
Gene Expression Profile Tests for Early Stage Breast Cancer

Version 1.0 Published June 2009

Authors

✈ Steve Goodman, MD, PhD
✈ Kay Dickerson, PhD
✈ Renee Wilson, MS

Johns Hopkins Center for Clinical Trials

Johns Hopkins Sidney Kimmel Cancer Center

Goals and Objectives

▪ Describe clinical studies that provide decision makers with reasonable confidence a medical technology improves health outcomes

▪ Provide a technology-specific methodological roadmap for the design of prospective comparative effectiveness research.

▪ Facilitate a dialog among patients, clinicians, payers, policy makers, methodologists, and industry about evidentiary standards within a specific medical technology area

© 2009 Center for Medical Technology Policy
Unauthorized use or distribution prohibited.
All rights reserved
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preface to the Effectiveness Guidance Document</td>
<td>i</td>
</tr>
<tr>
<td>Purpose</td>
<td>i</td>
</tr>
<tr>
<td>Complementarity with EGAPP</td>
<td>ii</td>
</tr>
<tr>
<td>Funding</td>
<td>iii</td>
</tr>
<tr>
<td>Introduction and Background</td>
<td>1</td>
</tr>
<tr>
<td>Gene Expression Technology</td>
<td>1</td>
</tr>
<tr>
<td>Use of Gene Expression Profiling Tests in Breast Cancer</td>
<td>1</td>
</tr>
<tr>
<td>Regulatory Framework for Multivariate Prediction Technology</td>
<td>2</td>
</tr>
<tr>
<td>Pre- and Post-marketing Requirements</td>
<td>3</td>
</tr>
<tr>
<td>Current Technical Documents</td>
<td>4</td>
</tr>
<tr>
<td>Reporting Standards</td>
<td>5</td>
</tr>
<tr>
<td>Guidance Recommendations</td>
<td>6</td>
</tr>
<tr>
<td>Statement of Test Purpose,</td>
<td>6</td>
</tr>
<tr>
<td>Clinical Context and Target Population</td>
<td>6</td>
</tr>
<tr>
<td>Algorithm Development and Validation</td>
<td>7</td>
</tr>
<tr>
<td>Analytic Validity</td>
<td>10</td>
</tr>
<tr>
<td>Statistical Assessment of Prognostic Ability</td>
<td>14</td>
</tr>
<tr>
<td>Assessment of Predictive Ability</td>
<td>16</td>
</tr>
<tr>
<td>Reporting</td>
<td>18</td>
</tr>
</tbody>
</table>
Preface

Purpose

Effectiveness Guidance Documents (EGDs) provide specific recommendations to product developers and clinical researchers about the design of clinical studies that will produce the evidence desired by patients, clinicians and payers. The goal is to describe clinical studies that would provide these “post-regulatory” decision makers with a reasonable level of confidence that the technology improves health outcomes. In this respect, they are intended to provide technology-specific methodological roadmaps for the design of prospective comparative effectiveness research.

The primary target audiences for EGDs are product developers and clinical researchers who are interested in designing clinical studies that will be informative to patients/consumers, clinicians and payers. They should also be useful to these decision makers themselves as they assess and appraise research that has already been conducted. They will be able to compare the design of available clinical studies to the recommendations contained in the EGD.

Each EGD focuses on a specific category of health care technologies, for example, radiation therapy for cancer, cardiac imaging for diagnosis of coronary disease, gene expression profiling for breast cancer, or mechanical interventions for chronic wounds. Methodological considerations for the design of clinical studies will be specific to defined categories of technologies so that the recommendations for study designs are more concrete and specific when targeted to a well-defined group of related technologies. For therapeutic interventions, the primary focus will be on evidence of comparative clinical effectiveness. For diagnostic intervention the primary focus will be on comparative clinical utility.

EGDs are intended to be analogous to FDA guidance documents, which are also targeted to product developers and clinical researchers, and provide guidance on the design of clinical studies that are intended to support regulatory decision making. EGDs will serve a comparable function for product developers and clinical researchers, but are focused on the design of clinical studies to support post-regulatory decision making. These post-regulatory decisions include individual clinical decisions made by patients/consumers, clinical recommendations made by clinicians, clinical policies generated by medical professional societies, and reimbursement decisions made by payers. Since the FDA has regulatory oversight over all health care technologies, that organization is naturally positioned to develop guidance documents providing recommendations on studies intended for regulatory approval. Because there has not been a single organization that represents the universe of post-regulatory decision makers, CMTP is providing a forum in which study design recommendations can be generated reflecting the perspective of key decision makers (patients, clinicians, payers, industry and policy makers) in the design of comparative effectiveness research.

By including the relevant FDA regulatory experts in the EGD development process, it is hoped that EGDs will reflect optimal alignment between study design elements intended
for regulatory approval and study design elements targeted to clinical and health policy
decision making. This may help to reduce the need for multiple separate studies to address
these different evidentiary purposes.

The EGD recommendations are not intended to describe the design characteristics of “gold
standard” clinical studies. The recommendations aim to achieve a balance across a number
of important considerations associated with these studies, including scientific validity,
feasibility, time requirements cost. In other words, EGDs seek to achieve a balance
between scientific consideration and practical considerations, recognizing that there is an
inevitable trade-off between the level of certainty that can be achieved through clinical
research with the cost/time/burden required to achieve that level of certainty. EGDs aim
to strike a balance of a reasonable level of certainty at a reasonable level of burden. In
order to determine what constitutes a reasonable balance, the process used by CMTP to
produce EGDs (described in more detail below) integrates the perspectives of all
knowledgeable and affected experts and stakeholders.

