Paper PP24

Beyond Log-Rank: Advanced Survival Analysis for Non-Proportional Hazards in Oncology Trials

Jagdish Nikam, Phastar, India

ABSTRACT

Traditional survival analysis in oncology relies on the log-rank test and Cox model, which assume proportional hazards. However, modern therapies like immunotherapies often violate this assumption, leading to delayed effects or crossing curves. This paper explores alternative methods—Restricted Mean Survival Time (RMST), MaxCombo tests, weighted log-rank tests, and time-varying Cox models—to address non-proportional hazards. Using a simulated study with SAS programming, we compare these methods and offer guidance on selecting the appropriate approach based on trial characteristics. The aim is to help statisticians apply more robust, clinically meaningful tools beyond traditional log-rank tests in oncology trials.

INTRODUCTION

Time-to-event endpoints such as overall survival and progression-free survival are central to the evaluation of treatment efficacy in oncology clinical trials. The log-rank test and Cox proportional hazards model are widely used standard methods for analyzing such endpoints due to their simplicity and efficiency. However, both approaches rely on the proportional hazards assumption, which implies that the treatment effect on event risk remains constant over time.

In practice, this assumption is frequently violated in oncology studies. Modern therapies, including immuno-oncology agents and targeted treatments, often exhibit delayed treatment effects, durable responses in a subset of patients, or treatment switching after disease progression. These features can lead to non-proportional hazards, reducing statistical power and potentially producing misleading estimates when traditional methods are applied.

Moreover, hazard ratios, the primary estimand from Cox models—represent instantaneous relative treatment effects and may be difficult to interpret clinically, particularly for patients and healthcare decision-makers who are more interested in absolute or time-averaged measures of benefit.

To address these limitations, alternative and more flexible statistical approaches have been proposed. Methods such as restricted mean survival time (RMST), weighted log-rank tests, and models allowing time-varying treatment effects provide robust and clinically interpretable assessments of treatment benefit when proportional hazards assumptions do not hold.

FDA- INITIATED COLLABORATION

Recognizing the increasing prevalence of non-proportional hazards in oncology clinical trials and the limitations of traditional survival analysis methods under such conditions, the U.S. Food and Drug Administration (FDA) highlighted this issue during the *Public Workshop on Oncology Clinical Trials in the Presence of Non-Proportional Hazards*. The workshop emphasized the need for greater clarity and consistency in the selection, interpretation, and reporting of statistical methods when the proportional hazards assumption is violated.

In response, the FDA initiated a collaborative dialogue with industry statisticians, academic researchers, and other key stakeholders to collectively address these methodological challenges. A series of meetings held in 2016 and subsequently in 2017 underscored the consensus that a systematic, structured, and evidence-based evaluation of available statistical approaches was necessary. Participants agreed that ad hoc or single-method solutions were insufficient given the diverse patterns of non-proportional hazards observed across oncology trials.

The primary objective of this collaborative effort is to identify and evaluate appropriate analytical methods tailored to different non-proportional hazard scenarios, such as delayed treatment effects, diminishing effects over time, or the presence of long-term survivors. Importantly, the group acknowledged that no single statistical method is universally optimal, and that method selection should be guided by the underlying hazard structure, clinical context, and trial objectives.

This initiative is explicitly non-product-specific and designed to foster cross-industry collaboration. Multiple pharmaceutical companies and academic partners contributed expertise and data, while the FDA actively participates in discussions, methodological evaluations, and interpretation of findings. The overarching goal is to promote robust, transparent, and clinically meaningful survival analyses that support sound regulatory decision-making and improve the interpretability of oncology trial results.

CONSIDERATIONS FOR PRIMARY ANALYSIS UNDER ICH E9

The ICH E9 guideline emphasizes the critical importance of pre-specifying all key elements of a clinical trial's design, conduct, and statistical analysis in a protocol finalized before trial initiation. This includes clear definition of the primary estimand, analysis population, statistical methods, and handling of intercurrent events. Adherence to these pre-specified plans and execution of the primary analysis as planned are fundamental to ensuring the credibility, interpretability, and regulatory acceptability of trial results.

Pre-specification of the primary analysis is particularly important for time-to-event endpoints, as deviations from the planned analysis may introduce bias or raise concerns regarding data-driven decision-making. Regulatory confidence in trial conclusions is strengthened when the protocol-defined analysis is followed and supported by appropriate sensitivity analyses.

