Biomarkers and surrogate end points —the challenge of statistical validation

Marc Buyse, Daniel J. Sargent, Axel Grothey, Alastair Matheson and Aimery de Gramont

Abstract | Biomarkers and surrogate end points have great potential for use in clinical oncology, but their statistical validation presents major challenges, and few biomarkers have been robustly confirmed. Provisional supportive data for prognostic biomarkers, which predict the likely outcome independently of treatment, is possible through small retrospective studies, but it has proved more difficult to achieve robust multi-site validation. Predictive biomarkers, which predict the likely response of patients to specific treatments, require more extensive data for validation, specifically large randomized clinical trials and meta-analysis. Surrogate end points are even more challenging to validate, and require data demonstrating both that the surrogate is prognostic for the true end point independently of treatment, and that the effect of treatment on the surrogate reliably predicts its effect on the true end point. In this Review, we discuss the nature of prognostic and predictive biomarkers and surrogate end points, and examine the statistical techniques and designs required for their validation. In cases where the statistical requirements for validation cannot be rigorously achieved, the biological plausibility of an end point or surrogate might support its adoption. No consensus yet exists on processes or standards for pragmatic evaluation and adoption of biomarkers and surrogate end points in the absence of robust statistical validation.

Buyse, M. *et al. Nat. Rev. Clin. Oncol.* **7**, 309–317 (2010); published online 6 April 2010; doi:10.1038/nrclinonc.2010.43

Introduction

Biomarkers and surrogate end points have an increasingly important role in both cancer research and clinical practice. Biomarkers can be used to assess prognosis and to predict how individual patients will respond to specific treatments, whereas surrogate end points potentially enable the effectiveness of new interventions to be assessed more rapidly, and at times with greater accuracy, than classic end points (such as survival) in clinical trials. However, the challenges that must be overcome in the adoption of biomarkers and surrogate end points are numerous,[1] and range from discovery, verification and assay qualification through to statistical validation, successful use in clinical trials and, ultimately, routine use in the clinic. Here, we focus on one of the most demanding stages in this process, the challenge of statistical validation.

Definitions

The definitions of biomarkers and end points used in this Review are summarized in Table 1. According to the definition adopted by the Biomarkers Definitions Working Group,[2] a biomarker is defined as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention". Within this broad category, we focus our discussion on biomarkers that forecast future states, namely prognostic and predictive biomarkers, as opposed to pharmacokinetic or pharmacodynamic biomarkers that are used in early drug development. Prognostic biomarkers predict the likely course of disease, irrespective of treatment; for instance, lymph-node involvement predicts a poor outcome in the management of solid tumors, even though treatment can prolong survival of patients with or without evidence of nodal involvement. By contrast, predictive biomarkers forecast the likely response to treatment; for example, hormone receptor status predicts response to endocrine therapies in breast cancer. Furthermore, some biomarkers, such as hormone receptor status in breast cancer, are both prognostic and predictive.

Biomarkers can be contrasted with clinical end points, which capture information on how patients would feel, function or survive.[3] Surrogate end points, which might themselves be based upon a biomarker, aim at replacing a clinical end point with a faster and more-sensitive evaluation of the effect of experimental treatments. In this Review, we discuss the challenges of achieving statistical validation for prognostic and predictive biomarkers, and finally surrogate end points.

The term 'validation' can itself be confusing and has been subjected to different usages within the literature; there is a need for greater standardization in the nomenclature for all the stages of biomarker discovery and adoption.[4] Some authors and regulatory authorities have used

Competing interests
M. Buyse declares an association with the following institution: International Drug Development Institute. D. J. Sargent declares associations with the following companies: Almac, DiagnoCure, Exiqon, Genomic Health, Precision Therapeutics. See the article online for full details of the relationships. The other authors declare no competing interests.

International Drug Development Institute, 30 Avenue Provinciale, 1340 Louvain-la-Neuve, Belgium (**M. Buyse**). Cancer Center Statistics, Mayo Clinic Cancer Center (**D. J. Sargent**), and Department of Medical Oncology, Mayo Clinic College of Medicine (**A. Grothey**), 200 First Street SW, Rochester, MI 55905, USA. Fondation ARCAD, 22 Rue Malher, 75004 Paris, France (**A. Matheson**). Hôpital Saint-Antoine, Pavillon Moïana, 184 rue du Faubourg Saint-Antoine, 75012 Paris, France (**A. de Gramont**).

Correspondence to:
M. Buyse
marc.buyse@iddi.com

Key points

- Candidate prognostic biomarkers are relatively easy to identify, but multi-site validation has rarely been done
- Predictive biomarkers require extensive data for validation, based on large randomized clinical trials and meta-analyses
- Surrogate end points require data demonstrating both that the surrogate is prognostic of the true end point, and that the effect of treatment on the surrogate correlates with that of the true end point
- The biological plausibility of a biomarker or surrogate might support its adoption even in cases where full statistical validation is lacking
- No consensus exists on the best approach for pragmatic evaluation and adoption of biomarkers and surrogate end points when robust statistical validation is lacking

Table 1 | Definitions of biomarkers and surrogate end points

Term	Definition
Biomarker	A characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention[2]
Prognostic biomarker	Biomarker that forecasts the likely course of disease irrespective of treatment
Predictive biomarker	Biomarker that forecasts the likely response to a specific treatment
Clinical end point	Measurement providing information on how a patient feels, functions or survives[3]
Surrogate end point	Measurement providing early and accurate prediction of both a clinical end point, and the effects of treatment on this end point
Validation	Confirmation by robust statistical methods that a candidate prognostic biomarker, predictive biomarkers or surrogate end point fulfills a set of conditions that are necessary and sufficient for its use in the clinic

the term 'qualification' for the process of establishing the credibility of a biomarker or surrogate end point.[5–7] The terms 'validation' and 'qualification' have also both been applied to the process of confirming the effectiveness of a biomarker assay.[1,6] For the purposes of the present discussion, however, we will ignore these important aspects of assay validation, and refer to a validated biomarker as one that has been demonstrated by robust statistical methods to be associated with a given clinical end point (prognostic biomarkers), to predict the effect of a therapy on a clinical end point (predictive biomarkers), or to be able to replace a clinical end point to assess the effects of a therapy (surrogate end points).

