The Oncologist®

Is Molecular Profiling Ready for Use in Clinical Decision Making?

John P. A. Ioannidis

Oncologist 2007;12;301-311
DOI: 10.1634/theoncologist.12-3-301

The Oncologist®
The Community Oncologist

Is Molecular Profiling Ready for Use in Clinical Decision Making?

JOHN P. A. IOANNIDIS

Clinical and Molecular Epidemiology Unit, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, and Biomedical Research Institute, Foundation for Research and Technology-Hellas, Ioannina, Greece and Institute for Clinical Research and Health Policy Studies, Tufts University School of Medicine, Boston, Massachusetts, USA

Key Words. Molecular profiling • Microarrays • Clinical use • Clinical practice • Prediction Prognosis

LEARNING OBJECTIVES

After completing this course, the reader will be able to:

1. Discuss the current status of translational research on molecular profiling for cancer.
2. Highlight the steps and difficulties and biases involved in moving molecular profiling from the bench to the bedside.
3. Propose potential solutions to the challenges of clinical use of this new technology.

CME Access and take the CME test online and receive 1 *AMA PRA Category 1 Credit*™ at CME.TheOncologist.com

ABSTRACT

Molecular profiling, the classification of tissue or other specimens for diagnostic, prognostic, and predictive purposes based on multiple gene expression, is a technology that holds major promise for optimizing the management of patients with cancer. However, the use of these tests for clinical decision making presents many challenges to overcome. Assay development and data analysis in this field have been largely exploratory, and leave numerous possibilities for the introduction of bias. Standardization of profiles remains the exception. Classifier performance is usually overinterpreted by presenting the results as *p*-values or multiplicative effects (e.g., relative risks), while the absolute sensitivity and specificity of classification remain modest at best, especially when tested in large validation samples. Validation has often been done with suboptimal attention to methodology and protection from bias. The postulated classifier performance may be inflated compared to what these profiles can achieve. With the exception of breast cancer, we have little evidence about the incremental discrimination that molecular profiles can provide versus classic risk factors alone. Clinical trials have started to evaluate the utility of using molecular profiles for breast cancer management. Until we obtain data from these trials, the impact of these tests and the net benefit under real-life settings remain unknown. Optimal incorporation into clinical practice is not straightforward. Finally, cost-effectiveness is difficult to appreciate until these other challenges are addressed. Overall, molecular profiling is a fascinating and promising technology, but its incorporation into clinical decision making requires careful planning and robust evidence. *The Oncologist* 2007;12:301–311

Disclosure of potential conflicts of interest is found at the end of this article.

Correspondence: John P.A. Ioannidis, M.D., Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina 45110, Greece. Telephone: 302651097807; Fax: 302651097867; e-mail: jioannid@cc.uoi.gr Received November 1, 2006; accepted for publication January 10, 2007. ©AlphaMed Press 1083-7159/2007/$30.00/0 doi: 10.1634/theoncologist.12-3-301

Introduction

Multidimensional measurement of biological processes has become feasible over the last decade with the advent of suitable technological platforms [1–3]. In theory, this knowledge may lead to optimized, individualized management of patients. Toxicity may be decreased by avoiding unnecessary therapy in patients who have an otherwise excellent prognosis without treatment and in those for whom it can be inferred that they will not respond to available regimens. Efficacy may also be optimized by selecting for treatment those patients who would experience the maximal benefit. The term molecular profiling is used here to collectively describe molecular approaches that concomitantly measure the expression of multiple genes on tissue or other biological samples. The typical application is gene-expression profiling using microarrays.

Cancer applications represent the most prominent medical domain of this exponentially increasing literature [4]. Moreover, breast cancer is the first field in which molecular profiling has been approved and reimbursed for clinical use. Several other molecular profiles are also close to clinical application in cancer patients. The interest is intense: several of the most-cited articles across all medicine in the last decade pertain to molecular profiling [5]. This review examines whether molecular-profiling technologies are likely to affect clinical decision making in the routine management of cancer patients. I discuss what the difficulties are and how they could possibly be bypassed.

Failed Research on Cancer Markers: What Is Different Now?

Cancer markers have a very long research history with many thousands of publications in the last three decades. Nevertheless, despite hundreds of markers being tested, only a handful have moved into clinical practice to date. Some authors have questioned whether this may reflect an inflexibility of the regulatory and translational process [6]. If so, we should find ways to approve more markers and make use of them. However, approval may not be the main obstacle [7]. Even those markers that have been approved have a rather modest performance and limited role, apparently, and this holds true even for the most promising ones, for example, human epidermal growth factor receptor (HER)-2 [8, 9].

The relative failure of cancer markers to date can teach us several lessons that may be useful to also consider for complex molecular profiles. Until now, tumor marker research has depended mostly on small underpowered studies. Most of them represent opportunistic analyses performed post hoc based on readily available databases and assays. Validation of claims has been uncommon, fragmented, and incomplete [10, 11].

What is different now? Studies on molecular profiling don't have larger sample sizes than past one-marker-at-a-time studies [12]. However, they are much larger in terms of complexity and volume of information. In theory, they should capture biological complexity more comprehensively [13]. However, this complexity requires even more heightened attention to robust design, methodological detail, and avoidance of bias. Table 1 shows some prerequisites to making a cancer marker useful for the clinic. One may examine whether the current status of new molecular profiles satisfies these criteria and whether we can bypass the problems of the past.

