FRIENDS of CANCER RESEARCH

Enhancing Study Designs and Interpretation of Interim Overall Survival Data in Oncology Trials

Friends of Cancer Research White Paper | 2024

Executive Summary

In oncology drug development, early endpoints such as progression-free survival (PFS) and objective response rate (ORR) are commonly used to support expedited development of therapies by facilitating earlier efficacy readouts and regulatory review. This can help provide timely access to potentially life-saving treatments. Challenges arise when there are limited overall survival (OS) data available at the time of this early assessment, leaving uncertainty about the true benefit-risk profile of a drug. In these settings, interim OS data may be evaluated as a safety endpoint to assess potential harm. However, the interpretation of interim OS data can be challenging due to small event numbers, limited duration of follow up, and trial dynamics such as patient crossover.

A collaborative, multidisciplinary working group outlined key considerations for improving the analysis and interpretation of interim OS data in oncology clinical trials. These include a proposal for a multi-step approach to be incorporated into trial designs to guide both qualitative and quantitative evaluations, ensuring a more complete understanding of the data. Taken together, an improved design and more structured interpretation of interim OS data will lead to better informed decision-making during drug development.

Key Insights and Recommendations

- Early assessments of OS can provide unreliable results due to small event counts, limited follow-up, and overall immature data. We examined case studies, which highlight how patient subgroups, treatment crossover, and other trial design factors complicate interim OS data interpretation.
- A carefully considered study design can enhance the reliability of interim OS data interpretations. This includes pre-specifying criteria for patient crossover, planning for sufficient follow-up duration, and simulating potential scenarios to inform analysis timing and threshold setting. Design elements, paired with a structured analysis approach, can ensure more accurate and timely decision-making during drug development.
- A structured, multi-step approach to interpreting interim OS data is proposed:
    - Perform a qualitative descriptive analysis, including a review of event counts, patient comorbidities, the timing of adverse events, rates of dose interruptions or reductions, and subsequent therapies. This provides essential clinical context for early signals of harm or efficacy.
    - Apply a streamlined/comprehensive quantitative framework that balances the risk of mistakenly concluding that a treatment is harmful (false positive) or missing a true safety issue (false negative) when interpreting interim OS data. This includes calculating hazard ratios (HRs) and their confidence intervals, setting thresholds for identifying potential harm, and using predictive models taking into account the data maturity to assess whether the final OS outcome is likely to show benefit, harm, or no difference.

These insights emphasize the importance of integrating careful design and comprehensive analysis of interim OS data to help ensure that oncology trials can better balance early efficacy signals with expected long-term survival outcomes.

Authors

Arunava Chakravartty, Novartis

Andy Chi, Takeda

Tai-Tsang X. Chen, GSK

Ruthanna Davi, Medidata Solutions

Maura Dickler, Genentech

Alex Fleishman, Amgen

Sebastiano Gattoni-Celli, Daiichi Sankyo, Inc.

Thomas Gwise, Medidata Solutions

Lisa Hampson, Novartis

Philip He, Daiichi Sankyo, Inc.

Laura Huggins, EMD Serono

Qi Jiang, Pfizer, Inc.

Michael Leblanc, Fred Hutchinson Cancer Center

Jianchang Lin, Takeda

Ray Lin, Genentech

Andrew Lithio, Eli Lilly and Company

Qiang Liu, Amgen

Briggs Morrison, Crossbow Therapeutics

Pabak Mukhopadhyay, AstraZeneca

Abderrahim Oukessou, Bristol Myers Squibb

Antony Sabin, GSK

Sabeen Shah, Johnson and Johnson Innovative Medicine

Mark Stewart, Friends of Cancer Research

Michael Sweeting, AstraZeneca

Yevgen Tymofiyev, Johnson and Johnson Innovative Medicine

Chris White, Patient Advocate

Nevine Zariffa, NMD Group

Helen Zhou, GSK

Table of Contents

Background

In oncology drug development, early endpoints such as progression-free survival (PFS) and objective response rate (ORR) are commonly used to provide early indications of a drug's efficacy. These endpoints help to expedite drug development, addressing unmet medical needs by enabling timely regulatory approvals and patient access to potentially beneficial therapies through Accelerated Approval, or traditional approval in certain circumstances. While overall survival (OS), the gold standard for assessing clinical benefit of cancer therapies, is traditionally evaluated as an efficacy endpoint at the end of a trial interim looks at OS data at the time of ORR or PFS assessment can serve as a safety endpoint, providing additional context for assessing the benefit-risk profile of new cancer drugs. However, interim OS data can be challenging to interpret due to the immaturity of the data at the time of an early endpoint readout.

In some cases, a statistically significant effect observed in PFS or ORR efficacy results may be overshadowed by a potential risk of harm based on interim OS data (e.g., an observed hazard ratio above 1.0). This scenario poses a conundrum due to potential conflicting data on the benefit-risk assessment of a drug. At early looks, like any endpoint, statistical estimates of endpoint readouts can be highly variable due to small sample size, limited follow up, low information fraction, and potentially delayed treatment effects. As the data mature, these fluctuations can stabilize to provide a clearer picture of the true treatment effect.[1] If interim OS analyses lead to the erroneous conclusion that a drug is harming patients, its approval may be unduly delayed, depriving patients of potential benefits based on a false conclusion of harm. However, if interim OS data correctly identify a potential safety issue early, a potentially harmful drug is kept off the market, thus protecting patients from adverse outcomes. Differentiating between these two scenarios requires careful planning and robust data, raising important questions about how best to minimize the risk of drawing false conclusions from interim OS data. To navigate these challenges, a robust, standardized, and context-specific framework can help guide the analysis and interpretation of interim OS data, ensuring reliable evaluation and good regulatory decision-making.

Friends of Cancer Research established a multidisciplinary working group to address these challenges and to develop best practices for assessing interim OS data in oncology trials. This white paper outlines key design and analysis considerations when interim OS data are evaluated in a trial and proposes a strategy for simulation studies that could provide data driven insights, with a goal of improving the understanding and application of interim OS data.

