Neuro-Oncology Advances

3(1), 1–10, 2021 | doi:10.1093/noajnl/vdab007 | Advance Access date 15 February 2021

Extent of radiological response does not reflect survival in primary central nervous system lymphoma

Matthijs van der Meulen, Alida A. Postma, Marion Smits, Katerina Bakunina, Monique C. Minnema, Tatjana Seute, Gavin Cull, Roelien H. Enting, Marjolein van der Poel, Wendy B.C. Stevens, Dieta Brandsma, Aart Beeker, Jeanette K. Doorduijn, Samar Issa, Martin J. van den Bent, and Jacoline E.C. Bromberg

Department of Neuro-Oncology, Erasmus MC Cancer Institute, Brain Tumor Center, University Medical Center Rotterdam, Rotterdam, The Netherlands (M.v.d.M., M.J.v.d.B., J.E.C.B.); Department of Radiology and Nuclear Medicine, Maastricht University Medical Center, School for Mental Health and Sciences, Maastricht, The Netherlands (A.A.P.); Department of Radiology and Nuclear Medicine, Erasmus MC Cancer Institute, Brain Tumor Center, University Medical Center Rotterdam, Rotterdam, The Netherlands (M.S.); Department of Hematology, HOVON Data Center, Erasmus MC Cancer Institute, Rotterdam, The Netherlands (K.B.); Department of Hematology, University Medical Center Utrecht, Utrecht, The Netherlands (M.C.M.); Department of Neurology, University Medical Center Utrecht, Utrecht, The Netherlands (T.S.); Haematology Department, Sir Charles Gairdner Hospital and PathWest Laboratory Medicine, Nedlands, Australia (G.C.); University of Western Australia, Crawley, Australia (G.C.); Department of Neurology, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands (R.H.E.); Department of Hematology, University Medical Center, Maastricht, The Netherlands (M.v.d.P.); Department of Hematology, Radboud University Medical Center, Nijmegen, The Netherlands (W.B.C.S.); Department of Neuro-Oncology, Netherlands Cancer Institute-Antoni van Leeuwenhoek, Amsterdam, The Netherlands (D.B.); Department of Hematology, Spaarne Gasthuis, Haarlem, The Netherlands (A.B.); Department of Hematology, Erasmus MC Cancer Institute, University Medical Center Rotterdam, Rotterdam, The Netherlands (J.K.D.); Department of Haematology, Middlemore Hospital, Auckland, New Zealand (S.I.)

Corresponding Author: Matthijs van der Meulen, MD, Department of Neuro-Oncology, Erasmus MC Cancer Institute, University Medical Center Rotterdam, Dr. Molewaterplein 40, 3015 GD, Rotterdam, The Netherlands (m.vandermeulen.2@erasmusmc.nl).

Abstract

Background. In primary central nervous system lymphoma (PCNSL), small enhancing lesions can persist after treatment. It is unknown whether a difference in response category (complete response [CR], complete response unconfirmed [CRu], or partial response [PR]) reflects survival. We aimed to determine the value of a central radiology review on response assessment and whether the extent of response influenced progression-free and/or overall survival.

Methods. All patients in the HOVON 105/ALLG NHL 24 study with at least a baseline MRI and one MRI made for response evaluation available for central review were included. Tumor measurements were done by 2 independent central reviewers, disagreements were adjudicated by a third reviewer. Crude agreement and interobserver agreement (Cohen's kappa) were calculated. Differences in progression-free and overall survival between different categories of response at the end-of-protocol-treatment were assessed by the log-rank test in a landmark survival-analysis.

Results. Agreement between the central reviewers was 61.7% and between local and central response assessment was 63.0%. Cohen's kappa's, which corrects for expected agreement, were 0.44 and 0.46 (moderate), respectively. Progression agreement or not was 93.3% (kappa 0.87) between local and central response assessment. There were no significant differences in progression-free and overall survival between patients with CR, CRu, or PR at the end-of-protocol-treatment, according to both local and central response assessment.

Conclusions. Reliability of response assessment (CR/CRu/PR) is moderate even by central radiology review and these response categories do not reliably predict survival. Therefore, primary outcome in PCNSL studies should be survival rather than CR or CR/CRu-rate.

Key Points

- Reliability of response assessment is moderate in PCNSL.
- The extent of response at the end-of-treatment does not reflect survival.
- Progression-free survival and overall survival are more reliable endpoints than complete response rate.

Importance of the Study

In primary central nervous system lymphoma (PCNSL), small enhancing lesions can persist after treatment. It is unknown whether a difference in response category reflects survival. We calculated the interobserver agreement between 2 central reviewers and between local and central review. Then, in a landmark-analysis, we compared progression-free (PFS) and overall survival (OS) between patients with different responses at end-of-protocol-treatment. Interobserver agreement was excellent (kappa 0.87) in defining progression versus no progression. Interobserver agreement for each response category was moderate (kappa 0.46) and similar for local and central response assessment. The added value of central radiology review in clinical studies on PCNSL therefore seems limited. Furthermore, no significant differences were found in PFS and OS between patients categorized as complete response, complete response unconfirmed, or partial response. This suggests that radiological response is not an adequate surrogate endpoint for survival in PCNSL, and studies should have survival as the primary endpoint.

