PANEL 2

Improved Insights into Effects of Cancer Therapies

Raymond DuBois, MD, M.D. Anderson Cancer Center
Donald Berry, PhD, M.D. Anderson Cancer Center
Jim Doroshow, MD, FACP, National Cancer Institute
Paolo Paoletti, MD, Glaxo SmithKline
Richard Pazdur, MD, Food and Drug Administration
Nancy Roach, C3: Colorectal Cancer Coalition

The need for clarity on efficacy endpoints

In an effort to accelerate safe and effective cancer drug development and to decrease the time to drug approval, the oncology community has long sought endpoints other than overall survival (OS) to evaluate new agents. Measures of disease progression, health-related quality of life, patient-reported symptoms, and biomarkers have been proposed and tested in clinical studies, but consensus has not been reached on the role of these endpoints in determining the overall clinical benefit of a therapy. One auxiliary endpoint that has been the focus of particularly intense discussion is progression-free survival (PFS), which employs the RECIST criteria to determine the progression of cancer based on imaging.[1] PFS is the length of time during and after treatment in which a patient is living with a disease that does not worsen, according to the established criteria. RECIST defines progression as an increase of 20 percent in a single dimension on computed tomogram or magnetic resonance imaging. However, concern about potential biases in the measurement of disease progression by radiographic imaging has resulted in debate over the current use of PFS in clinical trials.

Two recent cancer drug approvals were based solely on evidence of PFS, underscoring the need to reach agreement on how this endpoint is defined and interpreted. In December 2005, the U.S. Food and Drug Administration (FDA) approved sorafenib for the treatment of advanced kidney cancer based on an increase in progression-free survival, despite the absence of a statistically significant benefit in overall survival. More recently, the FDA granted accelerated approval for bevacizumab in patients with breast cancer. Data showed that a combination of bevacizumab and paclitaxel nearly doubled median PFS compared with paclitaxel alone, but the secondary endpoint of overall survival in this trial did not reach statistical significance. A review of the data by the FDA will be required for the accelerated approval to be converted into a full approval by the end of 2008. If upcoming trials do not show a survival benefit, then the accelerated approval for breast cancer could be revoked or curtailed until more data are collected.

Lack of clarity around the appropriate use of PFS and other auxiliary endpoints can be a barrier to efficient clinical cancer research, as well as to the review and approval of cancer therapies. This paper will brief conference attendees on the complex issues surrounding measurement of treatment efficacy and propose a set of principles to guide the evaluation and use of auxiliary endpoints. Finally, these principles will be illustrated using PFS as a case study.

Issues around the use of auxiliary endpoints

The gold standard for clinical effectiveness of a given agent is an improvement in a defined endpoint in a randomized clinical trial.[2,3] In oncology, overall survival is most often seen as the best single agreed-upon endpoint. However, randomization rarely occurs in the Phase II setting, as is common in other areas of drug development. Instead, Phase II trials measure the rate of complete and partial response to given agents prior to progression to randomized Phase III trials. Single-arm, historically-controlled Phase II trials are rarely employed outside of oncology. This issue is multifactorial, but this may contribute to the high failure rate of drugs proceeding from the Phase II to the Phase III setting in cancer drug development. Thus, there is a need for endpoints that can quickly detect drug efficacy or failure, in order to avoid unnecessary resource allocation to drugs that will ultimately fail to exhibit a patient benefit.

Overall survival is objectively measured and not prone to the potential investigator biases associated with endpoints that require clinical judgment. However, using overall survival as a primary endpoint significantly slows the rate of cancer drug development. As approved therapies have become increasingly effective at prolonging survival, so too have they prolonged the duration of trials designed to detect that endpoint. Delays in recruitment and follow-up ultimately serve to prolong the regulatory review and approval of newer agents that could provide needed options for cancer patients. Furthermore, overall survival is a crude instrument for measuring the effects of many targeted therapies, which may be designed to work in a subset of patients with specific molecular targets. Thus, a great need exists to define and validate alternative markers of effect.[4]

While the term “surrogate endpoint” has been more commonly used in the literature, “auxiliary endpoint” is a preferable term because the endpoints under investigation are not meant to supplant more conventional endpoints, but rather to be evaluated in conjunction with other endpoints. We define auxiliary endpoints to include the collection of endpoints – other than overall survival – 1used to infer the effects of cancer therapies from clinical trials. Auxiliary endpoints may be primary or secondary endpoints within a trial, and may include progression-free survival (time to progression), response rate, patient-reported outcomes (e.g., quality of life), and biomarkers (e.g., tumor size, circulating tumor cells, and tumor-specific markers). Clearly defining the strengths, limitations, and appropriate uses of auxiliary endpoints could accelerate the development, review, and approval of new treatments.