Because they do not describe “gold standard” studies, EGD recommendations are not
intended to imply that further and more rigorous studies should not be done on any
specific technology. Many important questions are likely to remain even after studies are
completed that are consistent with the study design recommendations of the EGD, and
further research will be valuable to pursue for most technologies. In some cases, coverage
with evidence development will be a useful policy tool to facilitate additional studies of
technologies for which there is evidence that meets the EGD recommendations.

Complementarity with EGAPP

Like CMTP, the Evaluation of Genomic Applications in Practice and Prevention (EGAPP)
initiative, established by the Office of Public Health and Genomics at the Centers for Disease
Control and Prevention, is working to improve evidentiary standards for genetic tests. An
independent panel supporting the EGAPP initiative, known as the EGAPP Working Group,
has commented on the quality of evidence for gene expression profiling (EGAPP Working
Group, 2009) and also offered guidance about the factors they consider in rating the quality
of evidence on genetic tests in general (Teutsch et al., 2009).

While both CMTP and EGAPP aim to improve the quality of evidence supporting the use of
generic tests, there are important distinctions between Effectiveness Guidance Documents
and the work of the EGAPP Working Group. Perhaps the most important distinction is that
our Effectiveness Guidance Documents are live, evolving standards, which aim through
broad stakeholder input to narrow the chasm between what product developers consider
to be feasible evidence to produce and gold standards for clinical research. The EGAPP
Working Group recognizes that gold standards for evidence are often not attainable for
genetic tests for many valid reasons, including the indirect link between outcomes of
interest and the actions of patients or their clinicians who receive the test results, and the
rapidly evolving competitive market for genetic tests (Teutsch et al., 2009). For this
reason, they note indirect evidence may plausibly suggest selected gene expression profiling tests could improve clinical outcomes (EGAPP Working Group, 2009), and modeling may be useful for examining this indirect evidence (Teutsch et al., 2009). Another important distinction between our work and that of EGAPP is the main target audience for Effectiveness Guidance Documents is product developers and clinical researchers. For that reason, our Effectiveness Guidance Documents aim to provide sufficient detail so that these groups could design studies that produce a level of evidence that is minimally sufficient for clinical adoption and coverage. While we see many complementarities between our work and that of EGAPP, we hope by providing a neutral, objective forum, and a mechanism for a dialogue, we might bridge the gap in communication between health care decision makers and product developers and change the way studies are designed in the future.

**Funding**

The Center for Medical Technology Policy (CMTP) is a private, non-profit organization that provides a trusted, transparent, and neutral forum to promote discussion and to develop practical tools and strategies that improve the quality and efficiency of clinical research for decision-making in clinical and health policy. Although CMTP actively seeks input from all stakeholders in a transparent process, the organization and primary author retain complete independence and responsibility for the conclusions and recommendations set forth in the EGD. CMTP receives funding through a diverse combination of sources, but it does not accept financial support from product developers or private insurers to develop any individual EGDs. Since its inception, CMTP has received funding from the following sources:

- Aetna
- Agency for Healthcare Research and Quality
- Blue Shield of California Foundation
- California Healthcare Foundation
- The Commonwealth Fund
- Institute of Medicine
- Johnson & Johnson
- Kaiser
- National Pharmaceutical Council
- Pfizer
- United Healthcare Foundation

More information on CMTP’s mission and products can be found at: [http://www.cmtpnet.org](http://www.cmtpnet.org)
Introduction and Background

Gene Expression Technology

Gene expression profiling technology exploits the transcription process of genetic information from DNA to messenger RNA (mRNA), and the subsequent translation of mRNA into protein. “Gene expression” refers to the translation of the information encoded in a gene into an RNA transcript (Marchionni et al., 2008). "Gene expression profiling" refers to any genomic technique that measures the degree of expression of multiple genes from an individual's tissue sample. "Gene expression pattern" and "gene expression signature" are equivalent terms to "gene expression profile." Gene expression profiling tests allow us, for the first time, to get a dynamic view of the tumor's molecular biology via a functional “fingerprint,” providing biological clues regarding the properties and clinical behavior of an individual's tumor (Marchionni et al., 2007).

Uses of Gene Expression Profiling Tests in Breast Cancer

Gene expression profiling can be used for prognosis or for prediction of a clinical benefit of a treatment (Marchionni et al., 2007; Marchionni et al., 2008). Throughout this report, predictive and prognostic indicators will be defined and used in the following manner. Predictive indicators are defined as those that predict response or non-response to a specific therapy or combination of therapies. Prognostic indicators predict the frequency of clinical outcomes (such as disease relapse, or death) independent of future treatment effects. There are many prognostic indicators (e.g. age) that may not predict the magnitude of treatment benefit, and some predictive indicators (e.g. estrogen receptor) which have a much stronger relationship to the efficacy of a therapy (e.g. tamoxifen) than they do to prognosis in the absence of therapy. Some indicators have both prognostic and predictive properties.

In the arena of genetic testing, another terminology is often used that has parallel meaning: clinical validity is defined as the accuracy with which a test predicts a patient's prognosis and clinical utility is defined as the accuracy with which a test predicts treatment benefit (Marchionni et al., 2008). The latter has been redefined by some to mean that applying the test in a clinical setting would result in better overall outcomes for the patient than if the test were not available (Hunter, 2008). These descriptors can be confusing, because knowing one's prognosis can have considerable clinical utility, in the plain language sense, and “clinical validity” suggests something different than prognosis. The prognostic/predictive terminology widely used in the biostatistical realm is more linguistically precise, and does not attach an adjective that suggests which property is useful.