Prognostic and predictive biomarkers

Biomarkers can be image-based or physiological indicators, but with the advent of the targeted-therapy era, cellular, molecular and genetic biomarkers have become increasingly important. In oncology, studies in breast cancer have pioneered the search for cellular, molecular and genetic biomarkers. In particular, the Early Breast Cancer Trialists' Collaborative Group (EBCTCG) has accumulated long-term data on the recurrence and mortality hazard rates over time for hormone receptor-positive tumors in comparison with receptor-negative tumors.[8] The analyses conducted by the EBCTCG have confirmed that hormone receptor status is a prognostic marker for outcome in breast cancer. They have also confirmed that endocrine therapies (such as tamoxifen and aromatase inhibitors) are only beneficial in tumors expressing hormonal receptors—in other words, hormone receptor status is also a predictive marker for therapeutic response.[8] Current research in breast cancer is exploring the molecular heterogeneity of the disease in greater detail,[9,10] and these studies might, in turn, lead to the discovery of further biomarkers and refine the use of those already identified. Other than in breast cancer, the number of validated prognostic and predictive biomarkers remains sparse, but the accumulating knowledge in this field suggests that advances in molecular oncology will ultimately revolutionize patient selection as well as cancer treatments.

Biological considerations have a key role in the initial identification of prognostic and predictive biomarkers, and remain important—alongside statistical analysis of clinical trials—during their evaluation and adoption into clinical practice. For instance, HER2/*neu* overexpression is integral to the biology of some forms of breast cancer, and accordingly HER2/*neu* status is expected to predict the clinical effectiveness of agents targeting the HER2 pathway, as has been confirmed in recent trials.[11–15] Overexpression of HER2/*neu* in breast cancer, mutations in *KIT* in gastrointestinal stromal tumors, and the presence of the fusion gene *BCR–ABL* (Philadelphia chromosome) in chronic myelogenous leukemia are all examples of both prognostic (linked to tumor biology) and predictive (linked to the treatment effect) biomarkers. In addition, all of these genes are involved in driving the aggressive phenotype of the malignancy, and as such, might provide targets for therapeutic interventions.[16,17] By contrast, putative biomarkers that are not integral to the disease process are less likely to have a prognostic impact, but in some cases they might predict a lack of benefit (rather than enhanced benefit). For instance, *KRAS* mutations in colorectal cancer are not disease-critical, but predict a lack of benefit of anti-EGFR monoclonal antibodies, and as such are highly useful as a predictive biomarker for patient selection.[18,19]

Importantly, while biological considerations can strengthen the case for the adoption of a biomarker, they must be interpreted with caution, and might in some cases prove misleading. For instance, at least one recent study suggests that HER2-directed therapies can be clinically beneficial even in the absence of HER2/*neu* overexpression,[20,21] a finding that would not be anticipated on the basis of HER2 biology as currently understood. If confirmed, this finding might reveal other mechanisms of action of HER2-directed therapies separate from direct interaction with their primary molecular target. Ultimately, biology cannot substitute for the validation of biomarkers through clinical trials and statistical analysis.

Validating prognostic biomarkers

The process of validating biomarkers and surrogates from a statistical standpoint typically begins with an initial demonstration that a correlation exists between the marker and the outcome of interest, followed by independent statistical validation of the relationship. For a biomarker to be validated as prognostic, an association must be demonstrated between the presence and

Table 2 | Examples of prognostic and predictive biomarkers and surrogate end points

Type of biomarker	Uses in management and clinical trials	Identification	Validation	Examples
Prognostic biomarker	Treatment choice, patient selection and stratification	Easy, but often flawed or biased	Frequent, but often inadequate because of regression to the mean or flaws in the initial identification study	Poor performance status, elevated hepatic enzymes, multi-site metastases in advanced colorectal cancer.
Predictive biomarker	Treatment choice, patient selection and stratification	Difficult, requires randomized trial	Uncommon, requires large randomized trial	*KRAS* mutation predictive of lack of activity of cetuximab and panitumumab in colon cancer.[18,19] Hormone receptor status predictive of effect of tamoxifen and aromatase inhibitors in breast cancer.[79] HER2/*neu* amplification predictive of effect of trastuzumab and lapatinib in breast cancer.[11–15] *EGFR* mutations predictive of effect of erlotinib and gefitinib in non-small-cell lung cancer.[80]
Surrogate end point	Treatment choice, treatment evaluation	Very difficult, requires meta-analysis or large randomized trial	Rare, requires meta-analysis or large randomized trial	Progression-free survival as a surrogate for overall survival for fluoropyrimidine-based regimens in treating colon cancer.[66] Hematologic complete remission for time to disease progression in patients with leukemias.[46,47]

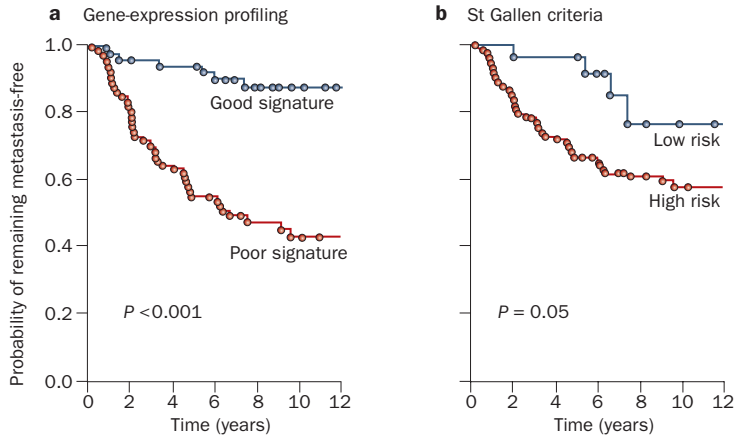
absence of the marker at baseline, or changes in the biomarker over time, and a treatment-independent clinical end point (Table 2).[22] This is a relatively straightforward requirement from a statistical standpoint, and does not require any specific study design; indeed, small retrospective studies can often be a sufficient source of data. The number of events required to identify a genuine prognostic marker decreases as the hazard ratio for the outcome becomes more extreme (differs more from unity), and as the percentage of patients who have the marker approaches 50%.