However, planning an optimal primary analysis becomes more complex in the presence of potential non-proportional hazards (NPH). In many oncology trials, the underlying hazard structure is difficult to predict at the design stage. The timing, onset, and durability of treatment effects—especially for novel therapies such as immuno-oncology agents—may not be fully understood prior to trial initiation. As a result, it is challenging to prospectively select a single statistical method that is guaranteed to perform well across all possible NPH scenarios.

This uncertainty highlights the need for thoughtful protocol planning that anticipates potential departures from proportional hazards. Strategies may include pre-specification of alternative or complementary estimands, inclusion of sensitivity analyses using methods robust to NPH, or adoption of flexible analytical frameworks that preserve type I error while maintaining clinical interpretability. Such approaches align with the principles of ICH E9 by balancing methodological rigor with the practical realities of modern oncology trial designs.

PRIMARY ANALYSIS METHODS UNDER NON-PROPORTIONAL HAZARDS

When non-proportional hazards (NPH) are anticipated or observed, careful selection of primary analysis methods is essential to ensure statistical validity, adequate power, and clinical interpretability. Several classes of methods are available, each addressing different hazard patterns and trial objectives. These methods can be considered for primary or supportive analyses, depending on the pre-specified estimand and regulatory context.

1. Rank-Based Tests

Rank-based tests remain widely used in time-to-event analyses due to their simplicity and well-understood properties. The standard log-rank test (LRT) is optimal when the proportional hazards assumption holds and provides maximum power under constant relative treatment effects over time. However, in the presence of NPH, particularly with delayed or diminishing effects, the LRT may suffer from reduced power or fail to detect clinically meaningful differences.

Weighted log-rank tests extend the classical log-rank framework by incorporating time-dependent weights, allowing emphasis on early or late survival differences. These tests are particularly useful when the anticipated treatment effect is expected to occur after a delay (e.g., immuno-oncology therapies) or predominantly early in follow-up. The choice of weighting scheme should be clinically motivated and pre-specified to avoid data-driven inference.

2. Combination Tests

Combination of testing approaches, such as the max-combo test, address uncertainty in the underlying hazard pattern by combining multiple weighted log-rank tests, each targeting a different treatment effect profile (e.g., early effect, proportional hazards, delayed effect). The test statistic is derived from the maximum of these component tests while maintaining overall type I error control.

Max-combo tests are robust across a wide range of NPH scenarios and have gained increasing acceptance in regulatory guidance for use as a primary analysis method when the hazard structure is unknown at trial design. By protecting power under multiple plausible alternatives, combination tests offer a pragmatic and pre-specifiable solution consistent with ICH E9 principles.

The Breslow (generalized Wilcoxon) test is another rank-based method that assigns greater weight to early events. This approach may be appropriate when early treatment effects are clinically meaningful, such as in aggressive disease settings with rapid event accrual.

3. Kaplan–Meier–Based Methods

Kaplan–Meier (KM)–based approaches provide intuitive and transparent summaries of survival experience. Weighted KM tests allow formal comparison of survival curves while accounting for time-dependent weighting, making them useful as supportive or sensitivity analyses under NPH. These methods are particularly valuable for visualizing treatment effects over time and supporting clinical interpretation.

Restricted mean survival time (RMST) has emerged as a key alternative estimand in the presence of NPH. RMST compares the average survival time up to a pre-specified truncation time and does not rely on the proportional hazards assumption. As an absolute measure of treatment benefit, RMST is often more interpretable for clinicians and patients than hazard ratios. RMST-based analyses can be pre-specified as a primary or key secondary analysis, particularly when delayed or long-term treatment effects are anticipated.

Advantage and disadvantages of advanced Methods for Non-Proportional Hazards

1. Restricted Mean Survival Time (RMST)
RMST measures the average survival time up to a pre-specified truncation time and compares this quantity between treatment groups, providing an absolute measure of treatment benefit.

Pros:

- Does not rely on the proportional hazards assumption
- Clinically intuitive and easy to communicate (e.g., months gained)
- Robust under delayed or changing treatment effects

Cons:

- Requires pre-specification of truncation time
- Results depend on the chosen time horizon
- Less efficient than Cox model when PH holds

2. Weighted Log-Rank Test
Extends the standard log-rank test by applying time-dependent weights to emphasize survival differences during specific periods, such as early or late follow-up.

