DOI:10.1093/jnci/djt282
Advance Access publication October 17, 2013

COMMENTARY

Statistical and Practical Considerations for Clinical Evaluation of Predictive Biomarkers

Mei-Yin C. Polley, Boris Freidlin, Edward L. Korn, Barbara A. Conley, Jeffrey S. Abrams, Lisa M McShane

Manuscript received February 28, 2013; revised August 8, 2013; accepted August 13, 2013.

Correspondence to: Mei-Yin C. Polley, PhD, Biometric Research Branch, National Cancer Institute, National Institutes of Health, 9609 Medical Center Dr, Rm 5W638, MSC 9735, Bethesda, MD 20892–9704 (e-mail: polleymc@mail.nih.gov).

Predictive biomarkers to guide therapy for cancer patients are a cornerstone of precision medicine. Discussed herein are considerations regarding the design and interpretation of such predictive biomarker studies. These considerations are important for both planning and interpreting prospective studies and for using specimens collected from completed randomized clinical trials. Specific issues addressed are differentiation between qualitative and quantitative predictive effects, challenges due to sample size requirements for predictive biomarker assessment, and consideration of additional factors relevant to clinical utility assessment, such as toxicity and cost of new therapies as well as costs and potential morbidities associated with routine use of biomarker-based tests.

J Natl Cancer Inst;2013;105:1677–1683

Precision medicine aims to guide therapy informed by biological characterization of a patient's disease. For patients with cancer, these characterizations are typically achieved by molecular analysis of tumor biomarkers or sometimes by examination of host characteristics, such as variations in germline DNA. Two classes of biomarkers in oncology are prognostic markers and predictive markers (1): prognostic markers inform about likely disease outcome independent of the treatment received, and predictive markers provide information about likely outcomes with application of specific interventions. Therefore, predictive markers can help select among two or more therapy options. Predictive markers are of particular importance for targeted therapies, which often are expected to benefit only patients whose disease is characterized by presence of a biomarker.

When there is strong evidence that a targeted agent benefits only patients whose tumors have a particular biomarker, then evaluation of the targeted therapy in a trial that enrolls only patients whose tumors are positive for that marker would be appropriate. This type of trial design is called an enrichment design (2), and the marker is classified as an enrichment or selection marker (3). If there is uncertainty about the predictive biomarker and reasonable possibility that the therapy could benefit a broad population of patients, one can evaluate a candidate predictive biomarker by studying a group of unselected patients treated with the targeted therapy and another group of patients treated with some alternative therapy. The biomarker is predictive if the relative efficacy of the two treatments is different for the biomarker-positive patients than for the biomarker-negative patients. (Note that just demonstrating that biomarker-positive patients have better clinical outcome than biomarker-negative patients on a cohort of patients uniformly treated with the targeted therapy is insufficient to demonstrate that the biomarker is predictive because it may solely be prognostic.) A familiar example of a tumor marker that is both prognostic and predictive is estrogen receptor status in breast cancer, for which a positive status is associated with a more favorable prognosis generally and with specific benefit from endocrine therapy. Although much of the discussion here is framed around evaluation of predictive biomarkers for targeted therapies in retrospective studies, the same general principles would apply for evaluation of a predictive biomarker for selection between two different standard (nontargeted) therapies and for designing prospective clinical trials to evaluate predictive biomarkers.

The best setting in which to evaluate a predictive biomarker for an experimental targeted therapy is a randomized clinical trial (RCT) of the targeted therapy vs a standard treatment, where the biomarker status is obtained on the patients but not used to direct treatment (4). (Other possible RCT designs to evaluate biomarkers are possible but not as efficient as this design (2).) Ideally, a biomarker would be assessed prospectively in an RCT of the targeted agent. However, biomarker development often lags behind therapeutic development. The reasons for this asynchrony may include an incomplete understanding of the mechanism of action of a drug, the uncertainty about what form of a marker is most relevant (eg, DNA mutation, mRNA expression, protein expression), and technical difficulties with marker assay development (5). These reasons contribute, in part, to why biomarker studies are frequently conducted retrospectively on archived specimen collections several years after a cancer therapy has been developed. With a careful prospective design, such a retrospective analysis can provide convincing evidence in support of a predictive biomarker (6).

We assume that the assay for the biomarker or biomarker signature (a collection of biomarkers that are combined through some mathematical model and linked to a biological or clinical outcome) yields a fully specified binary marker [or is continuous and has been dichotomized with a cutoff based on appropriate statistical procedures (7,8)] that has demonstrated acceptable

analytical performance and sufficient robustness against influences of preanalytical factors (eg, warm ischemia time or duration of time that the paraffin blocks are stored before analysis) (9). Preanalytic factors may be of special concern because of the desire to extrapolate from retrospective biomarker analyses to how the biomarker will work in the clinic when the specimens are processed contemporaneously (6).

Under the assumptions that the biomarker is biologically relevant and analytically validated, we first define and describe in the next section predictive biomarkers that display clinical qualitative and quantitative interactions with the treatments. In the following section, we discuss the sample sizes required to reliably assess these interactions and discuss how to appropriately interpret estimates of predictive effect. This is followed by a discussion of the clinical utility of a predictive biomarker, which goes beyond whether or not a statistically significant interaction is present. We end with a brief discussion.

