Tumor Markers

FROM LABORATORY TO CLINICAL UTILITY*

Anne-Sofie Schrohl‡, Mads Holten-Andersen‡, Fred Sweep§, Manfred Schmitt¶, Nadia Harbeck¶, John Foekens‖, and Nils Brünner‡, on behalf of the European Organisation for Research and Treatment of Cancer (EORTC) Receptor and Biomarker Group**

A very broad definition of a tumor marker is: a tool that enables the clinician to answer clinically relevant questions regarding a cancer disease (1). However, most researchers in this field would probably prefer the following more specific definition of a tumor marker: a molecule, a process, or a substance that is altered quantitatively or qualitatively in precancerous or cancerous conditions, the alteration being detectable by an assay (2). Alterations can be produced either by the tumor itself or by the surrounding normal tissue as a response to tumor cells (2). Regardless of which definition is preferred, the tumor marker itself can be DNA, mRNA, protein, or processes (apoptosis, angiogenesis, proliferation, etc.) measured quantitatively or qualitatively by an appropriate assay. In addition, the types of specimen in which the tumor marker is detected can be different; tissue, blood (plasma/serum), saliva, urine, etc. are all used. The tumor marker assays can be of very different formats ranging from complex animal models to immunohistochemical test kits. The most commonly used format is probably the immunoassay, which is a well-characterized methodology. However, this field is progressing rapidly, and new and advanced assays such as microarrays and mass spectrometry are becoming established technologies in tumor marker research.

The first known tumor marker was described in 1846, when Henry Bence-Jones reported the precipitation of a protein in acidified urine from patients with multiple myeloma. Detection of the monoclonal immunoglobulin light chain in this disease is still in use, and since then numerous potential tumor markers have been reported on in the literature (1). Examples of such markers in clinical use are: alpha-fetoprotein for tumors of the liver, testis, and other germ cell line tumors, CA125 for ovarian cancer, prostate specific antigen (PSA)[1] for prostate cancer, and steroid hormone receptors (estrogen and progesterone receptor) used in management of breast cancer. However, as the field of tumor markers has expanded rapidly over the last two decades with a concomitant increase in published reports, it has become increasingly apparent that a strong need exists for establishment of consensus guidelines for development and use of tumor markers. Such guidelines should be internationally accepted if any of these potential new markers are ever to reach a stage where they will benefit the patients. The guidelines should define the potential specific clinical uses of tumor markers, define specific requirements for the technical development of tumor marker assays, and state specific requirements that are to be fulfilled before clinical implementation of a tumor marker. Suggestions for such guidelines have been made; in 1996, a tumor marker expert panel convened by the American Society of Clinical Oncology proposed a framework to be used for evaluation of tumor marker studies: the tumor marker utility grading system (TMUGS), which also includes a framework for rating published evidence (2). The TMUGS framework is further discussed in "Clinical Testing." However, work in this field is still ongoing, and some important aspects to consider in the process of designing such guidelines will be covered in this review.

The possible clinical uses of tumor markers are manifold, and several categories of markers can be defined. A **diagnostic tumor marker** is a marker that will aid in detection of malignant disease in an individual. Preferably, the marker should be tissue specific and not be influenced by benign diseases of the particular tissue/organ. Thus, a diagnostic marker should exhibit both high levels of diagnostic sensitivity and specificity (see below) to be of clinical value, especially if the marker is to be used for (mass) screening purposes. A fundamental prerequisite for development of any diagnostic (screening) tumor marker lies in the nature of the disease

From the ‡Department of Pharmacology and Pathobiology, Royal Veterinary and Agricultural University, Ridebanevej 9, DK-1870 Frederiksberg C, Copenhagen, Denmark, §Department of Chemical Endocrinology, University Medical Centre Nijmegen, P.O. Box 9101, Geert Groteplein 10, NL-6500 HB Nijmegen, The Netherlands, ¶Clinical Research Unit, Department of Obstetrics & Gynaecology, Technical University of Munich, Ismaninger Strasse 22, D-81675 München, Germany, and ‖Rotterdam Cancer Institute (Daniel der Hoed Klinik), Josephine Nefkens Building, Nr. BE 426, Dr. Molewaterplein 50, NL-3015 GE Rotterdam, The Netherlands

Received, June 9, 2003

Published, MCP Papers in Press, June 17, 2003, DOI 10.1074/mcp.R300006-MCP200

[1] The abbreviations used are: PSA, prostate specific antigen; TMUGS, tumor marker utility grading system; CEA, carcino-embryonic antigen; uPA, urokinase-type plasminogen activator; PAI-1, plasminogen activator inhibitor type-1; EORTC, European Organisation for Research and Treatment of Cancer; RBG, Receptor and Biomarker Group; ER, estrogen receptor; PgR, progesterone receptor; HCG, human chorionic gonadotropin; QC, quality control; LOE, level of evidence.

because this must be relatively prevalent, the biology of the disease must be known, and, naturally, an effective treatment should be at hand (3). Measurement of a **prognostic marker** gives the clinician a tool for estimating the risk of disease recurrence and/or cancer-related death for an individual patient following the initial surgical removal of the cancer but without administration of adjuvant therapy. In contrast, a **predictive tumor marker** will foretell how the patient is going to respond to a given therapy. Tumor markers for detection of recurrence or remission are classified as **monitoring markers** and are used during follow-up of patients who do or do not receive anticancer therapy. Finally, a new and potentially important area includes the use of tumor markers for localizing tumors and for targeting of cytotoxic agents (1).

Thus, the uses of tumor markers are numerous. However, regardless of the type of tumor marker, in order for a marker to be considered for routine implementation it needs to be demonstrated that measurement of the marker ultimately impacts on clinical management of the malignant disease either by improving patient outcome or quality of life or by lowering costs of care (2).

