What should physicians look for in evaluating prognostic gene-expression signatures?

Jyothi Subramanian and Richard Simon

Abstract | Most cancer treatments benefit only a minority of patients. This has led to a widespread interest in the identification of gene-expression-based prognostic signatures. Well-developed and validated genomic signatures can lead to personalized treatment decisions resulting in improved patient management. However, the pace of acceptance of these signatures in clinical practice has been slow. This is because many of the signatures have been developed without clear focus on the intended clinical use, and proper independent validation studies establishing their medical utility have rarely been performed. The practicing physician and the patient are thus left in doubt about the reliability and medical utility of the signatures. We aim to provide guidance to physicians in critically evaluating published studies on prognostic gene-expression signatures so that they are better equipped to decide which signatures, if any, have sufficient merit for use, in conjunction with other factors in helping their patients to make good treatment decisions. A discussion of the lessons to be learned from the successful development of the Onco*type* DX® genetic test for breast cancer is presented and contrasted with a review of the current status of prognostic gene-expression signatures in non-small-cell lung cancer.

Subramanian, J. & Simon, R. *Nat. Rev. Clin. Oncol.* advance online publication 27 April 2010; doi:10.1038/nrclinonc.2010.60

Introduction

A biomarker is 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 are biological measurements that may be used for a diverse range of medical purposes; for example, diagnostic, prognostic, predictive, pharmacodynamic or a surrogate end point. Effective development and validation of biomarkers depends critically on the intended use of the marker. A gene-expression signature is a biomarker in which the expression levels of multiple genes are combined in a defined manner to provide either a continuous score or a categorical classifier. Continuous scores are usually converted to classes for ease of therapeutic decision making. Two common applications for gene-expression signatures are as prognostic or predictive biomarkers.

Prognostic signatures are baseline measurements that provide information about the long-term outcome for untreated patients or those receiving standard treatment. The motivation for studying untreated patients is to understand the biology of the disease unperturbed by treatment. Since cancer patients are usually treated, studying completely untreated patients is usually not an option. Prognostic signatures developed by studying cancer patients undergoing a defined standard treatment can be used to identify patients with a poor prognosis who receive that treatment and hence require more-aggressive treatment. Alternatively, prognostic signatures can also be used to identify patients with sufficiently good prognosis with the standard treatment that they would not require additional treatment. Prognostic signatures, however, need to be developed and validated with a defined intent of use to be useful for therapeutic decision making. For example, the Onco*type* DX® (Genomic Health, Redwood City, CA) recurrence score was validated on node-negative breast cancer patients with estrogen receptor (ER)-positive tumors who received only hormonal therapy in order to identify patients whose prognosis was sufficiently good that they may not require chemotherapy in addition to hormonal therapy.[2]

Predictive signatures are baseline measurements that identify patients who are likely or unlikely to benefit from a specific treatment. For example, *HER2* amplification is a predictive signature for benefit from trastuzumab (Herceptin®, Genentech Inc. South San Francisco, CA) and perhaps also from doxorubicin[3] and paclitaxel.[4] A predictive signature could also be used to identify patients who are poor candidates for a particular drug; for example, colorectal cancer patients whose tumors have *KRAS* mutations seem to be poor candidates for treatment with *EGFR* inhibitors.[5]

When validated, prognostic and predictive signatures can guide personalized treatment decisions for the individual patient. This results in a better balance between therapeutic efficacy and toxic side-effects, which leads to improved patient outcomes. There is substantial literature on the process of developing prognostic classifiers based on high-dimensional gene-expression data,[6–10] which is beyond the scope of this Review. Analysis of high-dimensional data requires the careful

Competing interests
The authors declare no competing interests.

Biometric Research Branch, National Cancer Institute, 9000 Rockville Pike, Bethesda, MD 20892-7434, USA (**J. Subramanian**, **R. Simon**).

Correspondence to: R. Simon
rsimon@mail.nih.gov

© 2010 Macmillan Publishers Limited. All rights reserved

Key points

- Though many gene-expression-based prognostic signatures have been reported in the literature, very few are used in clinical practice
- Developmental studies on prognostic signatures should be designed and analyzed to address a clearly defined, medically important use for such signatures to become useful for improving patient treatment decisions
- Prognostic signatures should be evaluated in independent validation studies before use in clinical practice
- Validation studies should be prospectively planned focused evaluations of whether a previously defined signature improves patient outcome by informing therapeutic decision making compared with use of current practice standards
- The gold standard for establishing clinical utility of a prognostic signature is its validation in a prospective clinical trial to evaluate the medical utility of the proposed signature
- In some cases, focused analysis using archived specimens from multiple suitable clinical trials, if performed under strict conditions, can provide a high level of evidence of clinical utility

usage of sophisticated statistical techniques, including adequate control for false-positive differentially expressed genes, unbiased estimation of the prediction accuracy and use of appropriate statistical tests to evaluate the improvement in prediction accuracy relative to standard covariates. The review by Dupuy and Simon presents a list of common flaws found in publications of gene-expression profiles related to clinical outcomes and provides a set of guidelines for the statistical analysis and reporting of microarray studies for clinical outcomes.[11] The BRB-Array Tools software (Biometric Research Branch, National Cancer Institute, Bethesda, MD) provides extensive resources for development of a wide range of classifiers based on gene-expression data,[12] and can be downloaded.[13] In this Review, we focus on guidelines that physicians could refer to when evaluating studies on prognostic gene-expression signatures. For ease of reference, a checklist of the key points addressed in this article is summarized in Box 1.

