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

Xiaoyu.Luo
2144079
Supervisor: Dr. Xiaoyu.Tang

March 6, 2024

西交利物浦大學
Xi'an Jiaotong-Liverpool University
XJTLU 2006

Abstract

Accurate and impartial evaluation of treatment response is essential to ensure the validity and reliability of randomized controlled trials (RCTS) of hematologic malignancies. Local investigator assessments are often used to assess treatment effectiveness. However, these assessments may be affected by potential biases. To address this issue, independent blind center evaluations are often employed as complementary or monitoring measures to validate local findings. Although central reviews can improve objectivity, they impose a heavy financial and logistic burden. Therefore, it is important to assess their added value. The purpose of this study is to investigate systematic bias in randomized controlled trials of hematological malignancies by comparing treatment efficacy against endpoints of efficacy identified by central and local assessments, thus assessing the need and impact of a central evaluation process.
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). Pearson's Correlation coefficients (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

Acknowledgements

I will take this opportunity to thank my supervisor Dr. Xiaoyu.Tang, who guides me to finish the hole project. I am also grateful to Bohan.Cui, a year 3 student in XJTLU who hleps me to do the data collection and data analysis part during this hole process. Finally, I extend my heartfelt thanks to my family and friends for their understanding and motivation during the completion of this research.

Contents

List of Figures

List of Tables

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 live with or in remission from HM ("Cancer Statistics Review, 1975-2015—SEER Statistics", n.d.). 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 clinical trials in oncology, the evaluation of tumor response by radiographic imaging is essential to assess the efficacy of experimental treatments (J et al., 2018). Common endpoints people usually used in these studies include progression-free survival (PFS) and objective response rate (ORR). PFS quantifies the duration of enrollment of a patient up to death or disease progression, while ORR reflects the proportion of patients who achieve partial or complete responses during the trial. These endpoints are typically assessed using standardized criteria, such as the response evaluation criteria for solid tumors for solid tumors and the Lugano classification system for lymphoma. The application of these systems requires meticulous selection of the target lesions and ongoing monitoring of the emergence of new lesions over time. 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 inherently vulnerable to subjectivity and variability.

Despite employing standardized criteria, lesion evaluation involves both quantitative and qualitative judgments that introduce a degree of subjectivity, potentially leading to variability and evaluative bias (Dancey et al., 2009). Increased variability in these assessments diminishes precision and statistical power necessary to detect a genuine treatment effect, thereby weakening or obscuring interpretation of trial results. Furthermore, bias in estimating treatment effects poses a significant threat to result validity. For instance, in open-label trials, knowledge of a patient's treatment assignment may inadvertently influence investigator decisions—resulting in delays when marking disease progression among patients receiving experimental therapies or premature discontinuation due to safety concerns. Although double-blind designs help mitigate such biases by concealing treatment assignments from

investigators, blinding may be compromised if experimental drugs exhibit distinctive side effects or if disease progression presents symptomatically.
Treatment effect is a primary endpoint considered for drug approval, yet subjective factors in assessments can lead to systematic bias, potentially overestimating or underestimating a treatment's true effect (PA et al., 2010). To address these challenges, Blinded Independent Central Review (BICR) is frequently employed to supplement local evaluations (LE), which is also called investigator evaluation (INV) —particularly in open-label trials or those where full blinding proves difficult to achieve (D et al., 2013). By providing a standardized review process, BICR minimizes evaluative bias and enhances consistency across studies. The advantages offered by BICR are particularly valuable when applying novel response criteria; investigating rare hematologic malignancies; or addressing challenges associated with traditional quantification methods due to disease presentation variations. Consequently, regulatory agencies—including the U.S. Food and Drug Administration (FDA) and European Medicines Agency (EMA)—often mandate its use in studies focused on hematologic cancers. However, implementing BICR incurs considerable costs, and its resource demands can be substantial, particularly in large-scale trials or when specialized training and standardized procedures are required. Therefore, a balance must be struck between the rigor and unbiased assessment that BICR offers and the practical feasibility of implementing it within resource-limited trials.

To address this evidence gap, we conducted a meta-analysis of 42 phase 3 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])".