Learnings from Recent Clinical Trials on Interim OS Data

Recent clinical trials provide insights into the challenge of interpreting interim OS data, particularly in relation to early endpoints like PFS and ORR. These trials demonstrate how factors such as treatment crossover, the immaturity of interim OS data, and patient subgroups can affect the interpretation of OS results. By examining these factors, we can gain insights into how to optimize the design and interpretation of endpoints in future trials, particularly in terms of balancing early efficacy signals with long-term survival outcomes.

The following case studies provide further context, exploring how these insights apply to individual trials and offering lessons for future study designs (**Appendix 1** summarizes study details and outcome data at interim analyses, when available):

1. PSMAFore Trial (177Lu-PSMA-617 in Metastatic Castration-Resistant Prostate Cancer)

The PSMAFore trial explored the use of radioligand therapy in 468 patients with metastatic castration-resistant prostate cancer (mCRPC).[2] Patients who progressed on standard therapies in the control arm were allowed to crossover to the experimental arm. The radiographic PFS (rPFS) results were highly favorable at the primary analysis (cutoff: ~7 months median time from randomization until cutoff; Hazard Ratio; HR = 0.41, 95% Confidence Interval; CI: 0.29–0.56), showing a strong treatment effect. The interpretation of interim OS analyses was complicated by high rates of patient crossover, making it difficult to accurately assess the long-term survival benefits of the therapy. By the time of the second interim OS analysis, over half of all patients randomized to the control arm (123 of 234 patients) had crossed over to the experimental arm, and the unadjusted OS HR at this analysis (targeting a treatment policy estimand, not adjusting for crossover) was 1.16 (95% CI: 0.83-1.64). At the time of the third interim OS analysis (cutoff: ~24 months median time from randomization until cutoff), 134 of 234 patients randomized to control had crossed over. This analysis showed an unadjusted OS HR of 0.98 (95% CI: 0.75–1.28). The final OS analysis is pending.

Key Insight: High crossover rates complicate OS interpretation and it is often necessary to evaluate the OS data under a variety of sensitivity and supplementary analyses to investigate the robustness of the results. These can include statistical estimation approaches, such as those based on the rank preserving structural failure time (RPSFT) model used in this study, though interpretation of these analyses can still be challenging. This trial highlights a common issue across oncology studies, where patient crossover allowed under the protocol design can make it harder to see the true effect of the new treatment relative to the standard treatment than if patients had not been permitted to crossover.

2. monarchE Trial (Abemaciclib + ET in Early HR+/HER2- Breast Cancer)

The monarchE trial examined adjuvant abemaciclib, a CDK4/6 inhibitor, in combination with endocrine therapy (ET) for patients with high-risk of recurrence early-stage HR+/HER2- breast cancer.[3] The trial included a large population of 5,637 patients. Though the trial previously demonstrated a significant benefit in invasive disease-free survival (IDFS), the immaturity of the OS data was still evident at the first interim analysis of OS (cutoff: 36 months from study start). While the IDFS readout was statistically significant (HR = 0.696, 95% CI: 0.588–0.823), the initial interim OS analysis showed an OS HR of 1.091 (95% CI: 0.818–1.455), favoring the control arm. The initial FDA approval was limited to patients at high risk of recurrence and high Ki-67 expression.[4] This was based on careful consideration of prespecified subgroups and additional analyses including a gated hierarchical testing strategy that included the additional endpoint of IDFS in patients with a KI-67 score ≥20% which also demonstrated a statistically significant IDFS (HR=0.626, 95% CI: 0.488, 0.803) and an interim OS analysis showed an OS HR of 0.767 (95%CI: 0.511, 1.152) favoring the abemaciclib arm. A subsequent interim analysis provided additional information on the OS effect. At the time of the interim analysis (cutoff: 51 months from study start), the IDFS in the intent-to-treat (ITT) population remained statistically significant and the observed OS HR was 0.929 (95% CI: 0.748–1.153). Upon review of updated data at this second interim OS analysis, FDA broadened the approved population by removing the requirement for high Ki-67 expression. Although still pending final analysis, these results indicate a more favorable OS trend with further follow-up.

Key Insight: Interim OS data, especially when based on a small proportion of events relative to a large trial population, may not provide sufficient insight into clinical benefit. Benefit was initially observed in patients with high Ki-67 expression. As the data matured, including the observed OS HR dropping below 1, FDA determined that "… although OS remains immature and not statistically significant, a potential detriment in survival was no longer observed for the ITT population."[5] The indication was subsequently expanded to remove the requirement of a Ki-67 score of ≥20%. The original indication in the Ki-67 ≥20% population was only granted because the population was prespecified in the statistical hierarchy and had an OS HR <1, highlighting the importance of prespecifying subgroups in the statistical analysis plan. The indication was expanded after further follow up, highlighting the importance of ensuring long-term data collection to support broader treatment decisions.

3. MONALEESA-2 Trial (Ribociclib + Letrozole in HR+/HER2- Metastatic Breast Cancer)

In the MONALEESA-2 trial, ribociclib, a CDK4/6 inhibitor, was tested in combination with letrozole in 668 patients with HR+/HER2- metastatic breast cancer.[6–8] The PFS data indicated a statistically significant outcome (HR = 0.556, 95% CI: 0.429–0.720). However, the initial interim OS analysis (cutoff: 24 months from study start) revealed a hazard ratio greater than 1 (HR = 1.128, 95% CI: 0.619–2.055). At that time, only 43 OS events had been observed, with an information fraction of 11%. A pre-planned OS analysis one year later (cutoff: 36 months from study start) showed improved OS results with an HR of 0.746 (95% CI: 0.517–1.078). These updated data were submitted to regulators and considered in the initial approval. The final OS analysis (cutoff: 78 months from study start) also showed a statistically significant survival benefit (HR = 0.76, 95% CI: 0.63–0.93).