Primary central nervous system lymphoma (PCNSL) is a rare non-Hodgkin lymphoma confined to the brain, leptomeninges, spinal cord, and eyes without manifestations outside the central nervous system. For response assessment in PCNSL the criteria from the International Primary CNS Lymphoma Collaborative Group (IPCG), are commonly used.[1] These response criteria are based on radiological, ophthalmologic, and spinal fluid cytology examination, and the use of corticosteroids. The MRI response evaluation defines the following categories: complete response (CR): no signs of abnormal gadolinium-based contrast agent enhancement, complete response unconfirmed (CRu): a small but persistent contrast enhancement abnormality likely related to biopsy or focal hemorrhage, partial response (PR): a reduction of ≥50% of the contrast-enhancing lesion, stable disease (SD): <50% reduction and ≤25% increase of the contrast-enhancing lesion, progressive disease (PD): >25% increase in contrast-enhancing lesion, relapse: a new contrast-enhancing lesion after prior CR or CRu.[1] These response criteria do not take nonenhancing lesions into account. Recent findings and earlier reports suggest that these lesions might, however, be considered as tumor as well.[2,3]

The correlation of radiological response with survival endpoints (progression-free survival [PFS] and overall survival [OS]) is uncertain: one study showed that patients with a CR at the end of induction chemotherapy had a better OS than those who did not reach CR, but in this study PR, SD, and progression were combined.[4] In another study highly variable outcomes were found in patients with PR at the end-of-treatment,[2] and a third study did not show a survival difference between those who attained CR compared to those who did not reach CR at the end of induction treatment.[5] Thus, it is questionable whether in PCNSL the extent of radiological response is relevant for predicting OS, the golden endpoint in oncology studies. It is also unclear whether interobserver variation exists in assessing response in PCNSL, which if present, will affect the reliability of that endpoint.

In the HOVON 105/ALLG NHL 24 trial, the primary endpoint was event-free survival (EFS). Events were defined as "not reaching complete response" or "complete response unconfirmed at the end-of-treatment," or "progression or death after response."[6] Because event-free survival includes a radiological evaluation as endpoint (ie, achieving CR or CRu), based on local assessment, central MRI review is important for the trial analysis. The aim of the present study was to review the local assessment by central radiology review and to assess whether CR, CRu, and PR reflect PFS and OS. In addition, we evaluated the relevance of nonenhancing lesions at baseline and after treatment.

Methods

Patient Selection

The HOVON 105/ALLG NHL 24 study is a phase III randomized controlled trial, in which between 2010 and 2016 199 patients were recruited from Dutch, Australian, and New Zealand hospitals. The treatment protocol and primary outcome results have been published before.[6] In short, immunocompetent patients with a newly diagnosed, CD20 positive B-cell PCNSL aged 18–70 years with WHO/ECOG performance status 0–3 were included. Patients were randomized for 2 courses of high-dose methotrexate (HD-MTX)-based chemotherapy: methotrexate, teniposide, BCNU, and prednisolone (MBVP) versus MBVP with rituximab (R-MBVP). This was followed by HD-cytarabine (Ara-C) chemotherapy. Patients ≤60 years-old only subsequently received 30Gy whole-brain radiotherapy (WBRT). A simultaneous focal boost of 10Gy was given to the original enhancing tumor in patients who only achieved PR.

Patients were included for the central MRI review if they gave informed consent for central radiology review, and if a baseline MRI as well as at least one follow-up MRI was available for central review. Additionally, a measurable brain lesion had to be present at baseline in order to be able to assess response. Patients were excluded if only a CT was available at baseline and/or if only CTs were used for evaluation at subsequent time points. The baseline MRI had to have been made within 21 days before initiation of protocol treatment. Those patients with a progressive disease or a relapse outside the brain parenchyma were excluded from the time of progression and for the landmark analysis.

The study was approved by the ethics committee of all participating centers. All participants signed informed consent for the randomized controlled trial and separately for the central radiology review.

Radiological Follow-up

According to protocol, MRI evaluations were performed before the initiation of chemotherapy (baseline), after the second (R-)MBVP course, after Ara-C, and after WBRT, if applicable. Follow-up MRIs were made every 3 months in the first 2 years after treatment, followed by every 6 months up to 5 years after treatment and yearly thereafter.

At least the following MRI sequences were performed: an axial T1 weighted scan before and after gadolinium-based contrast agent administration, and an axial T2 weighted and/or fluid-attenuated inversion recovery (FLAIR) scan. If locally possible additional sagittal or coronal T1 weighted scans with gadolinium-based contrast agent administration were also performed or reconstructed. All MR images were acquired on a 1.5 or 3.0 Tesla scanner.

Central Radiology Review

Scans made at baseline and after each treatment component was centrally reviewed to evaluate response of the tumor to treatment. In case of relapse or progression, the brain MRI on which this was diagnosed according to the local physician, as well as the last MRI made before progression were also centrally reviewed, to verify progression and to make sure true progression had not occurred earlier than locally ascertained. Scans made in follow-up were not reviewed if response was not changed according to the local evaluation. PD was defined as relapse or progression at any site (brain, spinal cord, cerebrospinal fluid, or eyes). In case progression was located outside the brain parenchyma, the brain MRI was not included in the central radiology review.

At the end of the study, all MR images were submitted for review on DVD or CD and stored on a secured central server. Except for the baseline scan, locally assessed response rates were collected for all MRIs performed for this study. Local physicians were not blinded for treatment arm and/or other clinical information.

Central evaluation of response was performed retrospectively, in parallel by 2 reviewers (M.M. and A.A.P.). In case of disagreement on the response between these reviewers, an adjudicator (M.S.) finalized the central response category. The central reviewers and the adjudicator were blinded for study-arm and clinical information.

Single evaluations were excluded if (1) both central reviewers considered an MRI not assessable, (2) no MRI with gadolinium-based contrast agent administration was made, or (3) the MRI was made outside ±3 weeks around the planned evaluation moment during the treatment period.