Principles for the evaluation and use of auxiliary endpoints

We propose three basic principles to consider when selecting auxiliary endpoints for a given trial. First, a strong biological rationale should support the potential auxiliary endpoint as a marker of treatment effectiveness. For example, biomarkers that predict variability in survival time may be preferred endpoint candidates. Second, the potential auxiliary endpoint should be shown to explain variability in treatment outcomes in terms of survival for treated patients.[5] Third, ideal auxiliary endpoints should accurately assess the efficacy of the drug being evaluated with minimal risk of subjectivity or bias. Where the possibility of bias exists, the trial design should compensate by seeking to minimize potential bias.[6]

The development of new drugs for AIDS patients closely follows this model of auxiliary endpoint development. CD4 count and viral load were validated as auxiliary endpoints in trials in the late 1980s and early 1990s, allowing for an explosion in the available therapies for AIDS patients. Applying this model to the oncology community will be more difficult. Since cancer involves many heterogeneous disease processes, many auxiliary endpoints will need to be developed according to these criteria.

Toward a rational and valid process for evaluating progression-free survival

Progression-free survival (PFS) is a desired endpoint in many settings, but it is not a surrogate for overall survival. Advantages of PFS as a primary endpoint include a more rapid clinical trial and the elimination of confounding effects when evaluating experimental therapies in diseases with existing effective therapies. For example, if a patient enters a clinical trial after four failed conventional therapies and later discontinues that trial due to progression, numerous other approved therapies may be available. The patient could survive long after a given trial ceased accruing patients, and other therapies could contribute to his or her demise, minimizing the effect of the experimental therapy. Further, non-trivial improvements in PFS are considered a clinical benefit in some settings. Patients may see a benefit in a lack of progression in their tumor burden, irrespective of the benefits in overall survival.

Although PFS has many advantages, it is not without limitations. Unlike OS, the precise timing of PFS is not known. This leads to the potential for evaluation-time bias, which produces biased estimates of treatment effectiveness when the evaluation times for progression status differ by treatment arm.[7] Further, elements of subjectivity remain in spite of efforts, such as RECIST, to standardize the evaluation of progressive disease.[8] Indeed, a non-trivial number of discrepancies between radiologists evaluating the progression status of the same patients are to be expected. These discrepancies can come from multiple sources. At the start of trials, baseline lesions are usually defined, but occasionally, these lesions are altered or ignored in the course of a trial. Radiographic scans can be misplaced, leading to clinical judgments based on varied amounts of radiographic information. In addition, radiologists may have different interpretations of the available scans. In a recent trial, the discrepancy rate between two expert radiologists blinded to the treatment assignment was 34 percent. When these discrepancies are unrelated to treatment, they are a source of measurement variability, which results in attenuated estimates of effect sizes. Measurement variability reduces the power to detect a true difference but will not lead to invalid conclusions when the experimental therapy is truly ineffective. In other words, measurement variability alone will not result in ineffective therapies entering the oncology community. However, if the variability is large enough, it could preclude effective therapies from being revealed.

The most significant concern about discrepancies in assessment arises when progressive disease evaluations are influenced by an investigator's lack of objectivity about the therapies under study. The potential for evaluations to be influenced by knowledge of treatment assignment, combined with pre-existing views about their relative effectiveness, has led to the introduction of Blinded Independent Central Reviews (BICR) as a suggested means of validating efficacy in trials with PFS endpoints. However, the use of BICR is problematic and may lead to invalid analyses, as it does not always provide an unbiased estimate of a treatment's effectiveness. Specifically, BICR analyses for PFS are likely subject to the presence of informative censoring, which invalidates standard analyses. The methodology relies on the assumption that censoring is independent of factors associated with progression or survival.

