6-1-2017

Investigator and independent review committee exploratory assessment and verification of tumor response in a non-Hodgkin lymphoma study

Robert R Ford

Robert W Ford

Michael O'Neal

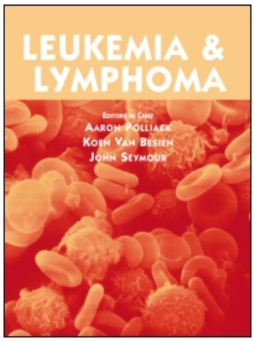
Brad S Kahl

Ling Chen

See next page for additional authors

Authors

Robert R Ford, Robert W Ford, Michael O'Neal, Brad S Kahl, Ling Chen, Mihaela Munteanu, and Bruce D Cheson



LEUKEMIA & LYMPHOMA

Editors in Chief
Aaron Polliack
Koen Van Besien
John Seymour

Taylor & Francis
Taylor & Francis Group

Leukemia & Lymphoma

ISSN: 1042-8194 (Print) 1029-2403 (Online) Journal homepage: https://www.tandfonline.com/loi/ilal20

Investigator and independent review committee exploratory assessment and verification of tumor response in a non-Hodgkin lymphoma study

Robert R. Ford, Robert W. Ford, Michael O'Neal, Brad S. Kahl, Ling Chen, Mihaela Munteanu & Bruce D. Cheson

To cite this article: Robert R. Ford, Robert W. Ford, Michael O'Neal, Brad S. Kahl, Ling Chen, Mihaela Munteanu & Bruce D. Cheson (2017) Investigator and independent review committee exploratory assessment and verification of tumor response in a non-Hodgkin lymphoma study, Leukemia & Lymphoma, 58:6, 1332-1340, DOI: 10.1080/10428194.2016.1233535

To link to this article: https://doi.org/10.1080/10428194.2016.1233535

Published online: 10 Oct 2016.

Article views: 6385

LEUKEMIA & LYMPHOMA, 2017
VOL. 58, NO. 6, 1332–1340
http://dx.doi.org/10.1080/10428194.2016.1233535

Taylor & Francis
Taylor & Francis Group

ORIGINAL ARTICLE: CLINICAL

OPEN ACCESS

Investigator and independent review committee exploratory assessment and verification of tumor response in a non-Hodgkin lymphoma study

Robert R. Ford[a], Robert W. Ford[b], Michael O'Neal[c], Brad S. Kahl[d], Ling Chen[e], Mihaela Munteanu[e] and Bruce D. Cheson[f]

[a]Clinical Trials Imaging Consulting LLC, Belle Mead, NJ, USA; [b]Department of Radiology, Thomas Jefferson University Hospital, Philadelphia, PA, USA; [c]Medical Imaging and Cardiovascular Solutions Management, BioClinica Inc, Princeton, NJ, USA; [d]Division of Oncology, Washington University School of Medicine in St. Louis, St. Louis, MO, USA; [e]Teva Branded Pharmaceutical Products R&D Inc, Frazer, PA, USA; [f]Division of Hematology and Oncology, Georgetown University, Washington, DC, USA

ABSTRACT

Interpretation of endpoints (e.g. overall response rate) in clinical trials depends on the accurate and reliable measurement and identification of tumors. Regulatory agencies recommend blinded reviews of imaging data by independent review committees (IRCs). Differences in response outcomes that arise between IRCs and site investigators raise regulatory/sponsor concerns. Here, we evaluate discrepant tumor response assessments by the IRC and unblinded investigators (complete versus partial response, respectively) occurring in 52 (13% of 393 IRC-assessed responders) of 447 enrolled patients with treatment-naïve non-Hodgkin lymphoma from a randomized study. The IRC and investigators were 'likely correct' in 73% and 25% of cases, respectively ($p < .001$). Investigators were more likely to make errors by misinterpreting lymph node data and not utilizing PET results. This *post hoc* finding suggests a possible role for post-training site evaluation/audit, with retraining as needed, and a specialized consensus committee for concurrent blinded review of site/central data.

ARTICLE HISTORY
Received 25 May 2016
Revised 22 August 2016
Accepted 31 August 2016

KEYWORDS
Independent review committee; investigator; non-Hodgkin lymphoma; tumor response

Introduction

Overall survival (OS), although widely considered the gold standard for treatment efficacy of oncology agents, needs considerable time for data maturation, and can be confounded by sequential therapy, non-cancer deaths, and patient loss to follow-up.[1] To address these and other concerns, alternate endpoints, including progression-free survival (PFS), time to progression, and overall response rate (ORR), are used as efficacy (even as primary) endpoints.[1] These measures, however, are based on interpretation of tumor response, which has multiple dependencies including measurement precision, reader perception, inherent tumor characteristics, manifestations of treatment effect, image quality, underlying patient condition, and the response criteria used for assessment.[2,3] Further, both inter- and intra-reader variability could lead to 'noise' in tumor response results.[2] Readers have been found to differ in the selection of a target lesion and tumor-margin measurement, particularly with poorly defined tumors.[4–6]