A developmental or a validation study

In assessing reports on prognostic signatures, it is very important to distinguish developmental studies from validation studies. A developmental study is one that develops a signature; it is analogous to a phase II clinical trial. A validation study is analogous to a phase III clinical trial where a completely specified signature, developed previously, is tested in a prospective manner under conditions that simulate clinical application of the signature to determine whether the use of the signature results in patient benefit. The majority of the published studies on gene-expression signatures are developmental studies. Since appropriately designed validation studies are expensive to conduct in terms of time and resources, an important objective of a developmental study is to provide unbiased evidence to help decide whether the signature seems sufficiently promising to justify initiating further validation studies. Although to some extent developmental studies can be exploratory in nature, validation studies need to be more focused. Below, we highlight key requirements for developmental and validation studies of prognostic signatures that can serve as a guide for physicians to better evaluate publications on such signatures.

Key points for developmental studies

Most developmental studies of prognostic signatures are retrospective studies conducted on a convenience sample of patients with available tissue. Although a developmental study will generally utilize tissues from patients previously diagnosed, treated and followed, the developmental study should itself be 'prospectively designed' to address a clearly defined, medically important intended use; the patient selection criteria, sample size and the analysis plan should be driven by the intended use of the signature. Moreover, the study should report a completely specified signature and provide an unbiased estimate of the prognostic power of the new signature. These points are elaborated further below. The absence of prospective planning for an intended use has been one of the most common and serious deficiencies in the literature of prognostic markers. The development of Onco*type* DX®, however, provides an illustration of the benefits of focused prospective planning.

Addressing a clinically relevant intended use

Standard guidelines for treatment selection are normally based on easily measurable clinico-pathological factors. For a new prognostic signature to be accepted and widely used it should provide therapeutically relevant information beyond what could be obtained using these practice standards. For example, in the case of non-small-cell lung cancer (NSCLC), post-resection adjuvant treatment decisions are based on tumor stage. For patients with completely resected stage IA NSCLC, adjuvant chemotherapy is considered only in the presence of risk factors that include poor differentiation, vascular invasion, wedge resection and minimal margins.[14] Disease relapse rates, however, are as high as 30% even for these patients with early-stage disease. Hence, a clinically useful objective would be the discovery of gene-expression-based prognostic signatures that could reliably identify high-risk stage IA patients who would benefit from adjuvant chemotherapy.[15] It is also possible that subsets of patients with stage IB and II disease exist who are at a low risk of recurrence so that adjuvant chemotherapy could be withheld. Hence, a second clinically useful objective would be the discovery of gene-expression-based prognostic signatures that identify stage IB or stage II patients who would not require aggressive chemotherapy.

Appropriate patient selection and sample size

Inclusion/exclusion criteria for selecting patients for the gene-expression study and defining the sample size of the study should be carefully planned based on the intended use of the prognostic signature. The principal reason that new signatures are not being used in clinical practice is that most developmental studies on gene-expression signatures are conducted based on a convenience sample of patients with available tissue without adequate emphasis on selecting patients appropriate for

© 2010 Macmillan Publishers Limited. All rights reserved

the intended use. As a result, these studies often include a heterogeneous mix of patients. For example, in a review of prognostic gene-expression studies in NSCLC, most of the studies were found to include patients from stage I up to stage III with the consequence that some studies had a mixed patient population who did and did not receive adjuvant treatments.[16] Showing that a new signature is prognostic for such a heterogeneous group is unlikely to have value for therapeutic decision making and such signatures are rarely used in practice.[17] It is thus highly desirable even for developmental studies to use only patients from a single clinical trial who are homogeneous with respect to treatment received. The sample size for the developmental study should be planned so that there is sufficient number of patients for both developing the classifier and for obtaining precise unbiased estimates of its predictive accuracy.[18]

Box 1 | Guidelines for reports on prognostic signatures

Objectives
- Is the study a developmental or validation study?
- Does the study address a clinically important intended use?

Patient selection
- Is the patient selection and sample size appropriate for the study objectives?
- Is the number of appropriate patients sufficient to provide narrow confidence limits for the predicted outcome in risk groups?