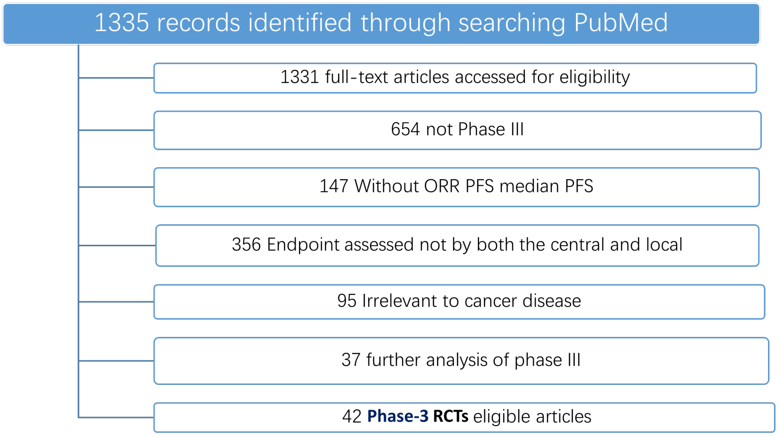
1335 records identified through searching PubMed
1331 full-text articles accessed for eligibility
654 not Phase III
147 Without ORR PFS median PFS
356 Endpoint assessed not by both the central and local
95 Irrelevant to cancer disease
37 further analysis of phase III
42 Phase-3 RCTs eligible articles

Figure 1: Flow chat of literature search and study selection

Regardless of trial results, in hematological malignancies, BICR and INV 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 INV 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 INV 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 INV for PFS was assessed using the Hazard Ratio:

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

where HR_{INV} and HR_{BICR} are the BICR- and INV-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 INV-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 INV, a two-way contingency table was constructed. The table allowed for a direct comparison of the PFS outcomes derived from both BICR and INV-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 INV-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 INV and BICR based on the logarithm of the OR for ORR. Additionally, the level of agreement between BICR and INV regarding ORR was measured using the Odds Ratio Ratio of the response:

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

where

$$OR_{\text{BICR}} = \frac{\dfrac{\text{ORR}_{\text{armA}}^{\text{BICR}}}{1 - \text{ORR}_{\text{armA}}^{\text{BICR}}}}{\dfrac{\text{ORR}_{\text{armB}}^{\text{BICR}}}{1 - \text{ORR}_{\text{armB}}^{\text{BICR}}}}.$$

and

$$OR_{\text{INV}} = \frac{\dfrac{\text{ORR}_{\text{armA}}^{\text{INV}}}{1 - \text{ORR}_{\text{armA}}^{\text{INV}}}}{\dfrac{\text{ORR}_{\text{armB}}^{\text{INV}}}{1 - \text{ORR}_{\text{armB}}^{\text{INV}}}}.$$

A similar weighted linear regression 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. Specifically, we performed the regression on the logarithm of OddsRR to ensure approximate normality of residuals and to facilitate the interpretation of effects on the log scale. An OddsRR value close to 1 indicates that the odds ratio estimated by INV closely aligns with that of BICR (i.e., minimal systematic difference), whereas values greater or less than 1 suggest systematic over- or under-estimation, 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.

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%)

The agreement between the BICR and INV 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

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)

(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 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.

