわちDay1のイベントとして扱うこととした

EFS評価における患者経過の例

患者A



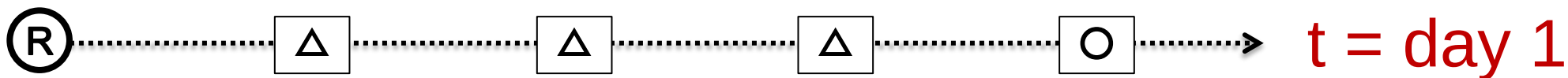
患者B



患者C



患者D



$\bigcirc$:寛解 Δ :寛解に至らず $\times$:再発 $\square$:再発に至らず

事例②に対する論点

番号	論点
論点②-1	治療不成功をEFSイベントに含めることは、CQと整合するか。 治療不成功をDay 1でイベントとする場合、どのような治療効果を推定していることになるか。 治療不成功は、複合ストラテジーとして扱うべきか、それとも他のストラテジーや別の評価項目等として扱うべきか。
論点②-2	24週までに寛解しないことを、どの時点のイベントとして扱うべきか。 Day 1、week 24、あるいは別の時点イベント発生時点とする場合、臨床的解釈はどう変わりうるか。
論点②-3	寛解有無、再発、死亡、治療不成功の内訳をどのように収集・提示すべきか。
論点②-4	主解析では治療不成功をどのように取り扱うか。
論点②-5	治療不成功をDay 1イベント、week 24イベント、打ち切り等とした感度解析は有用か。 EFSに加えて、寛解割合、寛解までの時間、再発までの時間、OS等を補足的に提示すべきか。
論点②-6	治療不成功がEFSイベントの大部分を占める場合、HRを主な要約指標とすることは妥当か。

以下のスライドに含まれる議論結果は当日の4グループの結果を要約したものとなります。
これらの内容はDSRT参加者の経験に基づく意見及び見解であり、各所属組織の公式のものではありません。

論点②-1：

治療不成功をEFSイベントに含めることは、CQと整合するか。治療不成功をDay 1でイベントとする場合、どのような治療効果を推定していることになるか。治療不成功は、複合ストラテジーとして扱うべきか、それとも他のストラテジーや別の評価項目等として扱うべきか。

- ・治療不成功をEFSイベントに含めることは、CQと整合するか。

治療不成功をEFSイベントに含めること自体の考え方は理解できるものの、Time-to-eventエンドポイントとして評価する場合にはCQとの整合性に違和感があるとの意見があった。特に、期間を評価する指標であるEFSにおいて、24週時点までの治療不成功をDay 1でイベントとして扱うことは解釈が難しく、EFSを過小評価する可能性もあるのではないかと指摘があった。

また、治療不成功をDay 1でイベントとするのではなく、24週時点でイベントとする方が自然ではないかとの意見や、イベント発生時点を変更した感度解析を補足的に実施することが有用ではないかとの提案もあった。治療不成功をどの時点でイベントとみなすかにより推定対象となる治療効果の意味が大きく変わるため、CQとの整合性を明確にした上で定義すべきとの認識が共有された。

- ・治療不成功をDay 1イベントとする場合、どのような治療効果を推定していることになるか

寛解に至らないことを臨床的に極めて重い失敗として位置付けるのであれば、Day 1でイベントとして扱う考え方にも一定の合理性があるのではないかと意見もみられた。さらに、本事例では寛解割合のような二値評価と、再発までの期間のような時間評価が混在しているように見えるとの指摘があり、治療不成功の重み付けをどのように考えるかが重要な論点として挙げられた。

- ・治療不成功は、複合ストラテジーとして扱うべきか、それとも他のストラテジーや別の評価項目等として扱うべきか
治療不成功をEFSに組み込むのであれば複合変数（Composite）ストラテジーとして捉える考え方が示された一方、寛解割合などを別の評価項目として扱う方が解釈しやすいのではないかと意見もあった。

論点②-2：

24週までに寛解しないことを、どの時点のイベントとして扱うべきか。

Day 1、week 24、あるいは別の時点イベント発生時点とする場合、臨床的解釈はどう変わりうるか。

- ・ 24週までに寛解しないことを、どの時点のイベントとして扱うべきか。

24週までに寛解しないことをどの時点のイベントとして扱うべきかについて、Day 1とweek 24で大きく解釈が異なるとの意見があった。week 24をイベント時点とする方法は自然な選択肢として考えられる一方で、「寛解しなかったこと」をある特定時点のイベントとして表現すること自体に違和感があるとの指摘もあった。

また、患者にとっては寛解の達成や寛解期間の延長が重要であり、その観点からはDay 1イベントとして扱う考え方にも一定の合理性があるとの意見があった。特に、「寛解しないこと」を最も重い治療不成功と位置付けるのであれば、Day 1イベントとして扱うことも臨床的には解釈可能ではないかとの見解が示された。

Day 1、week 24、あるいは別の時点イベント発生時点とする場合、臨床的解釈はどう変わりうるか。EFSは本来期間を評価するエンドポイントであり、治療不成功をDay 1でイベントとして組み込むことで、時間評価と治療不成功の評価が混在し、解釈が難しくなる可能性が指摘された。そのため、Day 1とweek 24のいずれを採用するかに加え、イベント時点を変更した補足的解析や感度解析を実施し、結果の頑健性を確認することも有用ではないかとの意見があった。

論点②-3：

寛解有無、再発、死亡、治療不成功の内訳をどのように収集・提示すべきか。

・ 寛解有無、再発、死亡、治療不成功の内訳をどのように収集・提示すべきか。

寛解率と再発・死亡を同一のEFSの中で評価するのではなく、それぞれを分けて収集・提示した方が解釈しやすいのではないかとの意見があった。補足的な評価として、寛解までの時間（Time to remission）を感度解析として提示することも有用ではないかとの意見があった。これにより、単に治療不成功をEFSイベントとして組み込むだけでは把握しにくい治療効果を、多面的に評価できる可能性があると考えられた。

論点②-4：

主解析では治療不成功をどのように取り扱うか。

- ・ 主解析では治療不成功をどのように取り扱うか。

治療不成功をEFSイベントとして主解析に含めることについて、発生割合が少ない場合は影響が限定的である一方、発生割合が多い場合には、Day 1へのイベントの集中やEFSイベントの内訳ごとの治療効果の違いなどにより、比例ハザード性の成立やハザード比の解釈可能性が問題となる場合があり、イベントとして組み込むことには慎重な意見があった。そのため、治療不成功は主解析のEFSイベントとするのではなく、副次評価項目として別途評価したり、Co-Primary endpoint（複数主要評価項目）として取り扱ったりする方法も考えられるのではないかとの意見が示された。

また、治療不成功を主解析に含める場合には、試験計画段階で発生頻度を十分に予測し、例数設計やシミュレーションを通じて解析への影響を検討しておくことが重要との認識が共有された。実際の事例では、治療不成功をDay 1でイベントとして扱った結果、両群で早期イベントが多く発生し、中央値が1日となったことも紹介され、そのような状況を事前に想定できていたのかという点も議論された。

論点②-5：

治療不成功をDay 1イベント、week 24イベント、打ち切り等とした感度解析は有用か。
EFSに加えて、寛解割合、寛解までの時間、再発までの時間、OS等を補足的に提示すべきか。

- ・ 治療不成功をDay 1イベント、week 24イベント、打ち切り等とした感度解析は有用か。
治療不成功をDay 1イベント、week 24イベント、打ち切りなど異なる設定で扱った場合、結果の解釈が変わり得ることから、EFSの頑健性を確認するために多様な感度解析を実施することが有用との意見があった。また、EFSのみでは治療効果の全体像を十分に把握できない可能性があるため、寛解割合、寛解までの時間、再発までの時間、OSなどの補足的な評価項目も併せて提示し、総合的に判断すべきとの認識が共有された。
- ・ EFSに加えて、寛解割合、寛解までの時間、再発までの時間、OS等を補足的に提示すべきか。
EFSについても時間経過に基づく解析だけでなく、特定時点（例：week 24）の評価や二値エンドポイントとしての評価を補足的に示す案が挙げられた。一方で、補足的解析はあくまで参考情報であり、主解析と同等の強い結論を導くものではないとの意見もあった。ただし、主解析では捉えきれない治療効果を補足的解析が示唆する可能性はあるため、試験の目的に応じて適切に活用することが重要との認識が共有された。また、承認取得を目的とする試験と、承認後に最適な治療戦略を検討する試験とでは、重視すべき評価項目が異なる可能性があるとの意見もあった。

論点②-6：

治療不成功がEFSイベントの大部分を占める場合、HRを主な要約指標とすることは妥当か。

治療不成功がEFSイベントの大部分を占める場合、HRを主な要約指標とすることは妥当か。

治療不成功がEFSイベントの大部分を占める場合、HRを主な要約指標とすることの妥当性について検討が必要との意見があった。特に、治療不成功が頻繁に発生する場合には比例ハザード性が成立しない可能性があり、まずKM曲線による結果の確認や比例ハザード性の検証を行うべきとの指摘があった。一方で、治療不成功の発生割合が少ない場合には、HRを主要な要約指標として用いることも一定程度妥当ではないかとの意見が示された。

また、HRでは情報が十分に反映されない可能性があるため、代替指標としてRMSTを用いる方が望ましいのではないかと意見もあった。ただし、これまで多くの試験でHRが広く用いられてきた経緯があり、実務上はHRを完全に否定することは容易ではないとの認識も共有された。

さらに、どの治療効果を評価したいのかというCQを明確にすることが重要であり、場合によっては治療不成功例を含めた集団だけでなく、寛解に到達した患者集団に限定したCQなどPrincipal stratum strategyなどの別のestimandや、解析上の検討が必要であるとの意見があった。

議論ポイント2

Terminal event(死亡など)を踏まえたPROの取扱い

なぜPRO/QOLを議論するのか

- がん治療では、生存期間や腫瘍縮小だけでなく、症状・機能・QOLの維持・改善も重要である
- PRO/QOLは、医療者評価では捉えにくい患者視点の症状・機能・生活への影響を補完する
- 規制当局や医療技術評価の場でも、治療価値を判断する情報としてPRO/QOLへの関心が高まっている