Developmental studies
- Has the signature been internally validated in an unbiased manner?
  1. Does the study state explicitly that internal validation was conducted using either split-sample or complete cross-validation?
  2. If the study uses cross-validation, does the study explicitly state that the model was built from scratch for each training set, including gene selection?
  3. If split-sample is used for internal validation, is there sufficient number of patients in the validation set to give narrow confidence intervals for the predicted outcomes within risk groups?
- Does the study report *P*-values from an appropriate statistical test to demonstrate that the prediction accuracy for the signature is better than chance?
- Does the signature demonstrate significantly better prediction accuracy than standard risk factors using separation of Kaplan–Meier curves within each level of standard prognostic factors?
- Does the study provide a completely specified signature for others to independently validate?

Validation studies
- Is the assay standardized and analytically validated?
- Is the validation conducted using a prospective clinical trial? If yes, is the design of the trial appropriate and efficient?
- Is the validation conducted using specimens archived from a clinical trial? If yes, is the study carried out under strict conditions listed in Box 2?

Unbiased internal validation of signatures

An important objective of a developmental study is to provide an estimate of whether the signature is promising enough to warrant a validation study. Hence, internal validation of the signature is required even in the developmental phase. In order to obtain an unbiased estimate of the prediction accuracy of a signature from gene-expression studies, where typically one deals with high-dimensional data (that is, the number of genes [variables] is much larger than the number of patients [samples]), it is necessary to separate the data used for developing the signature (training set) from the data used for estimating its predictive accuracy (validation set).[19] Estimates of prediction accuracy computed on the same training set that was used to develop the signature are called 'resubstitution estimates'. In high-dimensional settings, the model can be optimized so as to fit the training set perfectly, but with poor ability to predict for new data. Hence resubstitution estimates of prediction accuracy are highly optimistically biased and should not be reported in publications.

Unbiased estimates of prediction accuracy can be obtained using the methods of split-sample or cross-validation.[20] In the split-sample method, the data in the developmental study is randomly partitioned into training and validation sets. The training set is used for developing the signature. After a single completely specified signature is developed using the training set, it is applied to the cases in the validation set and the prediction accuracy is estimated. In contrast to this simple approach, cross-validation methods utilize the available data more efficiently. In cross-validation, the available data is partitioned into *K* subsets. Each subset is in turn left out during training. The model is trained on the union of the remaining *K* minus one subsets and predictions are obtained for the left out subset. After the *K* rounds of training and testing are complete, all the test set predictions are used to estimate the accuracy. In cross-validation, it is imperative that for each training-test partition, the model is developed from scratch based only on the training set, otherwise, again, considerable optimistic bias will be introduced.[19] In particular, it is invalid to select the genes beforehand using all the data and then to simply cross-validate the model-building process for that restricted set of genes. Proper application of these techniques gives a nearly unbiased estimate of the prediction accuracy and only these estimates should be reported in publications.

Demonstrating improvement in prediction accuracy

Standard adjuvant treatment guidelines are based on easily measurable clinico-pathological prognostic factors. New genomic signatures are best used in conjunction with prognostic factors that represent the standard of care, and developmental studies should establish that the new signature substantially improves the predictive accuracy compared with the use of standard prognostic factors alone. Hazard ratios and statistical significance of regression coefficients in a multivariate analysis do not accomplish this objective because they do not adequately measure predictive accuracy.[21] The *P*-values corresponding to a test of significance of hazard ratios essentially test whether the hazard ratio is equal to 1 and not whether the signature improves upon the prediction obtained through standard prognostic factors.[22]

To demonstrate that the new signature has significantly better prediction accuracy than standard prognostic factors, Kaplan–Meier survival curves showing

© 2010 Macmillan Publishers Limited. All rights reserved

the separation of the risk groups predicted by the new signature within each level of the standard prognostic factors should be calculated. The prediction accuracy required to justify validation studies for a new signature depends on the standard of care for the disease and the intended use of the signature. For example, in the case of node-negative, ER-positive breast cancer patients, the standard of care following resection and radiation is chemotherapy and hormonal therapy and a clinically important use for a new genomic signature is the identification of subsets of these patients who could be spared cytotoxic chemotherapy. In such situations, even demonstrating a statistically significant separation of the risk groups is not sufficient for indicating that the predicted risk of recurrence for the low-risk group is sufficiently small to justify withholding cytotoxic chemotherapy. The recurrence rate for the low-risk group should be sufficiently small in absolute terms and precisely determined for the signature to be useful for this purpose.

If a combination of standard risk factors is available, the change in the area under the receiver operating characteristic (ROC) curve (denoted by AUC) can be analyzed to test whether the new signature has statistically significantly better prediction accuracy than the combination of standard prognostic factors.[22] The ROC curve is a graph of the sensitivity of a test versus one minus the specificity of that test as a function of the threshold for positivity of the test. ROC curves are frequently used with binary clinical outcome data to avoid having to pre-select positivity thresholds for the two prognostic tests compared. Larger AUCs imply better predictions. For survival data, the change in the area under the time-dependent ROC curve can be used as a measure of improvement in prediction accuracy.[23] AUCs evaluated at various time points can used to estimate the predictive accuracy of the signature over time.