Informative censoring arises in the following manner: Patients who progress by investigator assessment may not have the same assessed time of progression under the BICR. Once a patient has progressed according to the investigator, he or she will be taken "off protocol" and further follow-up is not likely. If the BICR does not determine that a patient has progressed by the time the patient is off protocol, the patient is censored for the purpose of analysis. This patient, however, is more likely to have progressed, as assessed by BICR, sooner than those remaining in the at-risk cohort. This violates the standard assumptions for censoring subjects and, as a result, survival-analysis estimates are biased. Further, although methods for modeling informative censoring exist, these methods cannot conclusively eliminate the potential effects of informative censoring. Dodd et al. provide a more detailed discussion of this issue with an example from a clinical trial.[9]

In a review of Phase III oncology trials published in the last five years that had BICRs as a component of assessment, no cases were shown to have substantial differences between analyses between the BICR and investigator assessments.[10] (See Tables 1 and 2.) The lack of differences is striking in light of the seemingly high discrepancy rates between BICR and local review, which range from 36 to 53 percent. However, these discrepancies are likely due in large part to random, rather than systematic, differences between the clinicians who evaluate the radiographic imaging scans. This variation in assessment between two independent reviewers is a well-studied phenomenon in many therapeutic areas.[11] Further, there was no trend that would indicate that either BICR or local review resulted in a stronger treatment effect.

In conclusion, BICR does not necessarily provide a less biased estimate of a treatment's effectiveness than local review, and situations in which the BICR conclusions differ from those based on the investigators' assessments result in an ambiguous situation. The discrepancy may be caused by measurement variability, informative censoring, or true evaluation bias. Methods that effectively reduce evaluation bias where it is most likely to affect trial outcomes are needed. Four approaches are worthy of consideration.

Proposals for the auditing of PFS

Matter for clarification: No BICR when trials are double-blinded.

Blinding of treatment assignment would eliminate systematic bias in PFS evaluation related to knowledge of treatment assignment. Therefore, there should be no requirement for central review in double-blinded trials, except in the case where an extreme imbalance between treatment arms in the incidence of side effects could lead to a considerable level of unblinding. This level of imbalance would be characterized by the majority of patients in the treatment arm experiencing a particular side effect with a virtual absence of this same side effect in the control arm.

Case 1: An open-label superiority trial with an BICR-based audit of progression.

Detection of meaningful evaluation bias will be gauged via an audit of progression determinations. BICR could be performed in both arms of a trial on a subset of cases. A sample size for the audit would be specified in advance (for example, 10 percent of participants or a minimum number of cases). If bias is suspected, then the audit would expand to a larger proportion of cases. The goal of the audit would be to determine whether there is a meaningful difference in hazard ratios between the local review and BICR. It is recognized that, given the potential biases present with both BICR and local review, a discrepancy in assessments would make a conclusion about a treatment's efficacy more difficult.

Large effect sizes will likely be robust to small discrepancies between treatment arms, while smaller effect sizes will be quite sensitive to small discrepancies. Therefore, when effect sizes are large, relatively smaller audits may be necessary to detect the amount of bias needed to alter the trial conclusion substantively. However, in some cases, no audit, no matter its size, can rule out evaluation bias.

Because the goal of the audit is to detect actual bias, measurement variability should be controlled. Technologies that enable synchronization, allowing patients to be followed by the BICR as a trial is ongoing, are strongly encouraged.

It is recognized that data-driven analyses are necessary to develop the scientific justification for the selection of the recommended audit size. The Phamaceutical Researchers and Manufacturers of America (PhRMA) and the National Cancer Institute (NCI) have begun research projects to address this specific issue. The NCI will be collecting patient-level data from multiple large clinical trials with data from both central and local reviews to better inform the audit process. Since it not expected that these data sets will contain meaningful bias, such bias will be introduced into the data so that the auditing strategy can be tested. Simulation studies, based on an understanding of the trial data, will also inform recommendations. Clearly, an understanding of what is an important level of bias for a particular study given an observed effect size is needed.

Case 2: An open-label superiority trial with large effect size.

When treatment effect sizes are large enough, an audit is not necessary, since evaluation bias is not expected to be of a magnitude that would meaningfully impact the observed effect size. As part of this proposal, increased monitoring of the protocol-specified imaging procedures at the local site could be undertaken. It is expected that the investigator is the greatest potential source for bias in a PFS assessment. It should also be noted that a local radiologist is frequently unaware that patients are participating in a clinical trial, so a procedure that records the measurements or progression assessments of both the radiologist and the investigator is recommended. Whenever the investigator overrides the radiologist's determination, the reasons for this will be documented. When this occurs more frequently in one treatment arm and the reasons are not easy to verify objectively, concern about bias will arise.

Case 3: PFS evaluation at two time points with auditing at these evaluation times.