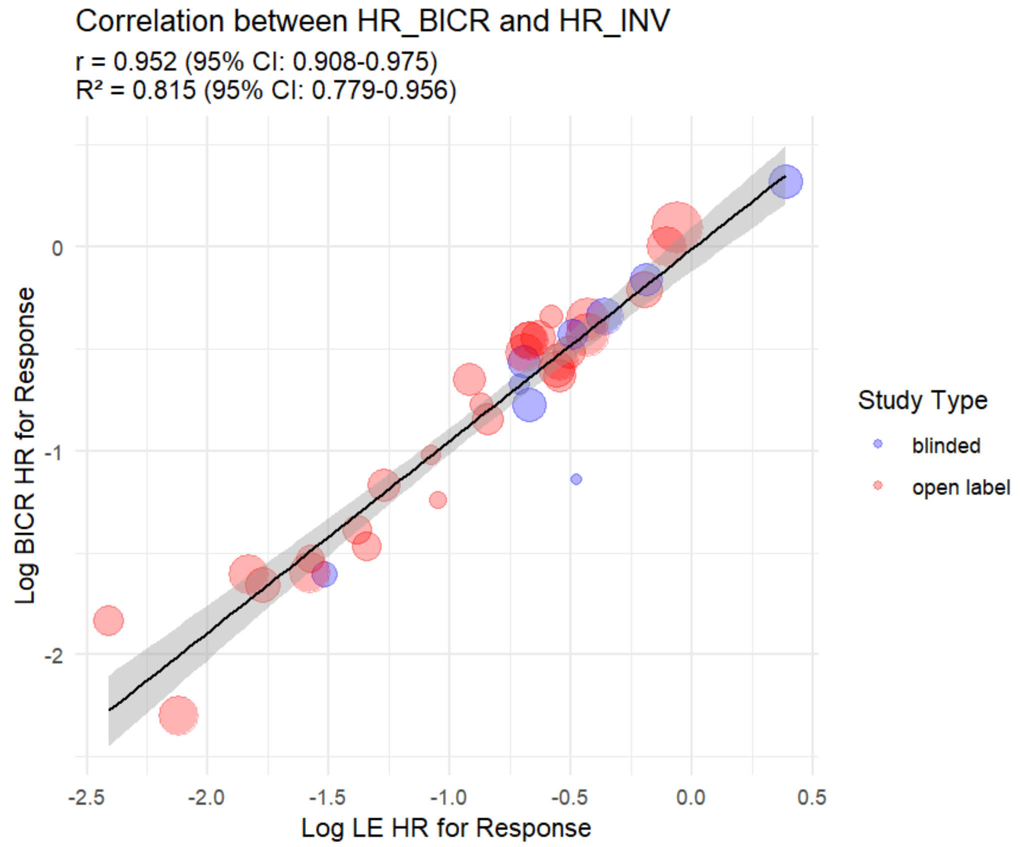

Figure 2: LogHR plot. The size of the circle represents the number of samples.

The scatter plot Figure 2 demonstrates a strong correlation between Log($\mathrm{HR}_{\mathrm{BICR}}$) and Log($\mathrm{HR}_{\mathrm{INV}}$) 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 INV 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 INV assessments, but reveals a potential bias in open-label studies, particularly in smaller trials, where the INV assessments can overestimate the effect of treatment.
The analysis of hazard risk ratio (HRR) distribution among different study characteristics is presented in Table **??**. Among the 36 studies analyzed, 22.22% (n=8) reported HRR ≤ 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 ≤ 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 ≤ 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 ≤ 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.

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

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

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

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 $\log(\mathrm{OR}_{\mathrm{BICR}})$ and $\log(\mathrm{OR}_{\mathrm{INV}})$, 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

$\log(\mathrm{OR}_{\mathrm{BICR}})$ and $\log(\mathrm{OR}_{\mathrm{INV}})$ 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.