Is the reported signature completely specified?

Even if internal validation in a developmental study shows that the new gene-expression signature has excellent predictive accuracy, this is not sufficient evidence for its clinical application. This is in part because developmental studies are often conducted on patients available at a small number of centers and the assay is typically performed at one time in one research laboratory. As a result of this, many sources of variation occurring in the clinical setting due to differences in sample collection, tissue handling and assay performance that may influence the predictive accuracy of a signature classifier are not taken into account in the internal validation.[24] Hence, an independent validation study that simulates the application of the signature in a clinical trial setting should follow a successful developmental phase. Developmental studies also typically demonstrate that a signature is prognostic, but a validation study is necessary to demonstrate that it is actionable and results in patient benefit. That is, developmental studies demonstrate 'clinical validity' but not 'medical utility'. Consequently, validation studies should always follow a successful development. To enable independent validation of a new signature by other investigators, the developmental study should report a completely specified signature-based 'classifier' that translates the gene-expression profile to risk groups. Complete specification of the signature includes not just the list of significant genes, but also the mathematical form used to combine expression levels for the genes used in the signature, weights for relative importance of genes, and cut-off values for forming the risk groups.

Key points for validation studies

The ASCO Tumor Markers Guidelines Committee recommended five levels of evidence (LOE) that could be used to evaluate a tumor marker.[25] The levels range from I to V with level I (LOE I) being definitive evidence. Level I evidence for establishing the clinical utility of a new signature is obtained through compelling results from a prospective clinical trial that is specifically designed to test the signature. In this article, we discuss trial designs for prognostic signatures. For a thorough discussion on trial designs for predictive signatures, readers can refer to two published reviews.[26,27] Before the initiation of validation studies the assay should be standardized and undergo analytical validation. Also, the signature should be completely specified and should not be one of many signatures to be evaluated in an exploratory analysis.

Prospective trials for validation of signatures

The marker strategy design is a direct, though often inefficient design for prospectively validating the medical utility of a prognostic signature.[28] In this trial design, patients are randomized to be tested using the signature or not. For patients who are not tested, treatment decisions are made using standard prognostic factors and practice guidelines. For patients who are assigned to be tested, the signature is used, in conjunction with standard prognostic factors for treatment decisions. The trial is evaluated by comparing outcomes overall for the two randomization groups.

The marker strategy design, however, is often very inefficient. Since many patients may receive the same treatment regardless of the randomization group to which they are assigned, a huge sample size may be needed to allow adequate statistical power to detect differences in outcome for the subset of patients for whom treatment assignment is changed by use of the signature.[26] Outcomes for that subset cannot be evaluated directly using the marker strategy design because the signature is not measured for the standard of care group. Moreover, if the analysis is to demonstrate that withholding chemotherapy for patients predicted to be low-risk by the signature is not inferior to adding chemotherapy, this inefficiency is amplified, and even with a huge sample size, the results are not likely to be convincing. The marker strategy design can also be poorly informative in cases where the treatment implications of the marker are complex because subset analyses are not possible since the signature is not measured for the controls.

A modified marker strategy design can be used to avoid the deficiencies of the marker strategy design. In

© 2010 Macmillan Publishers Limited. All rights reserved

this modified design the signature is measured in all patients and patients are only randomized when the treatment assignment that is based on the signature differs from treatment assignment that is based on the standard of care. To illustrate this strategy, consider the case of NSCLC where the standard of care for stage IA patients is to withhold chemotherapy. To validate a prognostic signature for stage IA NSCLC, a trial may be designed in the following manner: The signature is measured in all eligible stage IA NSCLC patients. Patients predicted to be low-risk by the signature are taken off the study. Patients predicted to be high-risk are randomly assigned to receive either chemotherapy or no chemotherapy and outcomes are compared. This design presumes, however, that the standard of care, as a function of standard prognostic variables, is determined. This strategy of testing all patients up-front is used by the Trial Assigning Individualized Options for Treatment (Rx) (TAILORx) study for evaluating the Onco*type* DX® recurrence score in patients with breast cancer.[29]

A primary objective of the TAILORx study is to determine whether adjuvant hormonal therapy alone is not inferior to adjuvant chemotherapy and hormonal therapy in women who meet established clinical guidelines for adjuvant chemotherapy and have an intermediate Onco*type* DX® recurrence score of 11–25. A second objective of the trial is to determine the long-term prognosis and efficacy of hormonal therapy alone in the low-recurrence group (recurrence score <11). The recurrence score ranges used in the initial validation studies of Onco*type* DX® were: low-risk (<18), intermediate-risk (18–30), and high-risk (≥31). The ranges for defining the groups for the TAILORx trial were lowered (low-risk [<11], intermediate-risk [11–25], and high-risk [≥25]) to minimize the potential for undertreatment in both the intermediate-risk and the high-risk groups.[2] The new cutoffs were pre-specified in the trial protocol and would not be changed in an exploratory manner during the trial. The design of this trial is based on a modified marker strategy design and all eligible patients are tested for the signature upfront. Patients with low predicted risk are assigned to hormonal therapy alone and patients with high predicted risk are assigned to chemotherapy and hormonal therapy. Patients with intermediate predicted risk are randomized to either hormonal therapy or chemotherapy and hormonal therapy. If the recurrence score is accurate, the relapse rate for the low-risk group would be very low and hence the potential benefit of adding chemotherapy would be very small. For the intermediate risk group, an absolute 3% decrease in the 5-year disease-free survival (DFS) rate from 90% with chemotherapy to 87.0% or lower on hormonal therapy alone would be considered unacceptable.[29]