Qualitative and Quantitative Interactions

A biomarker is predictive if the treatment effect (experimental *E* compared with standard *S*) is different for the biomarker-positive patients compared with the biomarker-negative patients. This is known as a (statistical) interaction between the treatment effect and biomarker status. An interaction can be qualitative or quantitative (10). Qualitative interaction can help guide treatment choice: *E* is better than *S* for the biomarker-positive patients and *E* is not better than *S* for the biomarker-negative patients (for which it could be equally efficacious or worse). With a quantitative interaction, *E* is better than *S* for both biomarker groups, but the amount of the treatment benefit is different for the two biomarker groups (otherwise there would be no interaction of any kind).

Examples of qualitative interactions are given in Figure 1. Gefitinib is better than chemotherapy as first-line treatment in epidermal growth factor receptor (*EGFR*) mutation–positive non–small cell lung cancer (NSCLC) patients (hazard ratio [HR] =0.48; 95% confidence interval [CI] = 0.36 to 0.64; $P < .001$) (Figure 1A) but worse than chemotherapy in *EGFR* mutation–negative patients (HR = 2.85; 95% CI = 2.05 to 3.98; $P < .001$) (Figure 1B); the interaction is statistically significant with *P* less than .001 for the progression-free survival endpoint (11). A second example of qualitative interaction is given by the RCT of cetuximab + chemotherapy vs chemotherapy alone for first-line treatment of *EGFR* immunohistochemistry (IHC)–positive NSCLC patients (12). Based on the observed *EGFR* and outcome data, a division of the patients into low and high *EGFR* expression demonstrated a qualitative interaction: the addition of cetuximab improved overall survival for the patients with tumors with high *EGFR* expression (HR =0.73; 95% CI = 0.58 to 0.93; $P = .01$) (Figure 1C) but offered no benefit for patients with tumors with low *EGFR* expression (HR = 0.99; 95% CI = 0.84 to 1.16; $P = .88$) (Figure 1D); the interaction is statistically significant with *P* equal to .04 (13). (Because the *EGFR* cutpoint was not prespecified, these results would need to be validated on a new dataset.)

Examples of quantitative interactions are given in Figure 2. Pazopanib improves progression-free survival (PFS) relative to placebo in locally advanced or metastatic renal cell carcinoma patients with high interleukin 6 (IL-6) plasma concentrations (Figure 2A) and low IL-6 concentrations (Figure 2B), but more so for the high IL-6 subgroup (HR = 0.31; 95% CI = 0.31 to 0.58; $P < .0001$) than for the low IL-6 subgroup (HR = 0.55; 95% CI = 0.28 to 0.71; $P < .002$); the interaction is statistically significant ($P = .009$) (14). A second example of a quantitative interaction is given by the interaction between maintenance treatment of NSCLC with erlotinib and tumor *EGFR* mutation status: wild-type (Figure 2C) vs mutant (Figure 2D) (15). There is an interaction (statistically significant with $P < .001$), and the interaction is quantitative: erlotinib improves PFS for patients with wild-type *EGFR* tumors (HR = 0.78; 95% CI = 0.63 to 0.96; $P = .02$) and for patients with *EGFR*-mutant tumors (HR = 0.10; 95% CI = 0.04 to 0.25; $P < .0001$), but much more for the patients with *EGFR*-mutant tumors.