Assay Development and Standardization

Any high-tech measurement may still have measurement error. Microarray technology is particularly susceptible, because a long chain of decisions on sampling, preprocessing, processing, calibration, and analysis is required. During assay development, in particular, the chain of decisions is long and convoluted. Errors and biases may involve the sampling of the specimens, their quality, the amount of tissue obtained, storage, clinical or laboratory processing, the timing of processing, fixation (paraffin versus frozen tissue), plating, and readout of microchips. Influential design choices include the number of replicates, the control for extraneous factors, the use or not of pooled samples, and how pooling is analyzed. For more details, the reader is referred to excellent reviews [14, 15].

This complexity sensitized researchers early to the need for presenting information in a way that crucial aspects of the measurements are accurately conveyed. The minimum information about a microarray experiment (MIAME) guidelines largely serve this purpose [16]. Nevertheless, it is unclear whether the whole series of complex decisions can be fully captured even with best intentions. This does not pertain only to the laboratory component of the measurements. The analytical calibration and informatics analysis plan can also be very convoluted. Preprocessing of data can be cumbersome. Major decisions need to be made for transformation, normalization, data filtering, use of shrinkage methods, removal of technical artifacts, and whether background subtraction is to be used, to name a few decision nodes [17, 18]. For each decision node, the possibilities are numerous, and new methods and software appear almost weekly in the literature. Additional decisions are made on quality control measures to eliminate, or not, samples and readings as ineligible. Finally, analysis of data can use unsupervised or supervised methods with literally hundreds of minor or major variants on how exactly to arrive at

Table 1. Steps in making molecular profiling useful for clinical decision making

• Assay development and standardization
• Demonstration of diagnostic, prognostic, and predictive performance
• Validation of performance
• Provision of independent information beyond and above classic predictors
• Nonselective and transparent accumulation of evidence
• Demonstration of clinical effect (efficacy)
• Demonstration of benefit on routine clinical use (effectiveness)
• Integration in clinical care
• Cost-effectiveness

a molecular profile. Bias can lurk at each of these steps [19, 20].

The availability of routine informatics platforms does not necessarily improve the situation. Many of the analytical decisions are made with little understanding about what they entail. A side effect of the simplification of commercial statistical software has been the ability of nonexperts to apply them. This risk is magnified with bioinformatics software. Moreover, data may be analyzed with many different approaches, but only the "best" results may be shown.

It is not unfair to say that any molecular profile emerges eventually out of an abyss of experiments and analyses. This is still acceptable, provided that whatever emerges from the abyss is then standardized for further extended testing and practical use. Standardization means that the proposed profile is made definitive and considered fixed for further testing; detailed instructions are provided so that the profile can be measured and analyzed in exactly the same way in different collections of data.

As of now, several molecular profiles have reached the point of being fairly standardized. This is particularly the case for breast cancer, where at least four profiles exist based on a 21-gene recurrence score (Onco*type* DX®; Genomic Health, Redwood City, CA), 70-gene signature, 76-gene signature, and wound-response profile [21–24]. Some profiles for hematological malignancies may also be considered standardized [25–27]. For other malignancies, profiles are mostly in a more exploratory phase.

Standardization should not leave ambivalent points and subjective interpretation. For example, if a molecular profile is supposed to create three categories of high risk, intermediate risk, and low risk, then these three categories should be maintained separate and ordered in all subsequent analyses and evaluations of the profile. Onco*type* DX® has such a categorization. However, in some validation analyses the intermediate category has been merged with the low category, with the excuse that "their survival curves were not significantly different" [28]. Other analyses focus more on the comparison of the extreme (high versus low) categories [21]. Even for such a simple three-class categorization, there are at least seven contrasts that can be conceived of, and of these only the first one is properly representing the original categorization (Table 2). When continuous value cutoffs are involved, changing the cutoff can create a literature with an infinite number of variants of the classifier.

Demonstration of Diagnostic, Prognostic, and Predictive Performance

A molecular profile aims to classify samples appropriately. This could involve the classification of healthy versus diseased cases (diagnostic performance), good versus poor prognosis cases (prognostic performance), and cases with good versus poor outcome after an intervention (predictive performance). A profile is typically generated by training on some data. Unless the training is unsupervised (i.e., no knowledge of the correct class is involved), the performance of the molecular profile on the training dataset is entirely uninformative about its true performance. Analytical techniques are currently available that ensure that we may always reach 100% accuracy on training data, if we so desire. Thus a sensitivity of 100%, specificity of 100%, and $p = 10^{-100}$ in the training dataset means absolutely nothing. In fact, a profile that is perfectly (over)fit to the training data set is unlikely to work well elsewhere.

There are two approaches to test whether the proposed profile is an accurate classifier: cross-validation and independent validation [17]. Regardless of the approach, different metrics may be used to describe the classifier performance. One may present statistical testing measures (e.g., *p*-values), multiplicative effect measures (e.g., likelihood ratios or hazard ratios), or absolute effect measures (e.g., sensitivity and specificity). While all information has some utility, absolute effect measures are most meaningful from a clinical perspective (Table 3) [29].

Validation of Performance

Most studies use only cross-validation to examine the performance of a classifier. Cross-validation entails excluding one or more cases from the training data, training a profile on the remaining data, and then checking whether the resulting profile is an appropriate classifier for the remaining case(s); the process is repeated several times, until all the data points have been selected at some iteration for testing purposes. Cross-validation is a robust technique, because at all times the training and testing data remain distinct. How-

Table 2. Recategorizing a molecular profile that has three categories of risk: possible contrasts

High versus intermediate versus low (ordered categories)
High versus intermediate versus low (without consideration of order)
High versus low
High versus intermediate plus low combined
High versus intermediate
Intermediate versus low
High plus intermediate combined versus low

Table 3. Statistical significance and multiplicative and absolute measures of effect