MRI Tumor Measurement

For all enhancing lesions, the largest diameters on the axial post-gadolinium-based contrast agent T1 weighted images were measured as well as their perpendicular diameter on the same slice. The product of these measurements was used to define the size of the tumor. In case of multiple lesions, response assessment was based on the sum of all products, up to a maximum of 4 lesions.

Nonenhancing space-occupying lesions were measured by one of the central reviewers (M.M.) on FLAIR images if possible, and otherwise on the T2 weighted images. In patients experiencing recurrent disease, the localization of the recurrence was compared to the localization of the initial nonenhancing and enhancing lesions.

Landmark Analysis

To estimate the survival probability for responding patients in the different response categories (CR, CRu, or PR) in an unbiased way, a landmark analysis was performed.[7] Regardless of the type of last administered treatment on protocol (ie, MBVP, Ara-C, or WBRT), the response at the end-of-protocol-treatment was related to PFS and OS for those still at risk at that timepoint. PFS was defined as time from randomization to progression, relapse, or death from any cause, whichever came first. OS was defined as time from randomization to death from any cause. Patients still alive at the date of last contact were censored. Follow-up data were available up to October 1, 2019. In this landmark

analysis, all patients alive at the landmark timepoint who had an MRI at end-of-protocol-treatment and were classified as CR, CRu, or PR on that MRI were included. Thus patients who had less than PR at the end-of-treatment had relapsed (for PFS only), died, or did not have their end-of-treatment scan before the landmark were excluded from landmark analyses.

The reference time point (landmark) was set between 4 weeks after last treatment, but before the first follow-up MRI (ie, 3 months after end-of-protocol-treatment), in such a way that most patients could be included. The landmark analysis was performed for both local and central response assessment.

Statistical Analysis

Since the HOVON 105/ALLG NHL 24 study showed no differences in EFS, PFS, or OS between the 2 treatment arms we analyzed both arms together.[6] For the interobserver agreement between the central reviewers, and between central and local response evaluations we calculated the crude agreement and Cohen's kappa,[8] in which crude agreement is corrected for expected agreement (ie, the agreement that would occur "by chance"). First, interobserver agreement was assessed for all response categories separately; second, agreement was assessed for combined categories CR/CRu and PD/relapse, and third for 3 categories: response (CR/CRu/PR), SD, and progression (PD/relapse). Lastly, the crude agreement and kappa's interobserver agreement for progression versus no progression were calculated on the MRI on which progression was diagnosed and on the preceding MRI.

In the landmark analysis, the survival curves for PFS and OS were constructed using the Kaplan-Meier method for the different categories of response (ie, CR, CRu, and PR) according to central and to local response evaluation at the end-of-protocol-treatment. Differences by response were assessed with a log-rank test with a 5% significance level. All analyses were performed with Stata, version 15.0.

Results

Of the 199 trial patients, 115 were included in this study. Three patients were excluded because they did not give informed consent for the radiology review, 12 patients because no baseline MRI was present and 3 for whom only CT was available, 61 patients were excluded because baseline MRI was made outside the predefined time window, and 5 for other reasons (see CONSORT diagram, Figure 1). The median age of patients included in this study was 61 years (range: 38–70), 44% were female, and 73% had WHO performance score <2 see Table 1. On October 1, 2019 (last follow-up), in the central radiology review cohort 45 patients were alive without progression, and 15 were alive with progression.

Of these 115 included patients, 396 scans were centrally reviewed: 115 baseline MRIs, 235 after treatment, and 46 PD or last before PD scans. Scans were excluded if they were not received for central review (n = 154), were made outside the predefined time window (n = 7), or progression was not located in the brain parenchyma (n = 14).

Central Radiology Review

For the MRIs made during treatment (n = 235) the agreement between central reviewers 1 and 2 and between local and central response assessment was higher than the expected agreement by chance (P < .001). Between the central reviewers, the agreement for all response categories was 61.7%, with a kappa of 0.44, see Table 2. After adjudication, if necessary, the agreement between local and central response assessment was 63.0% with a kappa of 0.46, see Table 3.

When CR and CRu were combined into one category, and the categories PD and relapse were combined, the interobserver agreement and kappa values increased, but the latter remained in the moderate range. Between reviewers 1 and 2, agreement increased to 77.0% (kappa 0.57), and between local and central assessment agreement improved to 74.5% (kappa 0.54), see Supplementary Table S1 and Supplementary Table S2, respectively. When response categories were classified into response (CR, CRu, or PR), stable disease, or progression (PD or relapse) the agreement increased to 95.3% (kappa 0.40) between the central reviewers and 94.9% (kappa 0.41) between local and central response assessment. The kappa remained relatively low, because of the increased expected agreement.

The response assessment for the MRIs on which progression or relapse was diagnosed and the last MRI made before progression were analyzed separately. Progression could take place during treatment or follow-up; only patients with both PD and last before PD scans provided for central review were included in this analysis. Agreement on whether there was progression or relapse versus "no progression" was 96.7% between central reviewers 1 and 2 (kappa 0.93) and 93.3% between the local and central response assessment (kappa 0.87), both were significantly higher than expected agreement (P < .001), Table 4A and 4B.

