The Oncologist®
Cancer Treatment Reports

Blinded Independent Central Review of the Progression-Free Survival Endpoint

OHAD AMIT,[a] WILL BUSHNELL,[a] LORI DODD,[b] NANCY ROACH,[c] DANIEL SARGENT[d]

[a]GlaxoSmithKline, Collegeville, Pennsylvania, USA; [b]National Institute of Allergy and Infectious Diseases, Bethesda, Maryland, USA; [c]Colorectal Cancer Coalition, Alexandria, Virginia, USA; [d]Mayo Clinic, Rochester, Minnesota, USA

Disclosures: Ohad Amit: *Employment:* GlaxoSmithKline, *Ownership interest:* GlaxoSmithKline; **Will Bushnell:** *Employment:* GlaxoSmithKline, *Ownership interest:* GlaxoSmithKline; **Lori Dodd:** None; **Nancy Roach:** *Employment:* Colorectal Cancer Coalition, (volunteer) Chair, Board of Directors; **Daniel Sargent:** None.

The content of this article has been reviewed by independent peer reviewers to ensure that it is balanced, objective, and free from commercial bias. No financial relationships relevant to the content of this article have been disclosed by the independent peer reviewers.

INTRODUCTION AND BACKGROUND

Progression-free survival (PFS) is an endpoint of increasing use in phase III clinical trials. The primary appeal of the PFS endpoint is that, in contrast to the endpoint of overall survival (OS), it is measured prior to the use of alternative or subsequent anticancer therapies, thereby providing an estimation of the agent's biologic activity not confounded by other therapies. In addition, because progression is an event that occurs, in most cases, months or years before death resulting from cancer, clinical trials can be conducted more quickly with fewer patients than a trial designed using an OS endpoint. Although some have argued that PFS measures direct clinical benefit in some clinical settings, the benefits from delaying progression may be difficult to quantify. For the purposes of this panel, we accept that PFS can be a useful endpoint in some contexts, which will depend on the purpose of the trial, the magnitude of the PFS improvement expected, and the adverse event profile of the agent(s) under study.

When PFS is considered an appropriate endpoint for a trial, care must be taken to ensure that the PFS endpoint is reliably and reproducibly measured. Specifically, there are unique sources of bias related to PFS that must be considered. These include: evaluation-time bias, attrition bias, and reader-evaluation bias. Evaluation-time bias occurs when there are intentional or unintentional differences in the evaluation times by treatment arm [1, 2]. Specifically, when progression is evaluated more frequently in one arm, bias may result. For example, time of progression cannot be determined when attrition bias occurs as a result of lost-to-follow-up. This is unlike OS, for which a determination is usually possible. Reader-evaluation bias in unblinded trials, which is the focus of this panel, is of concern because of the potential for subjective elements to influence the disease progression evaluation.

In spite of objective criteria for determining progression [3], its evaluation is complicated by many factors. These complications include variation in tumor measurement, variation in the choice of target lesions to follow across time, failure to detect a new lesion, as well as differing interpretations about changes in nontarget and nonmeasurable lesions. These measurement issues can result in different determinations of a patient's status between evaluators at a given evaluation time. Because of this, a certain number of discrepancies is to be expected in any given trial (even in the absence of bias). However, the impact of these discrepancies on the evaluation of the treatment effect is an area of ongoing research.

Received February 9, 2010; accepted for publication February 15, 2010. ©AlphaMed Press 1083-7159/2010/$30.00/0 doi: 10.1634/theoncologist.2009-0261

When evaluations are made with knowledge of treatment assignment, there is a concern that assessments may be influenced by an evaluator's beliefs about a therapy. This knowledge creates the potential for intentional or unintentional actions to bias the estimate of the treatment effect, which is the main motivation for blinded independent central review (BICR) of locally evaluated (LE) progression times. BICR has been recommended in regulatory guidance documents for unblinded phase III clinical trials [3]. BICR is usually conducted by contract research organizations and is a large expense added to the already high cost of oncologic drug development. Although the motivation for BICR arises from variability in PFS assessments, the presence of reader-evaluation bias in the estimates of treatment effect based on LE progression times has not been, to date, documented in actual clinical trials. A paper by Dodd et al. [4] showed that, in a limited sample of clinical trials, there was generally consistent estimation of treatment effect between LE and BICR PFS, leading some to question the motivation for full BICR.

Additionally, Dodd et al. [4] describe a type of informative censoring that may bias the estimate of treatment effect based on BICR [4]. When an investigator has made an assessment of progression at a time point, the patient is typically withdrawn from the study and no further protocol scans are conducted. This means that if, upon review, the BICR does not determine progression for this patient at this time point, the patient's data are censored at this time point for statistical analysis based on the BICR data. Because this patient is more likely to have a BICR progression sooner than the remaining at-risk cohort, this censoring is informative. In other words, the standard statistical assumption that censoring is unrelated to prognosis is violated, and may bias estimates of treatment effect. Imbalance of this type of censoring between treatment groups is of particular concern.

