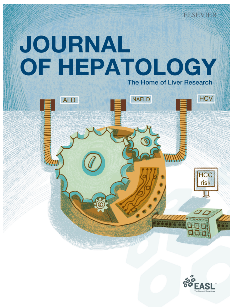


Strategies to address non-proportional hazards between survival curves - Lessons from phase III trials in hepatocellular carcinoma

Ezequiel Mauro, Tiago de Castro, Marcus Zeitlhoefler, Allan Hackshaw, MinJae Lee, Tim Meyer, Amit G. Singal, Josep M. Llovet

PII: S0168-8278(25)02492-4

DOI: https://doi.org/10.1016/j.jhep.2025.08.042

Reference: JHEPAT 10279

To appear in: *Journal of Hepatology*

Received Date: 31 January 2025

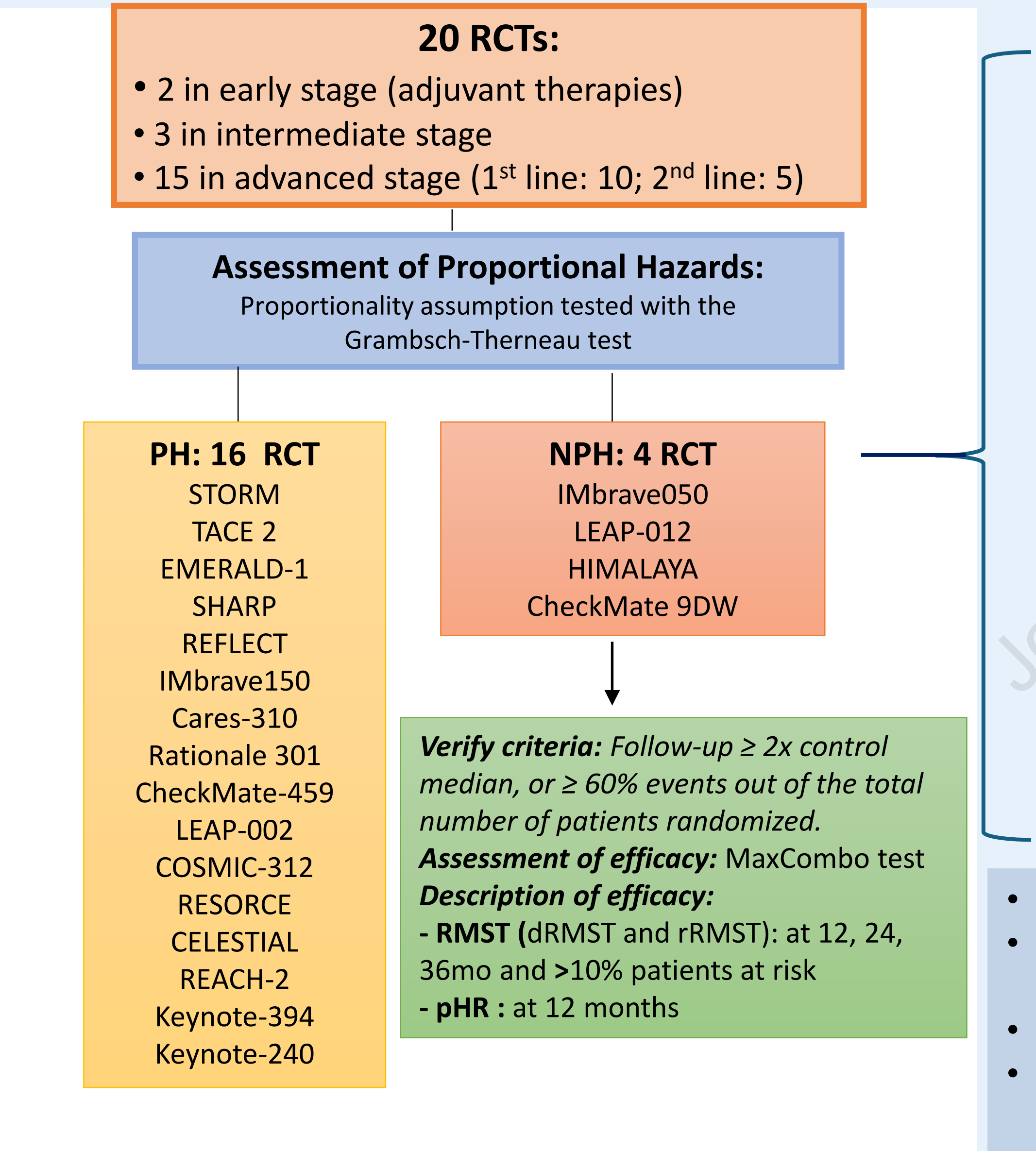
Revised Date: 21 August 2025

Accepted Date: 25 August 2025

Please cite this article as: Mauro E, de Castro T, Zeitlhoefler M, Hackshaw A, Lee M, Meyer T, Singal AG, Llovet JM, Strategies to address non-proportional hazards between survival curves - Lessons from phase III trials in hepatocellular carcinoma, *Journal of Hepatology*, https://doi.org/10.1016/j.jhep.2025.08.042.

Strategies to address non-proportional hazards (NPH) in phase III trials in hepatocellular carcinoma

Flowchart of 20 phase III HCC Trials



20 RCTs:

- 2 in early stage (adjuvant therapies)
- 3 in intermediate stage
- 15 in advanced stage (1st line: 10; 2nd line: 5)

Assessment of Proportional Hazards:

Proportionality assumption tested with the Grambsch-Therneau test

PH: 16 RCT

STORM
TACE 2
EMERALD-1
SHARP
REFLECT
IMbrave150
Cares-310
Rationale 301
CheckMate-459
LEAP-002
COSMIC-312
RESORCE
CELESTIAL
REACH-2
Keynote-394
Keynote-240

NPH: 4 RCT

IMbrave050
LEAP-012
HIMALAYA
CheckMate 9DW

Verify criteria: *Follow-up ≥ 2x control median, or ≥ 60% events out of the total number of patients randomized.*

Assessment of efficacy: MaxCombo test

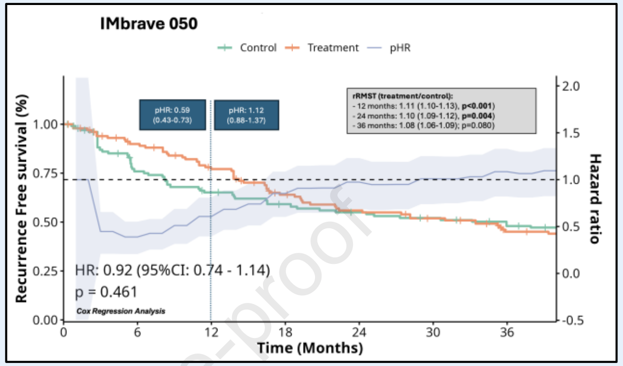
Description of efficacy:

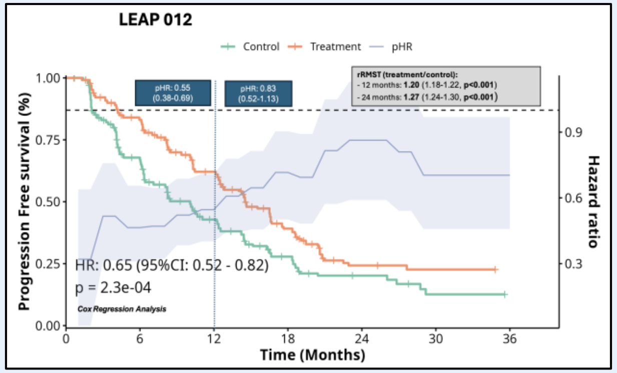
- RMST (dRMST and rRMST): at 12, 24, 36mo and **>**10% patients at risk

- pHR : at 12 months

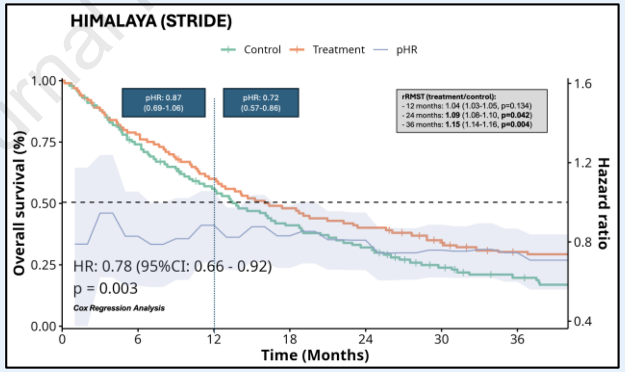
NPH patterns

- Diminishing Effects

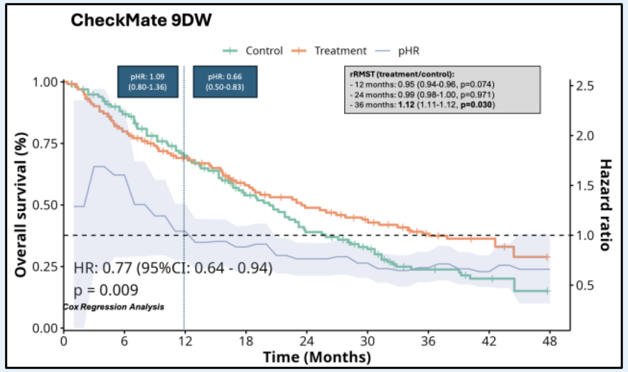




- Delayed Effects



- Crossing Hazards



- **NPH is observed in ~20% of pivotal phase III trials in HCC.**
- **In the presence of NPH, mature data (adequate follow-up and events) are required to establish trial positivity. Early stopping rules are discouraged.**
- **Specific statistical tools (MaxCombo test, RMST and pHR) are needed in NPH settings.**
- **Of 4 pivotal positive RCTs with NPH, 3 confirmed the results, whereas 1 lost significance whith adequate, robust methods were applied.**

Mauro E, et al. J Hepatol 2025

Title: Strategies to address non-proportional hazards between survival curves - Lessons from phase III trials in hepatocellular carcinoma

Short Title: HCC Trials and Non-proportional Hazards.

Authors: Ezequiel Mauro[1], Tiago de Castro[2], Marcus Zeitlhoefler[2], Allan Hackshaw[3], MinJae Lee[4], Tim Meyer[3], Amit G. Singal[5], Josep M. Llovet[1,2,6]

1. Liver Cancer Translational Research Group - Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), Liver Unit-Hospital Clínic, Universitat de Barcelona, Barcelona, Catalonia, Spain.
2. Mount Sinai Liver Cancer Program, Division of Liver Diseases, Department of Medicine, Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA
3. Royal Free Hospital and UCL Cancer Institute, University College London, London, UK.
4. Peter O'Donnell Jr. School of Public Health - Harold C. Simmons Comprehensive Cancer Center, University of Texas Southwestern, Dallas, TX, USA.
5. Department of Internal Medicine, UT Southwestern Medical Center, Dallas, TX, USA.
6. Institució Catalana de Recerca i Estudis Avançats, Barcelona, Catalonia, Spain.

Corresponding Author:

Josep M Llovet, MD, PhD
Liver Cancer Translational Research Group - Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS) - Liver Unit, Hospital Clínic Barcelona, Universitat de Barcelona.
Rosselló 153 08036 Barcelona Catalonia, Spain
E-mail: jmllovet@clinic.cat

ORCIDs:

Ezequiel Mauro: https://orcid.org/0000-0002-0757-7676

Tiago de Castro: https://orcid.org/0009-0005-8660-0983

Marcus Zeitlhoefler: https://orcid.org/0000-0003-4900-0290

Allan Hackshaw: https://orcid.org/0000-0002-5570-5070

MinJae Lee: https://orcid.org/0000-0002-4329-506X

Tim Meyer: https://orcid.org/0000-0003-0782-8647

Amit Singal: https://orcid.org/0000-0002-1172-3971

Josep M Llovet: https://orcid.org/0000-0003-0547-2667

Keywords: *hepatocellular carcinoma, non-proportional hazards, immunotherapy, clinical trials, MaxCombo test, restricted mean survival time, proportional hazards, piecewise hazard ratio.*

List of Abbreviations

HCC, hepatocellular carcinoma; OS, overall survival; ICI, immune checkpoint inhibitor; PH, proportional hazards; HR, hazard ratio; NPH, non-proportional hazards; RCT, randomized controlled trials; IA, interim analysis; *RFS, recurrence-free survival; CI, confidence interval; RMST, restricted mean survival time; pHR,* piecewise hazard ratios*; TACE, transarterial chemoembolization; ASCO, American Society of Clinical Oncology; ILCA, International Liver Cancer Association; ESMO, European Society for Medical Oncology; AASLD, American Association for the Study of Liver Diseases; KM, Kaplan-Meier; PFS, progression-free survival; G-T, Grambsch-Therneau; dRMST, difference in restricted mean survival time; rRMST, ratio of restricted mean survival time;*

***Disclosures*:**

- **EM** received travel funding from Roche.
- **TdC** received speaker honoraria from AstraZeneca and BMS, and travel support from MSD.
- **MZ** no disclosures.
- **AH** no disclosures.
- **ML** no disclosures.