Rejecting the null hypothesis at $p = 0.05$ simply says that the classifier performs better than chance; this says nothing about how much better than chance the performance is. Likelihood ratios and relative risks or hazard ratios are more useful, especially when accompanied by a measure of uncertainty (e.g., 95% confidence interval). However, they may still be clinically misleading, because they do not convey the absolute scale of classification. For example, what is considered a highly successful validated performance for the 70-gene signature in breast cancer for overall survival [29] corresponds to $p < .001$ and a hazard ratio of 2.2–2.5 (depending on the analysis), but a sensitivity of 90% coupled with a specificity of only about 40%, with the resulting area under the curve (AUC) being only 0.648 for survival. Such sensitivity, specificity, and AUC figures correspond to a poor-to-modest classifier: to identify correctly 90 of the 100 patients with poor prognosis (miss 10 of them), almost 60 of 100 good-prognosis patients are identified as having poor prognosis.

ever, empirical evidence suggests that cross-validation exercises that are reported in the literature tend to have very optimistic performances, even more so when very small studies are involved [12]. This could reflect selective reporting and a mixture of flawed approaches, summarily called incomplete or suboptimal cross-validation. Typical scenarios are shown in Table 4.

Even perfect and complete cross-validation suffers from unknown external validity. At the end of any cross-validation exercise, we don't know how well the proposed profile can perform outside the narrow limits of the data at hand. This is crucial for clinical practice. Moreover, there is the paradox that cross-validation, if performed correctly, cannot lead to a standardized profile, a prerequisite for moving toward clinical practice and broad use. By definition, in complete cross-validation, a different profile is built with each iteration. Therefore, truly assessing the classifier performance of the profile requires independent validation.

Independent validation means that a profile is generated in one data set and is then tested in one or more completely different data sets. It is not the same population being resampled, and there is no overlap between the training and testing datasets. In theory, independent validation is a most rigorous technique. If it is applied correctly, results should be reliable. However, there are many reasons why independent validation may not be performed or reported appropriately, as summarized and exemplified in Table 4. Validation may further be compromised by flexibility in definitions of outcomes, including even survival [30].

These bias threats are not just theoretical concerns. We know that they eventually have an impact in the circulating literature. While we cannot tell which of these several biases has played the key role(s) in each case, we have evidence that the validation performance of several proposed signatures in the literature is inflated. Michiels et al. [31], using a multiple random sampling approach, showed that of seven molecular profiles with proposed high classifier accuracy, five really should not have had classifier accuracy better than chance, if the training and validation had been performed truly without any bias. The other two had only modest classifier accuracy. Some published classifier accuracies were completely incompatible even with the 95% credibility boundary of what one would get based on 500 possible training–testing validations: the published results were far too good to be true. The gradual decline in classifier accuracy across subsequent studies is occasionally alarming. The first paper on the 70-gene signature proposed practically perfect accuracy, while the latest "validation" shows sensitivity of 90% and specificity of about 40% for time to distant metastasis at 5 years, and slightly worse performance for survival at 10 years [22, 29].

Another possible hint to bias with inflated results is the fact that the genes selected for each proposed profile are very unstable. Juxtaposition of the proposed profiles for breast cancer shows that their overlap is minimal to nonexistent. Different splits of the training and validation data result in very different genes being selected [32]. To have some certainty about the selected genes, the required sample size for the training process should be in the range of many thousands [33], that is, about 100-fold larger than the sample sizes that have been used to date. A counterargument, however, is that specific genes are not so important and what matters for a profile is to include some members of key pathways implicated in the biological behavior. Different genes may have interchangeable roles in different profiles [28].

While some interchangeable features make sense, one would be skeptical when nearly all proposed multigene profiles seem to work well. The predictions of the poorly over-

Table 4. Examples of incomplete, suboptimal, or inappropriate practices of cross-validation and independent validation

A. Cross-validation
• Not allowing for different genes to be selected each time; the set of genes may be determined based on the overall dataset, in which case the profile has drawn information from all data points and cross-validation is no longer meaningful.
• Performing multiple analytical models and presenting the best one based on cross-validation performance; this is a major threat, because it is difficult to ensure transparency of reporting all analyses performed.
• Identifying signatures with different numbers of genes and presenting the best one based on cross-validation performance; same reasons as above.
B. Independent validation
• Same team performing/analyzing both sets; while not reproachable per se, it is important to ensure that the data retain complete independence; this cannot be assured when the same investigators use all data sets; ideally independent validations should be performed with sealed protocols and they may also be run by competitors of the investigators who have proposed a profile in the first place.
• Sets may not be fully independent, especially if they come from the same sampling pools and sources; given that sources of specimens are limited, this is not an uncommon scenario; there are grades of independence: samples from the same study, e.g., the National Surgical Adjuvant Breast and Bowel Project (NSABP)-14 trial; samples from the same program/cohort/investigators, e.g., the NSABP; samples from completely different investigators and cohorts.
• Trying different splits of training and validation datasets and selecting the best split; this is one of the many ways in which the availability of all data at a single center can lead to misleading results; in the more generic form, any intended or unintended crosslooking at the two datasets may introduce bias.
• Multiple gene-selection machines validated, only the best results presented; this can introduce bias even when different teams hold the training and testing data; in fact, we usually do not know how many gene-selection machines have been tried in each case; if the number is too large, then one expects one of them to have the best performance, not necessarily because it is the best, but possibly also because of chance.
• Multiple gene sets with different numbers of genes (same machine) validated, only the best results presented; same reasons as above.
• Multiple datasets validated, only the best results presented; same reasons as above.
• Same dataset validated with different definitions and different adjustments, only the best results presented; same reasons as above.

lapping 70-gene signature and 21-gene recurrence score of Onco*type* DX® agree for about 80% of patients, and similar high concordance is seen between other distinct molecular profiles [28]. However, this is based on a dataset that was used in part for the training of three of the five compared profiles (including three of the four that show concordance). In general, keeping the training data into a combined training-plus-testing database spuriously inflates the accuracy and concordance of profiles. Training data should always be discarded in the independent validation process. Figure 1 shows another example, in which consideration of the training and testing data together can make a failed validation seem as if it were highly successful [34, 35].

