Evaluation of concordance between disease response assessments by investigator and by blinded independent central review (BICR) in hematology clinical trials

Xiaoyu Luo, Bohan Cui
School of Mathematical Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, 2144079, China
School of Mathematical Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, 2255981, China
Supervisor: Dr. Xiaoyu Tang

Corresponding author(s).
E-mail(s):xiaoyu.luo21@student.xjtlu.edu.cn
Bohan.Cui22@student.xjtlu.edu.cn

Abstract

Background: Blinded independent central review (BICR) is widely used in oncology trials to mitigate bias in endpoint assessment, but its added value in hematologic malignancies (HMs) remains debated due to disease-specific complexities in evaluating progression-free survival (PFS) and objective response rate (ORR).

Methods: We conducted a meta-analysis of 42 Phase III randomized controlled trials (2014–2024) across HMs (lymphoma, leukemia, myeloma) comparing PFS and ORR assessments by BICR versus investigators (INV). Correlation coefficients (Pearson's 'r'), hazard ratio ratios (HRR), and odds ratio ratios (ORR) were calculated. Subgroup

analyses assessed concordance by trial design (blinded/open label) and disease type.
Results: Among 36 PFS comparisons, BICR and INV showed strong correlation ($r = 0.952$, 95% CI: 0.907–0.975), with 95.3% variability explained ($R^2 = 0.953$). HRR indicated minimal divergence (mean difference: 6.5%; pooled HRR = 0.935, 95% CI: 0.892–0.980). Open-label trials exhibited slightly lower concordance (HRR = 0.920) than blinded trials (HRR = 1.009). For ORR (21 comparisons), correlation remained high ($r = 0.900$, $R^2 = 0.823$), with pooled ORR = 0.870 (95% CI: 0.685–1.105). Statistical inferences agreed in 75% of PFS comparisons (Cohen's $\kappa = 0.901$).
Conclusion: BICR and INV assessments demonstrate excellent agreement in HM trials, supporting the reliability of local evaluations. Open-label designs may introduce modest bias, but BICR's incremental value appears limited, particularly in large or blinded studies. These findings question the necessity of routine BICR in HM trials, suggesting context-specific use rather than universal mandate.
Keywords: hematologic malignancy, blinded independent central review, progression-free survival, objective response rate, meta-analysis

1 Introduction

Hematologic malignancy (HM) comprises a diverse group of conditions arising from cells in the bone marrow and lymphatic system (Miller et al. 2025). In the United States alone, approximately 1.4 million people are living with or in remission from HM (Rodriguez-Abreu, Bordoni, and Zucca 2007). Therapeutic efficacy evaluation in hematologic malignancies poses unique challenges compared to solid tumors, requiring tailored methodologies to reduce bias and ensure reliable trial outcomes.

In hematology clinical trials, endpoints like progression-free survival (PFS) and objective response rate (ORR) are widely used and often rely on integrated evaluations incorporating imaging, histopathology, and molecular diagnostics in blood-based cancers. For instance, response criteria like the Lugano classification for lymphomas (Cheson et al. 2014) and the International Myeloma Working Group (IMWG) criteria for multiple myeloma (Narita et al. 2018) require interpretations of radiographic findings, circulating biomarkers, and bone marrow biopsies. This process is inherently vulnerable to subjectivity and variability. While these frameworks aim to

standardize assessments, the assessment of progression is subject to measurement variability which may introduce error or bias (Dancey et al. 2009), particularly in open-label trials where investigator awareness of treatment allocation may influence outcome assessments. Subtle biases may lead to delayed identification of disease progression in patients receiving experimental therapies or premature discontinuation of treatment due to perceived safety concerns. These challenges are amplified in hematologic malignancies, where distinguishing disease progression from treatment-related complications can be clinically ambiguous. Even in double-blind trials, investigators' knowledge of treatment allocation may persist due to treatment-specific adverse effects or distinct clinical progression patterns, particularly in hematologic malignancies. Blinded independent central review (BICR) (Ford et al. 2009) improves objectivity through centralized analysis and standardized evaluation protocols, particularly valuable for novel therapies or rare hematologic malignancies lacking established response criteria.