The most important use of any prediction technology is to improve treatment decisions. In breast cancer, difficult treatment decisions face those with early stage disease. Adjuvant endocrine therapy and cytotoxic chemotherapy significantly improve breast cancer
survival in populations of women with invasive breast cancer (Berry, 2005; Early Breast Cancer Trialists’ Collaborative Group, 2005), but these treatments are not necessary for all individuals, with studies showing that a large proportion of early stage breast cancer patients go into remission with surgery (lumpectomy or mastectomy) and radiation alone (Early Breast Cancer Trialists’ Collaborative Group, 2005). If a woman knew that her chance of recurrent disease was so low (in other words that she has such a favorable prognosis), or alternatively that she has a relatively unfavorable prognosis but that the chances that the therapy would work were so low (in other words, that a predictive factor suggests resistance to therapy), she might conclude that the absolute chance of her benefitting is so small that she might choose to forego the discomfort and attendant risks of chemotherapy.

In breast cancer, prognostic factors that classify women at higher or lower risk for disease recurrence after surgery are already in wide use (Isaacs, 2001). These factors include nodal status, tumor size, and in many practices tumor grade. Furthermore, while estrogen and progesterone receptor (ER, PgR) and HER2 status provide modest prognostic information, they are potent predictors for benefit (or lack of benefit) from endocrine treatments such as tamoxifen or the aromatase inhibitors and anti-HER2 therapies, such as trastuzumab or lapatinib (Harris, 2007; Wolff, 2007). These factors are taken into account by guidelines bodies, such as the National Institutes of Health (NIH) Consensus Development Criteria (Eiffel et al., 2001; NIH, 2000), St. Gallen Guidelines (Goldhirsch et al., 2007), National Comprehensive Cancer Network (NCCN) Guideline (NCCN, 2008), Adjuvant! Online (Ravidin et al., 2001) and the American Society of Clinical Oncology (ASCO), and the College of American Pathologists (CAP) (Box A). Gene expression profiling offers the possibility of improving on these predictions by including information on a tumor’s biology that conventional predictors do not have (Cardoso et al., 2008; Pawitan, 2005).

### Regulatory Framework for Multivariate Prediction Technologies

#### Regulatory Roles of Federal Agencies

At the Federal level, the Food and Drug administration (FDA) and the Centers for Medicaid and Medicare Services (CMS) have the prominent oversight roles for genetic tests and testing sites.

Gene expression profiling tests developed and manufactured by a clinical laboratory for the laboratory’s own purposes (a laboratory-developed test), are not currently considered medical devices and therefore are not subject to medical device regulation requiring assessment of their clinical validity (http://www.genome.gov/19518345). They currently fall under the regulatory guidance of the CMS’s Clinical Laboratory Improvement Amendments of 1988 (CLIA) guidelines (FDA, 2007a), which govern laboratory operations in general and in particular the consistency of laboratory procedures used to implement the test (http://www.fda.gov/cdrh/clia/). CMS may initiate inspections: these inspections
may include a random, brief review of the required validation package for a laboratory-developed test, but such a review does not constitute an in-depth evaluation of test validation data (U.S. Department of Health and Human Service, 2008). The FDA has been involved with CLIA since it took over the responsibility for categorizing the complexity of certain diagnostic tests in 2000 (www.cms.hhs.gov/CLIA/10Categorization_of_Tests.asp#/).

If a gene expression profiling test is to be marketed as a “test kit” for commercial distribution to be implemented by labs, providers, or patients, the FDA considers it to be a medical device (http://www.fda.gov/cdrh/devadvice/312.html), and subject to regulation by the their Center for Devices and Radiological Health. The FDA defines regulatable kits as those that are "intended for use in the diagnosis of disease or other conditions, or in the cure, mitigation, treatment, or prevention of disease." Test kits can be classified as Class II or Class III medical devices (see Box B for definitions and related regulations). Depending on the device classification, either a premarketing notification or a premarketing application needs to be submitted for clearance or approval to the FDA. This process is discussed in more detail in the following section.

The FDA also recognizes a category of devices known as “in vitro diagnostic multivariate index assays” (IVDMIAs). IVDMIAs are tests that use a combination of patient laboratory data (i.e., analysis of cancer tissue or body fluids such as blood or urine) plus an “interpretative” analysis to yield a “score” or “index” useful for diagnosis and treatment decisions. FDA’s IVDMA draft guidance (U. S. Department of Health and Human Services, 2007) intended as a policy guidance document suggests that IVDMIAs would be subject to pre- and post-market FDA regulations regardless of whether they are developed for “in house” or for multi-site marketing. FDA approval under this guidance (if it moves past the “Draft Guidance” stage) will remain voluntary. Although the draft IVDMA guidance does not directly address the design of studies necessary to demonstrate clinical validity or clinical utility, it suggests contacting the Office of In Vitro Diagnostic Device and Safety Evaluation (OIVD) for the type of information to submit for pre-market clearance or approval on a test by test basis.

**Pre- and Post-marketing Requirements**

CMS has no authority to conduct post-market reviews, or adverse event reporting (U. S. Department of Health and Human Services, 2008), though they do have inspections, violation audits, sanctioning and licensing revocation authority under CLIA. The FDA has the authority to require pre-marketing certification (see http://www.fda.gov/cdrh/devadvice/312.html) and to perform post-market review of medical devices, including gene expression profiling tests, and has a system in place for adverse event reporting. The FDA has the authority to require data that demonstrate the safety and effectiveness of laboratory-developed tests (U. S. Department of Health and Human Service, 2008). Until recently, the FDA has not exercised this regulatory power via enforcement discretion. Two guidance (or draft guidance) documents have been generated that may place certain laboratory tests under closer scrutiny by the FDA: Guidance for
Industry and FDA Staff Commercially Distributed Analyte Specific Reagents (ASRs) (FDA, 2007b), and In Vitro Diagnostic Multivariate Index Assays (FDA, 2007a).