- **TM** received institutional grants from the MSD, Boston Scientific, and Bayer. He received personal consulting fees from Signant Health, Roche, AstraZeneca, Guerbet, Geneos, Ipsen and Grey Wolf.
- **AS** has served as a consultant or on advisory boards for Genentech, AstraZeneca, Bayer, Eisai, Exelixis, Merck, Elevar, Boston Scientific, Sirtex, HistoSonics, FujiFilm Medical Sciences, Exact Sciences, HelioGenomics, Roche, Abbott, Glycotest, Universal Dx, DELFI, and IMCare.
- **JML** is receiving research support from Bayer HealthCare Pharmaceuticals, Eisai Inc, and consulting fees from Eisai Inc, Merck, Roche, Genentech, Glycotest, AstraZeneca, Bayer HealthCare Pharmaceuticals, Bristol-Myers Squibb, AbbVie, Inc, Sanofi Aventis

Acknowledgments:

- **EM** was supported by Andrew K. Burroughs Short-Term Training Fellowship 2021 from EILF-EASL and Reto AECC 70% Supervivencia: RETOS245779LLOV).
- **TdC** was supported by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation; 549129192).
- **MZ** was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation; 531006414) and Schweizer Nationalfond (SNSF, Swiss National Science Foundation; P500PM_222183).
- **TM** is supported by the NIHR Royal Free Clinical Research Facility and NIHR UCLH Biomedical Research Centre
- **JML** was supported by grants from the NIH (R01-CA273932-01, R01DK56621, and R01DK128289), European Commission (Horizon Europe-Mission Cancer, THRIVE, Ref. 101136622), by an Accelerator Award from Cancer Research UK, Fondazione per la Ricerca sul Cancro (AIRC) and Fundación Científica de la Asociación Española Contra el Cáncer (FAECC) (HUNTER, Ref. C9380/A26813), by the Samuel Waxman Cancer Research Foundation, the Spanish National Health Institute (Project PID2022-139365OB-I00, funded by MICIU/AEI/10.13039/501100011033 and FEDER); the Asociación Española Contra el Cáncer (Proyectos Generales: PRYGN223117LLOV; Reto AECC 70% Supervivencia: RETOS245779LLOV), the Generalitat de Catalunya (AGAUR, 2021-SGR 01347), the “la Caixa” Banking Foundation and the Acadèmia de Ciències Mèdiques i de la Salut de Catalunya i Balears.

Data availability statement: Data supporting the findings of this study are available upon request from the corresponding author (JML).

CRediT Authorship contributions:

EM: Conceptualization - Lead, Data curation - Lead, Formal Analysis - Lead, Methodology - Lead, Writing: Original Draft - Lead.

TdC: Conceptualization - Equal, Data curation - Supporting, Formal Analysis - Equal, Methodology - Equal, Writing: Original Draft - Supporting.

MZ: Conceptualization - Equal, Data curation - Supporting, Formal Analysis - Equal, Methodology - Equal, Writing: Original Draft - Supporting.

AH: Conceptualization - Supporting, Data curation - Equal, Formal Analysis - Supporting, Methodology - Supporting, Writing: Original Draft - Supporting.

ML: Conceptualization–supporting, data curation–equal, formal analysis–supporting, methodology–supporting, writing: original draft–supporting.

TM: Conceptualization - Supporting, Data curation - Equal, Formal Analysis - Equal, Methodology - Supporting, Writing: Original Draft - Supporting.

AS: Conceptualization - Supporting, Data curation - Equal, Formal Analysis - Equal, Methodology - Supporting, Writing: Original Draft - Supporting.

JML: Conceptualization - Lead, Data curation - Lead, Formal Analysis - Lead, Methodology - Lead, Writing: Original Draft - Lead.

Impact and Implications

Non-proportional hazards (NPH) impact phase III RCTs in hepatocellular carcinoma (HCC), particularly in immunotherapy trials, potentially causing discrepancies between the interim and final analyses. In fact, halting trials at the interim analysis can be premature when NPH is present. Thus, we propose a framework to ensure study maturity based on follow-up duration and event accruals to optimize interim analyses in the presence of NPH. Whenever NPH is identified, distinct statistical tools should be used to assess reliable differences between arms (MaxCombo) and to assess the effect size [restricted mean survival time (RMST) and piecewise hazard ratios (pHR)] for regulatory decisions and clinical guidance. Implementing these strategies can improve trial design, and better support decision-making for HCC management.

Highlights:

- NPH was present in 4/20 (20%) Phase III trials in HCC, all involving immunotherapies, and three patterns were identified: 1) diminishing effects, 2) delayed effects, and 3) crossing hazards.
- NPH caused discrepancies in IMbrave050’s interim and subsequent efficacy analyses, thus explaining the distinct outcomes reported.
- Robust interim analysis in HCC requires a minimum follow-up duration or number of events before prematurely stopping an RCT with NPH.
- Whenever NPH is identified, distinct statistical tools should be used to assess reliable differences between arms (MaxCombo), and assessment of the effect size (RMST, and pHR) for regulatory decisions, and clinical guidance.
- These strategies are proposed to refine the trial design and enhance treatment decisions in HCC.

ABSTRACT

Background and Aims: Non-proportional hazards (NPH) can cause discrepancies between interim (IA) and final analyses (FA) in randomized clinical trials (RCT) in hepatocellular carcinoma (HCC). We analyzed the impact of NPH in pivotal HCC trials and proposed strategies for a robust analysis.

Methods: Pivotal phase III HCC RCTs (2008-2024) were selected. The Grambsch-Therneau test assessed proportional hazards. For NPH, we proposed an optimal IA timing (twice the estimated median of primary endpoint in the control group) or event size (≥60%). MaxCombo test, restricted mean survival time (rRMST) and piecewise HR (pHR) were used in the NPH scenario.

Results: NPH was present in 4/20 (20%) phase III trials in HCC, all involving immunotherapies, and displayed 3 patterns: 1) diminishing effects, 2) delayed effects, and 3) crossing hazards. Two RCTs (IMbrave050, LEAP-012) reported positive IA results with diminishing effects. In IMbrave050, discrepancies were observed when comparing IA and FA using MaxCombo analysis (p=0.02 and p=0.33, respectively), rRMST at 12 and 36 months [1.11 (p<0.001) and 1.08 (p=0.08), respectively] and pHR prior/after 12 months [0.59 (95%CI 0.43-0.73) and 1.12 (95%CI 0.88-1.37)]. In LEAP-012, consistently results at both 12 and 24 months were observed (MaxCombo test p<0.001) and rRMST [1.20 (p<0.001) and 1.27 (p<0.001)]. HIMALAYA and Checkmate 9DW reported positive results, which were confirmed by MaxCombo test. HIMALAYA showed delayed effects [rRMST at 12 and 36 months: 1.04 (p=0.13) and 1.15 (p=0.004)], while CheckMate 9DW displayed crossing hazards [rRMST at 12 and 36 months: 0.95 (p=0.07) and 1.12 (p=0.03)].

Conclusion: NPH caused discrepancies in IMbrave050's interim and subsequent efficacy analyses. Robust IA requires a minimum follow-up duration or number of events before prematurely stopping an RCT with NPH.

Introduction

Hepatocellular carcinoma (HCC) is a major cause of global cancer mortality, significantly affecting overall survival (OS) due to poor adherence to surveillance programs and frequent late-stage diagnoses.[1,2] Advances in immune checkpoint inhibitor (ICI)-based therapies have reshaped the therapeutic landscape, particularly in advanced stages of disease, and have even shown promising results at earlier stages HCC.[3–5]

Two well-known considerations in clinical trials have been the proportional hazards (PH) assumption when time-to-event endpoints are used, and the timing and interpretation of interim analyses (IA). The PH assumption suggests that the hazard ratio (HR) remains constant over time.[6] Deviations from this assumption, referred to as non-proportional hazards (NPH), can influence the interpretation of statistical significance and the perceived magnitude of the treatment effect.[7] Patterns of NPH in ICI-based oncology trials include diminishing effects over time, delayed treatment effects, and crossing hazards.[8] This issue has gained considerable attention in the HCC field following the IMbrave050 phase III randomized controlled trial (RCT), in which the interim analysis (IA) demonstrated a positive outcome for recurrence-free survival (RFS) (HR: 0.72; 95% CI: 0.53-0.98) with adjuvant atezolizumab-bevacizumab versus surveillance after resection in high-risk HCC.[9] These results led to its adoption in American Association for the Study of Liver Diseases (AASLD) clinical practice guidelines.[10] However, subsequent analysis with more mature RFS data failed to confirm this benefit (HR: 0.90; 95% CI: 0.72-1.12)[11], resulting in the withdrawal of this recommendation from the major guidelines. This inconsistency in results may be related to the presence of NPH, which poses challenges for determining the optimal timing and robustness of IA.[12]

To address NPH, advanced statistical tools, such as the MaxCombo test[13] and restricted mean survival time (RMST) analysis[14], offer robust alternatives to assess statistical significance and magnitude of benefit compared with traditional log-rank and Cox and HR models.[15] The MaxCombo test combines several weighted log-rank tests, enhancing detection of treatment effects across different NPH types by providing greater statistical power.[8,13,15] RMST measures the average survival up to a specified point, akin to a patient's life expectancy within that period.[16] This method offers a clear, intuitive summary of survival, relying less on the assumption of proportional hazards and providing a more interpretable estimate of treatment benefits when HR fluctuate over time.[17] Another complementary method to address NPH is the use of piecewise hazard ratios (pHR), which focus on estimating the magnitude of treatment effects within specific time intervals, allowing for a more detailed evaluation of how treatment efficacy evolves over time.[13]

Given the potential implications of NPH in the interpretation of pivotal phase III trials, we aimed to assess the prevalence and potential impact of NPH. We focused on evaluating the treatment significance and effect magnitude while proposing a robust strategy to ensure rigorous interpretation of interim analyses in HCC.

Methods

Literature Search and Data Sources

A comprehensive literature search was conducted from January 2008 to September 15, 2024, using Medline (Ovid), Embase (Elsevier), CENTRAL, and Web of Science. This investigation focused on phase III pivotal RCTs involving adjuvant therapy for early-stage HCC with sorafenib or ICI, combination therapies of transarterial chemoembolization (TACE) with sorafenib or ICI-based regimens, and systemic treatments (1st and 2nd line) for HCC. Furthermore, a manual review of the reference lists from relevant conference materials, including abstracts and posters, from the American Society of Clinical Oncology (ASCO), International Liver Cancer Association (ILCA), and European Society for Medical Oncology (ESMO) until September 15, 2024, was performed. The search strategy employed both thesaurus terms and keywords in the titles and abstracts. For the propose of trial screening and selection, we adhered to the criteria outlined by the American Association for the Study of Liver Diseases (AASLD) Consensus Conference on Trial Design and Endpoints[18] to establish optimal selection parameters for the target population and control arms, including only trials with a modified Jadad score of ≥8.[19] This was not intended as a formal systematic review, but rather a focused selection of high-quality phase III RCTs with potential clinical impact. Comprehensive details of the search strategy and trial selection are provided in the open labeled **Supplementary Material**.