Key Insight: The timing of interim OS analyses is critical, as early assessments may not reflect the true treatment benefits. This underscores a broader challenge in drug development, where a small number of events and/or low information fractions can lead to misleading conclusions. It is crucial to consider the HR in combination with its confidence interval (and width), which can characterize the uncertainty that is present at interim analyses.

4. Bellini Trial (Venetoclax + Bortezomib in Relapsed/Refractory Multiple Myeloma)

The Bellini trial investigated the combination of venetoclax, a targeted BCL-2 inhibitor, with bortezomib in patients with relapsed or refractory multiple myeloma.[9] The trial included 291 patients. The PFS analysis showed a statistically significant outcome (HR = 0.63, 95% CI: 0.44–0.90). However, OS at the first interim analysis (cutoff: 18 months from study start) showed a hazard ratio greater than 1 (HR = 2.03, 95% CI: 1.04–3.95), suggesting detriment when analyzed in the overall ITT population. This risk of increased mortality was added to the FDA label for venetoclax under the Warnings and Precautions section, and a partial clinical hold was placed on clinical trials of venetoclax in patients with multiple myeloma. The subgroup of patients with the t(11;14) translocation did not show the same level of OS detriment, however, meaningful conclusions were not possible given the small size of the population (n=35). The final analysis of OS (33 months from study start) showed an OS HR of 1.19 (95% CI: 0.80-1.77), suggesting a lack of benefit and risk of increased mortality in the ITT population, but with a wide confidence interval. A Phase 3 trial (CANOVA) of venetoclax plus dexamethasone compared to pomalidomide plus dexamethasone was subsequently conducted in patients with t(11;14)-positive multiple myeloma; however, the trial failed to demonstrate statistical significance on the primary endpoint of PFS superiority.

Key Insight: Significant improvement in an early endpoint was observed, but with an OS detriment in the ITT population, which was later confirmed with more mature data. OS data indicated a possible benefit in a subgroup of patients, but the assessment was limited by the small sample size in this subgroup. This highlights the need for careful preplanning of interim analyses to assess harm, as well as appropriate powering of subgroup analyses to identify both potential benefits and risks within distinct patient populations. This approach ensures that meaningful effects are not overlooked and that potential detriment in other subgroups is properly addressed.

5. PI3K Inhibitors in Hematological Malignancies

Phosphatidylinositol 3-kinase (PI3K) inhibitors have been explored for their therapeutic potential in hematological malignancies. Four PI3K inhibitors—idelalisib, copanlisib, duvelisib, and umbralisib—received FDA approval for indications involving relapsed or refractory indolent non-Hodgkin lymphoma (NHL) or chronic lymphocytic leukemia (CLL). Despite showing promising results in terms of durable ORR or PFS, significant concerns emerged regarding their OS outcomes and tolerability.[10]

These drugs demonstrated substantial toxicities, including severe immune-mediated side effects such as hepatotoxicity, pneumonitis, colitis, and increased risk of infections. For example, idelalisib had halted trials in untreated CLL and indolent NHL due to increased deaths and severe adverse events. The UNITY-CLL trial of umbralisib showed an interim OS HR of 1.23, suggesting a potential increase in mortality compared to control.[11] Similar trends were observed across trials with other PI3K inhibitors, leading to safety concerns and voluntary withdrawals of certain indications.

Key Insight: The class-wide issues with PI3K inhibitors highlight the importance of evaluating both efficacy and safety comprehensively, particularly when using early endpoints like ORR to support initial approval. While these drugs improved ORR, the interpretation of their impact on OS is complicated by substantial toxicities that may negate the benefits. This underscores the need for careful assessment of the benefit-risk balance and ongoing OS monitoring, especially in cases where early endpoints show benefit but OS data suggest harm.

Frameworks and Strategies for Interpreting Interim OS Data

Interim OS data can provide early insights into both the potential benefits and risks associated with a treatment, but as noted, they are often challenging to interpret due to data maturity, leading to variable HRs and wide CIs. Over-reliance on point estimates, which do not reflect underlying uncertainties about harms or benefits, may lead to misinterpretation. More robust approaches for designing and interpreting interim OS data are needed to ensure reliability and accuracy.

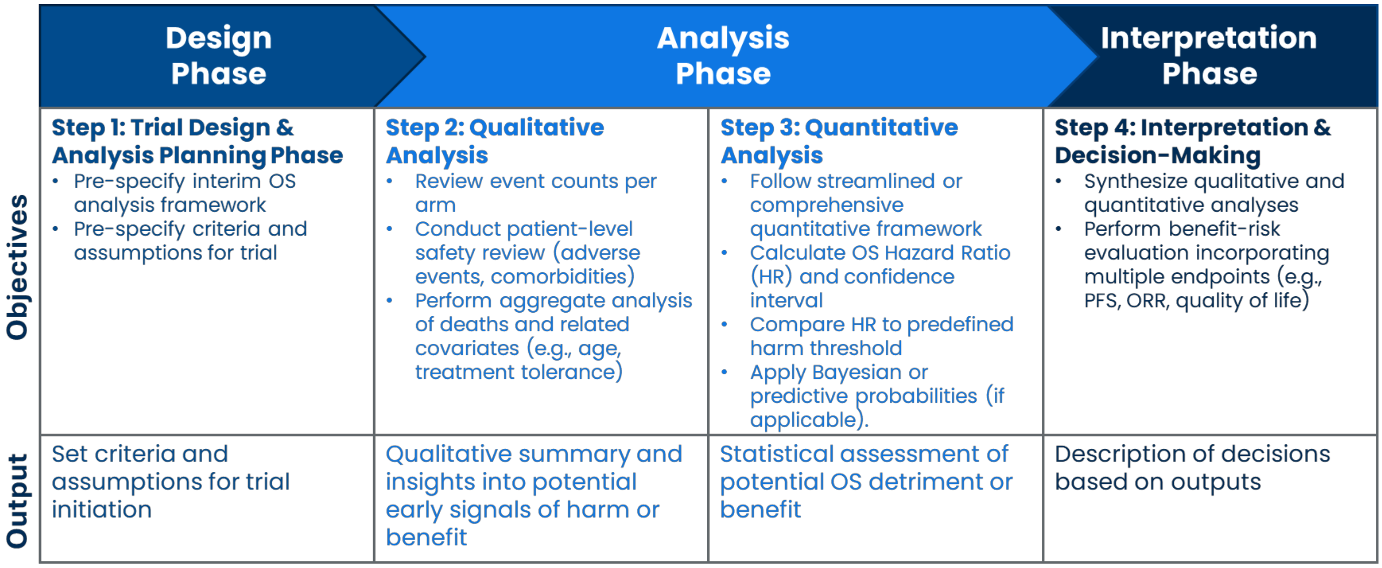
Current study design considerations may not adequately account for the complexities of interim OS analyses. Recent FDA-sponsored discussions and subsequent external publications emphasize the need for more thoughtful trial designs and comprehensive planning to consider these complexities when assessing potential harm using interim OS data.[12, 13] To address these challenges effectively, we consider two quantitative frameworks for interpreting interim OS data. Additional work to refine each and establish standards for trial sponsors and regulators regarding their practical implementation is desirable. These quantitative frameworks are briefly summarized below:

