The 10th Data Science Roundtable Meeting - Estimand in Oncology-

PMDA
Center for Medical Health Data Science, The Institute of Statistical Mathematics
JPMA Data Science subcommittee

Facilitators:
Shunsuke Shiota, M.MSc. (PMDA)
Shogo Nomura, Ph.D. (The University of Tokyo)
Ryosuke Kato, M.E. (National Cancer Center)
Hiroshi Kuriki, M.E. (Chugai Pharmaceutical Co., Ltd)
Sho Sakui, M.S. (Takeda Pharmaceutical Company Limited)

Discussion point 1
Handling of intercurrent events using Progression-Free survival as an example

Discussion points to be addressed

Based on the specific Clinical Questions (CQs) and estimands of interest, the following issues will be considered:

1. What intercurrent events may occur? How should intercurrent events be handled?
   - What strategy will be applied to the anticipated intercurrent events?
   - If an event/censoring is defined at the time a certain intercurrent event occurs, what kind of estimand is of interest?

2. How will data collection be conducted (especially, how will data related to endpoints and intercurrent events be collected)?
   - Can the estimand discussed in discussion point 1 be evaluated with the data collection policy discussed in discussion point 2?

3. What will the primary analysis be?
   - If there is an intercurrent event leading to censoring, can the event be treated as noninformative?
   - What sensitivity analyses should be conducted to examine the assumptions made at the study design?
   - What supplementary analyses can be considered?

Case study 1:
A solid tumor trial with PFS by central review as the primary endpoint. Various trial constraints exist.

Phase III, double-blind, placebo-controlled, randomized trial evaluating the efficacy and safety of Drug X in patients with a certain solid tumor	
Anticipated CQ 1 (This CQ will be the basis of our consideration this time)	We want to evaluate the effect of Drug X in delaying disease progression (PD) or death throughout the entire course of treatment, including subsequent therapy and crossover.
Anticipated CQ 2	We want to evaluate the effect of Drug X in delaying PD or death prior to the implementation of subsequent therapy or crossover.
Primary endpoint	PFS by central review
Definition of the primary endpoint	Time from the date of randomization to the earliest of PD by central review or death

Key point to consider: If a case is judged as non-PD by central review after an investigator-assessed PD, a lack of subsequent central review data complicates the interpretation of central review PFS.

Trial constraints (open to revision based on discussion)

Common setting

- Tumor assessment is performed every 8 weeks.
- Subsequent cancer therapy can be administered even before event observation if deemed necessary by the investigator.
- Unblind the patient after investigator-assessed PD.

Handling after investigator-assessed PD (by group)

	Drug X	**Placebo**
Study drug administration	May continue if deemed to have clinical benefit	Discontinued / Crossover to Drug X is permitted
Tumor assessment	Central review is also continued if study drug administration is continued	Investigator assessment is not performed either (due to practical burden and crossover)

Case 1: Discussion points

	Discussion points
①-1	How should Investigator-assessed PD, crossover, initiation of subsequent therapy, lack of central review assessment, and other events be handled as intercurrent events? Is the proposed handling aligned with the clinical question/estimand for which BICR-assessed PFS is the primary endpoint?
①-2	To what extent can the estimand be evaluated under the current trial constraints? Which aspects of data collection can be compromised, and which cannot?
①-3	In the primary analysis, which time point should be treated as an event or censoring? If BICR data are missing, can such censoring be regarded as non-informative?
①-4	What sensitivity and supplementary analyses would be useful, including Investigator-assessed PFS, BICR-assessed PFS, analyses with alternative censoring rules, and summaries of patients with BICR-assessed non-PD after Investigator-assessed PD?

The discussion results presented in the following slides are a summary of the outcomes from the four groups participated in DSRT meeting. Please note that these contents reflect the opinions and perspectives based on the experiences of the DSRT participants, and do not represent the official positions of their affiliated organizations.

Discussion point①-1
How should Investigator-assessed PD, crossover, initiation of subsequent therapy, lack of central review assessment, and other events be handled as intercurrent events? Is the proposed handling aligned with the clinical question/estimand for which BICR-assessed PFS is the primary endpoint?

Potential intercurrent events and proposed handling:
One proposal was to handle Investigator-assessed PD, crossover, and initiation of subsequent therapy using a treatment policy strategy, while handling the lack of imaging assessment—particularly missing central review assessments due to imaging not being performed after Investigator-assessed PD, or missing two or more consecutive assessments—using a hypothetical strategy.

Other events were also raised as potential intercurrent events, including initiation of subsequent therapy, initiation of crossover treatment, treatment discontinuation, poor adherence, discontinuation due to toxicity, adverse events, unblinding, and withdrawal of consent.