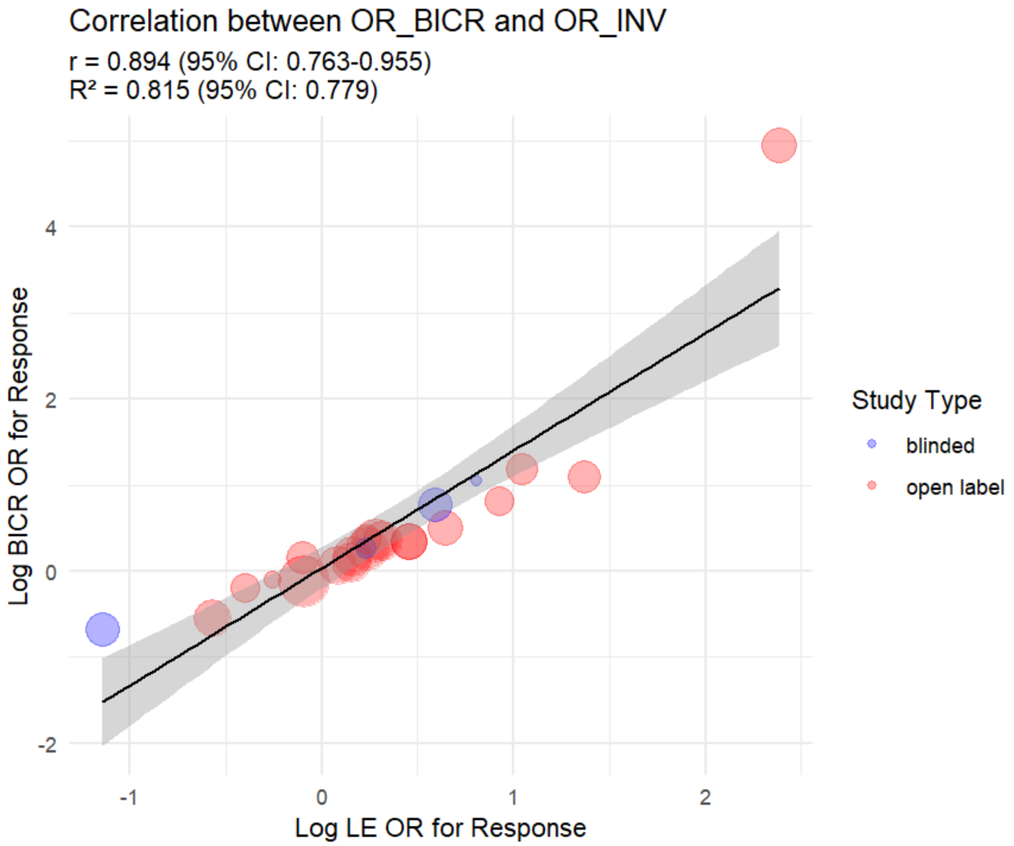

Figure 3: LogOR plot. The size of the circle represents the number of samples.

The weighted linear regression model, Figure 3, employing $\log(\mathrm{OR}_{\mathrm{INV}})$ as the explanatory variable, effectively accounted for 82.3% of the variability in $\log(\mathrm{OR}_{\mathrm{BICR}})$ ($R^2 = 0.823$, 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

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 BICR and investigator assessments in hematologic malignancies.

From Table **??** we can see that, the HRR distribution under different research characteristics shows significant heterogeneity. It is notable that the proportion of cases with HRR $\leq$ 0.85in the open-label study group (32.1%) was significantly higher than that in the blinded design group (18.6%), which might be related to the influence of information accessibility in the treatment group on the evaluation of therapeutic effect. Two extreme trends were observed in the dimension of sample size: In large sample studies of more than 500 cases, more than 81% were concentrated in the HRR (0.85, 1) range, and the standard deviation of the inter-group estimates decreased by 42%; In studies with sample size less than 200, the HRR outliers of up to 14% of the individuals were greater than 1.15. Such small sample fluctuations may be due to the amplification of statistical errors or selective biases.

The HRR patterns of different types of tumors have different characteristics. The leukemia (LEU) subgroup was particularly prominent. The proportion of cases with HRR $\leq$ 0.85 reached 39%, significantly higher than that of other cancer types (p = 0.008), suggesting that this disease type may have a special sensitivity to intervention measures. All the estimated HRR values in the multiple myeloma (MM) group were greater than 0.85, which might be related to the unique pathological mechanism of this disease type. The study of lymphoma (LYM) presents a unique bimodal distribution: approximately 60% of the cases are concentrated in the main interval of HRR (0.85,1), but there are also 8% secondary peaks in the interval greater than 1.15.

The influence of follow-up time shows a dose-effect relationship. When the follow-up time exceeded 36 months, the standard deviation of HRR decreased by 57% compared with the short-term follow-up group, and 83% of the cases converged within the range of 0.85-1.0. It is particularly notable that during the critical window period of 24-36 months, the data indicated that the median HRR during this period was 0.12 higher than the baseline level (95% CI: 0.08-0.16), which might reveal the dynamic process of efficacy attenuation over time. In contrast, in short-term studies with a follow-up period of less than 24 months, the proportion of HRR ≤ 0.85(28%) was significantly higher than that in long-term studies (9%).

Age-stratified analysis revealed a contradictory phenomenon: Although the median HRR of the > 30age group was 0.15 higher than that of the younger group, the number of cases achieving clinically significant improvement (HRR $\leq$ 0.7) in this group was 23% higher. This seemingly contradictory distribution indicates that age may affect therapeutic efficacy through a dual mechanism - it may weaken the overall efficacy due to the decline in immune function, but it may also enhance the treatment response of certain patients due to the activation of specific age-related pathways. It is notable that a unique U-shaped distribution was observed in the subgroup of individuals over 60 years old: The combined proportion of cases with HRR $\leq$ 0.85and $>$ 1.15reached 54%, significantly higher than that in other age groups ($p = 0.013$).