A **streamlined quantitative framework** can provide a more straightforward, predefined approach focusing on a standardized set of criteria for interim OS interpretation. This framework may be optimal in trials where patient crossover is not permitted and an assumption of proportional hazards is plausible, meaning that the underlying treatment effect is expected to remain consistent over time. In such cases the focus is on quantifying the uncertainty around the potential for unacceptable harm. Pre-specified thresholds, such as a minimum number of events and information fraction, and an upper limit for the HR CI, guide the interpretation and are particularly efficient when limited variability is expected in interim OS outcomes. However, it is important to tailor these thresholds to reflect the clinical considerations specific to each trial, including factors such as the disease setting, expected survival on the control therapy, and unmet medical need.

Alternatively, a **comprehensive quantitative framework** may be necessary for trials with more complexity, such as those involving non-proportional hazards or patient crossover.[4] This approach would incorporate a broader set of tools to enable deeper analysis when interim OS results are expected to be less conclusive or when complex trial dynamics may make it harder to get a reliable estimate of the treatment's true effect. This may include probabilistic assessments to quantify the likelihood of harm or benefit and the integration of qualitative factors such as patient comorbidities and subsequent therapies. This framework ensures that early signals of potential harm or benefit are not overlooked due to the complexity of the trial design or the mechanism of action of the novel drug.

A Multi-Step Approach for Study Design and Interpretation of Interim OS Data

These two proposed quantitative frameworks can be integrated into a multi-step approach for interpreting interim OS data. This multi-step approach incorporates a descriptive evaluation with the quantitative evaluation to provide a structured methodology for comprehensively understanding the data and ensuring decisions are evidence-based and aligned with trial objectives. This approach can also be used for interpretation of results and to prospectively align on the trial design features that can make the interpretation more reliable. The proposed multistep approach is summarized in Figure 1.



	Design Phase	Analysis Phase		Interpretation Phase
Objectives	**Step 1: Trial Design & Analysis Planning Phase** • Pre-specify interim OS analysis framework • Pre-specify criteria and assumptions for trial	**Step 2: Qualitative Analysis** • Review event counts per arm • Conduct patient-level safety review (adverse events, comorbidities) • Perform aggregate analysis of deaths and related covariates (e.g., age, treatment tolerance)	**Step 3: Quantitative Analysis** • Follow streamlined or comprehensive quantitative framework • Calculate OS Hazard Ratio (HR) and confidence interval • Compare HR to predefined harm threshold • Apply Bayesian or predictive probabilities (if applicable).	**Step 4: Interpretation & Decision-Making** • Synthesize qualitative and quantitative analyses • Perform benefit-risk evaluation incorporating multiple endpoints (e.g., PFS, ORR, quality of life)
Output	Set criteria and assumptions for trial initiation	Qualitative summary and insights into potential early signals of harm or benefit	Statistical assessment of potential OS detriment or benefit	Description of decisions based on outputs

Figure 1. Proposed Multi-Step Approach for Evaluating Interim OS Data in Oncology Trials.

While randomization allows for unbiased comparison across treatment arms, immature data and other factors previously noted make interpreting interim OS data difficult. In this context, it may be useful to consider alternative summary measures as supplementary analyses for the comparison of the interim OS data, rather than just the hazard ratio. Adequately powered, randomized comparisons are still the best approach for generating reliable effect estimates, but early insights can be gained by supplementing these comparisons with contextual analyses. Thus, the first step of interim OS data interpretation involves qualitative assessment, based on a structured descriptive summary of the available data. Sponsors could provide a review of the number of events, establishing a rate per person per year in each arm, a case-by-case examination of each death including precursor safety findings. A more in-depth patient-level assessment could explore whether adverse events (AEs) or lab abnormalities were related to, or led to, death. An analysis of comorbidities and dose considerations could be conducted at the patient and at the aggregate population level, which would involve determining if dose interruptions or reductions occurred in response to these AEs and whether they resolved after the changes. Evaluating patient baseline characteristics, such as age, existing comorbidities, and other risk factors, can provide additional insights into how these factors may have influenced outcomes, as comorbidities could exacerbate AEs or affect treatment tolerance. Further, examining pharmacokinetic (PK) exposure data may help identify whether unusually high drug exposure contributed to toxicity or death. While a case-by-case patient-level assessment might be necessary in some scenarios, a more practical approach for larger Phase 3 trials, may involve conducting an aggregate-level analysis that compares patients with different outcomes, such as those who survived and those who did not, to identify any meaningful differences. However, additional assumptions or alternative estimators may be needed to establish whether these observed differences reflect a harmful or beneficial causal effect of the novel drug.[14]

These descriptive insights can help provide necessary clinical context and identify any evidence of excess mortality or early signals of harm and confounding factors. Such descriptive analyses can help determine whether the interim OS results warrant deeper quantitative exploration, if there are sufficient data to do so.