Landmark Analysis on CR, CRu, and PR

In total 91 "end-of-protocol-treatment" MRIs were available and were locally and centrally assessed. The landmark, aiming to include as many patients as possible after the end-of-protocol-treatment MRI but before first follow-up MRI, was positioned at 6.9 months after randomization. Only those with a CR, CRu, or PR at the end-of-protocol-treatment were included in this analysis. Two patients who had not had their end-of-protocol-treatment scan yet were excluded. For the PFS analysis, we also excluded those who had less than PR (n = 7 according to central response and 8 according to the local response assessment) at the end-of-protocol-treatment, had progression before the landmark (n = 9 in central and n = 6 in local response assessment) or died without progression (n = 1). For the OS analysis, we excluded those who had less than PR (see above) or died (n = 2). Since we analyzed survival according to local as well as to central response assessment, the number of patients the analyses were based on differed: survival analysis was performed on 72 (central assessment) and 74

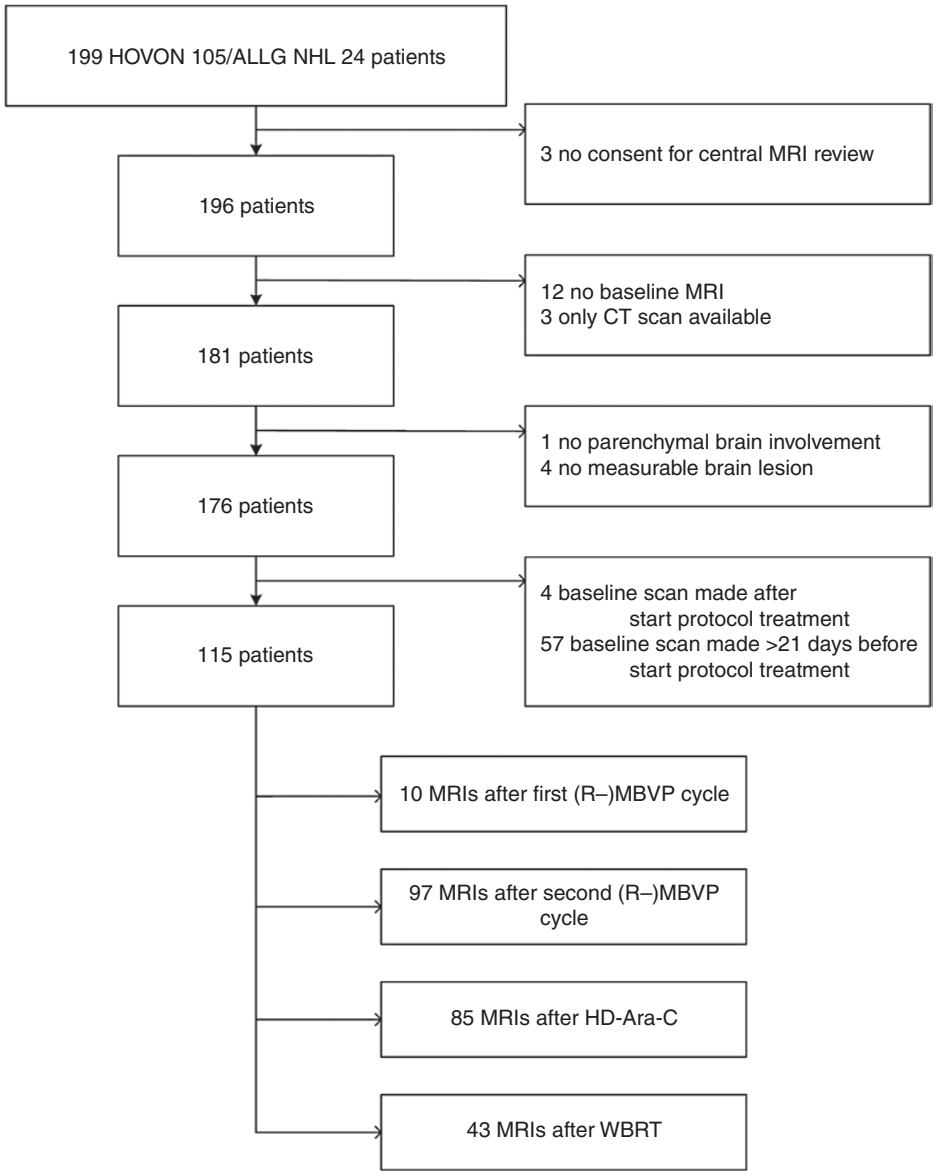

Figure 1. CONSORT diagram. 235 MRIs from 115 patients were assessed. (R-) MBVP = (rituximab), methotrexate, teniposide, BCNU and prednisolone, HD-Ara-C = high-dose cytarabine, WBRT = whole-brain radiotherapy.

(local assessment) patients for PFS and 80 (central assessment) and 79 (local assessment) patients for OS.

There was no statistically significant difference in PFS (Figure 2A and B) between those judged as CR, CRu, or PR, both in central response assessment ($P = .97$) and according to local judgment ($P = .76$). Similar results were found for overall survival (Figure 2C and D), for central ($P = .69$) and local ($P = .16$) response assessment. There were no significant differences in extent of response between those who received WBRT and those who did not, neither in the PFS and the OS analyses (Supplementary Table S4).

Nonenhancing Lesions

At baseline 7 patients were identified with nonenhancing space-occupying lesions. Baseline characteristics in these patients were similar to the total trial population, Supplementary Table S3. After chemotherapy, in 5 of the 7 patients, the lesions diminished with ≥50%.

Four of these patients relapsed, in 2 patients this was at the same location as the original enhancing lesion. None of the patients had a relapse at the location of the nonenhancing lesions.