The challenges of moving from the initial establishment of a correlation to robust validation are considerable, as illustrated by the MammaPrint® (Netherlands Cancer Institute and Agendia BV, Amsterdam, The Netherlands) gene signature.[23–25] This signature is a microarray-based assay of 70 genes developed with the goal of predicting outcome in breast cancer (Figure 1).[23] In a retrospective analysis involving a relatively small sample of 78 patients, this signature was identified as a strong prognostic marker for the occurrence of metastases within 5 years of resection. Patients with a poor prognosis based on their MammaPrint® signature were found to have an odds ratio of 15.0 (95% CI 3.3–56, $P<0.001$) for distant metastases within 5 years, when compared with patients who had a good prognosis.[24]

It is worth noting that while the sensitivity of prediction for an unfavorable outcome using the 70-gene signature was high (91% of patients who developed metastatic disease had the poor-prognosis signature), the specificity was modest (only 59% of patients who did not develop metastatic disease had the good-prognosis signature).[24] Looking at the predictive accuracy of the signature, the positive predictive value of the signature was 0.63 (about two thirds of the patients with a poor-prognosis signature would be expected to develop metastases within 5 years) and the negative predictive value was 0.9 (only one patient in 10 with a good-prognosis signature would be expected to develop metastases within 5 years). These findings indicate that while the MammaPrint® signature might potentially be useful to help avoid aggressive chemotherapy in patients with a good prognosis, it is not a sufficiently accurate predictor of which patients will, or will not, develop metastases to provide the sole basis for a treatment decision.

The results from the MammaPrint® study are instructive, because they demonstrate that a highly significant *P*-value and odds ratio are not necessarily adequate conditions to identify a prognostic marker with wide clinical utility.[26,27] Indeed, a clear visual separation between the marker-positive and marker-negative Kaplan–Meier curves and a highly significant log-rank test might prove to be a misleading assessment of the worth of a prognostic marker. Even if a putative prognostic marker has a highly significant impact on a particular clinical outcome of interest after adjustment for classic clinical factors (either through stratification or in a multivariate model), it does not imply that the predictive accuracy of the marker is sufficient to justify its use in clinical practice. Indeed, measures of predictive accuracy and of explained variation are generally required, but seldom reported.[27,28]

Initial identification of a prognostic biomarker should be followed up by multicenter validation or by cross-validation using re-sampling techniques if only one dataset is available. In the case of the MammaPrint® gene signature, a large validation study was conducted involving independent samples contributed by centers in Villejuif, Paris, Oxford, London and Stockholm, with central pathological review performed in Milan, statistical analysis in Brussels and microarray analysis in Lausanne.[29,30] In this analysis, no significant variation was found in the prognostic utility of the signature between centers, but hazard ratios were found to be less impressive than in the original Amsterdam studies. For instance, the hazard ratio for time to distant metastases was 2.13 in the validation series compared with 6.07 in the original series (the survival hazard ratios were 2.63 and 17.46, respectively).[28] Of the various possible reasons for this disparity, one of



Number at risk							
Good signature	60	57	54	45	31	22	12
Poor signature	91	72	55	41	26	17	9

Number at risk							
Low risk	22	22	21	17	9	5	2
High risk	129	107	88	69	48	34	19

Figure 1 | Prognostic markers: initial findings with the Amsterdam 70-gene signature in breast cancer. Copyright © 2002 Massachusetts Medical Society. All rights reserved.

the most interesting was the duration of follow-up, which was twice as long (13.6 years) in the validation series than in the original series (6.7 years). This finding suggests that the prognostic impact of the gene signature is highly time-dependent, that is, the signature is very good at identifying patients at high risk of early disease progression, as opposed to those at risk of later disease progression.[29,30] The validation study concluded that the MammaPrint® gene signature does provide some additional prognostic information to that derived from known clinical and pathological factors (including age, tumor size and grade, estrogen receptor status and nodal status), but its overall clinical utility remains to be confirmed.

Even when multicenter statistical confirmation for a biomarker has been achieved, the ultimate proof of its usefulness in the clinic still requires randomized, prospective evidence in clinical trials. In particular, prospective studies are required to clarify the utility of the biomarker in patients for whom the optimal course of treatment is not apparent from classic parameters. For instance, patients with a low risk of disease recurrence might require only standard therapy, whereas individuals at high risk of disease recurrence require experimental therapy, but for patients with intermediate risk (based on the biomarker and/or clinico-pathological factors) there might be uncertainty regarding the treatment decision. Such patients could be randomized to either standard or experimental treatment in prospective studies to clarify the role of the biomarker in determining treatment. Examples of such biomarker-based treatment trials in early breast cancer include the ongoing MINDACT[31] and TAILORx[32] trials.

Validating predictive biomarkers

The case of the MammaPrint® signature illustrates the difficulties of statistically validating a promising prognostic biomarker. Predictive markers, however, present even greater challenges, both with respect to initial demonstration of a correlation of the outcome with the marker measured and subsequent robust validation (Table 2).

A biomarker can be considered to be predictive when the baseline value, or changes in the value of the biomarker over time, forecasts the efficacy or toxicity of a treatment, as assessed by a defined clinical end point.[19] Statistical identification of predictive biomarkers requires data from randomized trials that include patients with both high and low levels of the biomarker. The highest level of evidence derives from trials with an 'interaction' design, in which all patients are stratified by biomarker level and then randomized to one of two treatments; this approach to validation is currently being used in the ongoing Marker Validation of Erlotinib in Lung Cancer (MARVEL) trial, in which patients are tested for EGFR-status and then randomized between erlotinib (Tarceva®, F. Hoffmann-La Roche Ltd, Basel, Switzerland) or pemetrexed (Alimta®, Eli Lilly and Company, Indianapolis, IN) as second-line treatment of non-small-cell lung cancer.[33] In this trial, the analysis is being conducted separately in marker-positive and marker-negative patients, with the use of an interaction test aimed at showing that the treatment effects differ in these two groups. Large numbers of events, and hence patient populations, are generally required for the reliable detection of interactions.[34] Realistically, therefore, 'interaction' trials capable of validating predictive markers are likely to be few in number.