SCREENING MARKERS

Screening markers belong to the subclass of diagnostic markers. The major issue when developing new markers for screening of populations for presence of cancer is the specificity and sensitivity of the marker with regard to the diagnosis. Specificity is defined as the proportion of negatives that are correctly identified by the test. By subtracting the percentage of correctly identified negatives ("true negatives") from 100, the percentage of false-positive test results can be calculated. Sensitivity is the proportion of positives that are correctly identified by the test. Positive predictive value is the proportion of patients with positive test results who are correctly diagnosed, and, accordingly, it is calculated by dividing the number of "true positives" by the total number of positives (true and false). Similarly, the negative predictive value is the proportion of patients with negative test results who are correctly diagnosed and it is calculated as the number of "true negatives" divided by the total number of negatives (true and false). In order for a screening test to find acceptance in the clinical setting, both specificity and sensitivity must be high. However, depending on the cancer disease screened for, the demands for specificity and sensitivity can vary. For example, screening for colon cancer demands a high specificity, because all individuals with a positive test result should subsequently undergo colonoscopy, a time-consuming and therefore expensive procedure. On the other hand, a breast cancer screening test could be acceptable even with a low specificity as long as its sensitivity is high because such a test could be regarded as a premammography test, and all individuals testing positive could subsequently be offered mammography, a noninvasive and easy procedure. Such a test would significantly reduce the number of individuals who participate in mammography screening (or delay intervals between mammographies), and at the same time the high sensitivity of the breast cancer test would secure that almost all of the diseased individuals would be referred to mammography and further work-up.

The desired or required specificity is also dependent on the prevalence of a particular cancer disease in the study population. For example, by studying a group of individuals with a particularly high risk of disease, *e.g.* due to family history, an increase in positive predictive value is gained, because a number of the false-positive individuals will be among the nonstudied population. Screening for colon cancer is a good example: If the prevalence of the disease is considered 1:1000 in a population of 50- to 80-year-old individuals, and the test has a specificity of 95%, on average 50 colonoscopies have to be performed in order to detect one colon cancer patient. By preselecting patients at high risk for colon cancer, *e.g.* one or more family members with prior colon cancer, patients with prior colon adenoma, or patients with inflammatory bowel diseases, the prevalence of colon cancer increases to for example 1:100. If the prevalence increases ten times, one only has to perform one-tenth the number of colonoscopies in order to find one colon cancer. In the present example this means 5 colonoscopies. It should be emphasized that selecting a high-risk population for screening has no impact on the sensitivity of the test. The sensitivity could be increased by, for instance, including an additional screening marker in the test, which is independent of the first marker and thus provides additional information.

A receiver operating curve can graphically illustrate the relationship between specificity and sensitivity. The ideal curve is the one giving the maximum possible area under the curve. Fig. 1 shows a receiver operating curve from one of our studies of plasma concentrations of tissue inhibitor of metalloproteinases-1 (TIMP-1) in healthy individuals and colon cancer patients (4). 1-Specificity is depicted on the *x* axis and sensitivity shown on the *y* axis. We added information from measurements of serum concentrations of another protein, carcino-embryonic antigen (CEA) and were able to show that this added significantly to the area under the TIMP-1 curve.

There are only two approved protein-based cancer screening markers: PSA, which is measured in blood when screening for prostate cancer, and hemoglobin measured in feces (fecal occult blood test) when screening for colon cancer. While PSA has a high sensitivity, specificity is rather low. This has been found acceptable, because subsequent biopsy of the prostate is considered a simple procedure. However, many attempts are being put into discovery of proteins that can be used in combination with PSA in order to increase specificity. The more recent commercially available fecal occult blood tests have very high specificity but only medium sensitivity.

Lack of patients' compliance is a major problem in cancer screening. In order to assure good patient compliance, a screening test should be as noninvasive as possible, and

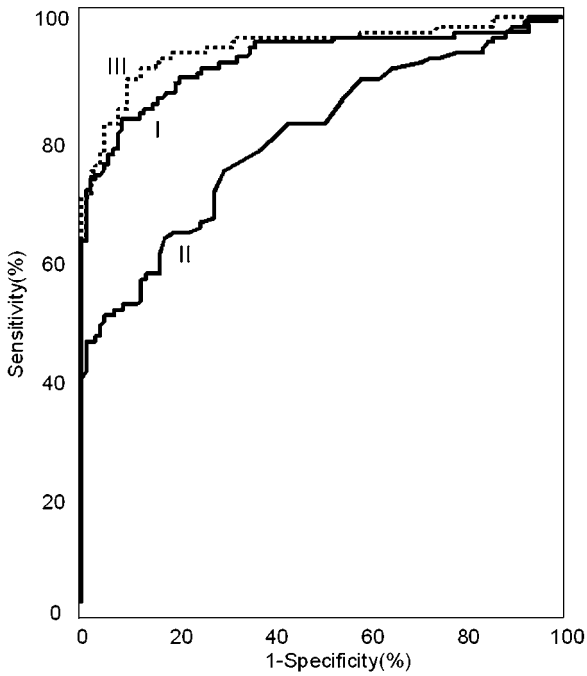

FIG. 1. **Receiver operating curves (colon cancer, right sided, all Dukes' stages) for total plasma TIMP-1 (*I*), serum CEA (*II*), and TIMP-1/CEA combined (*III*).** Area under the curve for each curve: *I*, 93%; *II*, 80%; *III*, 95%. Reproduced from Ref. 4 with copyright permission from *Clinical Cancer Research*.

ideally the assay should enable testing either by the patient himself or by the family physician.

Another crucial point that has to be considered for any screening method is that it should primarily detect early stage disease where the chance of cure by subsequent standard treatment is high.

As for any marker study, sufficient evidence for the clinical use of the marker in question has to be provided. For screening markers, this will include randomized controlled studies as well as systematic reviews or meta-analyses of published and hitherto unpublished data.

PROGNOSTIC MARKERS

Prognostic markers give information about patient outcome and thus about the aggressiveness of the disease. Usually, prognostic markers are determined at the time of primary therapy in order to predict the future course of the disease in an individual patient.

A good example of a prognostic marker, which was thoroughly developed according to the criteria initially suggested by McGuire *et al.* (5) (Table I) and which is now being used in clinical practice, is the urokinase-type plasminogen activation system with its components urokinase-type plasminogen activator (uPA) and plasminogen-activator inhibitor type-1 (PAI-1) in breast cancer. Invasion factors uPA and its inhibitor PAI-1 are the first novel tumor biological prognostic factors that have been validated at the highest level of evidence (to be discussed later) with regard to their clinical utility in breast cancer.