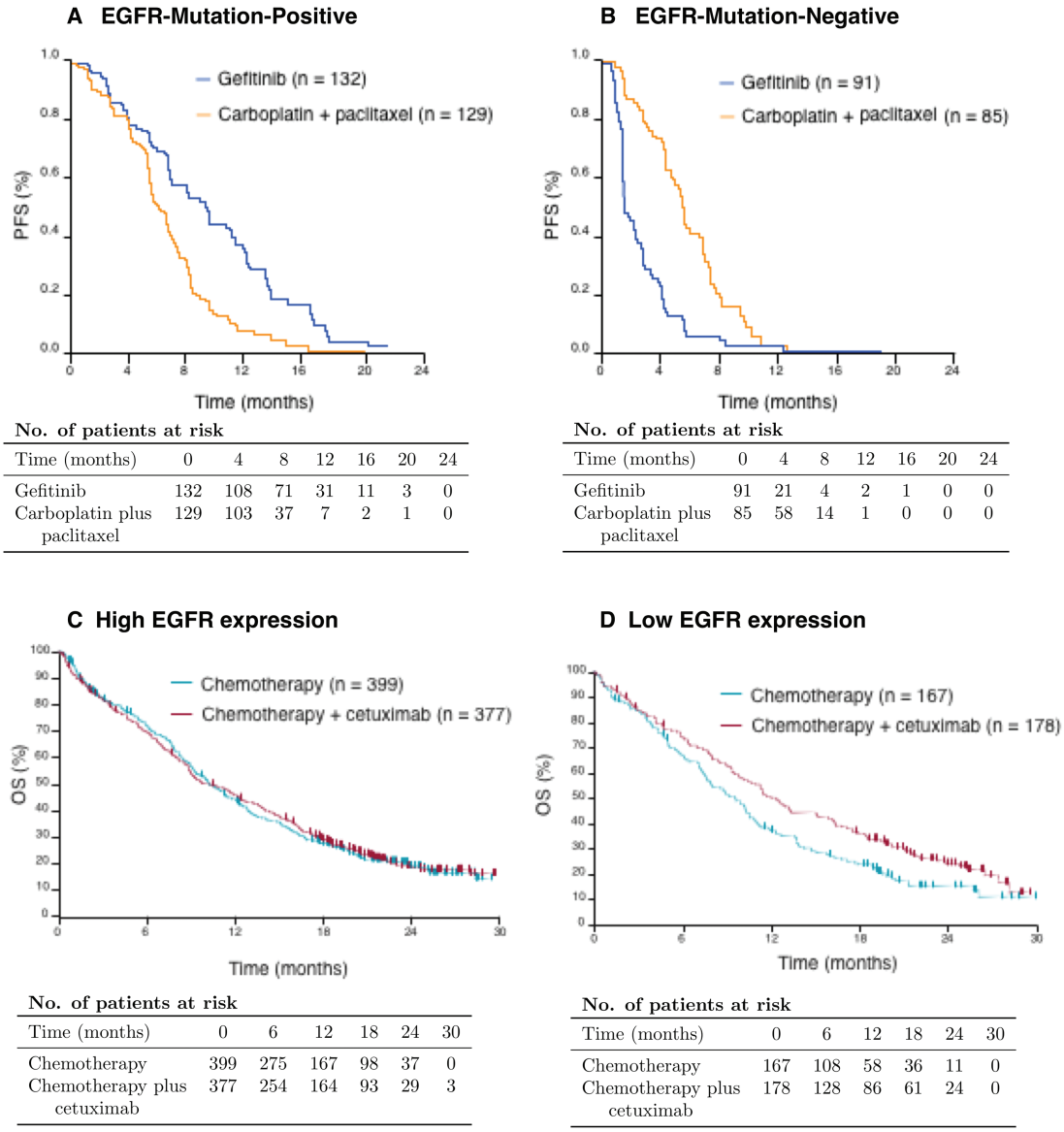
Qualitative interactions can provide a clear indication of treatment choice: patients who are biomarker-negative should get the standard treatment (*S*), and, when the experimental treatment (*E*) is sufficiently better than *S* in the biomarker-positive subgroup, these patients should receive *E*. Whether a biomarker that has quantitative interaction with treatment would be useful in directing treatment is a more difficult question and is discussed below.

Sample Sizes and Ability to Draw Conclusions

In designing an RCT to assess an experimental treatment vs a control treatment, one typically designates a minimally clinically interesting treatment difference (for example, a target hazard ratio of 0.75) that one would not want the trial to miss detecting. One then chooses the sample size of the trial so that there will be a high probability of having a statistically significant result at the end of the trial (the power of the trial) if the true treatment difference is this target treatment difference or more extreme.

One approach for testing whether a biomarker is a predictive biomarker is to estimate the statistical interaction and to assess whether that interaction is statistically significant. Unfortunately, the required sample size to test whether an interaction is statistically significant can be much larger than the required sample size to assess a treatment difference. For example, with a biomarker with 50% positive prevalence, it requires approximately four times the sample size to detect a 0.75 interaction (defined, for example, by a hazard ratio of 0.75 in a biomarker-positive subgroup and a hazard ratio of 1.00 in a biomarker-negative subgroup) as it does to detect a treatment difference with a hazard ratio of 0.75 (16). If the biomarker positivity prevalence is not 50%, the required sample size will be even larger. For example, if 20% of the patients have positive biomarkers and 80% have negative biomarkers, then the required sample size to detect the 0.75 interaction would be more than six times the required sample size to detect the same size treatment difference. If the investigator is using data and tissue from a retrospective RCT, not all patients will have tumor tissue available and the assay will not be successful for all tissues, further reducing the effective sample size.

A potential solution to the sample size problem is to pool biomarker data from multiple treatment trials that are asking similar questions. This approach is consistent with published guidelines for conducting prospective–retrospective studies that include a requirement that similar results for a biomarker be observed in



No. of patients at risk

Time (months)	0	4	8	12	16	20	24
Gefitinib	132	108	71	31	11	3	0
Carboplatin plus paclitaxel	129	103	37	7	2	1	0

No. of patients at risk

Time (months)	0	4	8	12	16	20	24
Gefitinib	91	21	4	2	1	0	0
Carboplatin plus paclitaxel	85	58	14	1	0	0	0