Data Source

A reverse-engineering method was employed to extract event times and censoring information from the published Kaplan-Meier (KM) survival curves of the selected studies. For the three primary end points assessed in the studies, OS, RFS, or progression-free survival (PFS), curves were digitized utilizing WebPlotDigitizer v.4.6 software[20], available at https://automeris.io. The digitized curves were processed, and the original survival data were reconstructed using the reconstructKM v.0.3.0 package. To ensure data quality, survival curves were regenerated with survival v.3.4-0 and survminer v.0.4.9 packages, and HRs were recalculated.(**Supplementary Material**)

Statistical Methods

The proportionality assumption was tested using the Grambsch-Therneau test (G-T test)[21] and visually examined using Schoenfeld residual plots. In phase III RCTs in which NPH was detected using the G-T test, we evaluated the efficacy using the MaxCombo test and quantified the effect size through the RMST and pHR specifically assessed at 12 months. The MaxCombo test was implemented using Fleming-Harrington weighted log-rank tests with (ρ, γ) combinations of (0,0), (1,0), and (0,1), allowing sensitivity to early, constant, and late treatment effects.To analyze the significance of the treatment by RMST, we used the last time point at which more than 10% of the patients remained at risk.[22] Furthermore, we estimated the difference in RMST (dRMST: treatment - control, absolute effect in months)[14] and the ratio of RMST (rRMST: treatment/control, relative effect); an rRMST>1 indicates that the treatment group exhibits a longer average survival time compared to the control group, with, for instance, an rRMST of 1.2 suggesting a 20% increase in average survival time in the treatment group; values equal to 1 imply no relative difference in average survival times between groups, while rRMST< 1 indicate a shorter average survival time in the treatment group relative to the control group.[23] Additional summary measures included dRMST or rRMST between the treatment arms at landmark survival estimates at 12, 24, and 36 months (when estimable), which were selected based on their clinical relevance. Differences and ratios of RMST were tested using the asymptotic normal approximation method. For each study, if both tests being compared yielded two-sided p values <0.05 (nominally significant) or ≥0.05 (not nominally significant), the results were deemed concordant; otherwise, the results were classified as discordant. All analyses utilized a significance level of α=0.05 and were conducted using R v.4.2.1.

Optimal Timing for Interim Analyses in RCTs with NPH

For trials exhibiting NPH, we evaluated IA timing using previously suggested cut-offs: either achieving a minimum follow-up (FU) duration of at least twice the estimated median of the primary endpoint for the control group or an occurrence of ≥60% of events out of the total number of patients randomized.[8,12,24] When either criterion was met, efficacy was assessed using the MaxCombo test, and effect size was quantified through RMST and pHR at predefined time horizons.(**Figure 1**)

Results

Study Selection and Characteristics

Twenty pivotal phase III RCTs encompassing various HCC treatment settings were included in our analysis. The main characteristics of the included phase III RCTs are summarized in **Table 1**. In the adjuvant setting following surgery or ablation, we included principal studies with RFS as the primary endpoint, specifically the STORM[25] and IMbrave050 trials. For IMbrave050, both the initial positive results from IA[9] and subsequent negative findings from the more mature final analysis[11] were evaluated to capture the evolution of treatment efficacy over time. Regarding the combinations of TACE with sorafenib or ICI-based treatments, we included trials such as the TACE 2 (TACE plus sorafenib)[26], the EMERALD-1[4], and the LEAP-012 trials[5]. In these studies, the primary endpoint was PFS. In the first-line systemic treatment setting, we incorporated trials demonstrating superiority in OS, including SHARP[27], IMbrave150[28], HIMALAYA (STRIDE arm)[29,30], CheckMate 9DW[31], and CARES 310[32] trials. Additionally, non-inferiority trials were included, such as REFLECT[33], HIMALAYA (durvalumab monotherapy arm)[29], and RATIONALE 301[34], as well as negative ICI-based trials, specifically CheckMate-459[35], COSMIC-312[36] and LEAP-002[37]. Finally, for second-line systemic treatment, four positive trials were included, RESORCE[38], CELESTIAL[39], REACH-2[40], and KEYNOTE-394[41], and one negative trial, KEYNOTE-240[42].

Assessment of Non-proportional Hazards

We assessed the proportionality assumption using the G-T test among the 20 phase III RCTs analyzed, four (20%) of which exhibited evidence of NPH, all of which assessed immunotherapies (**Figure 2**). The IMbrave050 and LEAP-012 trials showed diminishing treatment effects over time (**Figure 3A-B**). IMbrave050 displayed a progressive reduction in the observed benefits as the FU advanced for RFS at both IA and FA (**Figure 3A**). In contrast, the HIMALAYA trial demonstrated a delayed treatment effect for OS at FA, characterized by overlapping survival curves during the initial ~6 months period, followed by a subsequent separation, favoring the treatment arm (**Figure 3C**). The CheckMate 9DW trial displayed a crossover of hazards for OS between treatment arms, where the survival curves intersected, reflecting a reversal of the treatment effect at a specific time point (**Figure 3D**).

Evaluation of Efficacy in the Presence of NPH

In phase III RCTs with NPH identified by the G-T test, treatment efficacy was evaluated using the log-rank test, MaxCombo test, and RMST analysis (calculated up to the last time point with more than 10% of patients at risk in the original KM curve[22]). Interestingly, despite the limitations of the log-rank test in the presence of NPH, all RCTs demonstrated consistent statistical significance across the log-rank test, the MaxCombo test, and RMST (**Table 2**).

In the IMbrave050 trial, the interim analysis, conducted with a median FU of 17.5 months and a limited number of events (33% in the treatment arm and 40% in the surveillance arm), demonstrated statistically significant differences in RFS across all statistical methods employed: log-rank test (p=0.012), MaxCombo test (p=0.018), and RMST (p=0.026). Nevertheless, upon the second final analysis, conducted at a median FU of 35.1 months with a 49% event rate, the statistical significance of the therapeutic strategy effect did not maintain significance, including the MaxCombo test (p=0.326) and RMST (p=0.08).

Magnitude of Effect and Proposed Interim Analysis Strategies in the Presence of NPH

<u>NPH with diminishing effects</u>

IMbrave050

The IMbrave050 trial is a prominent example of diminishing treatment effects over time as a pattern of NPH, with maximal separation of curves at 12 months (RFS 78% vs. 65%) and convergence by 24 months, highlighting the limitations of HR interpretation alone in trials with non-proportional hazards. The key characteristics of this study are summarized in **Table 1**. The primary endpoint was RFS assessed by an independent review with a hierarchical interim analysis design that first evaluated RFS, followed by OS, if RFS was positive. The first interim analysis for RFS was planned at 236 events, which represented 73% of the 323 events estimated for the final analysis, with a significance threshold of p≤ 0.019 and a target HR of 0.73 as a boundary for a positive trial outcome. The protocol-defined expected median RFS in the control arm was 20 months. At the prespecified analysis, the median FU was 17.4 months. A total of 243 events were reported (33% in the treatment arm and 40% in the surveillance arm), with an HR of 0.72 (95% CI: 0.53–0.98) and a log-rank p-value of 0.012. Based on the O'Brien & Fleming stopping rules for superiority, this study was deemed positive. The dRMST at 12 and 24 months were 1.12 months (95% CI: 1.03-1.21, p<0.001) and 1.85 months (95% CI: 1.74-1.98, p=0.004), respectively. In relative terms, the rRMST at 12 and 24 months was 1.11 (95% CI: 1.10-1.13, p<0.001) and 1.10 (95% CI: 1.09-1.12, p=0.004), respectively, indicating a 10-11% increases in mean RFS.

A subsequent analysis was further conducted with more mature data including a median FU of 35.1 months and an overall 49% event rate, resulting in a HR of 0.90 (95% CI: 0.72-1.12).[11] At 36 months, the dRMST was 1.84 months (95% CI: 1.67-2.02; p=0.08), while the rRMST was 1.08 (95% CI: 1.06-1.09; p=0.08). Additionally, the pHR assessed prior and after 12 months was 0.59 (95% CI: 0.43-0.73) and 1.12 (95% CI: 0.88-1.37), respectively.(**Figure 3A**)

In this study, the proposed analysis regarding the significance and magnitude of benefit in cases of NPH was demonstrably inapplicable at the IA and remains so in the subsequent analysis, as neither the ≥60% event threshold nor a FU period of at least 40 months (double the expected median RFS of 20 months for the control arm) was achieved.(**Figure 1**) Nonetheless, it is evident that the diminishing effect pattern driving the NPH is not reversible at this stage and is unlikely to alter the negative outcome, even when the assumptions required for applying our proposed approach are eventually met.

LEAP-012

The main characteristics of LEAP-012 are summarized in **Table 1**. PFS and OS were the primary endpoints analyzed using a hierarchically tested primary endpoint strategy, where PFS was tested first, and statistical significance was allocated to OS only if PFS met the predefined threshold. The IA served as the final analysis for PFS, with an initial one-sided alpha of 0.025 allocated to PFS and a target HR of 0.68. The median FU at IA was 25.6 months, more than twice the expected median of 8 months for the control group, with events occurring in 56% and 63% of the patients in the treatment and control arms, respectively. The HR was 0.66 (95% CI: 0.51-0.84, p<0.001).[5]

The dRMST at 12 and 24 months were 1.63 months (95% CI: 1.56-1.70, p<0.001) and 3.07 months (95% CI: 3.03-3.12, p<0.001), respectively. Correspondingly, the rRMST at 12 and 24 months were 1.20 (95% CI: 1.18-1.22, p<0.001) and 1.27 (95% CI: 1.24-1.30, p<0.001), respectively, indicating 20% and 27% increases in mean PFS at these time points. Additionally, the pHR assessed before and after 12 months was 0.55 (95% CI: 0.38-0.69) and 0.83 (95% CI: 0.52-1.13), respectively.(**Figure 3B**)

The proposed analysis of the significance and magnitude of benefit in NPH at IA is deemed applicable (≥60% of events and a FU period exceeding 16 months, twice the expected median of 8 months for the control group). This analysis demonstrated a significant benefit of the therapeutic strategy (MaxCombo test of p<0.001), with the magnitude of the effect being adequately represented by RMST and pHR demonstrating a dominant effect. (**Figure 1**)

NPH with delayed effect

HIMALAYA

HIMALAYA exemplifies NPH because of delayed treatment effects or the presence of long-term survivors. The current analysis is based on mature (82% events rate) after a median FU of 61.2 months, yielding a HR: 0.76 (95% CI: 0.65-0.89, p<0.001).[43]

Regarding the magnitude of the effect over time, the dRMST at 12, 24, and 36 months were 0.38 months (95% CI: 0.36-0.39, p=0.134); 1.22 months (95% CI: 1.21-1.23, p=0.042); and 2.58 months (95% CI: 2.48-2.67, p=0.004), respectively. Correspondingly, the rRMST at these time points were 1.04 (95% CI: 1.03-1.05, p=0.134); 1.09 (95% CI: 1.08-1.10, p=0.042); and 1.15 (95% CI: 1.14-1.16, p=0.004), respectively. It is noteworthy that at 12 months, RMST did not exhibit significant differences. However, at 24 and 36 months, significant increases in mean OS of 9% and 15%, respectively, were observed, demonstrating the delayed benefit of ICI therapy in this RCT. Additionally, the pHR at 12 months was 0.87 (0.69-1.06) and 0.72 (0.57-0.86) thereafter. (**Figure 3C**)

In this case, our proposed analysis for NPH is applicable (≥60% of events and a FU period exceeding 20 months, twice the protocol-defined expected median for the control group) and demonstrated a significant benefit of the therapeutic strategy (MaxCombo test of p<0.001). The magnitude of the benefit was adequately represented through RMST and pHR analyses, specifically after 12 months, highlighting its predominantly late effect. (**Figure 1**)

Recently, extended FU data were published, with median FU of 62.5 (59.5-64.8) months and 59.9 (58.3–61.5) months for the STRIDE and sorafenib arms, respectively.[30] At the 5-year data cut-off, 309 patients (78.6% OS data maturity) in the STRIDE arm and 332 patients (85.3%) in the sorafenib arm had died. In this mature dataset, the magnitude of the effect was further supported by long-term RMST analyses conducted at a timepoint where over 20% of patients in the experimental arm remained at risk, showing a dRMST of 4.35 months at 48 months (95% CI: 3.93-4.76, p<0.001) and rRMSTs of 1.32 (95% CI: 1.32-1.33, p<0.001).

NPH with crossing hazards

CheckMate 9DW

CheckMate 9DW exemplifies NPH, which is characterized by crossing the HRs. In the sole and interim analysis available, with an FU of 35.2 months, OS events occurred in 58% and 68% of the patients in the treatment and control arms, respectively. The study reported an HR of 0.79 (95% CI: 0.65-0.96, p=0.018), favoring the treatment group.[31] Notably, the KM curves revealed a clear change in the direction of treatment efficacy for nivolumab plus ipilimumab versus lenvatinib (85%) or sorafenib (15%) from 12 months onward. The dRMST at 12, 24, and 36 months were -0.46 months (95% CI: -0.51- -0.41, p=0.074), -0.02 months (95% CI: -0.09-0.05, p=0.971), and 2.73 months (95% CI: 2.50-2.96, p=0.03), respectively. Correspondingly, the rRMST at these time points were 0.95 (95% CI: 0.94-0.96, p=0.074), 0.99 (95% CI: 0.98-1.00, p=0.971), and 1.12 (95% CI: 1.11-1.12, p=0.030), respectively. These results indicate a significant 12% increase in the mean OS at 36 months, but no benefit at 12 or 24 months. Furthermore, the pHR was1.09 (95% CI: 0.80-1.36) prior to 12 months and 0.66 (95% CI: 0.50-0.83) thereafter, respectively, demonstrating a significant late effect. (**Figure 3D**)

In this study, our proposed analysis for NPH is valid (≥60% of estimated events and a FU period exceeding 28 months, twice the protocol-defined expected median for the control group) and demonstrated a significant benefit of the therapeutic strategy (MaxCombo test, p=0.003). The magnitude of the benefit was appropriately quantified through RMST and pHR analyses, demonstrating a non-significant trend toward greater efficacy of lenvatinib/sorafenib over nivolumab-ipilimumab during the initial 12 months, followed by an inverse significant effect in favor of nivolumab-ipilimumab, particularly evident after 24 months, indicating a shift in the direction of treatment benefit.