Therefore, the validation of most emerging biomarkers intended to inform a binary treatment decision employ alternative approaches to 'interaction' trials. Frequently, a 'selection' design has been adopted, in which only marker-positive patients enter the validation trial.[35,36] Such trials have the capacity to confirm the usefulness of the marker in indentifying a population in which there is a treatment benefit, but they do not imply that the marker is truly predictive, since they provide no information with respect to the lack of benefit among marker-negative patients. A key example of such a situation is the effect of trastuzumab (Herceptin®, Genentech, San Francisco, CA) in delaying or preventing recurrence of breast cancer. In patients with HER2/*neu* amplified tumors, the benefit of trastuzumab treatment has been established by several large randomized trials.[11–15] However, it has been suggested that treatment might have similar effects in patients with HER2/*neu* non-amplified tumors.[20,21] Specifically, for about 10% of patients entered in the NSABP B-31[20] and NCCTG 9831[21] trials, tumors had been assessed as HER2/*neu* amplified in local laboratories, but not when retested in central laboratories. An analysis of this subset of patients suggested that they enjoyed the same benefit from trastuzumab as patients confirmed to be HER2/*neu*-amplified by central laboratories. Several explanations were offered to explain this finding: HER2 might be overexpressed without gene amplification, trastuzumab might exert beneficial effects not mediated by the currently known HER2 alterations, there might be laboratory artifacts in assessing HER2/*neu* amplification, or the results (based on relatively small subsets) might simply be due to chance. Whatever the explanation turns out to be in this particular case, it

suggests that interaction trials (with a suitably small marker-negative group) may be indicated even when the prior assumption is that treatment will work only in marker-positive patients.

The 'selection' approach is being used in a number of current trials, notably in studies that were initiated to test EGFR inhibitors in patients with *KRAS* wild-type colorectal cancer as a treatment of advanced disease and as an adjuvant treatment, including the ongoing trials N0147[37], PETACC-8[38], and C80405[39]. A further example is the ECOG E5202[40] trial in patients with stage II colon cancer. In this trial, patients whose tumors have microsatellite instability (a putative predictive biomarker of resistance to fluoropyrimidines), and a normal *18q* chromosome (a prognostic biomarker) will not receive adjuvant therapy, whereas patients whose tumors have microsatellite stability and *18q* chromosomal abnormality will be randomized to receive a standard regimen of adjuvant therapy with 5-fluorouracil, leucovorin and oxaliplatin with or without bevacizumab. Sargent and colleagues[35,36,41] have offered a detailed discussion of the limitations of these trial designs to prospectively validate predictive biomarkers; in short, the selection design cannot confirm predictive effects while the interaction design may lack statistical power to do so.

Given the challenges of performing randomized studies to validate predictive biomarkers, retrospective analyses of completed randomized trials might yet prove to be the most important source of evidence. However, to yield convincing evidence, such retrospective analyses need to be planned in a prospective protocol that provides analytical details (cut-points, statistical methods, etc.) as well as interpretational guidelines (level of statistical significance, magnitude of effect, etc.). If several series are available, the results should be concordant across them. This 'retrospective-prospective' approach was used recently in advanced colorectal cancer, where *KRAS* mutation status was consistently shown in multiple trials to predict for a complete lack of effect of two EGFR-directed monoclonal antibodies, cetuximab (Erbitux®, Merck KgaA, Darmstadt, Germany) and panitumumab (Vectibix®, Amgen, Thousand Oaks, CA).[18,19,42–44]

Validating surrogate end points

At present, there are few accepted surrogate end points in clinical oncology, and none based on tumor response, molecular or genetic markers in solid tumors.[45] By contrast, in non-solid tumors, hematological complete remission has long been considered a surrogate for time-to-disease progression and overall survival.[46,47] Statistical validation of a surrogate end point is a strenuous process, with respect to both initial demonstration of an apparent relationship between a surrogate and the clinical end point, and subsequent robust confirmation (Table 2). General literature reviews on surrogate end points are available from both Weir and Walley[48] and Lassere.[49]

Prentice initially proposed that for a surrogacy relationship to be established the surrogate should predict the clinical end point, treatment should have a significant effect on both the candidate surrogate and the clinical end point, and the treatment effect on the surrogate should capture the full effect of treatment on the clinical end point.[50] This latter requirement has proved unrealistic and, consequently, alternative paradigms have been developed. While validation criteria are still an area of intense statistical research, the current consensus is that validation can be based on a 'correlation approach'. This involves demonstration in either randomized trials or meta-analysis that, on the one hand, the surrogate is prognostic for disease outcome and, on the other hand, that the effect of intervention on the surrogate is sufficiently correlated with the effect on the true end point.[48,49,51–57]

Thus, to achieve validation, the candidate surrogate marker must first be shown to forecast outcome in the same fashion as a prognostic marker, without reference to specific interventions. This aspect of surrogacy is generally referred to as 'individual-level' surrogacy (an alternative term might be 'outcome' surrogacy), which means that for individual patients, the marker or surrogate outcome must correlate well with the final outcome of interest, such as survival. Secondly, and more critically, the effect of treatment on the candidate surrogate marker must be closely correlated with the effect of treatment on the true clinical end point. This aspect of surrogacy is generally referred to as 'trial-level' surrogacy (an alternative term might be 'effect' surrogacy) since it must be demonstrated for a group of patients in a clinical trial.

Statistically, individual-level surrogacy and trial-level surrogacy are independent of one another for normally distributed end points and, as such, they require separate validation.[58] Standard correlation coefficients are the most commonly used approach to quantify statistical associations, but alternative measures derived from information theory are also available and hold promise as a unifying paradigm.[59] The individual-level surrogacy can be demonstrated in a single trial, but a shortcoming of trial-level surrogacy is that it must typically be based on a meta-analysis of several randomized trials.[60–62] However, when an insufficient number of trials are available to conduct a meta-analysis, it might be possible to break the results of large trials down into smaller units of analysis, such as individual countries or study centers. This approach has been used to show that prostate-specific antigen should not be considered a surrogate for survival in patients with advanced prostatic cancer.[63,64]

With respect to trial-level surrogacy, the concept of a 'surrogate threshold effect' (STE) was recently introduced. The STE is defined as the minimum treatment effect on the surrogate necessary to predict an effect on the true end point.[65] This approach affords a natural interpretation of surrogacy from a clinical point of view, since treatments that are able to induce effects larger than the STE on the surrogate would be expected to also induce a proportionally greater effect on the clinical outcome. In advanced colorectal cancer, progression-free survival (PFS) has been shown to be an acceptable surrogate end point for overall survival with respect to fluoropyrimidine-based therapies, with PFS and overall survival being highly correlated ('individual-level' surrogacy) and effects of treatment on PFS and overall survival

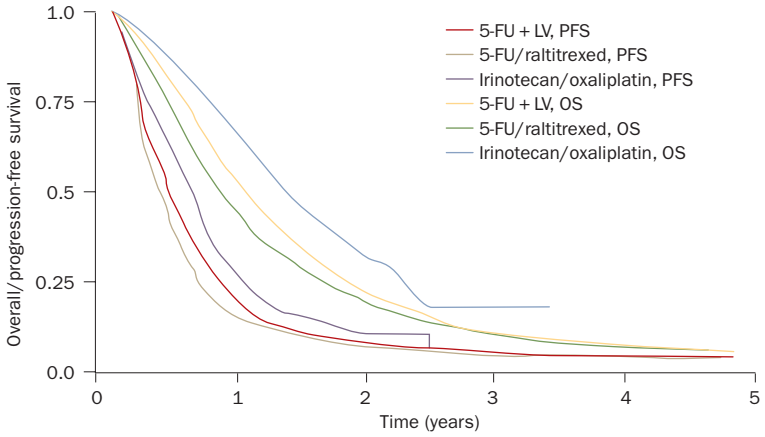


Figure 2 | Progression-free survival (PFS) and overall survival (OS) in advanced colorectal cancer.