Table 1. Baseline Characteristics of the Patients Included in This Study, Those Who Were Excluded, and for the Total Study Population

	Included Patients $n = 115$	Excluded Patients $n = 84$	Total $n = 199$
Sex (*n*, % males)	64 (56%)	45 (54%)	109 (55%)
Age (median, range)	61 (38–70)	61 (26–70)	61 (26–70)
WHO performance score (*n*, %)			
0	27 (23%)	16 (19%)	43 (22%)
1	57 (50%)	44 (53%)	101 (51%)
2	17 (15%)	17 (20%)	34 (17%)
3	14 (12%)	7 (8%)	21 (10%)
Comorbidities ($n > 2$, %)	60 (52%)	44 (52%)	104 (52%)
Solitary lesions (*n*, %)	66 (57%)	37 (44%)	103 (52%)
Missing/ NA	1 (1%)	18 (21%)	19 (10%)
Deep lesion (*n*, %)	83 (72%)	42 (50%)	125 (63%)
Periventricular (*n*, %)	61 (53%)	35 (42%)	96 (48%)
Basal ganglia (*n*, %)	8 (7%)	6 (7%)	14 (7%)
Cerebellar (*n*, %)	22 (19%)	8 (10%)	30 (15%)
Brain stem (*n*, %)	10 (9%)	2 (2%)	12 (6%)
Spinal (*n*, %)	2 (2%)	–	2 (1%)
Lobar (*n*, %)	58 (50%)	37 (44%)	95 (48%)
Study drug exposure			
High-dose cytarabine (*n*, %)	98 (85%)	63 (75%)	161 (81%)
WBRT (*n*, %)	48 (42%)	22 (26%)	70 (35%)
Radiation boost given (*n*, %)	24 (21%)	15 (18%)	39 (20%)
Intrathecal treatment given (*n*, %)	12 (10%)	4 (5%)	16 (8%)

NA = not applicable in case of no brain lesion; WBRT = whole-brain radiotherapy.

Table 2. Level of Agreement Between Central Reviewer 1 and Central Reviewer 2 in All 235 Scans Made After Each Treatment Module for All Response Categories

	Reviewer 2						
Reviewer 1	CR	CRu	PR	SD	PD	Relapse	Total
CR	32	34	11	1	0	0	78
CRu	0	38	15	0	0	1	54
PR	0	17	74	4	0	0	95
SD	0	0	2	1	1	0	4
PD	0	0	0	1	0	1	2
Relapse	0	0	1	0	1	0	2
Total	32	89	103	7	2	2	235

Agreement 62%, kappa 0.44. CR = complete response; CRu = complete response unconfirmed; PR = partial response; SD = stable disease; PD = progressive disease

Discussion

We found an excellent crude agreement (96.7%) and kappa score (0.93) between the central reviewers and between local and central radiological evaluations (crude agreement 93.3%, kappa 0.86) in differentiating progression from no progression. However, for response assessment after treatment, interobserver agreement was moderate at best. Furthermore, the crude interobserver agreement (62%) and kappa statistics between the 2 central reviewers and between local and central radiology response assessment after each treatment component ($n = 235$) were almost identical (local

Table 3. Level of Agreement Between Local And Central Assessment in All 235 Scans Made After Each Treatment Module for All Response Categories

	Local						
Central	CR	CRu	PR	SD	PD	Relapse	Total
CR	42	12	8	0	0	1	63
CRu	15	28	30	0	0	0	73
PR	1	9	74	1	0	1	86
SD	0	0	4	1	1	0	6
PD	0	0	2	0	3	0	5
Relapse	0	0	2	0	0	0	2
Total	58	49	120	2	4	2	235

Agreement 63%, kappa 0.46. CR = complete response; CRu = complete response unconfirmed; PR = partial response; SD = stable disease; PD = progressive disease.

Table 4. Level of Agreement (A) Between Both Central Reviewers: Agreement 96.7%, Kappa 0.93 and (B) Between Local and Central Assessment: Agreement 93.3%, Kappa 0.87 in All Scans Which Confirmed PD and Made "Last Before PD"

A	Reviewer 2			B	Local		
Reviewer 1	No PD	PD		Central	No PD	PD	
No PD	16	1	17	No PD	14	1	15
PD	0	13	13	PD	1	14	15
	16	14	30		15	15	30

PD = progressive disease, including relapses.

vs central kappa 0.46 and both central reviewers 0.44). This illustrates the difficulty of defining and delineating residual abnormalities after treatment and suggests that there is little added value of a central radiology review in PCNSL patients. Crude interobserver agreement increased when response categories were combined, but the kappa statistics remained in the range of moderate agreement. This is most likely due to increased expected agreement since Cohen's kappa statistic is the agreement found, corrected for expected agreement to occur by chance. Thus, our data show that although the presence of response is well agreed upon, judgement regarding the extent of response is less reliable. This suggests, together with the excellent agreement regarding the moment of progression, that PFS and OS are more reliable endpoints than specific and more detailed response categories and that they also better reflect patient benefit. It is not clear whether the low predictive value of the categories of response is the result of the interobserver variability in scoring, whether, as in systemic lymphoma, some but not all residual abnormalities represent active disease, or whether response is actually a continuum in which the categories are artificial and do not truly represent a different prognostic value.

But because it is clinically relevant to differentiate responding form not responding patients we would advocate to simplify the response criteria into response (decrease of the enhancing lesions of >50%), stable disease, or progression.

To the best of our knowledge, a central radiology review in PCNSL with assessment of the interobserver agreement has not been described before. Several studies assessed interobserver agreement in glioma patients.[9–13] Our excellent agreement on PD versus no PD contrasts with the interobserver agreement in standard radiology assessment for progression in glioma.[9,12]This might be explained by the rapid evolution of most PCNSL and its easily recognizable appearance on the MR images: PCNSL, at relapse or progression as well as primary presentation, generally appears as a homogenously enhancing, circumscribed space-occupying lesion, rather than the ill-defined mass and irregular enhancement in high-grade glioma.