PRO/QOLは「補足的情報」から、治療価値を患者視点で解釈するための重要な情報になりつつある



PRO/QOL評価を臨床試験で活用する課題

- PRO/QOL評価を臨床試験に組み込む目的の明確化
- 使用する調査票・尺度の妥当性
- スコア変化を臨床的ベネフィットとしてどう解釈するか
- PRO endpoint、欠測、ICEの扱いを含む統計解析の標準化
- PRO/QOL評価を行った臨床試験の報告方法

本日の焦点

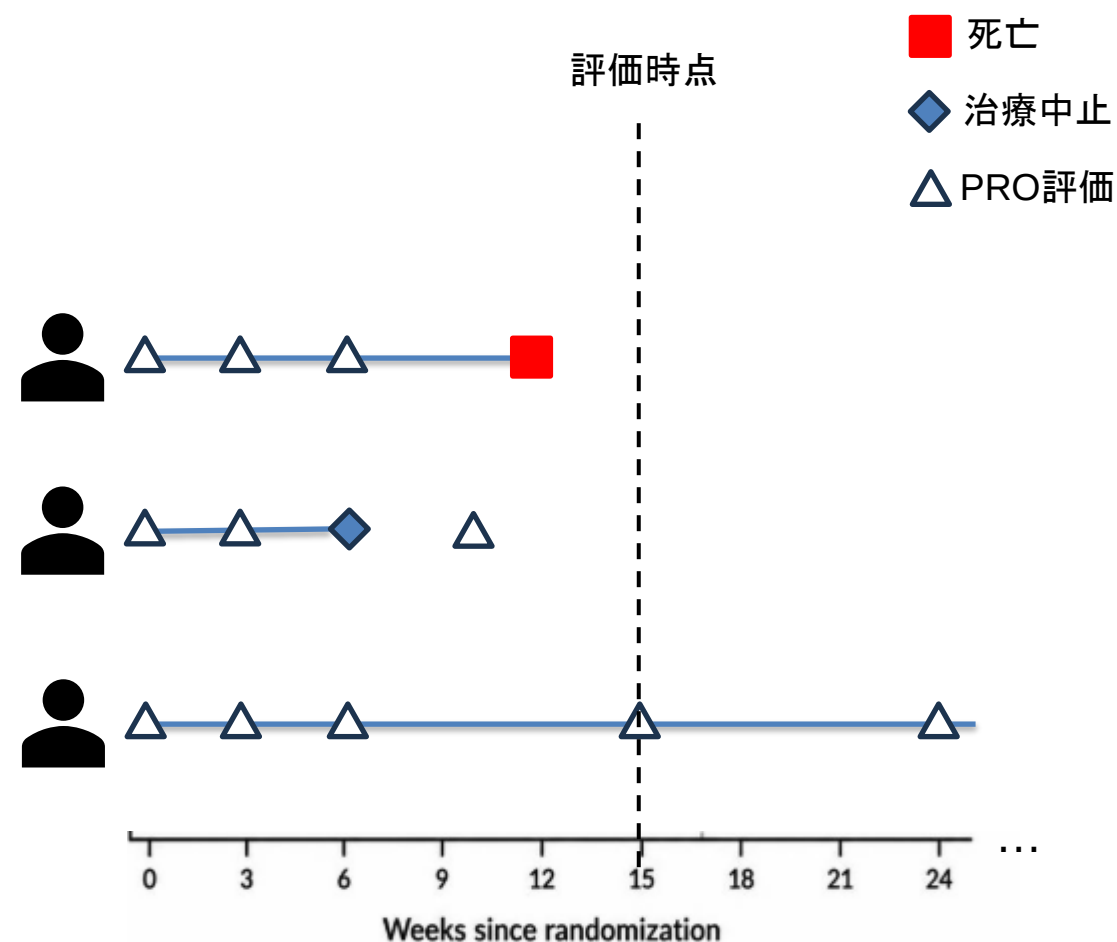
PRO/QOL評価を推進・否定する場ではなく、既存のPRO解析事例を材料に、PRO endpoint・ICE/欠測・解析方法・結果提示の整合性を議論する

PROを用いた臨床試験の計画から報告までに用いられるガイドライン

目的	ガイドライン
調査票の選択・妥当性確認	COSMIN
プロトコルへの組み込み	SPIRIT-PRO
統計解析	SISAQOL
結果の報告	CONSORT-PRO

事例紹介: KEYNOTE-024

- 試験デザイン: 第3相無作為化非盲検並行群間比較試験
- 対象疾患: 未治療の進行/転移性 NSCLC、PD-L1 TPS $\geq 50\%$ 、EGFR/ALK陰性
- 治療群: 試験治療: pembrolizumab 200 mg q3w、対照: プラチナ製剤
- PROの位置づけ: 探索的解析
- PRO評価項目:
 - QLQ-C30 GHS/QOL の baselineからweek 15 までの変化量
 - QLQ-LC13 の cough / chest pain / dyspnoea composite の time to deterioration
- PRO収集時点: サイクル 1-3 day 1。以後 9週ごと(治療中止後は治療中止時点及び中止後30日のフォローアップ時点で測定するが以降は収集しない)
- 連続量の解析: cLDA (constrained longitudinal data analysis)
- time-to-deteriorationの解析: 10点以上の悪化をイベント基準、Kaplan-Meier / stratified log-rank / Cox model、死亡は打ち切りとして扱う
- Responder割合の解析: 記述的な解析、欠測はMIで補完



事例紹介: KEYNOTE-024

	Pembrolizumab group (n=151)	Chemotherapy group (n=148)
QLQ-C30		
Baseline	145 (96%)	137 (93%)
Week 3	127 (84%)	122 (82%)
Compliance	127/144 (88%)	122/138 (88%)
Week 6	120 (79%)	110 (74%)
Compliance	120/138 (87%)	110/131 (84%)
Week 15	109 (72%)	92 (62%)
Compliance	109/129 (84%)	92/117 (79%)
Week 24	98 (65%)	75 (51%)
Compliance	98/111 (88%)	75/92 (82%)

Compliance was defined as the proportion of patients who completed the indicated questionnaire among those who were expected to complete it at each timepoint (excluding those patients missing by design).

	Pembrolizumab group	Chemotherapy group
Baseline		
Number of patients	145	137
Mean score (SD)	62.2 (22.3)	59.9 (22.3)
Week 15		
Number of patients	109	92
Mean score (SD)	71.0 (21.2)	63.7 (20.6)
Change from baseline*		
Number of patients	150	147
Least-squares mean score (95% CI)	6.9 (3.3 to 10.6)	-0.9 (-4.8 to 3.0)
QLQ-C30=European Organisation for the Research and Treatment of Cancer Quality of Life Questionnaire Core 30. GHS=global health status. QOL=quality of life. *Based on a constrained longitudinal data analysis model with the patient-reported outcome scores as the response variable, and treatment by study visit interaction and stratification factors as covariates.		
Table 2: Change in QLQ-C30 GHS/QOL score from baseline to week 15		

At week 15, we observed an improvement from baseline of 6.9 points (95% CI 3.3 to 10.6) for pembrolizumab and a decrease of 0.9 points (-4.8 to 3.0) for chemotherapy, **for a difference in least-squares means of 7.8 points (2.9–12.8; two sided nominal p=0.0020; table 2).**

事例紹介: KEYNOTE-024

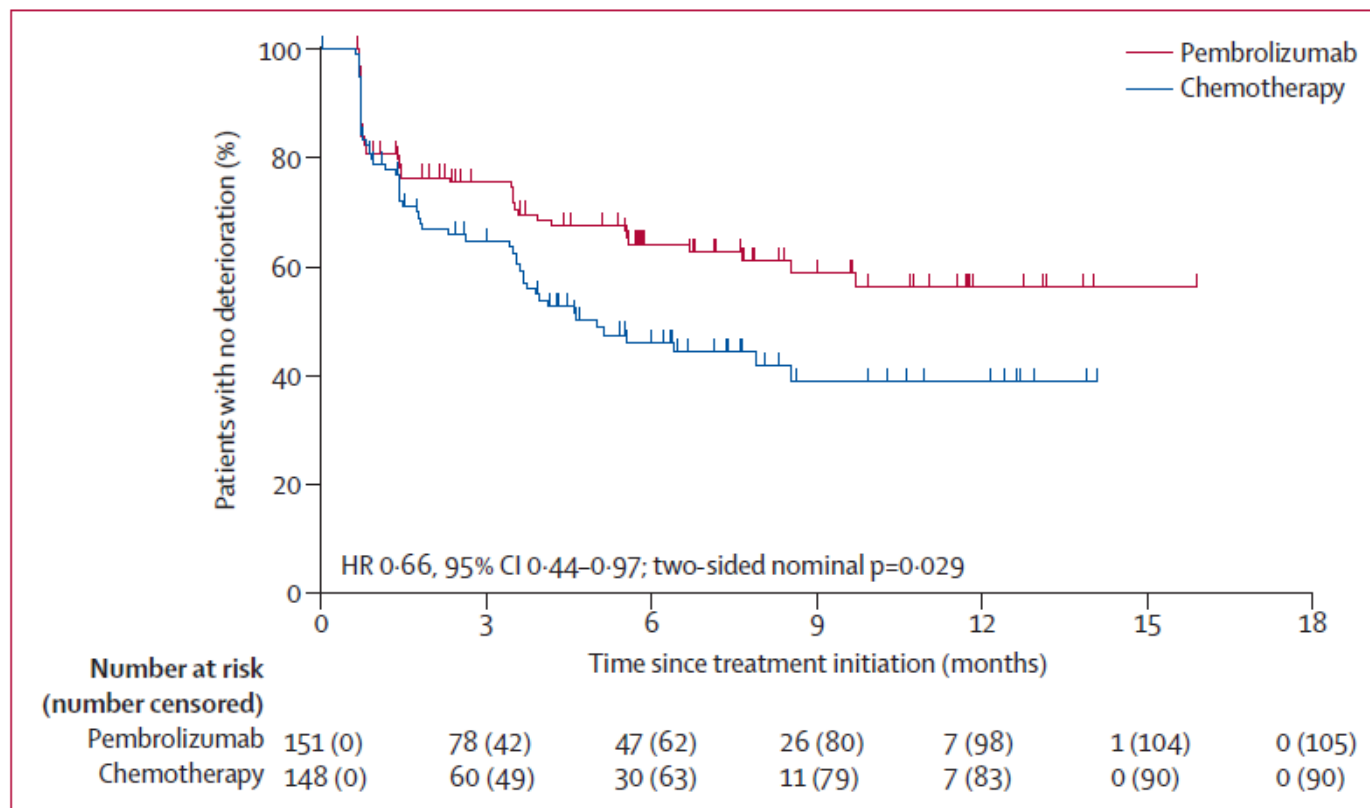


Figure 3: Time to deterioration of the composite of cough, chest pain, and dyspnoea in the QLQ-LC13. Deterioration defined as time to the first onset of a 10-point or greater decrease from baseline in cough, chest pain or dyspnoea, confirmed by a second adjacent 10-point or greater decrease from baseline in any of these three symptoms. QLQ-LC13=Quality of Life Questionnaire Lung Cancer 13 items. PRO=patient-reported outcome. HR=hazard ratio.

