Towards validation of statistically reliable biomarkers

Marc Buyse

IDDI, Louvain-la-Neuve, Belgium

Introduction

Biomarkers play an increasing role in the development of new cancer treatments. A biomarker can be formally 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' [1]. Biomarkers can include biochemical markers, cellular markers, cytokines, genetic markers, gene expression profiles, imaging markers, or physiological markers [2]. They can be measured once before a treatment is administered, or repeatedly before, during and after the treatment is administered, in which case interest focuses on changes in the biomarkers over time. In contrast to biomarkers, clinical endpoints directly measure 'how a patient feels, functions or survives' [3].

Uses of biomarkers

As outlined in a related chapter in this volume, biomarkers can prove extremely useful for patient management [4]. In terms of clinical development of new treatments, biomarkers have the potential of increasing the sensitivity (or statistical power) of clinical trials, and of accelerating the outcome of these trials if they are observed well before the clinical endpoints of interest. Specifically, biomarkers can be used to select patients eligible for clinical trials, to stratify patients at entry in clinical trials, to monitor patients and guide treatment decisions, or to substitute for a clinical endpoint in the evaluation of the effects of new treatments. These purposes are quite different from each other, and imply different conditions for the biomarkers. It is convenient to categorise biomarkers in three broad groups, depending on their intended use: (1) *prognostic* biomarkers, which predict the outcome of individual patients or groups of patients in terms of a clinical endpoint. (2) *predictive* biomarkers, which predict the effect of a specific treatment on a clinical endpoint for individual patients or groups of patients. (3) *surrogate* biomarkers, which replace a clinical endpoint in clinical trials carried out to evaluate the effect of a specific treatment on individual patients or groups of patients.

Prognostic and predictive biomarkers

Figure 1 schematically represents a simple situation that makes the distinction between prognostic and predictive biomarkers clear. For simplicity, we shall assume throughout that patients can be biomarker positive (blue line) or negative (red line), and can receive a standard treatment or an experimental treatment.

Figure 1 can be interpreted as follows: Panel A shows a biomarker that would be of no utility, since its status does not modify patient outcome, regardless of treatment.

Panel B shows a purely prognostic biomarker, for which the clinical outcome of the patients depend on their biomarker status. In this figure, it has been

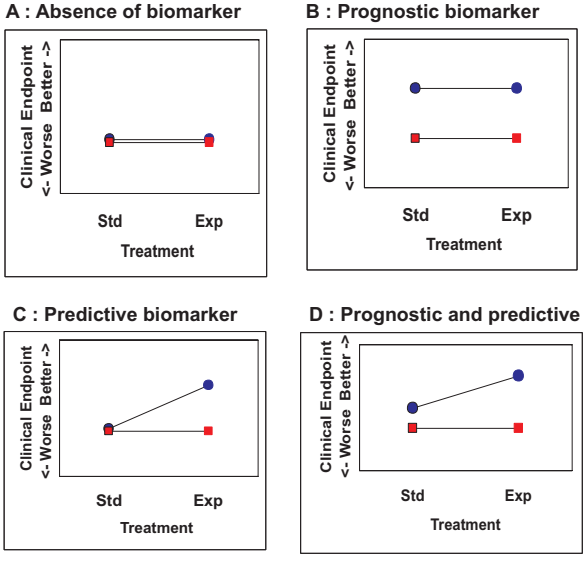

Fig. 1. Distinction between prognostic and predictive biomarkers. Patients can be biomarker positive (blue line) or negative (red line), and can receive a standard treatment (Std) or an experimental treatment (Exp).

Table 1
Main potential reasons for conflicting results of published biomarker studies.

Clinical	Different endpoints (response, time to recurrence, overall survival, etc.) Different patient populations (stage, treatments received, etc.)
Technical	Different assays or measurement techniques Different specimens (fresh-frozen, fixed tissue, serum, etc.)
Statistical	Studies under-powered (small sample sizes and few clinical events) Data over-analysed (multiple endpoints, cutpoint optimisation, model overfitting, subset analyses, etc.) Publication bias (positive studies published first)

assumed that the clinical outcome tends to be better when the biomarker is positive than when it is negative. Prognostic biomarkers are quite common: for example, in patients with early breast cancer, lymph node involvement is a bad prognostic factor regardless of treatment.