The processes for pre-market certification of medical devices as defined in **Box B** are detailed on the "device advice" web page and will not be elaborated upon here. Additionally, the "Ensuring the Safety of Marketed Medical Devices: CDRH’s Medical Device Postmarket Safety Program" (FDA, 2006a) provides further details and a graphic representation of the process that can be helpful to developers.

According to the “Guidance for Industry; Reports on the Status of Postmarketing Study Commitments...”(FDA, 2006b), postmarketing regulations as defined in sections 522 of the FDA Modernization Act of 1997, do not apply to medical devices. However, the CDHR internal postmarket safety workgroup has developed a report that discusses a medical device postmarket safety framework (FDA 2006a) to address risks and adverse events associated with medical devices available on the market. The document, “Ensuring the Safety of Marketed Medical Devices CDRH’s Medical Device Postmarket Safety Program” outlines the problem identification, assessment, and resolution processes utilized by CDHR. Based on the discussions of surveillance activities, science-based information, enforcement actions, information/education tools, and best practices within the organization the report provides a foundation for program improvement with a focus on prioritizing efforts and maximizing available resources (FDA 2006a). This report also provides a useful graphic representation of the post-marketing processes. It was followed up by the “Report of the Postmarket Transformation Leadership Team: Strengthening FDA’s Postmarket Program for Medical Devices” (FDA, 2006c), which sets forth recommendations regarding medical device safety and postmarketing requirements. Thus far, no other guidelines have been issued by the FDA and there are no postmarketing requirements in effect.

**Current Technical Documents**

There are a variety of technical documents available to individuals, institutions, or companies developing gene expression profiling tests, produced both privately and by government agencies. These cover only the initial technical steps in the test developmental pathway. This is partly due to the fact that the regulatory standard is vague, and that proving clinical value is not currently a requirement to bring a test to market. However, clinical value is important to professional organizations, payers, patients and physicians who will be utilizing the tests. An annotated bibliography has been added to this document (Appendix B) to assist developers and users in identifying important documents of interest.
Reporting Standards

Although the topic of reporting standards for studies of diagnostic and prognostic tests is not directly related to the topic of this guidance document, confusion about the difference between reporting standards and test development standards merits clarification. Our focus in this document is to provide specific direction on how to design clinical investigations that will produce the types of evidence needed for decision-making about uptake and coverage of gene expression profile tests. Clinical investigations undertaken as part of this process should be reported in the literature, regardless of the strength and direction of the findings. Fortunately, there are a number of reporting standards now available that can assist product developers in reporting results in a transparent and complete manner. REporting recommendations for tumor MARKer prognostic studies (REMARK) suggest relevant information related to the study design, methods, results and discussion sections to be included in articles about the prognostic ability of biomarkers (McShane et al., 2006). The Standards for Reporting of Diagnostic Accuracy (STARD) initiative was initiated to improve the accuracy and completeness of reporting in studies of diagnostic accuracy (Bossuyt et al., 2003; Bossuyt et al., 2004; Bossuyt et al., 2006).
Guidance Recommendations

Statement of Test Purpose, Clinical Context and Target Population

1. The therapeutic action or decision that the test is intended to affect must be pre-specified.

Clear specifications of the therapeutic decisions to be impacted by a given test are arguably the most important and most neglected aspect of predictor development. In breast cancer, initial decisions can include which type of surgery is best (e.g., lumpectomy or mastectomy), which type of axillary lymph node evaluation is indicated (none, sentinel node assessment, axillary nodal dissection), whether to have local-regional radiation after surgery, and whether to treat with adjuvant hormonal and/or cytotoxic chemotherapy. Each of these decisions involves incurring risk or morbidity for possible benefit. Clear specification of the decision implicitly defines the population of interest and is relevant to the choice of outcome. For example, many current guidelines strongly recommend adjuvant chemotherapy if a patient has evidence of axillary nodal involvement, based on poor prognosis. If the therapeutic decision of interest is whether to add adjuvant chemotherapy to surgery after initial diagnosis, one might not include women with regional or distant spread, for whom the decision to treat may already be determined based on these guideline recommendations. Similarly, one might exclude women who had a local recurrence or were previously treated. However, given the rapidly advancing field of cancer therapeutics, classes of patients for which there previously not therapeutic choices can change.

2. The population in whom this action or decision will be taken must be fully and clearly specified.

This includes clear specification of the patient population of interest, including their demographics (e.g. age, menstrual status), tumor characteristics and treatment history (size, grade, hormone receptor status, HER2 status), the clinical decisions facing them (e.g., addition of cytotoxic therapy to hormonal therapy), their treatment options at this point in their course, the efficacy of those treatment options, and the alternatives to testing (i.e., what decisions would be made in the absence of the test) (Nuyten et al., 2008).

3. There must be evidence that the treatment being considered can materially change prognosis in the target population.

Benefit from the assay can only be achieved in those situations where the resulting decision is likely to materially change the patient’s outcome, i.e. not when a favorable outcome is either highly likely or extremely unlikely regardless of treatment (Henry, 2006; Simon, 2005a).
4. The decision must be characterized in terms of all treatment alternatives under consideration at the time of the test (e.g., cytotoxic therapy in addition to hormonal therapy), along with concomitant treatments that might be unaffected by the test (Nuyten et al., 2008).