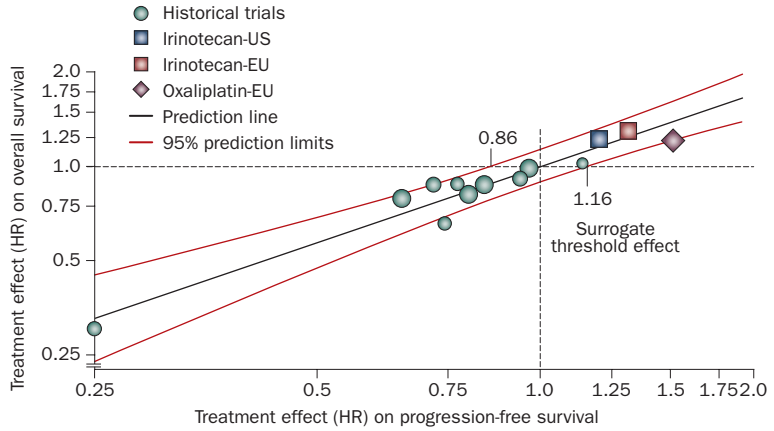


Figure 3 | Surrogate end point in advanced colorectal cancer: 'trial level' (effect) surrogacy and surrogate threshold effect.

being correlated ('trial-level' surrogacy) (Figures 2 and 3).[66] On the basis of a historical series of 10 randomized trials evaluating fluoropyrimidine-based treatment, the surrogate threshold effect was equal to 0.86, or 0.77 after elimination of a highly influential trial, indicating that if a new treatment reduced the hazard of tumor progression by at least 23%, it would be very likely to produce a benefit on survival (Figure 3).[65]

A major difficulty for the validation of surrogate end points, however, arises from the fact that they are validated with respect to a specific treatment or set of treatments. For a new treatment with a novel mechanism of action, it is uncertain if the same surrogacy relationship is applicable to that demonstrated for previous treatments. For instance, in first-line trials in advanced colorectal cancer, PFS has not yet been demonstrated to be a surrogate for overall survival with respect to novel targeted therapies such as bevacizumab (Avastin®, Genentech, San Francisco, CA), panitumumab and cetuximab. The question arises as to whether it is reasonable to assume that a surrogacy that was demonstrated for prior therapeutics can legitimately be treated as a surrogate in clinical assessment of every new treatment that emerges. A further difficulty arises from the fact that treatment options evolve with time. For instance, PFS was validated as a surrogate for classical 5-fluorouracil-based chemotherapy in colorectal cancer before the introduction of novel cytotoxics and targeted therapies, which now provide a greater range of salvage therapies. Had such therapies been available at an earlier stage in the evolution of colorectal cancer treatment, it is unlikely that the surrogacy of PFS for overall survival would have been demonstrated for 5-fluorouracil-based chemotherapy.[66]

Indeed, as standards of care in clinical oncology evolve, the difficulties of demonstrating surrogacy between proximal end points and overall survival will inevitably mount as the number of active treatment options increase and survival is extended. A recent study in the area of advanced breast cancer found that although tumor response, PFS, disease control and time-to-disease progression were all associated with overall survival at an individual-level, none had a sufficiently strong association at trial-level to qualify as a validated surrogate end point.[67] An example of this situation was demonstrated in a large trial of bevacizumab in advanced breast cancer whereby bevacizumab treatment was associated with a highly significant PFS benefit, but no overall survival benefit.[68]

These observations suggest that it might be difficult to formally establish PFS as a surrogate for overall survival in solid tumors for which several lines of treatment are currently available, but this does not imply that PFS does not have utility as an end point in its own right. Indeed, PFS might be the only sensitive (and realistic) end point for drug evaluation, given the availability of multiple active therapeutic lines (all of which have the potential to improve overall survival).[69]

Strict or pragmatic validation?

The current shortage of validated predictive and surrogate biomarkers in oncology reflects not only the statistical challenges discussed in this article, but difficulties at every stage of the discovery and evaluation process.[70] The US National Cancer Institute's Early Detection Research Network has proposed five distinct phases for the development of biomarkers for early cancer detection.[70,71] In Table 3, we adapt these five phases to the development of any biomarker, and outline the current status of MammaPrint® with respect to these phases as an example. One of the greatest challenges of validation is the lack of availability of both high-quality biological samples and standardized measures of response from all major trials, whether the trials are run by government-funded agencies or by industry. Regulatory authorities, such as the European Medicines Agency (EMEA) and the FDA should consider making stipulations to alleviate this problem to their industry and academic partners. For example, the generation of multi-trial tissue banks and databases would accelerate the search for biomarkers and provide a resource for retrospective analysis. The Foundation for the NIH Biomarkers Consortium in the USA represents a welcome but modest step in this direction.[72]

Despite the difficulties involved, the next few years are likely to see the accumulation of an increasing number of biomarker candidates with varying degrees of statistical

validation. Adherence to standards of reporting should improve the quality of validation studies.[73] It is also likely that in addition to individual biomarkers, groups of biomarkers will be developed and used collectively (for instance in multivariate analysis or other predictive modeling) to predict outcomes and responses despite the limited predictive power of individual candidates. Biological considerations will help increase the plausibility, and hence the practical acceptability, of predictive biomarkers despite incomplete validation. For example, HER2/*neu* status is currently accepted as a predictive biomarker for the efficacy of trastuzumab treatment, despite the absence of a formal test of interaction between the biomarker and the effects of this therapeutic agent.[74] This is mainly because the benefit of trastuzumab is large and well-established in patients with HER2/*neu*-positive tumors, while there is as yet no convincing evidence for a benefit in patients with HER2/*neu*-negative tumors. The combined use of information from several biomarkers is also likely to be of pragmatic utility in clinical decision-making, for instance by considering data for a biomarker that when present has a high sensitivity for an outcome mandating a particular therapy (most patients having the marker should be treated), alongside data for a biomarker that when absent mitigates with high specificity against treatment (most patients not having the marker should not be treated).