Provision of Independent Information Beyond and Above Classic Risk Factors

A molecular profile that is standardized, developed in an unbiased manner, and validated to have high classifier accuracy may still have no clinical utility, unless it can offer additional information beyond what classic predictors can already tell us. Classic predictors may include variables that are routinely available and require no effort or little extra cost to obtain them, and others that require some special testing. For example, age, sex, tumor size, and possibly grade and lymph node status, largely belong to the first category. Conversely, estrogen receptor and HER-2 status belong to the second category. To evaluate the incremental information offered by the molecular profile, we need multivariate models that include all the classic risk factors. We also need models that consider both the classic risk factors and the gene-expression profile. These models should be tested in datasets that are entirely independent from those in which the molecular profile was developed, as discussed above. Moreover, in these datasets, the classic risk factors should show the discriminating accuracy that we are used to.

For most applications of molecular profiling to date, such investigation of incremental classification ability is not performed at all [12]. Large datasets are required to do this [36]. Other excuses include the lack of information or incomplete information on classic risk factors, poor standardization of classic risk factors, and even a lack of consensus about which classic risk factors are important to consider. Many classic risk factors and tumor markers are supported by unreliable evidence.

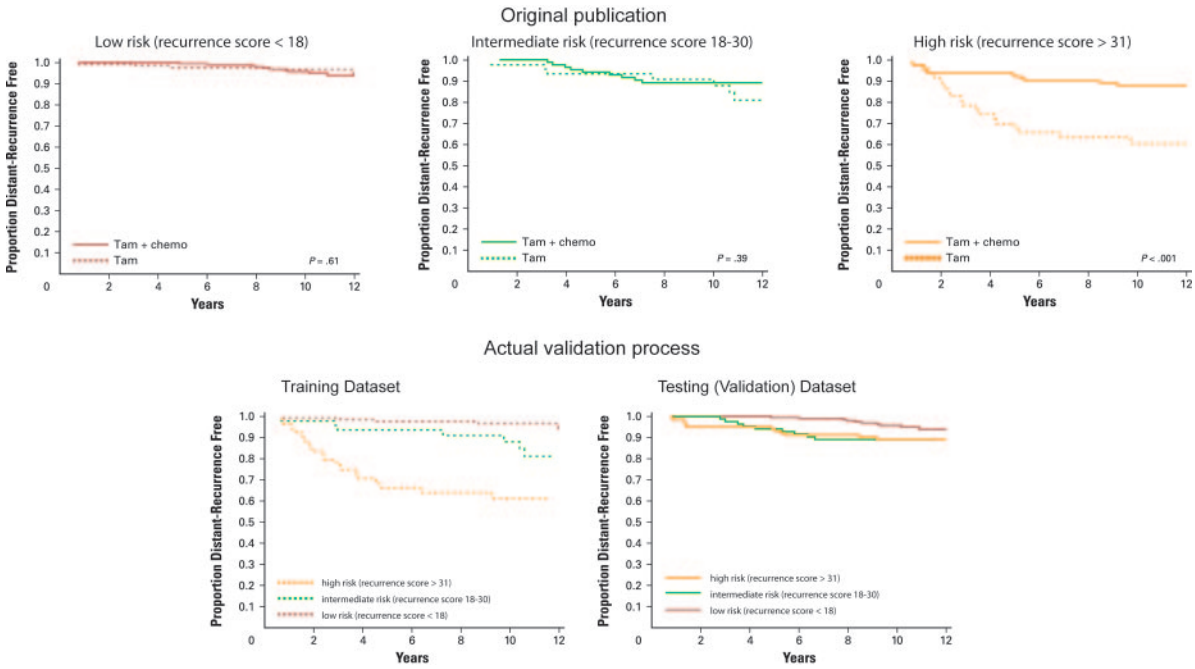

Figure 1. Mixing training and test datasets: a failed validation presented as a successful validation with extension of the clinical indications. The development of the 21-gene recurrence score (Onco*type* DX®) involved training in data from the patients of the tamoxifen arm of the National Surgical Adjuvant Breast and Bowel Project (NSABP)-20 trial. At a subsequent stage, the investigators also tested the Onco*type* DX® classification on the other arm of the same trial, in which the patients had also received chemotherapy in addition to tamoxifen. Then the investigators compared the distant recurrence-free survival in the two arms separately for patients with low recurrence score, intermediate recurrence score, and high recurrence score. The top three panels show these comparisons, as published by the investigators (from Paik S, Tang G, Shak S et al. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. J Clin Oncol 2006;24:3726–3734, reprinted with permission from the American Society of Clinical Oncology). As shown, patients with a high recurrence score have a clear benefit from chemotherapy, while those with a low or intermediate score show no significant difference. The investigators claimed that the 21-gene score not only offers prognostic information but also predicts the treatment effect from chemotherapy, that is, an extended clinical indication. However, the tamoxifen data were the training dataset for the 21-gene recurrence score, so the ability to separate patients at different risks is only too easy to anticipate even if the signature has no true classification accuracy. The true test (validation) is the independent data of the chemotherapy arm. The lower two panels show the same exact data in an appropriate framework. The tamoxifen arm data (training set) are shown on the left side and the chemotherapy arm data (test set) are shown on the right side. As shown, these data suggest, in fact, that the 21-gene recurrence score fails to discriminate the risk of recurrence in the true test dataset (chemotherapy).