Alignment with the clinical question/estimand:
Regarding alignment with the clinical question for which BICR-assessed PFS is the primary endpoint, some participants considered the proposed handling to be ideally aligned. However, because Investigator-assessed PD is used in actual clinical decision-making, another view was that greater emphasis should be placed on Investigator-assessed PD if the objective is to evaluate the effect of a treatment strategy.

Discussion point①-2 (1/2)
To what extent can the estimand be evaluated under the current trial constraints? Which aspects of data collection can be compromised, and which cannot?

To what extent can the estimand be evaluated under the current trial constraints?
Many participants commented that, if the clinical question in this case is to evaluate the overall effect of the treatment strategy, including crossover and subsequent therapy, i.e., effectiveness, follow-up should be continued until PD is confirmed, even after initiation of subsequent therapy or crossover, and data should be collected as much as possible.

In particular, when a treatment policy strategy is adopted, it would be ideal to continue collecting information, including after Investigator-assessed PD. There was a shared understanding that, if such data are not collected, it may no longer be possible to adequately evaluate the originally intended estimand.

Discussion point①-2 (2/2)
To what extent can the estimand be evaluated under the current trial constraints? Which aspects of data collection can be compromised, and which cannot?

Which aspects of data collection can or cannot be compromised?
On the other hand, participants also noted that, in practice, continued follow-up may impose substantial operational burden and budgetary constraints. Possible pragmatic approaches included reducing the frequency of data collection, extending the assessment interval after Investigator-assessed PD, or collecting data in a subset of patients.

However, it was emphasized that even supplementary analyses cannot be performed without data. Therefore, it was considered important to collect as much information as possible in both treatment groups. It was also pointed out that, if unblinding after Investigator-assessed PD or the lack of imaging assessment is handled as censoring at a certain time point, such as the last imaging assessment date, this may lead to informative censoring. If PD occurs during a period in which imaging assessments are not performed, this could introduce bias in the direction of overestimating PFS.

Furthermore, although the rationale for using BICR as the primary endpoint was understood, especially because treatment assignment may be inferred from adverse events even in a double-blind trial, some participants noted that if discordance from Investigator assessment makes evaluation difficult, the use of Investigator assessment may also warrant consideration.

Discussion point①-3
In the primary analysis, which time point should be treated as an event or censoring? If BICR data are missing, can such censoring be regarded as non-informative?

Event and censoring definitions in the primary analysis:
Some participants suggested that follow-up should continue as much as possible even after Investigator-assessed PD, crossover, or initiation of subsequent therapy, and that the time point when PD or death is ultimately confirmed should be treated as the event. For cases without imaging assessment, censoring at an appropriate assessment time point was proposed. Kaplan-Meier curves and Cox regression were also noted as commonly used methods for PFS analysis.

Non-informative censoring when BICR data are missing:
Participants shared the view that missing BICR data cannot easily be regarded as non-informative censoring. Therefore, reasons for missing data should be collected as much as possible, including the background reasons for withdrawal of consent where feasible. Crossover and initiation of subsequent therapy were discussed as intercurrent events, whereas lack of BICR assessment or skipped assessments may be more appropriately classified as missing data. Consecutive missed assessments may lead to unobserved PD and could bias PFS upward.

Participants also emphasized that the clinical question should clarify whether the objective is to evaluate effectiveness or efficacy. The decision on continued follow-up and censoring should then be aligned with that objective. Rather than censoring patients too readily, assessments should be continued as much as possible, including after subsequent therapy or crossover, and results should be interpreted in light of the reasons for missing data.

Discussion point①-4：
What sensitivity and supplementary analyses would be useful, including Investigator-assessed PFS, BICR-assessed PFS, analyses with alternative censoring rules, and summaries of patients with BICR-assessed non-PD after Investigator-assessed PD?

To assess the robustness of the primary analysis, participants suggested collecting and analyzing Investigator-assessed PFS, and conducting sensitivity analyses in which Investigator-assessed PD is treated as an event. Assessing the concordance between Investigator-assessed PD and BICR-assessed PD, and summarizing patients with BICR-assessed non-PD after Investigator-assessed PD, may help clarify discrepancies between assessments and their impact on the results.

Participants also emphasized the importance of collecting detailed reasons for censoring and presenting their breakdown as supplementary information. Such information would help evaluate the plausibility of the assumptions underlying the primary analysis.

The extent to which sensitivity analyses should be prespecified in the SAP was also discussed. In practice, a predefined set of sensitivity analyses is often planned using templates. While methods such as modeling the censoring mechanism and applying IPCW could be considered to address informative censoring, these approaches may increase analytical complexity and can be difficult to prespecify.