These specifications guide all subsequent steps in the development/evaluation pathway, and without them, that pathway may produce evidence that is not completely coherent, and thereby not convincing to regulators or prospective users of the test.

Algorithm Development and Validation

5. The manner in which the test or algorithm is developed has little or no relevance to approval standards, as the latter are based almost exclusively on external validation (test set) results.

Features of the test development process that maximize the chance that the test will validate well are the use of internal validation procedures (e.g., bootstrap, cross-validation), and of populations and endpoints similar to those on whom the test will be validated.

6. Minimal sample sizes required for reliable test development will be between approximately 50 and 80 subjects (Dobbin, 2007; Dobbin, 2008).

The purpose of initial profile/algorithm development is to develop a tool that distinguishes reliably between subjects who will experience the outcome of interest (e.g., 5 year survival) versus those who will not. The goals of this phase are first gene selection, followed by score development and choice of cutoffs. Cutoffs should be chosen to reflect probability classes likely to result in different therapeutic decisions.

Initial development of the profile is an exploratory exercise, and great flexibility is allowed in the choice of patient population, endpoints, gene selection algorithm, and gene expression measurement technology. However, since the goal is to develop a tool for use in a broader population of patients, these initial parameters must be chosen to maximize the chance that the discrimination seen in the “training” or “development” dataset is maintained in an independent set of patients who make up the “validation” or “testing” set. In the case of OncotypeDX™, this meant starting with a pool of 250 candidate genes that were known to play a prognostic role or mechanistic role in breast cancer, reduced to a set of 21 that demonstrate the most significant associations with recurrence (16 cancer-related, 5 reference) used in the final index. The development of Mammaprint™ was more of a data mining exercise, in which more than 5,000 genes were initially probed for their relationship with 5-year distant recurrence, 273 were initially selected using a supervised cluster algorithm, and a final pool of 70 genes were selected for use in the predictor after internal validation. The final gene sets of
OncotypeDX™ and Mammaprint™ shared only one gene but three biological pathways (proliferation, ER, and HER2)

The algorithms by which the expression levels are measured, combined, and used to classify risk are similarly diverse, with no specific classifier form that has been recognized as generally optimal. Typically, regression or non-parametric methods of risk categorization (e.g., CART) are employed. Zucknick et al. has compared a range of classification methods (univariate comparison to machine learning algorithms) (McShane et al., 2006). In all such exercises, internal validation techniques must be used to minimize the possibility that genes are selected due to chance. Standard methods of internal validation include bootstrapping, jackknife, and cross-validation. In the development set, classification performance can appear to increase with the number of genes in the development set, but at the risk of poorer validation in independent populations (Baker and Kramer, 2006)

Development studies should be large enough so that they can incorporate validation techniques and reduce chance findings (Steyerberg et al., 2001). Simulations (Dobbin, 2007; Dobbin 2008) show that under a variety of reasonable assumptions, optimum sample size ranges roughly from 50 to 100 subjects, if the proportion of the population with an event is 50%. Sample size can increase if that proportion is significantly below 25%. Interestingly, for the purposes of algorithmic development, traditional statistical parameters such as threshold significance level, number of genes, classifier structure and gene correlations are not relevant. Sample size depends on three factors and can be calculated using an online resource from the NCI (linus.nci.nih.gov/brb/samplesize/):

1. the largest standardized fold change, as measured by the difference in average expression between the classes divided by the within-class standard deviation of expression of that gene (on the log scale).
2. the number of genes or features on the microarrays
3. the proportion of cases and controls in the population.

**Tissue samples**

7. **Identify a source of tissue samples upon which the test will be developed/validated.**

It is essential that these samples represent a dataset that can be used to address the clinical question of interest, not just a broad question of “prognosis.” It is also important that the dataset provide clinical data that are relevant to the endpoints of interest and that follow up is sufficient to be informative. For example, if it is desired to develop a test that will predict 10 year survival after chemotherapy, that cannot be achieved with a dataset that has data only on 5 year recurrence rates. Similarly, if the target population is newly diagnosed women, cohorts of pre-treated women will not be useful. The control samples require similar care in characterization. Selection of the source of tissue samples and the specification of the clinical decision to be informed by the test can be an iterative process, with the clinical context determining the samples that are
sought, and the nature of the available samples in turn affecting the clinical decision that can be addressed, ultimately reflected in the indications claimed for the test.

8. **Ascertain and report data regarding tissue handling and preservation method and fraction of usable samples.**

Pre-analytical issues are of paramount concern. Does the assay work in the type of tissues that are available in the archived bank? Does it work equally well in archival tissue as it does in “real time”—since once the assay is marketed, it is likely to be used within a very short period of the specimen’s being collected. Some of the specifications outlined above might be constrained by the nature of the tissue samples available to the test developer, which needs to be established early in the process.

For the tissues that are used, including any control samples, the REMARK guidelines require the reporting of methods of preservation and storage (McShane, 2006). A useful guidance on storage and handling of biospecimens have been issued by the NCI Office Of Biorepositories and Biospecimen Research (biospecimens.cancer.gov).

The quality of data needed for test development and validation are typically only gathered under the auspices of research protocols, although sometimes are available as part of patient registries and tumor banks collected in the course of clinical care. The developer needs to ascertain the degree of missing data, whether all the data needed for risk adjustment have been gathered, and how the accuracy of data has been verified (for more discussion, see “Test validation”).