Once a descriptive understanding is established, the prespecified streamlined or comprehensive quantitative framework can then be applied to further evaluate the data. Evaluations may include calculating the HR point estimate and an associated CI. One can envisage setting a threshold for harm and assessing the upper end of the CI to quantify the degree of uncertainty around potential for harm, as is done in classical statistical frameworks. As the data mature, the evidentiary threshold required to rule out harm may become more stringent, and it may help to consider two-sided confidence intervals less than 95% at interim analyses for assessment of harm.[12] Assessment of the risk of erroneous conclusion with regard to unacceptable OS detriment (False Positive and False Negative), under different assumptions, should be provided to determine the reliability of the results and facilitate a more transparent trade-off of risks associated with any decision making based on such interim OS data. Conditional probabilities or Bayesian predictive probabilities, based on current data and external evidence, may help predict whether the final OS outcome is likely to be neutral, beneficial, or detrimental. This quantitative step may support assessments of how early OS detriment can be established with high certainty, if it exists, and how frequently the wrong conclusions may be drawn.

The final step may involve synthesizing the descriptive and quantitative findings into a broader benefit-risk evaluation, considering multiple endpoints. This can include a totality of evidence approach, incorporating not only OS but also other endpoints such as PFS and ORR as well as safety, tolerability, and quality of life endpoints beyond OS.

Design Stage Considerations

Effective study design is crucial to ensuring reliable interim OS data interpretations. This section outlines key considerations to manage factors such as patient crossover or non-proportional hazards in the design stage. To allow for appropriate data capture and analysis, the decision to use a streamlined or comprehensive quantitative framework should be pre-specified during the design stage of the trial. This decision should consider factors such as risks for non-proportional hazards (e.g. the potential for delayed treatment effect or subgroups with heterogeneous treatment effects), early patient crossover, data maturity, duration of follow-up, and overall study power. Robust trial designs plan for adequate follow-up duration to ensure sufficient data collection and maturity at interim analyses as well as pre-specify criteria for patient crossover and minimize/manage missing data.

If heterogeneous treatment effects are expected in subgroups, the study would need to be sized appropriately to enable a thorough benefit-risk assessment in each of the subgroups. Leveraging historical data from similar therapies and/or patient populations is one strategy that can help estimate relationships between early endpoints and OS, as well as predict HRs and whether hazards are proportional throughout allowing for more informed predictions of potential outcomes. Evaluating how control groups perform on key endpoints can also help set expectations and provide context for interpreting OS findings. However, gaps remain in standardizing methods to determine the impact of early safety events on OS and in using historical data to establish specific thresholds for defining harm.

Including simulation of expected survival curves and determining the operating characteristics over a range of plausible assumptions and aligning these with the planned OS assessment criteria are important in the trial design stage. Such simulations can help set the timings for analyses and optimize the study's power, especially given that the timing of this OS interim analysis is often driven by PFS or ORR analysis milestones. Furthermore, it is important to plan for the collection of OS data even after final PFS analyses are completed, and to pre-specify OS interim analysis milestones and approaches for handling patient crossover. Many analysis techniques that adjust for crossover make an assumption of no unmeasured confounding. To support this assumption, trials would need to capture detailed information on potential (fixed or time-varying) confounders—specifically, patient-level covariates linked to both prognosis and the likelihood of crossover.[15] Collecting data on factors such as comorbidities, baseline characteristics, and treatment-related considerations can help mitigate confounding, enhancing the reliability of analytical assumptions. The impact of non-proportional hazards can also be anticipated and planned for at the design stage through simulations of various patterns such as delayed separation or early small excess harm followed by benefit. When non-proportional hazards are expected, additional metrics beyond the traditional OS HR, such as restricted mean survival time (RMST) or milestone survival rates (i.e., KM estimate at 1, 2, or 3 years), or piecewise hazard ratios (e.g., HRs from 0-6 months and after 6 months) may be considered and prespecified as supplementary analyses.[16] If these supplementary analyses are being considered to address non-proportional hazards, the design stage is the appropriate time to set the analysis interval cutoffs.

Timing is another critical consideration in interim OS analyses. The timing should balance the need for early decision-making with the risk of making incorrect decisions based on incomplete data. Conducting an analysis too early may lead to uncertain conclusions if there are not enough events to provide reliable information. This can be partially avoided through the pre-specification and agreement of a harm threshold for the interpretation of early OS data, and the level of evidence required at each analysis time point to rule out harm.

It is not only the timing of interim analyses that matters, but also the overall event accumulation rate for OS. In some scenarios, low event rates and the associated power for OS may mean that even waiting longer may not lead to significantly improved probability of detecting an OS treatment effect. The design stage can also be used to assess false positive and false negative rates based on the harm threshold and alternative (OS benefit) threshold, either through simulations or in some simple settings through modified power calculations.

Interpretation Stage Considerations

With a robustly established design which includes OS as a safety endpoint, a comprehensive and methodical interpretation of OS can begin at the primary endpoint assessment (e.g., the early endpoint). The interpretation stage can benefit from both descriptive and quantitative evaluations described above. Accurately interpreting interim OS data requires a range of analytical methods and metrics.

When efficacy trends in a subgroup differ from the overall study population or other subgroups, it is important to determine whether the observed OS detriment is likely due to chance or is plausible from a scientific, biological, or clinical perspective. For instance, does the difference align with the treatment's mechanism of action? Do patients in the subgroup have distinct clinical or biological characteristics that predispose them to a higher risk of adverse effects? Could variations be attributed to differences in clinical practices across sites or regions? It's also essential to examine the totality of the data in subgroups, including other safety and efficacy endpoints. If the detrimental OS is associated with higher incidences of serious or high-grade adverse events or lab abnormalities, this suggests a potential concern rather than a chance finding—this holds true whether observed in specific subgroups or across the overall trial population. Similarly, if the detriment is observed consistently across multiple endpoints, such as ORR, PFS, and OS, this further raises safety concerns.