Panel C shows a purely predictive biomarker, for which the clinical outcome of the patients who receive a certain treatment depends on the biomarker status. In this figure, it has been assumed that the clinical endpoint tends to better when the biomarker is positive when patients receive the experimental treatment, but not when they receive the standard treatment. Predictive biomarkers have been known for a long time in patients with early breast cancer, for whom the hormonal receptor status predicts the effect of endocrine therapies such as tamoxifen or aromatase inhibitors. Note that a biomarker is predictive for a given treatment or class of treatments. The biomarker can be predictive of efficacy or of safety, with clinical relevance in both cases.

Panel D shows a biomarker that is both prognostic and predictive. In this figure, it has been assumed that the clinical endpoint tends to be better when the biomarker is positive than when it is negative, but the difference in clinical outcome is larger for patients receiving the experimental treatment than for those receiving the standard treatment. Overexpression of the HER2-neu gene in patients with early breast cancer provides an example of a biomarker that has both prognostic value (patients with HER2-neu overexpression having a worse prognosis) and predictive value for herceptin (patients with HER2-neu overexpression deriving a benefit from this treatment).

Although prognostic biomarkers can be useful, it is mostly predictive biomarkers that have the potential of changing oncology practice in the near future. Indeed, the presence of unknown predictive factors in a patient population can profoundly affect the statistical power of trials aimed at showing the benefit of new treatments for these patients [5]. Conversely, if a biomarker could reliably identify a subset of patients who derive the most benefit (or the least toxicity) from a new treatment, then clinical trials could be restricted to this subset, and the treatment of future patients could be targeted at this subpopulation only.

Surrogate biomarkers

The search for prognostic or predictive biomarkers has been intense over the last years, but the real long-term hope is to identify a surrogate biomarker, i.e. a biomarker capable, in and of itself, to show that a new treatment has the desired effect on the ultimate endpoint of interest. Currently, no such biomarkers have been identified in oncology. One of the most intensely studied cancer biomarkers is prostate-specific antigen (PSA), a glycoprotein found almost exclusively in normal and neoplastic prostate cells. Changes in PSA often antedate changes in bone scan, and they have long been used as an indicator of response in patients with androgen-independent prostate cancer [6,7]. Notwithstanding the usefulness of PSA for patient management, it was recently shown that PSA could not be considered a surrogate for long-term clinical endpoints [8,9]. Similarly, tumour shrinkage was shown not to be an acceptable surrogate endpoint for overall survival in advanced colorectal cancer [10]. In contrast, disease-free survival was shown to be an acceptable surrogate endpoint for overall survival in resectable colorectal cancer [11]. Admittedly, disease-free survival itself is a clinical endpoint and not a biomarker, but the statistical issues in validating a potential surrogate biomarker are exactly the same as in validating a surrogate endpoint [12].

Requirements for biomarkers

The clinical literature is replete with examples of biomarkers that have not been properly identified and validated, resulting in a large number of false claims for putative biomarkers that turn out to be of limited

Table 2
Different uses of biomarkers, and minimum requirements for their statistical validation (a biomarker can be the value of a characteristic at a given time or a change in this characteristic over time)

Use of biomarker	Definition	Minimum requirements for statistical validation	
		Study design	Sample size
Prognostic	Biomarker predicts clinical outcome	Case-control or cohort study	>100 patients
Predictive	Biomarker predicts treatment effect on clinical outcome	Large randomised trial	>500 patients
Surrogate	Treatment effect on biomarker predicts treatment effect on clinical outcome	Several randomised trials, or a large trial with several units of analysis (e.g. countries)	>10 units of analysis >1000 patients

usefulness either for patient management or to assess the efficacy and safety of new therapies [13]. Table 1 summarises some of the reasons why biomarker studies have tended to show divergent results.