Incorporation of blinded independent review committees (IRCs) aims to improve objectivity and reliability of clinical data that might be subject to observer bias and variability.[7] A prevalent concern is that site investigators (INVs) may be subject to unintentional bias by additional patient information or the expected outcome of the trial.[8] The IRC review process is frequently structured to gain agreement between 2 radiologists well-versed in the study protocol, with disagreement between initial reviewers adjudicated by a third reviewer (i.e. '2 + 1' design).[9] Improvements in electronic communication facilitate prompt transmission of imaging data; in some cases, real-time evaluation of imaging data by an imaging core facility can provide feedback to trial sites.[10]

Recent studies and meta-analyses comparing INV and IRC imaging assessments found broad consistency between reviews.[2,8,10–14] However, among the ongoing debates is whether a threshold rate of discrepancy between INV and IRC assessments, if available, is an adequate indicator of validity of trial results,

CONTACT Robert R. Ford, MD rfordmd@gmail.com Principal, Clinical Trials Imaging Consulting LLC, 9 Raymond Lane, Belle Mead, NJ 08502, USA

and whether there are specific types of trials for which an IRC is particularly critical, and those for which an independent audit may be sufficient.[9,15] Moreover, understanding how differences in data interpretation arise could identify future opportunities for improved processes to minimize inter-reader differences. Some of the process improvement may include developing additional guidelines, for use across all sites, that address specific ambiguous scenarios.

As part of this effort, we conducted an exploratory review of a subset of discrepant responses (partial responses [PR] assessed by INVs and complete responses [CR] assessed by IRC) from the primary data of a recent first-line clinical trial in non-Hodgkin lymphoma (NHL). The randomized, noninferiority, global, phase 3 study compared the efficacy and safety of bendamustine-rituximab (B-R) against standard chemotherapy regimens.[16] Analysis of this study provided an opportunity to evaluate how differences could arise between INV and IRC in an active-controlled trial for which both INVs and IRC received training for objective response measures, specifically the International Working Group (IWG) response criteria.[16]

Materials and methods

Study design and objectives

The phase 3 study (NCT00877006) was an open-label, active-controlled, randomized clinical trial designed to assess noninferiority of B-R compared with standard treatment regimens of rituximab plus cyclophosphamide, doxorubicin, vincristine, and prednisone (R-CHOP) or rituximab plus cyclophosphamide, vincristine, and prednisone (R-CVP) in the first-line treatment of patients with CD-20–positive indolent NHL or mantle cell lymphoma (MCL).[16] The study design, enrollment eligibility criteria, treatment plan, and statistical analysis have been previously described.[16] The global study was submitted to each institution's Independent Ethics Committee or Institutional Review Board, and all patients submitted written informed consent.[16,17]

All efficacy response analyses were performed by the INV and an IRC, and local readers received training from the sponsor (Appendix). Tumor response was assessed by the revised IWG response criteria for malignant lymphoma.[18] This multidisciplinary assessment of response, including radiology and clinical reviews, incorporates qualitative findings on physical examination with tumor assessments by computed tomography (CT) scan, magnetic resonance imaging, and [^{18}F] fluorodeoxyglucose positron emission tomography (PET) when available, and results from immunohistochemistry and flow cytometry on tumor pathologic specimens from bone marrow biopsies and aspirates.[18] CR rates were also compared among treatment groups by preplanned subgroups defined by Follicular Lymphoma International Prognostic Index (FLIPI) risk score and bulky-disease status (i.e. tumor diameter ≥3 cm in ≥3 regions or diameter ≥7 cm in 1 region).[19] Patient baseline characteristics are summarized in Table 1.

Table 1. Baseline demographics and clinical characteristics.

Characteristic	Patients ($n = 52$)
Age, median, years (range)	60.5 (37.0–82.0)
Sex (male/female, %)	37/63
Race, *n* (%)	
White	47 (90)
Asian	1 (2)
Other	4 (8)
Histologic classification, *n* (%)	
Lymphoplasmacytic	1 (2)
Marginal zone	4 (8)
Mantle cell	11 (21)
Follicular, grade 1	17 (33)
Follicular, grade 2	19 (37)
Missing	1 (2)
FLIPI category for patients with follicular lymphoma, n (%)	
Low risk	11 (21)
Intermediate risk	10 (19)
High risk	15 (29)
FLIPI score, *n* (%)	
0	2 (4)
1	9 (17)
2	10 (19)
3	12 (23)
4	2 (4)
5	1 (2)
IPI category, *n* (%)	
Low risk	18 (35)
Low-intermediate risk	18 (35)
High-intermediate risk	14 (27)
High risk	2 (4)
B symptom present, *n* (%)	
Yes	18 (35)
No	32 (62)
Unknown	2 (4)
Age at onset, median, years (range)	59.0 (37.0-82.0)
Median time from diagnosis, months (range)	1.5 (0.1-86.2)
Baseline ECOG performance status, n (%)	
0	37 (71%)
1	15 (29)
≥2	0
Median time from most recent biopsy (not bone marrow), months (range)	1.0 (0.1–5.4)
PET data available	16 (31)

ECOG: Eastern Cooperative Oncology Group; FLIPI: Follicular Lymphoma International Prognostic Index; IPI: International Prognostic Index; PET: positron emission tomography.