In our landmark analyses, we found no difference in PFS or OS for the different types of response CR, CRu, or PR. The lack of difference in outcome between CR and CRu patients is in line with the current response criteria,[1] which state regarding CRu lesions that if the type of abnormality does not change or slowly involutes over time without therapy or corticosteroids, it is reasonable to categorize these lesions as CR. However, we also found that PR was associated with a similar PFS and even a similar OS, suggesting that these response categories do not translate into meaningful differences in outcome and are therefore not reliable surrogate endpoints in PCNSL. A few other studies compared survival for different response categories.[2,4,5,14] Only one of these studies[2] used a landmark analysis, resulting in selection bias in the other studies (ie, immortal time bias) since response and

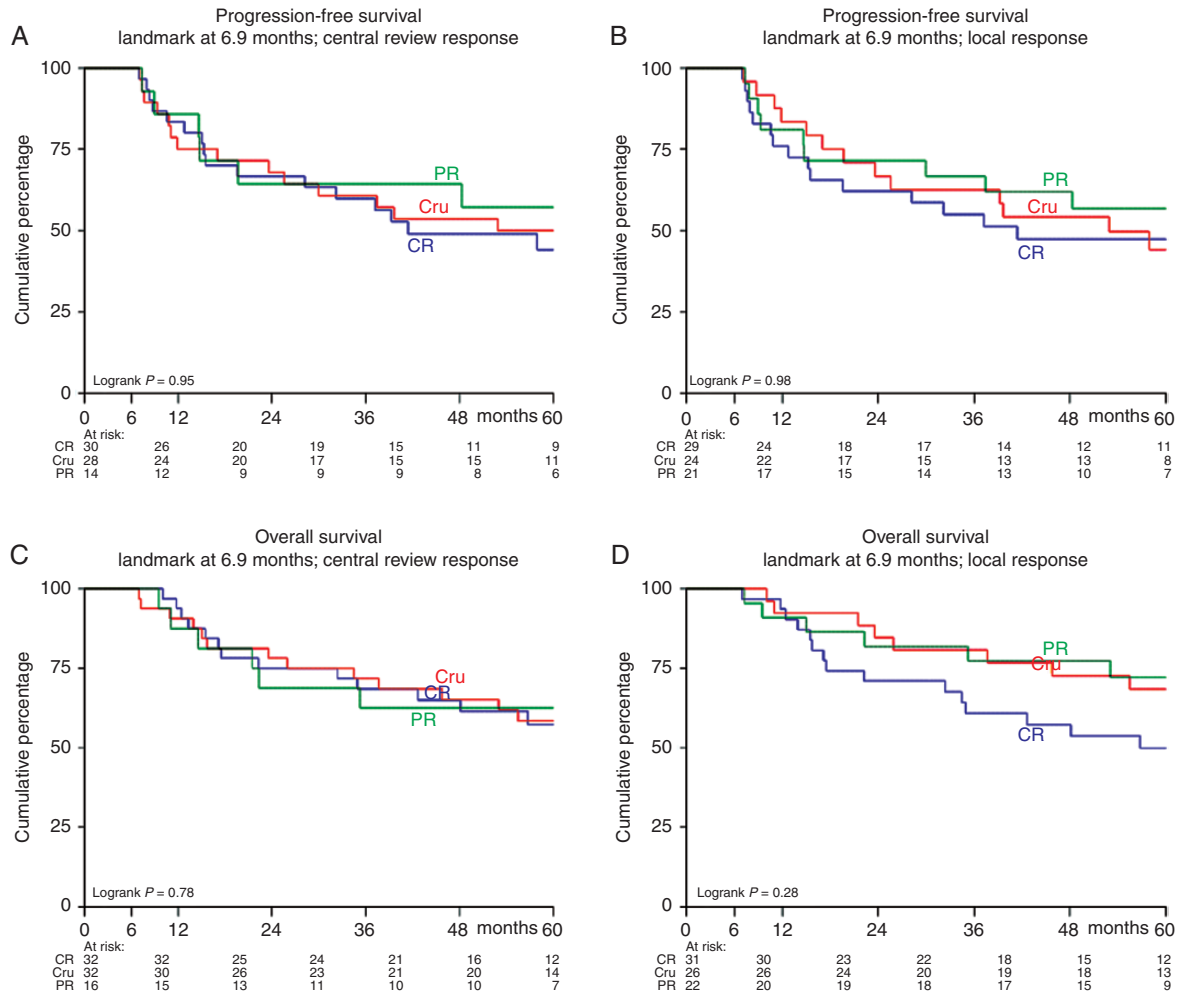

Figure 2. Progression-free survival (A and B) and overall survival (C and D) for those patients who had a partial response (PR), complete response (CR) or complete response unconfirmed (CRu) at the end of treatment MRI, according to central (A and C) and local response (B and D) assessment.

survival are influenced by the passing time. If survival analysis is done after end-of-protocol-treatment, regardless of when the MRI was made, those who have had a later MRI would have had more chance to achieve CR. One large, prospective study ($n = 511$) showed a significant difference between CR versus no CR (PR, SD, and PD combined) for OS (39 vs 22 months; $P < .0001$) and PFS (36 vs 6 months).[4] In that study, CR was defined as complete resolution of contrast-enhancing lesions on MRI or on CT. The latter radiological examination might have missed small contrast-enhancing foci. Furthermore, combining PR with nonresponding and progressive patients does not allow conclusions regarding the PR patients. Similarly, in a retrospective analysis of a phase II study in 85 patients, differences in survival rates between patients with CR, PR, SD, or progression after the end of chemotherapy were calculated using a single log-rank test. A significant difference was found for OS ($P < .001$), and a nearly significant difference for PFS ($P = .076$).[2] Again, due to the comparison of all groups including nonresponding or progressive patients this analysis does not allow comparison between patients with different extents of response. Lastly, a small retrospective single-center series evaluated patients after chemotherapy. Those with CR after the completion of chemotherapy ($n = 10$) had no better PFS or OS than those with no CR ($n = 30$).[5] In that study, however, patients without CR subsequently received additional treatment: radiotherapy or autologous stem cell transplantation.