Some of the best work on addressing the incremental discriminatory ability of molecular profiles has been done in breast cancer. The 21-gene recurrence score and the 70-gene signature may provide independent information after multivariate adjustment for classic risk factors [29, 34]. Even here, though, the verdict is not final, and other investigators have reached opposite conclusions when gene expression is compared with a strong classic predictive system (e.g., the Nottingham Prognostic Index) or optimized combinations of conventional markers [37, 38].

Molecular profiles may have modest to strong correlation with some classic risk factors, for example, tumor size and grade. Therefore, when both the molecular profile and the classic risk factors are included in multivariate models, coefficients may be affected by collinearity and become unstable. Generally, a variable that carries information from many subvariables (the typical molecular profile) may outperform one that is more narrowly defined, when both are included in the same model. Collinearity may damage more the coefficient of the classic risk factor.

Sometimes the molecular profile coefficient remains formally significant, while other classic risk factors are no longer statistically significant in the multivariate model. However, this is clinically meaningless when we have routinely available classic risk factors. Age, sex, tumor size, and lymph node status are known without any special effort. It makes no sense to say that if one can measure the molecular profile, then we don't need to know the age, sex, tumor size, and lymph node status. The real question is the absolute increment in discrimination offered by a model with the molecular profile plus classic routine risk factors versus a model with classic risk factors alone. This question

is usually not addressed at all. The few presented data suggest that this incremental benefit is small in the best successes to date. For the 70-gene signature for breast cancer, the area under the curve (in the curve showing the interchange between specificity and sensitivity) for metastasis at 5 years improves from 0.659 with the classic risk factors (Adjuvant! Online) to only 0.681 using the molecular profile [29].

Paradoxically, the selection of appropriate study populations for which a molecular profile is most needed may sometimes underestimate the importance of classic risk factors. For example, the dataset of 295 women with breast cancer that was used for the training of several of the breast cancer signatures was chosen in a way that ensured that classic predictors had little to offer [39]: eligibility criteria included a strict cutoff to ensure small tumor size, a strict cutoff for young age (52 years), and no infiltration of apical lymph nodes. Within this narrow range, classic risk factors are already reduced to have little influence, so even a mediocre profile would add incremental information. Would this profile be able to add equally incremental information in a more general population in which the classic predictors carry more information? Once in clinical practice, generalization of the use of tumor markers beyond their original training and validation setting ("transportability") is likely to happen [40]. The broader population in which a predictive tool is eventually applied may not necessarily be plausibly related to what was studied in the training and validation setting.

Nonselective and Transparent Accumulation of Evidence

The wider molecular literature, not only on profiles, probably suffers from considerable selective reporting and publication bias [41, 42]. The recent REporting recommendations for tumor MARKer prognostic studies (REMARK) statement [43] is making an effort to standardize the reporting of tumor prognostic marker studies and hopefully will help to improve transparency in the future. REMARK arose out of the realization that the reporting of prognostic marker studies is often highly deficient. Almost all studies on tumor markers report statistically significant results. Selective reporting strongly favors the dissemination of data and analyses that claim statistical significance. In one example, retrieval of additional fine-print published data and unpublished data from the investigators changed the conclusions of a meta-analysis entirely. Whereas TP53 seemed to be a very strong predictor of survival in head and neck cancer based on readily available published data, this ended up having no effect at all on the risk for death when these additional data were considered [44].

The exact extent of selective reporting biases in gene-expression profiling research is unknown. However, selection biases are unavoidable when the prevailing mentality remains that these studies should remain data-rich exercises on a few subjects [13]. Failure to act on this front may generate unreliable literature [45]. At a minimum, studies with large sample sizes and those that reach close to clinical translation should enjoy full transparency. The poststudy odds of a research finding being true are small when effect sizes are small, when studies are small, when a field is "hot" (many teams working on it), when there is strong interest in the results, when databases are large, and when analyses are more flexible [45]. Molecular profiling research fulfils all these criteria.

Demonstration of Clinical Effect (Efficacy)

Randomized trials have been the typical way to demonstrate clinical efficacy for all interventions. Should molecular profiling be validated for clinical use through clinical trials? Should each proposed molecular profile be tested through clinical trials? Is one clinical trial sufficient or are several trials needed?

Answering these questions is not easy. For example, Onco*type* DX® was approved for use before any results from clinical trials were obtained. Clinical Laboratory Improvement Amendments (CLIA) approval and reimbursement through the Centers for Medicare & Medicaid Services were granted considering only the good performance of retrospective data from available datasets (with the caveats discussed above) and extrapolating that this would also translate into a clinical benefit. Nevertheless, a large clinical trial, the Trial Assigning IndividuaLized Options for Treatment (Rx) (TAILORx), is already also under way to validate the utility of the profile prospectively. Another large trial, the Microarray In Node negative Disease may Avoid ChemoTherapy (MINDACT) trial, has started with the aim to test the utility of the 70-gene signature [15].

Both trials test the utility of the respective molecular signature only in selected strata of patients. In TAILORx, patients are randomized to receive or not receive chemotherapy only if they have an Onco*type* DX® score between 11 and 25. Patients with higher scores are all given chemotherapy and patients with lower scores are not given chemotherapy. All patients receive hormonal therapy. In the MINDACT trial, patients whose prediction is "high risk" according to both the traditional prognostic tool (Adjuvant! Online) and the 70-gene signature are given chemotherapy, patients whose prediction is "low risk" according to both tools are not given chemotherapy, while patients who have discordant predictions (high risk by Adjuvant! Online and

low risk by the 70-gene signature) are randomized to either receive or not receive chemotherapy.