Pros:

- Improved power when treatment effects are delayed or early
- Maintains familiar log-rank framework
- Straightforward to implement

Cons:

- Requires correct anticipation of timing of treatment effect
- Choice of weights may be subjective
- Limited interpretability beyond hypothesis testing

3. Max-Combo Test (Combination of Fleming–Harrington Tests)
Combines multiple weighted log-rank tests, each targeting different hazard patterns (early, proportional, delayed). The maximum standardized statistic is used with multiplicity adjustment.

Pros:

- Robust across a wide range of NPH scenarios
- Preserves power when hazard pattern is unknown
- Recognized in FDA/EMA guidance
- Easily pre-specified, aligning with ICH E9 principles

Cons:

- Provides a global test without a direct effect estimate
- More complex to explain to non-statistical audiences
- Requires additional supportive analyses for interpretation

4. Milestone Survival Analysis
Compares survival probabilities between treatment groups at fixed, clinically meaningful time points (e.g., 12 or 24 months).

Pros:

- Highly interpretable for clinicians and regulators
- Effective for communicating long-term benefit
- Particularly useful in immunotherapy trials

Cons:

- Ignores survival information before and after the milestone
- Sensitive to censoring around the milestone time
- Not suitable as a standalone primary analysis

5. Time-Varying Cox Models (Piecewise Cox Model)
Divides follow-up time into pre-defined intervals and estimates separate hazard ratios for each period, allowing treatment effects to vary over time.

Pros:

- Explicitly models changing treatment effects
- Provides detailed insight into effect evolution
- Flexible and informative under NPH

Cons:

- Requires pre-specification of time intervals
- Increased model complexity
- Multiple hazard ratios may complicate interpretation

Study Design

This analysis employs a simulation-based example to illustrate the challenges posed by non-proportional hazards, particularly in situations where survival curves cross over time. Such scenarios are increasingly relevant in oncology trials, for instance when treatment effects are delayed or evolve differently across patient subgroups.

Survival times were generated for two treatment groups with distinct hazard patterns:

- **Group A:** Survival times were simulated from an exponential distribution with an added delay to induce a delayed treatment effect. This design produces poorer survival in the early follow-up period, followed by improved outcomes at later time points, reflecting the behavior of treatments such as immuno-oncology agents.
- **Group B:** Survival times were simulated from a standard exponential distribution without any delay, resulting in earlier events and better early survival compared with Group A.

A sample size of 50 subjects per group was chosen to ensure sufficient events for demonstrating NPH patterns, while keeping the simulation tractable for illustrative purposes.

This data-generating mechanism was intentionally designed to produce crossing Kaplan–Meier curves, which directly violate the proportional hazards assumption. Consequently, traditional methods such as the standard Cox proportional hazards model or log-rank test may provide misleading results or reduced power. This motivates the use of alternative analytical approaches, including piecewise hazard ratios, restricted mean survival time (RMST), weighted log-rank tests, and max-combo tests, which are more robust and interpretable under NPH.

By simulating this scenario, the analysis provides a clear, reproducible framework to explore how different methods perform when the treatment effect changes over time, offering practical insights into trial design and primary analysis planning in the presence of non-proportional hazards.

```sas
data survdata;
    call streaminit(123);

    do id = 1 to 50;  /* Group A */
        group = "A";
        /* Exponential survival times + delay to create crossing */
        time = rand("Exponential", 5) + 2;
        event = rand("Bernoulli", 0.7);
        output;
    end;

    do id = 51 to 100; /* Group B */
        group = "B";
        time = rand("Exponential", 5);  /* earlier events */
        event = rand("Bernoulli", 0.6);
        output;
    end;
run;
```

Interpretation of Overall and Piecewise Hazard Ratios

1. Overall Hazard Ratio

- Estimate: HR = 1.434
- 95% Confidence Interval (CI): 0.857 – 2.416

Statistical Interpretation:

- The overall HR compares the instantaneous risk of the event in Group B versus Group A across the entire follow-up period.
- HR > 1 suggests higher risk in Group B overall.
- However, the 95% CI includes 1 (0.857–2.416), meaning the difference is not statistically significant at the conventional α = 0.05 level.
- We cannot conclude there is a definitive overall difference in survival between Group A and B when using a single, time-averaged hazard ratio.

Limitation:

- Overall HR assumes proportional hazards over time. If the hazard ratio changes with time (non-proportional hazards), this single summary can mask important differences in specific time intervals.