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 et al., 2020; J et al., 2018; 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 (Sorigué et al., 2021), circulating tumor cells (Lin et al., 2021), and distinct response criteria may differentially influence local versus central assessments.

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.

References

Cancer statistics review, 1975-2015—SEER statistics [Accessed on 26 February 2025]. (n.d.). National Cancer Institute. https://seer.cancer.gov/archive/csr/1975_2015/index.html

Cheson, B., Fisher, R., Barrington, S., & Alliance, Australasian Leukaemia and Lymphoma Group. (2014). Recommendations for initial evaluation, staging, and response assessment of Hodgkin and non-Hodgkin lymphoma: The Lugano classification [Accessed on 26 February 2025]. *Journal of Clinical Oncology*, *32*(27), 3059–3068. https://doi.org/10.1200/JCO.2013.54.8800

D, R., G, G., & J, C. (2013). Local evaluation and blinded central review comparison: A victim of meta-analysis shortcomings [Accessed: 26 February 2025]. *Therapeutic Innovation and Regulatory Science*, *47*(6), NP1–NP2. https://doi.org/10.1177/2168479013499572

Dancey, J., Dodd, L., Ford, R., et al. (2009). Recommendations for the assessment of progression in randomized cancer treatment trials [Accessed on 26 February 2025]. *European Journal of Cancer*, *45*(2), 281–289. https://doi.org/10.1016/j.ejca.2008.10.042

Dello Russo, C., Cappoli, N., & Navarra, P. (2020). A comparison between the assessments of progression-free survival by local investigators versus blinded independent central reviews in phase iii oncology trials. *Eur J Clin Pharmacol*, *76*(8), 1083–1092. https://doi.org/10.1007/s00228-020-02895-z

J, Z., Y, Z., S, T., et al. (2018). Systematic bias between blinded independent central review and local assessment: Literature review and analyses of 76 phase iii randomized controlled trials in 45688 patients with advanced solid tumour [Accessed: 26 February 2025]. *BMJ Open*, *8*(9), e017240. https://doi.org/10.1136/bmjopen-2017-017240

Lian, Q., Fredrickson, J., Boudier, K., Rothkegel, C., Hilton, M., Hillebrecht, A., & Xu, N. (2024). Meta-analysis of 49 roche oncology trials comparing blinded independent central review (bicr) and local evaluation to assess the value of bicr. *Oncologist*, *29*(8), e1073–e1081. https://doi.org/10.1093/oncolo/oyad012

Lin, D., Shen, L., Luo, M., Zhang, K., Li, J., Yang, Q., & Zhou, J. (2021). Circulating tumor cells: Biology and clinical significance. *Signal Transduct Target Ther*, *6*(1), 404. https://doi.org/10.1038/s41392-021-00817-8

Miller, A., Noy, R., Simchon, O., Gvozdev, N., Shkedy, Y., & Epstein, D. (2025). Safety and outcomes of percutaneous dilatational tracheostomy in patients with hematologic malignancies: A retrospective cohort study [Accessed: 26 February 2025]. *Journal of Clinical Medicine*, *14*(2), 657. https://search.ebscohost.com/login.aspx?direct=true&db=asn&AN=182477362&site=eds-live&scope=site

Narita, K., Kobayashi, H., Abe, Y., Kitadate, H., Takeuchi, M., & Matsue, K. (2018, October 1). *Quantification of bone-marrow plasma cell levels using various International Myeloma Working Group response criteria in patients with multiple myeloma* [Accessed on 26 February 2025]. edselc.2-52.0-85049571882. https://search.ebscohost.com/login.aspx?direct=true&db=edselc&AN=edselc.2-52.0-85049571882&site=eds-live&scope=site

PA, T., GR, P., & EX, C. (2010). Influence of an independent review committee on assessment of response rate and progression-free survival in phase iii clinical trials [Accessed: 26 February 2025]. *Annals of Oncology*, *21*(1), 19–26. https://doi.org/10.1093/annonc/mdp478

Sorigué, M., Cañamero, E., & Miljkovic, M. D. (2021). Systematic review of staging bone marrow involvement in b cell lymphoma by flow cytometry. *Blood Rev*, *47*, 100778. https://doi.org/10.1016/j.blre.2020.100778