The REMARK publication provides useful guidance for the conduct and reporting of biomarker studies [14].

Basic methodological requirements for the identification and use of biomarkers are that their measurements be accurate, standardised and reproducible. An essential additional requirement is that biomarkers be validated for their intended use. A related chapter in this volume discusses the clinical issues involved in the validation of biomarkers [15,16]. We expand here on the statistical aspects of the validation.

Table 2 shows the three categories of biomarkers discussed in the previous section, and the minimum requirements for the statistical validation of biomarkers in each category.

Prognostic biomarkers

For a prognostic biomarker, the baseline value of the biomarker, or changes in the biomarker over time, should be correlated with the clinical endpoint in untreated or in treated patients. This condition is straightforward to establish statistically, and does not require any particular study design: the prognostic impact of a biomarker can be investigated retrospectively in any series of patients providing a sufficient number of clinical events are available. In addition, a prognostic biomarker will be of clinical interest only if its impact on the clinical endpoint of interest is large enough, hence the number of patients required will often be smaller than what is required to establish or confirm a treatment benefit. The ideal setting to identify and validate a prognostic biomarker is a randomised therapeutic trial, in which all procedures to measure the biomarker are specified in the protocol (biological material collected, methods of preservation, assays and quantitation methods used, quality control procedures, reproducibility assessments, etc.) and in which all patients fulfil well-defined characteristics and are treated and followed according to a pre-determined schedule.

One issue that has not received sufficient attention is the choice of appropriate statistic(s) to quantify the impact of a prognostic biomarker on the clinical outcome of interest. The *P*-value of a test comparing the clinical outcome in two groups of patients, one with and another without the biomarker, provides insufficient evidence that the biomarker is of any clinical interest. Indeed, the *P*-value depends on the sample size so that in large series of patients, factors having a small, perhaps negligible impact on the patient prognosis, may well reach statistical significance. Measures that quantify the magnitude of effect of the biomarker on the clinical endpoint, such as the odds ratio (for tumour response and other dichotomous endpoints) or the hazard ratio (for survival time and other time-related endpoints), along with their confidence limits, are far more informative. However, these statistics are not sufficient to gauge the performance of the prognostic biomarker [17]. Measures of predictive accuracy are needed as well, and these are usually provided by the sensitivity and specificity of the biomarker, or equivalently by its positive and negative predictive value. If the biomarker is a continuous measurement, ROC (Receiver Operating Characteristics) curves can appropriately be used. Finally, a biomarker is of interest only if it provides *additional* prognostic value, over and above that of all easily measured clinical and pathological characteristics of the patients [18].

Predictive biomarkers

For a predictive biomarker, the baseline value of the biomarker, or changes in the biomarker over time,

should be correlated with the effect of treatment on the clinical endpoint of interest. This condition is far more difficult to establish than a mere prognostic impact of the biomarker on the endpoint. It is a common misconception that a biomarker that has prognostic value in a group of treated patients is predictive, *even if* the biomarker does not have prognostic value in another group of untreated patients. Indeed, the lack of impact in one group and the impact in another may be due to different selection criteria and other confounding factors that make such an indirect comparison untenable. The most reliable way to formally identify and validate a predictive factor is through a randomised trial, where patients are randomly allocated to treatment A or treatment B, as in Fig. 1 (typically to a standard treatment and an experimental treatment for which a predictive biomarker is thought to exist), since such a design makes treated and untreated patients comparable.

The statistical evidence required to establish that a biomarker is truly predictive is an open question. The most convincing situation would be one in which an interaction test between the effect of treatment and the biomarker status reaches statistical significance. The null hypothesis of interest in this test is that the effect of treatment is the same whether the biomarker is positive or negative. If the null hypothesis is rejected, one can safely conclude that the biomarker status modulates the effect of treatment – which is exactly equivalent to concluding that the biomarker is predictive. The major problem of interaction tests is that they lack power, so that a very large trial would be required for the test to reach significance. As a rule of thumb, the sample size required to show an interaction is at least four times larger than that required to detect main effects of the same magnitude [19]. Often biological considerations will alleviate the need for definite statistical evidence. For instance, it is acceptable to consider HER2-neu status as a predictive biomarker for herceptin, even though no study was carried out to formally test for an interaction between the biomarker and the targeted agent.