However, BICR's applicability in hematologic research remains debated. While radiographic progression drives endpoint determination in solid tumors, hematologic malignancies frequently require supplementary diagnostic approaches—including flow cytometry, cytogenetic testing, and minimal residual disease (MRD) assessment—to comprehensively evaluate disease status. This complexity limits BICR's effectiveness given its primary focus on imaging analysis. Practical challenges such as review delays and the resource-intensive nature of cross-functional expert panels further complicate cost-benefit considerations for BICR in time-sensitive hematologic trials. Recent meta-analyses in solid tumors suggest minimal divergence between BICR and investigator (INV) for PFS and ORR. In 2024, Jacobs et al. found no significant differences in PFS by either local assessment or BICR from 24 studies enrolling 13,168 patients (Jacobs et al. 2024). Several studies have evaluated discordance rates between local investigators and BICR-assessed PFS in solid tumor trials (W. Liang et al. 2016; J. Zhang, Y. Zhang, Tang, Jiang, et al. 2018; J. Zhang, Y. Zhang, Tang, H. Liang, et al. 2017), but none have specifically addressed hematologic malignancies, where endpoint definitions are more heterogeneous and may yield different conclusions.

To address this evidence gap, we conducted a meta-analysis of 42 clinical trials across hematologic malignancies, comparing treatment effect estimates for PFS and ORR from BICR versus investigator assessments. We included pivotal Phase III trials, diverse disease subtypes (including lymphomas, leukemias, and multiple myeloma) and emerging therapies such as

CAR-T cells and bispecific antibodies. We evaluated the reliability of local evaluations and the added value of central review in detecting meaningful differences in treatment effects.

2 Materials and Methods

2.1 Searching Strategy

The meta-analysis was performed using comprehensive test-level information from all phase III randomized controlled trials (RCTs) of hematoma studies(N = 42) from 2014 to 2024. Eligible studies we included were Phase III randomized controlled trials designed to assess the effectiveness of anticancer treatments in patients with hematological malignancies. Additionally, tumor response or progression was evaluated through imaging assessments conducted by both central reviewers and local investigators. A comprehensive search strategy was implemented to identify relevant studies for this meta-analysis. We search PubMed up to December 4, 2024, for articles using the search terms: "((Lymphoma [Mesh] OR Myeloma[Mesh] OR Leukemia[Mesh] OR Lymphoma [tiab] OR Myeloma[tiab] OR Leukemia[tiab] OR Hematologic [Mesh] OR Hematologic[tiab] AND (progression free survival [Mesh] OR disease progression[Mesh] OR progression free survival[tiab] OR PFS[tiab] OR objective response[tiab] OR ORR[tiab] OR investigator OR independent review) AND ("Phase 3" OR "Phase III"OR"Phase 2" OR "Phase II") (English) AND (y_10[Filter])) AND (randomized controlled trial [pt] OR controlled clinical trial [pt] OR randomized [tiab] OR placebo [tiab] OR drug therapy [sh] OR randomly [tiab] OR trial [tiab] OR groups [tiab] AND (y_10[Filter])) AND (y_10[Filter])) AND ("2014/01/01"[Date - Publication] : "2024/11/27"[Date - Publication])".

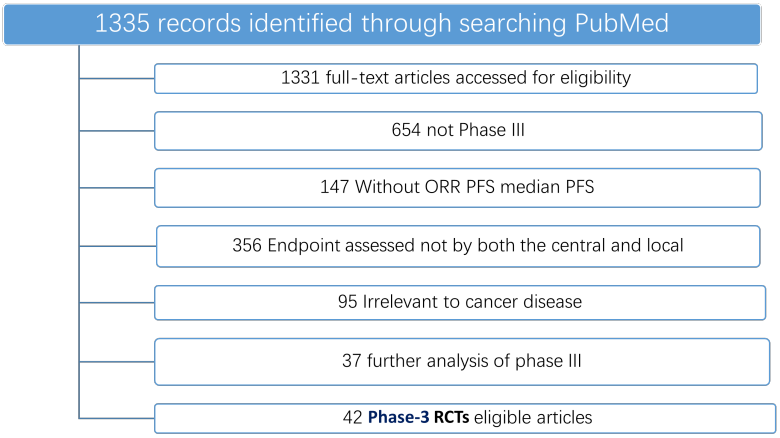

Figure 1: Flow chat

Regardless of trial results, in hematological malignancies, BICR and LE assessment results reported PFS and/or ORR, respectively. Eligible trials were Phase III randomized controlled trials that directly evaluate the efficacy of anticancer drugs in patients with hematoma. These trials played a crucial role in advancing cancer treatment by providing robust evidence on the effectiveness and safety of new therapies. In such studies, researchers selected participants who meet specific criteria to ensure the validity and reliability of the results. The primary objective of these Phase III trials was to compare the new anticancer drugs against existing standard treatments or placebos. This comparison helped determine whether the new drugs offer significant advantages in terms of tumor shrinkage, progression-free survival, overall survival, and quality of life for patients. The inclusion of patients with hematoma ensures that the focus of the study is on those individuals who are most likely to benefit from these new treatments.