Another consideration is whether the tumor samples have been stored in a manner that permits gene expression analysis using the technology that will serve as the basis of the test being developed. Issues such as age of the sample, preservation method (e.g., FFPE or FF), and analyzable tumor percentage within the tissue sample can impact gene expression profiling test results. Another issue that will be discussed in the validation section is whether the nature of tissue sample preservation or processing mimics how the test will be administered in practice.
Analytic Validity

9. Analytic validity needs to be established and reported per CLIA, IVDMIA Guidance, CLSI guidance and FDA Special Controls Document for Gene Expression Profiling Test Systems (FDA, 2007c).

   Key analytic variability elements include:
   o Accuracy (determine if the assay measure the analyte accurately, if this can be ascertained independently of the technology used)
   o Reproducibility (within patient)
   o Percent of successful assays
   o Variability of risk classification

10. Laboratories should describe how they plan to monitor and report quality control/quality assessment and analytic performance data. {EGAPP, 2009}

   There is no gold standard for gene expression beyond the technology used for the assays (i.e. RT PCR and DNA microarray). Imperfect analytic validity reflects itself in the predictive properties of the test, so these already incorporate some elements of variability due to the assay procedures. However, as treatment decisions are contingent on test properties, it is important that the variability not occur at the individual level, i.e. that multiple assays from the same subject or tumor are not so different as to produce different clinical decisions. It is also important to ascertain the degree to which predictive properties reported in research studies on the gene signature apply to the marketed test. (See validation section). Finally, understanding the determinants of analytic validity, or reproducibility, are critical for quality control procedures to maintain the predictive values seen in the research setting.

Study Design: Test Validation

Choice of population

11. The population used to validate the prognostic algorithm (the “test” set) must be completely independent from the one used to develop the algorithm (the “training” set) (Simon 2005a, Teutsch 2009).

12. The validation population must reflect the population for whom the test is recommended, both in terms of therapeutic choices and clinical characteristics.

13. The validation population should be “therapeutically homogeneous,” i.e., that they are all facing the same therapeutic decision at the same point in the disease process (e.g., after initial diagnosis, or after first recurrence) (Simon 2005b).

14. The validation population must not have known prognostic heterogeneity beyond what is incorporated in measured risk factors, e.g. tumor histology, size, grade, hormone-receptor status, lymph node status, degree of spread, patient
age, and menopausal status. If the population is heterogeneous, population averages may apply to the group taken as a whole, but not for individual group members.

If the population is composed of two or more groups of clearly different prognosis (e.g., treated and untreated patients), the resulting numbers may not clearly apply to any individual in the group; it will be an average that is not wrong for the group taken as a whole, but is recognizably wrong for each group member. Since treatment decisions are made on an individual basis, the numbers from the group must plausibly apply to each member of that group. Prognostic homogeneity can be achieved using traditional prognostic and clinical factors currently used to select systemic adjuvant treatment of breast cancer and judge the likelihood of disease progression e.g. tumor histology, size, grade, hormone-receptor status, lymph node status, degree of spread, patient age, and menopausal status (Modlich et al., 2006). Combination prognostic indices noted earlier are a good tool as well to define such groups. If co-morbidities affect treatment decisions, then they must similarly affect eligibility criteria for validation (Aapro, 2001; McShane et al., 2006). In summary, the population on which the test should be validated must be chosen on the basis of similar treatment options and similar risks for the outcomes of interest (McShane et al., 2006; Bossuyt et al., 2004).

15. For the FDA, it is preferred that the validation dataset should consist of clinical samples collected from at least three different clinical sites in different geographical locations. If the studies are conducted outside the U.S, the relevance of studies to U.S. clinical practice and demographics must be documented (FDA, 2007b).

Test performance

16. The test, including the complete algorithm, created in the development or discovery phase cannot be altered in the validation phase. (If it is, additional independent data must be used to validate the modified algorithm.) (Simon, 2005a, Teutsch 2009)

17. Assays must be performed masked to the study end point and subjects with and without the outcome of interest should be not be clustered in ways such that a “batch effect” could bias predictive ability.

If the assays are not masked, there is no assurance that assays were not performed in such a way as to enhance predictive ability. Similarly, if many “cases” (i.e., subjects who experience the primary outcome) are processed together, separately from “controls”, any changes in assay properties between batches will falsely appear to be a property of the assay itself. There should be no correlation between case-control status and any temporal or procedural aspect of test performance.

18. Some of the validation must be of the entire test procedure (i.e. not just the performance of the expression “signature” as measured in research settings).
This means that patient samples must be sent, as they would in clinical practice, to the same lab and subject to the same procedures as will be used for the marketed test.

19. If the bulk of the validation information derives from the use of the signature and not the test, the results obtained from the marketed test must be compared - on the same individuals - to those obtained from the assay used in the studies to see how often they agree.

A key parameter is the percentage of women whose clinical decision about therapy, based on a proposed risk threshold (e.g., 5-10% recurrence risk) would have been changed if they took the marketed test. This agreement study should be done in a population that includes women from all risk strata, and should be large enough to reliably detect a 10% or greater chance of a changed decision, which requires the sample N in the range of at least 80-100 (depending on how many fall into different risk strata).

Outcomes

20. Valid outcomes or surrogates for breast cancer prognosis include distant recurrence at 5 or 10 years, disease-free survival, disease-specific mortality or overall survival.