Interpreting trial data alongside external evidence—such as literature, prior trials, or real-world data—may provide additional valuable insights, particularly when the trial sample size is limited or the data are immature. If the observed OS trend aligns with findings from prior trials of the same agent or others with a similar mechanism of action, this trend is less likely to be due to random chance.

When interpreting immature OS data at the time of an early endpoint analysis, revisiting design assumptions based on accrued information can offer valuable insights. Viewing this data in a Bayesian framework, where assumptions range from implausible to more likely scenarios, can help reviewers better visualize uncertainty and refine their expectations for future events. Additionally, other trial monitoring methods may be adaptable for evaluating OS as an early safety indicator. Characterizing error rates (e.g., false positives and negatives) and using tipping point analyses or other methods (e.g., Bayesian) to evaluate the robustness of interim OS results can help account for potential future variations.

Table 1 provides a summary of evolving strategies, outlining both current approaches and emerging best practices for improving the design and interpretation of interim OS data. It serves as a starting point to help navigate the complexities of using interim OS as a safety endpoint, managing trial design considerations, and handling data immaturity at interim analyses.

Table 1. Evolving Strategies for the Design and Interpretation of Interim OS Data.

Category	Current Approaches and Emerging Best Practices
Clarification of OS as a Safety Endpoint - OS is frequently used as a safety endpoint when evaluating early endpoint data.	Current Approach: - Typically, specify OS as a co-primary or secondary efficacy endpoint when feasible and clinically relevant. Emerging Best Practices: - Pre-specify OS analysis plans for safety, including clear definitions of OS detriment and thresholds for defining harm informed by discussions with regulatory authorities.
Trial Design Considerations - Factors like non-proportional hazards, patient crossover, data maturity and completeness, duration of follow-up, evolving standard of care, and study power can complicate interim OS interpretation.	Current Approach: - Some consideration of evaluation of design factors such as non-proportional hazards, crossover, data maturity and completeness aiming to minimize bias of OS assessment at OS planned interim analyses. Emerging Best Practices: - Emphasis on OS data collection beyond approval milestones. - In cases where cross-over is unavoidable, additional data collection on key baseline and time-varying covariates associated with patient prognosis and likelihood to crossover. - Systematic application of a quantitative framework for the transparent trade-off of risk of false negative vs false positive assessment of potential OS detriment - Monitor design assumptions closely and avoid deviations when possible.
Handling OS Data Immaturity at Interim Analysis - OS data is often immature at interim analyses, leading to variability and potential misinterpretation.	Current Approach: - Focus on OS driven interim analysis frameworks, which are most often group-sequential in nature. Emerging Best Practices: - A thorough assessment is performed to quantify the degree of uncertainty around potential for unacceptable detriment in OS. - Incorporate both qualitative and quantitative assessments to provide context for immature data. Engage patients through patient preference studies to define acceptable margins for potential OS detriment in specific settings. - Focus on standard key analyses (to be defined) such as comparing HR and CI to predefined harm thresholds and conducting qualitative patient-level assessments. - Use simulations and/or Bayesian models to refine predictions of final OS results, conditional on existing data and/or using external data if appropriate. - Use of tipping point or Bayesian framework to assess the robustness of interim OS data with respect to any future potential risk of detriment.

Future Considerations for Tool Development and Best Practice Alignment

This section lays out potential strategies for tool development and further best practice development that may support stakeholders in designing and interpreting interim analyses. These tools would be intended to streamline trial processes, improve decision-making, and enhance the robustness of OS data interpretation.

Use of Simulations for Enhanced Decision-Making

Simulations (see Appendix 2 for a concept proposal) can provide a powerful means to predict and explore the various outcomes at the interim stages of a trial, quantifying the operating characteristics of clinically driven decision-making thresholds, and ensuring robustness in trial design. Simulations can be particularly useful in oncology trials where data immaturity and crossover effects can obscure true treatment effects, providing a means to model outcomes under different conditions. Specifically:

- Simulations may help in understanding the possible trajectories of survival outcomes under different scenarios, such as varying treatment effects, patient heterogeneity, crossover, and information fractions.
- Simulations could be used to define the thresholds for potential harm or benefit and evaluate their operating characteristics, assess the impact of treatment crossover and non-proportional hazards, and inform the timing of interim analyses. They may also be valuable for identifying scenarios where low power for OS could result in less reliable conclusions even with extended follow-up.
- Simulations could be used to predict future outcomes based on current study data in various endpoints and integration of relevant external data if appropriate.

Development of Tools for Design and Interpretation

In addition to simulations, practical tools could be developed to guide sponsors and researchers in designing trials and interpreting interim OS results more effectively. These may include:

- A structured assessment aid could be developed to assist sponsors during the trial design stage. This tool may consist of a uniform set of questions to help guide thinking around key aspects such as patient crossover, pre-specifying OS analysis milestones, determining adequate follow-up durations, data collection on important patient covariates associated with prognosis and key intercurrent events, and patient heterogeneity.
- Providing standardized methodology for key decisions that impact quality and completeness of data (OS and other data) collection could help ensure consistency across trials and provide clear guidance on how to mitigate bias and enhance the reliability of interim OS data.
- Bayesian modeling approaches could be incorporated as a complementary tool to simulations and standard conditional probabilities. These models may provide probabilistic statements regarding the magnitude of the final OS treatment effect (e.g., HR) based on the observed interim data. Priors may be informed by historical relationships between early endpoints (e.g., PFS, ORR) and OS,

allowing for a more nuanced and evidence-based interpretation. Using Bayesian frameworks could allow for the integration of new information as it becomes available as well as data outside of the study, thus improving the precision of interim OS estimates and supporting better-informed decisions.

Best Practices Alignment

To facilitate consistent application of best practices in designing and interpreting interim OS data, aligned best practices could be developed and uniformly adopted by stakeholders, reducing variability in approaches to interpreting interim OS data.