Discussion

The introduction of immunotherapies has transformed the therapeutic landscape of HCC, extending its impact beyond advanced stages.[3–5] At early stages of the disease, the first positive RCT in 40 years, IMbrave050, paved the way to improve the outcome of patients undergoing resection/local ablation, when the disease can be cured. The proposed regimen, atezolizumab plus bevacizumab improved RFS vs surveillance[9], and thus it was adopted by AASLD clinical practice guidelines[10]. However, a subsequent efficacy analysis with longer FU showed a reduction in the magnitude of clinical benefit, and thus the trial did no longer met the primary endpoint[11]. The inconsistency between the interim and subsequent efficacy analyses prompted us to conduct a thorough analysis to understand the reason for such an important switch in the outcome.

The current study identified that inconsistencies in the outcome between the interim and subsequent analyses in the IMbrave050 trial were directly attributable to the presence of NPH, calling for a broader exploration of pivotal practice-changing RCT reported in HCC. Importantly, our aim was not to reinterpret or challenge the conclusions of individual trials, but to illustrate how statistical significance and effect size may differ under NPH conditions. To this end, we propose a framework based on data maturity and appropriate analytical tools (MaxCombo, RMST, and pHR) to improve the reliability of efficacy assessments.

Similar to other cancers, our findings demonstrated that immunotherapies for HCC are associated with NPH issues in 20% of RCTs (IMbrave050[11], LEAP-012[5], HIMALAYA[29,30] and CheckMate 9DW[31]). This phenomenon highlights the necessity of employing additional statistical tools to accurately evaluate treatment efficacy and appropriately describe the magnitude of effect over time under different patterns of NPH.[7,13,15,44]

To address NPH issues in these trials, we adopted an integrative approach (**Figure 1**) that highlights both the need for more mature data (either by ensuring enough FU or number of events)[8,12,24] and the application of sensitivity analyses using statistical tools for these scenarios[13–15]. In this regard, we defined mature data as either the double of the median of the estimated outcome of the primary endpoint for the control arm, or a number of events providing robust information (i.e., ≥60% of events out of the total number of patients randomized). These criteria are essential to ensure that the analysis captures the full impact of the treatment, particularly under the influence of distinct NPH patterns, in which the timing of hazard changes can significantly affect the observed outcomes.[8,12,24] In settings with low event rates, the ≥60% threshold may not be achievable; in such cases, reaching at least double the median FU of the control group remains a valid proxy for data maturity under NPH.

In terms of statistical tools, aside of the G-T test to define the presence of NPH, we use the MaxCombo test -instead of the log-rank test- and the RMST analyses instead of the HR[15,45], since the latter fails to properly capture the dynamics of treatment benefits under NPH conditions.[7] As emphasized by the ICH E9 (R1) addendum[12,46], RMST offers a robust alternative, with dRMST quantifying the absolute benefit in time units and rRMST expressing relative benefit.[16,47,48] Additionally, we used pHRs, which allows a dynamic representation of how treatment benefits evolve over time, offering a nuanced perspective that is particularly relevant in scenarios with varying hazard patterns.[13] However, it is important to highlight that MaxCombo and RMST improve analysis under NPH but are not reliable without adequate maturity. Without sufficient FU or events, even these methods may yield misleading results.

Using this approach, we defined 4 cases of NPH, 2 of which had a diminishing effect pattern (IMbrave050 and LEAP-012), one with a delayed effect pattern (HIMALAYA trial) and one with a crossover effect pattern (CheckMate 9DW). In all studies, assessment of Max Combo test matched the same effect as long-rank test. Nonetheless, three trials were stopped at the interim analysis (IMbrave050, LEAP-012 and CheckMate 9DW), and one at the final analysis (HIMALAYA). When exploring the robustness of information, only IMbrave050 stands-out as immature for assessing benefit effects in the setting of NPH. Other trials either showed maturity in terms of median FU (LEAP012) or >60% of events (HIMALAYA and Checkmate 9DW). Thus, as a whole, the only practice-changing pivotal trial that has been affected by immature interim analysis has been IMbrave050. This has implications for the future of trial design in HCC, as exploring the status of NPH should be mandatory in all interim analyses. In the case of immature data, the study is not recommended to be stopped, even in cases where the long-rank test hits the primary endpoint, and thus, stopping rules, such as those of O'Brien-Fleming, would not apply in these scenarios.

Therefore, our recommendation is that for cases of NPH at IA, evaluating both the significance and magnitude of treatment benefit requires careful consideration of key parameters, including the minimum FU duration (at least twice the estimated median time of the primary endpoint for the control group) or sufficient events (≥60% of events out of the total number of patients randomized). In our analysis, all four trials with confirmed NPH showed consistent significance across log-rank, MaxCombo, and RMST tests. Notably, MaxCombo consistently yielded smaller p-values, suggesting greater sensitivity in detecting time-varying effects. This observation aligns with prior meta-analyses supporting MaxCombo as the preferred test under NPH conditions.[15] In addition, the application of RMST and pHR analyses allowed us to accurately quantify the magnitude and temporal dynamics of this effect.

The PH assumption is rarely reported in RCTs, even when ICI regimens are tested.[6]. Recent analysis of 66 RCT relied on KM curves and the log-rank test (88%), while only 12%

checked for NPH.[49] In NPH scenarios, the HR depends on the censoring distribution and thus loses clear interpretability, while log-rank tests may fail to detect genuine effects. In support of this, a study found that out of 18 trials with inconclusive long-rank tests, 9 found differences when by applying NPH-specific methods.[50] A similar finding was observed in a RCT assessing acute lymphoblastic leukemia therapies.[51,52] These insights align with IMbrave050 and underscore the importance of robust FU, sufficient event percentages, and appropriate statistical methods under NPH. Consequently, many recent ICI-based RCTs now incorporate RMST or pHR as sensitivity analyses.[31,53,54]

Anticipating NPH is essential given its frequent association with ICI yet predicting its pattern or whether it arises from treatment effects or population heterogeneity remains challenging.[8,12] Notably, both trials with a delayed effect (HIMALAYA and CheckMate 9DW) included CTLA-4 blockade, suggesting that similar delayed effects may be expected in studies with this mechanism of action. Proactive strategies that account for NPH, such as incorporating sensitivity analyses, enable adjustments in sample size and IA timing, ultimately strengthening study power and reliability.[8,12] Accounting for NPH from the design phase is key not only to ensure statistical validity but also to optimize sample size, FU duration, and IA schedules. The approaches proposed here can help avoid premature decisions based on incomplete data, as evidenced in the reanalysis of IMbrave050, and further emphasize the critical role of robust trial designs tailored to NPH scenarios.

Addressing the potential challenges posed by NPH in phase III RCTs, particularly in ICI-based regimens, may warrant a closer look at how clinical benefits are currently evaluated. While the methodological approach to NPH is consistent across settings, the clinical interpretation of benefit depends on disease stage and endpoint. For instance, the European Society for Medical Oncology Magnitude of Clinical Benefit Scale (ESMO-MCBS) grading system[55] does not incorporate statistical methods tailored to NPH, relying instead on median-based differences, Cox-derived hazard ratios, or landmark timepoints (e.g., 2-3 years) to define meaningful survival gains. In its latest version (v2.0), ESMO-MCBS has introduced stricter criteria for tail-of-the-curve credit, requiring that at least 20% of patients in the experimental arm remain at risk at the evaluation timepoint to ensure robustness.[56] While such timepoints may partially capture delayed effects, landmark analyses exclude censored patients, lack formal statistical inference, and are sensitive to the number at risk, limitations that reduce their reliability under NPH.[57,58] In contrast, RMST offers a more robust and interpretable metric in this context.

Adapting these guidelines to better account for NPH scenarios could be worthwhile, especially given that ESMO-MCBS scores have been found to correlate significantly with favorable health technology assessment (HTA) outcomes, potentially increasing the likelihood of positive HTA decisions and reducing the time between marketing authorization and HTA recommendations.[59]

Our study has several limitations. The absence of individual patient-level data in RCTs required the reconstruction of KM curves from published data, which may introduce estimation inaccuracies. Our re-analysis, however, were consistent with the original findings, providing validation for the methodology employed. Importantly, our proposal based on defining maturity as either twice the median of the estimated outcome for the control arm or the occurrence of ≥60% of events among randomized patients, is not intended as a definitive standard, but rather as a structured, literature-based framework that should be prospectively validated in future phase III trials with time-varying treatment effects. We propose specific statistical tests for NPH scenarios, the MaxCombo test, although useful, may disproportionately weigh events at specific intervals, potentially affecting interpretation in the presence of complex hazard patterns. Similarly, since the MaxCombo and log-rank tests rely on different assumptions, their p-values are not directly comparable. However, previous simulations have shown that MaxCombo test performs better in scenarios with time-varying treatment effects.[13] Additionally, RMST is constrained to a predefined time horizon and potentially fails to capture long-term effects or delayed treatment benefits, whereas pHR can be challenging to interpret when treatment effects vary significantly over time, particularly in cases of crossing hazards or delayed effects. These considerations underscore the need for the careful interpretation of these statistical tools in the clinical context of NPH.[60]

In conclusion, the presence of NPH in ICI-based trials requires a paradigm shift in the approach to statistical analysis, particularly at interim analysis bounded by stopping rules. Our proposed strategy focuses on adapting statistical methodologies to address the challenges posed by NPH, ensuring that the IA provides robust and consistent data. By incorporating advanced techniques such as MaxCombo, RMST, and pHR, this approach enables a more accurate detection of treatment effects over time and minimizes the risk of premature or misleading conclusions based on insufficient FU or event maturity. This approach ensures that trial outcomes reflect the true clinical impact of novel therapies and optimizes decision-making during the trial process. Furthermore, this strategy improves the communication of treatment benefits to the scientific community, healthcare professionals, and key stakeholders, ultimately fostering informed decisions that directly benefit HCC patients.

References:

[1] Singal AG, Kanwal F, Llovet JM. Global trends in hepatocellular carcinoma epidemiology: implications for screening, prevention and therapy. Nat Rev Clin Oncol 2023;20:864–84. https://doi.org/10.1038/S41571-023-00825-3.

[2] Llovet JM, Kelley RK, Villanueva A, et al. Hepatocellular carcinoma. Nat Rev Dis Primers 2021;7. https://doi.org/10.1038/S41572-020-00240-3.

[3] Llovet JM, Pinyol R, Yarchoan M, et al. Adjuvant and neoadjuvant immunotherapies in hepatocellular carcinoma. Nat Rev Clin Oncol 2024;21:294–311. https://doi.org/10.1038/S41571-024-00868-0.

[4] Sangro B, Kudo M, Erinjeri JP, et al. Durvalumab with or without bevacizumab with transarterial chemoembolisation in hepatocellular carcinoma (EMERALD-1): a multiregional, randomised, double-blind, placebo-controlled, phase 3 study. Lancet 2025. https://doi.org/10.1016/S0140-6736(24)02551-0.

[5] Kudo M, Ren Z, Guo Y, et al. Transarterial chemoembolisation combined with lenvatinib plus pembrolizumab versus dual placebo for unresectable, non-metastatic hepatocellular carcinoma (LEAP-012): a multicentre, randomised, double-blind, phase 3 study. Lancet 2025. https://doi.org/10.1016/S0140-6736(24)02575-3.

[6] Rahman R, Fell G, Ventz S, et al. Deviation from the Proportional Hazards Assumption in Randomized Phase 3 Clinical Trials in Oncology: Prevalence, Associated Factors, and Implications. Clin Cancer Res 2019;25:6339–45. https://doi.org/10.1158/1078-0432.CCR-18-3999.

[7] Uno H, Claggett B, Tian L, et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. J Clin Oncol 2014;32:2380–5. https://doi.org/10.1200/JCO.2014.55.2208.