TABLE I
Criteria for evaluation of prognostic factors (modified from Ref. 5)

1.	Biological hypothesis
2.	Simple, standardized, and reproducible determination method
3.	Biostatistical planning of data analysis
4.	Correlation to established factors
5.	Optimized cut-off values for distinction between low- and high-risk group
6.	Univariate and multivariate analysis (independence of factors/weighing)
7.	Validation (different patient collective/different research group)
8.	Clinical study, impact on therapy
9.	Transfer into clinical practice

1. Biological Hypothesis—Abundant experimental evidence demonstrates a key role of uPA and PAI-1 in tumor invasion and metastasis; they are involved in focal proteolysis, adhesion, and migration (6).

2. Standardized Method—Antigen levels of both factors are determined by standardized, quality assured enzyme-linked immunosorbent assays in extracts of primary tumor tissue (7).

3–6. Demonstration of Clinical Impact—Numerous international studies have shown that patients with low levels of uPA and PAI-1 have a significantly better survival rate than patients with high levels of either factor (reviewed in Ref. 8).

7–8. Validation and Therapy Trial—Recently, these data have been validated by a multicenter prospective, randomized therapy trial in node-negative breast cancer ("Chemo N_0") (9) and a European Organization for Research and Treatment of Cancer (EORTC) Receptor and Biomarker Group (RBG) pooled analysis comprising more than 8300 breast cancer patients (10).

The particular combination of both factors, uPA/PAI-1 (both low *versus* either or both factors high), outperforms the single factors as well as other, established prognostic factors with regard to risk group assessment, particularly in node-negative breast cancer.

9. Transfer to Clinical Practice—Node-negative breast cancer patients with low levels of uPA and PAI-1 have a very good prognosis and may thus be candidates for being spared the burden of adjuvant chemotherapy. In contrast, node-negative patients with high uPA/PAI-1 are at substantially increased risk of relapse, comparable to that of patients with three or more involved axillary lymph nodes. First results from the Chemo N_0 trial indicated that these high-risk patients benefit from adjuvant chemotherapy. The prospective results have recently been substantiated by a large retrospective analysis indicating that breast cancer patients with high uPA/PAI-1 may derive particular benefit from adjuvant chemotherapy (11). A new therapy trial (NNBC-3, node-negative breast carcinoma) will now investigate the optimal chemotherapy for high-risk node-negative patients with high uPA/PAI-1. Last but not least, uPA and PAI-1 are promising targets for tumor biological therapy, and novel therapeutic approaches are cur-

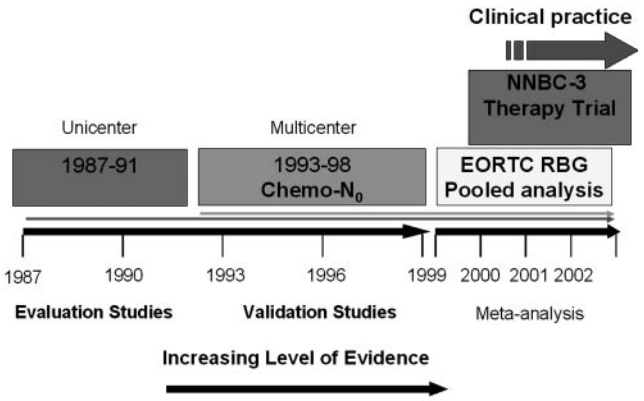

FIG. 2. **Time course for validation of the clinical utility of uPA and PAI-1 in primary breast cancer.** See text (part regarding "Prognostic Markers") for details.

rently being evaluated in preclinical models and early phase clinical trials (12).

Next to uPA/PAI-1, such consistent evaluation of clinical utility for a prognostic factor in breast cancer has only been demonstrated for the proliferation marker thymidine labeling index in Italy (13–15). As discussed above for uPA/PAI-1, thorough evaluation of a prognostic marker may be a long process comparable time-wise to that of developing a novel therapeutic agent (Fig. 2).

A couple of caveats need to be considered before drawing conclusions about potential clinical utility of prognostic markers. First, prognostic markers are best evaluated in patients who did not receive any systemic therapy after standard loco-regional treatment because the course of a disease can be altered quite substantially by systemic therapy. However, if only patients with systemic treatment are available for marker evaluation, a homogeneous collective with regard to therapy is preferable for assessment of a prognostic marker. For example, high levels of a given marker may be strongly associated with tumor aggressiveness. In a cohort of untreated patients, this association will translate to a significant difference in the survival estimates. Yet, in a cohort of patients with systemic therapy, this impact on survival may be lost if patients with high marker levels respond significantly better to the administered therapy than patients with low marker levels. The prognosis, *i.e.* the course of the disease, of the high-risk patients would then be favorably influenced by the administered therapy, and the survival difference seen in the untreated patients might be obscured by the treatment effect. Thus, in the graphical representation, the two Kaplan-Meier survival curves may be close together. Of course, similar considerations need to be taken into account when calculating optimized cut-off levels. Again, uPA/PAI-1 is an example of this impact of therapy on patient outcome: The substantial prognostic impact of uPA/PAI-1 in untreated breast cancer patients is lost in patients who did receive adjuvant systemic therapy (16). Thus, in particular when analyzing markers in archived material, detailed knowledge about administered therapy that may alter the course of disease is a prerequisite for analysis of prognostic markers in clinical samples.

Second, prognostic markers that are significantly correlated to a particular course of the disease may not necessarily be clinically useful. The clinical utility of prognostic, and other, markers depends on the fact whether the result of marker determination alters clinical decision making, *i.e.* whether additional diagnostic tests or treatments will be advised or not or whether closer or less frequent follow-up will be recommended. For example, a marker may significantly split the patient cohort into two groups, yet if the percentage of patients within the high- or low-risk group is too small or too large in order to warrant any clinical consequences, the marker may be clinically useless. Similarly, if patient prognosis in the high- or low-risk group is too poor in order to forgo any additional therapy, the marker may also not be suitable for clinical application. Thus, successful validation of prognostic markers for clinical utility requires an interdisciplinary effort of basic researchers, clinicians, pathologists, and biostatisticians.