Case study 2*: A hematological cancer trial with investigator-assessed EFS as the primary endpoint. Consider the issues surrounding the handling of "treatment failure."

*Case study 2 was reconstructed as a hypothetical case based on the review report for ivosidenib.

Phase III, double-blind, placebo-controlled, randomized trial evaluating the efficacy and safety of Drug X in patients with a certain hematological cancer	
Anticipated CQ 1	Evaluate the effect of delaying relapse after remission or death
Anticipated CQ 2 (This CQ will be the basis of our consideration this time)	In addition to CQ 1 above, failing to achieve remission itself (“treatment failure”) is also evaluated as a lack of treatment effect
Primary endpoint	Investigator-assessed EFS
Definition of the primary endpoint	Time from the date of randomization to the earliest of: a. Relapse after remission b. Death c. Treatment failure (described later)

- Assess the presence or absence of remission and relapse every 8 weeks according to pre-specified evaluation criteria.
- Bone marrow examination, peripheral blood examination, etc., are used for the evaluation.
- EFS events are defined as relapse after remission, death, and treatment failure.
  - **If remission is never achieved by week 24, it is defined as treatment failure. In this study, treatment failure was treated as an event on the first day of administration, i.e., Day 1.**

Case 2: Discussion points

	Discussion points
②-1	Is including treatment failure as an EFS event aligned with the clinical question? If treatment failure is treated as a Day 1 event, what treatment effect is being estimated? Should treatment failure be handled using a composite strategy, or should it be addressed using another strategy or as a separate endpoint?
②-2	At which time point should failure to achieve remission by Week 24 be treated as an event? How might the clinical interpretation differ if the event is placed at Day 1, Week 24, or another time point?
②-3	How should the breakdown of remission status, relapse, death, and treatment failure be collected and presented?
②-4	How should treatment failure be handled in the primary analysis?
②-5	Are sensitivity analyses useful in which treatment failure is handled as a Day 1 event, a Week 24 event, censoring, or in other ways?
②-6	If treatment failure accounts for the majority of EFS events, is it appropriate to use the hazard ratio as the primary summary measure?

The discussion results presented in the following slides are a summary of the outcomes from the four groups participated in DSRT meeting. Please note that these contents reflect the opinions and perspectives based on the experiences of the DSRT participants, and do not represent the official positions of their affiliated organizations.

Discussion point②-1 (1/2)

Is including treatment failure as an EFS event aligned with the clinical question? If treatment failure is treated as a Day 1 event, what treatment effect is being estimated? Should treatment failure be handled using a composite strategy, or should it be addressed using another strategy or as a separate endpoint?

Alignment between including treatment failure as an EFS event and the clinical question

Participants understood the rationale for including treatment failure as an EFS event. However, some expressed concern about its alignment with the clinical question when EFS is evaluated as a time-to-event endpoint. In particular, because EFS is an endpoint intended to evaluate time, treating treatment failure up to Week 24 as an event occurring on Day 1 was considered difficult to interpret. It was also pointed out that this approach may underestimate EFS. Some participants suggested that it may be more natural to treat treatment failure as an event at Week 24 rather than on Day 1. Others proposed that supplementary sensitivity analyses using different event times could be useful.

There was a shared understanding that the meaning of the treatment effect being estimated can change substantially depending on the time point at which treatment failure is regarded as an event. Therefore, the definition should be established after clarifying its alignment with the clinical question.

Discussion point②-1 (2/2)

Is including treatment failure as an EFS event aligned with the clinical question? If treatment failure is treated as a Day 1 event, what treatment effect is being estimated? Should treatment failure be handled using a composite strategy, or should it be addressed using another strategy or as a separate endpoint?

What treatment effect is being estimated if treatment failure is treated as a Day 1 event?:
Some participants noted that, if failure to achieve remission is regarded as a clinically very serious failure, there may be some rationale for treating it as a Day 1 event. It was also pointed out that this case appears to combine a binary assessment, such as remission rate, with a time-to-event assessment, such as time to relapse. How much weight should be assigned to treatment failure was therefore raised as an important discussion point.

Should treatment failure be handled using a composite strategy, another strategy, or a separate endpoint?:
One view was that, if treatment failure is incorporated into EFS, it could be considered as part of a composite strategy. On the other hand, some participants suggested that it may be easier to interpret treatment failure if it is evaluated as a separate endpoint, such as remission rate, rather than being incorporated into EFS.

Discussion point②-2
At which time point should failure to achieve remission by Week 24 be treated as an event? How might the clinical interpretation differ if the event is placed at Day 1, Week 24, or another time point?