What outcomes should these trials have? Survival is the most important, patient-relevant outcome [46], but some investigators have argued against survival being the key outcome in cancer trials [47], and trials with survival as the primary endpoint require extremely long follow-up. Moreover, molecular profiling may have no impact on survival, but may still affect other patient-relevant outcomes such as avoidance of drug toxicity and improvement in quality of life. More rational use of treatments may also reduce costs, with less chemotherapy being administered and fewer hospitalizations resulting from adverse events. The clinical utility of potential outcomes should be carefully graded [48].

No matter what outcomes are selected, one has to ensure that the design of the trials carries objectivity. For example, blinding may not be feasible given the nature of the intervention, but allocation concealment should be guaranteed. Knowledge of the intervention assignment may affect the interpretation of soft, subjective outcomes and decisions such as quality of life, use of treatment, and hospital admission.

Of the two ongoing trials, the primary outcomes of TAILORx are disease-free survival, distant recurrence-free interval, recurrence-free interval, and overall survival. The trial expects to follow-up patients for up to 20 years. For the MINDACT trial, the primary endpoint is distant metastasis-free survival, but the power calculations are not based on the comparison of the two randomized arms. The trial is powered to reject the null hypothesis that the 5-year distant metastasis-free survival rate is 92% in the discordant-test patients (low risk molecular, high risk clinical prediction), who receive no chemotherapy, if the true rate is 95%. Thus, it aims to prove that molecular prediction can safely be used to spare chemotherapy and thus its toxicity and cost in selected patients.

The number of available profiles may escalate geometrically in the future. Should each new profile be tested with one or more clinical trials? This will make the approval and use of these tests very cumbersome and their development and testing prohibitively costly. In this regard, proof-of-concept trials may be recommended. However, the decision of whether or not a new profile can fit into the already tested concept is difficult. For treatments at least, we have been repeatedly misled into believing that testing one intervention suffices and all others would be "similar" [49]. Certainly most available tests that are already used in everyday clinical practice have not been evaluated for clinical utility in the setting of clinical trials. Yet should this be an argument in favor of continuing this laxity toward diagnostic and predictive evidence as opposed to therapeutic interventions?

Perhaps the answer to these questions should be given on a case-by-case basis. We should consider each time what the prospects are for improving outcomes for a specific disease, stage, and setting. For example, for a disease and stage for which no effective treatment exists, one may argue that molecular markers may be able to identify people who do benefit, despite no average effect. While the concept is attractive, we hardly have any examples in which this approach has worked to date to transform an ineffective treatment into an effective one for a subgroup. More likely the benefits should be sought in individualizing treatments that are already known to have some average benefit [50].

Demonstration of Benefit on Routine Clinical Use (Effectiveness)

Clinical trials may give the opportunity to test the benefit of using molecular profiling in very special settings of patient populations that have the optimal eligibility criteria and are receiving optimal care in highly experienced centers that are thoroughly familiar with the technology. However, even if a benefit is shown under such specialized, knowledgeable conditions, there is no guarantee that it would be possible to extrapolate the benefit into routine clinical care.

Given the funding limitations for conducting clinical research, clinical trials should try to replicate as closely as possible the wide variety of conditions and settings in which a particular profile may be employed. Both TAILORx and the MINDACT trial, the two ongoing clinical trials, have struggled to introduce pragmatic designs [51], as we discussed above. Still, TAILORx is performing a randomized comparison in a stratum of intermediate-risk patients according to the molecular predictor score without any effort being made to compare this with a traditional clinical prediction.

Perhaps more importantly, it is difficult to model in a clinical trial the extent of misuse of a technology that will occur when it is available on the market. By definition, even the most pragmatic trials do the best possible to ensure that the diagnostic or predictive test is used and interpreted appropriately. Simplifying these tests to the maximum possible extent before introduction into clinical use is also critical. Some of the proposed profiles are already simplified. However, others are using more complex classifications [52] that may escape the average oncology specialist. Interestingly, the split of the Onco*type* DX® score into three categories of risk was different in the studies that led to the training and validation of this molecular predictor (<18, 18–30, >30) from the three categories examined in the "pragmatic" TAILORx (<11, 11–25, >25). When there is

such inconsistency even in the research phase, moving the goalpost (perhaps inappropriately) may be even more frequent in clinical applications.

Misuse may take different forms. These include employing a test when it is not likely to be helpful—then the results will lead to misleading reassurance about the test-driven decisions—and not using a test when it could provide useful information. These mishaps go beyond the errors in management that would result from simple misclassification due to inherently imperfect test performance.

Integration in Clinical Care

Some of the early problems encountered during assay development reappear on a grand scale when molecular profiles move into the clinic. Are clinical units able to obtain and process samples and apply the assays in a way that will ensure their classifier ability is maintained? This entails training not only the oncologists, but also all other staff involved in obtaining, processing, and analyzing the needed samples and specimens. mRNA is not a very stable molecule, and differences in the sampling and processing of the specimens could introduce considerable noise. This is worse than the noise introduced by the same factors in the discovery phase, in which highly expert teams are involved.

A series of other challenges follows for which we have no clear answers yet. What are some minimal quality criteria that should exist and how are they to be enforced? How will these tests be introduced in the clinical care routine? Who will order them? Would they require a minimum of specialization and knowledge-based training? Should they be used only by "experts"? If so, what level of expertise is warranted and required? Should the use of these tests be audited? If so, what should be the audit criteria? With the rapidly increasing number of tests, should expertise be continuously re-evaluated and reinforced?

Perhaps the above concerns are exaggerated. Multicenter experience should give us a clearer picture [53–56]. As medicine evolves toward a more patient- rather than physician-centered decision-making environment, patients may be prime motivating forces advocating the use of these tests. The prospect of individualized medicine may empower people to use technology directly. However, perhaps use of these complex technologies by patients without expert input may complicate or even worsen care and outcomes.