With regard to surrogate end points, the difficulties encountered in validating PFS or similar measures of disease control are likely to increase with the availability of more treatment options, even for treatments capable of major extensions in both disease control and overall survival. Paradoxically, while end points that can be observed early are required to increase the speed of clinical trials, they are least likely to achieve the status of surrogates for overall survival in situations where new treatments are most effective. This paradox arises because such new treatments will likely be given to most, if not all, patients once first-line trial treatments are completed, and this will make the benefits of the new treatment on overall survival hard to demonstrate statistically. Trials that compare strategies including several lines of treatment have been proposed for a better assessment of the true impact of a new agent, for instance to decide whether the agent should be given as part of first-line treatment, or later therapies.[75,76] However, the confounding effect of giving a very effective new agent after failure of a strategy remains an issue in the analysis of overall survival.

These difficulties call for a pragmatic approach, in which candidate surrogates are evaluated not only on the basis of statistical validity, but also, as suggested by the International Conference on Harmonization, with respect to their biological plausibility and usefulness demonstrated in clinical trials.[77] This raises the question of how surrogate end points can be evaluated systematically, robustly and to common standards within a broader pragmatic framework. Lassere has proposed a formal schema for numerically assessing the strength of the relationships between surrogate biomarkers and end points, based on a weighted evaluation of biological, epidemiological, statistical, clinical trial and risk-benefit evidence.[49] Lathia *et al.*[78] have also discussed the multi-faceted evaluation of candidate surrogates, and argued for a flexible approach to the adoption of surrogates. Further clarification and standardization might be required; in the mean time, individual-level and trial-level data supported by biological considerations should remain the cardinal criteria by which the validity of candidate surrogates is assessed.

Table 3 | The phases of biomarker development and validation[71]

Phase	Title	Purpose	Example of MammaPrint®
1	Preclinical exploratory studies	Identify promising biomarkers	Not applicable
2	Clinical assay development	Develop and validate the clinical assay used to measure the biomarker	Initial series of 78 patients[23]
3	Retrospective validation	Quantify the biomarker impact using available patient series, tumor banks and other stored material	Retrospective series of 234 patients from same institution[24] and 326 patients from five other institutions[29]
4	Prospective validation	Confirm the biomarker impact in prospective trials	Prospective ongoing trial[81]
5	Clinical utility	Show the clinical utility of the biomarker in prospective trials	Prospective ongoing trial[81]

Conclusions

Prognostic and predictive biomarkers as well as surrogate end points are all required in oncology, but few have been confirmed, and evidence of their effectiveness in trials and in the clinic remains limited. Even for prognostic biomarkers, which require comparatively modest retrospective data for initial identification, many candidates turn out to be flawed after independent validation studies. Predictive markers require large multicenter randomized trials for validation, while surrogate end points have proved the most challenging of all biomarkers to identify, and require meta-analyses of randomized trials for validation. Notwithstanding these challenges, many trials are in progress to identify and validate biomarkers and surrogate end points in oncology, and as the search for effective targeted therapies continues many further candidates are likely to emerge. Realistically, the adoption of biomarkers and surrogate end points cannot rely on exhaustive statistical validation in all circumstances, but instead should be based on evidence utilizing a combination of statistical, clinical and biological considerations. This in turn raises questions about how the broader process of biomarker and end point adoption should be structured and standardized.

Review criteria

The PubMed database was searched for articles published between the period of 1st January 2004 to 1st April 2009. The search terms used were "surrogate endpoint", "biomarker", "validation" and "qualification" within the article title or abstract. Results were screened for relevant articles and PubMed-designated related articles. The authors contributed further articles to the search results based on their personal knowledge and experience.