Event timing for failure to achieve remission by Week 24:
Participants noted that the interpretation would differ markedly between treating failure to achieve remission by Week 24 as a Day 1 event and treating it as a Week 24 event. Week 24 was considered a natural event time. However, some participants felt that representing “ not achieving remission” as an event at a specific time point is itself somewhat conceptually difficult. From the patient perspective, achieving remission and prolonging remission duration are important. Therefore, if failure to achieve remission is regarded as the most serious form of treatment failure, treating it as a Day 1 event may also be clinically interpretable.

Clinical interpretation by event timing:
Because EFS is fundamentally a time-to-event endpoint, treating treatment failure as a Day 1 event may mix the evaluation of time with the evaluation of treatment failure, making the interpretation challenging. In addition to choosing between Day 1 and Week 24, supplementary or sensitivity analyses using alternative event times may be useful to assess the robustness of the results.

Discussion point②-3
How should the breakdown of remission status, relapse, death, and treatment failure be collected and presented?

Some participants suggested that it may be easier to interpret the results if remission rate, relapse, and death are collected and presented separately, rather than being evaluated together within a single EFS endpoint.

It was also suggested that time to remission could be presented as a supplementary assessment or sensitivity analysis. By doing so, it may be possible to evaluate the treatment effect from multiple perspectives, including aspects that may be difficult to capture simply by incorporating treatment failure as an EFS event.

Discussion point②-4
How should treatment failure be handled in the primary analysis?

Participants noted that including treatment failure as an EFS event in the primary analysis may have limited impact if its incidence is low. However, if treatment failure accounts for a large proportion of events, incorporating it as an event requires caution. Concentration of events on Day 1 and differences in treatment effects across EFS event components may raise concerns about the proportional hazards assumption and the interpretability of the hazard ratio. Therefore, some participants suggested that treatment failure could be evaluated separately as a secondary endpoint, rather than being included in the primary EFS analysis. Another possible option would be to consider it as a co-primary endpoint.

If treatment failure is included in the primary analysis, its expected incidence should be carefully assessed at the trial planning stage, and its impact should be evaluated through sample size planning and simulations. An example was discussed in which treating treatment failure as a Day 1 event led to many early events in both groups and a median EFS of 1 day, raising the question of whether such a situation could have been anticipated prospectively.

Discussion point②-5
Are sensitivity analyses useful in which treatment failure is handled as a Day 1 event, a Week 24 event, censoring, or in other ways?

Sensitivity analyses using alternative handling of treatment failure:
Because the interpretation of the results may differ depending on whether treatment failure is handled as a Day 1 event, a Week 24 event, censoring, or in another way, participants considered various sensitivity analyses useful for assessing the robustness of EFS. Participants also shared the view that EFS alone may not fully describe the overall treatment effect. Therefore, supplementary endpoints such as remission rate, time to remission, time to relapse, and OS should be presented and interpreted together.

Supplementary presentation beyond EFS:
For EFS, in addition to time-to-event analyses, supplementary assessments at specific time points, such as Week 24, or analyses using binary endpoints were also proposed. Some participants noted that supplementary analyses are supportive in nature and should not be used to draw conclusions as strong as those from the primary analysis. However, such analyses may suggest treatment effects that are not fully captured by the primary analysis. They should therefore be used appropriately according to the trial objective. The endpoints to be emphasized may also differ between trials intended to support regulatory approval and post-approval trials aimed at exploring optimal treatment strategies.

Discussion point②-6 If treatment failure accounts for the majority of EFS events, is it appropriate to use the hazard ratio as the primary summary measure?
Participants noted that, when treatment failure accounts for most EFS events, the appropriateness of using the hazard ratio as the primary summary measure should be carefully considered. If treatment failure occurs frequently, the proportional hazards assumption may not hold. Therefore, Kaplan-Meier curves should first be examined, and the proportional hazards assumption should be assessed. In contrast, if the incidence of treatment failure is low, the hazard ratio may still be reasonably appropriate as the primary summary measure. Some participants suggested that RMST may be preferable as an alternative measure, since the hazard ratio may not fully reflect the relevant information. However, because the hazard ratio has been widely used in many trials, it may not be easy in practice to completely move away from it. Participants also emphasized the importance of clarifying the clinical question, including which treatment effect is of interest. Depending on the objective, alternative estimands or analytical approaches may be needed, such as a principal stratum strategy focused on patients who achieve remission.

Discussion point 2
Handling of PROs considering terminal events

PRO/QOL Analyses with Terminal Events and Missing Data

Discussion background for interpreting patient-reported outcomes in oncology trials

Why this discussion is needed

PROs capture symptoms, functioning, and quality of life directly from the patient perspective.

In oncology trials, progression, treatment discontinuation, subsequent therapy, and death can affect both PRO availability and interpretation.

PRO after death is not merely missing — it is undefined. Its handling should be driven by the clinical question and estimand.