Surrogate biomarkers

The requirements for a biomarker to be considered a valid surrogate have been a theme of intense debate in the statistical literature over the last years. In early days of the search for surrogate endpoints, a common misconception was that an association between the surrogate endpoint and the clinical endpoint was sufficient to establish surrogacy. It was later demonstrated that correlation alone does not imply surrogacy. Formal validation criteria were then proposed [20], but they were too strict to be useful and were also criticised theoretically [21]. Currently, the preferred approach to validate a biomarker for use in drug intervention trials consists of establishing the following [22]: a strong association between values of the biomarker and the clinical endpoint: this is called the '*individual-leve"*association; a strong association between the effects of treatment on the biomarker and the clinical endpoint, as assessed in one or more randomised trial(s): this is called the '*trial-level'* association.

To be accepted, a surrogate that has biological plausibility would require to be validated at both the individual-level and the trial-level. Individual-level surrogacy deals with the ability of the potential surrogate to predict the clinical endpoint in an individual patient, while trial-level surrogacy deals with the ability of the treatment effect on the surrogate biomarker to predict the treatment effect on the clinical endpoint. The appropriate statistics to quantify the strength of the associations, both at the individual-level and at the trial-level, are a matter of debate. Standard correlation coefficients are currently most popular, but new measures derived from information theory have been recently proposed [23].

It is important to underline that the strength of the association between a biomarker and a clinical endpoint does not provide information about the relationship between the effects of treatment on the biomarker and the clinical endpoint, i.e. between treatment-induced changes in the biomarker and corresponding changes in the risk of the clinical endpoint. Mathematically, these two relationships are independent of each other, which may seem somewhat counterintuitive. However, to see the independence of the two levels of associations, one may think of two hypothetical situations: (1) one extreme situation in which a biomarker is perfectly correlated with a clinical endpoint, but the effect of treatment on the clinical endpoint is not entirely captured by the biomarker. In this situation, the trial-level correlation could be low if the effect of treatment on the clinical endpoint was mostly mediated by other intermediate endpoints independent of the biomarker. (2) the other extreme situation in which there is no individual-level correlation between the biomarker and the clinical endpoint, yet there is a perfect trial-level correlation because of a confounding factor leading to a spurious correlation between the treatment effects on the biomarker and the clinical endpoint.

In most situations of clinical interest, one would not expect to encounter situations as extreme as those just

described. The treatment would be expected primarily to 'shift' patients from a high risk group to a low risk group, and the difference in prognosis between the two risk groups would be expected to reasonably predict the impact of treatment on the clinical endpoint of interest. For example, there is some empirical evidence that treatment-induced cholesterol reductions lead to reductions in major cardiovascular events that parallel predictions made from population-based epidemiological studies. It remains to be seen whether such predictions can successfully be identified in oncology.

In summary, for a surrogate to be valid, it must show a strong association with the true endpoint at both the individual-level (association between biomarker and clinical endpoint) and the trial-level (association between treatment effects on biomarker and clinical endpoint). The best theoretical setting to test both conditions consists of several large-scale randomized clinical trials in the context of a meta-analysis, or a large randomised trial that can be broken down in smaller, clinically meaningful units, such as countries or clinical sites. In order for a surrogate biomarker to be used in the clinical development of a new drug, surrogacy would have to be demonstrated across a range of treatments testing multiple new drugs, or one new drug compared with multiple comparators. This again suggests recourse to a meta-analysis of several randomised trials [24].

Study designs for prospective biomarker validation

The previous section discussed the evidence required to identify various types of biomarkers. A biomarker which is suggested by tumour biology or by the mode of action of a new treatment will generally need to be validated in a prospective experiment. This last section briefly discusses designs appropriate to validate prognostic or predictive biomarkers. Prospective validation of surrogate biomarkers is a long way down the road and will not be discussed here.