Brahmer et al. *Lancet Oncol.* 2017;18:1600–1609.

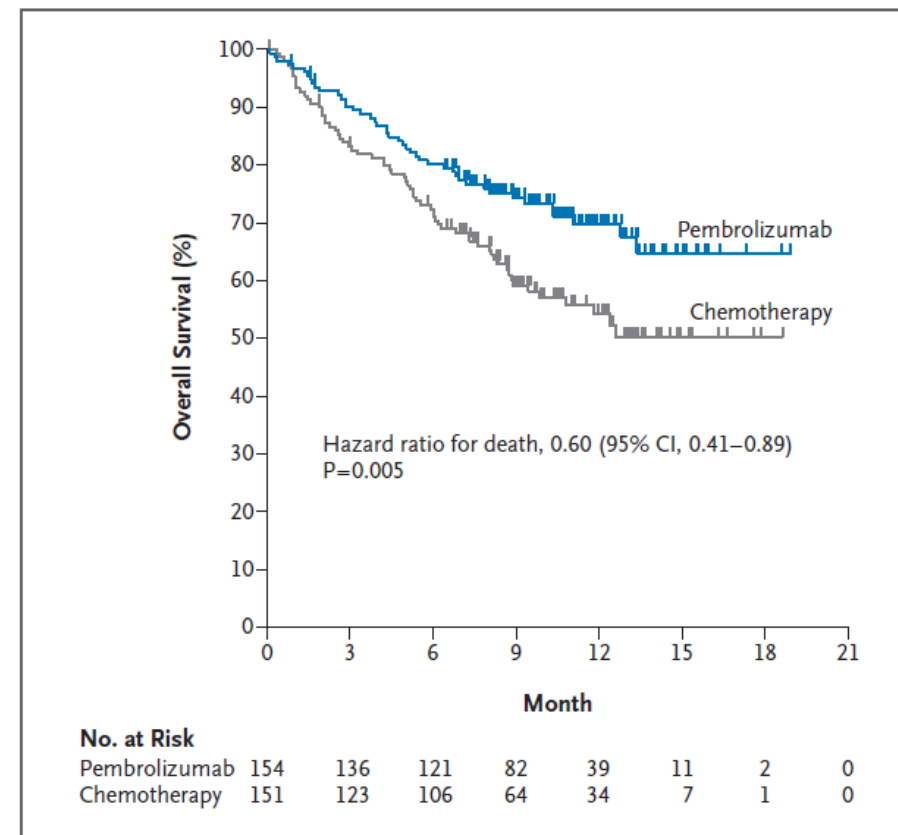


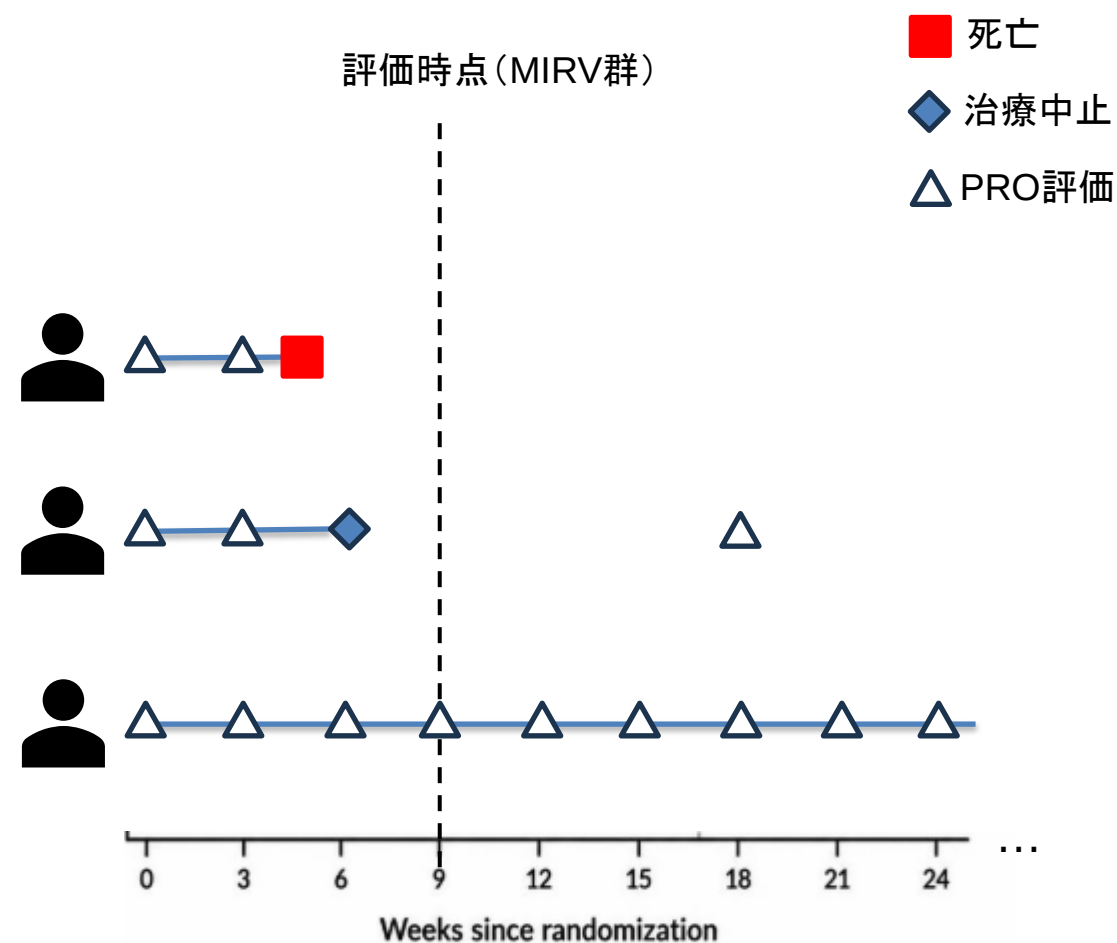
Figure 2. Overall Survival in the Intention-to-Treat Population.

Shown are Kaplan–Meier estimates of overall survival, according to treatment group. Tick marks represent data censored at the last time the patient was known to be alive. The intention-to-treat population included all patients who underwent randomization.

Reck M et al. *N Engl J Med* 2016; 375: 1823–1833.

事例紹介: MIRASOL

- 試験デザイン: 第3相無作為化非盲検並行群間比較試験
- 対象疾患: FR α 高発現のプラチナ抵抗性再発卵巣がん(高悪性度漿液性を含む、卵巣がん・原発性腹膜がんも含む)。
- 治療群: 試験治療: MIRV、対照: 化学療法
- PROの位置づけ: 副次評価項目/探索的評価項目
- PRO評価項目:
 - 副次評価項目: QLQ-OV28 腹部・消化器症状スコアが week 8 or 9 で15.0点以上改善した患者の割合
 - 他の評価項目: 他のQLQ-OV28サブスケール等
- PRO収集時点: 各治療サイクルのday 1(MIRV3週毎、化学療法4週毎)。24週以降は12週毎、治療中止後は3か月ごとに収集
- Responder割合の解析: CMH検定、評価時点前(9週)の後治療開始/PD/死亡はNon-Responderとして扱う
- 連続量の解析: MMRM
- time to definitive clinically meaningful deteriorationの解析: Kaplan-Meier / stratified log-rank / Cox model、死亡は打ち切りとして扱う



事例紹介: MIRASOL

- The primary patient-reported outcome endpoint responder analysis found that the proportion of patients who reported an improvement in QLQ-OV28 abdominal and gastrointestinal scores at week 8 or 9 was 34 (21.0%; 95% CI 15.0–28.1) of 162 patients in the MIRV group versus 23 (15.3%; 10.0–22.1) of 150 patients treated with investigator's choice of chemotherapy (OR 1.5 [95% CI 0.8–2.6]; $p=0.26$); these analyses included patients who did not have week 8 or 9 assessments but had either progression or death before day 71.

(A) Proportion of patients in each treatment group with improved abdominal and gastrointestinal symptoms based on either a 15.0-point threshold (primary analysis) or 11.1-point threshold (sensitivity analysis). Due to the constraints of the QLQ-OV28 abdominal and gastrointestinal symptom subscale, the smallest actual value above 15.0 is 16.67 points, a 3-unit raw score change.

*Nominal p value calculated from the Cochran-Mantel-Haenszel test stratified by the randomisation strata.