PREDICTIVE MARKERS

In the design of studies on predictive markers, it is important to understand the differences between a predictive and a prognostic marker. A predictive marker predicts response or resistance to a specific therapy, whereas a prognostic marker, as described above, predicts relapse or progression independently of future treatment effects (2, 17). Many markers may have both a prognostic and a predictive value. In breast cancer, the most widely used, and studied, cell biological predictive markers are steroid hormone receptors. Estrogen receptor (ER) and progesterone receptor (PgR) are determined in order to predict response to endocrine therapy. ER is a direct target for hormonal agents such as the anti-estrogen tamoxifen, and PgR is a target for anti-progestins. In the adjuvant setting, tamoxifen is effective in preventing breast cancer recurrences in ER-positive patients (18–20), and the benefit to endocrine therapy shows a positive relationship with the level of ER in the primary tumor (19). Similarly, in the advanced setting, clinical benefit is more pronounced in patients whose primary tumors display higher levels of ER and PgR (21), see Fig. 3.

One needs to keep in mind that there are severe limitations with regard to retrospective studies on the predictive impact of tumor-associated cell biological factors. Often, the predictive value of a factor is studied for its relationship with the efficacy of treatment that was given in the adjuvant setting. The end point in these studies is not response-to-therapy, but the occurrence of a relapse in patients who were treated with an intention to cure, *i.e.* to eliminate occult micro-metastases at time of primary surgery by systemic treatment. However, with currently available systemic (combination) treatments, cure will only be achieved in a certain percentage of patients.

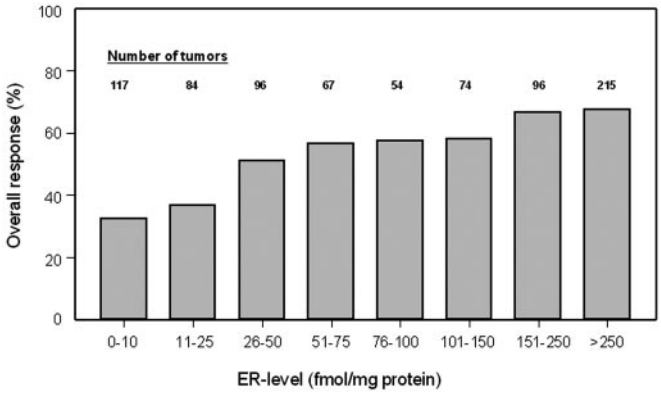

FIG. 3. **Overall response to endocrine therapy in patients with metastatic breast cancer, grouped according to the ER level of the primary tumor.**

The actual treatment benefit as a result of a certain predictive factor is therefore rather difficult to interpret, even more so if tumor bank samples were selected from patients who did not exclusively participate in randomized adjuvant trials. Moreover, because patients outside clinical trials were treated based on the preference of the patient or the physician, or according to guidelines in force at that time, confounding factors are introduced in retrospective analyses of the relationship between a marker and the efficacy of systemic treatment. Furthermore, different adjuvant treatment policies are employed in different centers. Therefore, it is difficult to come to definite conclusions on whether a marker is a pure predictive marker, a prognostic marker, or both when retrospectively analyzing the predictive value of a cell biological factor. Ideally, predictive information in the adjuvant setting should be obtained from a prospective study either designed as a marker study or having the predictive impact as a secondary objective in a therapy trial. However, for ethical reasons, new studies in most cancer types cannot include control groups of patients without adjuvant systemic therapy anymore. Nonetheless, as argued before (11), "for factors that do not strongly correlate with treatment decisions, the problem of confounding can be reduced by various methods, in particular by appropriate use of multivariate analyses and stratification." Thus, introduction of a proper multivariate statistical scoring model to minimize the effects of confounding factors when analyzing predictive markers in the adjuvant setting, *e.g.* in patients with primary breast cancer, is imperative (11). Taking the proper precautions into consideration, the valuable information that is stored in large retrospective data sets and tumor banks may be used for addressing the predictive value of cell biological factors, such as we have shown for uPA and PAI-1 (11).

A more direct way to study the predictive value of a tumor-associated factor is analysis of response-to-therapy in patients with locally advanced or metastatic cancer. In these studies, the end point is the directly measurable effect of treatment on the size of the recurrence(s) or the development of new relapses. This is different from studies performed in the adjuvant setting in which potentially cured cancer patients are evaluated with relapse occurrence as the primary end point, which, as discussed above, in nonrandomized studies not necessarily reflects an association between a marker and efficacy of adjuvant treatment. Furthermore, in contrast to studies performed in the recurrent setting, the type of response (as complete remission, partial remission, stable disease, progressive disease) cannot be studied in adjuvant studies because, by definition, there is no measurable tumor mass and there may be only nonmeasurable occult micrometastases present. Advantages of studying the predictive value of a marker in the metastatic setting are that the follow-up period needed for response assessment may only be less than 1 year and that, because of the yes or no answer, the patient number may be relatively low. Furthermore, tumor samples of nonrandomized patient cohorts can be studied retrospectively because the treatment effect is directly measurable by the size of the recurrence or occurrence of new metastases. In contrast, adjuvant studies need to be randomized, require accrual of hundreds of patients, and require many years of follow-up to give an answer with respect to the predictive value of the marker studied. However, there is one main disadvantage of measuring cell biological factors in the primary tumor, and then later correlating their expression levels to the type of response to systemic therapy that is given for a recurrence up to 10 years after primary therapy. The recurrent tumor, which is used for response assessment, may be phenotypically or genetically different from the primary tumor in which the markers were originally analyzed. Thus, this approach only works for markers that are not known to change their expression between primary lesions and metastases, even while withstanding adjuvant systemic therapy. Moreover, as stated before for the adjuvant setting, proper statistical methods need to be used in order to adjust for confounding factors such as different types of adjuvant therapy, etc.

Nevertheless, for the past 20 years, presence of ER and PgR in primary breast tumors has been the major guide for physicians to treat recurrent breast cancer patients successfully with endocrine therapy. However, the steroid hormone receptor status of the primary tumor does not fully predict which patients will fail or respond to endocrine therapy (Fig. 3). This also holds true with regard to adjuvant endocrine treatment of ER-positive primary breast cancer patients, showing benefit in only 20–25% of patients. Therefore, many potential cell biological predictive markers, which may also serve as potential targets for future therapeutic intervention, have been studied (22). The availability of currently employed high-throughput technologies to simultaneously assess the expression levels of tens of thousands of genes will hopefully lead to identification of new and powerful predictive markers, and possibly to the development of new therapeutic strategies.

MONITORING MARKERS

Markers for monitoring can be useful in a number of clinical settings. First, valid markers may be used for monitoring efficacy of or response to a given therapy. That is, a change in disease status during treatment should be reflected by a change in the tumor marker status. Second, monitoring of markers can be employed in the process of follow-up after administration of primary therapy with the goal of monitoring onset and extent of recurrent disease.