PROs/QOL are often specified as exploratory endpoints, and estimands have not been sufficiently discussed in Japan. Therefore, a clearer framework for PRO/QOL estimands is needed.

Typical data flow

What the group discussions addressed

1. Terminal events

Handling death across change scores, responder endpoints, and time-to-deterioration endpoints.

2. Missing data and post-ICE PRO

Classifying reasons for unavailable PRO data and considering observed PRO after intercurrent events.

3. MIRASOL case study

Evaluating whether responder analysis, MMRM, and TTD address both improvement and maintenance of symptoms/QOL.

Discussion framework

Clinical question → PRO endpoint → ICE strategy → Analysis → Interpretation

The aim was to discuss alignment, assumptions, limitations, and transparent reporting rather than to recommend a single approach for all trials.

Case 1: Discussion points

How should PROs after treatment initiation be compared and interpreted, considering terminal events (e.g., death)?

	Discussion points
1-1	When PRO after death does not exist, what should be evaluated?
1-2	How might the strategy for terminal events differ depending on the PRO endpoint? For each clinical question listed in Discussion point 1-1, discuss what strategy would be applied when the data type is specified, such as continuous, binary, or time-to-event data.
1-3	Can the selected strategy be implemented using the planned analysis method?
1-4	What information, in addition to QOL scores, should be presented to support interpretation of the results?

The discussion results presented in the following slides are a summary of the outcomes from the four groups participated in DSRT meeting. Please note that these contents reflect the opinions and perspectives based on the experiences of the DSRT participants, and do not represent the official positions of their affiliated organizations.

Discussion point 1-1
When PRO after death does not exist, what should be evaluated?

Participants agreed that the handling of PRO after death is not merely an analytical question of what value should be imputed for patients who have died. Rather, the key question is what patient benefit should be evaluated, including death.

Possible clinical questions for handling terminal events in PRO evaluation included:
- How does PRO/QOL change, or improve, among patients who are alive?
- How does PRO/QOL change, or improve, as part of an overall patient state that includes death?
- How would PRO/QOL have changed, or improved, under a hypothetical scenario in which patients had not died?
- How does PRO/QOL change, or improve, among patients who would be able to survive for a certain period?

Discussion point 1-2

How might the strategy for terminal events differ depending on the PRO endpoint?
For each clinical question listed in Discussion point 1-1, discuss what strategy would be applied when the data type is specified, such as continuous, binary, or time-to-event data.

Continuous endpoint, such as change from baseline

When PRO is analyzed as a continuous outcome, the key issue is how to quantify PRO after death, since no score exists after death. With methods such as MMRM, PRO values that are undefined after death are effectively handled through the missing-data assumptions of the model, such as MAR. This may lead to an interpretation close to a hypothetical strategy. This may be interpreted as being close to a hypothetical strategy, evaluating what QOL would have been if the patient had not died. If death is assigned a score of 0 or the worst possible score, this may be considered a composite strategy. However, the clinical validity of the assigned score and the interpretation of the results require caution.

Binary endpoint, such as responder status

For responder analyses, participants generally considered it acceptable to treat patients who had not achieved a response before death as non-responders, corresponding to a composite strategy. This does not necessarily equate death with “non-improvement” in the PRO score itself, but rather interprets death as an indicator of worsening patient status. Because non-responders may include patients with death, PD, initiation of subsequent therapy, operational missingness, or simple lack of improvement, the reasons for non-response should be presented as a breakdown.

Time-to-event endpoint, such as time to worsening

For time to worsening, the handling of death directly defines the estimand. If death is treated as an event, the endpoint becomes time to QOL worsening or death, corresponding to a composite strategy. If death is treated as censoring, the analysis is closer to a while-alive strategy, focusing on observed QOL worsening among patients who are alive. However, whether death can be regarded as non-informative censoring is a key concern. If death is treated as a competing risk, QOL worsening and death can be evaluated separately, improving transparency. In this case, cumulative incidence functions and event breakdowns should be presented. However, interpretation may be difficult if the proportion of deaths differs between treatment groups.

Discussion point 1-3
Can the selected strategy be implemented using the planned analysis method?

Continuous outcomes (e.g., MMRM)

For continuous outcomes, possible approaches included using observed values up to immediately before death or assigning a score of 0 after death. However, participants noted that careful consideration is needed regarding whether the missing-data assumptions of models such as MMRM are consistent with how post-death data are handled.

Responder proportions

For responder analyses, the selected strategy could be implemented relatively clearly by using the full analysis population as the denominator. By prespecifying whether death is treated as non-response, the intended estimand, such as a composite strategy or while-on-treatment strategy, can be reflected relatively directly in the analysis.