Prognostic biomarkers

Simple cohort studies are sufficient to confirm the impact of prognostic biomarkers. In practice, however, such biomarkers will often be validated in the context of a randomised trial, in which case their predictive ability will also be analysed. Assuming a randomised trial is contemplated to compare an experimental treatment with a standard one, the simplest design randomises all patients to receive either treatment either without taking any account of the biomarker (a 'completely randomised design'), or after stratification for the biomarker (a 'stratified randomised design'), as shown in Fig. 2A. The latter design is almost always preferable to the former, since it balances treatment allocation in patients with biomarker positive and negative, but neither design uses the biomarker status to modify the treatment allocation. As such these designs are useful primarily to identify new biomarkers, rather than to validate a biomarker which has a high likelihood of being truly predictive. Of note, an imbalanced allocation could be used in order to assign a larger proportion of patients to either treatment depending on the biomarker status.

Sometimes interest focuses on showing that a simple biomarker such as a gene profile has better prognostic value than the clinical and pathological characteristics of the patients, and as such should be used to identify high risk patients who need more aggressive treatment (assumed here to be the experimental treatment). In this case, the best design consists of randomising patients whose risk classification is discrepant whether it is based on traditional clinico-pathological characteristics or on the biomarker [25, 26]. Patients whose risk assessment is the same using traditional clinico-pathological characteristics as with the biomarker receive the treatment adapted to their risk. This approach is illustrated schematically in Fig. 2B and was adopted for the MINDACT trial ("Microarray In Node-negative Disease may Avoid Chemotherapy Trial"), a clinical trial for women with early breast cancer carried out under the auspices of the European Organisation for Research and Treatment of Cancer (EORTC) [27].

Predictive biomarkers

For biomarkers that are already well documented but still await formal validation, various more complex designs have been suggested. These designs aim at validating the biomarker while taking advantage of their potential for predicting clinical outcome, thus maximizing the potential for patient benefit. The first design, shown in Fig. 2C, is called the 'marker-based strategy design'. Patients are randomly assigned to have their treatment determined by their marker status or to receive treatment independent of their marker status [28,29]. In this design, patients in the marker-based group receive the experimental treatment if they have the marker and the standard treatment otherwise, while all patients in the non-marker-based