[8] Ananthakrishnan R, Green S, Previtali A, et al. Critical review of oncology clinical trial design under non-proportional hazards. Crit Rev Oncol Hematol 2021;162. https://doi.org/10.1016/j.critrevonc.2021.103350.

[9] Qin S, Chen M, Cheng AL, et al. Atezolizumab plus bevacizumab versus active surveillance in patients with resected or ablated high-risk hepatocellular carcinoma (IMbrave050): a randomised, open-label, multicentre, phase 3 trial. Lancet 2023;402:1835–47. https://doi.org/10.1016/S0140-6736(23)01796-8.

[10] Singal AG, Llovet JM, Yarchoan M, et al. AASLD Practice Guidance on prevention, diagnosis, and treatment of hepatocellular carcinoma. Hepatology 2023;78:1922–65. https://doi.org/10.1097/HEP.0000000000000466.

[11] Yopp A, Kudo M, Chen M, et al. LBA39 Updated efficacy and safety data from IMbrave050: Phase III study of adjuvant atezolizumab (atezo) + bevacizumab (bev) vs active surveillance in patients (pts) with resected or ablated high-risk hepatocellular carcinoma (HCC). Annals of Oncology 2024;35:S1230. https://doi.org/10.1016/J.ANNONC.2024.08.2279.

[12] Roychoudhury S, Anderson KM, Ye J, et al. Robust Design and Analysis of Clinical Trials With Nonproportional Hazards: A Straw Man Guidance From a Cross-Pharma Working Group. Stat Biopharm Res 2023;15:280–94. https://doi.org/10.1080/19466315.2021.1874507.

[13] Lin RS, Lin J, Roychoudhury S, et al. Alternative Analysis Methods for Time to Event Endpoints Under Nonproportional Hazards: A Comparative Analysis. Stat Biopharm Res 2020;12:187–98. https://doi.org/10.1080/19466315.2019.1697738.

[14] Royston P, Parmar MKB. Restricted mean survival time: An alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. BMC Med Res Methodol 2013;13:1–15. https://doi.org/10.1186/1471-2288-13-152/FIGURES/3.

[15] Mukhopadhyay P, Ye J, Anderson KM, et al. Log-Rank Test vs MaxCombo and Difference in Restricted Mean Survival Time Tests for Comparing Survival Under Nonproportional Hazards in Immuno-oncology Trials: A Systematic Review and Meta-analysis. JAMA Oncol 2022;8. https://doi.org/10.1001/jamaoncol.2022.2666.

[16] Dehbi HM, Royston P, Hackshaw A. Life expectancy difference and life expectancy ratio: Two measures of treatment effects in randomised trials with non-proportional hazards. BMJ (Online) 2017;357. https://doi.org/10.1136/bmj.j2250.

[17] Kloecker DE, Davies MJ, Khunti K, et al. Uses and limitations of the restricted mean survival time: Illustrative examples from cardiovascular outcomes and mortality trials in type 2 diabetes. Ann Intern Med 2020;172:541–52. https://doi.org/10.7326/M19-3286.

[18] Llovet JM, Villanueva A, Marrero JA, et al. Trial Design and Endpoints in Hepatocellular Carcinoma: AASLD Consensus Conference. Hepatology 2021;73 Suppl 1:158–91. https://doi.org/10.1002/HEP.31327.

[19] Jadad AR, Moore RA, Carroll D, et al. Assessing the quality of reports of randomized clinical trials: Is blinding necessary? Control Clin Trials 1996;17. https://doi.org/10.1016/0197-2456(95)00134-4.

[20] Author: Ankit Rohatgi, Website: https://automeris.io/WebPlotDigitizer, Version: 4.6. https://automeris.io/WebPlotDigitizer/citation.html

[21] Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika 1994;81:515–26. https://doi.org/10.1093/BIOMET/81.3.515.

[22] Pocock SJ, Clayton TC, Altman DG. Survival plots of time-to-event outcomes in clinical trials: good practice and pitfalls. Lancet 2002;359:1686–9. https://doi.org/10.1016/S0140-6736(02)08594-X.

[23] Trinquart L, Jacot J, Conner SC, et al. Comparison of Treatment Effects Measured by the Hazard Ratio and by the Ratio of Restricted Mean Survival Times in Oncology Randomized Controlled Trials. J Clin Oncol 2016;34:1813–9. https://doi.org/10.1200/JCO.2015.64.2488.

[24] Rufibach K, Grinsted L, Li J, et al. Quantification of follow-up time in oncology clinical trials with a time-to-event endpoint: Asking the right questions. Pharm Stat 2023;22:671–91. https://doi.org/10.1002/PST.2300.

[25] Bruix J, Takayama T, Mazzaferro V, et al. Adjuvant sorafenib for hepatocellular carcinoma after resection or ablation (STORM): A phase 3, randomised, double-blind, placebo-controlled trial. Lancet Oncol 2015;16:1344–54. https://doi.org/10.1016/S1470-2045(15)00198-9.

[26] Meyer T, Fox R, Ma YT, et al. Sorafenib in combination with transarterial chemoembolisation in patients with unresectable hepatocellular carcinoma (TACE 2): a randomised placebo-controlled, double-blind, phase 3 trial. Lancet Gastroenterol Hepatol 2017;2. https://doi.org/10.1016/S2468-1253(17)30156-5.

[27] Llovet JM, Ricci S, Mazzaferro V, et al. Sorafenib in advanced hepatocellular carcinoma. N Engl J Med 2008;359:378–90. https://doi.org/10.1056/NEJMOA0708857.

[28] Cheng AL, Qin S, Ikeda M, et al. Updated efficacy and safety data from IMbrave150: Atezolizumab plus bevacizumab vs. sorafenib for unresectable hepatocellular carcinoma. J Hepatol 2022;76:862–73. https://doi.org/10.1016/J.JHEP.2021.11.030.

[29] Abou-Alfa GK, Lau G, Kudo M, et al. Tremelimumab plus Durvalumab in Unresectable Hepatocellular Carcinoma. NEJM Evidence 2022;1. https://doi.org/10.1056/EVIDOA2100070.

[30] Rimassa L, Chan SL, Sangro B, et al. Five-year overall survival update from the HIMALAYA study of tremelimumab plus durvalumab in unresectable HCC. J Hepatol 2025. https://doi.org/10.1016/J.JHEP.2025.03.033.

[31] **Yau T**, **Galle PR**, Decaens T, et al. Nivolumab plus ipilimumab versus lenvatinib or sorafenib as first-line treatment for unresectable hepatocellular carcinoma (CheckMate 9DW): an open-label, randomised, phase 3 trial. Lancet 2025. https://doi.org/10.1016/S0140-6736(25)00403-9.

[32] Qin S, Chan SL, Gu S, et al. Camrelizumab plus rivoceranib versus sorafenib as first-line therapy for unresectable hepatocellular carcinoma (CARES-310): a randomised, open-label, international phase 3 study. The Lancet 2023;402:1133–46. https://doi.org/10.1016/S0140-6736(23)00961-3.

[33] Kudo M, Finn RS, Qin S, et al. Lenvatinib versus sorafenib in first-line treatment of patients with unresectable hepatocellular carcinoma: a randomised phase 3 non-inferiority trial. Lancet 2018;391:1163–73. https://doi.org/10.1016/S0140-6736(18)30207-1.

[34] Qin S, Kudo M, Meyer T, et al. Tislelizumab vs Sorafenib as First-Line Treatment for Unresectable Hepatocellular Carcinoma: A Phase 3 Randomized Clinical Trial. JAMA Oncol 2023;9:1651–9. https://doi.org/10.1001/jamaoncol.2023.4003.

[35] Yau T, Park JW, Finn RS, et al. Nivolumab versus sorafenib in advanced hepatocellular carcinoma (CheckMate 459): a randomised, multicentre, open-label, phase 3 trial. Lancet Oncol 2022;23:77–90. https://doi.org/10.1016/S1470-2045(21)00604-5.

[36] Kelley RK, Rimassa L, Cheng AL, et al. Cabozantinib plus atezolizumab versus sorafenib for advanced hepatocellular carcinoma (COSMIC-312): a multicentre, open-label, randomised, phase 3 trial. Lancet Oncol 2022;23:995–1008. https://doi.org/10.1016/S1470-2045(22)00326-6.

[37] Llovet JM, Kudo M, Merle P, et al. Lenvatinib plus pembrolizumab versus lenvatinib plus placebo for advanced hepatocellular carcinoma (LEAP-002): a randomised, double-blind, phase 3 trial. Lancet Oncol 2023;24:1399–410. https://doi.org/10.1016/S1470-2045(23)00469-2.

[38] Bruix J, Qin S, Merle P, et al. Regorafenib for patients with hepatocellular carcinoma who progressed on sorafenib treatment (RESORCE): a randomised, double-blind, placebo-controlled, phase 3 trial. Lancet 2017;389:56–66. https://doi.org/10.1016/S0140-6736(16)32453-9.

[39] Abou-Alfa GK, Meyer T, Cheng A-L, et al. Cabozantinib in Patients with Advanced and Progressing Hepatocellular Carcinoma. N Engl J Med 2018;379:54–63. https://doi.org/10.1056/NEJMOA1717002.

[40] Zhu AX, Kang YK, Yen CJ, et al. Ramucirumab after sorafenib in patients with advanced hepatocellular carcinoma and increased α-fetoprotein concentrations (REACH-2): a randomised, double-blind, placebo-controlled, phase 3 trial. Lancet Oncol 2019;20:282–96. https://doi.org/10.1016/S1470-2045(18)30937-9.

[41] Qin S, Chen Z, Fang W, et al. Pembrolizumab Versus Placebo as Second-Line Therapy in Patients From Asia With Advanced Hepatocellular Carcinoma: A Randomized, Double-Blind, Phase III Trial. J Clin Oncol 2023;41:1434–43. https://doi.org/10.1200/JCO.22.00620.

[42] Finn RS, Ryoo BY, Merle P, et al. Pembrolizumab As Second-Line Therapy in Patients With Advanced Hepatocellular Carcinoma in KEYNOTE-240: A Randomized, Double-Blind, Phase III Trial. J Clin Oncol 2020;38:193–202. https://doi.org/10.1200/JCO.19.01307.

[43] Rimassa L, Chan SL, Sangro B, et al. 947MO Five-year overall survival (OS) and OS by tumour response measures from the phase III HIMALAYA study of tremelimumab plus durvalumab in unresectable hepatocellular carcinoma (uHCC). Annals of Oncology 2024;35:S656. https://doi.org/10.1016/J.ANNONC.2024.08.1007.

[44] Alexander BM, Schoenfeld JD, Trippa L. Hazards of Hazard Ratios - Deviations from Model Assumptions in Immunotherapy. N Engl J Med 2018;378:1158–9. https://doi.org/10.1056/NEJMC1716612.

[45] Chen TT. Statistical issues and challenges in immuno-oncology. J Immunother Cancer 2013;1:18. https://doi.org/10.1186/2051-1426-1-18.

[46] European Medicines Agency. ICH E9 (R1) addendum on estimands and sensitivity analysis in clinical trials to the guideline on statistical principles for clinical trials - Step 2b. EMA 2017;44.

[47] Pak K, Uno H, Kim DH, et al. Interpretability of Cancer Clinical Trial Results Using Restricted Mean Survival Time as an Alternative to the Hazard Ratio. JAMA Oncol 2017;3. https://doi.org/10.1001/jamaoncol.2017.2797.

[48] Liang F, Zhang S, Wang Q, et al. Treatment effects measured by restricted mean survival time in trials of immune checkpoint inhibitors for cancer. Annals of Oncology 2018;29:1320–4. https://doi.org/10.1093/annonc/mdy075.

[49] Jachno K, Heritier S, Wolfe R. Are non-constant rates and non-proportional treatment effects accounted for in the design and analysis of randomised

controlled trials? A review of current practice. BMC Med Res Methodol 2019;19. https://doi.org/10.1186/s12874-019-0749-1.

[50] Dormuth I, Liu T, Xu J, et al. Which test for crossing survival curves? A user’s guideline. BMC Med Res Methodol 2022;22. https://doi.org/10.1186/s12874-022-01520-0.

[51] Korn EL, Freidlin B, Mooney M. Stopping or reporting early for positive results in randomized clinical trials: The national cancer institute cooperative group experience from 1990 to 2005. Journal of Clinical Oncology 2009;27. https://doi.org/10.1200/JCO.2008.19.5339.

[52] Salzer WL, Devidas M, Carroll WL, et al. Long-term results of the pediatric oncology group studies for childhood acute lymphoblastic leukemia 1984-2001: A report from the children’s oncology group. Leukemia 2010;24. https://doi.org/10.1038/leu.2009.261.