No. of patients at risk

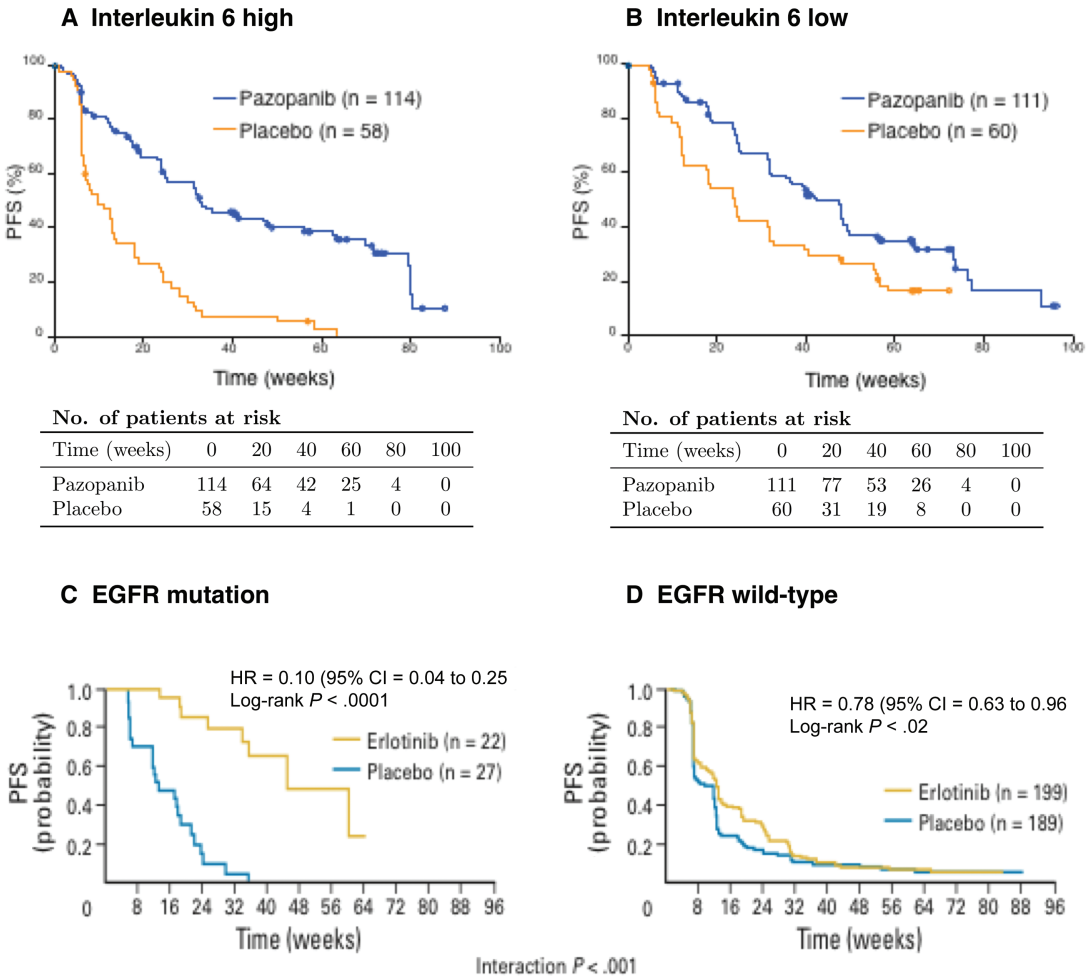
Time (months)	0	6	12	18	24	30
Chemotherapy	399	275	167	98	37	0
Chemotherapy plus cetuximab	377	254	164	93	29	3

No. of patients at risk

Time (months)	0	6	12	18	24	30
Chemotherapy	167	108	58	36	11	0
Chemotherapy plus cetuximab	178	128	86	61	24	0

Figure 1. Examples of qualitative interactions. Gefitinib vs carboplatin + paclitaxel for first-line treatment of non–small cell lung cancer patients with *EGFR* mutation–positive tumors (**A**) and *EGFR* mutation–negative tumors (**B**) [adapted from Figure 2 of Mok et al (11). Reprinted with permission. Copyright 2009 Massachusetts Medical Society.]. Cetuximab + chempotherapy vs chemotherapy for first-line treatment of non–small cell lung cancer patients with high-expressing *EGFR* immunohistochemistry (IHC)–positive tumors (**C**) and low-expressing *EGFR* IHC–positive tumors (**D**) [adapted from Figure 4 of Pirker et al. (13). Reprinted with permission. Copyright 2012 Elsevier]. PFS = progression-free survival.

at least two comparable studies for a biomarker to reach level IB evidence for medical utility (6). For example, based on a meta-analysis of multiple trials of the anti-*EGFR* monoclonal antibodies cetuximab or panitumumab for metastatic colorectal cancer, it was shown that these agents provide no benefit for patients with *KRAS*-mutated tumors (17). However, caution in pooling is required because a biomarker may have a qualitative interaction with treatment in one setting but not in another "similar" setting. For example, *EGFR* mutation status has a qualitative interaction with the treatment choice of the *EGFR* tyrosine kinase inhibitor gefitinib vs chemotherapy in first-line treatment of NSCLC (Figure 1, A and B) but only a quantitative interaction with the choice of *EGFR* tyrosine kinase inhibitor erlotinib vs placebo as maintenance therapy for NSCLC (Figure 2, C and D). Another cautionary example is the situation when a predictive biomarker in one disease setting may not be predictive in another disease setting. For example, although *KRAS* mutation has been shown to be a negative predictive factor for benefit from anti-*EGFR* monoclonal antibodies in colorectal cancer (17), data to date have not supported its predictive value in NSCLC (18,19). Although because of limited sample sizes there will not be sufficient power to detect the minimally clinically interesting treatment difference in each biomarker subgroup, there may be power to detect a larger interaction effect that is still considered plausible. For example, suppose a targeted agent is tested in an unselected population with the overall treatment hazard ratio of 0.75. One expects the treatment only to work in the biomarker-positive patients that comprise one-half of the population and not to work in the one-half of the