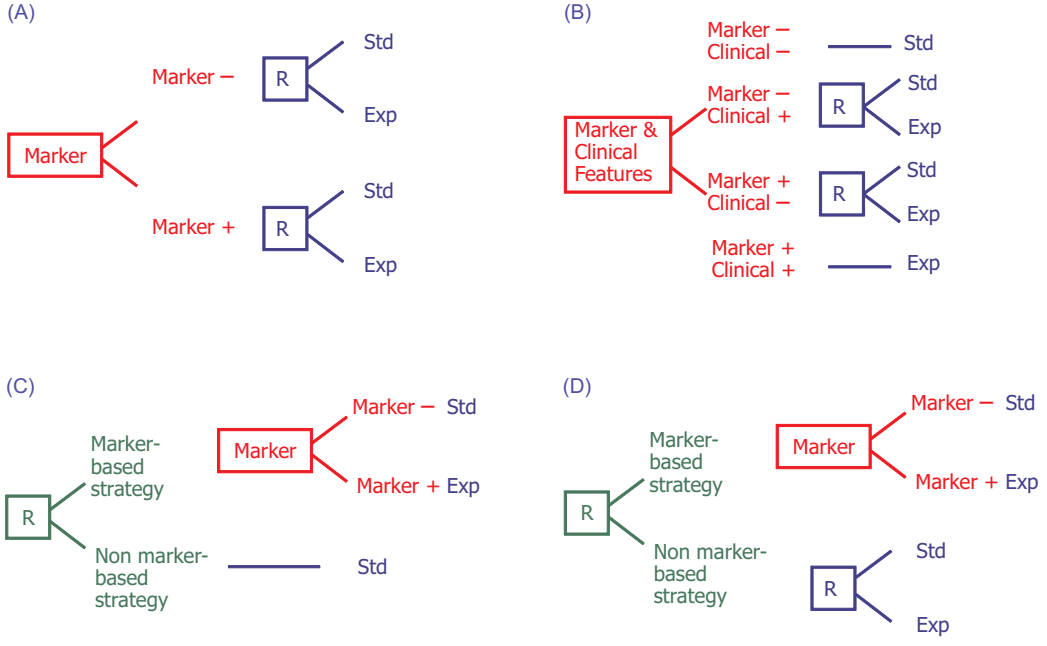

Fig. 2. Clinical trial designs for prospective biomarker validation. (A) Stratified randomised design: patients are stratified according to biomarker status (positive or negative) and randomised to either standard or experimental treatment. (B) Discrepant case randomised design: patients whose biomarker status (positive or negative) differs from clinico-pathological status are randomised to either standard or experimental treatment. (C) Marker-based strategy design: patients are randomly assigned to have their treatment determined by their marker status or to receive standard treatment. (D) modified marker-based strategy design: patients are randomly assigned to have their treatment determined by their marker status or to be randomised again to either standard or experimental treatment.

group receive the standard treatment (i.e. the treatment they would have received outside of the trial). In this design, the marker-based group might do better if the experimental treatment had a true effect in all patients, regardless of marker status. A second design, called the 'modified marker-based strategy design', reduces this problem by randomising patients in the non-marker-based group to receive either the experimental treatment or the standard treatment [28,29]. This design is shown in Fig. 2D.

Recently, a two-stage design has been proposed to identify a predictive gene expression profile, and to validate it in a single prospective trial [30]. In stage 1, the gene expression profile is identified to predict whether a patient is more likely to benefit from the experimental treatment compared with the standard one (patients 'sensitive' to the experimental treatment). The gene expression profile is prospectively applied to identify the subset of sensitive patients among the stage 2 patients, rather than to restrict the entry of stage 2 patients. The final analysis of the trial consists of a comparison of the experimental treatment with the standard treatment in the whole trial, as well as in the subset of the stage 2 sensitive patients [30], with proper adjustment of the significance level of each test to keep the overall significance level under an acceptable value such as 0.05. It is likely that adaptive designs will increasingly be used to take advantage of knowledge that emerges during patient accrual, so as to facilitate a rapid and efficient evaluation of new cancer treatments.

Conflict of interest statement

None declared.

References

1 Lesko LJS, Atkinson AJ. Use of biomarkers and surrogate end-points in drug development and regulatory decision making: criteria, validation, strategies. *Ann Rev Pharmacol Toxicol* 2001, **41**, 347–366.
2 Biomarker Definition Working Group (2001). Biomarkers and surrogate end-points: preferred definitions and conceptual framework. *Clin Pharmacol Therapy* 2001, **69**, 89–95.
3 Temple, RJ. A regulatory authority's opinion about surrogate endpoints. In: Nimmo WS, Tucker GT, eds. *Clinical measurement in drug evaluation*. New York, John Wiley & Sons, 1995, 3–22.
4 Desmedt C, Sotiriou C. When should I start using a new biomarker? *Eur J Cancer* 2007, (this volume).
5 Betensky RA, Louis DA, Cairncross JG. Influence of unrecognized molecular heterogeneity on randomised clinical trials. *J Clin Oncol* 2002, **20**, 2495–2499.

6 Sridhara R, Eisenberger MA, Sinibaldi VJ, *et al.* Evaluation of prostate-specific antigen as a surrogate marker for response of hormone-refractory prostate cancer to suramin therapy. *J Clin Oncol* 1995, **13**, 2944–2953.

7 Smith DC, Dunn RL, Stawderman MS, *et al.* Change in serum prostate-specific antigen as a marker of response to cytotoxic therapy for hormone-refractory prostate cancer. *J Clin Oncol* 1998, **16**, 1835–1843.