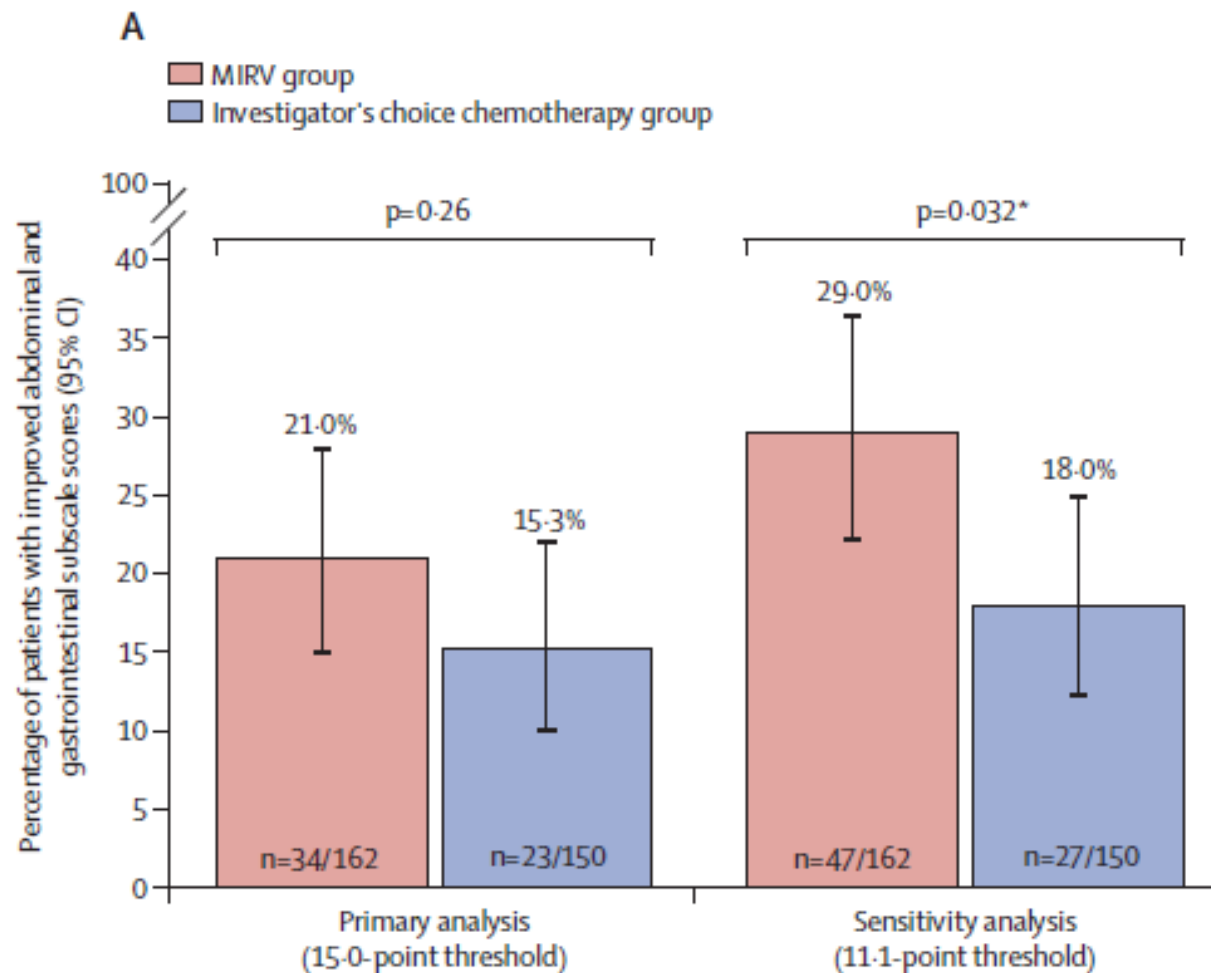


Figure 1: Responder analysis at week 8 or 9 and mixed model repeated measures change from baseline scores for QLQ-OV28 abdominal and gastrointestinal symptom subscale

事例紹介: MIRASOL

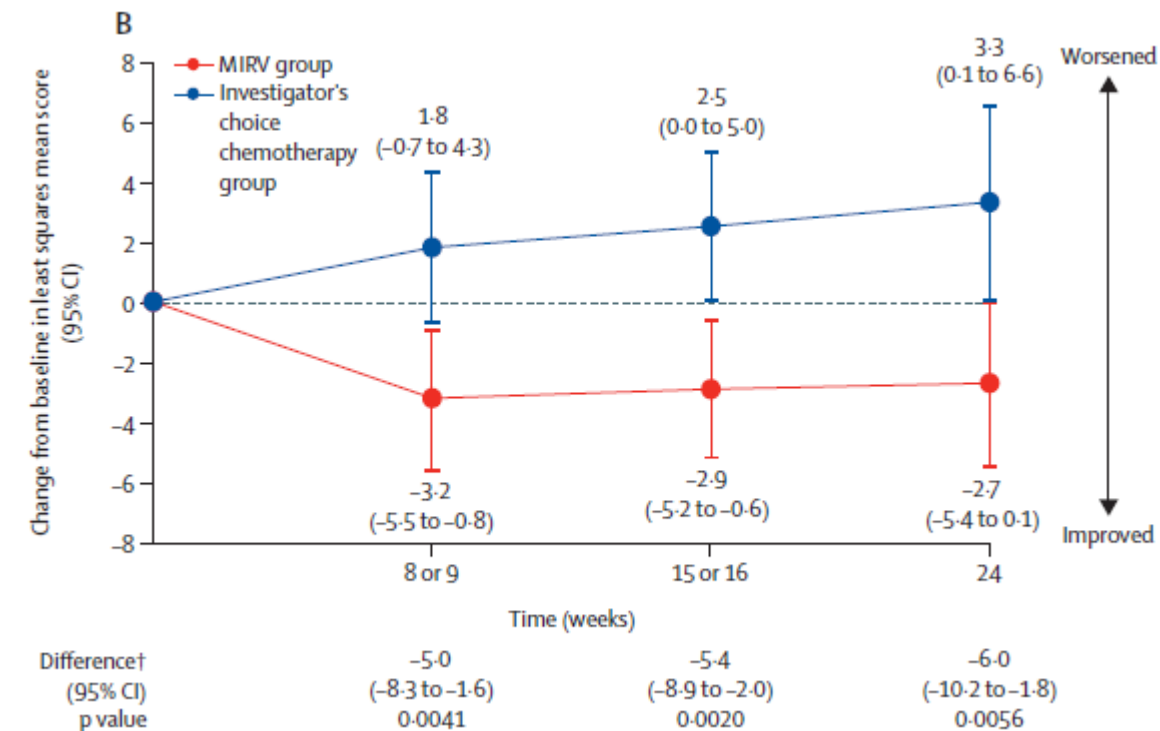


Figure 1: Responder analysis at week 8 or 9 and mixed model repeated measures change from baseline scores for QLQ-OV28 abdominal and gastrointestinal symptom subscale

(B) Mixed model repeated measures change from baseline scores of the QLQ-OV28 abdominal and gastrointestinal subscale by treatment group in the intention-to-treat population; decreased LS mean score from baseline indicates improved level of symptoms

†Difference is between LS mean for MIRV versus investigator's choice of chemotherapy.

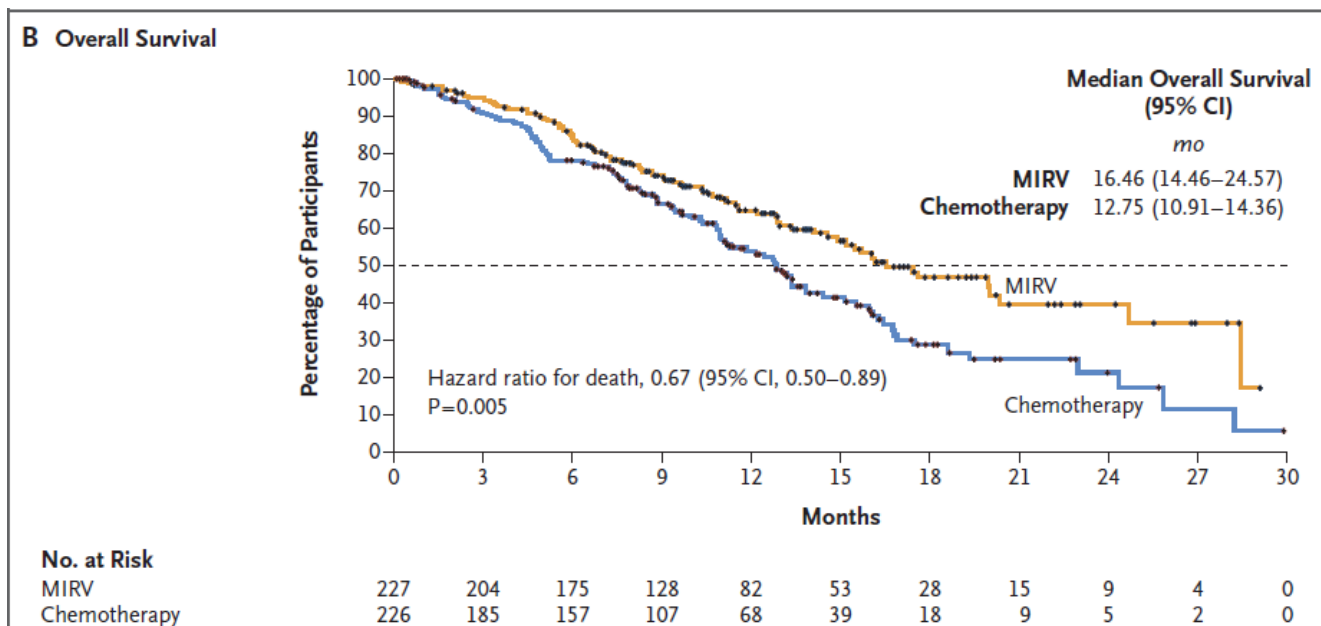


Figure 1. Efficacy Findings with MIRV as Compared with the Investigator's Choice of Chemotherapy.

Shown are the Kaplan–Meier estimates of investigator-assessed progression-free survival (the primary end point [Panel A]) and of overall survival (a key secondary end point [Panel B]) in the mirvetuximab soravtansine-gynx (MIRV) group and the chemotherapy group. The restricted mean progression-free survival at 12 months was 6.13 months (95% CI, 5.62 to 6.64) in the MIRV group and 4.72 months (95% CI, 4.21 to 5.23) in the chemotherapy group. Analyses were performed in the intention-to-treat population.

Moore KN, et al. *N Engl J Med* 2023; **389**: 2162–74.

Van Gorp et al. *Lancet Oncol.* 2025;26:503–515.