1. Rifai, N., Gillette, M. A. & Carr, S. A. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat. Biotechnol.* **24**, 971–983 (2006).
2. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin. Pharmacol. Ther.* **69**, 89–95 (2001).
3. Temple, R. J. A regulatory authority's opinion about surrogate endpoints. In *Clinical Measurement in Drug Evaluation* (Eds Nimmo, W. S. & Tucker, G. T.) 17 (Wiley, New York, 1995).
4. Ransohoff, D. F. Rules of evidence for cancer molecular-marker discovery and validation. *Nat. Rev. Cancer* **4**, 309–314 (2004).
5. Goodsaid, F. M., Frueh, F. W. & Mattes, W. Strategic paths for biomarker qualification. *Toxicology* **245**, 219–223 (2008).
6. Wagner, J. A., Williams, S. A. & Webster, C. J. Biomarkers and surrogate end points for fit-for-purpose development and regulatory evaluation of new drugs. *Clin. Pharmacol. Ther.* **81**, 104–107 (2007).
7. Goodsaid, F. & Frueh, F. Biomarker qualification pilot process at the US Food and Drug Administration. *AAPS J.* **9**, E105–E198 (2007).
8. Clarke, M. Meta-analyses of adjuvant therapies for women with early breast cancer: the Early Breast Cancer Trialists' Collaborative Group overview. *Ann. Oncol.* **17,** 59–62 (2006).
9. Sørlie, T. *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA* **98**, 10869–10874 (2001).
10. Sørlie, T. Molecular classification of breast tumors: toward improved diagnostics and treatments. *Methods Mol. Biol.* **360**, 91–114 (2007).
11. Slamon, D. J. *et al.* Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *N. Engl. J. Med.* **344**, 783–792 (2001).
12. Piccart-Gebhart, M. J. *et al.* Trastuzumab after adjuvant chemotherapy in HER-2 positive breast cancer. *N. Eng. J. Med.* **353**, 1659–1672 (2005).
13. Romond, E. H. *et al.* Trastuzumab plus adjuvant chemotherapy for operable HER-2 positive breast cancer. *N. Eng. J. Med.* **353**, 1673–1684 (2005).
14. Slamon, D. *et al.* Phase III randomized trial comparing doxorubicin and cyclophosphamide followed by docetaxel with doxorubicin and cyclophosphamide followed by docetaxel and trastuzumab with docetaxel, carboplatin and trastuzumab in HER2 positive early breast cancer patients: BCIRG 006 study. In *Proc. 28th Annual San Antonio Breast Cancer Symp. 1* (San Antonio, Texas, USA 2005).
15. Joensuu, H. *et al.* Adjuvant docetaxel or vinorelbine with or without trastuzumab for breast cancer. *N. Engl. J. Med.* **354**, 809–820 (2006).
16. Benjamin, R. S. *et al.* Gastrointestinal stromal tumors II: medical oncology and tumor response assessment. *Semin. Oncol.* **36**, 302–311 (2009).
17. Gora-Tybor, J. & Robak, T. Targeted drugs in chronic myeloid leukemia. *Curr. Med. Chem.* **15**, 3036–3051 (2008).
18. Amado, R. G. *et al.* Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. *J. Clin. Oncol.* **26**, 1626–1634 (2008).
19. Di Fiore, F. *et al.* Clinical relevance of KRAS mutation detection in metastatic colorectal cancer treated by Cetuximab plus chemotherapy. *Br. J. Cancer* **96**, 1166–1169 (2007).
20. Paik, S. *et al.* Benefit from adjuvant trastuzumab may not be confined to patients with IHC 3+ and/or FISH-positive tumors: Central testing results from NSABP B-31. *J. Clin. Oncol. (Meeting abstracts)* **25**, 511 (2007).
21. Perez, E. A. *et al.* Updated results of the combined analysis of NCCTG N9831 and NSABP B-31 adjuvant chemotherapy with/without trastuzumab in patients with HER2-positive breast cancer. *J. Clin. Oncol. (Meeting abstracts)* **25**, 512 (2007).
22. Buyse, M. Towards the validation of statistically reliable biomarkers. *Eur. J. Cancer* **41** (Suppl. 1) 89–95 (2007).
23. van 't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).
24. van de Vijver, M. J. *et al.* A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* **347**, 1999–2009 (2002).
25. Sotiriou, C. & Pusztai, L. Gene-expression signatures in breast cancer. *N. Engl. J. Med.* **360**, 790–800 (2009).
26. Hayes, D. F., Trock, B. & Harris, A. L. Assessing the clinical impact of prognostic factors: when is "statistically significant" clinically useful? *Breast Cancer Res. Treat.* **52**, 305–319 (1998).
27. Pepe, M. S., Janes, H., Longton, G., Leisenring, W. & Newcomb, P. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am. J. Epidemiol.* **159**, 882–890 (2004).
28. Royston, P., Parmar, M. K. & Altman, D. G. Visualizing length of survival in time-to-event studies: a complement to Kaplan–Meier plots. *J. Natl Cancer Inst.* **100**, 92–97 (2008).
29. Buyse, M. *et al.* Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J. Natl Cancer Inst.* **98**, 1183–1192 (2006).
30. Desmedt, C. *et al.* Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clin. Cancer Res.* **13**, 3207–3214 (2007).
31. US National Library of Medicine. *ClinicalTrials.gov* [online]. http://www.clinicaltrials.gov/ct2/show/NCT00433589?term=NCT00433589&rank=1 (2009).
32. US National Library of Medicine. *ClinicalTrials.gov* [online]. http://www.clinicaltrials.gov/ct2/show/NCT00310180?term=NCT00310180&rank=1 (2009).
33. US National Library of Medicine. *ClinicalTrials.gov* [online]. http://www.clinicaltrials.gov/ct2/show/NCT00738881?term=NCT00738881&rank=1 (2009).
34. Peterson, B. & George, S. L. Sample size requirements and length of study for testing interaction in a 2 × k factorial design when time-to-failure is the outcome. *Control. Clin. Trials* **14**, 511–522 (1993).
35. Mandrekar, S. J. & Sargent, D. J. Clinical trial designs for predictive biomarker validation: one size does not fit all. *J. Biopharm. Stat.* **19**, 530–542 (2009).
36. Mandrekar, S. J. & Sargent, D. J. Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. *J. Clin. Oncol.* **27**, 4027–4034 (2009).
37. US National Library of Medicine. *ClinicalTrials.gov* [online]. http://www.clinicaltrials.gov/ct2/show/NCT00079274?term=NCT00079274&rank=1 (2009).
38. US National Library of Medicine. *ClinicalTrials.gov* [online]. http://www.clinicaltrials.gov/ct2/show/NCT00265811?term=NCT00265811&rank=1 (2009).
39. US National Library of Medicine. *ClinicalTrials.gov* [online]. http://www.clinicaltrials.gov/ct2/show/NCT00265850?term=NCT00265850&rank=1 (2009).
40. US National Library of Medicine. *ClinicalTrials.gov* [online]. http://www.clinicaltrials.gov/ct2/show/NCT00217737?term=NCT00217737&rank=1 (2009).
41. Sargent, D. J., Conley, B. A., Allegra, C. & Collette, L. Clinical trial designs for predictive marker validation in cancer treatment trials. *J. Clin. Oncol.* **23**, 2020–2027 (2005).
42. Karapetis, C. S. *et al.* K-ras mutations and benefit from cetuximab in advanced colorectal cancer. *N. Engl. J. Med.* **359**, 1757–1765 (2008).
43. Bokemeyer, C. *et al.* Fluorouracil, leucovorin, and oxaliplatin with and without cetuximab in the first-line treatment of metastatic colorectal cancer. *J. Clin. Oncol.* **27**, 663–671 (2009).
44. Van Cutsem, E. *et al.* KRAS status and efficacy in the first-line treatment of patients with metastatic colorectal cancer (mCRC) treated with FOLFIRI with or without cetuximab: The CRYSTAL experience. *J. Clin. Oncol. (Meeting abstracts)* **26**, 2 (2008).
45. Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D. & Geys, H. The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* **1**, 49–67 (2000).
46. Estey, E. H., Shen, Y. & Thall, P. F. Effect of time to complete remission on subsequent survival and disease-free survival time in, AML, RAEB-t, and RAEB. *Blood* **95**, 72–77 (2000).
47. Kern, W. *et al.* Early blast clearance by remission induction therapy is a major independent prognostic factor for both achievement of complete remission and long-term outcome in acute myeloid leukemia: data from the German AML Cooperative Group (AMLCG) 1992 Trial. *Blood* **101**, 64–70 (2003).
48. Weir, C. J. & Walley, R. J. Statistical evaluation of biomarkers as surrogate endpoints: a literature review. *Stat. Med.* **25**, 183–203 (2006).
49. Lassere, M. N. The Biomarker-Surrogacy Evaluation Schema: a review of the biomarker-surrogate literature and a proposal for a criterion-based, quantitative, multidimensional hierarchical levels of evidence schema for evaluating the status of biomarkers as surrogate endpoints. *Stat. Methods Med. Res.* **17**, 303–340 (2008).
50. Prentice, R. L. Surrogate endpoints in clinical trials: definition and operational criteria. *Stat. Med.* **8**, 431–440 (1989).
51. Sargent, D. J. *et al.* Disease-free survival versus overall survival as a primary end point for adjuvant colon cancer studies: individual patient data from 20,898 patients on 18 randomized trials. *J. Clin. Oncol.* **23**, 8664–8670 (2005).
52. Burzykowski, T., Molenberghs, G. & Buyse, M. (Eds) *The Evaluation of Surrogate Endpoints* (Springer, New York, 2005).
53. Buyse, M. & Molenberghs, G. Criteria for the validation of surrogate endpoints in randomized experiments. *Biometrics* **54**, 1014–1029 (1998).
54. Buyse, M. *et al.* Relation between tumor response to first-line chemotherapy and survival in advanced colorectal cancer: a meta-analysis. *Lancet* **356**, 373–378 (2000).
55. Alonso, A., Molenberghs, G., Geys, H., Buyse, M. & Vangeneugden, T. A unifying approach for surrogate marker validation based on Prentice's criteria. *Stat. Med.* **25**, 205–221 (2006).
56. Buyse, M., Burzykowski, T., Michiels, S. & Carroll, K. Individual- and trial-level surrogacy in colorectal cancer. *Stat. Methods Med. Res.* **17**, 467–475 (2008).
57. Prentice, R. L. Surrogate and mediating endpoints: current status and future directions. *J. Natl Cancer Inst.* **101**, 216–217 (2009).