Time to worsening

For time to worsening, treating death as censoring was considered consistent with a while-on-treatment strategy and implementable using the planned analysis method. However, appropriate interpretation requires examining the breakdown and occurrence of death and other censoring events. If the clinical question is clearly to evaluate PRO among patients who are alive, not treating death as an event may be a natural choice.

Discussion point 1-4
What information, in addition to QOL scores, should be presented to support interpretation of the results?

The following supplementary information was proposed across the group discussions:

Survival-related information
- OS
- Proportion of deaths at each time point

Treatment status
- Subsequent therapy status at each time point
- Treatment completion rate at each time point
- Balance in total dose and other treatment-related factors between treatment groups

Terminal events
- Number of terminal events, such as deaths
- Breakdown of causes of death

Missing data and censoring
- Reasons for missing data
- Reasons for censoring, including withdrawal of consent
- Proportion censored

PRO data availability
- Compliance rate at each time point

Patient characteristics
- Review of patient background characteristics

QOL trajectory and visualization
- QOL change up to each time point among patients who are alive and those who have died
- Individual QOL trajectory plots
- Listings including reasons for missing data

Responder analysis
- Breakdown of reasons for non-response

Case 2: Discussion points

While the main focus is on PROs during treatment, treatment discontinuation can also impact symptoms and QOL; therefore, we consider scenarios in which post-discontinuation PROs are collected to a certain extent.

	Discussion points
2-1	To what extent should reasons for missing PRO data be collected in the CRF, and what are the barriers to collection?
2-2	How should reasons for missing PRO data be classified at a level of granularity that can be used for analysis and interpretation?
2-3	How should difficult-to-classify reasons for missing PRO data, non-monotone missingness, and PRO data collected after intercurrent events be handled in the primary analysis, sensitivity analyses, and supplementary analyses?
2-4	How should the breakdown of reasons for missing PRO data, data availability over time, and the availability of PRO data after intercurrent events be presented to support interpretation of QOL score results?

The discussion results presented in the following slides are a summary of the outcomes from the four groups participated in DSRT meeting. Please note that these contents reflect the opinions and perspectives based on the experiences of the DSRT participants, and do not represent the official positions of their affiliated organizations.

Discussion point 2-1
To what extent should reasons for missing PRO data be collected in the CRF, and what are the barriers to collection?

Reasons for missing PRO data are important for interpreting results and evaluating missing data, and should therefore be collected to the extent feasible. PRO data may be missing due to factors related to the treatment course, such as adverse events, laboratory abnormalities, or deterioration in patient condition. Collecting the reasons for missing PRO data may help assess whether disease progression or adverse events are associated with QOL deterioration or missing PRO assessments.

However, detailed collection may increase the burden on patients and clinical sites. Potential reasons for missing PRO data include physician judgment due to poor patient condition, adverse events, inability to visit the site, lost questionnaires, forgetting to complete records, patient refusal, and ePRO system issues. Collecting all of these reasons in detail may not be realistic. Increasing the number of collection items may also increase patient burden, reduce response quality, and lead to further missing data. Adjusting the collection frequency or assessment interval may therefore be needed to reduce burden. Overall, data collection items should be defined by balancing the level of detail needed for analysis and interpretation with patient burden.

At a minimum, it would be useful to distinguish missing PRO data due to death from missing PRO data due to other reasons. A practical approach would be to prespecify major reasons as selectable options and collect other reasons under “Other.” Free-text entries under “Other” could potentially be classified and organized in the future using algorithms or similar methods.

Discussion point 2-2
How should reasons for missing PRO data be classified at a level of granularity that can be used for analysis and interpretation?
Examples include operational missingness, missing PRO data related to treatment course or patient condition, which may be potential intercurrent events, and difficult-to-classify reasons.

Operational missingness:
Operational missingness may include system errors, incomplete entries, lost questionnaires, forgetting to complete records, and ePRO system issues. Cases identified through CRC review could also be classified as missingness not directly related to patient condition. Discontinuation of PRO collection for budgetary reasons may also be included in this category.

Missing PRO data related to treatment course or patient condition:
This category may include death, disease progression, treatment discontinuation, initiation of subsequent therapy, toxicity or adverse events, and physician judgment due to deterioration in patient condition. Missing PRO data due to deterioration in patient condition is particularly important for interpretation, as it may be related to both efficacy and safety. Reasons anticipated in advance should preferably be prespecified as selectable options to enable prospective classification.

Difficult-to-classify reasons:
Patient refusal, withdrawal of consent, transfer to another hospital, end of site visits, and physician judgment were considered difficult to classify. For example, patient refusal may be due to worsening symptoms or simply to the burden of completing the questionnaire. Similarly, transfer or end of site visits may reflect either changes in patient condition or personal circumstances. For reasons that cannot be classified prospectively, clinicians and statisticians could review the details during data review and assign analytically appropriate categories.