A second clinical trial that uses the modified marker strategy design is the MINDACT (Microarray in Node-Negative Disease May Avoid Chemotherapy) trial, which is designed to prospectively evaluate MammaPrint® (Agendia, Amsterdam, The Netherlands), a 70-gene prognostic signature for guiding adjuvant treatment decisions in node-negative breast cancer patients.[30]

Box 2 | Conditions to be satisfied by studies on archived specimens

- The prospective clinical trial with archived specimens that serves as the basis of the validation study should have essentially the same eligibility criteria and structure as would have been used if it were prospectively designed to evaluate the prognostic signature. The number of patients included will, however, in many cases be less.
- The validation study should evaluate a single completely specified signature. If any of the parameters for the signature are optimized during validation, the study is no longer a proper validation study unless special analysis methods are utilized to adjust for the optimization.
- The assay should be analytically validated for use with archived tissue. Variation due to both pre-analytical factors, such as tissue collection, processing, storage, and preparation, as well as analytical factors, such as reagent choice, incubation time and conditions, and method of readout should be minimized.
- The number of patients in each clinical trial used for the evaluation should be large and specimens should be available on a large predominance of the patients in the clinical trial.
- The archived tissues should not be assayed until a new protocol has been written that focuses on the evaluation of the new signature with a completely specified statistical analysis plan.
- The assay should be performed blinded to the clinical data.
- Two or more individually adequate clinical trials should be available for evaluation of archived specimens in the manner described above.

Signature validation using archived specimens

Although the gold standard for establishing the clinical utility of a new signature is through compelling results from one or more prospective clinical trials, such trials may not always be feasible because they may sometimes involve withholding standard therapy. This situation arises, for example, if the objective is to identify low-risk, stage II breast cancer patients for whom the standard of care chemotherapy could be withheld. Prospective clinical trials also require many thousands of patients, can take many years to conduct and can be quite expensive. The TAILORx study, for example, opened in 2006 and the results of this trial are expected to be reported only by 2013.[29] Hence there is a real possibility that other developments might make the test being evaluated obsolete by the time the trial is completed.

In cases where uniformly collected specimens are available for patients on multiple large appropriately designed clinical trials, non-exploratory focused analysis of those trials with regard to an analytically validated assay for a single prospectively (before analysis) defined signature using the archived specimens can sometimes provide the same high level of evidence of clinical utility and can speed up the incorporation of signatures into clinical practice. A refinement to the previously published LOE scale has been recently proposed that sets the requirement criteria for the design and analysis of validation studies using archived specimens.[24] To avoid any form of bias, these studies need to be conducted under strict conditions as outlined in Box 2.

The results from a validation study using archived specimens should be confirmed using specimens from a second study based on archived tissue from a different trial designed, conducted, and analyzed in a manner similar or identical to the initial one. Both validation

© 2010 Macmillan Publishers Limited. All rights reserved

Box 3 | Key points of the development of the Onco*type* DX® assay

Focus on an important intended clinical use
Identification of ER-positive, node-negative breast cancer patients receiving endocrine therapy for whom the risk of recurrence is sufficiently low that chemotherapy can be avoided.

Analytical validation of the assay
The analytical properties of the Onco*type* DX® assay were investigated and it was concluded that the operational performance specifications defined for the Onco*type* DX® assay allow reporting of quantitative recurrence score values for individual patients with a standard deviation within two recurrence score units on a 100-unit scale. Furthermore, an analysis of study design showed the assay imprecision contributed by instrument, operator, reagent, and day-to-day baseline variation to be low.[40]

Validation on a large, therapeutically relevant, separate group of patients than those used for developing the signature
The 21-gene signature was validated on 668 ER-positive, node-negative patients from the tamoxifen arm of the NSABP B-14 trial. The results of this validation study showed that the Kaplan–Meier estimate for the proportion of patients free of distant recurrence at 10 years was 93.2% and 69.5% for the predicted low-risk and high-risk groups, respectively ($P<0.001$), demonstrating the prognostic value of the signature.[41] The recurrence rate for the low-risk group (93.2%) was low enough to make the signature therapeutically relevant.

Convenience
The assay requires only formalin-fixed paraffin-embedded tissue specimens making its use practical in standard practice.