58. Molenberghs, G. *et al.* Statistical challenges in the evaluation of surrogate endpoints in randomized trials. *Control. Clin. Trials* **23**, 607–625 (2002).
59. Alonso, A. & Molenberghs, G. Surrogate marker evaluation from an information theory perspective. *Biometrics,* **63**, 180–186 (2007).
60. Buyse, M. Contributions of meta-analyses based on individual patient data to therapeutic progress in colorectal cancer. *Int. J. Clin. Oncol.* **14**, 95–101 (2009).
61. Shi. Q. & Sargent, D. J. Meta-analysis for the evaluation of surrogate endpoints in cancer clinical trials. *Int. J. Clin. Oncol.* **14**, 102–111 (2009).
62. Piedbois, P. & Buyse, M. Endpoints and surrogate endpoints in colorectal cancer: a review of recent developments. *Curr. Opin. Oncol.* **20**, 466–471 (2008).
63. Buyse, M. *et al.* Validation of biomarkers as surrogates for clinical endpoints. In *Biomarkers in Clinical Drug Development* (Eds Bloom, J. C. & Dean, R. A.) 149–168 (Marcel Dekker, New York, 2003).
64. Collette, L. *et al.* Is prostate-specific antigen a valid surrogate endpoint for survival in hormonally treated patients with metastatic prostate cancer? Joint research of the European Organization for Research and Treatment of Cancer, the Limburgs Universitair Centrum, and AstraZeneca Pharmaceuticals. *J. Clin. Oncol.* **23**, 6139–6148 (2005).
65. Burzykowski, T. & Buyse, M. Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. *Pharm. Stat.* **5**, 173–186 (2006).
66. Buyse, M. *et al.* Progression-free survival is a surrogate for survival in advanced colorectal cancer. *J. Clin. Oncol.* **25**, 5218–5224 (2007).
67. Burzykowski, T., Buyse, M., Sargent, D., Sakamoto, J. & Yothers, G. Exploring and validating surrogate endpoints in colorectal cancer. *Lifetime Data Anal.* **14**, 54–64 (2008).
68. Miller, K. *et al.* Paclitaxel plus bevacizumab versus paclitaxel alone for metastatic breast cancer. *N. Engl. J. Med.* **357**, 2666–2676 (2007).
69. Sargent, D. J. & Hayes, D. F. Assessing the measure of a new drug: is survival the only thing that matters? *J. Clin. Oncol.* **26**, 1922–1923 (2008).
70. Ransohoff, D. F. How to improve reliability and efficiency of research about molecular markers: roles of phases, guidelines, and study design. *J. Clin. Epidemiol.* **60**, 1205–1219 (2007).
71. Pepe, M. S. *et al.* Phases of biomarker development for early detection of cancer. *J. Natl Cancer Inst.* **93**, 1054–1061 (2001).
72. Altar, C. A. The Biomarkers Consortium: on the critical path of drug discovery. *Clin. Pharmacol. Ther.* **83**, 361–364 (2008).
73. McShane, L. M. *et al*. Reporting recommendations for tumor marker prognostic studies (REMARK). *Nat. Clin. Pract. Oncol.* **2**, 416–422 (2005).
74. Masood, S. & Bui, M. M. Prognostic and predictive value of HER2/neu oncogene in breast cancer. *Microsc. Res. Tech.* **59**, 102–108 (2002).
75. Tournigand, C. *et al.* FOLFIRI followed by FOLFOX6 or the reverse sequence in advanced colorectal cancer: a randomized GERCOR study. *J. Clin. Oncol.* **22**, 229–237 (2004).
76. Allegra, C. *et al.* End points in advanced colon cancer clinical trials: a review and proposal. *J. Clin. Oncol.* **25**, 3572–3575 (2007).
77. Green, E., Yothers, G. & Sargent, D. J. Surrogate endpoint validation: statistical elegance versus clinical relevance. *Stat. Methods Med. Res.* **17**, 477–486 (2008).
78. Lathia, C. D. *et al.* The value, qualification, and regulatory use of surrogate end points in drug development. *Clin. Pharmacol. Ther.* **86**, 32–43 (2009).
79. Rastelli, F. & Crispino, S. Factors predictive of response to hormone therapy in breast cancer. *Tumori* **94**, 370–383 (2008).
80. Jackman, D. M. *et al.* Impact of epidermal growth factor eceptor and KRAS mutations on clinical outcomes in previously untreated non-small cell lung cancer patients: results of an online tumor registry of clinical trials. *Clin. Cancer Res.* **15**, 5267–5273 (2009).
81. Bogaerts, J. *et al.* Gene signature evaluation as a prognostic tool: challenges in the design of the MINDACT trial. *Nat. Clin. Pract. Oncol.* **3**, 540–551 (2006).