2. Piecewise Hazard Ratios

To explore time-dependent effects, the follow-up was split into three intervals:

Interval (Months)	HR (B vs A)	95% CI	Interpretation
0–5	1.532	0.815 – 2.910	Slightly higher hazard in B early on; not significant
5–10	0.537	0.077 – 2.509	Lower hazard in B mid-period; not significant
10–15	1.993	0.648 – 7.362	Higher hazard in B late-period; not significant

Statistical Interpretation:

Interval 0–5 months (HR = 1.532, 95% CI: 0.815–2.910):

- Group B had a slightly higher risk of events early on.
- CI includes 1 → not statistically significant, meaning the observed difference could be due to chance.
- Reflects poorer early survival in Group B, consistent with crossing survival curves.

Interval 5–10 months (HR = 0.537, 95% CI: 0.077–2.509):

- Group B's hazard appears lower than Group A during the mid-period.
- Wide CI indicates high uncertainty, likely due to fewer events in this interval.
- Again, not statistically significant, so no definitive conclusion can be drawn.

Interval 10–15 months (HR = 1.993, 95% CI: 0.648–7.362):

- Group B shows a trend toward higher risk late in follow-up.
- CI is wide → not statistically significant.
- Suggests late crossing effect, but uncertainty remains due to limited events.

3. Pattern Suggests Non-Proportional Hazards (NPH)

- The piecewise HRs vary substantially over time: early (HR >1), mid (HR <1), late (HR >1).
- This changing treatment effect violates the proportional hazards assumption, which underlies the standard Cox model and overall HR.
- Key implication: the overall HR (1.434) may be misleading, as it averages these time-dependent effects and masks the fact that Group B performs better in some periods and worse in others.

4. Clinical Interpretation

- Early survival: Group B may experience slightly worse outcomes (HR > 1), potentially due to delayed treatment benefit or early toxicity.
- Mid follow-up: Group B may have better survival (HR < 1), suggesting a delayed treatment effect emerges.
- Late follow-up: HR again increases, indicating heterogeneity in long-term effects or small sample variability.
- Treatment effect is time-dependent, so relying solely on the overall HR could lead to incorrect clinical conclusions.

5. Statistical Recommendations

Given the observed NPH pattern:

- Piecewise Cox models or time-varying Cox models are appropriate to capture changing hazard ratios over time.
- Restricted mean survival time can provide an absolute, interpretable summary of survival difference that does not rely on proportional hazards.
- Weighted log-rank or max-combo tests could be used to detect differences robustly, especially when delayed effects are expected.

Note: Results are based on simulated data and are presented for illustrative purposes only. The objective is to demonstrate time-varying hazard ratios under non-proportional hazards rather than statistical significance.

Conclusion

Non-proportional hazards are increasingly common in oncology clinical trials, particularly in the presence of delayed treatment effects and long-term survivors. In such settings, reliance on a single hazard ratio from a proportional hazards model may obscure important time-dependent treatment effects and limit clinical interpretability.

Using a simulated example with crossing survival curves, this work illustrates how hazard ratios can vary across follow-up intervals, demonstrating violation of the proportional hazards assumption. The observed lack of statistical significance is expected given the illustrative nature of the simulation and does not detract from the key message that treatment effects may change over time.

Advanced analytical approaches—such as restricted mean survival time, weighted and combination log-rank tests (e.g., max-combo), milestone survival analysis, and time-varying Cox models—provide robust and interpretable alternatives when proportional hazards do not hold. These methods align with regulatory expectations and can be pre-specified at the design stage, consistent with ICH E9 principles.

Overall, careful consideration of non-proportional hazards at the trial design and analysis stages is essential to ensure valid inference, improve clinical interpretability, and support regulatory decision-making in modern oncology trials.

REFERENCES

Center for Health Policy is part of Duke University, Public Workshop: Oncology Clinical Trials in the Presence of Non-Proportional Hazards

https://healthpolicy.duke.edu/events/public-workshop-oncology-clinical-trials-presence-non-proportional-hazards

An Overview of Introductory and Advanced Survival Analysis Methods in Clinical Applications: Where Have we Come so far?

https://ar.iiarjournals.org/content/44/2/471

ICH E9 statistical principles for clinical trials - Scientific guideline

https://www.ema.europa.eu/en/ich-e9-statistical-principles-clinical-trials-scientific-guideline

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:
Author Name: Jagdish Nikam
Company: Phastar
Email: jagdish.nikam@phastar.com