PRO/QOLの結果を解釈する上で困ること

どのPRO endpointで評価しているか？

- 変化量 / Responder割合 / Time to deteriorationで、見ているものが異なる
- 「**改善**」を見ているのか、「**維持**」を見ているのか
- 特定時点なのか、期間全体なのか

どの患者状態のPROを見ているか？

- **死亡後のPROは存在しない**
- PD後、治療中止後、後治療開始後のPROを含めるか
- While-alive / while-treatmentでは、**評価可能な患者集団に依存**し得る

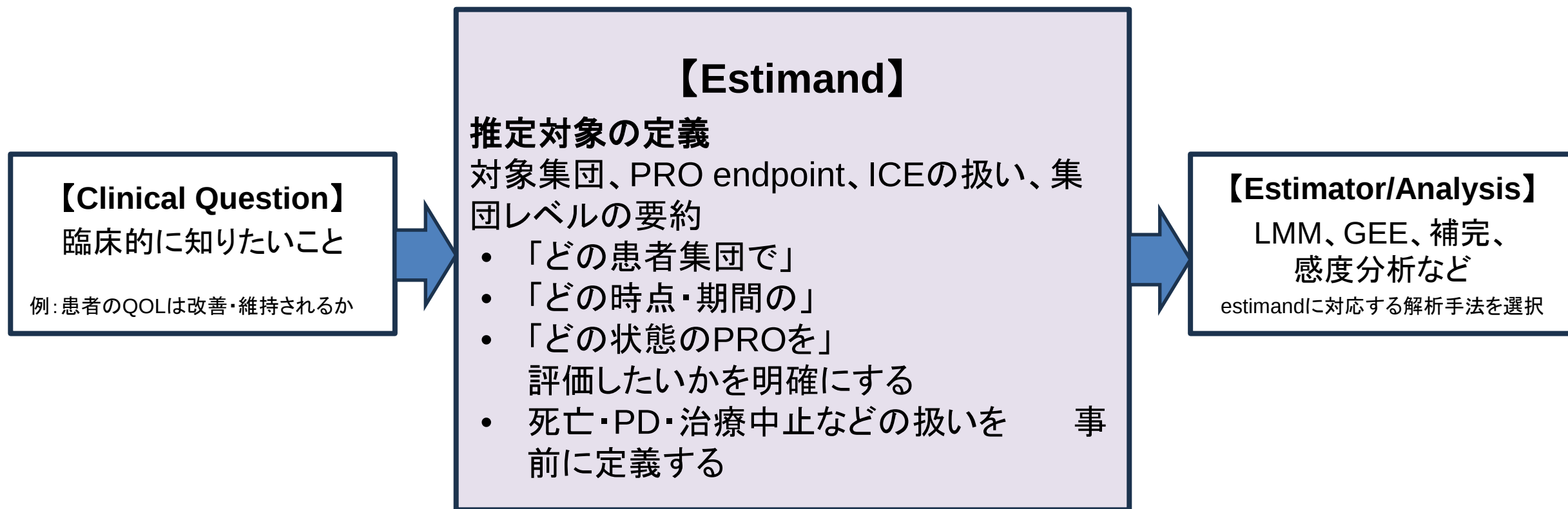
データがない理由は何か？

- **運用上の欠測**か、**ICEに関連する未取得**か
- 単調脱落か、非単調欠測（一時的未取得）か
- completion rate / available data rateの低下をどう読むか



CQ / estimand に戻って、評価時点・endpoint・ICE
未取得理由・結果提示を整理する必要がある

PRO解析でも「何を推定したいか」から始める



定義したestimandに対応するように、
解析方法・欠測データの扱い・感度分析を選択する

SISAQOL-IMI: PRO評価の国際的コンセンサス

Setting International Standards in Analysing Patient-Reported Outcomes and Quality of Life Endpoints in Cancer
Clinical Trials–Innovative Medicines Initiative

Interactive Table

研究目的とPRO変数に応じて、関連する推奨を参照できるWebツール



規制当局・アカデミア・企業・患者団体等、41組織の専門家が参画



146件のコンセンサス
推奨事項を整備

Guidebook

推奨事項の開発背景とツールの使用方法を詳述した包括的なガイド



Plain language materials

患者・非専門家向けに専門用語を平易に解説したチェックリスト等の資料



Glossary

PRO及びestimandに関連する用語の定義を統一したオンライン辞書



RCTおよび単群試験におけるPROデータの収集・解析・可視化について、**論**
点を体系的に確認するための推奨事項が整備されている

SISAQOL-IMI Interactive Table:

推奨を確認するためのツール

Randomised controlled trial Single arm trial

Confirmatory Objective:
Evaluate clinical benefit
Superiority Equivalence / Non-inferiority

Descriptive Objective:
Describe patient perspective

Magnitude of PRO (change) score at time t	Recommendations →	Recommendations →	Recommendations →
Responder with PRO improvement at time t	Recommendations →	Recommendations →	Recommendations →
Responder with PRO worsening at time t	Recommendations →	Recommendations →	Recommendations →
Time to PRO improvement	Recommendations →	Recommendations →	Recommendations →
Time to PRO worsening	Recommendations →	Recommendations →	Recommendations →
Overall mean or median PRO scores over a specified time frame	Recommendations →	Recommendations →	Recommendations →

Confirmatory - Superiority - Magnitude of PRO (change) score at time t

For RCTs, the PRO objective is to evaluate clinical benefit by showing that patients in the treatment arm reported better PRO (change) scores at a specific time point (time t) than the reference arm.

Estimands framework

- The choice of each estimand should depend on the PRO objective of the clinical trial and on the research question. [EstFrame1_GEN_RCT]

Explanations

Population

Treatment

PRO variable of interest

Handling of intercurrent events

Population-level summary

PRO score interpretation thresholds

Study design considerations

Analysis considerations

Results communication and visualisation

SISAQOL-IMI Interactive Table

<https://www.sisaqol-imi.org/interactive-table/>

💡 ただし、適用には事前の整理が必要

① 何を評価したいか:

PROを評価する理由/PRO概念/
評価時点・期間/PRO endpoint

② どう解釈可能にするか:

ICEと欠測の扱い/
結果解釈に必要な補足情報

本日のDSRTで扱う統計的論点

本日の主な想定: 優越性検証を目的とするRCT

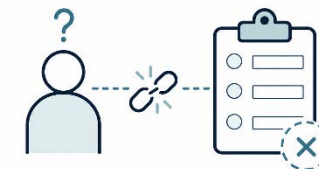
terminal eventの取扱い

terminal event
(死亡など)後のPRO
をどう扱うか



PRO未取得・ICE後PRO

未取得理由とICE後データ
の扱いを整理する



PRO endpoint

PROで何を比較・評価
するか



ケーススタディ

endpoint・strategy・未取
得理由・結果提示を統合し
て検討する



SISAQOL-IMI全体を網羅するのではなく、
PRO estimandに関する統計的論点に絞って議論する。

PROで何を比較・評価するか

PRO endpointの選択により、必要な仮定・解析方法・結果の解釈が変わる

PRO endpoint	評価したいこと	主な論点
特定時点の変化量	ある時点でのPRO差	その時点まで評価可能な患者に依存
Responder割合	臨床的に意味のある改善・悪化の割合	閾値設定が必要
Time to worsening	PRO悪化までの時間	死亡・増悪・欠測/打ち切りの扱い
期間平均/中央値	一定期間全体のPRO推移	観察期間差・欠測の影響



本日は実例で用いられているPRO endpointを主な題材とし、terminal event・PRO未取得・解析方法・結果提示を議論する

論点1: terminal eventをどう扱うか

terminal event後のPROは存在しないことも踏まえ、strategyと解析方法等を考える必要あり

Strategy	Clinical Question	主な論点
Hypothetical	死亡しなかった場合のPROはどうだったか	強い仮定が必要
Composite	死亡も含めた患者状態をどう評価するか	QOL悪化と死亡の影響が混在
While alive	生存中のPROはどう推移したか	(治療)群間で生存集団に違いがみられる場合、解釈に注意
Principal stratum	一定期間生存し得る患者集団でのPROはどうだったか	対象集団の定義・識別が難しい



本日は、PRO endpointごとにterminal eventをどのstrategyで扱うとCQに整合するか、その実装方法と限界を議論する。

論点1: terminal eventをどう扱うか

治療開始後のPROを、死亡などのterminal eventを踏まえて
どのように比較・解釈すべきか？

※PD、治療中止、後治療開始などの死亡以外のICEは論点2以降で議論する

- 論点1-1
 - 死亡後のPROが存在しない状況で、何を評価対象とするか？
- 論点1-2
 - PRO endpointに応じて、terminal eventのstrategyはどう変わり得るか？
 - ✓ 特定時点の変化量、Responder割合、Time to worsening etc.
- 論点1-3
 - 選択したstrategyは、予定している解析方法で実装できているか？
- 論点1-4
 - 結果を解釈するために、QOLスコア以外に何を併せて提示すべきか？

論点1: terminal eventをどう扱うか

手法	解析手法の概要	死亡の扱い例
Time To Deterioration	悪化の閾値を設定して、Time-to-event型の解析(Cox回帰、ログランク検定等)を行う。	死亡とQOL悪化を同列のイベントとして扱う。死亡を打ち切りとして処理する場合もある。
Semi-competing risk analysis	悪化の閾値を設定して、競合リスクに対するTime-to-event型の解析を行う。 原因別(死亡/QOL悪化)にハザードを推定して、累積発生関数で要約する。 累積発生関数を直接モデル化するFine-Gray modelやGray検定も利用できる。	死亡を競合事象として扱い、QOL悪化を別に見る
MMRM, cLDA, GLMM	悪化/改善の閾値を設定して二値データ型に変換(GLMM)またはPROのスコアをそのまま解析に用いる(MMRM, cLDA)。 各時点で別々に評価するのではなく、経時的な推移を考慮して解析を行う。	死亡後データはMARの仮定のもとで暗黙に扱うが、死亡後のQOLは本質的に定義不能
Responder Analysis	悪化/改善の閾値を設定して二値データ型に変換する。 経時的な推移は考慮せず、特定の時点で改善割合の差やオッズ比等の指標で比較する。	Non-Responderとして扱う