Evaluation of patient-level divergent interpretations in tumor response

A *post hoc* analysis was performed on a subset of patients in the study with best response assessed as

PR by site INVs and CR by IRC. There were relatively few differences in assessments for other types of response, such as stable disease and progressive disease, and they were not assessed in this review due to resource constraints. The purpose of this review was to understand the reasons for the divergent interpretations and to identify opportunities for improvement so as to minimize differences in the future. This retrospective subanalysis was performed across all treatment groups by a consensus of 3 independent reviewers, 2 of whom each have 30 years of experience in the field of clinical trials using oncology imaging, and a postgraduate year-3 radiology resident who were blinded to treatment (to avoid potential bias) and were not part of the IRC, study, or participating sites. Available data included: index and non-index lesions and corresponding anatomic site codes, as well as lesion measurements; response assessment at each time point; the presence/absence of new lesions; and selected clinical information (e.g. bone marrow biopsy results, clinical symptoms, presence of hepatosplenomegaly, or individual lesions selected by the INV during the physical exam). Collection of PET data was optional as per the protocol, but, if performed, INV and IRC radiologists were required to include these data in assessing response. Index lesions were selected independently by the INV and IRC. Annotated files containing images marked with measurements were created and stored by IRC radiologists; however, the measurements provided by trial sites were not recorded with the image to indicate exactly what was measured. Therefore, inaccurate measurements and lesion selection errors could not be detected for INV, which could have biased this analysis against the IRC.

A complete review of all lesions and measurements was not performed, and therefore, lesion selection and measurements were initially assumed to be accurate as provided, and the nomenclature of 'likely correct' best response was adopted rather than 'true correct best response' given the available data and review limitations. Despite this planned methodology, obvious errors in lesion measurement were noted during this imaging review, and any corresponding changes in best response were acknowledged. In cases where both the INV and IRC reached a logically sound, but discrepant, best response based on data they included in their assessment, an additional image review was performed to determine if either assessment could be considered more accurate. This largely involved reviewing different anatomic site codes to determine if persistent abnormal lesions existed in those locations, which in turn would prevent a PR from being upgraded to CR. Specifically, if the image review identified any abnormal lymph node in the discrepant anatomic location, the PR was assessed as the 'likely correct' response, which was the most common reason for an IRC-assessed CR to be downgraded to PR.

Several types of error were considered for inclusion in this review. Random error cannot be controlled and is assumed to be similar in the IRC and INV groups. Human error can include, but is not limited to, lesions that are missed during the assessment, incorrectly measured, and/or incorrectly selected as index lesions by either the IRC or INV. The extent of human error cannot be fully assessed, however, without a complete review of all time points for accuracy in all aspects of outcome assessment (essentially determining a 'gold standard'). As this approach was beyond the scope of this review, the contribution of human error was not assessed and was assumed to be similar between the IRC and INV radiologists. This study therefore aims to isolate the degree to which INV and IRC review processes are themselves subject to error.

Multiple types of errors fell under the umbrella term of 'Process error'. A 'Process' error, defined as a data inclusion, application, and/or conclusion error, serves as a comparative measure between IRC and INV review methodologies. 'Data inclusion' errors were defined as a failure to incorporate available image data or clinical information into the response assessment. An 'Application' error was defined as incorrect application of the study protocol, response criteria, or response assessment (e.g. index/non-index disease does not meet inclusion criteria; only the percentage change in sum of product of the diameters (SPD) was considered instead of appropriate criteria for lesion type/number; failure to consider if lymph nodes returned to normal size before assigning best response; and failure to consider clinical data in best response assignment). A 'Conclusion' error was considered as a subset of application error, and results from arriving at the incorrect response conclusion based on the intrinsic data considered.

Results

Patient-level divergent interpretations in tumor response

A *post hoc* analysis of the study data set identified 52 patients whose best tumor response was categorized differently by IRC (CR) and INV (PR); these 52 patients represented 13% of the 393 IRC-assessed responders. Review of these cases based on available data found the IRC was 'likely correct' in 73% of discordant cases,

Table 2. Attributions of patient-level divergent interpretations ($n = 52$) in tumor response.

	INV	IRC	p Value[c]
'Likely Correct Best Response' in discordant cases[a]	25%	73%	<.0001
Process error[b]	79%	4%	<.0001
Data inclusion error	31%	0	<.0001
Application error	56%	6%	<.0001
Conclusion error	41%	4%	<.0001

INV: investigator; IRC: independent review committee.

[a] In one case, both the INV and IRC were 'likely correct', depending on which lesions were measured.

[b] More than one error was observed in some cases. 'Process errors' are defined as errors of inclusion, application, and/or conclusion. They are generally characterized as any errors primarily attributed to the process of image acquisition, distribution, and review, as well as response criteria application, but they would not include such errors as differences related to lesion choice, lesion/mass measurement, random chance, or failure to assess minimal residual disease in the bone marrow/blood. 'Data inclusion error' results from failure to incorporate available data/information into response assessment; primarily derived from failure to include positron emission tomography data when available. 'Application error' is defined by incorrect application of study protocol, response criteria, and response assessment. 'Conclusion error', a subset of application error, results from arriving at incorrect conclusion of response based on the intrinsic data considered.

[c] Fisher's exact test.

and the INV were 'likely correct' in 25% of discordant cases ($p < .0001$). In one discordant case (2%), the outcome difference was driven by alternative index lesion selection, without subsequent obvious errors in measurement, data acquisition, or the application of response criteria. As a result, neither the IRC nor the INV outcome could be judged as being incorrect.