These potential complications with both BICR and LE estimates of treatment effect have resulted in a dilemma for regulatory agencies in deciding which of the two estimates should be referenced in product labeling. In this document, we summarize two separate efforts addressing concerns related to BICR. The first was undertaken by the Pharmaceutical Research and Manufacturing Association (PhRMA) PFS Working Group. The second was undertaken by the National Cancer Institute (NCI), in collaboration with Eastern Cooperative Oncology Group and Genentech statisticians. Before describing these results, we review the outcomes from the 2008 Brookings session on PFS outcomes.

2008 Brookings Session on PFS Outcomes

At the Brookings Institute conference on cancer research in 2008, the primary conclusions included: (a) confirmation that, in truly double-blind clinical trials, BICR is not needed, which is consistent with U.S. Food and Drug Administration (FDA) guidance [5], and (b) a consensus that the method for auditing LE by obtaining BICR in a subset of patients needs to be developed. It was hoped that such an auditing method would replace the full independent review in confirmatory phase III trials. Researchers within the NCI and within the PhRMA PFS Working Group were requested to develop a sample-based audit of the investigator's assessment of progression that would be able to provide assurance of a lack bias in estimating treatment effects or to identify such a bias when present.

Summary of PhRMA PFS Working Group Data Collection and Analysis

As background to the audit methods that were presented, the PhRMA Working Group felt that the most important metric through which to understand the underlying agreement of investigator and BICR estimates of treatment in the case of PFS is the hazard ratio (HR) comparing the control with the experimental arm of a clinical trial. The primary goal of the audit was to understand how discordance (disagreement at the patient level between the investigator and BICR) affects how well the PFS HRs based on the BICR and local investigator agree.

The Independent Review subteam of the PhRMA Working Group undertook a data collection project to understand the operating principles of BICR in randomized oncology clinical trials. The team summarized HRs from 23 oncology clinical trials that used BICR to assess PFS, via a literature review. In addition, this team performed a data collection exercise to further evaluate the relationship between discordance and the agreement of BICR and investigator HRs. They investigated discordance by treatment group to determine how differential discordance results in potential bias of the BICR HR. The results from the literature review and data collection exercise were confirmed through simulation.

Preliminary results suggest that there is strong agreement between the investigator and BICR estimates of treatment effect. Further, there is evidence to suggest that the overall level of discordance is not related to the reliability of either investigator or BICR estimates of treatment effect. However, a difference between arms in discordance does appear to correlate with more divergent estimates of treatment effect between the BICR and investigator.

Summaries from the literature review and detailed data

collection will be presented. It is important to understand the strong agreement demonstrated in the analysis of 23 clinical trials as a background to understanding the need and threshold for detecting bias in an independent review audit.

PhRMA PFS Working Group Audit Methodology

The PhRMA Independent Review team took the approach of developing and using measures of discordance as the foundation of their audit methodology. It is acknowledged that the ultimate measure of interest is the HR; however, it is less sensitive as a tool for detecting bias and therefore was not explored as part of the audit methodology. Bias in treatment effect in this setting could be caused by two behaviors. The first behavior that could cause bias is when an investigator either knew or suspected that a patient was in the control group, felt the patient was not doing well, and declared progressive disease based on clinical symptoms with no substantiating radiologic evidence. Conversely, an investigator who knew or suspected a patient was in the experimental arm and felt that the patient was doing well despite meeting technical protocol criteria for progression could make the decision to keep the patient on treatment. Simulations have demonstrated that both these actions would result in inflated estimates of treatment effect and would increase the chances of a false-positive finding for the study. In addition, the magnitude of the difference between arms in certain discordance rates is markedly greater in the presence of bias. It is critical therefore that the audit mechanism proposed be sensitive to detection of either of these two possible biases.

The independent review team developed and evaluated multiple audit-based measures of discordance. The team generated, through simulation, multiple scenarios to represent the breadth of possible examples from oncology clinical trials.

The criteria for choosing the measure of discordance to be used in the audit were based on a high probability of detecting bias in a simulated scenario and to likewise have properties that resulted in a low probability of falsely declaring bias. The candidate discordance measure had to have stable performance regardless of the event rate in the trial, the differential event rate between arms, and the sample size of the trial. The discordance measures of interest and their performance will be discussed.

Some recommendations for consideration include having a central repository for all scans. This repository can then be a source for a random sample of subjects on which to perform BICR. The sample size of central review would depend on the sensitivity and specificity of differential discordance measures.

NCI Audit Methodology

Although BICR is potentially afflicted by informative censoring, agreement between the LE and BICR HRs provides reassurance that any positive treatment effect obtained by evaluations at local sites is not a result of reader-evaluation bias. Different distributions of discrepancies in PFS times between LE and BICR by treatment arm is an indication of reader-evaluation bias. However, because of censoring (administrative and otherwise), such an analysis is complicated.

Because the HR is ultimately the measure of interest in determining whether a treatment is efficacious, the efforts of this team focused on using BICR to estimate a HR that would have been obtained with a BICR, but in an efficient way not requiring a full-sample BICR. The audit strategy can be summarized as follows:

1. When the LE HR indicates a clinically meaningful and statistically significant effect, BICR will be conducted on a subset.