No. of patients at risk

Time (weeks)	0	20	40	60	80	100
Pazopanib	114	64	42	25	4	0
Placebo	58	15	4	1	0	0

No. of patients at risk

Time (weeks)	0	20	40	60	80	100
Pazopanib	111	77	53	26	4	0
Placebo	60	31	19	8	0	0

Figure 2. Examples of quantitative interaction: pazopanib vs placebo for locally advanced or metastatic renal cell carcinoma patients with high interleukin 6 (IL-6) values (**A**) and low IL-6 values (**B**) [adapted from Figure 2 of Tran et al. (14). Reprinted with permission. Copyright 2012 Elsevier]. Erlotinib maintenance therapy vs placebo for non–small cell lung cancer patients with *EGFR* mutation–positive tumors (**C**) and *EGFR* wild-type tumors (**D**) [adapted from Figure 3 of Brugger et al. (15). Reprinted with permission. Copyright 2011 American Society of Clinical Oncology]. Note that data were not available from Brugger et al. (15) to provide the number of patients at risk for (**C**) and (**D**). CI = confidence interval; HR = hazard ratio; PFS = progression-free survival.

patients who are biomarker negative (ie, hazard ratio = 1.0 in the biomarker-negative subgroup). Then, the treatment hazard ratio in the biomarker-positive patients would have to be approximately 0.56 for the overall hazard ratio to be 0.75 (log 0.75 = [log 0.56 + log 1.00] / 2). This corresponds to an interaction of 0.56, and a required sample size that is the same as the treatment trial (not four times larger) to achieve the same power. However, note that the analysis in this situation will not have much power to detect an interaction corresponding to a clinically meaningful treatment effect in the biomarker-positive patients (HR = 0.75) with no effect in the biomarker-negative patients. Therefore, it is important to describe the sample size limitations in reports of retrospective biomarker analyses, which are best conveyed by confidence intervals for the treatment effects in the biomarker subgroups.

Example of Difficulties Arising From Limited Sample Sizes

We consider the RCT of temozolomide + radiotherapy vs radiotherapy alone for glioblastoma multiforme (20), for which the biomarker *MGMT* methylation status was studied on a subset of the patients for whom tumor specimens and assay results were available (n = 206 patients of the 573 randomized) (21). The biomarker-stratified results are displayed in Table 1 and show a statistically significant PFS benefit of the addition of temozolomide for both biomarker strata. There was a statistically significant survival benefit from temozolomide in the *MGMT* methylated subgroup (P = .007 based on two-sided log-rank test), but this benefit did not reach statistical significance in the *MGMT* unmethylated subgroup (P = .06 based on two-sided log-rank test). However, because patients with unmethylated *MGMT* did relatively poorly regardless of treatment and the temozolomide benefit is less for patients with unmethylated *MGMT* (ie, larger hazard ratios), Hegi et al. (21) conclude, "Patients with glioblastoma containing a methylated *MGMT* promoter benefited from temozolomide, whereas those who did not have a methylated *MGMT* promoter did not have such a benefit." Given the wide confidence interval for the hazard ratio in the unmethylated group, one cannot confidently reach a conclusion from these data alone (21) about the potential value of *MGMT* methylation status as a predictive marker for benefit from temozolomide in patients with glioblastoma. Further

Table 1. Outcome of randomized clinical trial of temozolomide + radiotherapy (T+R) vs radiotherapy (R) for glioblastoma patients stratified by *MGMT* methylation status*

MGMT status	No. of events/ No. of patients (T+R vs R)	Hazard ratio (95% CI)	*P*†
Overall survival			
Methylated	27/46 v 38/46	0.51 (0.41 to 0.84)	.007
Unmethylated	52/60 v 53/54	0.69 (0.47 to 1.02)	.06
			$P_{interaction}$ = .29‡
Progression-free survival			
Methylated	40/46 v 45/46	0.48 (0.31 to 0.75)	.001
Unmethylated	53/60 v 54/54	0.62 (0.42 to 0.92)	.02
			$P_{interaction}$ = .38‡

* Results in this table are from Hegi et al. (21) except the interaction *P* value for progression-free survival. CI = confidence interval.

† *P* values are based on two-sided log-rank tests.

‡ The interaction *P* value for overall survival was reported in Table 2 of Hegi et al. (21) based on a two-sided test of interaction between treatment and *MGMT* methylation status in the Cox proportional hazards model. The two-sided interaction *P* value for progression-free survival was calculated by us based on the asymptotic normality of the log ratio of the hazard ratios (16).

clinical follow-up on these patients resulted in demonstration of a statistically significant overall survival benefit in the unmethylated *MGMT* group [HR = 0.6; 95% CI = 0.4 to 0.8; Table 2 in Stupp et al. (22)], with an apparent long-term survival benefit (Figure 3). A rough calculation based on the information provided in Stupp et al. (22) suggests that the hazard ratio in the methylated *MGMT* group is approximately 0.6 (95% CI = 0.4 to 0.9). These calculations are consistent with Stupp et al. (22) who reported, "Survival was significantly longer in patients treated with temozolomide and radiotherapy than in patients treated with radiotherapy alone, both in patients with a methylated and unmethylated *MGMT* promoter." Although *MGMT* methylation status is highly prognostic, currently available data do not provide sufficient evidence that *MGMT* methylation status is predictive. Recognizing that the nature of the treatment effects in the two *MGMT* subgroups are quite different (Figure 3), evaluation of this biomarker in other studies will be needed to provide further clarity to its potential as a predictive biomarker. This example highlights the challenges associated with establishing the predictive value of a biomarker using data from a previously completed RCT. Specifically, such studies often lack statistical power to reliably ascertain whether there is a treatment effect in the biomarker-negative subgroup. As such, statistical significance in one biomarker subgroup but not in the other biomarker subgroup is insufficient for establishing the predictive value of a biomarker. Such evaluations should be done in conjunction with a critical examination of the estimate and the confidence interval of the treatment difference in both biomarker subgroups, with the recognition that *P* values are influenced by available sample size and number of events.