2.2 PFS Meta-analysis

The first primary outcome of interest is progression-free survival(PFS). This quantity is defined as the time from entry into a clinical trial until progression or death, whichever comes first. The relationship between BICR and LE in terms of log(HR of PFS) was assessed using Pearson's correlation coefficient (r). To determine the extent to which variability in BICR estimates

is accounted for by LE estimates, the coefficient of determination (R^2) was derived from a linear regression model. This model was based on log(HR) values and weighted according to sample size. The level of concordance between BICR and LE for PFS was assessed using the Hazard Ratio:

$$HRR = \frac{HR_{\text{LE}}}{HR_{\text{BICR}}},$$

where HR_{LE} and HR_{BICR} are the BICR- and LE-based PFS HRs, respectively. The Hazard Ratio Ratios (HRRs) were calculated for each individual comparison and subsequently categorized into four distinct intervals: ≤ 0.85, $(0.85, 1]$, $(1, 1.15]$, and > 1.15. These intervals were then analyzed in relation to key study characteristics to explore potential associations between these characteristics and the HRRs. This stratification allowed for a more detailed understanding of how study design factors, such as treatment regimens, patient population, and assessment methodologies, may influence the magnitude of agreement between the BICR and LE-based hazard ratios.
For studies where the protocol stipulated that progression-free survival (PFS) comparisons should be formally tested, statistical significance was determined by comparing the P-value to a pre-specified alpha boundary. Based on this comparison, statistical inferences were categorized as either "statistically significant" or "not statistically significant." This approach ensured that only those comparisons meeting the threshold for significance were included in the final analysis.
To assess the consistency of statistical inferences between BICR and LE, a two-way contingency table was constructed. The table allowed for a direct comparison of the PFS outcomes derived from both BICR and LE-based assessments. Cohen's kappa coefficient was then calculated to quantify the level of agreement between the two methods, providing a measure of inter-rater reliability. A kappa value closer to 1 would indicate a strong agreement, whereas a value closer to 0 would suggest a lack of agreement, thus offering a clearer picture of how consistently BICR and LE-based evaluations align in terms of statistical significance.

2.3 ORR Meta-analysis

The second primary outcome we considered is objective response rate(ORR). Similar to the PFS meta-analysis, Pearson's correlation coefficient (r) and

the coefficient of determination (R^2) were computed to assess the relationship between LE and BICR based on the logarithm of the OR for ORR. Additionally, the level of agreement between BICR and LE regarding ORR was measured using the Odds Ratio Ratio of the response:

$$\text{OddsRR} = \frac{OR_{\text{LE}}}{OR_{\text{BICR}}},$$

where

$$OR_{\text{BICR}} = \frac{ORR_{\text{armA BICR}}(1 - ORR_{\text{armB BICR}})}{ORR_{\text{armB BICR}}(1 - ORR_{\text{armA BICR}})},$$

and

$$OR_{\text{LE}} = \frac{ORR_{\text{armA LE}}(1 - ORR_{\text{armB LE}})}{ORR_{\text{armB LE}}(1 - ORR_{\text{armA LE}})}.$$

A similar random effects model used in the PFS meta-analysis was applied to estimate the Odds ratio ratio(ORR) and its 95% CI for all comparisons and for double-blind and open-label comparisons, respectively.

3 Results

Based on article identification and selection Figure 1, we include total 42 phase-3 RCTs eligible articles to do the analysis of the consistency of data analysis between BICR and investigators, involving 36 PFS comparisons and 21 ORR comparisons. The Table 1 shows the summarization of the characteristics of the investigated comparisons, such as the PFS comparisons and the ORR comparisons.

The agreement between the BICR and LE PFS results can be observed in Table 2. The comprehensive analysis encompassing 36 pairwise comparisons demonstrated a remarkably strong correlation between log (HR BICR) and log (HR INV), with an overall Pearson's correlation coefficient of r = 0.952 (95% CI, 0.907-0.975). The results of the subgroup analyses in the open-label (n = 27) and blinded (n = 9) comparisons agreed with the overall study. In the blinded and open-label groups, the degrees of correlation were 0.908 (95% CI, 0.613, 0.981) and 0.966 (95% CI, 0.927, 0.985), respectively. The disease-specific analyses showed strong correlations between log (HR BICR) and log (HR INV) across hematologic malignancies: lymphoma (n=26, r=0.913, 95% CI 0.799-0.964), leukemia (n=7, r=0.972, 95% CI 0.756-0.997), and myeloma (n=4, r=0.953, 95% CI -0.096-0.999). The limited sample size in myeloma