Outcomes utilized in other studies have included interim outcomes such as residual disease or pathologic complete response (PCR) after neo-adjuvant systemic therapy (Natowicz et al., 2008), disease-free survival (DFS) (also disease recurrence (Conlin and Seidman, 2007) and recurrence-free survival (RFS)), and distant metastases (Hudis et al., 2007; Weigelt et al., 2005). Virtually all of these serve as surrogates for overall survival (OS), the outcome of most interest to patients (Goetz et al., 2006). However, not all retrospective datasets have sufficient data on survival. Distant recurrence (e.g., metastatic disease) is the best surrogate marker for survival in breast cancer. Sensitivity or resistance to a specific chemotherapeutic agent has also been assessed (Anguiano and Potti, 2007). Although this is not a clinical outcome, it can be a useful short-term proxy if the tests results are to be used for purposes other than patient decision-making.

The clinical outcome of interest should be specified prior to validating the marker; otherwise the developer could choose the best among several, which would require further validation. While the surrogate outcomes above tend to be correlated, it is difficult to know to what extent a test predictive of one will predict another; this depends on what aspect of treatment response or tumor biology is captured by the genes in the signature. Conversely, Onco-typeDX™, which incorporated no information about predicting treatment effect in its development phase, had significant ability to predict treatment benefit when applied to data from an RCT.
The time frame for the outcome measure must also be established, although this is typically constrained by the available data. In general, it is best to validate the test against relevant outcomes over as long a time horizon as the data allow, although the longer that time horizon, the more that subsequent treatments may confound the comparison, a particular concern in observational designs (see Study Design).

**Design**

21. **For purely prognostic studies, one-arm cohort validation studies are adequate.**

These one arm studies are strongest if they are based on a well characterized cohort originally assembled for research purposes, but similar quality evidence can be garnered from representative samples of patients from clinical care settings, if the quality of clinical and outcome information is high. If banked specimens are available, with reliable patient follow-up with the population well characterized in terms of prognostic factors and treatments, non-concurrent studies based on banked samples can provide reliable estimates of prognostic ability. It is necessary to document that these populations and specimens are not subject to collection, information or selection bias (Marchionni et al., 2007).

If a test is to inform decision-making, then presumably it must provide information relevant to the selection of subsequent treatments. Improved prognostication is of unclear value if this improvement is not accompanied by a therapeutic change that in turn improves net patient outcomes. Although some patients and providers regard improved prognostication by itself as providing a meaningful psychological benefit even without an effect on treatment or disease outcomes, knowledge alone has not generally been considered a medical health outcome by payers or regulators.

22. **For assessment of predictive value (aka clinical utility, treatment benefit), single arm designs can sometimes be of value.**

Ordinarily, information about prognosis by itself tells us little or nothing about the efficacy of treatment. However, this prognosis can sometimes provide upper bounds on the efficacy of treatment, which is sometimes sufficient information to make a decision. For example, if a test could identify a subset of women not treated with chemotherapy whose recurrence risk was very low, since recurrence risk could be lowered by therapy at most to zero, those women would know that the maximum benefit of therapy was small even if their cancer is sensitive to the effects of therapy. So if the long term recurrence without treatment was 5%, we would know that treatment would have at most a 5% absolute benefit, and probably less. If a woman had felt that the treatment benefit had to exceed 7% to make chemotherapy worthwhile for her, then knowledge that she was at most 5% risk of recurrence without chemotherapy might be sufficient information for her to forego it. Conversely, if the test identified a group receiving chemotherapy that nevertheless had an extremely poor prognosis, the low survival rate while on therapy would represent a (low) ceiling on the benefit of
treatment. So at the extremes, purely prognostic, non-comparative information can be useful to patients in making treatment decisions.

**Statistical Assessment of Prognostic Ability (Clinical Validity)**

23. Non-parametric (e.g., Kaplan-Meier curves) and semi- or fully parametric survival modeling techniques (e.g., Cox or Weibull models) are sufficient to assess the predicted risks of either a continuous or categorized risk score based on gene expression.

24. Measures of absolute risk and test discrimination are the basis on which to assess the clinical value of the test, not hazard ratios and p-values.

25. Risk prediction models must be both properly calibrated in the risk regions or cutpoints of interest and show adequate discrimination to be useful.

Calibration is a measure of how well the predicted probability from the gene expression score agrees with the observed probability (Steyerberg et al., 2001). If the scoring algorithm is not based on an actual predicted probability, but instead is a kind of concordance index (e.g., the basis of Mammaprint™), then the risks corresponding to various categories of the score must be calculated. Decision making requires absolute risks, so measures that do not clearly relate a gene-expression score to actual levels of risk will not be useful in a clinical context. The probability regions of greatest interest tend to be those at the extremes, i.e. either above or below a threshold risk for treatment. If the threshold risk is very low, as it is in the case of chemotherapy for breast cancer, then the score must accurately predict probabilities in that low risk region. This can be difficult, as it is often at the extremes of risk prediction that risk scores perform most poorly. But if the score is very accurate in the region of greatest interest (e.g. in the lowest risk subjects), but is poorly calibrated for higher risk subjects whose inaccurate risk prediction may not result in mistreatment, this may not be of concern from the perspective of decision-making.

Discrimination is the property that represents how well the score separates those who have an event from those who do not have an event. It can be globally measured via the area under receiver operating characteristic (ROC) curves and C-statistics, which both measure the probability that a randomly chosen subject who had an event will have a higher score than someone who will not (Steyerberg et al., 2001). However, these global measures of discrimination are not good measures of how many subjects will be impacted by the implementation of the score. For that, reclassification tables are needed (see below).

26. If there are to be cutpoints (e.g., “low” or “high” risk), fully describe the criteria used to establish them, and the absolute risks they correspond to. For distant recurrence endpoints, the maximum “low risk” threshold is generally considered
to be 10% by established clinical practice and medical professional societies. The cutpoints should be established in the development set and validated in an independent sample.