Evaluation of treatment effectiveness could be based on the proportion of patients whose cancer has progressed at two time points, rather than using an analysis based on a survival model. Two time points for imaging assessments would be determined prospectively, corresponding to the approximate median PFS and approximately twice median PFS of the control arm or conventional therapies. Summary statistics would include the proportion alive and progression-free at each time point. Progressions that have been documented prior to the designated imaging assessment time would be counted as an event for the rate of progression or death, and images would be audited at the two time points. For patients who progress prior to the designated scan times, the audit would be based on the scan that determined progressive disease.

This two-point approach reduces evaluation-time bias and results in a simpler trial design.[12] Since the approach limits the focus to the two imaging assessments, the issues of compliance, timing, rigor, and consistency are easier to maintain or verify. Further, central review of two time points should be easier to implement. While one might have concerns about a loss in power of the trial design as compared to a log-rank analysis, the loss in power with two time points is less than that from a single time point. Indeed, Freidlin et al. demonstrate that there is little risk in major power loss from this approach.[13] The trade-off for some loss in power, however, is decreased susceptibility to bias.

Conclusion

This paper has presented a proposal for auditing PFS in three different scenarios. Establishing such auditing procedures can help build confidence in PFS as an indicator of clinical benefit. The suggestions listed above also hint at a way forward for improving the reliability of information produced by other auxiliary endpoints. If the cancer research community can determine how to most effectively utilize auxiliary endpoints – without compromising the quality of safety and efficacy data – cancer patients will benefit greatly.

Table 1. Trials that have used retrospective blinded independent central reviews[a]

Disease	Trial	Sample size	Hazard ratio and 95% confidence interval per central review	Hazard ratio and 95% confidence interval per local review
Renal Cell Carcinoma	sorafenib vs. placebo[14], b	903	0.44 (0.35-0.55)	0.51 (0.43-0.60)
	sunitinib vs. interferon alpha[15]	750	0.42 (0.32-0.54)	0.42 (0.33-0.52)
Colorectal cancer	panitumumab plus best supportive care vs. best supportive care[16]	463	0.54 (0.44-0.66)	0.39 (0.32-0.48)
Breast Cancer	lapatinib plus capecitabine vs. capecitabine[17]	324	0.49 (0.34-0.71)	0.59 (0.42-0.84)
	bevacizumab plus capecitabine vs. capecitabine[18]	462	0.98 (0.77-1.25)	0.90 (0.72-1.12)
	ixabepilone plus capecitabine vs capecitabine[19]	752	Median PFS[c] 5.8(5.45-6.97) vs 4.2(3.81-4.50) months	Median PFS[c] 5.3 vs 3.8 months[d]
	bevacizumab plus paclitaxel vs paclitaxel[20, 21]	722	0.48 (0.39, 0.61)	0.42 (0.34,0.52)

Notes for Table 1

[a] We reviewed the literature and searched PubMed for studies in breast cancer, colorectal cancer, lung cancer and renal cell carcinoma. Search terms included, "progression free survival" or "time to progression," with filters of "randomized controlled trial" and "published in last five years." This revealed 209 manuscripts, of which only six reported having a central review of progression. The bevacizumab plus paclitaxel trial in breast cancer (last row) was included separately because it generated much discussion during an FDA Oncologic Drug Advisory Committee meeting on Dec. 5, 2007. All of these trials implemented a retrospective BICR. The panitumimab trial allowed cross-over at the time of locally determined progression amongst patients receiving the control treatment. As a result, patients for whom progression was not confirmed centrally continued to be evaluated centrally for progression.

[b] Double-blinded trial

[c] Hazard ratios not reported for local review.

[d] Difference statistically significant ($p<0.0011$). 95% CI for median PFS not reported.

Table 2. Discrepancy rates for three trials with central review

	Discrepancy Rate in Assignment of Progression/ Censoring Datea	Discrepancy Rate in Assignment of PFS Status	Per Central Review		Per Local Review	
			HR	95% CI	HR	95% CI
Lapatinib plus capecitabine v capecitabine[22]	Lapatinib plus capecitabine 87 of 163 = 53% Capecitabine, 69 of 161 = 43%	Lapatinib plus capecitabine, 40 of 163 = 25% Capecitabine, 40 of 161=25%	0.49	0.34 to 0.71	0.59	0.42 to 0.84
Bevacizumab plus paclitaxel v paclitaxel[23, 24]	Bevacizumab plus paclitaxel 118/330 = 35.7% Paclitaxel, 114/319 = 35.7%b, 25	Bevacizumab plus paclitaxel 90 of 368 = 24.5% Paclitaxel, 84 of 354 = 23.7%c	0.48	0.39 to 0.61	0.42	0.34 to 0.52
Bevacizumab plus capecitabine v capecitabine[25]	Bevacizumab plus capecitabine, 88/232=38%, Capecitabine, 99/230=43%27	Both arms combined: 105 of 462 = 23%	0.98	0.77 to 1.25	0.90[28]	0.72 to 1.12[29]

Notes for Table 2

[a] Computed as agreement in date of progression or date of censoring.