Some discordant cases were associated with more than one error (Table 2). In 48% of discordant cases, the INV failed to consider that pathologic lymph nodes had returned to normal size, and erroneously assigned a best response of PR instead of CR according to the response criteria. Additionally, in 31% of discordant cases, the INV did not apply available PET data, which was required by the protocol. Relevant clinical data were applied incorrectly by the IRC in 6% of discordant cases. Incorrect lesion measurements were incidentally detected during additional image review in 12% and 17% of discordant cases for the IRC and INV, respectively.

The most common errors were 'Application Errors' (56% and 6% of INV and IRC tumor response interpretations, respectively) and occurred when the original pathologic index lesions returned to normal size at follow-up as defined by the study criteria, resulting in a 'likely correct' assessment of CR by the IRC. In these cases, the INV typically incorrectly assessed PR based on a reduction of greater than 50% in the SPD of the index lesions, while not accounting for the fact that all pathologic lesions returned to normal size. In 4 application-error cases, the IRC failed to downgrade CR to PR when bone marrow was involved at baseline but bone-marrow biopsy was not repeated at clinical CR. 'Data Inclusion Error' was another common error (31% and 0% of INV and IRC tumor response interpretations, respectively) and was seen when there were PET-negative residual nodal masses at follow-up that retained pathologic measurements. In these 15 cases, PR was incorrectly assessed by the INV despite the availability of relevant PET data, which would necessitate assigning CR to lesions that demonstrate resolution of hypermetabolic activity regardless of size.

Although it was not the main goal of the study, human error was observed during image review. As an example, a patient was enrolled with an abnormal supraclavicular lymph node at screening. Though the lymph node returned to normal size by cycle 6 as confirmed on image review, it was measured as abnormal by the INV, thus resulting in an incorrect assignment of PR as opposed to a valid CR. Another example involved a patient with extensive retroperitoneal/perivascular disease at screening. Although consensus imaging review confirmed that disease had resolved at cycle 8, multiple INV measurement errors incorrectly resulted in INV classification as PR instead of the correct outcome of CR as assessed by the IRC. As an example of an IRC human error, a patient had splenomegaly with an abnormal heterogeneous enhancement pattern at screening. The spleen later demonstrated a normal CT enhancement pattern at follow-up but remained enlarged. Though the IRC erroneously assigned a time point response of CR, the 'likely correct' response was determined to be PR due to persistent splenomegaly.

Consideration of clinical relevance of discordant assessments: trends in IRC and INV assessments of tumor response in the full study population

The full efficacy and safety analyses from this study have been previously described.[16,17] Across the full data set from the study, the IRC study assessed the CR rate at 31% (95%CI 25.3–38.2%) in the B-R group and 25% (95%CI 19.5–31.7%) in the R-CHOP/R-CVP groups. In the primary analysis for noninferiority (margin 0.88), the p value was .022, indicating statistical noninferiority between the two treatments. Site INVs reported fewer CR in the R-CHOP/R-CVP group, increasing the margin between the study groups: 31% (95%CI 24.8–37.7%) in the B-R group and 19% (95%CI 14.2–25.5%) in the R-CHOP/R-CVP group. Analysis of these data demonstrated noninferiority ($p = .002$), which paralleled the IRC findings.[16] The p value for

Table 3. Best overall response by INV compared with IRC.

Evaluable patients (%) in the study														
	BR treatment group ($n = 213$)							R-CHOP/R-CVP group ($n = 206$)						
	IRC assessment (%) (read down)							IRC assessment (%) (read down)						
	CR	PR	SD	PD	CPD	UN	INV total	CR	PR	SD	PD	CPD	UN	INV total
Investigator assessment (%) (read across)														
CR	19	*12*	*<1*	*0*	*0*	*0*	*31*	12	*6*	*1*	*0*	*0*	*0*	*19*
PR	**12**	52	*2*	*<1*	*0*	*0*	*66*	**13**	56	*4*	*0*	*0*	*0*	*74*
SD	**<1**	**1**	0	*0*	*0*	*0*	*2*	**0**	**3**	2	*0*	*0*	*0*	*6*
PD	**0**	**0**	**0**	0	*0*	*0*	*0*	**0**	**0**	**<1**	0	*0*	*0*	*<1*
CPD	**0**	**0**	**0**	**<1**	0	*0*	*<1*	**0**	**0**	**0**	**0**	0	*<1*	*<1*
UN	**0**	**<1**	**0**	**0**	**0**	0	*<1*	**0**	**0**	**0**	**0**	**0**	0	*0*
IRC Total	**31**	**65**	**3**	**<1**	**0**	**0**		**25**	**66**	**9**	**0**	**0**	**<1**	

BR: bendamustine and rituximab; CPD: clinical progressive disease; CR: complete response; INV: investigator; IRC: independent review committee; PD: progressive disease; PR: partial response; R-CHOP: rituximab plus cyclophosphamide, doxorubicin, vincristine, and prednisone; R-CVP: rituximab plus cyclophosphamide, vincristine, and prednisone; SD: stable disease; UN: unknown.

superiority was .005, which was significant, in contrast to the IRC analysis.[16]