Measuring tumor markers for evaluation of treatment effects is often used as a surrogate end point for treatment efficacy. Such markers are of obvious value, because patients who do not respond to the applied treatment could at a very early time point either be shifted to another treatment or treatment could be stopped and thereby the patient would escape potential toxic side effects of the treatment. In germ cell line tumors, alpha-fetoprotein and human chorionic gonadotropin-beta are both being used to follow treatment efficacy.

In patients with gastrointestinal cancers, CEA and CA19.9 are used to follow the patients after primary therapy in order to detect disease recurrence at an early stage. Similarly, in patients with ovarian cancer, CA125 is a frequently used serum marker for monitoring of recurrent disease. Besides being used for monitoring of treatment efficacy as mentioned above, alpha-fetoprotein and human chorionic gonadotropin-beta are also used to monitor patients with germ cell line tumors in order to detect recurrent disease. It should be emphasized, though, that in order to have clinical value, it has to be demonstrated that early detection of recurrent disease has an impact on patient survival. For example, in breast cancer, no survival benefit has been linked to any tumor marker measurement in regular follow-up after primary therapy. Thus, tumor markers have disappeared form the respective guidelines; yet, they are useful monitoring tools during systemic therapy in those breast cancer patients experiencing recurrence of disease.

QUALITY ASSURANCE OF BIOMARKERS

Different Test Procedures May Yield Different Assay Results—Determination of biological markers in tumor tissue is becoming increasingly important, and the repertoire of potential markers is increasing steadily as is the variety of methods used for their measurements. Obviously, one cannot necessarily assume that one method for the assessment of a marker present in blood or tissue will provide the same results as another method. Moreover, assay results are often quite heterogeneous, depending on the composition of the specimen, way of tissue processing, design and specificity of an assay, type of antibodies used in immunometric assays, and, as important, statistical methods used for data evaluation. The inherent problem of standardizing immunometric assays is that different test kits may generate different test results. This is due to differences in specificity and/or affinity of antibodies used in enzyme-linked immunosorbent assay, immunohistochemistry, or Western blot and the use of different standards provided with the kits. Also, biomarkers may occur in different molecular forms, and such variant molecular forms may be present in different types of cancer. This endogenous variation will even be greater when different tissue extraction methods are employed, *e.g.* by including or excluding nonionic detergents in the extraction buffer (23). Furthermore, the clinical significance of a marker present in the cell membrane fraction may be different from that of the marker present in the cytosol fraction.

Standardization of Laboratory Tests—Although immunometric methods are extensively used in clinical research settings, standardization and quality control is often lacking. Still, assay results from unvalidated markers are being made available to clinicians, but guidelines on how the results should be interpreted are often missing. As most of the clinical trials involve multicenter cooperation, special emphasis should be put on the quality and reproducibility of the assays performed in the different laboratories. An excellent example of what can be reached with proper quality assessment and assurance is measurement of steroid hormone receptors in breast cancer tissue extracts. Determination of ER and PgR content of human breast tumors is widely used for prediction of response to endocrine therapy (24) as described above, and also for indicating patient prognosis. Within Europe, a multitude of translational multicenter cancer studies have been coordinated by the EORTC. Within the EORTC, the RBG serves to this consortium to research and advise on common methodologies for biomarker assays and to ensure that appropriate external quality assessment schemes are available. In the past 20 years, large-scale external quality assessment trials for ER and PgR, amounting to 165 participating laboratories from 18 countries, have been carried out organized by the RBG (25, 26). During several workshops, adequate and uniform methodologies for ER and PgR assays performed were defined and applied (27–29). By use of calibration vials, the initial between-laboratories coefficients of variation of 45% were substantially reduced to less than 15% (25).

FINAL TEST RESULTS DEPEND ON TOTAL OF PREANALYTICAL, ANALYTICAL, AND POSTANALYTICAL ASPECTS

Preanalytical Aspects—Before a tumor specimen or blood sample enters the process of analyte quantification, several crucial steps have been passed outside the laboratory. Collection of tumor specimens should be done as representatively as possible with regard to storage temperature of the tumor tissue during transport to the pathologist and the tumor bank (on ice pack), size of the piece of tumor, as well as content of tumor cells, nonmalignant cells, extracellular matrix, fat, and/or presence of necrotic areas in the tumor specimen. Because of this obvious heterogeneity, sampling bias may occur leading to different assay results if different areas of a tumor are analyzed. In addition, selection bias may occur

as very often only tissue pieces of relatively large tumors are given to the tissue banks, the smaller tumor specimens (<1 cm in diameter) being used by the pathologist for primary diagnosis. Transport of the tissue from the operating theater to the laboratory should be done in a standard manner and as quickly as possible (<1 h). Upon receipt of the specimen by the laboratory, the material should be placed on ice and immediately be processed or snap-frozen in liquid nitrogen. Long-term storage should be in low temperature-controlled containers (–80 °C freezers or liquid nitrogen tank) and freeze-thawing cycles avoided. Disintegration/extraction of tissue samples should be done according to internationally agreed protocols. Tissue extract aliquots (50 μl) should be snap-frozen in liquid nitrogen, and storage should be done at low temperature (–80 °C freezers or liquid nitrogen tank). Once the aliquots have been thawed, they should be used up and not be refrozen anymore. Blood samples or bodily fluids should be collected under standardized conditions (*e.g.* type of anticoagulant employed, application of a tourniquet, time of day, condition of the patient, etc.). Plasma or serum should be prepared, aliquoted, and stored at low temperature (–80 °C freezers or liquid nitrogen tank).

Analytical Aspects—Prior to producing test results used for clinical application, for each of the methods a (test) laboratory must verify or establish performance specifications of the test. The following analytical specifications have to be assessed: type of standard/reference material, recovery rate, accuracy, precision, sensitivity, specificity, linearity, and interferences.

A standard (reference material) is used to relate the reading of an assay to the quantity of measured analyte. Regarding the standard employed, in immunometric assays one should take care of the stability of the standard, buffer composition, and affinity between the standard and antibody.