Table 1: Comparison of PFS and ORR Characteristics

Characteristics	**PFS comparisons (n=36)**	**ORR comparisons (n=21)**
Masking		
Blind	5 (13.89%)	2 (9.52%)
Double-blind	4 (11.11%)	2 (9.52%)
Open-label	27 (75.00%)	17 (80.95%)
Sample size		
≤ 100	3 (8.33%)	3 (14.29%)
100 to 500	26 (72.22%)	12 (57.14%)
≥ 500	7 (19.44%)	6 (28.57%)
Cancer type		
LYM	23 (63.89%)	12 (63.16%)
LEU	6 (16.67%)	4 (21.05%)
MM	5 (13.89%)	3 (10.53%)

and the combined lymphoma/leukemia subgroup (n=3) precluded reliable interpretation. Overall, the weighted linear regression model using log (HR INV) as the explanatory variable explains 95.3% ($\mathbf{R^2} = 0.953$ [95% CI, 0.926, 0.980]) of the variability in the log (HR BICR), confirming the strong agreement between BICR and INV. The estimated overall HRR from the random effects model was 0.935 (95% CI: 0.892, 0.980), indicating an average difference of only 6.5% between HR BICR and HR INV. The estimated HRR in the open-label group was numerically slightly lower than that in the double-blind group (0.920 vs. 1.009), but both were close to 1 indicating a high degree of agreement in the PFS HR estimates overall.

Table 2: Agreement Assessment of PFS Between BICR and INV

Number of Comparisons	**Log (HR)** r^2 **(95% CI)**	**Log (HR)**/R^2 **(95% CI)**	**HRR (95% CI)**
Overall:36	0.952 (0.907, 0.975)	0.953 (0.926, 0.980)	0.935 (0.892, 0.980)
Open-label:27	0.966 (0.927, 0.985)	0.950 (0.929, 0.981)	0.920 (0.870, 0.972)
Blinded:9	0.908 (0.613, 0.981)	0.942 (0.909, 0.975)	1.009 (0.900, 1.132)

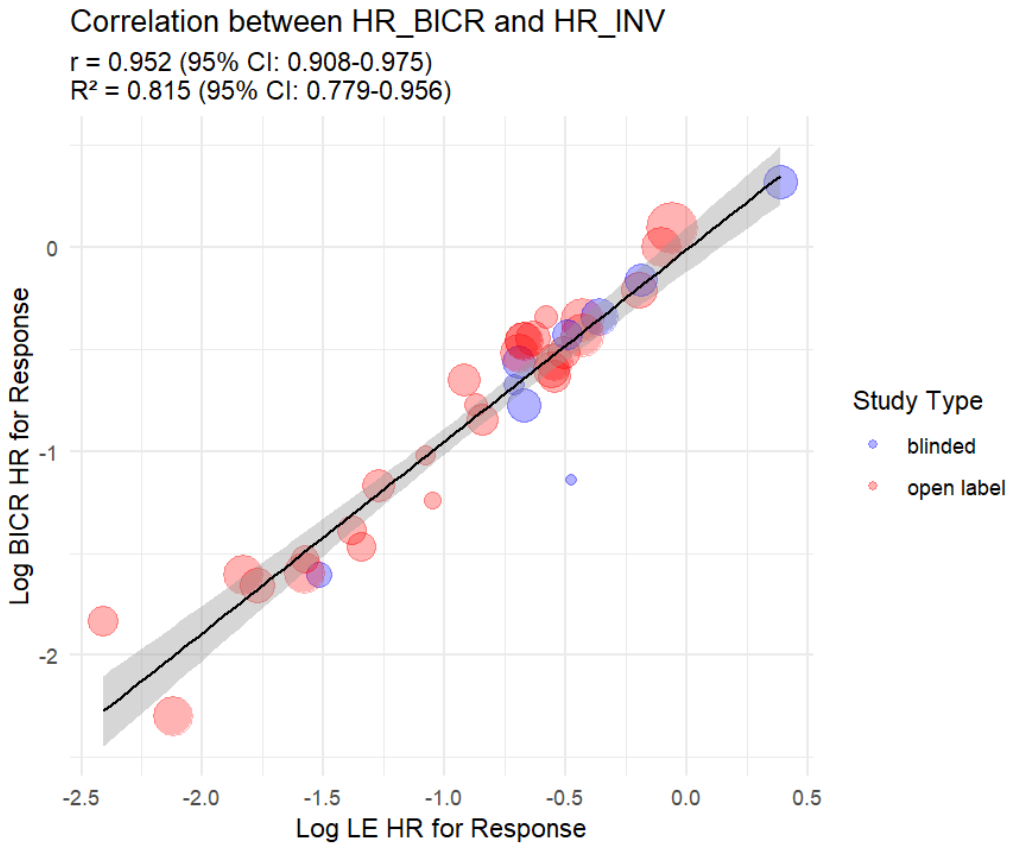

Figure 2: LogHR plot. The size of the circle represents the number of samples.