There was greater difference between the INV- and IRC-assessed tumor responses in the R-CHOP/R-CVP group than in the B-R group. In the R-CHOP/R-CVP group, the IRC measured more CR than INV (25% vs. 19%, respectively). Correspondingly, the PR rates determined by IRC and INV were 66% and 74%, respectively, for the patients receiving standard therapy. In the B-R group, both the INV and IRC assessed the CR rate to be 31%, and the PR rates were 66% and 65% by the INV and IRC, respectively. The proportions of patients with better tumor response when measured by IRC compared with INV were 15% in the B-R group and 17% in the R-CHOP/R-CVP group (Table 3, bold numbers). INV-assessed CR was downgraded in the subanalysis to PR in 12% of patients receiving B-R and 6% of patients receiving R-CHOP/R-CVP. The proportions of patients with a poorer tumor response when measured by the IRC compared with INV were 15% in the B-R group and 12% in the R-CHOP/R-CVP group (Table 3, italicized numbers). The IRC upgraded INV-assessed response to CR in 12% of patients receiving B-R and 13% of patients receiving R-CHOP/R-CVP.

Discussion

The objective of this exploratory analysis was to evaluate causes of discordance between INV and IRC tumor assessments from a phase 3 study in advanced indolent NHL or MCL. Although the INV and IRC assessed similar rates of tumor response in the trial, this case-by-case comparison of the 52 patients whose data were interpreted differently (out of 447 patients) provides a qualitative look at how errors are introduced into clinical trial data. More precisely, in this response data set, site INVs categorized patients as PR and the IRC categorized patients as CR. Overall, we found that site INVs made significantly more process errors than the IRC ($p < .0001$). Errors occurred in the IRC process as well, although to a smaller extent. Although the differences affected a small proportion of the study population, the potential exists for these errors to affect data analysis and interpretation.

In this analysis, we have focused on the types of error that occur in interpreting clinical and imaging data for the determination of tumor response in clinical trials. Although the data set reported here was not suitable to completely and prospectively analyze human and random errors, these errors likely occur at similar frequencies in both review processes, and would presumably contribute equally to the data set. Our analysis indicates that INVs are more prone to process errors than a blinded IRC, with INV errors occurring in approximately 73% of discrepant cases. The majority of these errors reflected the lack of incorporation of available critical data (in particular, available PET data) or incorrect application of response criteria (in particular, only considering reduction in the SPDs and failure to include that lymph node returned to normal size; thus, a best response of PR rather than CR is assigned per response criteria), and thus the INVs did not arrive at the correct conclusion during the response assessment. These process errors may be attributed to the incorrect selection or misinterpretation of imaging data, and may reflect selection bias. Measures to address these errors might include additional risk-based monitoring, without breaking the blind, to minimize inter-reader differences. Whether the errors were equally distributed across all INV sites or clustered at a few sites was not assessed, but could be seen as an opportunity for an additional risk-based monitoring. It should be pointed out that as the measurements provided by trial sites were not recorded with the image to indicate exactly what were measured, inaccurate measurements and lesion selection errors could not be detected for INV, which could have biased this analysis against the IRC.

In this study, both INV and IRC results supported the primary endpoint. Of note, however, there was a trend for more discordance between INV and IRC assessments in the R-CHOP/R-CVP group than in the B-R group. Although the cause of this difference is unknown, possible factors might include that tumor progression is subject to time bias, and studies with different schedules of disease assessments may be more prone to variability between treatment arms.[20] In this study, the treatment cycles were 28 and 21 days for the B-R and R-CHOP/R-CVP regimens, respectively.[16] Unblinded observer bias is an additional possibility.

These data about process errors add granularity to the previously published analyses comparing the data reported by INVs and IRCs from other cancer trials.[8,11–13] Meta-analyses have shown modestly, but generally higher PFS rates reported by INVs compared with IRCs,[11,13] but one, which also compared response rates, reported that aggregate results across studies were similar between INVs and IRCs, although there were wide variations among the component trials.[11] Because the response rate analysis did not evaluate the differences in the qualification of the response in the 18 trials with response rate as the primary endpoint in the meta-analysis, a direct comparison between that paper and our analysis cannot be made.[11] However, the authors of these analyses have generally recommended the use of a blinded IRC if the primary endpoint is changes in tumor response, or in cases where there are potential local INV biases due to the nature of the trials (e.g. unblinded trial where a small effect on PFS is observed).[8,11,13] This observation, however, was not extended to PFS. Median PFS reported by the INVs was equally likely to be longer or shorter than the reported IRC data in the trials included in the meta-analyses, suggesting little to no bias by the INV, although this finding may vary between tumor types and available treatments.[11,12] This finding has been interpreted to indicate an inherent variability in the process of measuring PFS at the patient level. Regulatory authorities currently consider the relative treatment effect across the study population, which modulates the variability, rather than analyzing patient-level data.[12] However, high-quality raw data are necessary to properly assess the efficacy of oncology treatments, particularly as more agents become available through development pipelines. The impact of data errors on the treatment effects reported by INV without an IRC review is unknown.