Conclusion

The interpretation of interim OS data in oncology trials poses unique challenges. Early endpoints, such as ORR and PFS, are often the basis for accelerated approvals, allowing timely access to potentially beneficial therapies. However, instances in which immature OS data conflicts with efficacy signals detected with early endpoints can lead to uncertainty around the treatment's benefit-risk profile. This white paper highlights the importance of carefully evaluating and considering interim OS data, so it provides meaningful data to support evaluation of new drugs.

The case studies provided illustrate how factors such as immature OS data, patient subgroups, early patient crossover, and information fractions impact the interpretation of results. These examples reinforce the need for a framework that integrates descriptive and quantitative analyses, supported by thorough preplanning, to ensure accurate conclusions.

Future Directions

To address these challenges and optimize the use of interim OS data in oncology drug development, several next steps should be considered:

1. Adopting a structured, multi-step approach for interpreting interim OS data, starting with descriptive assessments and followed by quantitative analyses tailored to the therapy and clinical context.
2. Prioritizing robust trial designs that pre-specify OS milestones, harm thresholds, and strategies for incomplete data handling. Simulations can be used to predict outcomes and optimize designs.
3. Fostering continued collaboration among regulators, sponsors, and statisticians to harmonize methods for evaluating interim OS data. Future efforts should focus on refining simulation methods, threshold setting, and predictive modeling, tailored to oncology trials.

By addressing these key areas, the interpretation of interim OS data can be improved, leading to more accurate, timely decisions that benefit patients.

References

1. Zhang JJ, Blumenthal GM, He K, et al: Overestimation of the Effect Size in Group Sequential Trials. Clin Cancer Res 18:4872–4876, 2012. Available from: https://pubmed.ncbi.nlm.nih.gov/22753584/

2. Morris MJ, Castellano D, Herrmann K, et al: 177Lu-PSMA-617 Versus a Change of Androgen Receptor Pathway Inhibitor Therapy for Taxane-Naive Patients with Progressive Metastatic Castration-Resistant Prostate Cancer (PSMAfore): A Phase 3, Randomised, Controlled Trial. Lancet 404:1227–1239, 2024. Available from: https://pubmed.ncbi.nlm.nih.gov/39293462/

3. Johnston SRD, Toi M, O'Shaughnessy J, et al: Abemaciclib Plus Endocrine Therapy for Hormone Receptor-Positive, HER2-Negative, Node-Positive, High-Risk Early Breast Cancer (Monarche): Results From a Preplanned Interim Analysis of a Randomised, Open-Label, Phase 3 Trial. Lancet Oncol 24:77–90, 2023. Available from: https://pubmed.ncbi.nlm.nih.gov/36493792/

4. Royce M, Osgood C, Mulkey F, et al: FDA Approval Summary: Abemaciclib with Endocrine Therapy for High-Risk Early Breast Cancer. J Clin Oncol 40:1155–1162, 2022. Available from: https://pubmed.ncbi.nlm.nih.gov/35084948/

5. Royce M, Mulkey F, Osgood C, et al: US Food and Drug Administration Expanded Adjuvant Indication of Abemaciclib in High-Risk Early Breast Cancer. J Clin Oncol 41:3456–3457, 2023. Available from: https://pubmed.ncbi.nlm.nih.gov/37104738/

6. Hortobagyi GN, Stemmer SM, Burris HA, et al: Ribociclib as First-Line Therapy for HR-Positive, Advanced Breast Cancer. N Engl J Med 375:1738–1748, 2016. Available from: https://pubmed.ncbi.nlm.nih.gov/27717303/

7. Hortobagyi GN, Stemmer SM, Burris HA, et al: Updated Results from MONALEESA-2, A Phase III Trial of First-Line Ribociclib Plus Letrozole Versus Placebo Plus Letrozole in Hormone Receptor-Positive, HER2-Negative Advanced Breast Cancer. Ann Oncol 29:1541–1547, 2018. Available from: https://pubmed.ncbi.nlm.nih.gov/29718092/

8. Hortobagyi GN, Stemmer SM, Burris HA, et al: Overall Survival with Ribociclib plus Letrozole in Advanced Breast Cancer. N Engl J Med 386:942–950, 2022. Available from: https://pubmed.ncbi.nlm.nih.gov/35263519/

9. Kumar SK, Harrison SJ, Cavo M, et al: Venetoclax or Placebo in Combination with Bortezomib and Dexamethasone in Patients with Relapsed or Refractory Multiple Myeloma (BELLINI): A Randomised, Double-Blind, Multicentre, Phase 3 Trial. Lancet Oncol 21:1630–1642, 2020. Available from: https://pubmed.ncbi.nlm.nih.gov/33129376/

10. Meeting of the Oncologic Drugs Advisory Committee Meeting Announcement - 04/21/2022 | FDA. Available from: https://www.fda.gov/advisory-committees/advisory-committee-calendar/updated-information-april-21-22-2022-meeting-oncologic-drugs-advisory-committee-meeting-announcement

11. TG Therapeutics Announces Voluntary Withdrawal of the BLA/sNDA for U2 to Treat Patients with CLL and SLL | TG Therapeutics, Inc. Available from: https://ir.tgtherapeutics.com/news-releases/news-release-details/tg-therapeutics-announces-voluntary-withdrawal-blasnda-u2-treat

12. Fleming TR, Hampson L V, Bharani-Dharan B, et al: Monitoring Overall Survival in Pivotal Trials in Indolent Cancers. 2023. Available from: https://arxiv.org/abs/2310.20658v3

13. Rodriguez LR, Gormley NJ, Lu R, et al: Improving Collection and Analysis of Overall Survival Data. Clin Cancer Res 30:OF1–OF9, 2024 Available from: https://pubmed.ncbi.nlm.nih.gov/39037364/

14. Bas BB, Groenwold RHH: Identification of Causal Effects in Case-Control Studies. BMC Med Res Methodol 22, 2022. Available from: https://pubmed.ncbi.nlm.nih.gov/34996386/

15. Watkins C, Huang X, Latimer N, et al: Adjusting Overall Survival for Treatment Switches: Commonly Used Methods and Practical Application. Pharm Stat 12:348–357, 2013. Available from: https://pubmed.ncbi.nlm.nih.gov/24136868/

16. Lin RS, Lin J, Roychoudhury S, et al: Alternative Analysis Methods for Time to Event Endpoints under Non-proportional Hazards: A Comparative Analysis. Stat Biopharm Res 12:187–198, 2019. Available from: https://www.tandfonline.com/doi/full/10.1080/19466315.2019.1697738

Appendices

Appendix 1. Summary of Select Oncology Clinical Trials: Early Endpoint and Interim Overall Survival Readouts.