The scatter plot Figure 2 demonstrates a strong correlation between $\text{Log}(\text{HR}_{\text{BICR}})$ and $\text{Log}(\text{HR}_{\text{LE}})$ for PFS comparisons, with a correlation coefficient of $r = 0.952$ (95% CI: 0.907–0.975) and a coefficient of determination of $R^2 = 0.953$ (95% CI: 0.919–0.977). The data points cluster closely around the solid reference line, indicating a high overall agreement between the two assessment methods. However, distinct patterns emerge when considering the study type and sample size. In open-label studies, the points generally fall above the reference line, suggesting that LE assessments may overestimate the treatment effect compared to BICR. This trend is particularly noticeable in smaller trials, where the deviation from the reference line is more pronounced. In contrast, larger open-label studies show better agreement, with points align-

ing more closely to the reference line. In blinded studies, including both double-blind and blinded trials, the points are evenly distributed around the reference line, indicating no systematic bias and a consistent agreement between the two methods, regardless of sample size. Larger blinded studies exhibit even stronger consistency, with points clustering more tightly around the reference line. In general, the graph highlights a high level of agreement between the BICR and LE assessments, but reveals a potential bias in open-label studies, particularly in smaller trials, where the LE assessments can overestimate the effect of treatment.

The analysis of hazard risk ratio (HRR) distribution among different study characteristics is presented in Table 3. Among the 36 studies analyzed, 22.22% (n=8) reported HRR $\leq$ 0.85, while 38.89% (n=14) had HRR between 0.85 and 1. The proportion of studies with HRR between 1 and 1.15 was 30.56% (n=11), and only 8.33% (n=3) exhibited HRR $>$ 1.15. Among the five blind studies, none showed HRR $\leq$ 0.85, with the majority (60.00%, n=3) falling within the (0.85, 1] range, while two studies (40.00%) had HRR in the (1, 1.15] range. For double-blind studies (n=4), HRR was evenly distributed between $\leq$ 0.85 (0%), (0.85, 1] (50.00%), (1, 1.15] (25.00%), and $>$ 1.15 (25.00%). In contrast, open-label studies (n=27) had a relatively higher proportion of HRR $\leq$ 0.85 (29.63%) compared to blind and double-blind studies, with 33.33% in the (0.85, 1] range, 29.63% in the (1, 1.15] range, and 7.41% in the $>$ 1.15 category.

Table 4 summarizes the consistency of statistical inferences between BICR and INV based on PFS results. We denote a statistically significant PFS difference based on BICR assessment as "BICR+" and a not statistically significant difference as "BICR−"; "INV+" and "INV−" are analogously defined for the INV-based assessment. Of the 36 comparisons, 28 were alpha-controlled, leading to statistical inferences. For the majority (75.0% [27/36]) of the comparisons, BICR agreed with INV in terms of the resulting statistical inferences: 78.6% (22/36) of the comparisons led to an INV+/BICR+ result and 17.9% (5/36) to an INV−/BICR− result. Discordant results, i.e., INV+/BICR− and INV−/BICR+ combinations, were observed in 1 (2.2%) and 0 (0%) of the comparisons, respectively. Cohen's Kappa was 0.901, indicating a substantial agreement between BICR and INV. Overall, a high agreement between BICR and INV estimates of the PFS treatment effect was observed in the meta-analysis, while the agreement was slightly stronger in the double-blind subgroup than in the open-label group. At the individual-trial level, BICR and INV gave consistent statistical inferences

Table 3: Characteristics and HRR Distribution

Characteristics	**n**	≤ 0.85	**(0.85, 1]**	**(1, 1.15]**	> 1.15
Overall, n (%)	36	8 (22.22%)	14 (38.89%)	11 (30.56%)	3 (8.33%)
Masking					
Blind	5	0	3 (60.00%)	2 (40.00%)	0
Double-blind	4	2 (50.00%)	1 (25.00%)	1 (25.00%)	0
Open-label	27	8 (29.63%)	9 (33.33%)	8 (29.63%)	2 (7.41%)
Sample size					
≤ 100	3	0	1 (33.33%)	0	2 (66.67%)
100 to 500	26	7 (26.92%)	9 (34.62%)	10 (38.46%)	0
≥ 500	7	1 (14.29%)	4 (57.14%)	1 (14.29%)	1 (14.29%)
Cancer type					
LEU	6	2 (33.33%)	1 (16.67%)	3 (50.00%)	0
LEU/LYM	2	0	1 (50.00%)	0	1 (50.00%)
LYM	23	6 (26.09%)	10 (43.48%)	5 (21.74%)	2 (8.70%)
MM	5	0	2 (40.00%)	3 (60.00%)	0
Follow-up duration					
< 24	9	3 (33.33%)	4 (44.44%)	2 (22.22%)	0
24 to 36	15	2 (13.33%)	6 (40.00%)	5 (33.33%)	2 (13.33%)
> 36	12	1 (16.67%)	4 (33.33%)	4 (33.33%)	3 (25.00%)
< 30	20	5 (25.00%)	9 (45.00%)	5 (25.00%)	1 (5.00%)
> 30	10	3 (30.00%)	3 (30.00%)	5 (50.00%)	1 (10.00%)