Although this analysis is *post hoc*, these data provide insight into the strengths and weaknesses of INV and IRC outcomes, and provide an opportunity to consider methods to reduce error. We recommend the incorporation of an IRC for most, if not all, large clinical trials studying tumor response, as well as a consensus committee where both the site and central data are reviewed concurrently when possible. This approach could be used to monitor site performance, while providing context for observed divergent interpretations for INVs, sponsors, and regulators. A higher level site training may reduce rates of process errors; however, site staff turnover is at a potential complication. Application errors could be addressed by better application of the study protocol and staging and response criteria. Similarly, training may reduce errors that occur in the IRC process. Monitoring reader performance is an important part of the IRC process. Improvements in imaging may reduce variability as well. The 2014 Lugano classification system emphasizes accurate imaging with PET-CT scans, which can improve the accuracy of staging treatment selection and measuring treatment response for patients with NHL.[21] Future clinical trials that incorporate these standards, especially when images are taken with high-quality calibrated scanners, may have higher consistency between INV and IRC assessments.

Large multicenter clinical trials are the best mechanism for evaluating the efficacy and safety of oncology drugs. The decentralized designs, however, are only as strong as the quality of each study site. Our data emphasize the importance of an IRC in oncology trials, and for additional review of data to evaluate site performance and identify points for improvement.

Acknowledgments

This research was sponsored and conducted by Teva Branded Pharmaceutical Products R&D, Inc., Frazer, PA, USA. Funding for editorial support was provided by Teva Branded Pharmaceutical Products R&D to The Curry Rockefeller Group, LLC, Tarrytown, NY, USA.

Potential conflict of interest: Disclosure forms provided by the authors are available with the full text of this article online at http://dx.doi.org/10.1080/10428194.2016.1233535.

References

[1] Villaruz LC, Socinski MA. The clinical viewpoint: definitions, limitations of RECIST, practical considerations of measurement. Clin Cancer Res. 2013;19:2629–2636.

[2] Ford R, Schwartz L, Dancey J, et al. Lessons learned from independent central review. Eur J Cancer. 2009;45:268–274.

[3] Petrick N, Kim HJ, Clunie D, et al. Evaluation of 1D, 2D and 3D nodule size estimation by radiologists for spherical and non-spherical nodules through CT thoracic phantom imaging. In: Summers RM, van Ginneken B, editors. Proceedings from the 2011 SPIE medical imaging conference; 2011 Feb 12–17; Lake Buena Vista, Florida: SPIE, The International Society for Optical Engineering; 2011.

[4] Hopper KD, Kasales CJ, Van Slyke MA, et al. Analysis of interobserver and intraobserver variability in CT tumor measurements. AJR Am J Roentgenol. 1996;167:851–854.

[5] Thiesse P, Ollivier L, Di Stefano-Louineau D, et al. Response rate accuracy in oncology trials: reasons for interobserver variability. Groupe Francais d'Immunotherapie of the Federation Nationale des Centers de Lutte Contre le Cancer. J Clin Oncol. 1997;15:3507–3514.

[6] Erasmus JJ, Gladish GW, Broemeling L, et al. Interobserver and intraobserver variability in measurement of non-small-cell carcinoma lung lesions: implications for assessment of tumor response. J Clin Oncol. 2003;21:2574–2582.

[7] Dancey JE, Dodd LE, Ford R, et al. Recommendations for the assessment of progression in randomised cancer treatment trials. Eur J Cancer. 2009;45:281–289.

[8] Amit O, Mannino F, Stone AM, et al. Blinded independent central review of progression in cancer clinical trials: results from a meta-analysis. Eur J Cancer. 2011;47:1772–1778.

[9] Goldmacher GV, Raunig D. The imaging core lab perspective on glioblastoma imaging and response assessment in clinical trials. Neuro Oncol. 2014;16:vii48–vii50.

[10] Hamilton TE, Barnhart D, Gow K, et al. Inter-rater reliability of surgical reviews for AREN03B2: a COG renal tumor committee study. J Pediatr Surg. 2014;49: 154–158.

[11] Tang PA, Pond GR, Chen EX. Influence of an independent review committee on assessment of response rate and progression-free survival in phase III clinical trials. Ann Oncol. 2010;21:19–26.

[12] Zhang JJ, Zhang L, Chen H, et al. Assessment of audit methodologies for bias evaluation of tumor progression in oncology clinical trials. Clin Cancer Res. 2013;19:2637–2645.

[13] Zhang JJ, Chen H, He K, et al. Evaluation of blinded independent central review of tumor progression in oncology clinical trials: a meta-analysis. Drug Inf J. 2013;47:167–174.

[14] Floquet A, Vergote I, Colombo N, et al. Progression-free survival by local investigator versus independent central review: comparative analysis of the AGO-OVAR16 Trial. Gynecol Oncol. 2015;136:37–42.

[15] FDA Briefing Document. Oncologic Drugs Advisory Committee Meeting: evaluation of radiologic review of progression-free survival in non-hematologic malignancies [Internet]. [cited 2012 Jul 24]. Available from: http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/OncologicDrugsAdvisoryCommittee/UCM311141.pdf.

[16] Flinn IW, van der Jagt R, Kahl BS, et al. Randomized trial of bendamustine-rituximab or R-CHOP/R-CVP in first-line treatment of indolent NHL or MCL: the BRIGHT study. Blood. 2014;123:2944–2952.

[17] Kahl BS, Bartlett NL, Leonard JP, et al. Bendamustine is effective therapy in patients with rituximab-refractory, indolent B-cell non-Hodgkin lymphoma: results from a multicenter study. Cancer. 2010;116:106–114.