From Interaction to Clinical Utility

The most important clinical question that needs to be answered is whether the biomarker can classify patients into biomarker subgroups for which different choices of treatment regimen may be indicated. Although the presence of a treatment-by-biomarker interaction provides some insights into whether differential degrees of treatment benefit exist in two biomarker-defined subgroups (ie, the biomarker is predictive), it does not by itself inform us which specific treatment is superior for each biomarker subgroup. In particular, the recommended treatment in the biomarker-negative subgroup may not be clear in the case of a quantitative interaction because there will be treatment benefit in that subgroup but it may be relatively modest compared with the benefit seen in the biomarker-positive group. In this case, it is useful to incorporate into the decision-making process other practical factors such as toxicities induced by the therapies, the cost of the therapies, and the morbidity/cost associated with routine use of a biomarker assay to identify the biomarker-negative patients. For example, in the study by Brugger et al. (15), erlotinib maintenance therapy improved PFS compared with placebo for NSCLC patients with both mutant *EGFR* tumors and wild-type *EGFR* tumors. However, although statistically significant, the degree of benefit was less in the *EGFR* wild-type subgroup. In this situation, it is useful to weigh the added benefit from erlotinib against its cost and toxicity profile, the availability of other treatments, and issues related to routine testing for *EGFR* mutation.

An additional consideration in using archived specimens from clinical trials is that the clinical utility of the biomarker is being explicitly examined only for the subset of patients in the RCT who have specimens available. For example, the biomarker may only be evaluable on larger tumors in a retrospective analysis where the specimens are being used for multiple research purposes. How these patients may be different from the patients for whom eventual clinical use of the biomarker is being contemplated should be considered when evaluating the clinical utility of the biomarker.

To address the treatment question, it is best to estimate the hazard ratio with its confidence interval for each of the biomarker subgroups. This information, along with presentation of the survival curves for each biomarker subgroup and the practical considerations mentioned above, can help decide the recommended treatment for each subgroup if the confidence intervals for the hazard ratios are sufficiently narrow. If the recommended treatment is the same for both biomarker subgroups (eg, treat with experimental therapy), then the biomarker is not clinically useful, regardless of its interaction with the treatment. For example, one would not

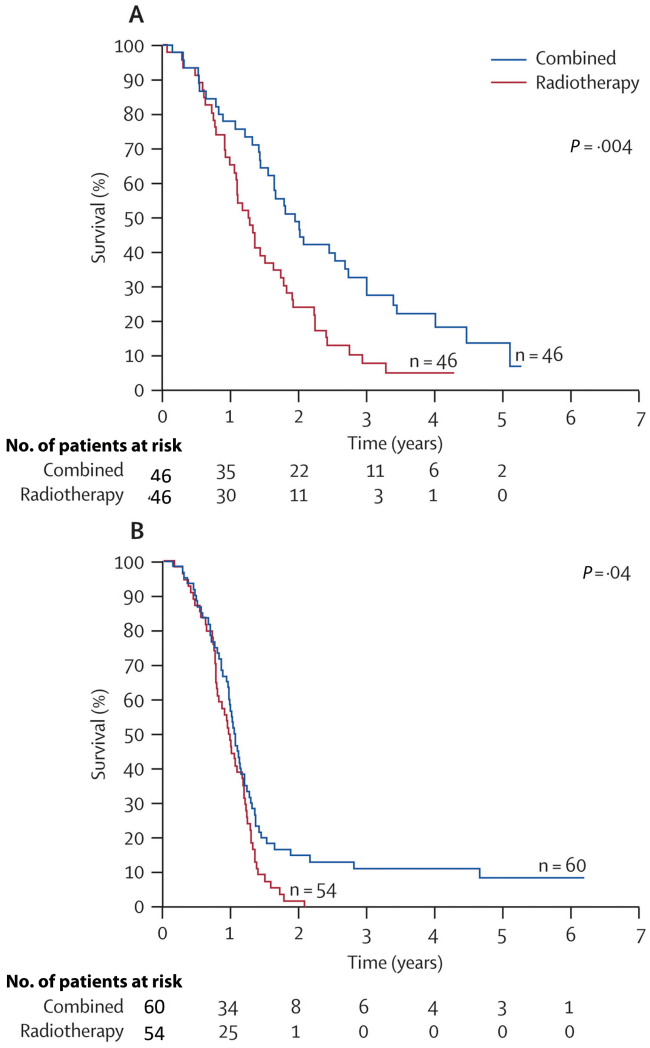


Figure 3. Combined therapy (temozolomide + radiotherapy) vs radiotherapy for glioblastoma patients with methylated *MGMT* (**A**) or unmethylated *MGMT* (**B**). The numbers of patients at risk are below the graphs. *P* values are based on two-sided log-rank tests [reprinted from Figure 4 of Stupp et al. (22). Reprinted with permission. Copyright 2009 Elsevier].

want to withhold pazopanib to renal carcinoma patients with low IL-6 values, based on Figure 2, A and B, and one might argue that temozolomide should be used for glioblastoma patients regardless of *MGMT* methylation status based on long-term survivors, as seen in Figure 3B (23,24). Whether erlotinib maintenance therapy for NSCLC is appropriate regardless of *EGFR* mutation status (Figure 2, C and D) is also a difficult issue (25).