2. The HR from the BICR audit will be estimated, and, using a statistically efficient estimator, confidence intervals will be estimated.

3. An hypothesis test of whether the BICR HR is statistically significant will be undertaken, as well as an evaluation of whether it is of clinically meaningful size.

This general strategy was applied to data from a study in breast cancer, which conducted a full BICR to confirm a large and significant improvement in PFS. Results from that application indicate that a strategy of conducting an audit in 20% of the total study population would conclude that the BICR HR is statistically significant 88% of the time. This supports the view that large treatment effects will likely require small BICR audits. Additional simulations indicated that, for moderate effect sizes that are statistically significant, larger audits are needed. Further, when treatment effects are small but statistically significant, the additional variability introduced by BICR may make assurance of a treatment effect through the use of a (complete) BICR impossible.

Conclusion

PFS as an endpoint in oncology is increasingly being employed. Measures to validate and efficiently determine biases inherent in studies employing PFS will greatly enhance the rapid development of new therapies.

FDA Response

PFS, defined as the time from randomization until objective tumor progression or death, is increasingly being used in the approval of oncology drugs and biologics. Compared with the use of OS as a primary endpoint, the use of PFS as a trial endpoint usually allows for the study of smaller patient populations and shorter follow-up. PFS is assessed prior to the introduction of subsequent therapies; hence, differences observed between treatment arms of a randomized trial will not be confounded if crossover occurs at the time of disease progression and the start of new therapies. Disease progression is usually the basis for a change in therapy.

Toxicities of most oncology drugs preclude the effective use of blinding. Disease progression is frequently assessed by an investigator's review of radiological examinations and bias can be introduced if effective blinding is not present. To evaluate if any bias has occurred, blinded, independent review committees (BIRCs) have been used to determine the potential presence of bias, rather than to simply note random discrepancies in disease progression dates between the investigator and the BIRC. Random measurement errors tend to obscure the demonstration of superiority, making "false-positive" conclusions in a clinical trial evaluation less likely.

In the PhRMA PFS Working Group presentation, an audit methodology to examine directional evaluation bias was discussed. Directional evaluation bias is of concern when an investigator systematically records progression early or late for one treatment arm of a randomized trial. For example, false-positive conclusions regarding the efficacy of a treatment resulting from bias would be observed if the investigator consistently called disease progression early for the control arm and/or late in the experimental treatment arm. In either case, this would potentially lead to a falsely optimistic evaluation of the experimental treatment.

Large differential discordance rates between treatment arms (i.e., differences between the investigator's and the BIRC's evaluation of disease progression) raise the suspicion of systematic evaluation bias. The presence of this bias is of concern in clinical trials relying on investigator-determined PFS evaluation in situations in which the success of blinding of the trial is uncertain, as well as in unblinded trials.

In blinded trials, FDA has not recommended the use of a BIRC, since evaluation bias is unlikely to be introduced. In trials where blinding cannot be used or when there is uncertainty of the blinding, the use of a BIRC has been recommended. These blinded reviews usually result in the re-examination of all the disease progression events of all patients.

Strategies examining disease progression events in a limited sample of patients at selected sites, in contrast to all patients at all sites, were looked at by the PhRMA PFS Working Group. The intent of this limited evaluation was to examine differential discordance in reading PFS events between treatments. The absence of any differential discordance would suggest that there is no systematic evaluation bias; that is, the local investigator evaluation provides a reliable estimate of treatment effect. However, if there is a differential discordance, the potential for evaluation bias would need to be considered and further evaluated by comparing a larger sample of the BIRC- and investigator-determined PFS evaluations.

The present strategies for limited evaluation of disease progression events have been examined in simulations and retrospective analyses of completed trials. Pilot studies are being planned to evaluate the prospective implementation of limited evaluations of PFS events by the BIRC to examine differential discordance. These pilot studies will further examine and refine how to select subjects and sites for review, the number of subjects and sites needed for a BIRC review, and the procedures to implement these limited evaluations prior to making recommendations for their use for regulatory purposes.

Acknowledgments

FDA Response provided by Richard Pazdur. Jeff Allen and Rasika Kalamegham of Friends of Cancer Research assisted with drafting.

References

1 Freidlin B, Korn EL, Hunsberger S et al. Proposal for the use of progression-free survival in unblinded randomized trials. J Clin Oncol 2007;25:2122–2126.

2 Williams G, He K, Chen G et al. Operational bias in assessing time to progression (TTP) [abstract 975]. Proc Am Soc Clin Oncol 2002;20:244a.

3 Response Evaluation Criteria in Solid Tumors 1.1 (RECIST 1.1), Updated January 2009. Available at http://www.recist.com/recist-in-practice/01.html, accessed September 3, 2009.

4 Dodd LE, Korn EL, Freidlin B et al. Blinded independent central review of progression-free survival in phase III clinical trials: important design element or unnecessary expense? J Clin Oncol 2008;26:3791–3796.

5 U.S. Department of Health and Human Services. FDA Guidance for Industry: Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics, May 2007. Available at http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM071590.pdf, accessed September 3, 2009.