8 Bloom JC, Dean RA, eds. *Biomarkers in clinical drug development*. New York, Marcel Dekker, 2003.

9 Collette L, Burzykowski T, Carroll KJ, *et al.* PSA is not a valid surrogate endpoint for overall survival in patients with metastatic prostate cancer. *J Clin Oncol* 2005, **23**, 6139–6148.

10 Buyse M, Thirion P, Carlson RW, *et al.* Relation between tumour response to first-line chemotherapy and survival in advanced colorectal cancer: a meta-analysis. *Lancet* 2000, **356**, 373–378.

11 Sargent D, Wieand S, Haller DG, *et al.* Disease-free survival (DFS) vs. overall survival (OS) as a primary endpoint for adjuvant colon cancer studies: Individual patient data from 20,898 patients on 18 randomized trials. *J Clin Oncol* 2005, **23**, 8664–8670.

12 Burzykowski T, Molenberghs G, Buyse M, eds. *The evaluation of surrogate endpoints*. New York, Springer, 2005.

13 Hammond MEH, Taube SE. Issues and barriers to development of clinically useful tumor markers: A development pathway proposal. *Semin Oncol* 2002, **29**, 213–221.

14 McShane LM, Altman DG, Sauerbrei W, *et al.* Statistics Subcommittee of the NCI-EORTC Working Group on Cancer Diagnostics. Reporting recommendations for tumor marker prognostic studies. *J Clin Oncol* 2005, **23**, 9067–9072.

15 Hayes DF, Trock B, Harris AL. Assessing the clinical impact of prognostic factors: when is "statistically significant" clinically useful? *Breast Cancer Res* 1998, **52**, 305–319.

16 Hayes DF. Tumor marker development: towards validation of clinically useful biomarkers. *Eur J Cancer* 2007, (this volume).

17 Pepe MS, Janes H, Longton G, *et al.* Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Am J Epidemiol* 2004, **159**, 882–890.

18 Buyse M, Loi S, van't Veer L, *et al.* Independent validation and clinical utility of a 70-gene prognostic signature for patients with node-negative breast cancer. *J Natl Cancer Instit* 2006, **98**, 1183–1192.

19 Peterson B, George SL. Sample size requirements and length of study for testing interaction in a $2 \times k$ factorial design when time-to-failure is the outcome. *Control Clin Trials* 1993, **14**, 511–522.

20 Prentice RL. Surrogate endpoints in clinical trials: definitions and operational criteria. *Statist in Med* 1989, **8**, 431–440.

21 Buyse M, Molenberghs G. Criteria for the validation of surrogate end-points in randomized experiments. *Biometrics* 1998, **54**, 1014–1029.

22 Buyse M, Molenberghs G, Burzykowski T, *et al.* The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics* 2000, **1**, 49–68.

23 Alonso A, Molenberghs G. Surrogate marker evaluation from an information theory perspective. *Biometrics* 2006, **63**, 180–186.

24 Daniels MJ, Hughes MD. Meta-analysis for the evaluation of potential surrogate markers. *Statist in Med* 1997, **16**, 1515–1527.

25 Simon R. Validation of pharmacogenomic biomarker classifiers for treatment selection. *Disease Markers* 2005, **21**, 1–8.

26 Simon R. Roadmap for developing and validating therapeutically relevant genomic classifiers. *J Clin Oncol* 2005, **23**, 1–9.

27 Bogaerts J, Cardoso F, Buyse M, *et al.* Gene signature evaluation as a prognostic tool: challenges in the design of the MINDACT trial. *Nature Clinical Practice Oncol* 2006, **3**, 540–551.

28 Sargent D, Allegra C. Issues in clinical trial design for tumor marker studies. *Semin Oncol* 2002, **29**, 222–230.

29 Sargent DJ, Conley BA, Allegra C, Collette L. Clinical trial designs for predictive marker validation in cancer treatment trials. *J Clin Oncol* 2005, **23**, 2020–2027.

30 Freidlin B, Simon R. Adaptive signature design: an adaptive clinical trial design for generating and prospectively testing a gene expression signature for sensitive patients. *Clin Cancer Res* 2005, **11**, 7872–7878.