in the majority of the comparisons, which is in line with the meta-analysis result.

	Consistency of Statistical Inference per PFS among α-Controlled Comparisons	
	Statistically significant (BICR+)	**Not statistically significant (BICR−)**
Statistically significant (INV+)	22 (78.6%)	1 (3.6%)
Not statistically significant (INV−)	0	5 (17.9%)
N = 28		

Table 4: Consistency of statistical inference per PFS among α-controlled comparisons.

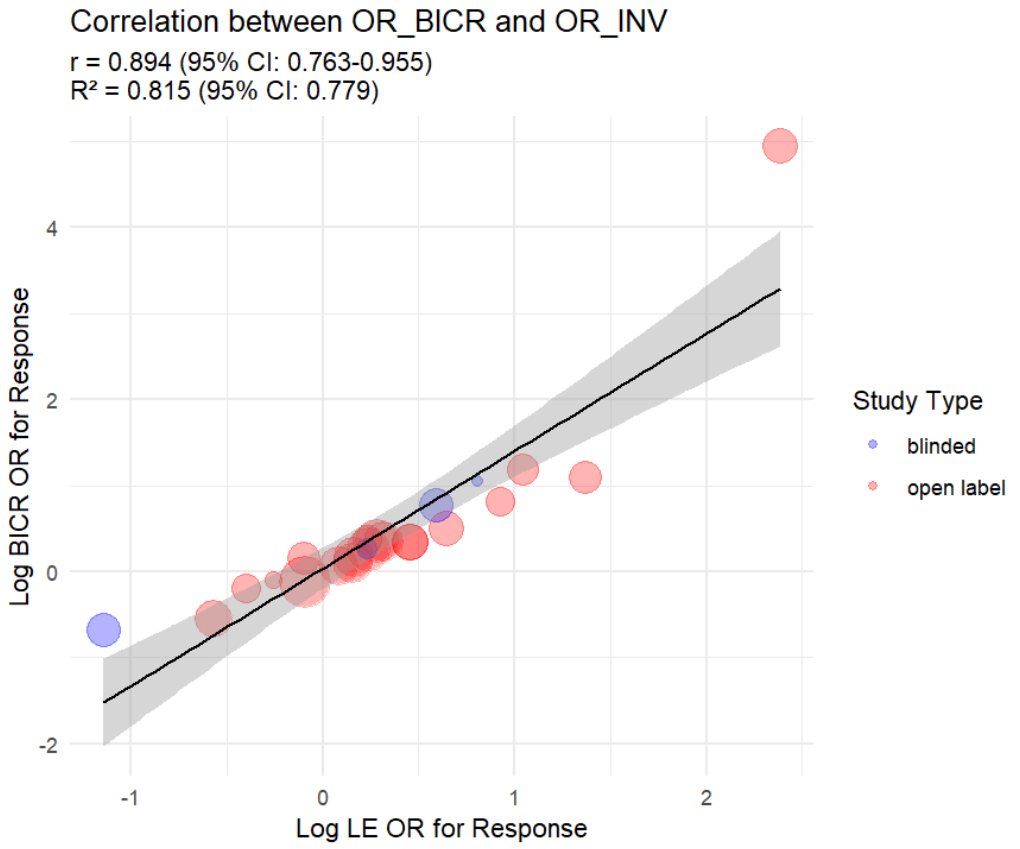

Figure 3: LogOR plot. The size of the circle represents the number of samples.