It is worth mentioning that analytes extracted from tumor tissue may be different in nature from those present in the peripheral blood circulation of a patient. Therefore, an assay designed for measurement of an analyte in tissue extracts may not always be suitable for assaying the same analyte in plasma or serum. In recovery experiments, a known amount of standard is added to samples with a known amount of biomarker and then recovery of the added marker is calculated. This will provide information on the nature of the analyte *versus* standard and/or on any interfering process. The accuracy of an assay is the agreement between the best estimate of a quantity and its true value. Still, only for analytes for which a reference method is available is such comparison possible. The precision of an assay is defined by the agreement between replicate measurements. For validation of an assay, at least the intra-sample, intra-assay precision performance should be included. The precision profile is an ideal tool to assess this. Two types of sensitivity are of interest in immunoassays. The limit of detection (analytical sensitivity) is defined by the lowest concentration detected that is significantly different from zero. The limit of quantification (functional sensitivity) is the lowest concentration at which a test result can be reliably measured with a coefficient of variation of <20%. The specificity of immunometric assays strongly depends on antibody characteristics. Polyclonal or monoclonal antibodies or mixtures of both are applied in different test kits. In general, when polyclonal antibodies are used there is increased sensitivity but also an increased risk that one of these antibodies will recognize an epitope on a different antigen, resulting in decreased specificity. In contrast, monoclonal antibodies are directed to a single epitope and have higher specificity. Most assay procedures demand that samples be diluted to within a specified range of protein content prior to the assay, but values multiplied by the dilution factor should give the same results, irrespective of extent of dilution (*i.e.* parallelism/linearity studies). One has to be aware of, however, that blood from patients treated with immunotherapy may contain antibodies against these therapeutic antibodies, which may interfere in sandwich assays leading to false-positive or false-negative test results. Such heterophilic antibodies can also occur through frequent contact of the patients with animals. Moreover, blood from patients with infectious diseases may contain high amounts of IgM molecules, causing nonspecific reactions in sandwich-type immunoassays. Assays should be designed to reduce these potential interferences.

Postanalytical Aspects—It is of crucial importance to establish specific guidelines for interpretation of assay results if the marker is to be used in the clinical setting, as numerous statistical approaches are available to process the assay read-outs and interpolate them in the standard curves. Also, for multicenter studies, at each of the centers, the same statistical approach to processing of assay data should be used. Reference intervals for the tumor marker are needed for the specification of "high" and "low" levels, according to which patients can be divided into relevant risk groups if the marker is used for prognostic stratification, and also a reference range can help the clinician in identification of patients with cancer disease if the marker is validated for diagnostic purposes. Because of population sampling errors and biological variation, every laboratory should establish its own reference values, if appropriate.

Standardization of Total Protein Assay—In general, the amount of a biological marker detected in a tumor tissue extract is related to the total protein content of the sample. Therefore, it is important to standardize the protein assay as well, as high within- and between-laboratory coefficients of variation of protein assays have been reported (25).

Quality Monitoring—New assays should be thoroughly validated upon first use. Design of an adequate control procedure should start with a definition of quality requirements weighing acceptable error against needed clinical decision levels. Every assay consists of a measurement procedure to determine analyte levels of a biomarker and a control procedure in which, by measurement of control samples, the validity of the measurement of the samples can be checked.

Comparison of the analyte values of control samples against predefined limits should always be an integral part of the assay procedure. Control (or reference) preparations should be time and temperature resistant with little or no vial-to-vial variation, homogeneous, similar in buffer/matrix composition to the test material, available at concentrations that cover the physiological range expected in the experimental material, and available in sufficient quantity.

For internal quality control (QC) purposes, the laboratory must include samples of different concentrations of control material (tissue extracts, serum, plasma, bodily fluids, etc.). Repeated measurement of control samples allows determination of the imprecision of the assay system. The long-term trend in assay performance should be checked regularly in order to detect any shift or drift.

For external QC purposes, preparations distributed by a reference laboratory should be included. External QC programs serve to monitor long-term assay performance within a laboratory. Moreover, they provide comparison of assay results between laboratories. This enables the external QC organization to assess systematic errors between laboratories. Systematic differences in test results pave the way for calibration, and successful normalization of data from different laboratories form one of the cornerstones of valid multicenter studies on the potential value of biomarkers. Normalization can only be achieved, however, if a marker is homogeneous in nature, preferentially with only one molecular form present.

It should be noted that in the case of using lyophilized cytosols as external QC samples, no conclusions can be derived with regard to preanalytical issues as the use of external controls only covers reproducibility of the analytical assay procedure and subsequent computation of data. Providing proper instructions is the only feasible way to monitor (between-hospital) variations in sample treatment conditions. Because most clinical trials are carried out on a multicenter basis, the importance of between-laboratory QC cannot be overemphasized. Therefore, all steps in the procedure from taking biopsies to reporting assay results to the clinician, including the preanalytical ones, should be subjected to strict handling. These handlings should be described in standard operating procedures.

CLINICAL TESTING

Before any routine clinical use of a marker, extensive and elaborate studies on performance and robustness of the test kit have to be carried out. As described above, procedures for "preclinical" testing, including pre-, peri-, and postanalytical assay performance, retrospective studies, meta-analyses, and prospectively performed clinical trials should all be part of such systematic and detailed studies. Also, based on such studies, cut-points, or reference intervals, for the markers should be established making use of the marker practical for daily clinical routine. The evaluation and use of uPA and PAI-1 as prognostic markers in breast cancer is an excellent example illustrating the route for a marker from the laboratory to the clinical setting.

In 1996, the TMUGS framework was proposed by the American Society of Clinical Oncology (2). In this framework, all available published data for the tumor marker in question forms the basis of an evaluation and subsequent scoring of the marker. However, in recognition of the great variability in quality and therefore in validity of reported tumor marker studies, a system for classification of published data into certain levels of evidence (LOE) was incorporated in the TMUGS framework. Herein, the published papers evaluated are categorized according to evidence levels ranging from V to I depending on study design and size. The lowest level of evidence is LOE V, where the evidence for the tumor marker has been gained from small pilot studies. LOE IV and III are descriptive for retrospective studies of either small (IV) or large size (III); most tumor marker studies are of these two levels. LOE II comes from prospective therapeutic studies, where the primary question is therapeutic; marker testing is a secondary goal. LOE I studies are either high-powered prospective studies with the primary aim of testing the clinical validity of the tumor marker or meta-analyses of (several) studies of lower LOE stages.