Prospective validation
Onco*type* DX® is being further prospectively evaluated in the ongoing TAILORx clinical trial to determine whether adjuvant hormonal therapy alone is as effective as adjuvant chemohormonal therapy in women who meet established clinical guidelines for adjuvant chemotherapy and have a low or intermediate Onco*type* DX® recurrence score.[29]

studies need to be performed using the same or similar assay, should address the same end point and the end point should reflect medical utility. In order to reach the level of evidence required to change clinical practice, the results of both validation studies must be consistent and provide equally compelling results. If the results of the validation studies are inconsistent, it can be at the most considered only level II evidence for the signature.

A validation study using specimens archived from the SWOG-8814 trial was recently conducted to determine if the Onco*type* DX® recurrence score could also be used to identify ER-positive breast cancer patients with node-positive disease, for whom chemotherapy offered little benefit.[31] The results of this study showed that women with low recurrence scores got little or no benefit when anthracycline-based chemotherapy was added to tamoxifen, while those with higher scores derived a substantial benefit, independent of the number of positive nodes. A prospective randomized clinical trial for a validation study in this situation would have been very difficult as it would involve withholding standard of care chemotherapy for some node-positive breast cancer patients.

Prognostic signatures in NSCLC

Current guidelines for treatment decisions in NSCLC are based on the TNM staging system and certain additional clinico-histopathological parameters.[14,32] As outlined earlier, a new gene-expression-based prognostic signature for NSCLC might be considered clinically useful in the following circumstances: first, if it is more effective than standard risk factors in identifying high risk, completely resected stage IA patients who might benefit from adjuvant chemotherapy or, second, it identifies stage IB or stage II patients who have low risk of recurrence without chemotherapy.

We recently conducted a review to critically evaluate studies reporting the development and/or validation of prognostic gene-expression signatures in NSCLC.[16] The objective of the review was to determine whether the studies were planned and conducted in a manner that provided evidence of clinical utility for the reported signature beyond that obtained by standard practice guidelines. For this purpose, the studies were evaluated on the basis of criteria representing the various points presented in the previous sections discussed above. The three major criteria used for the evaluation were: appropriateness of the study protocol, statistical validation of the prognostic models and presentation of results, and finally demonstration of medical utility for the signature.

All the studies reviewed were developmental studies reporting a new prognostic signature. It was evident from the review that most of the studies failed to place adequate importance to either patient selection or sample size planning. More than 50% of the studies presented biased resubstitution results where the same data used to develop the prognostic signature was used to evaluate the accuracy of the predictions. Also, most studies failed to report completely specified models that would facilitate further independent validation of the signatures. Most importantly, although most studies presented validation results on data not used for developing the prognostic signatures, these validation attempts failed to present sufficient evidence for the usefulness of the new signature in making improved treatment decisions in NSCLC disease stages IA, IB or II, and failed to establish the usefulness of the gene-expression signatures over and above known risk factors.

Thus far, no prognostic signature for NSCLC has demonstrated sufficient medical utility to be incorporated into standard treatment guidelines. Recently, a prospective clinical trial, CALGB-30506, was launched to compare adjuvant chemotherapy versus observation in treating patients with stage I NSCLC. The lung metagene signature will be used as a stratification factor in this trial to evaluate the potential utility of this signature for identifying stage I patients who benefit from adjuvant chemotherapy.[33,34]

Lessons learned from Onco*type* DX®

Onco*type* DX® is a diagnostic test comprised of a 21-gene assay applied to formalin-fixed, paraffin-embedded breast cancer tissue. The result of this assay is expressed as a recurrence score. Following development, the assay was analytically validated for reproducibility and robustness on paraffin preserved tissue and clinically validated on an independent set of ER-positive, node-negative patients from the tamoxifen arm of the NSABP B-14 trial.[35,36]

© 2010 Macmillan Publishers Limited. All rights reserved

Onco*type* DX® is currently being prospectively evaluated in the TAILORx trial. Key considerations that went into selecting Onco*type* DX® for evaluation in this trial were as follows: first, the assay was analytically validated, second, the assay only required formalin-fixed paraffin-embedded tissue specimens routinely processed in clinical pathology laboratories, third, the assay was developed for and validated in breast cancer patients with hormone-receptor positive, node-negative disease receiving tamoxifen, who represented a clinically important group. Finally, initial validation results indicated that the risk of recurrence of patients with a low recurrence score was sufficiently low to be used for informing treatment decisions.[2,29]

Numerous prognostic signatures have been developed for breast cancer and four signatures (MammaPrint®, Onco*type* DX®, Theros Breast Cancer Index℠ [Bio-Theranostics, San Diego, CA] and MapQuant Dx™ [Ipsogen, Marseilles, France and New Haven, CT, USA]) are commercially available.[37] In a comparative study of gene signatures in breast cancer, Fan *et al.*[38] found that although the signatures had little overlap in terms of gene identity, there was a high concordance in their outcome predictions for individual patients. Wirapati *et al.*[39] conducted a meta-analysis involving 2,833 breast tumors from several publicly available breast cancer gene-expression studies. The results of this study showed that all nine prognostic signatures that were compared exhibited similar prognostic performance in this large dataset and that their prognostic capabilities were mostly driven by proliferation-related genes.