When the confidence intervals for the hazard ratios are wide for one or both of the biomarker subgroups, then it may be impossible to make an informed treatment recommendation for each of the subgroups. In this case, it will be impossible to say whether the biomarker has any clinical utility. One solution to this problem is to obtain more information—for example, by pooling data from related trials as described previously or increasing follow-up time to observe more events and increase precision of the estimated treatment effects. Moreover, additional biological understanding of the biomarker and mechanism of action of the targeted agent may be obtained through further preclinical studies that increase confidence in the strength of the biomarker's predictive effect.

Discussion

A key part of designing a study to evaluate a putative predictive biomarker using specimens from an RCT is to specify what one will potentially be able to conclude from the study with sample size planned for the new trial or available from the previously conducted trial. In particular, although one may be able to conclude that the experimental treatment works much better than the standard treatment in the biomarker-positive subgroup (if it truly does), one may not be able to conclude whether the treatments are equally efficacious in the biomarker-negative subgroup (even if they truly are). This limitation is due partially to the inherently larger sample size required for testing an interaction compared with testing a treatment effect but is particularly important for analyses based on previously conducted trials where one does not have control of the sample size and because of the retrospective nature of the study specimens may not be available or be successfully assayed from all patients. In some circumstances, it may be reasonable to pool data from multiple trials to overcome the limited sample size in individual studies. Successful pooling relies on the clinical settings to be similar across the trials and requires that the biomarkers be assayed on all specimens using a common assay or assays that are sufficiently comparable. In addition, when evaluating the clinical utility of a biomarker, one needs to consider the possibility that biomarker assays may improve over time, so a treatment effect seen in the biomarker-negative subgroup may be partially due to some biomarker-positive patients being incorrectly classified as biomarker-negative; this treatment effect may no longer be present if a more accurate assay is used in future clinical practice. An example of a biomarker assay that has evolved over time is the HER2 assay in breast cancer. The immunohistochemical assay for HER2 that was used to screen for entry into the pivotal clinical trials of trastuzumab in metastatic breast cancer was replaced with a US Food and Drug Administration–approved companion diagnostic immunohistochemical assay at the time of approval of trastuzumab. Subsequently other immunohistochemical assays and fluorescence in situ hybridization tests for HER2 were widely adopted in clinical practice. Discordance rates of 20% or more were observed between different testing methods in clinical trial and community settings, prompting calls for HER2 testing standards (26).

In this commentary, we have focused on studies designed to demonstrate therapeutic superiority. In some settings, establishing noninferiority of a new therapy compared with the standard of care may justify change in clinical practice (eg, when the new therapy is less toxic and/or less costly than the standard one). However, because demonstration of noninferiority typically requires a larger sample size than that of superiority, the retrospective designs described here are unlikely to provide the precision needed to establish noninferiority unless they are based on previously conducted RCTs designed to demonstrate noninferiority. For example, consider a retrospective biomarker study based on data from a randomized trial designed to demonstrate superiority of a new therapy over the standard treatment. If the biomarker study shows that survival with the new treatment is statistically significantly better than with the standard treatment for the biomarker-positive subgroup, while the survival curves for the new and standard treatment arms are virtually identical to each other in the biomarker-negative subgroup, one may be tempted to conclude that the new therapy is at

least as good as the standard treatment across the entire population. However, because of the typically limited sample sizes in subgroup analyses performed retrospectively, estimates of treatment effects restricted to subgroups may lack precision. This limited precision may preclude one from ruling out a nontrivial survival detriment due to the new treatment in the biomarker-negative subgroup.

Routine collection and banking of specimens from clinical trials provides a rich resource for conducting retrospective analyses of promising biomarkers. However, specimen collection is expensive and time consuming, and these precious resources, once depleted, are usually nonrenewable. Similar to futility monitoring for a treatment effect in a prospective randomized clinical trial, one could perform biomarker assays in stages with the possibility to stop performing biomarker assays if it becomes clear early that the hypothesized biomarker effect will not be confirmed (27). In addition, patient specimens must be appropriately collected and stored and should be prioritized for well-designed studies that are most likely to delineate the clinical utility of promising biomarkers. For additional design considerations relevant to establishing clinical utility of biomarker assays or tests, readers are referred elsewhere (28).