Overall, reasons for missing PRO data should be prespecified as much as possible, while some cases may require individual judgment.

Discussion point 2-3
How should difficult-to-classify reasons for missing PRO data, non-monotone missingness, and PRO data collected after intercurrent events be handled in the primary analysis, sensitivity analyses, and supplementary analyses?

Participants noted that reasons such as withdrawal of consent or patient refusal are difficult to classify uniformly, because the underlying reasons are often unknown. Therefore, the handling of such cases should be prespecified as much as possible, and the primary analysis should follow the prespecified classification. Sensitivity analyses using alternative classification rules may be useful to assess the impact on the results.

Participants also discussed how to handle cases with no available data, as well as cases in which data were collected but are considered outside the scope of the primary analysis. From a conservative perspective, treating such cases as lack of efficacy under a worst-case scenario was raised as one possible approach. However, the analytical handling should be prespecified in all cases.

For non-monotone missingness, the primary analysis should follow the prespecified strategy for handling intercurrent events in the estimand. Similarly, PRO data collected after intercurrent events should be handled in the primary analysis according to the prespecified estimand. If needed, their impact should be explored through sensitivity or supplementary analyses. Overall, the handling of difficult-to-classify reasons, non-monotone missingness, and post-ICE PRO data should be clarified prospectively as much as possible. The primary analysis should follow that approach, and robustness should be assessed through sensitivity analyses.

Discussion point 2-4
How should the breakdown of reasons for missing PRO data, data availability over time, and the availability of PRO data after intercurrent events be presented to support interpretation of QOL score results?

To appropriately interpret QOL score results, participants suggested presenting supplementary information as much as possible. For reasons for missing PRO data, it would be useful to present not only the number of missing assessments, but also summaries by prespecified categories.

However, participants also noted that it may not be feasible to comprehensively prespecify all supplementary analyses and materials in advance. In practice, detailed additional analyses may not always be performed. Individual patient-level QOL trajectory plots and listings may help provide a more concrete understanding of missing data patterns and PRO data collected after intercurrent events.

ケーススタディ：主な議論点

Does the new treatment improve or maintain symptoms and QOL after the start of treatment compared to investigator's choice of chemotherapy?

- 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？

Case study3: Discussion points

Does the new treatment improve or maintain symptoms and QOL after the start of treatment compared to investigator's choice of chemotherapy?

	Discussion points
3-1	Can the responder proportion evaluate “improvement,” but not necessarily “maintenance”?
3-2	Is it aligned with the clinical question to include PD, death, and initiation of new anticancer therapy as non-response?
3-3	Are the MMRM and TTD analyses used in MIRASOL sufficient to evaluate “maintenance”? Are additional analyses or presentations needed?
3-4	How do missing PRO data due to death, PD, or treatment discontinuation affect the interpretation of MMRM and TTD results?
3-5	What information, in addition to QOL scores, should be presented to support interpretation of the results?

The discussion results presented in the following slides are a summary of the outcomes from the four groups participated in DSRT meeting. Please note that these contents reflect the opinions and perspectives based on the experiences of the DSRT participants, and do not represent the official positions of their affiliated organizations.

Discussion point 3-1
Can the responder proportion evaluate “improvement,” but not necessarily “maintenance”?

The responder proportion used in MIRASOL is a binary endpoint based on an improvement of at least 15 points. While it can evaluate “improvement,” many participants noted that it may not sufficiently evaluate “maintenance.” Patients who do not improve but also do not deteriorate are not counted as responders. Therefore, the current definition may miss patients whose condition is maintained. To evaluate maintenance, the range of score changes considered to represent “maintenance” should be prespecified, in addition to the threshold for improvement.

Participants also suggested that it may be important to evaluate not only whether patients improved, but also whether they deteriorated or did not deteriorate. Although a definition incorporating a worst-case approach may allow maintenance to be assessed, the target of evaluation should be clearly defined in any case.

Overall, the responder proportion alone may not fully address the clinical question of “improvement or maintenance.” If maintenance is also of interest, other endpoints, such as deterioration-based analyses or time to deterioration, should be interpreted together.

Discussion point 3-2
Is it aligned with the clinical question to include PD, death, and initiation of new anticancer therapy as non-response?

PD, death, and initiation of new anticancer therapy can be organized as non-response under a composite strategy. However, participants discussed whether uniformly treating these events as non-response is aligned with the clinical question of whether the new treatment improves or maintains symptoms/QOL compared with investigator's choice of chemotherapy.