[18] Cheson BD, Pfistner B, Juweid ME, et al. Revised response criteria for malignant lymphoma. J Clin Oncol. 2007;25:579–586.

[19] Solal-Celigny P, Roy P, Colombat P, et al. Follicular lymphoma international prognostic index. Blood. 2004;104:1258–1265.

[20] Bhattacharya S, Fyfe G, Gray RJ, et al. Role of sensitivity analyses in assessing progression-free survival in late-stage oncology trials. J Clin Oncol. 2009;27:5958–5964.

[21] Cheson BD, Fisher RI, Barrington SF, et al. Recommendations for initial evaluation, staging, and response assessment of Hodgkin and non-Hodgkin lymphoma: the Lugano classification. J Clin Oncol. 2014;32:3059–3068.

Appendix

Response definitions from the study protocol. The IRC conducted its reviews according to a charter. Per the nature of charters, it was more detailed than the following definitions in the protocol, which were provided to the investigators and the IRC.

Complete response (CR)

The designation of CR requires the following:

- There must be complete disappearance of all detectable clinical evidence of disease and disease-related symptoms, if present before therapy.
- PET scan assessment criteria are as follows:
  - If the pretreatment PET scan was FDG avid/positive, a residual mass of any size on a post-treatment CT is permitted when the corresponding PET scan is FDG negative.
  - If the pretreatment PET scan was not performed but the patient has a lymphoma subtype which is typically FDG avid/positive, a residual mass of any size on a post-treatment CT is permitted when the corresponding PET scan is FDG negative.
  - If the pretreatment PET scan was FDG negative, all lymph nodes and nodal masses must have regressed to normal size. All nodes/masses greater than 1.5 cm (long axis) pretreatment must have decreased to 1.5 cm (long axis) or less post-treatment. All nodes/masses of 1.1–1.5 cm (long axis) and greater than 1.0 cm (short axis) pretreatment must have decreased to 1.0 cm (short axis) or less post-treatment.
  - If the pretreatment PET scan was not performed and the patient has a lymphoma subtype for which FDG

avidity is either unknown or variable, all lymph nodes and nodal masses must have regressed to normal size. All nodes/masses greater than 1.5 cm (long axis) pretreatment must have decreased to 1.5 cm (long axis) or less post-treatment. All nodes/masses between 1.1–1.5 cm (long axis) and greater than 1.0 cm (short axis) pretreatment must have decreased to 1.0 cm (short axis) or less post-treatment.

- If the spleen and/or liver are enlarged on the basis of physical examination and/or anatomic imaging (CT and/or MRI) before treatment, the liver and/or spleen should be considered normal size on physical examination and by anatomic imaging after therapy, with disappearance of all nodules related to lymphoma.
- If the bone marrow was involved by lymphoma before treatment, the infiltrate must have cleared on subsequent bone marrow biopsies. The biopsy sample on which this determination is made must be adequate (with a goal of greater than 20 mm unilateral core). If the sample is indeterminate by morphology, it should be negative by immunohistochemistry. A sample that is negative by immunohistochemistry but that demonstrates a small population of clonal lymphocytes by flow cytometry will be considered a CR until data become available demonstrating a clear difference in patient outcome.

Partial response (PR)

The designation of PR requires the following:

- There must be at least a 50% decrease in the sum of the product of the diameters (SPD) of up to 6 of the largest dominant nodes/masses. (Nodes selected pretreatment should be clearly measurable in at least 2 perpendicular dimensions from disparate regions/anatomic sites including the nodes from the mediastinum and retroperitoneum when possible.)
- There must be at least a 50% decrease in the SPD of hepatic and splenic nodules in their greatest transverse diameter.
- There must be no increase in the size of the liver, spleen, and other nodes.
- There must be no measurable disease in organs other than the liver or spleen.
- Bone marrow assessment is irrelevant for determination of a PR if the sample was positive before treatment. A clinical CR with persistent morphologic bone marrow involvement will be considered a PR. A clinical CR with no post-treatment bone marrow evaluation will be considered a PR.
- No new sites of disease should be observed.
- Typically FDG-avid lymphoma: for patients with no pretreatment PET scan or if the PET scan was positive before therapy, the post-treatment PET should be positive in at least 1 previously involved site.
- Variably FDG-avid lymphomas/FDG avidity unknown: for patients without a pretreatment PET scan, or if the PET scan was positive before therapy, the post-treatment PET should be positive in at least 1 previously involved site.
- In patients with follicular lymphoma or mantle cell lymphoma, a PET scan is only indicated with 1 or at most 2 residual masses that have regressed by more than 50% on CT; those with more than 2 residual lesions are unlikely to be PET negative and should be considered partial responders.

Stable disease (SD)

The designation of SD requires the following:

- A patient is considered to have SD when he or she fails to attain the criteria needed for a CR or PR, but does not fulfill those for progressive disease.
- Typically FDG-avid lymphomas: the PET should be positive at prior sites of disease with no new areas of involvement on the post-treatment CT or PET.
- Variably FDG-avid lymphomas/FDG-avidity unknown: for patients without a pretreatment PET scan or if the pretreatment PET was negative, there must be no change in the size of the previous lesions on the post-treatment CT scan.