[53] Schmid P, Cortes J, Dent R, et al. Overall Survival with Pembrolizumab in Early-Stage Triple-Negative Breast Cancer. N Engl J Med 2024;391:1981–91. https://doi.org/10.1056/NEJMOA2409932.

[54] Blank CU, Lucas MW, Scolyer RA, et al. Neoadjuvant Nivolumab and Ipilimumab in Resectable Stage III Melanoma. New England Journal of Medicine 2024.

[55] Cherny NI, Dafni U, Bogaerts J, et al. ESMO-Magnitude of Clinical Benefit Scale version 1.1. Ann Oncol 2017;28:2340–66. https://doi.org/10.1093/ANNONC/MDX310.

[56] Cherny NI, Oosting SF, Dafni U, et al. ESMO-Magnitude of Clinical Benefit Scale version 2.0 (ESMO-MCBS v2.0). Ann Oncol 2025. https://doi.org/10.1016/J.ANNONC.2025.04.006.

[57] Li Y, Hwang WT, Maude SL, et al. Statistical Considerations for Analyses of Time-To-Event Endpoints in Oncology Clinical Trials: Illustrations with CAR-T Immunotherapy Studies. Clinical Cancer Research 2022;28. https://doi.org/10.1158/1078-0432.CCR-22-0560.

[58] Wang ZX, Wu HX, Xie L, et al. Correlation of Milestone Restricted Mean Survival Time Ratio With Overall Survival Hazard Ratio in Randomized Clinical Trials of Immune Checkpoint Inhibitors: A Systematic Review and Meta-analysis. JAMA Netw Open 2019;2. https://doi.org/10.1001/JAMANETWORKOPEN.2019.3433.

[59] Kanavos P, Visintin E, Angelis A. Use of the ESMO-Magnitude of Clinical Benefit Scale to guide HTA recommendations on coverage and reimbursement for cancer medicines: a retrospective analysis. Lancet Oncol 2024;25:1644–54. https://doi.org/10.1016/S1470-2045(24)00505-9.

[60] Approaches to Assessment of Overall Survival in Oncology Clinical Trials | FDA n.d. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/approaches-assessment-overall-survival-oncology-clinical-trials

Figure 1. Proposed Strategy for Addressing Non-proportional Hazards

Abbreviations: KM, Kaplan-Meier; PH, proportional hazards; NPH, non-proportional hazards; G-T test, Grambsch-Therneau test; HR, hazard ratio; FU, follow-up; dRMST, difference in restricted mean survival time; rRMST, ratio of restricted mean survival time; pHR, piecewise Hazard Ratio.

Figure 2. Flowchart of Proportional Hazards Assessment and Non-Proportional Hazards Detection in Phase III HCC Trials

Abbreviations: RCTs, Randomized Controlled Trials, HCC, Hepatocellular Carcinoma; PH, proportional hazards; NPH, non-proportional hazards; G-T test, Grambsch-Therneau test; HR, hazard ratio; FU, follow-up; dRMST, difference in restricted mean survival time; rRMST, ratio of restricted mean survival time; pHR, piecewise Hazard Ratio.

Figure 3. *Kaplan-Meier Curve with Effect Size Measures: Cox Regression Analysis, Time-Dependent Hazard Ratios, and Restricted Mean Survival Time*.

A) Diminishing treatment effects on IMbrave 050 (RFS)*.* B) Diminishing Treatment Effect in Leap 012 (PFS). C) Delayed treatment effects in HIMALAYA (OS). D) Crossing hazard treatment effect in CheckMate 9DW (OS).

Abbreviations: HR, hazard ratio; rRMST, ratio of restricted mean survival time; pHR, piecewise Hazard Ratio.

For all panels, Kaplan-Meier curves were generated. The hazard ratio (HR) was estimated using the Cox proportional hazards model. Piecewise hazard ratios (pHR) were calculated for predefined time intervals to assess time-varying effects. The ratio of restricted mean survival times (rRMST) was computed at clinically relevant timepoints to quantify average treatment benefit. All statistical tests were two-sided, and significance was defined as $p < 0.05$.

Table 1. Main Characteristics and Grambsch-Therneau test of Phase III Randomized Controlled Trials.

Study	**Treatment and number of patients**	**HR (95% CI)** ***Cox Regression Analysis***	**Primary Endpoint**	**Grambsch-Therneau test (p-value)**
Adjuvant therapies				
STORM[25]	Resection + Sorafenib (n=556) vs. Resection + placebo (n=558)	0.94 (0.78-1.13)	RFS	p=0.064
IMbrave050 (interim analysis)[9]	Resection or RF + Atezolizumab plus Bevacizumab (n=334) vs. Resection or RF + active surveillance (n=334)	0.72 (0.53–0.98)	RFS	**p<0.001**
IMbrave050 (final analysis)[11]	Resection or RF + Atezolizumab plus Bevacizumab (n=334) vs. Resection or RF + active surveillance (n=334)	0.90 (0.72-1.12)	RFS	**p<0.001**
TACE plus Sorafenib or IT				
TACE 2[26]	TACE + Sorafenib (n=157) vs. TACE +placebo (n=156)	0.99 (0.77-1.27)	PFS	p=0.918
EMERALD-1[4]	Durvalumab + Bevacizumab + TACE (n=204) vs. TACE (n=205)	0.77 (0.61-0.98)	PFS	p=0.809
LEAP-012[5]	Lenvatinib + Pembrolizumab + TACE (n=237) vs. TACE + Placebo (n= 243)	0.66 (0.51-0.84)	PFS	**p=0.008**
1st line Systemic treatment				
SHARP[27]	Sorafenib, (n=299) vs. placebo (n=303)	0.69 (0.55-0.87)	OS	p=0.666
REFLECT[33]*****	Lenvatinib (n=478); sorafenib (n=476)	0.92 (0.79-1.06)	OS	p=0.168
IMbrave150[28]	Atezolizumab + bevacizumab (n = 336) vs. sorafenib (n = 165)	0.66 (0.52-0.85)	OS	p=0.301

HIMALAYA - STRIDE[29]	Stride (n=393) vs. sorafenib (n=389)	0.78 (0.65-0.93)	OS	**p=0.014**
HIMALAYA - Durvalumab[29]*****	Durvalumab (n=389) vs. sorafenib (n=389)	0.86 (0.74-1.01)	OS	p=0.480
CheckMate 9DW[31]	Nivolumab+ipilimumab (n = 335) vs. sorafenib or lenvatinib (n = 333).	0.79 (0.65-0.96)	OS	**p<0.001**
Cares-310[32]	Camrelizumab+rivoceranib (n=272) vs. sorafenib (n=271)	0.62 (0.49-0.80)	OS	p=0.117
RATIONALE 301[34]*****	Tislelizumab (n=342) vs. sorafenib (n=332)	0.85 (0.71-1.02)	OS	p=0.070
CheckMate-459[35]	Nivolumab (n=371) vs. sorafenib (n=372)	0.85 (0.72-1.02)	OS	p=0.088
LEAP-002[37]	Lenvatinib + pembrolizumab (n=395) vs. lenvatinib (n=399)	0.84 (0.71-1.00)	OS	p=0.103
COSMIC-312[36]	Cabozantinib+ atezolizumab (n=432) vs. cabozantinib (n=188)	0.63 (0.44-0.91)	PFS	p=0.068
2nd line Systemic treatment				
RESORCE[38]	Regorafenib (n=379) vs. placebo (n=194)	0.63 (0.50–0.79)	OS	p=0.962
CELESTIAL[39]	Cabozantinib (n=470) vs. placebo (n=237)	0.76 (0.63–0.92)	OS	p=0.288
REACH-2[40]	Ramucirumab (n=197) vs. placebo (n=95)	0.71 (0.53-0.95)	OS	p=0.131
KEYNOTE-394[41]	Pembrolizumab (n=300) vs. placebo (n=153)	0.79 (0.63-0.99)	OS	p=0.066
KEYNOTE-240[42]	Pembrolizumab (n=278) vs.placebo (n=135)	0.78 (0.61-0.99)	OS	p=0.820

*Non-inferiority trial against sorafenib; non-inferiority margin upper limit of 95% CI for HR <1.08.

Hazard ratios and 95% confidence intervals were estimated using Cox proportional hazards regression. The proportional hazards assumption was assessed using the Grambsch-Therneau test. All p-values correspond to two-sided tests, with significance defined as p < 0.05.

Abbreviations: HR, hazard ratio; CI, confidence interval; RF, radiofrequency; TACE, Transarterial Chemoembolization; IT, Immunotherapy; RFS, recurrence-free survival; PFS, progression-free survival; OS, overall survival; STRIDE, durvalumab + tremelimumab.

Table 2. Summary of Primary Efficacy Outcomes in Phase III RCTs with Non-Proportional Hazards

Study	Outcome	Median Follow-Up	Protocol-defined expected median for the control group	Numbers of Events out of total patients	Patterns of NPH	Log-Rank (p-value)	MaxCombo test (p-value)	RMST*
	Adjuvant therapy following surgery or ablation							
IMbrave050 - interim analysis. [9]	RFS	17.4 months	20 months	treatment arm (33%); active surveillance (40%)	Diminishing effects	**0.012**	**0.018**	**0.026**
IMbrave050 - second analysis. [11]	RFS	35.1 months	20 months	49% of patients in both arms	Diminishing effects	NR	0.326	0.080
	TACE + IT							
LEAP-012 interim analysis. [5]	PFS	25.6 months	8 months	Treatment arm (56%); control arm (63%)	Diminishing effects	**<0.001**	**<0.001**	**<0.001**
	1º Line Systemic Treatment							
HIMALAYA - STRIDE final analysis. [29]	OS	33.2 months	10 months	Treatment arm (67%); control arm (75%)	Delayed effects	**0.003**	**0.002**	**0.001**

CheckMate 9DW interim analysis. [31]	OS	35.2 months	19 months	Treatment arm (58%) and control arm (68%)	Crossing of hazards	**0.018**	**<0.001**	**0.030**

The MaxCombo test and restricted mean survival time (RMST) were assessed. A p-value <0.05 was considered statistically significant.

**: Assessed at the final time point at which more than 10% of patients remained at risk.*

Abbreviations: RFS, Recurrence-Free Survival; PFS, Progression-Free Survival; OS, Overall Survival; FU, Follow-Up; NPH, Non-Proportional Hazards; RMST, Restricted Mean Survival Time; CI - Confidence Interval; TACE, Transarterial Chemoembolization; IT, Immunotherapy; NR, Not Reported; STRIDE, durvalumab + tremelimumab.