For death, treating it as non-response was considered to have some validity, although the interpretation may become closer to an evaluation focused on patients who are alive.

For PD, participants understood the rationale for treating it as non-response, since maintaining or improving symptoms/QOL is important from the patient perspective, in addition to survival and tumor control. However, because the meaning of PD may differ depending on the reason or clinical course leading to PD, endpoints may need to be considered separately where appropriate.

For initiation of new anticancer therapy, participants were cautious about uniformly treating it as non-response, as the interpretation may differ depending on the role of the therapy and the reason for initiation. Although patients with improving QOL may generally be unlikely to switch to a new therapy, initiation of new anticancer therapy may not always indicate non-response. Continued follow-up after treatment switching could also be considered.

If the clinical question is focused on symptom improvement, including these events as non-response may be reasonably justified. However, because non-responders may include heterogeneous reasons such as death, PD, and initiation of new anticancer therapy, the breakdown of reasons for non-response should be presented together with the overall proportion.

Discussion point 3-3
Are the MMRM and TTD analyses used in MIRASOL sufficient to evaluate “maintenance”? Are additional analyses or presentations needed?

From the perspective of evaluating maintenance, continuous analyses and time-to-event analyses may capture this concept more readily than responder proportions. Even with a binary endpoint, maintenance could potentially be assessed by appropriately defining deterioration. For example, if a decrease of at least 10 points is defined as deterioration, the proportion of patients without deterioration could be evaluated. However, maintenance of efficacy is difficult to fully assess using a single measure alone.

Additional presentations such as QOL trajectory plots at each time point and stacked bar charts showing improvement, maintenance, deterioration, and other states may be useful. It is also important to evaluate QOL together with OS, including the occurrence of deaths and QOL data up to shortly before death. A pairwise comparison approach evaluating patients who remain alive while maintaining QOL was also proposed.

For TTD, death may represent informative censoring, and interpretation therefore requires caution. For MMRM, limitations and assumptions should be shared with the project team to support appropriate interpretation. Because PRO data cannot be collected after death but can be collected after PD, the clinical question should consider whether PRO data after PD should be included in the evaluation.

A potential limitation of the current framework is that valid comparisons may only be feasible at early time points that are less affected by terminal events such as death. Some groups suggested considering approaches such as DOOR, or Desirability of Outcome Ranking, and generalized pairwise comparison.

Discussion point 3-4
How do missing PRO data due to death, PD, or treatment discontinuation affect the interpretation of MMRM and TTD results?

In MIRASOL, patients were censored at the last PRO assessment date after death or PD. However, participants noted that death, PD, and treatment discontinuation are unlikely to have no impact on PRO evaluation. QOL may decline after these events. If patients are censored just before PRO deterioration would have been observed, QOL may be biased upward. The magnitude of this impact may also differ depending on the control group and treatment characteristics, such as whether the comparison is against standard therapy or placebo. For MMRM, the analysis is based on patients with available data. Therefore, the results should be interpreted as reflecting the QOL of the patient population in whom data remained available.

Overall, missing PRO data due to death, PD, or treatment discontinuation may have a non-negligible impact. Interpretation should therefore consider both the appropriateness of censoring in TTD and the patient population evaluated by MMRM.

Discussion point 3-5
What information, in addition to QOL scores, should be presented to support interpretation of the results?

Reasons for missing data and censoring should be summarized by category. Detailed breakdowns of censoring and data availability over time should also be presented. Because non-responders may include patients with heterogeneous reasons, such as death, PD, and initiation of new treatment, the breakdown of reasons for non-response should be shown in addition to the overall non-responder proportion.

As supplementary analyses, time to deterioration and competing risk analysis treating death as a competing event were considered useful options. Presenting clinical outcomes such as OS together with QOL results may also support a more comprehensive evaluation of the patient's condition. Individual patient-level QOL trajectory plots and listings may help clarify clinical courses and missing data patterns that are difficult to understand from aggregate-level results alone.

Because it is difficult to comprehensively prespecify all supplementary analyses, it may be practical to prioritize information that is particularly important for interpreting the results.

Additional discussion points beyond those prespecified

If the objective is to evaluate the overall treatment effect, PD does not necessarily need to be treated as censoring in TTD analyses. An evaluation including PRO data after PD may also be considered. For interpreting PRO results, focusing on earlier time points was suggested as one possible approach, because imbalances in patient background characteristics and death status between treatment groups may be relatively limited at those time points.

PRO may be a useful measure for differentiating a new treatment from standard therapy and demonstrating treatment benefit. However, as discussed, interpretation is challenging because PRO results can be affected by terminal events, missing data, and intercurrent events. In single-arm trials, an appropriate comparator may not be available, and it may not always be clear how PRO results should be positioned and used.