論点1: terminal eventをどう扱うか

前提: 死亡後のPROは存在しないため、死亡をどう扱うかは補完方法ではなくCQの問題

戦略A: While-alive

【暗黙の仮定】

生存中のデータのみに
意味がある

【臨床的解釈】

治療開始後の「予想され
るQOL」を記述する目的
に整合する

【限界・リスク】

群間で生存率に差があ
る場合、生存バイアスが
生じ、純粋なQOLの比
較が困難になる

戦略B: Composite

【暗黙の仮定】

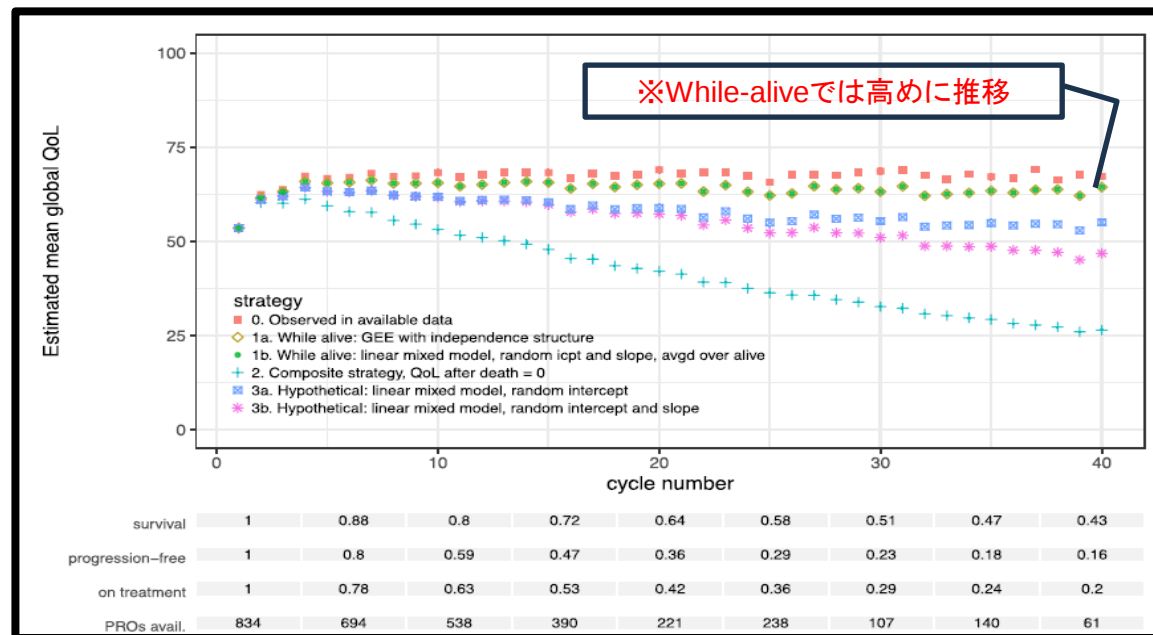
死亡をPROの一部(例:
スコア0)として扱う

【臨床的解釈】

生存とQOLの複合的な
評価(生存期間も加味し
た総合的ベネフィット)

【限界・リスク】

「死亡」という事象に特
定の単一スコアを割り当
てることの臨床的妥当
性に議論の余地あり



Thomassen et al. *BMC Med Res Methodol* 2024;24:290, Fig. 5.

- Compositeの一例として死亡を0点として扱う場合、平均QOLが急低下し得る
- 低下の要因が「QOL悪化」か「死亡増加」かは単独では判別しにくい
- 生存割合・PROデータ残存状況との併読が重要

論点2: PRO未取得理由・ICE後PROの整理



未取得理由を確認し、欠測・ICE関連データ・ICE後PROの扱いを整理したうえで、実装方法と結果提示を議論する

論点2: PRO未取得理由・ICE後PROの整理

主たる関心は治療中のPROだが、治療中止も症状・QOLと関連し得るため、
治療中止後PROも一定範囲で収集する場面を想定する

- 論点2-1
 - PRO未取得の理由をどこまでCRFで収集すべきであり、収集上の障壁はあるか？
- 論点2-2
 - PRO未取得理由を、解析・解釈に利用できる粒度でどう分類するか？
 - ✓ 例: 運用上の欠測／治療経過・患者状態に関連する未取得(ICE候補)／分類困難
- 論点2-3
 - 分類が難しい未取得理由、非単調欠測、ICE後に取得されたPROを、主解析・感度分析・補足的解析でどう扱うべきか？
- 論点2-4
 - QOLスコアの結果を解釈するために、未取得理由の内訳、データ残存状況、ICE後PROの取得状況をどう提示すべきか？

※一度PROが未取得となった後、次時点以降で再度取得される欠測パターン

ケーススタディ

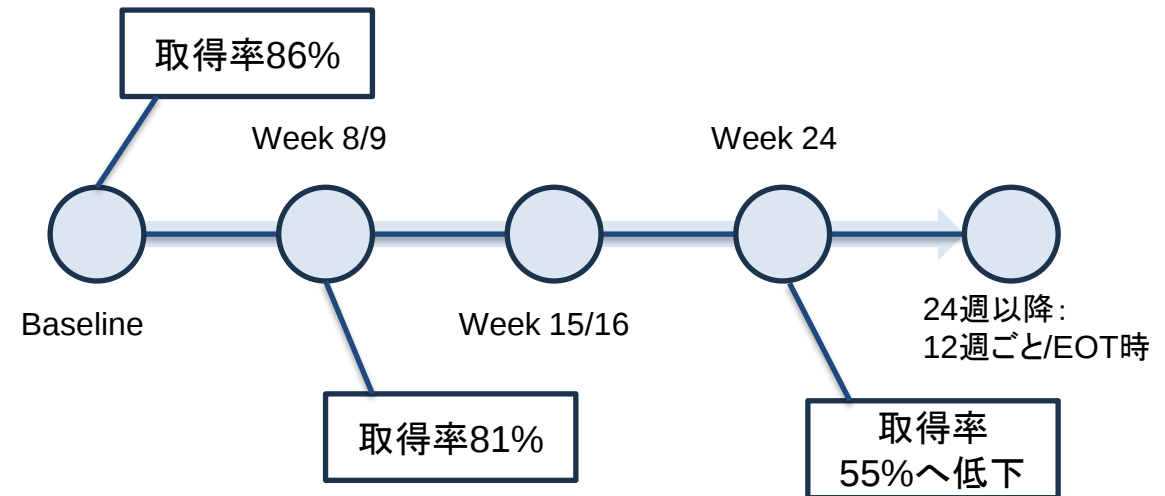
MIRASOLの事例をestimandの観点から

- 対象:
 - FR α 陽性プラチナ抵抗性卵巣がん
- 比較:
 - 新薬(MIRV群) vs 医師選択化学療法
- 試験デザイン:
 - 第Ⅲ相、open-label、RCT
- 疾患背景:
 - 予後不良であり、標準化学療法の有効性が低く、毒性が蓄積しやすい。
 - ベースラインで多くの患者が中等度～重度の腹部症状を抱える
- PROの意義:
 - 生存期間や腫瘍制御に加え、症状負担を悪化させず維持・改善できるかは、患者視点での治療価値を考えるうえで重要である

Van Gorp et al. *Lancet Oncol.* 2025;26:503–515.

評価ツール: EORTC QLQ-OV28 (腹部・消化器症状等)
QLQ-C30

※MIRV群は3週ごと、化学療法群は4週ごと



本ケースでの議論用のCQ

新治療は医師選択化学療法と比べて、治療開始後の症状・QOLを**改善または維持**できるか？

MIRASOLで用いられたPRO endpointと解析

PRO endpoint	評価したいこと	実施解析	論文・SAP上の扱い
Responder割合	week 8/9時点の症状改善	CMH検定	QLQ-OV28腹部/GI症状subscaleで 15.0点以上改善 をkey secondary PRO endpointとして評価。階層検定ではORR、OSの後にPRO endpointを検定。
変化量	各時点の平均スコア変化	MMRM	Restricted maximum likelihood-based MMRM。反応変数は各PRO評価時点のbaselineからの変化量。Week 8/9付近の群間差はweek 8.5で評価。
Time to deterioration	PROの臨床的に意味のある悪化までの時間	KM / stratified log-rank / Cox model	悪化未発生例は最終PRO評価日で打ち切り。 死亡・PDはdeterioration eventではない(打ち切り)。

欠測データの取扱い

- item missing: 質問票内の一部項目未回答。scoring manualに従い、算出可能な場合はsubscale scoreを算出
- form missing: 評価時点の質問票全体が未取得。form levelの補完は実施せず

SISAQOL/estimandの観点

PRO endpoint	論文・SAP上の扱い	主な議論点
Responder割合	Week 8/9改善。感度解析でPD/死亡/新規治療開始の扱いを変更	CQが「症状改善」なら、PD/死亡/新規治療開始を非改善に含めることは自然か。内訳提示は十分か。
変化量	観測データとモデル仮定に基づく平均変化	死亡・PD・治療中止による未取得がある場合、MMRMで推定される平均変化をどの患者集団・どの状態のPROとして解釈できるか。MAR仮定でよいのか。
Time to deterioration	死亡・PDはeventではなく、悪化未発生例は最終PRO評価日で打ち切り	死亡・PDをeventにしない解析は「PRO悪化が観測された患者における悪化までの時間」を中心に見ていることにならないか。死亡・PDによる打ち切りが非情報的とみなせるか。

ケーススタディ：主な議論点

新治療は医師選択化学療法と比べて、治療開始後の症状・QOLを改善または維持できるか？

- 論点3-1
 - Responder割合は「改善」は評価できるが、「維持」まで評価できるか？
- 論点3-2
 - PD・死亡・新規抗がん治療開始を非改善に含めることはCQと整合するか？
- 論点3-3
 - 「維持」まで評価するには、MIRASOLのMMRM・TTDで十分か？追加すべき解析・提示はあるか？
- 論点3-4
 - 死亡・PD・治療中止による未取得は、MMRM・TTDの結果解釈にどう影響するか？
- 論点3-5
 - 結果解釈のために、QOLスコア以外に何を併せて提示すべきか？

論点の再掲 及び 議論結果の要約

以下のスライドに含まれる議論結果は当日の4グループの結果を要約したものとなります。
これらの内容はDSRT参加者の経験に基づく意見及び見解であり、各所属組織の公式のものではありません。