Study	Patient Population	N	Early Endpoint (HR, 95% CI)	Early Endpoint Events	OS Readout	OS Events	OS HR (95% CI)	Cut-Off Date
PSMAFore Trial (177Lu-PSMA-617)	Metastatic Castration-Resistant Prostate Cancer (mCRPC)	468	rPFS HR = 0.41 (95% CI: 0.29–0.56)	166	Interim OS #2	134	1.16 (0.83, 1.64)	21-Jun-2023
					Interim OS #3	216	0.98 (0.75, 1.28)	27-Feb-2024
					Final OS	Pending	Pending	Pending
MonarchE Trial (Abemaciclib + ET)	Early HR+/HER2-Breast Cancer	5637	IDFS HR = 0.696 (0.588–0.823)	565	Interim OS #1	186	1.09 (0.82, 1.46)	1-Apr-2021
					Interim OS #2	330	0.93 (0.75, 1.15)	1-Jul-2022
					Interim OS #3	442	0.90 (0.75, 1.09)	3-Jul-2023
					Final OS	Pending	Pending	Pending
MONALEESA-2 (Ribociclib + Letrozole)	HR+/HER2-Metastatic Breast Cancer	668	PFS HR = 0.556 (0.429–0.720)	243	OS Interim #1	43	1.13 (0.62, 2.06)	29-Jan-2016
					OS Update	116	0.75 (0.52, 1.08)	2-Jan-2017
					Final OS	400	0.76 (0.63, 0.93)	10-Jan-2021
Bellini Trial (Venetoclax + Bortezomib)	Relapsed/Refractory Multiple Myeloma	291	PFS HR = 0.63 (0.44–0.90)	129	OS Interim #1	52	2.03 (1.04, 3.95)	26-Nov-2018
					Final OS	114	1.19 (0.80, 1.77)	15-Mar-2021

Appendix 2. Concept Plan and Future Directions for Interim OS Data Interpretation.

The simulation workplan is divided into distinct components, as described below. This initial work focuses solely on OS, largely independent of PFS. Future work can incorporate PFS directly into joint models or indirectly as part of a scenario regarding the totality of evidence across multiple endpoints.

1. Establish and evaluate various thresholds in a 'streamlined' criteria for harm based solely on the observed events available at the interim. This is done under the simplest assumptions to triangulate the initial set of criteria to be included in the evaluation. For example, we may find that a stringent criterion such as the upper bound of the confidence interval of the hazard ratio of 1.3 is almost equivalent to a test of efficacy, making it irrelevant to the intent of ruling out harm. Likewise, a value of 1.8 may be found to be too lenient, allowing obviously concerning scenarios to occur in an undesirably large proportion of simulation trials.
    a. It may be possible to determine a reasonable range of target operating characteristics equivalent to Type I and Type II error from the work above.
2. Evaluate existing frameworks and/or devise a new mathematical representation of the more complex scenarios that have occurred in practice, including considerations such as non-proportional hazards (e.g., early overlap of Kaplan-Meier curves followed by later separation, or early harm followed by separation), patients crossing over to the treatment arm at disease progression, dropout rates, information fraction, and the number of events available at the interim. This can be done by digitizing real examples, such as those described above, into piecewise hazard functions or by generating hypothetical scenarios. In either case, we can then evaluate the operating characteristics of the various frameworks, as outlined below.
3. Provisional Scenarios:
    a. Neutral effect on OS, proportional hazard of 1.0 throughout the trial.
    b. Separation of OS Kaplan-Meier curves, proportional hazard of modest scale (e.g., HR 0.9).
    c. Separation of OS Kaplan-Meier curves, proportional hazard of significant scale (e.g., HR 0.6).
    d. Delayed and modest separation of OS Kaplan-Meier curves after the interim (non-proportional hazard: 1.0 prior, 0.9 thereafter).
    e. Delayed and significant separation of OS Kaplan-Meier curves after the interim (non-proportional hazard: 1.0 prior, 0.6 thereafter).
    f. Small excess harm prior to the interim and small separation thereafter (non-proportional hazard: 1.15 prior, 1.0 for a period, 0.9 thereafter).
    g. Small excess harm prior to the interim and wide separation thereafter (non-proportional hazard: 1.15 prior, 1.0 for a period, 0.6 thereafter).

    h. Early modest benefit with later modest harm, with initial separation of OS Kaplan-Meier curves showing a proportional hazard of 0.9, followed by a reversal to show modest harm (HR 1.15 thereafter).
    i. Sustained modest harm throughout, reflected by a consistent proportional hazard of 1.15 maintained throughout the trial.
    j. Increasing harm over time, with initial modest harm (HR 1.15) that intensifies to a more significant level (HR 1.3 thereafter).

4. For each scenario of interest, use a variety of approaches to evaluate the following:
    a. How often do we incorrectly conclude harm when there isn't any?
    b. If there is harm, how often can we conclude correctly based on early data?
5. For both questions, establish whether there is a minimum amount of data (e.g., number of events) that is optimal for reasonably reliable decision making.
6. The concepts above are mainly applied to the interpretation of interim OS data; however, the true value lies in translating them to the design stage. We propose a few examples that will identify a recommended process flow for design considerations and the approach at each step.