All signatures currently commercially available for breast cancer have been shown to correlate with prognosis and the laboratories that perform the assays are CLIA (Clinical Laboratory Improvement Amendments) certified.[35] The MammaPrint® signature has also been cleared by the US FDA as a class 2, $510(k) product.[40] However, ASCO recommends only the Onco*type* DX® signature to predict risk of recurrence in node-negative, estrogen receptor-positive breast cancer patients treated with tamoxifen, as they believe that establishing clinical utility for a clearly defined intended use for the other signatures need further investigation.[41] The Onco*type* DX® assay only requires formalin-fixed paraffin-embedded tissue specimens making its use practical in standard clinical practice, compared with MammaPrint®, which requires snap-frozen tissue. The key points to be learned from the successful development of Onco*type* DX® are important for any gene-expression signature study and are further elaborated in Box 3.

Conclusions

Gene-expression signatures offer the promise of optimizing treatment decisions for individual cancer patients. However, the complexity of cancer biology, the complexity involved in the analyses of high-dimensional data, and lack of focus in the development and validation of prognostic signatures have proven to be formidable challenges in the move towards a more-predictive oncology. Prognostic signatures need to be developed and validated with focus on a therapeutically important intended use right from the start. Successful validation of a genomic signature in a developmental study is not sufficient evidence for its incorporation into clinical practice. Ideally, a new prognostic signature will be adopted into clinical practice only on the basis of evidence from a prospective randomized clinical trial. Two such clinical trials in breast cancer—the TAILORx trial evaluating the Onco*type* DX® signature and the MINDACT trial evaluating the MammaPrint® signature are currently in progress.[29,30] However, such prospective trials may not always be feasible or may not (in all cases) be the most-effective way to proceed in order to bring important prognostic decision tools into clinical practice.

Validation studies using archived specimens from suitable clinical trials, if performed under strict conditions, can also provide a high level of evidence of clinical utility.[24] In fact, one of the major aims of both the TAILORx and the MINDACT trials is the establishment of large, well annotated tissue banks.[29,30] These tissue banks would be a source of high-quality data for future research on breast cancer prognosis and ultimately for clinical management in the future. We hope that the guidelines presented in this Review will be helpful for physicians in evaluating publications reporting gene-expression signatures, and also for investigators planning new developmental and validation studies on gene-expression signatures.

Review criteria

Information was obtained by searching the PubMed database for articles published in English before 28 February 2010. The search terms included "developmental studies", "validation studies" in association with the terms "prognostic gene expression signatures"; "Onco*type* DX", "MammaPrint", "TAILORx", "MINDACT" and "breast cancer prognostic signatures". Where appropriate, reference lists of the primary articles were checked for additional relevant material. The review criteria for identifying NSCLC signatures are outlined in reference 16. No additional search for NSCLC signatures was conducted.

1. Lesko, L. J. & Atkinson, A. J. Jr. Use of biomarkers and surrogate endpoints in drug development and regulatory decision making: criteria, validation, strategies. *Annu. Rev. Pharmacol. Toxicol.* **41**, 347–366 (2001).
2. Sparano, J. A. & Paik, S. Development of the 21-gene assay and its application in clinical practice and clinical trials. *J. Clin. Oncol.* **26**, 721–728 (2008).
3. Hayes, D. F. Prognostic and predictive factors revisited. *Breast* **14**, 493–499 (2005).
4. Gennari, A. *et al.* HER2 status and efficacy of adjuvant anthracyclines in early breast cancer: a pooled analysis of randomized trials. *J. Natl Cancer Inst.* **100**, 14–20 (2008).
5. 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).
6. West, M. *et al.* Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc. Natl Acad. Sci. USA* **98**, 11462–11467 (2001).
7. Radmacher, M. D., McShane, L. M. & Simon, R. A paradigm for class prediction using gene expression profiles. *J. Comput. Biol.* **9**, 505–511 (2002).
8. Simon, R. *et al.* in *Design and Analysis of DNA Microarray Investigations* (Springer Verlag, New York, 2003).
9. Speed, T. P. (ed) *Statistical Analysis of Gene Expression Microarray Data* (Chapman and Hall, 2003).
10. Wang, Y., Miller D. J. & Clarke, R. Approaches to working in high-dimensional data spaces: gene