With the LOE tool, it is possible to attribute a semiquantitative score to a potential tumor marker based on the published evidence available. The possible marker scores of the TMUGS range from "0" to "+++" (2). A score of "0" signifies that the tumor marker in question is safely concluded to be of no clinical utility based on sufficient data. "NA" (not available) denotes a lack of data for the marker. The scores of "+/-" to "+++" correspond to grades from "investigational with only preliminary data" to "independent information for clinical decision-making". Only markers deemed to be "++" or "+++" should be implemented as standard practice in clinical management of cancerous disease.

Thus, the TMUGS is a framework that covers important aspects to evaluate and take into consideration when working with tumor markers. In particular, it illustrates the need for systematization and rating of evidence when transferring markers from the laboratory into clinical practice.

CONCLUSIONS

In conclusion, tumor markers are important tools that can aid clinicians in questions regarding early diagnosis, estimation of patient prognosis, prediction of therapy response, and disease monitoring. Nonetheless, in parallel to the comprehensive requirements for federal approval of therapeutic drugs, tumor markers should undergo extensive studies of validation and quality assessment at several levels prior to introduction into the clinical setting. The process should be systematic, and stringent evaluation criteria with regard to quality of published evidence and clinical utility need to be fulfilled before a marker can be transferred into clinical practice. Markers must prove useful in improving patient outcome,

or quality of life, or in lowering costs of care.

Quality assurance is an issue of crucial importance in biomarker research and when implementing biomarkers in the clinic. A major problem associated with evaluating biomarkers in tissues, blood, or bodily fluids is that different procedures (sample collection, sample storage, sample processing) and different assay formats may yield different results. Therefore, assays and procedures have to be standardized and standard operating procedures should be developed for each type of sample and assay format, and the quality of the biomarker assay results should be monitored by continuous between-laboratory proficiency testing of performance. Although for some markers considerable progress has been made in the standardization of methods and assay protocols, efforts should be continued as only the stringent application of quality control systems enables a consistent assessment of the clinical value of biomarkers.

* The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

** To whom correspondence should be addressed: Department of Pharmacology and Pathobiology, Royal Veterinary and Agricultural University, Ridebanevej 9, DK-1870 Frederiksberg C, Copenhagen, Denmark. Tel.: +45 35283130; Fax: +45 35353514; E-mail: nbr@kvl.dk.