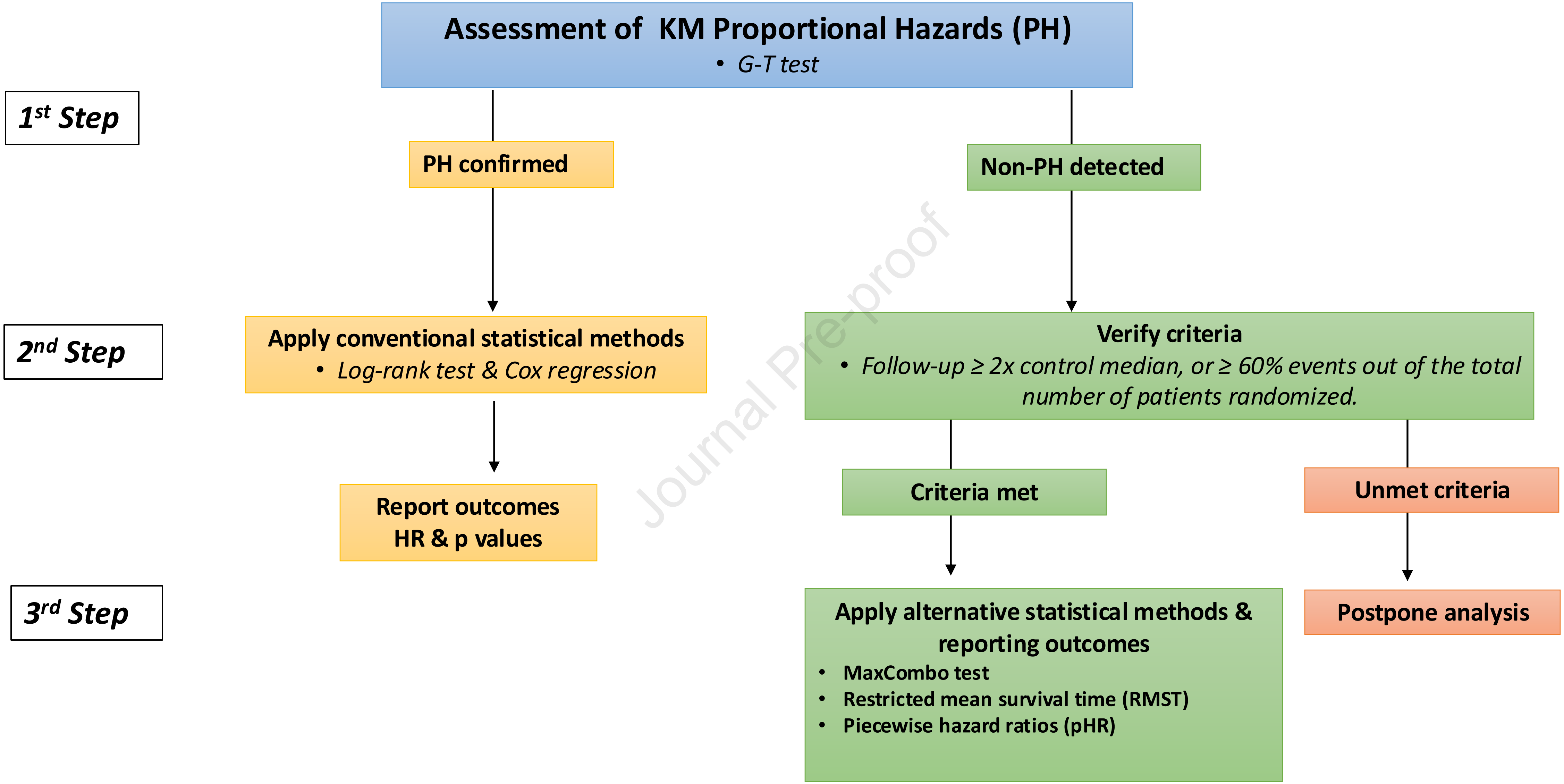
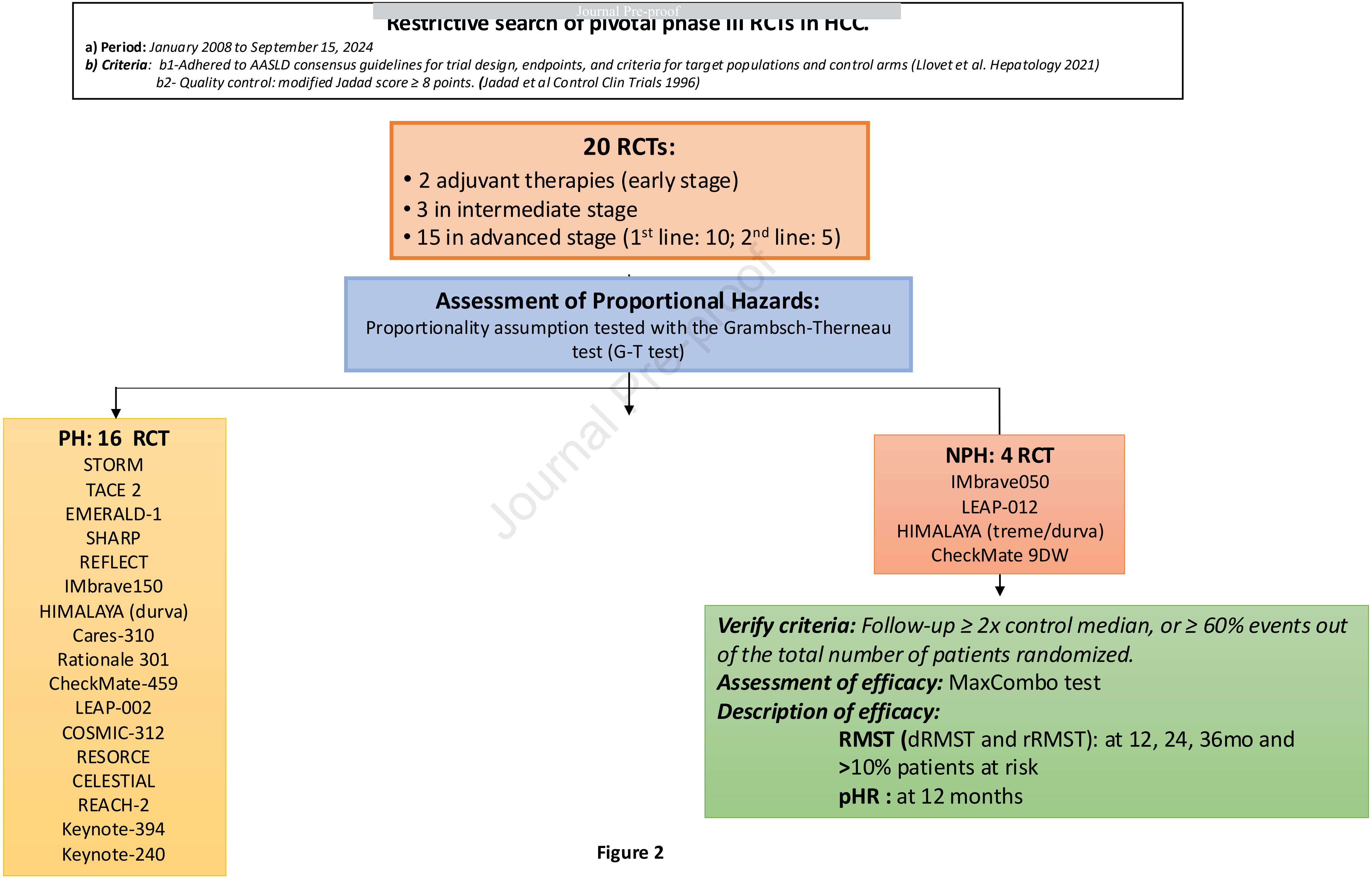

Assessment of KM Proportional Hazards (PH)
• G-T test
1st Step
PH confirmed
Non-PH detected
2nd Step
Apply conventional statistical methods
• Log-rank test & Cox regression
Verify criteria
• Follow-up ≥ 2x control median, or ≥ 60% events out of the total number of patients randomized.
Report outcomes
HR & p values
Criteria met
Unmet criteria
3rd Step
Apply alternative statistical methods & reporting outcomes
• MaxCombo test
• Restricted mean survival time (RMST)
• Piecewise hazard ratios (pHR)
Postpone analysis

Figure 1



Restrictive search of pivotal phase III RCTs in HCC.

a) Period: *January 2008 to September 15, 2024*

b) Criteria: *b1-Adhered to AASLD consensus guidelines for trial design, endpoints, and criteria for target populations and control arms (Llovet et al. Hepatology 2021)*

b2- Quality control: modified Jadad score ≥ 8 points. (Jadad et al Control Clin Trials 1996)

20 RCTs:

- 2 adjuvant therapies (early stage)
- 3 in intermediate stage
- 15 in advanced stage (1st line: 10; 2nd line: 5)

Assessment of Proportional Hazards:

Proportionality assumption tested with the Grambsch-Therneau test (G-T test)

PH: 16 RCT

STORM
TACE 2
EMERALD-1
SHARP
REFLECT
IMbrave150
HIMALAYA (durva)
Cares-310
Rationale 301
CheckMate-459
LEAP-002
COSMIC-312
RESORCE
CELESTIAL
REACH-2
Keynote-394
Keynote-240

NPH: 4 RCT

IMbrave050
LEAP-012
HIMALAYA (treme/durva)
CheckMate 9DW

Verify criteria: *Follow-up ≥ 2x control median, or ≥ 60% events out of the total number of patients randomized.*

Assessment of efficacy: MaxCombo test

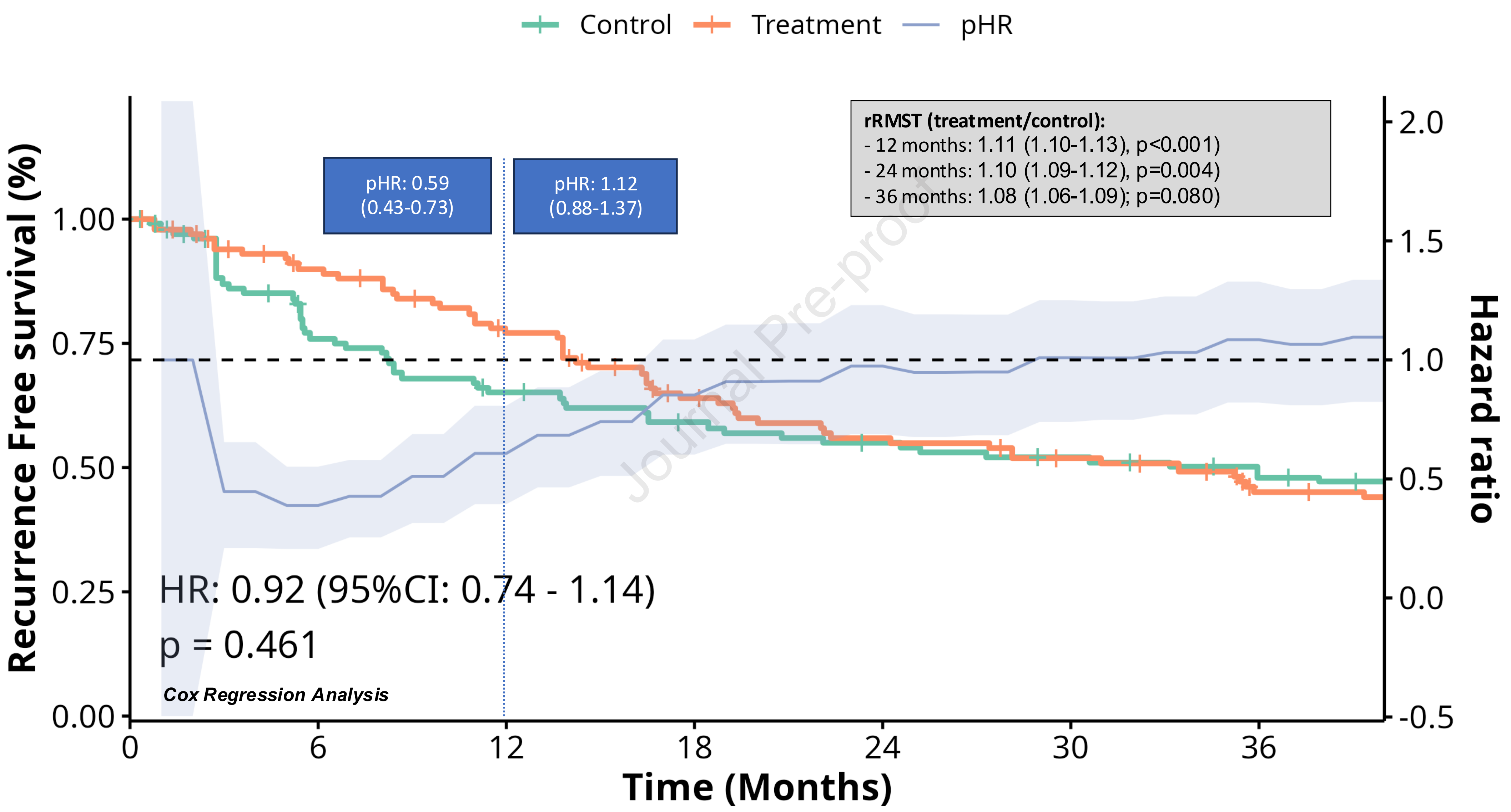
Description of efficacy:

RMST (dRMST and rRMST): at 12, 24, 36mo and **>**10% patients at risk

pHR : at 12 months

Figure 2

3A IMbrave 050



rRMST (treatment/control):
- 12 months: 1.11 (1.10-1.13), p<0.001)
- 24 months: 1.10 (1.09-1.12), p=0.004)
- 36 months: 1.08 (1.06-1.09); p=0.080)

pHR: 0.59 (0.43-0.73)

pHR: 1.12 (0.88-1.37)