Cost-Effectiveness

Cost-effectiveness depends primarily on the cost of the tests, the cost of the harms they can abort, the cost of the benefits they can help procure, and the relative costs of other interventions that form the management of cancer patients. Molecular profiling may induce or abort other interventions. Until now, treatment has been a much larger part of the cost of medical oncology than diagnosis; prognostic or predictive testing has accounted for a negligible component of the overall cost for the care of patients with cancer. This situation may change. The major potential of these tests is that by modulating therapeutic decisions and therapeutic toxicity they may affect the cost of overall care. However, these benefits need to be documented and replicated in clinical trials.

Until then, what is a suitable price for a molecular profile? A formal decision and cost-effectiveness analysis is premature in the absence of estimates of the true benefits of this technology. The current price of molecular profiles moving into the market is a few thousand dollars per test. This is expensive compared with most tests used in modern medicine. However, in the absence of the complete picture, it is impossible to tell whether this is worth it or not.

The Future Is Now

Molecular profiling is rapidly evolving from an interesting scientific concept to a clinical tool. The state of evolution in this process varies for different cancers. Overall, we have some promising early experience and a possibly fascinating future, but also several caveats to deal with. Caveats should be addressed on a case-by-case basis for each cancer and each clinical scenario. However, there are some common themes that probably apply to all situations in which this technology may help patients in the future. Attention to these themes may help improve prognostic marker research not only for current efforts in molecular profiling, but also for both traditional tumor markers and the future markers of the postgenomic era.

First, we need full transparency of the experimentation and results obtained in this field, and every effort should be made to avoid selective reporting. Second, we need larger datasets for all stages of development of these assays, including training, validation, proof of concept, and proof of clinical merit. Third, there is no excuse for anything less than full, extensive independent validation under completely independent conditions for any molecular profile that wants to be even a candidate for clinical use. Fourth, evaluation of classifier accuracy should consider all available classic risk factors, and this should be done in settings and with datasets for which classic risk factors are given a fair chance to show their merit. Replacing routine, free information with costly, convoluted biological signatures makes no sense. Fifth, although there is a debate about how many clinical trials we need, at the moment we have none

that have been completed and only a couple that are ongoing with quite novel designs. Therefore, we are certainly far from the point of saying that we have had enough randomized evidence. Trials should be designed carefully to minimize the potential for biased results and they should have a pragmatic outlook. Finally, we should keep thinking about how to best employ these tests when they emerge for widespread use into clinical practice, some time soon.

Disclosure of Potential Conflicts of Interest

The author indicates no potential conflicts of interest.

References

1 Lockhart DJ, Winzeler EA. Genomics, gene expression and DNA arrays. Nature 2000;405:827–836.

2 Alizadeh AA, Ross DT, Perou CM et al. Towards a novel classification of human malignancies based on gene expression patterns. J Pathol 2001;195:41–52.

3 Tefferi A, Bolander ME, Ansell SM et al. Primer on medical genomics. Part III: Microarray experiments and data analysis. Mayo Clin Proc 2002;77:927–940.

4 Schena M. Microarray Analysis. New York: Wiley-Liss, 2003:1–630.

5 Patsopoulos NA, Ioannidis JP, Analatos AA. Origin and funding of the most frequently cited papers in medicine: Database analysis. BMJ 2006;332:1061–1064.

6 Dalton WS, Friend SH. Cancer biomarkers–an invitation to the table. Science 2006;312:1165–1168.

7 Gutman S, Kessler LG. The US Food and Drug Administration perspective on cancer biomarker development. Nat Rev Cancer 2006;6:565–571.

8 Henry NL, Hayes DF. Uses and abuses of tumor markers in the diagnosis, monitoring, and treatment of primary and metastatic breast cancer. ***The Oncologist*** 2006;11:541–552.

9 Ludwig JA, Weinstein JN. Biomarkers in cancer staging, prognosis and treatment selection. Nat Rev Cancer 2005;5:845–856.

10 Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000;19:453–473.

11 Vergouwe Y, Steyerberg EW, Eijkemans MJ et al. Validity of prognostic models: When is a model clinically useful? Semin Urol Oncol 2002;20:96–107.

12 Ntzani EE, Ioannidis JP. Predictive ability of DNA microarrays for cancer outcomes and correlates: An empirical assessment. Lancet 2003;362:1439–1444.

13 Liu ET, Karuturi KR. Microarrays and clinical investigations. N Engl J Med 2004;350:1595–1597.

14 Pusztai L, Mazouni C, Anderson K et al. Molecular classification of breast cancer: Limitations and potential. ***The Oncologist*** 2006;11:868–877.

15 Abdullah-Sayani A, Bueno-de-Mesquita JM, van de Vijver MJ. Technology Insight: Tuning into the genetic orchestra using microarrays—limitations of DNA microarrays in clinical practice. Nat Clin Pract Oncol 2006;3:501–516.

16 Brazma A, Hingamp P, Quackenbush J et al. Minimum information about a microarray experiment (MIAME)–toward standards for microarray data. Nat Genet 2001;29:365–371.

17 Allison DB, Cui X, Page GP et al. Microarray data analysis: From disarray to consolidation and consensus. Nat Rev Genet 2006;7:55–65.

18 Simon R, Radmacher MD, Dobbin K. Design of studies using DNA microarrays. Genet Epidemiol 2002;23:21–36.

19 Simon R, Radmacher MD, Dobbin K et al. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 2003;95:14–18.

20 Ransohoff DF. Bias as a threat to the validity of cancer molecular-marker research. Nat Rev Cancer 2005;5:142–149.

21 Paik S, Shak S, Tang G et al. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. N Engl J Med 2004;351:2817–2826.

22 van't Veer LJ, Dai H, van de Vijver MJ et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530–536.