Number of Comparisons	**Log (OR) r (95% CI)**	**Log (OR) R^2 (95% CI)**	**OddsRR (95% CI)**
Overall: 21	0.900 (0.765, 0.995)	0.823 (0.798, 0.847)	0.984 (0.884, 1.103)
Open-label: 17	0.919 (0.785, 0.0971)	0.866 (0.700, 0.926)	0.988 (0.884, 1.104)
Blinded: 4	0.989 (0.457, 1.000)	0.990 (0.993, 0.996)	0.974 (0.864, 1.099)
Lymphoma: 12	0.961 (0.805, 0.989)	0.928 (0.887, 0.970)	0.992 (0.925, 1.063)
Myeloma: 2	*NA*	*NA*	*NA, NA*
Leukemia: 5	0.987 (0.811, 0.999)	0.977 (0.964, 0.991)	0.906 (0.881, 1.125)
Lymphoma/Leukemia: 2	*NA*	*NA*	*NA, NA*

Table 5: Agreement assessment of ORR between BICR and INV.

An analysis of ORR was conducted and it was carried out in a similar way to

that of PFS analysis (Table 5). This presented the agreement between BICR and INV assessments for response outcomes across various hematologic malignancies. We analyze 21 ORR comparisons, showing strong correlation between BICR and INV assessments, with Pearson's correlation coefficient (r): 0.900 (95% CI: 0.765-0.959). The Coefficient of determination (R^2) is 0.823 with 95% CI: 0.728-0.918, indicating INV assessments explain 82% of variability in BICR assessments. The Pearson correlation coefficient indicated a substantial connection between $\log(\text{OR}_{\text{BICR}})$ and $\log(\text{OR}_{\text{INV}})$, but slightly lower than that for PFS log(HR). The analysis results of the open-label group and the blinded group were consistent with the overall study. In the blinded group ($n = 4$), the correlation coefficient was 0.985 (95% CI: 0.457, 1.000), and in the open-label group ($n = 17$), it was 0.919 (95% CI: 0.786, 0.971). While both groups demonstrated strong concordance, the point estimate in blinded trials was 7.2% higher than in open-label studies (0.985 vs 0.919), possibly reflecting the potential impact of different trial designs on the results. Disease-specific analysis showed significant correlations between $\log(\text{OR}_{\text{BICR}})$ and $\log(\text{OR}_{\text{INV}})$ in lymphoma, leukemia, and myeloma among hematological malignancies. In the lymphoma group ($n = 12$), $r = 0.961$ (95% CI: 0.865, 0.989); in the leukemia group ($n = 5$), $r = 0.987$ (95% CI: 0.811, 0.999). However, the limited sample size in myeloma and the combined lymphoma/leukemia subgroup restricted the accuracy of the analysis.
The weighted linear regression model, Figure 3, employing $\log(\text{OR}_{\text{INV}})$ as the explanatory variable, effectively accounted for 82.3% of the variability in $\log(\text{OR}_{\text{BICR}})$ ($R^2 = 0.823$, 95% CI: 0.728, 0.918), further illustrating the strong correlation between BICR and INV assessment outcomes. The overall ORR estimated by the random-effects model was 0.870 (95% CI: 0.685, 1.105). The ORR estimate in the open-label group (0.885, 95% CI: 0.666, 1.175) was marginally superior to that in the blinded group (0.748, 95% CI: 0.581, 0.962), however both values approached 1, emphasizing the consistency of ORR assessment across various trial settings.

4 Discussion and Conclusion

To our knowledge, this is the first research article evaluating the differences between investigator evaluations (INV) and blinded independent central reviews (BICR) of PFS and ORR in phase III RCTs concerning hematologic malignancy. While several systematic reviews (Dello Russo, Cappoli, and

Navarra 2020; **zhang2018**; Lian et al. 2024, and references therein) have examined this issue in solid tumors, the hematologic malignancy literature consists primarily of isolated analyses from phase III clinical trials, leaving significant gaps in our understanding of disease-specific assessment concordance. This paucity of comprehensive evidence is particularly concerning given the unique challenges of response evaluation in hematologic cancers, where factors like bone marrow involvement (Sorigue, Cañamero, and Miljkovic 2021), circulating tumor cells (Lin et al. 2021), and distinct response criteria may differentially influence local versus central assessments.

The scatterplot demonstrates strong concordance between BICR and local evaluation (LE) assessments overall, with most data points adhering closely to the equivalence line. However, deeper analysis reveals important contextual variations. In open-label trial designs, particularly those with limited enrollment ($n < 100$), LE measurements systematically plot above the consensus line—a pattern suggesting potential inflation of treatment effects compared to blinded assessments. This divergence appears dose-dependent with sample size, as larger open-label trials ($n \geq 300$) show markedly improved alignment with central tendency.

The results demonstrate strong agreement between BICR and investigator assessments for response outcomes across hematologic malignancies, particularly in lymphoma and leukemia. The agreement appears slightly stronger in blinded studies compared to open-label studies, though both show excellent concordance. The limited data in myeloma and combined lymphoma/leukemia subgroups prevent meaningful conclusions for these populations. These findings complement the PFS agreement results shown in Table 2, showing that both time-to-event (PFS) and response outcomes demonstrate high concordance between central and investigator assessments in hematologic malignancies.