Relapsed disease (after complete response)/progressive disease (after partial response, stable disease)

Relapsed disease (after CR) and progressive disease (PD) (after PR or SD) requires the following:

- Lymph nodes should be considered abnormal if the long axis is greater than 1.5 cm regardless of the short axis. If a lymph node has a long axis of 1.1–1.5 cm, it should only be considered abnormal if its short axis is greater than 1.0 cm. Lymph nodes measuring 1.0 cm by 1.0 cm or less will not be considered as abnormal for relapse or progressive disease.
- There must not be any new lesion more than 1.5 cm in any axis during or at the end of therapy, even if other lesions are decreasing in size. Increased FDG uptake in a previously unaffected site should only be considered relapsed or progressive disease after confirmation with other modalities. In patients with no prior history of pulmonary lymphoma, new lung nodules identified by CT are mostly benign. Thus, a therapeutic decision should not be made solely on the basis of the PET without histologic confirmation.
- There must be at least a 50% increase from nadir in the SPD of any previously involved nodes, or in a single involved node, or the size of other lesions (e.g. splenic or hepatic nodules). To be considered progressive disease, a lymph node with a diameter of the short axis of less than 1.0 cm must increase by 2: 50% and to a size of 1.5 cm by 1.5 cm, or more than 1.5 cm in the long axis.
- There must be at least a 50% increase in the longest diameter of any single previously identified node more than 1 cm in its short axis.

- Lesions should be PET positive if observed in a typical FDG-avid lymphoma or the lesion was PET positive before therapy unless the lesion is too small to be detected with current PET systems (<1.5 cm in its long axis by CT). Measurable extranodal disease should be assessed in a manner similar to that for nodal disease. For these recommendations, the spleen is considered nodal disease. Disease that is only assessable (e.g. pleural effusions, bone lesions) will be recorded as present or absent only, unless, while an abnormality is still noted by imaging studies or physical examination, it is found to be histologically negative.

Assessment of response

Each investigator will assess disease response (CR, PR, SD, PD, or relapsed disease) at the end of cycles 3 and 6 and at the end of cycle 8, if applicable. The investigator should use the same modality used at baseline to assess both measurable and assessable disease throughout and at the end-of-treatment visit. Tumor assessments should incorporate findings from physical examination, CT scan, MRI, [^{18}F] FDG PET, immunohistochemistry, flow cytometry, molecular genetics when appropriate on tumor tissue, and bone marrow biopsies/aspirates.

The use of PET for response in this study is optional and at the discretion of the investigator. If used, PET scans may be obtained as a stand-alone scan or with CT or MRI integration. If PET scans are not utilized, response should be assessed as above, but only using CT scans. However, residual masses should not be assigned unconfirmed CR (CRu) status, but should be considered partial responses.

Computed tomography (CT) scans or magnetic resonance imaging (MRI)

CT scans or MRI of the neck, chest, abdomen and pelvis will be performed to assess extent of disease at baseline (CT scans or MRI performed during screening are acceptable as the baseline scan if completed within 6 weeks prior to the first study treatment), and any response or progression of disease at cycles 3, 6, and 8, if applicable, and at any time at the investigator's discretion.

All CT scans should be performed with intravenous (IV) contrast, and abdominal and pelvic CT scans should be performed with oral contrast. The CT scans may be performed only with oral contrast if a patient is allergic to IV contrast agents.

[^{l8}F]-fluorodeoxyglucose (FDG) positron-emission tomography (PET)

A PET scan with [^{l8}F] FDG extending from the neck through the mid-thighs may be performed to assess baseline disease at screening and to assess disease response at any time at the investigator's discretion.

Bone marrow biopsy and aspirate

A bone marrow aspirate and biopsy sample will be obtained up to 60 days prior to the first dose of study drug treatment. Initial bone marrow examinations should establish the presence of disease involvement, if any. Adequate immunophenotyping to establish disease in the bone marrow pretreatment (within 60 days prior to study drug) should be performed at baseline and with any subsequent bone marrow evaluations.

The bone marrow must be repeated at the time of a clinical CR, if the baseline bone marrow was positive (evidence of lymphoma), or was insufficient or indeterminate. If the bone marrow was involved at baseline and not repeated at the time of a clinical CR, the best possible response is a PR at that time point.

If a patient was known to have follicular lymphoma and B-cell lymphoma/leukemia 2 (BCL-2) positivity in the bone marrow at baseline, an assessment of BCL-2 on any subsequent bone marrow evaluations is suggested. Cytogenetics or fluorescence *in situ* hybridization (FISH) for t(14;18) and t(8;14) and/or variants and molecular genetic analysis to detect antigen gene receptor rearrangement/BCL-2 rearrangement is suggested but not required.

If a patient was known to have mantle cell lymphoma and cyclin D1 positivity in the bone marrow at baseline, an assessment of cyclin D1 on any subsequent bone marrow evaluations is suggested. Cytogenetics or FISH for t(11;14) and t (14;18) and/or variants and molecular genetic analysis to detect antigen gene receptor rearrangement/bcl-1 rearrangement is suggested but not required.

Standard bone marrow procurement procedures will be followed for the collection of tissue. The bone marrow should be reviewed by the hematopathologist/oncologist for morphologic assessment, flow cytometry, and cytogenetics.