[b] Estimated amongst the 649 (of 722) patients for whom images were available for central review. An agreement was counted if dates were within 6 weeks of one another. This is in contrast to the lapatinib plus capecitabine and bevacizumab plus capecitabine trials, in which exact date was used for agreement.

[c] A discrepancy was counted if either status assignment differed or if no image was available for central review. As a result, a total of 722 (and not 649) patients were included.

References

[1] Therasse, P. et al. New Guidelines to Evaluate the Response to Treatment in Solid Tumors. *J. Natl. Cancer Inst.* 92, 205-216 (2000).

[2] Freidlin, B. et al. Proposal for the Use of Progression-Free Survival in Unblinded Randomized Trials. *J Clin Oncol.* 25, 2122-2126 (2007).

[3] Ratain, M.J. et al. Recommended changes to oncology clinical trial design: Revolution or evolution? *Eur. J. Cancer.* 44, 8-11 (2008).

[4] Schilsky, R.L. End Points in Cancer Clinical Trials and the Drug Approval Process. *Clin. Cancer Res.* 8, 935-938 (2002).

[5] Ellenberg, S.S. Surrogate end points in clinical trials. *BMJ*. 302, 63-64 (1991).

[6] Pazdur, R. Endpoints for Assessing Drug Activity in Clinical Trials. *Oncologist* 13 suppl 2, 19-21 (2008).

[7] Freidlin, B. et al., 2007.

[8] Therasse et al., 2000.

[9] Dodd, L.E., et al. Blinded Independent Central Review of Progression-Free Survival in Phase III Clinical Trials: Important Design Element or Unnecessary Expense? *J. Clin. Oncol.* 26, 3791-3796 (2008).

[10] Dodd, L.E. et al., 2008.

[11] Feinstein, A.R. A bibliography of publications on observer variability. *J. Chronic Disease.* 38, 619-632 (1985).

[12] Freidlin, B. et al., 2007.

[13] Friedlin, B. et al., 2007.

[14] Escudier, B. et al. Sorafenib in advanced clear-cell renal-cell carcinoma. *N. Engl. J. Med.* 356, 125-134 (2007).

[15] Motzer, R.J. et al. Sunitinib versus interferon alfa in metastatic renal-cell carcinoma. *N. Engl. J. Med.* 356, 115-124 (2007).

[16] Van Cutsem, E. et al. Open-Label Phase III Trial of Panitumumab Plus Best Supportive Care Compared with Best Supportive Care Alone in Patients With Chemotherapy-Refractory Metastatic Colorectal Cancer. *J. Clin. Oncol.* 25, 1658-1664 (2007).

[17] Geyer, C.E. et al. Lapatinib plus capecitabine for HER2-positive advanced breast cancer. *N. Engl. J. Med.* 355, 2733-2743 (2006).

[18] Miller, K.D. et al., Randomized Phase III Trial of Capecitabine Compared with Bevacizumab plus Capecitabine in Patients with Previously Treated Metastatic breast Cancer. *J. Clin. Oncol.* 23, 792-799 (2005).

[19] Thomas, E.S. et al. Ixabepilone Plus Capecitabine for Metastatic Breast Cancer Progressing after Anthracycline and Taxane Treatment. *J. Clin. Oncol.* 25, 5210-5217.

[20] U.S. Food and Drug Administration. *FDA Briefing Document: Oncology Drug Advisory Committee Meeting BLA STN 125085/91.018 Avastin® (bevacizumab).* (2007).

[21] Genentech, Oncology Drugs Advisory Committee Meeting: 5 December 2007.

[22] Geyer, C.E. et al., 2006.

[23] U.S. Food and Drug Administration, 2007.

[24] Genentech, Oncology Drugs Advisory Committee Meeting: 5 December 2007.

[25] Personal communication. Suman Bhattacharya, PhD, Bio-oncology, Genentech.

[26] Miller, K.D. et al., 2005.

[27] Personal communication, Suman Bhattacharya.

[28] Personal communication, Suman Bhattacharya.

[29] Personal communication, Suman Bhattacharya.