論点1: terminal eventをどう扱うか

治療開始後のPROを、死亡などのterminal eventを踏まえて
どのように比較・解釈すべきか？

※PD、治療中止、後治療開始などの死亡以外のICEは論点2以降で議論する

- 論点1-1
 - 死亡後のPROが存在しない状況で、何を評価対象とするか？
- 論点1-2
 - PRO endpointに応じて、terminal eventのstrategyはどう変わり得るか？
 - ✓ 特定時点の変化量、Responder割合、Time to worsening etc.
- 論点1-3
 - 選択したstrategyは、予定している解析方法で実装できているか？
- 論点1-4
 - 結果を解釈するために、QOLスコア以外に何を併せて提示すべきか？

論点1-1：

死亡後のPROが存在しない状況で、何を評価対象とするか？

各グループともに死亡後PROの扱いは「死亡した患者のスコアとしてどのような値を代入するか」という解析上の問題ではなく、「死亡を含めて何を患者ベネフィットとして評価したいのか」を検討することが重要であるという点で一致していた。

PRO評価におけるterminal eventの扱いとして、以下のようなCQが挙げられた。

- ・ 生存中の患者のPRO/QOLがどの程度変化（改善）するか。
- ・ 死亡を含めた総合状態としてPRO/QOLを評価した際にどの程度変化（改善）するか。
- ・ 仮に死亡しなかった場合に、PRO/QOLはどのように変化（改善）するか。
- ・ 一定期間生存し得る患者でPRO/QOLはどのように変化（改善）するか。

論点1-2：

PRO endpointに応じて、terminal eventのstrategyはどう変わり得るか？

- ・変化量として扱う場合

連続量として扱う場合、死亡後のスコアが存在しないため、どのように数値化するかが最も議論となった。MMRM等を用いる場合は、死亡後の未観測値をMAR仮定のもとで暗黙に扱うことになり、実質的に「仮に死亡しなかった場合のQOL」を評価するhypothetical strategyに近い解釈となり得る。一方で、死亡を0点又は最悪値として扱う場合は、死亡をQOL評価に組み込むcomposite strategyと整理できる可能性があるが、そのスコアの臨床的妥当性や結果の解釈には慎重さが必要である。

- ・二値エンドポイント（例：Responder割合）として扱う場合

Responder解析では、死亡前にresponseが確認されなかった患者をnon-responderとして扱うcomposite strategyが当日参加者の中で最も受け入れられやすい方針となった。この取扱いは死亡を「PROスコア上の非改善」と同一視するというより、患者状態の悪化の指標として死亡を解釈したものと考えられる。ただし、このような場合にnon-responderには死亡、PD、後治療開始、運用上の欠測、単なる非改善が混在し得るため、集計表ではnon-responderの内訳を併せて示す必要があるのではないかという意見が挙がった。

- ・Time-to-event エンドポイント（例：time to worsening）として扱う場合

Time to worseningでは、死亡の扱いがestimandの意味を直接規定する。死亡をイベントとすれば「QOL悪化又は死亡までの時間」を評価するcomposite strategyとなり、死亡を大きな悪化又は治療不成功として含める解釈になる。一方、死亡を打切りとすれば「生存中に観測されたQOL悪化までの時間」を評価するwhile-alive strategyに近いが、死亡が情報のない打切りとみなせるかが課題となる。さらに、死亡を競合リスクとして扱えば、QOL悪化と死亡を分けて評価でき、結果解釈の透明性は高まるが、累積発生関数やイベント内訳などの提示が必要となるとの指摘があった。ただし競合リスクとして解析を実施する場合には、群間で死亡した被験者の割合に差があると解釈が難しくなるという限界も挙げられていた。

論点1-3：

選択したstrategyは、予定している解析方法で実装できているか？

- ・ 連続量としての解析（MMRM等）

連続量の解析では、死亡後の扱いとして「直前までの観測値を利用する」あるいは「死亡後を0として扱う」などの実装案が挙げられた。しかし、MMRMなどのモデルが前提とする欠測仮定と、死亡後データの補完方法との間に整合性があるのかは慎重に検討すべきという意見があった。

- ・ Responder割合の解析

Responder割合については、解析対象者数（N）をそのまま用いて評価する方法により、選択したstrategyを比較的明確に実装できるとの意見があった。特に、死亡をNon-responderとして扱うかどうかを事前に定義することで、Composite strategyやWhile on treatment strategyなど、想定するestimandを比較的直接的に解析へ反映できるとの認識が示された。

- ・ Time to worsening

Time to worseningで死亡を打ち切りとして扱う方法は、While on treatment strategyの考え方と整合しており、予定している解析手法で実装可能との意見があった。一方で、結果を適切に解釈するためには、死亡やその他の打ち切りの内訳・発生状況を併せて確認することが重要との指摘があった。また、「生存中の患者のPROを評価する」というClinical Question（CQ）が明確であれば、死亡をイベントとして扱わないことは自然な選択であるとの意見もあった。

論点1-4：

結果を解釈するために、QOLスコア以外に何を併せて提示すべきか？

各グループの議論では、補足情報として以下のような案が挙がった。

- ・ 生存に関する情報
OS、時点ごとの死亡割合
- ・ 治療状況に関する情報
時点ごとの後治療の状況、時点ごとの治療完遂割合、総投与量等が群間で異なっていないか
- ・ Terminal Eventに関する情報
Terminal Event（死亡）数、死因の内訳
- ・ 欠測、打ち切りに関する情報
欠測理由、打ち切りの理由（同意撤回など）、打ち切りの割合
- ・ PROデータ取得状況に関する情報
各時点のCompliance割合
- ・ 患者集団の特性・背景に関する情報
背景情報の確認
- ・ QOL推移の詳細・可視化に関する情報
生存している人と死亡した人でその時点までの変化量の推移
QOLの個別の推移図、欠測理由を含めた一覧表
- ・ Responder解析の補足情報
Non-responderの理由内訳

論点2: PRO未取得理由・ICE後PROの整理

主たる関心は治療中のPROだが、治療中止も症状・QOLと関連し得るため、
治療中止後PROも一定範囲で収集する場面を想定する

- 論点2-1
 - PRO未取得の理由をどこまでCRFで収集すべきであり、収集上の障壁はあるか？
- 論点2-2
 - PRO未取得理由を、解析・解釈に利用できる粒度でどう分類するか？
 - ✓ 例: 運用上の欠測／治療経過・患者状態に関連する未取得(ICE候補)／分類困難
- 論点2-3
 - 分類が難しい未取得理由、非単調欠測、ICE後に取得されたPROを、主解析・感度分析・補足的解析でどう扱うべきか？
- 論点2-4
 - QOLスコアの結果を解釈するために、未取得理由の内訳、データ残存状況、ICE後PROの取得状況をどう提示すべきか？

※一度PROが未取得となった後、次時点以降で再度取得される欠測パターン

論点2-1：

PRO未取得の理由をどこまでCRFで収集すべきであり、収集上の障壁はあるか？

参加者からは、PRO未取得の理由は結果解釈や欠測データの評価に重要であるため、可能な範囲で収集することが望ましいとの意見があった。特に、有害事象、検査値異常、患者状態の悪化など、治療経過と関連する要因によってPROが未取得となる可能性があり、増悪や有害事象がQOL低下やPRO未取得につながっている可能性を評価するためにも未取得理由の把握は有用であると考えられた。

一方で、未取得理由の詳細な収集は患者や医療機関の負担増加につながることから、運用上の限界も指摘された。実際に、患者状態不良による医師判断、有害事象、来院不能、質問票の紛失、記録忘れ、患者による回答拒否、ePROシステム上の問題など、多様な未取得理由が想定され、すべてを詳細に収集することは現実的ではないとの意見があった。また、収集項目の増加は患者負担の増大や回答品質の低下、さらなる欠測増加につながる可能性があり、収集頻度や評価間隔の調整によって負担軽減を図る工夫も必要と考えられた。

そのため、解析や解釈に必要な粒度を確保しつつ、患者負担とのバランスを考慮した収集項目の設定が重要との認識が共有された。最低限、死亡による未取得かそれ以外の未取得かを区別することが望ましいとの意見があり、実務的には主要な未取得理由を選択肢としてあらかじめ列挙し、それ以外は「Other」として収集する方法が提案された。また「Other」の自由記載情報については将来的にアルゴリズム等を活用して分類・整理することの有用性も議論された。

論点2-2：

PRO未取得理由を、解析・解釈に利用できる粒度でどう分類するか？

例：運用上の欠測／治療経過・患者状態に関連する未取得(ICE候補)／分類困難

運用上の欠測としては、システムエラー、記入漏れ、アンケート用紙の紛失、記録忘れ、ePROシステム上の問題などが挙げられた。また、CRCによる確認で発見されるケースも想定され、これらは患者状態とは直接関係しない欠測として整理できるとの意見があった。一方で、予算上の理由による収集中止なども運用上の欠測に含まれる可能性があると考えられた。

治療経過・患者状態に関連する未取得としては、死亡、病勢進行、治療中止、後治療開始、毒性・有害事象、患者状態の悪化による医師判断などが挙げられた。特に、患者状態の悪化による未取得は有効性や安全性とも関連する可能性があるため、結果解釈において重要との意見があった。また、事前に想定される理由については選択肢として規定し、あらかじめ分類できるようにしておくことが望ましいとの意見もあった。

一方で、患者拒否、同意撤回、転院、来院終了、医師判断などについては分類が難しいとの意見があった。例えば患者拒否は、症状悪化が背景にある場合もあれば単なる回答負担による場合もあり、理由によって解釈が異なる可能性がある。また、転院や来院終了についても、その背景に患者状態の変化があるのか個人的な事情によるものなのかが必ずしも明確ではないため、単純な分類が難しいとの指摘があった。

さらに、事前に分類できない事例については、Data Reviewの段階でClinicianとStatisticianが内容を確認し、解析上適切なカテゴリへ分類する運用も考えられるのではないかという意見が示された。全体としては、可能な限り事前定義を行いつつも、一部の未取得理由については個別の判断が必要になるとの認識が共有された。