References

1. Henry N, Hayes D. Uses and abuses of tumor markers in the diagnosis, monitoring, and treatment of primary and metastatic breast cancer. *Oncologist*. 2006;11(6):541–552.
2. Freidlin B, McShane L, Korn E. Randomized clinical trials with biomarkers: design issues. *J Natl Cancer Inst*. 2010;102(3):152–160.
3. US Food and Drug Administration (USFDA). *Guidance on Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products*. 2012. US FDA: Rockville, MD.
4. Freidlin B, Sun Z, Gray R, Korn EL. Phase III clinical trials that integrate treatment and biomarker evaluation. *J Clin Oncol*. 2013;31(25):3158–3161.
5. Taube S, Clark G, Dancey J, McShane L, Sigman C, Gutman S. A perspective on challenges and issues in biomarker development and drug and biomarker codevelopment. *J Natl Cancer Inst*. 2009;101(21):1453–1463.
6. Simon R, Paik S, Hayes D. Use of archived specimens in evaluation of prognostic and predictive biomarkers. *J Natl Cancer Inst*. 2009;101(21):1446–1452.
7. Jiang W, Freidlin B, Simon R. Biomarker-adaptive threshold design: a procedure for evaluating treatment with possible biomarker-defined subset effect. *J Natl Cancer Inst*. 2007;99(13):1036–1043.
8. Taylor JMG, Ankerst DP, Andridge RR. Validation of biomarker-based risk prediction models. *Clin Cancer Res*. 2011;14(19):5977–5983.
9. Moore H, Kelly A, Jewell S, et al. Biospecimen reporting for improved study quality (BRISQ). *J Proteome Res*. 2011;10(8):3429–3438.
10. Peto R. Statistical aspects of cancer trials. In: *Treatment of Cancer*. Halnan K (ed.), 1982:867–871. London: Chapman and Hall.
11. Mok T, Wu Y-L, Thongprasert S, et al. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *New Engl J Med*. 2009;361(10):947–957.
12. Pirker R, Pereira J, Szczesna A, et al. Cetuximab plus chemotherapy in patients with advanced non-small-cell lung cancer (flex): an open-label randomised phase III trial. *Lancet*. 2009;373(9674):1525–1531.
13. Pirker R, Pereira J, von Pawel J, et al. *EGFR* expression as a predictor of survival for first-line chemotherapy plus cetuximab in patients with advanced non-small-cell lung cancer: Analysis of data from the phase 3 flex study. *Lancet Oncol*. 2012;13(1):33–42.
14. Tran H, Liu Y, Zurita A, et al. Prognostic or predictive plasma cytokines and angiogenic factors for patients treated with pazopanib for metastatic renal-cell cancer: a retrospective analysis of phase 2 and phase 3 trials. *Lancet Oncol*. 2012;13(8):827–837.
15. Brugger W, Triller N, Blasinska-Morawiec M, et al. Prospective molecular marker analyses of *EGFR* and *KRAS* from a randomized, placebo-controlled study of erlotinib maintenance therapy in advanced non–small-cell lung cancer. *J Clin Oncol*. 2011;29(31):4113–4120.
16. Peterson B, George S. Sample size requirements and length of study for testing interaction in a 1 x k factorial design when time-to-failure is the outcome. *Control Clin Trials*. 1993;14(6):511–522.
17. Vale C, Tierney J, Fisher D, et al. Does anti-*EGFR* therapy improve outcome in advanced colorectal cancer? A systematic review and meta-analysis. *Cancer Treat Rev*. 2012;38(6):618–625.
18. Khambata-Ford S, Harbison C, Hart L, et al. Analysis of potential predictive markers of cetuximab benefit in BMS099, a phase III study of cetuximab and first-line taxane/carboplatin in advanced non–small-cell lung cancer. *J Clin Oncol*. 2010;28(6):918–927.
19. O'Byrne K, Bondarenko I, Barrios C, et al. Molecular and clinical predictors of outcome for cetuximab in non-small cell lung cancer (NSCLC): data from the flex study. *J Clin Oncol*. 2009;27:408(Suppl 415):abstract 8007.
20. Stupp R, Mason W, van den Bent M, et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *New Engl J Med*. 2005;352(10):987–996.
21. Hegi M, Diserens A-C, Gorlia T, et al. *MGMT* gene silencing and benefit from temozolomide in glioblastoma. *New Engl J Med*. 2005;352(10):997–1003.
22. Stupp R, Hegi M, Mason W, et al. Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. *Lancet Oncol*. 2009;10(5):459–466.
23. Weller M, Wick W, Hegi M, Stupp R, Tabatabai G. Should biomarkers be used to design personalized medicine for the treatment of glioblastoma? *Future Oncol*. 2010;6(9):1407–1414.
24. McDonald K, Aw G, Kleihues P. Role of biomarkers in the clinical management of glioblastomas: what are the barriers and how can we overcome them? *Front Neurol*. 2012;3(188):1–8.
25. Shepherd F. Maintenance therapy comes of age for non-small-cell lung cancer, but at what cost? *J Clin Oncol*. 2011;29(31):4068–4070.
26. Wolff AC, Hammond ME, Schwartz JN, et al. American Society of Clinical Oncology/College of American Pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer. *J Clin Oncol*. 2007;25(1):118–145.
27. Koopmeiners J, Vogel R. Early termination of a two-stage study to develop and validate a panel of biomarkers. *Stat Med*. 2012;32(6):1027–1037.
28. The Center for Medical Technology Policy. *Evaluation of Clinical Validity and Clinical Utility of Actionable Molecular Diagnostic Tests in Adult Oncology*. Baltimore, MD: CMTP; 2013. http://www.cmtpnet.org/wp-content/uploads/downloads/2013/05/CMTP_MDx_EGD05-01-2013.pdf. Accessed September 18, 2013.

Notes

The views expressed in this article are the personal opinions of the authors and do not necessarily reflect policy of the US National Cancer Institute.

Affiliations of authors: Biometric Research Branch (M-YCP, BF, ELK, LMS), Cancer Diagnosis Program (BAC), and Cancer Treatment and Evaluation Program (JSA), Division of Cancer Treatment and Diagnosis, National Cancer Institute, National Institutes of Health, Bethesda, MD.