REFERENCES

1. Diamandis, E. P. (2002) Tumor markers: Past, present, and future, in *Tumor Markers: Physiology, Pathobiology, Technology, and Clinical Applications* (Diamandis, E. P., Fritsche, H., Jr., Lilja, H., Chan, D., and Schwartz, M., eds.) pp. 3–8, AACC Press, Washington, D. C.
2. Hayes, D. F., Bast, R. C., Desch, C. E., Fritsche, H., Jr., Kemeny, N. E., Jessup, J. M., Locker, G. Y., Macdonald, J. S., Mennel, R. G., Norton, L., Ravdin, P., Taube, S., and Winn, R. J. (1996) Tumor marker utility grading system: A framework to evaluate clinical utility of tumor markers. *J. Natl. Cancer Inst.* **88,** 1456–1466
3. Chan, D. W., and Schwartz, M. (2002) Tumor markers: Introduction and general principles, in *Tumor Markers: Physiology, Pathobiology, Technology, and Clinical Applications* (Diamandis, E. P., Fritsche, H., Jr., Lilja, H., Chan, D. W., and Schwartz, M., eds.) pp. 9–18, AACC Press, Washington, D. C.
4. Holten-Andersen, M. N., Christensen, I. J., Nielsen, H. J., Stephens, R. W., Jensen, V., Nielsen, O. H., Sorensen, S., Overgaard, J., Lilja, H., Harris, A., Murphy, G., and Brunner, N. (2002) Total levels of tissue inhibitor of metalloproteinases 1 in plasma yield high diagnostic sensitivity and specificity in patients with colon cancer. *Clin. Cancer Res.* **8,** 156–164
5. McGuire, W. L. (1991) Breast cancer prognostic factors: Evaluation guidelines. *J. Natl. Cancer Inst.* **83,** 154–155
6. Andreasen, P. A., Kjoller, L., Christensen, L., and Duffy, M. J. (1997) The urokinase-type plasminogen activator system in cancer metastasis: A review. *Int. J. Cancer* **72,** 1–22
7. Sweep, C. G., Geurts-Moespot, J., Grebenschikov, N., de Witte, J. H., Heuvel, J. J., Schmitt, M., Duffy, M. J., Janicke, F., Kramer, M. D., Foekens, J. A., Brunner, N., Brugal, G., Pedersen, A. N., and Benraad, T. J. (1998) External quality assessment of trans-European multicentre antigen determinations (enzyme-linked immunosorbent assay) of urokinase-type plasminogen activator (uPA) and its type 1 inhibitor (PAI-1) in human breast cancer tissue extracts. *Br. J. Cancer* **78,** 1434–1441
8. Harbeck, N., Schmitt, M., Kates, R. E., Kiechle, M., Zemzoum, I., Janicke, F., and Thomssen, C. (2002) Clinical utility of urokinase-type plasminogen activator and plasminogen activator inhibitor-1 determination in primary breast cancer tissue for individualized therapy concepts. *Clin. Breast Cancer* **3,** 196–200
9. Janicke, F., Prechtl, A., Thomssen, C., Harbeck, N., Meisner, C., Untch, M., Sweep, C. G., Selbmann, H. K., Graeff, H., and Schmitt, M. (2001) Randomized adjuvant chemotherapy trial in high-risk, lymph node-negative breast cancer patients identified by urokinase-type plasminogen activator and plasminogen activator inhibitor type 1. *J. Natl. Cancer Inst.* **93,** 913–920
10. Look, M. P., van Putten, W. L., Duffy, M. J., Harbeck, N., Christensen, I. J., Thomssen, C., Kates, R., Spyratos, F., Ferno, M., Eppenberger-Castori, S., Sweep, C. G., Ulm, K., Peyrat, J. P., Martin, P. M., Magdelenat, H., Brunner, N., Duggan, C., Lisboa, B. W., Bendahl, P. O., Quillien, V., Daver, A., Ricolleau, G., Meijer-van Gelder, M. E., Manders, P., Fiets, W. E., Blankenstein, M. A., Broet, P., Romain, S., Daxenbichler, G., Windbichler, G., Cufer, T., Borstnar, S., Kueng, W., Beex, L. V., Klijn, J. G., O'Higgins, N., Eppenberger, U., Janicke, F., Schmitt, M., and Foekens, J. A. (2002) Pooled analysis of prognostic impact of urokinase-type plasminogen activator and its inhibitor PAI-1 in 8377 breast cancer patients. *J. Natl. Cancer Inst.* **94,** 116–128
11. Harbeck, N., Kates, R. E., Look, M. P., Meijer-van Gelder, M. E., Klijn, J. G., Kruger, A., Kiechle, M., Janicke, F., Schmitt, M., and Foekens, J. A. (2002) Enhanced benefit from adjuvant chemotherapy in breast cancer patients classified high-risk according to urokinase-type plasminogen activator (uPA) and plasminogen activator inhibitor type 1 (n = 3424). *Cancer Res.* **62,** 4617–4622
12. Schmitt, M., Wilhelm, O. G., Reuning, U., Krüger, A., Harbeck, N., Lengyel, E., Graeff, H., Gänsbacher, B., Kessler, H., Bürgle, M., Stürzebecher, J., Sperl, S., and Magdolen, V. (2000) The plasminogen activation system as a novel target for therapeutic strategies. *Fibrinolysis Proteolysis* **14,** 114–132
13. Silvestrini, R., Daidone, M. G., Luisi, A., Mastore, M., Leutner, M., and Salvadori, B. (1997) Cell proliferation in 3, 800 node-negative breast cancers: consistency over time of biological and clinical information provided by ^{3}H-thymidine labelling index. *Int. J. Cancer* **74,** 122–127
14. Amadori, D., Nanni, O., Marangolo, M., Pacini, P., Ravaioli, A., Rossi, A., Gambi, A., Catalano, G., Perroni, D., Scarpi, E., Giunchi, D. C., Tienghi, A., Becciolini, A., and Volpi, A. (2000) Disease-free survival advantage of adjuvant cyclophosphamide, methotrexate, and fluorouracil in patients with node-negative, rapidly proliferating breast cancer: a randomized multicenter study. *J. Clin. Oncol.* **18,** 3125–3134
15. Paradiso, A., Schittulli, F., Cellamare, G., Mangia, A., Marzullo, F., Lorusso, V., and De Lena, M. (2001) Randomized clinical trial of adjuvant fluorouracil, epirubicin, and cyclophosphamide chemotherapy for patients with fast-proliferating, node-negative breast cancer. *J. Clin. Oncol.* **19,** 3929–3937
16. Harbeck, N., Kates, R. E., and Schmitt, M. (2002) Clinical relevance of invasion factors urokinase-type plasminogen activator and plasminogen activator inhibitor type 1 for individualized therapy decisions in primary breast cancer is greatest when used in combination. *J. Clin. Oncol.* **20,** 1000–1007
17. Hayes, D. F., Isaacs, C., and Stearns, V. (2001) Prognostic factors in breast cancer: current and new predictors of metastasis. *J. Mammary Gland Biol. Neoplasia* **6,** 375–392
18. Early Breast Cancer Trialists' Collaborative Group (1992) Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy: 133 randomised trials involving 31,000 recurrences and 24,000 deaths among 75,000 women. *Lancet* **339,** 71–85
19. Early Breast Cancer Trialists' Collaborative Group (1998) Tamoxifen for early breast cancer: An overview of the randomised trials. *Lancet* **351,** 1451–1467
20. Early Breast Cancer Trialists' Collaborative Group (1992) Systemic treatment of early breast cancer by hormonal, cytotoxic, or immune therapy: 133 randomised trials involving 31,000 recurrences and 24,000 deaths among 75,000 women. *Lancet* **339,** 1–15
21. Ravdin, P. M., Green, S., Dorr, T. M., McGuire, W. L., Fabian, C., Pugh, R. P., Carter, R. D., Rivkin, S. E., Borst, J. R., and Belt, R. J. (1992) Prognostic significance of progesterone receptor levels in estrogen receptor-positive patients with metastatic breast cancer treated with tamoxifen: Results of a prospective Southwest Oncology Group study. *J. Clin. Oncol.* **10,** 1284–1291
22. Klijn, J. G. M., Berns, E. M. J. J., and Foekens, J. A. (2002) Prognostic and predictive factors and targets for therapy in breast cancer, in *Breast*

Cancer: Prognosis, Treatment and Prevention (Pasqualini J. R., ed.) pp. 93–124, Marcel Dekker, New York
23. Benraad, T. J., Geurts-Moespot, J., Grondahl-Hansen, J., Schmitt, M., Heuvel, J., deWitte, J. H., Foekens, J. A., Leake, R. E., Brunner, N., and Sweep, C. (1996) Immunoassays (ELISA) of urokinase-type plasminogen activator (uPA): Report of an EORTC/BIOMED-1 workshop. *Eur. J. Cancer* **32A,** 1371–1381
24. McGuire, W. L., and Clark, G. M. (1992) Prognostic factors and treatment decisions in axillary-node-negative breast cancer. *N. Engl. J. Med.* **326,** 1756–1761
25. Geurts-Moespot, J., Leake, R., Benraad, T. J., and Sweep, C. G. (2000) Twenty years of experience with the steroid receptor external quality assessment program—the paradigm for tumour biomarker EQA studies. On behalf of the EROTC Receptor and Biomarker Study Group. *Int. J. Oncol.* **17,** 13–22
26. Sweep, C. G., and Geurts-Moespot, J. (2000) EORTC external quality assurance program for ER and PgR measurements: Trial 1998/1999. European Organisation for Research and Treatment of Cancer. *Int. J. Biol. Markers* **15,** 62–69
27. Koenders, A., and Benraad, T. J. (1984) Standardization of steroid receptor analysis in breast cancer biopsies: EORTC receptor group. *Recent Results Cancer Res.* **91,** 129–135
28. EORTC Breast Cancer Cooperative Group (1980) Standards for the assessment of hormone receptors in human breast cancer. *Eur. J. Cancer* **9,** 379–381
29. EORTC Breast Cancer Cooperative Group (1980) Revisions of the standard for the assessment of hormone receptors in human breast cancer. *Eur. J. Cancer* **16,** 1513–1515