23 Wang Y, Klijn JG, Zhang Y et al. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. Lancet 2005;365:671–679.

24 Chang HY, Nuyten DS, Sneddon JB et al. Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. Proc Natl Acad Sci U S A 2005;102:3738–3743.

25 Ebert BL, Golub TR. Genomic approaches to hematologic malignancies. Blood 2004;104:923–932.

26 Lossos IS Czerwinski DK, Alizadeh AA et al. Prediction of survival in diffuse large-B-cell lymphoma based on the expression of six genes. N Engl J Med 2004;350:1828–1837.

27 Rosenwald A, Wright G, Chan WC et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. N Engl J Med 2002;346:1937–1947.

28 Fan C, Oh DS, Wessels L et al. Concordance among gene-expression-based predictors for breast cancer. N Engl J Med 2006;355:560–569.

29 Buyse M, Loi S, van't Veer L et al. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. J Natl Cancer Inst 2006;98:1183–1192.

30 Ioannidis JP. Microarrays and molecular research: Noise discovery? Lancet 2005;365:454–455.

31 Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: A multiple random validation strategy. Lancet 2005;365:488–492.

32 Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. Proc Natl Acad Sci U S A 2006;103:5923–5928.

33 Ein-Dor L, Kela I, Getz G et al. Outcome signature genes in breast cancer: Is there a unique set? Bioinformatics 2005;21:171–178.

34 Paik S, Tang G, Shak S et al. Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. J Clin Oncol 2006;24:3726–3734.

35 Ioannidis JPA. Gene expression profiling for individualizing breast cancer chemotherapy: success or not? Nat Clin Pract Oncol 2006;3:538–539.

36 Vergouwe Y, Steyerberg EW, Eijkemans MJ et al. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol 2005;58:475–483.

37 Eden P, Ritz C, Rose C et al. "Good Old" clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers. Eur J Cancer 2004;40:1837–1841.

38 Nimeus-Malmstrom E, Ritz C, Eden P et al. Gene expression profilers and conventional clinical markers to predict distant recurrences for premenopausal breast cancer patients after adjuvant chemotherapy. Eur J Cancer 2006;42:2729–2737.

39 van de Vijver MJ, He YD, van't Veer LJ et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 2002;347:1999–2009.

40 Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information, Ann Intern Med 1999;130:515–524.

41 Easterbrook PJ, Berlin JA, Gopalan R et al. Publication bias in clinical research. Lancet 1991;337:867–872.

42 Pan Z, Trikalinos TA, Kavvoura FK et al. Local literature bias in genetic epidemiology: An empirical evaluation of the Chinese literature. PLoS Med 2005;2:e334.

43 McShane LM, Altman DG, Sauerbrei W et al. Reporting recommendations for tumor marker prognostic studies (REMARK). J Natl Cancer Inst 2005; 97:1180–1184.

44 Kyzas PA, Loizou KT, Ioannidis JP. Selective reporting biases in cancer prognostic factor studies. J Natl Cancer Inst 2005;97:1043–1055.

45 Ioannidis JP. Why most published research findings are false. PLoS Med 2005;2:e124.

46 Ioannidis JP, Pavlidis N. Levels of absolute survival benefit for systemic therapies of advanced cancer: A call for standards. Eur J Cancer 2003;39: 1194–1198.

47 Di Leo A, Bleiberg H, Buyse M. Overall survival is not a realistic end point for clinical trials of new drugs in advanced solid tumors: A critical assessment based on recently reported phase III trials in colorectal and breast cancer. J Clin Oncol 2003;21:2045–2047.

48 Hayes DF, Bast RC, Desch CE et al. Tumor marker utility grading system: A framework to evaluate clinical utility of tumor markers. J Natl Cancer Inst 1996;88:1456–1466.

49 McAlister FA, Laupacis A, Wells GA et al. Users' Guides to the Medical Literature: XIX. Applying clinical trial results B. Guidelines for determining whether a drug is exerting (more than) a class effect. JAMA 1999;282: 1371–1377.

50 Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: Importance, indications, and interpretation. Lancet 2005; 365:176–186.

51 Bogaerts J, Cardoso F, Buyse M et al. Gene signature evaluation as a prognostic tool: Challenges in the design of the MINDACT trial. Nat Clin Pract Oncol 2006;3:540–551.

52 Perou CM, Sorlie T, Eisen MB et al. Molecular portraits of human breast tumours. Nature 2000;406:747–752.

53 Foekens JA, Atkins D, Zhang Y et al. Multicenter validation of a gene expression-based prognostic signature in lymph node-negative primary breast cancer. J Clin Oncol 2006;24:1665–1671.

54 Staal FJ, Cario G, Cazzaniga G et al. Consensus guidelines for microarray gene expression analyses in leukemia from three European leukemia networks. Leukemia 2006;20:1385–1392.

55 Zu Y, Steinberg SM, Campo E et al. Validation of tissue microarray immunohistochemistry staining and interpretation in diffuse large B-cell lymphoma. Leuk Lymphoma 2005;46:693–701.

56 Dobbin KK, Beer DG, Meyerson M et al. Interlaboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays. Clin Cancer Res 2005;11:565–572.

Harvey J. Kupferberg, Ph.D.

Kangamangus Gorge, New Hampshire

Is Molecular Profiling Ready for Use in Clinical Decision Making?

John P. A. Ioannidis

Oncologist 2007;12;301-311
DOI: 10.1634/theoncologist.12-3-301