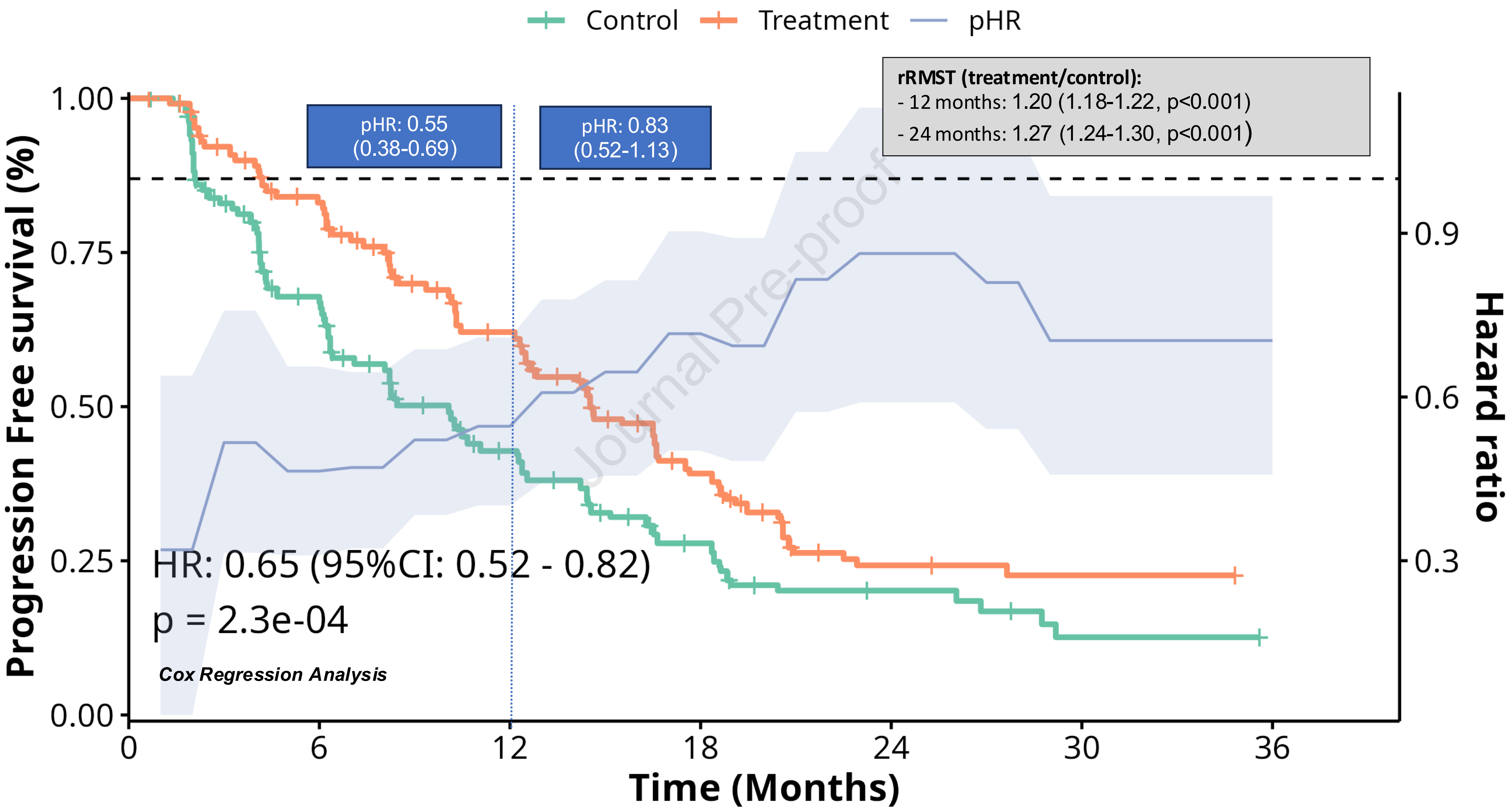
HR: 0.92 (95%CI: 0.74 - 1.14)

p = 0.461

Cox Regression Analysis

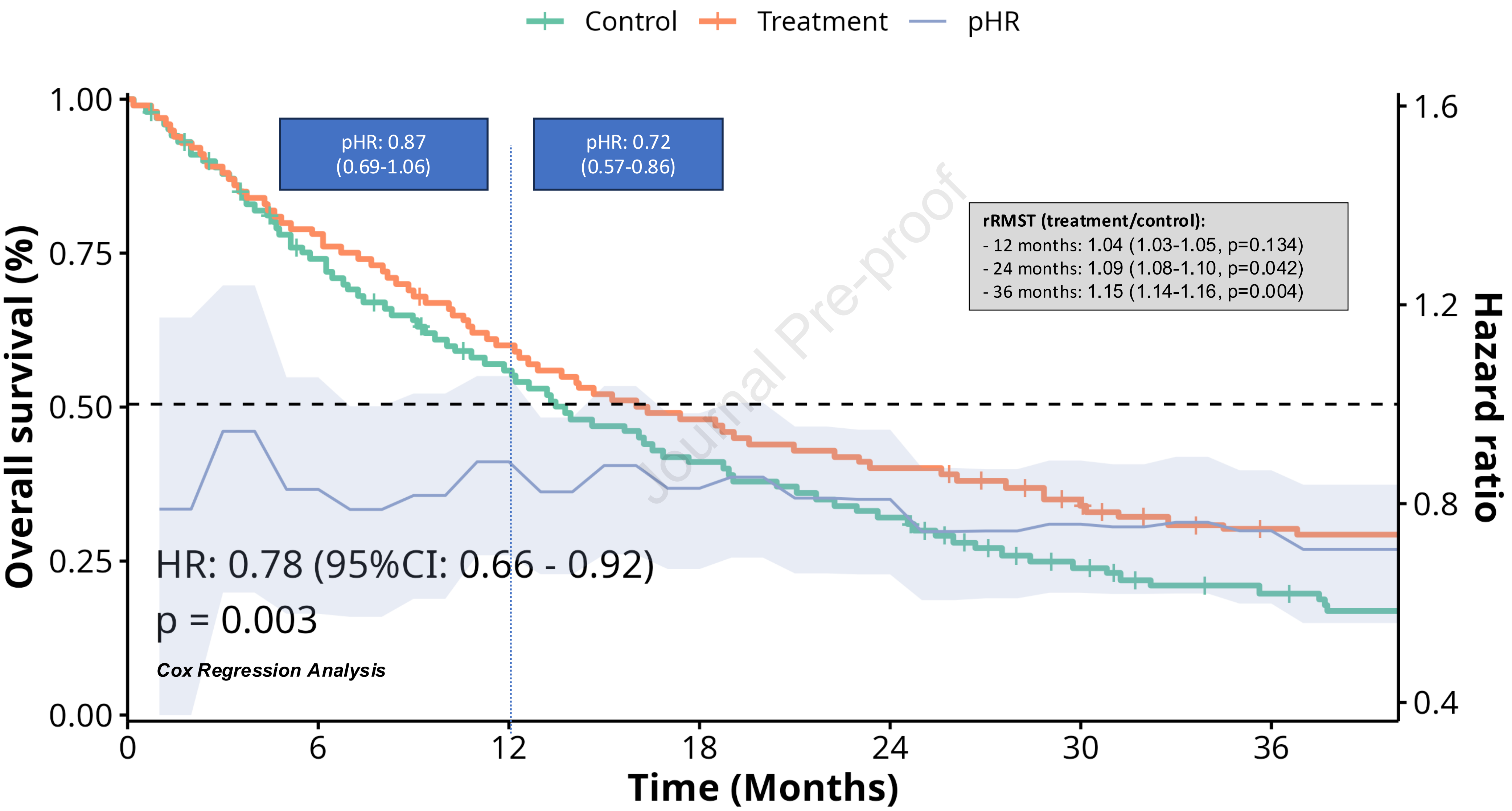
Atezo + bev	334	290	245	191	167	147	62
Active surveillance	334	247	207	185	170	145	63

3B



Lenva + Pembro	237	176	112	57	22	10	2
Dual placebo group	243	144	72	37	12	5	3

3C HIMALAYA (STRIDE)



pHR: 0.87 (0.69-1.06)

pHR: 0.72 (0.57-0.86)

rRMST (treatment/control):
- 12 months: 1.04 (1.03-1.05, p=0.134)
- 24 months: 1.09 (1.08-1.10, p=0.042)
- 36 months: 1.15 (1.14-1.16, p=0.004)

HR: 0.78 (95%CI: 0.66 - 0.92)

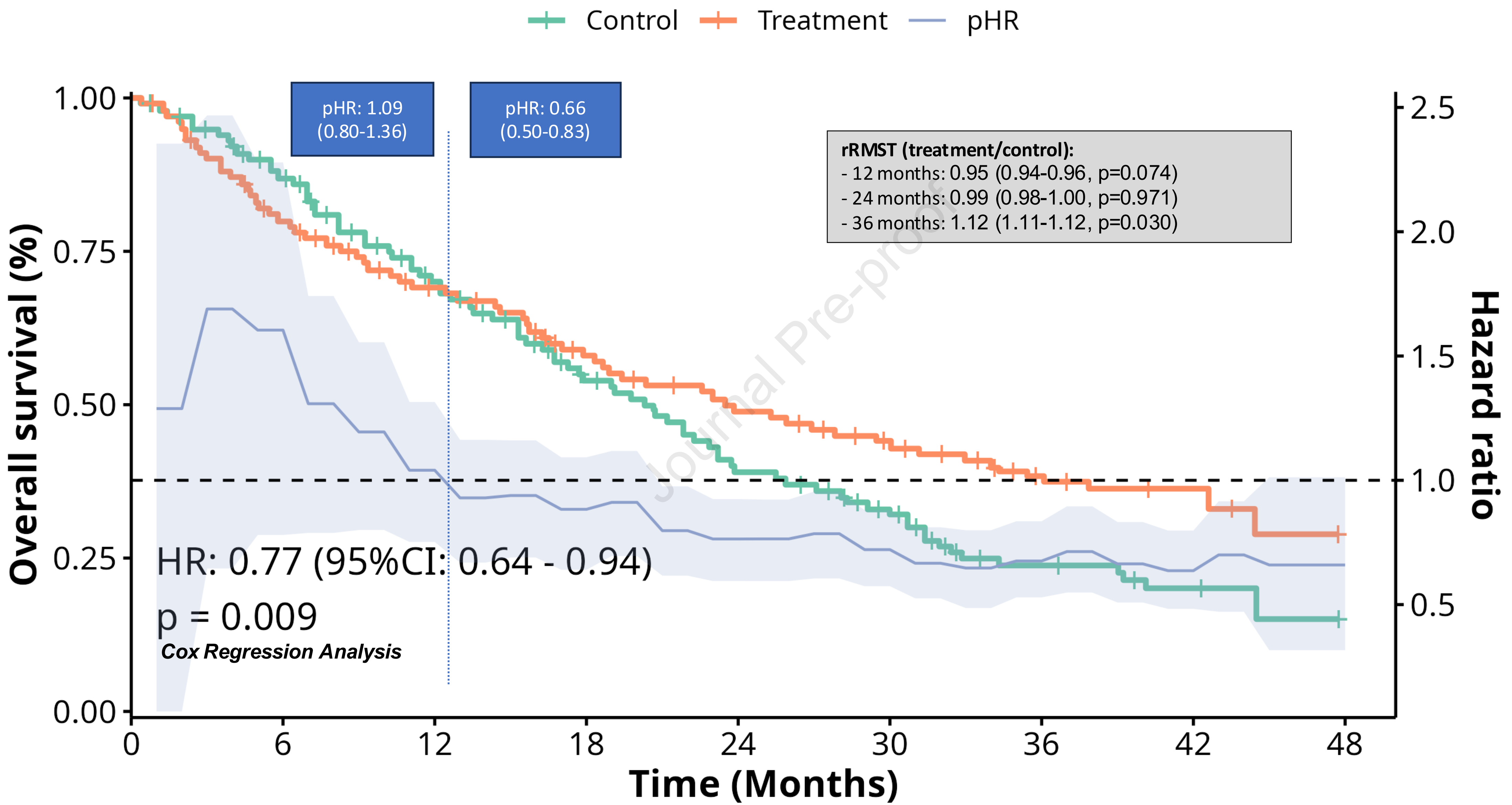
p = 0.003

Cox Regression Analysis

STRIDE	**393**	**308**	**235**	**190**	**158**	**98**	**32**
Sorafenib	389	283	211	155	121	62	21

3D

CheckMate 9DW



Nivo + Ipi	**335**	**264**	**220**	**179**	**150**	**104**	**42**	**11**	**0**
Lenvatinib or sorafenib	333	280	216	164	116	76	34	4	1

Highlights:

- NPH was present in 4/20 (20%) pivotal Phase III trials evaluated, all involving immunotherapies, and three patterns were identified: 1) diminishing effects, 2) delayed effects, and 3) crossing hazards.
- NPH caused discrepancies in IMbrave050's interim and subsequent efficacy analyses, thus explaining the distinct outcomes reported.
- Robust interim analysis in HCC requires a minimum follow-up duration or number of events before prematurely stopping an RCT with NPH.
- Whenever NPH is identified, distinct statistical tools should be used to assess reliable differences between arms (MaxCombo), and assessment of the effect size (RMST, and pHR) for regulatory decisions, and clinical guidance.
- These strategies are proposed to refine the trial design and enhance treatment decisions in HCC.