In systemic DLBCL, residual abnormalities on CT do not always consist of active disease and are not reliable markers of prognosis. Fluorodoxyglucose positron-emission tomography (FDG-PET) response evaluation has better prognostic value than CT.[15] The avidity of FDG-PET lesions are classified with the Deauville score (range 1–5), in which 1 is a truly negative lesion and 5 as truly positive or avid lesions.[16]

In PCNSL patients, few, relatively small studies have shown a possible prognostic effect of PET scans.[17,18] There was a high concordance between MRI and PET imaging, using the Deauville score. Although the interim FDG-PET

scan had no prognostic value, patients with a negative PET-scan at the end-of-treatment had a significantly prolonged PFS but not OS.[17,18] These results suggest that PET imaging might be more useful as prognostic instrument than the radiological extent of response at MRI, although validation of these results in larger cohorts is necessary. Possibly in future molecular markers in CSF or even plasma will prove to be more reliable markers of which patients are truly in remission after induction therapy.

Two-dimensional measurements are the golden standard in the current PCNSL response criteria,[1] and were therefore also applied in this study. This might, underestimate volumetric changes. In glioma, volumetric measurements, either manual or computerized, improved agreement regarding radiological response compared to 2D measurements in some studies,[13,19] and fully automated segmentation was significantly better in predicting OS ($P < .0001$) than the conventional 2D measurements.[13] In contrast, one other smaller study showed no differences in predicting OS between manual 2D and 3D measurement or computerized segmentation of the tumor.[20] Response in PCNSL is generally easily recognizable with reductions >50% being the rule, so small changes in the volume of the enhancing lesion are unlikely to influence response rates. However, small changes in residual abnormalities might result in a change in response category, between CRu and PR.

Lastly, we found a few ($n = 7$) patients with space-occupying nonenhancing lesions. In most of these patients, the lesions diminished in size after chemotherapy, suggesting that these nonenhancing lesions might also be part of the cerebral lymphoma, as was also suggested by Tabouret et al.[2] A decrease in size of these lesions is likely to be beneficial, in terms of prognosis. However, since even in patients with enhancing disease white matter changes frequently persist after successful treatment, it is even more challenging to value the response of these nonenhancing lesions. Larger series of patients with nonenhancing lesions, perhaps utilizing other imaging modalities, will be needed to be able to define response criteria for these patients.

Naturally, our study has some limitations: First, analyses were performed on a subgroup ($n = 115$) from a large clinical trial and inadvertent bias may have occurred in the selection of patients for this study. The reasons for exclusion of the 84 other patients were mostly no consent for central radiological review, availability of the correct baseline scan, and in a small fraction no brain involvement. These factors are not related to survival or response and the main prognostic factors age and performance status did not differ between the 115 patients in this study and the 199 patients in the main study. Nevertheless, given the selection, our results should be validated in a larger external cohort. Secondly, other prognostic factors were not evaluated for each response category (ie, CR, CRu, and PR) in the landmark analyses, because the number of events (progression or death) was too small, therefore, the results of the landmark analyses should be interpreted as a univariate prognostic evaluation. Lastly, the landmark analyses were performed for different categories of response based on the MRI made at the end-of-protocol-treatment, which included WBRT in patients under 60-years old. Both lower age and the receipt of WBRT may impact PFS.[4,21,22] However, it is not clear whether this impact will differ between patients in CR, CRu, or PR after treatment. In the overall survival landmark analysis, those who received WBRT had a better OS than those who did not receive WBRT (Supplementary Figure S1). However, this difference could also be explained by the difference in age between these groups, rather than the addition of WBRT.

In conclusion, our results suggest that at the end of protocol treatment, within responding patients, specific radiological response categories (CR, CRu, or PR) do not reliably predict survival in PCNSL patients, even after central radiology review, but that interobserver agreement in diagnosing relapse or progression is high. Therefore, until more reliable markers of true remission are available the primary outcome measure in PCNSL studies should be PFS or OS; as secondary outcome measure combined response rate (CR, CRu, and PR) is more reliable than CR or CR/CRu-rate.

Supplementary Data

Supplementary data are available at *Neuro-Oncology Advances* online.

Keywords

central radiology review | complete response | MRI | response evaluation | survival

Funding

The HOVON 105/ALLG NHL 24 study was funded by Roche, the Dutch Cancer Society, and Stichting STOPhersentumoren.nl.

Conflict of interest statement. None of the authors has any conflict of interest to declare.

Authorship Statement. Initiate and designed the study, and involved in data collection, data interpretation, and writing the manuscript: J.E.C.B., S.I., and J.K.D. Study design, data analysis and interpretation, and writing the manuscript: K.B. study design, data collection, data interpretation, and writing the manuscript: M.v.d.M., A.A.P., M.S., and M.J.v.d.B. Data collection, data interpretation, and writing the manuscript: M.C.M., T.S., G.C., M.v.d.P., W.B.C.S., D.B., and A.B.