論点2-3：

分類が難しい未取得理由、非単調欠測、ICE後に取得されたPROを、主解析・感度分析・補足的解析でどう扱うべきか？

参加者からは、論点2-2でも挙がっているように、同意撤回や患者拒否などの分類が難しい未取得理由については、撤回や拒否の背景理由が把握できない場合も多く、一律に分類することが難しいとの意見があった。そのため、可能な限り事前に対処を規定しておき、主解析では事前に定めた分類に従って解析を実施することが重要であるとの認識が共有された。また、感度分析において別の分類方法を適用し、結果への影響を確認することも有用であるとの意見があった。

データが存在しない場合や、主解析の対象外と考えられるもののデータ自体は取得されている場合の取扱いについても議論された。保守的な評価の観点から、すべてを有効性欠如として扱うworst-caseの考え方も一案として挙げられたが、いずれの場合も事前に解析上の取扱いを規定しておくことが重要であるとされた。

非単調欠測については、主解析では事前に定めたEstimandにおける中間事象の取扱い方針に従って解析を実施すべきとの意見があった。また、ICE後に取得されたPROについても、同様に事前に規定したEstimandに基づいて主解析での取扱いを決定し、必要に応じて感度分析や補足的解析でその影響を確認することが望ましいとの認識が共有された。

全体として、分類が難しい未取得理由や非単調欠測、ICE後PROについては、可能な限り事前に対処を明確化し、主解析はその方針に従って実施した上で、感度分析によって結果の頑健性を確認することが重要であるとの意見で概ね一致した。

論点2-4：

QOLスコアの結果を解釈するために、未取得理由の内訳、データ残存状況、ICE後PROの取得状況をどう提示すべきか？

参加者からは、QOLスコアの結果を適切に解釈するためには、可能な限り補足情報を併せて提示すべきとの意見があった。特に、未取得理由については単純な欠測数だけでなく、事前に定義したカテゴリごとの集計結果を併せて示すことが望ましいとの認識が共有された。

一方で、補足的解析や補足資料について、事前に網羅的に規定することには限界があるとの意見もあった。そのため、実務上は必ずしも詳細な追加解析を実施しない場合もあるとの指摘があった。

さらに、補足的な提示方法として、個別症例のQOL推移図や一覧表を用いることで、欠測やICE後のデータ取得状況をより具体的に把握できる可能性があるとの意見が示された。

ケーススタディ：主な議論点

新治療は医師選択化学療法と比べて、治療開始後の症状・QOLを改善または維持できるか？

- 論点3-1
 - Responder割合は「改善」は評価できるが、「維持」まで評価できるか？
- 論点3-2
 - PD・死亡・新規抗がん治療開始を非改善に含めることはCQと整合するか？
- 論点3-3
 - 「維持」まで評価するには、MIRASOLのMMRM・TTDで十分か？追加すべき解析・提示はあるか？
- 論点3-4
 - 死亡・PD・治療中止による未取得は、MMRM・TTDの結果解釈にどう影響するか？
- 論点3-5
 - 結果解釈のために、QOLスコア以外に何を併せて提示すべきか？

論点3-1：

Responder割合は「改善」は評価できるが、「維持」まで評価できるか？

MIRASOLで用いられているResponder割合は「15点以上の改善」を基準とした二値評価であり、「改善」を評価することはできるものの、「維持」を十分に評価できているとは言い難いのではないかと多くの意見が多く挙げられた。特に、改善はしていないが悪化もしていない患者はResponderに含まれないため、現在の定義では維持の状態が評価対象から漏れている可能性が指摘された。維持を評価するのであれば、改善だけでなく「維持」と判断するためのスコア変化の範囲を事前に定義する必要があるとの意見があった。さらに、「改善したかどうか」だけでなく、「悪化したか、悪化していないか」という観点からの評価も重要ではないかとの指摘があった。

Worst-caseの考え方を取り入れた定義にすることで維持も含めた評価が可能ではないかとの意見もあったが、いずれの場合も評価対象を明確に定義することが重要との認識が共有された。

全体としては、Responder割合のみでは「改善または維持」というClinical Questionを十分に評価できない可能性があり、維持まで確認したい場合には、Responder割合に加えて悪化の評価やTime to deteriorationなど、他のendpointの結果も併せて解釈する必要があるとの意見で概ね一致した。

論点3-2：

PD・死亡・新規抗がん治療開始を非改善に含めることはCQと整合するか？

PD、死亡、新規抗がん治療開始を非改善（non-responder）として扱う方法は、Composite strategyとして整理できるとの意見があった。一方で、「新治療は医師選択化学療法と比べて、治療開始後の症状・QOLを改善または維持できるか」というCQに対して、これらの事象を一律に非改善へ含めることが適切かについては議論があった。

死亡については、non-responderとして扱うことには一定の妥当性があるとの意見があったが、その場合には生存している患者を中心とした評価に近い解釈になる可能性が指摘された。PDについても、患者視点では生存期間や腫瘍制御に加えて症状やQOLの維持・改善が重要であることから、non-responderとして扱う考え方は理解できるとの意見があった。一方で、PDに至った理由や経緯によって意味合いが異なるため、必要に応じてエンドポイントを分けて考える余地もあるとの指摘があった。

新規抗がん治療開始については、治療の位置づけや開始理由によって解釈が異なるため、一律に非改善とすることには慎重な意見が示された。一般的にはQOLが改善している患者が新規治療へ切り替わる状況は考えにくいとの意見があった一方で、開始理由によっては必ずしも非改善を意味しない可能性もあり、切替後も追跡を継続する方法も考えられるとの意見が出された。

また、症状改善を評価するCQであれば、これらの事象を非改善に含めることは一定の合理性があるとの意見があった。しかし、non-responderには死亡、PD、新規抗がん治療開始など異なる理由が混在するため、単純な割合だけでなく、non-responderとなった理由の内訳を併せて提示することが重要であるとの認識が共有された。

論点3-3：

「維持」まで評価するには、MIRASOLのMMRM・TTDで十分か？追加すべき解析・提示はあるか？

「維持」の評価という観点では、Responder割合よりも連続量解析やTime-to-Event解析の方が捉えやすい可能性があるとの意見があった。一方で、二値評価であっても悪化の定義を適切に設定し、例えば「10点以上の低下」を悪化と定義した上で非悪化割合を評価することで、維持の状況を把握できるのではないかと意見もあった。ただし、有効性の維持を単一の指標のみで十分に評価することは難しいとの認識が共有された。

追加的な提示としては、各時点のQOL推移を示すプロットや、改善・維持・悪化等の状態を示す積み上げ棒グラフが有用ではないかと意見があった。また、QOLとOSを合わせて評価する視点も重要であり、死亡例の発生状況や死亡直前までのQOLデータを含めて検討する必要性が指摘された。さらに、「QOLを維持しながら生存している患者」を評価するようなペアワイズ比較の考え方も提案された。

一方で、TTDでは死亡が情報のある打ち切りとなる可能性があり、解釈には注意が必要との意見があった。また、MMRMについても限界や仮定に伴う懸念があるため、その点をプロジェクトチームに共有した上で結果を解釈すべきとの意見が示された。さらに、死亡後はデータ取得が不可能である一方、PD後はデータ収集が可能であることから、PD後のPROデータも含めて評価することをCQとして検討する必要があるとの認識が共有された。

現状の枠組み上、死亡などのterminal eventの影響を受けづらい早期時点でしか妥当な比較結果が得られないことが欠点であろうとの意見があった。

論点3-4：

死亡・PD・治療中止による未取得は、MMRM・TTDの結果解釈にどう影響するか？

MIRASOL試験では死亡やPD発生後に最終PRO評価日で打ち切りとして扱われているものの、死亡、PD、治療中止がPRO評価へ影響しないとは考えにくいとの意見があった。特に、これらの事象発生後はQOLが低下している可能性があり、結果としてPRO悪化の直前で打ち切られているような状況が生じることで、QOLを過大評価する方向のバイアスが生じる恐れが指摘された。また、こうした影響の大きさは、標準治療との比較かプラセボとの比較かなど、対照群や治療特性によっても異なる可能性があるとの意見があった。

さらに、MMRMについては、実際にデータが取得できた患者に基づく解析となるため、結果はデータが残存している患者集団のQOLを反映したものとして解釈する必要があるとの認識が共有された。全体として、死亡、PD、治療中止による未取得は無視できない影響を持つ可能性があり、TTDにおける打ち切りの妥当性や、MMRMが評価している対象集団を十分に考慮した上で結果を解釈することが重要であるとの意見で概ね一致した。

論点3-5：

結果解釈のために、QOLスコア以外に何を併せて提示すべきか？

未取得理由や打ち切り理由についてカテゴリ別の集計を示し、打ち切りの詳細な内訳や各時点におけるデータ残存状況を併せて提示することが重要との認識が共有された。また、Non-responderには死亡、PD、新規治療開始など異なる要因が含まれることから、単純な割合だけでなく内訳も明示すべきとの意見があった。

解析結果の補足としては、Time to deteriorationを提示することや、死亡を競合事象として扱う競合リスク解析を実施することも有用な選択肢として挙げられた。さらに、QOLの結果と併せてOSなどの臨床アウトカムを提示することで、患者状態をより包括的に評価できるのではないかとこの意見があった。

加えて、個別症例のQOL推移図や一覧表を提示することで、集団レベルの結果だけでは把握しにくい患者ごとの経過や欠測状況を理解しやすくなるとの指摘があった。一方で、補足的解析を事前に網羅的に規定することには限界があるため、実務上は結果解釈に特に重要な情報を優先して提示することが現実的ではないかとの認識が共有された。

ケーススタディ 議論時に挙げられたその他の論点

包括的な治療効果を評価したい場合には、TTD解析においてPDを必ずしも打ち切りとする必要はなく、PD後も含めた評価の可能性を検討してよいのではないかとの意見があった。また、PROの解釈においては、群間で患者背景や死亡状況の不均衡が比較的小さい早期の時点に着目して結果を評価することも一つの考え方として挙げられた。

一方で、標準治療との差別化や治療ベネフィットを示す観点からPROは有用な指標になり得るものの、これまで議論されたように、terminal eventや欠測、ICEの影響を受けるため解釈は容易ではないとの認識が共有された。また、単群試験では適切な比較対象が存在しない場合もあり、PRO結果をどのように位置付けて活用するかは必ずしも明確ではないとの意見もあった。