Our findings carry critical implications for trial design and regulatory practice in hematologic malignancies. The robust concordance between BICR and investigator assessments in lymphoma and leukemia questions the universal applicability of BICR in these malignancies. Given the substantial financial burden and operational delays associated with central reviews—often exceeding 6–8 weeks for complex cases—our data suggest that rigorous local evaluations, guided by consensus criteria, may suffice for trials evaluating established endpoints in routine settings. This paradigm shift could redirect resources toward patient-centric endpoints (e.g., quality of life, MRD monitoring) without compromising scientific validity.

References

Cheson, B. D. et al. (2014). "Recommendations for initial evaluation, staging, and response assessment of Hodgkin and non-Hodgkin lymphoma: the Lugano classification". In: *Journal of Clinical Oncology* 32.27, pp. 3059–3068. DOI: 10.1200/jco.2013.54.8800.

Dancey, J. E. et al. (2009). "Recommendations for the assessment of progression in randomised cancer treatment trials". In: *European Journal of Cancer* 45.2, pp. 281–289. DOI: 10.1016/j.ejca.2008.10.042.

Dello Russo, C., N. Cappoli, and P. Navarra (2020). "A comparison between the assessments of progression-free survival by local investigators versus blinded independent central reviews in phase III oncology trials". In: *European Journal of Clinical Pharmacology* 76.8, pp. 1083–1092. DOI: 10.1007/s00228-020-02895-z.

Ford, R. et al. (2009). "Lessons learned from independent central review". In: *European Journal of Cancer* 45.2, pp. 268–274. DOI: 10.1016/j.ejca.2008.10.031.

Jacobs, F. et al. (2024). "Progression-free survival assessment by local investigators versus blinded independent central review in randomized clinical trials in metastatic breast cancer: A systematic review and meta-analysis". In: *European Journal of Cancer* 197, p. 113478. DOI: 10.1016/j.ejca.2023.113478.

Lian, Q. et al. (2024). "Meta-Analysis of 49 Roche Oncology Trials Comparing Blinded Independent Central Review (BICR) and Local Evaluation to Assess the Value of BICR". In: *Oncologist* 29.8, e1073–e1081. DOI: 10.1093/oncolo/oyad012.

Liang, W. et al. (2016). "Comparison of assessments by blinded independent central reviewers and local investigators: An analysis of phase III randomized control trials on solid cancers (2010–2015)". In: *Unpublished or Incomplete Reference*.

Lin, D. et al. (2021). "Circulating tumor cells: biology and clinical significance". In: *Signal Transduction and Targeted Therapy* 6.1, p. 404. DOI: 10.1038/s41392-021-00817-8.

Miller, A. et al. (2025). "Safety and Outcomes of Percutaneous Dilatational Tracheostomy in Patients with Hematologic Malignancies: A Retrospective Cohort Study". In: *Journal of Clinical Medicine* 14.2, p. 12. DOI: 10.3390/jcm14020657.

Narita, K. et al. (2018). "Quantification of bone-marrow plasma cell levels using various International Myeloma Working Group response criteria in patients with multiple myeloma". In: *International Journal of Hematology* 108.4, pp. 371–374. DOI: 10.1007/s12185-018-2489-0.

Rodriguez-Abreu, D., A. Bordoni, and E. Zucca (2007). "Epidemiology of hematological malignancies". In: *Annals of Oncology* 18 Suppl 1, pp. i3–i8. DOI: 10.1093/annonc/mdl443.

Sorigue, M., E. Cañamero, and M. D. Miljkovic (2021). "Systematic review of staging bone marrow involvement in B cell lymphoma by flow cytometry". In: *Blood Reviews* 47, p. 100778. DOI: 10.1016/j.blre.2020.100778.

Zhang, J., Y. Zhang, S. Tang, L. Jiang, et al. (2018). "Systematic bias between blinded independent central review and local assessment: literature review and analyses of 76 phase III randomised controlled trials in 45 688 patients with advanced solid tumour". In: *BMJ Open* 8.9, e017240. DOI: 10.1136/bmjopen-2017-017240.

Zhang, J., Y. Zhang, S. Tang, H. Liang, et al. (2017). "Evaluation bias in objective response rate and disease control rate between blinded independent central review and local assessment: a study-level pooled analysis of phase III randomized control trials in the past seven years". In: *Annals of Translational Medicine* 5.24, p. 481. DOI: 10.21037/atm.2017.11.24.