References

1. Abrey LE, Batchelor TT, Ferreri AJ, et al.; International Primary CNS Lymphoma Collaborative Group. Report of an international workshop to standardize baseline evaluation and response criteria for primary CNS lymphoma. *J Clin Oncol.* 2005;23(22):5034–5043.
2. Tabouret E, Houillier C, Martin-Duverneuil N, et al. Patterns of response and relapse in primary CNS lymphomas after first-line chemotherapy: imaging analysis of the ANOCEF-GOELAMS prospective randomized trial. *Neuro Oncol.* 2017;19(3):422–429.
3. Küker W, Nägele T, Thiel E, Weller M, Herrlinger U. Primary central nervous system lymphomas (PCNSL): MRI response criteria revised. *Neurology.* 2005;65(7):1129–1131.
4. Thiel E, Korfel A, Martus P, et al. High-dose methotrexate with or without whole brain radiotherapy for primary CNS lymphoma (G-PCNSL-SG-1): a phase 3, randomised, non-inferiority trial. *Lancet Oncol.* 2010;11(11):1036–1047.
5. Kim YR, Kim SH, Chang JH, et al. Early response to high-dose methotrexate, vincristine, and procarbazine chemotherapy-adapted strategy for primary CNS lymphoma: no consolidation therapy for patients achieving early complete response. *Ann Hematol.* 2014;93(2):211–219.
6. Bromberg JEC, Issa S, Bakunina K, et al. Rituximab in patients with primary CNS lymphoma (HOVON 105/ALLG NHL 24): a randomised, open-label, phase 3 intergroup study. *Lancet Oncol.* 2019;20(2):216–228.
7. Anderson JR, Cain KC, Gelber RD. Analysis of survival by tumor response. *J Clin Oncol.* 1983;1(11):710–719.
8. Cohen J. A coeffeicient of agreement for nominal scales. *Educ Psychol Meas.* 1960; 20(1):37–46.
9. Kubben PL, Postma AA, Kessels AG, van Overbeeke JJ, van Santbrink H. Intraobserver and interobserver agreement in volumetric assessment of glioblastoma multiforme resection. *Neurosurgery.* 2010;67(5):1329–1334.
10. Berntsen EM, Stensjøen AL, Langlo MS, et al. Volumetric segmentation of glioblastoma progression compared to bidimensional products and clinical radiological reports. *Acta Neurochir (wien).* 2020;162(2):379–387.
11. Visser M, Müller DMJ, van Duijn RJM, et al. Inter-rater agreement in glioma segmentations on longitudinal MRI. *Neuroimage Clin.* 2019;22:101727.
12. Vos MJ, Uitdehaag BM, Barkhof F, et al. Interobserver variability in the radiological assessment of response to chemotherapy in glioma. *Neurology.* 2003;60(5):826–830.
13. Kickingereder P, Isensee F, Tursunova I, et al. Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. *Lancet Oncol.* 2019;20(5):728–740.
14. Pels H, Juergens A, Schirgens I, et al. Early complete response during chemotherapy predicts favorable outcome in patients with primary CNS lymphoma. *Neuro Oncol.* 2010;12(7):720–724.
15. Cheson BD, Fisher RI, Barrington SF, et al.; Alliance, Australasian Leukaemia and Lymphoma Group; Eastern Cooperative Oncology Group; European Mantle Cell Lymphoma Consortium; Italian Lymphoma Foundation; European Organisation for Research; Treatment of Cancer/Dutch Hemato-Oncology Group; Grupo Español de Médula Ósea; German High-Grade Lymphoma Study Group; German Hodgkin's Study Group; Japanese Lymphorra Study Group; Lymphoma Study Association; NCIC Clinical Trials Group; Nordic Lymphoma Study Group; Southwest Oncology Group; United Kingdom National Cancer Research Institute. Recommendations for initial evaluation, staging, and response assessment of Hodgkin and non-Hodgkin lymphoma: the Lugano classification. *J Clin Oncol.* 2014;32(27):3059–3068.
16. Barrington SF, Mikhaeel NG, Kostakoglu L, et al. Role of imaging in the staging and response assessment of lymphoma: consensus of the International Conference on Malignant Lymphomas Imaging Working Group. *J Clin Oncol.* 2014;32(27):3048–3058.
17. Jo JC, Yoon DH, Kim S, et al. Interim (18)F-FGD PET/CT may not predict the outcome in primary central nervous system lymphoma patients treated with sequential treatment with methotrexate and cytarabine. *Ann Hematol.* 2017; 96(9):1509–1515.
18. Birsen R, Blanc E, Willems L, et al. Prognostic value of early 18F-FDG PET scanning evaluation in immunocompetent primary CNS lymphoma patients. *Oncotarget.* 2018;9(24):16822–16831.
19. Kanaly CW, Mehta AI, Ding D, et al. A novel, reproducible, and objective method for volumetric magnetic resonance imaging assessment of enhancing glioblastoma. *J Neurosurg.* 2014;121(3):536–542.
20. Gahrmann R, van den Bent M, van der Holt B, et al. Comparison of 2D (RANO) and volumetric methods for assessment of recurrent glioblastoma treated with bevacizumab-a report from the BELOB trial. *Neuro Oncol.* 2017;19(6):853–861.
21. Abrey LE, Ben-Porat L, Panageas KS, et al. Primary central nervous system lymphoma: the Memorial Sloan-Kettering Cancer Center prognostic model. *J Clin Oncol.* 2006;24(36):5711–5715.
22. Ferreri AJ, Blay JY, Reni M, et al. Prognostic scoring system for primary CNS lymphomas: the International Extranodal Lymphoma Study Group experience. *J Clin Oncol.* 2003;21(2):266–272.