© 2010 Macmillan Publishers Limited. All rights reserved

expression microarrays. *Br. J. Cancer* **98**, 1023–1028 (2008).
11. Dupuy, A. & Simon, R. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J. Natl Cancer Inst*. **99**, 147–157 (2007).
12. Simon, R. *et al.* Analysis of gene expression data using BRB-Array Tools. *Cancer Inform.* **3**, 11–17 (2007).
13. National Cancer Institute *Biometric Research Branch Division of Cancer and Diagnosis* [online], http://brb.nci.nih.gov/ (2009).
14. National Comprehensive Cancer Network, Inc. *NCCN Clinical Practice Guidelines in Oncology™ Non-small cell lung cancer © 2009* Available at: www.nccn.org Vol. 2.2009.
15. Potti, A. *et al.* A genomic strategy to refine prognosis in early-stage non-small-cell lung cancer. *N. Engl. J. Med.* **355**, 570–580 (2006).
16. Subramanian, J. & Simon, R. Gene expression-based prognostic signatures in lung cancer: Ready for clinical use? *J. Natl Cancer Inst.* **102**, 464–474 (2010).
17. Bast, R. C. Jr *et al.* 2000 update of recommendations for the use of tumor markers in breast and colorectal cancer: clinical practice guidelines of the American Society of Clinical Oncology. *J. Clin. Oncol.* **19**, 1865–1878 (2001).
18. Dobbin, K. Zhao, Y. & Simon, R. How large a training set is needed to develop a classifier for microarray data? *Clin. Cancer Res.* **14**, 108–114 (2008).
19. Simon, R. *et al.* Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J. Natl Cancer. Inst.* **95**, 14–18 (2003).
20. Molinaro, A. M., Simon, R. & Pfeiffer, R. M. Prediction error estimation: a comparison of resampling methods. *Bioinformatics* **21**, 3301–3307 (2005).
21. 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).
22. Kattan, M. W. Evaluating a new marker's predictive contribution. *Clin. Cancer Res.* **10**, 822–824 (2004).
23. Heagerty, P. J., Lumley, T. & Pepe, M. S. Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* **56**, 337–344 (2000).
24. Simon, R., Paik, S. & Hayes, D. F. Use of archived specimens in evaluation of prognostic and predictive biomarkers. *J. Natl Cancer Inst.* **101**, 1446–1452 (2009).
25. Hayes, D. F. *et al.* Tumor marker utility grading system: a framework to evaluate clinical utility of tumor markers. *J. Natl Cancer Inst.* **88**, 1456–1466 (1996).
26. Simon, R. Clinical trial designs for evaluating the medical utility of prognostic and predictive biomarkers in oncology. *Personalized Med.* **7**, 33–47 (2010).
27. Simon, R. The use of genomics in clinical trial design. *Clin. Cancer Res.* **14**, 5984–5993 (2008).
28. Simon, R. & Wang, S. J. Use of genomic signatures in therapeutics development. *Pharmacogenomics J.* **6**, 166–173 (2006).
29. Zujewski, J. A. & Kamin, L. Trial assessing individualized options for treatment for breast cancer: the TAILORx trial. *Future Oncol.* **4**, 603–610 (2008).
30. Cardoso, F. *et al.* Clinical application of the 70-gene profile: the MINDACT trial. *J. Clin. Oncol.* **26**, 729–735 (2008).
31. Albain, K. S. Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: a retrospective analysis of a randomised trial. *Lancet Oncol.* **11**, 55–65 (2010).
32. Tanoue, L. T. Staging of non-small cell lung cancer. *Semin. Respir. Crit. Care Med.* **29**, 248–260 (2008).
33. Clinicaltrials.gov *Chemotherapy or observation in treating patients with stage I non-small cell lung cancer* [online], http://www.clinicaltrials.gov/ct2/show/record/NCT00863512?term=CALGB-30506&rank=1 (2010).
34. Ullmann, C. D. & McShane, L. Erratum to: Translating genomics into clinical practice: applications in lung cancer. *Curr. Oncol. Rep.* **11**, 413–496 (2009).
35. Cronin, M. *et al.* Analytical validation of the Oncotype DX genomic diagnostic test for recurrence prognosis and therapeutic response prediction in node-negative, estrogen receptor-positive breast cancer. *Clin. Chem.* **53**, 1084–1091 (2007).
36. Paik, S. *et al.* A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N. Engl. J. Med.* **351**, 2817–2826 (2004).
37. Sotiriou, C. & Pusztai, L. Gene-expression signatures in breast cancer. *N. Engl. J. Med.* **360**, 790–800 (2009).
38. Fan, C. *et al.* Concordance among gene-expression-based predictors for breast cancer. *N. Engl. J. Med.* **355**, 560–569 (2006).
39. Wirapati, P. *et al.* Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res*. **10**, R65 (2008).
40. Couzin, J. Diagnostics. Amid debate, gene-based cancer test approved. *Science*. **315**, 924 (2007).
41. Harris, L. *et al.* American Society of Clinical Oncology 2007 update of recommendations for the use of tumor markers in breast cancer. *J. Clin. Oncol.* **25**, 5287–5312 (2007).

© 2010 Macmillan Publishers Limited. All rights reserved