27. Standard regression techniques are necessary but not sufficient to show the added clinical value, i.e. comparative effectiveness, of a gene expression-based test over and above the standard combination prognostic indices, e.g., Adjuvant Online!®, St. Gallen criteria, and others. The statistical assessment must match the implementation of the tests, e.g. as a substitute, add-on or sequentially. Reclassification tables can be critical to demonstrate how many women would be placed into new risk categories that are likely to change a treatment decision (Pencina et al., 2007).

Reclassification tables are critical to the assessment of the decision impact of a given gene expression score. As implied by their name, they require that there be a comparator predictor, in that they measure exactly how the risk category changes with the addition of the new predictor, i.e. how subjects’ risk are “reclassified”. A reclassification table requires that it be clear whether the gene expression predictor will be used in combination with, or instead of, the comparator. If the comparator is based on freely available clinical information, as the breast cancer predictors are, then it makes the most sense to consider them in combination.

A reclassification table shows how many subjects are shifted into various risk categories, and what their observed risk is. Some of these shifts will correspond to changes in treatment decisions. It allows a direct quantification of how many subjects would have treatment altered by using the gene expression predictor in combination with existing predictors. For any projected treatment effect, the effects of this alteration could be directly modeled, so it would be possible to project whether the net benefit was outweighed by any projected increases in potential harms. These harms include both the physical harm of unnecessary treatment, neglected beneficial treatment, or adverse psychological effects (Fan et al., 2006; FDA, 2006a; Glas et al., 2006; Paik et al., 2006).

28. Sample size: There is no fixed guideline for sample size, but there can be for precision. The key determinants of this precision are not just overall sample size, but sample size and number of events in various risk strata. The precision needs to be sufficient so that high and low risk predictions – those with potentially therapeutic implications - in concert with established prognostic scores either via modeling or stratification have confidence intervals that do not appreciably overlap risk decision boundaries. For example, if a model predicts an 8% relapse risk, 95% CI, 1% to 17%, this does not provide high confidence that the true risk is less than 10%.
29. The proportion of unanalyzable, uninterpretable results and all missing data from the validation phase must be reported, as well as the reasons for all missingness. If these are appreciable (e.g., >10%), and if the reasons for missingness are possibly related to the value of the missing data, the possible bias produced by the missingness must be explored through sensitivity analyses. If indicated, appropriate missing data techniques (e.g., multiple imputation), should be employed.

**Assessment of Predictive Ability (Clinical Utility)**

30. An RCT, either concurrent or non-concurrent is the only design that provides an unbiased assessment of treatment benefit within subgroups of a risk score.

31. A non-concurrent assessment from an RCT that has banked, analyzable specimens and sufficient patient follow-up can provide very high quality evidence. The strength of this evidence depends on the strength of the underlying RCT, which needs to be assessed.

32. Treatment benefit is best assessed via an efficacy analysis stratified by a binary categorization of the index. However, if this binary split is dictated by the data, it is akin to an unplanned subgroup analysis and would require validation on independent datasets. Alternatively, if a continuous predictor of degree of benefit can be developed and validated as such.

33. A non-randomized comparative cohort study (with treated and untreated groups) can provide evidence about predictive ability, but the quality and acceptability of that evidence will depend critically on the degree to which the arms are deemed similar on all prognostic factors other than treatment.

34. Predictive ability encompasses both cost and Quality of Life (QOL) outcomes, so ideally these should be assessed as part of the development pathway (Teutsch 2009).

**Comparative cohort designs for predictive ability**

35. Comparative cohort designs are critical for assessing predictive ability, and randomization is highly desirable, although non-concurrent randomized cohorts can speed the evaluation process. The sufficiency of non-randomized cohorts must be very carefully scrutinized, on a case by case basis.

As noted, the ideal comparative cohort design is a randomized trial, performed either prospectively or non-concurrently. The advantages of using samples from a non-concurrent (aka “retrospective”) study is greatly reduced expense and time, as decades of follow-up may be available. The disadvantages are many. Tissue samples may have been lost, degraded or depleted over time, or the measurement of expression levels on
old samples may not be similar to recent ones. Co-treatments, patient monitoring or diagnostic procedures and supportive therapy may have changed over time, diminishing the relevance of patient survival data to today. Risk factors, tumor characteristics, or aspects of the clinical course relevant to the study may be missing or ascertained suboptimally in older records, due to the use of old technologies or for other reasons. However, there is nothing inherently flawed about the logic of using non-concurrent data; the issues are mainly the logistical challenges and the relevance of those results to the current day.

The validation of the predictive ability of the 21-gene recurrence score assay (OncotypeDx™) represents a model for how non-concurrent RCT data can be used to assess the ability to predict treatment benefit. They examined the comparative data from the NSABP B-20 trial (Paik, 2006). In that study, Paik and colleagues examined 10-year, distant recurrence-free survival in 651 patients with estrogen receptor positive, lymph node-negative disease who were randomly assigned to receive tamoxifen alone or tamoxifen with chemotherapy. They compared the 10 year recurrence risk between the two arms, stratifying by low and high values of the OncotypeDX™ risk score. The marked difference in estimated treatment effect between those two strata represented the strongest evidence thus far that gene expression tests can predict the magnitude of treatment benefit, at least for this particular treatment in this treatment context (i.e., newly diagnosed, early stage cancer).

The current design of the MindACT trial, which is testing the clinical utility of the Amsterdam 70-gene assay (Mammaprint™ in estrogen receptor rich, node negative breast cancer patients,) represents another model for prospective randomized assessment of the test’s value, but also exemplifies the difficulty of such evaluations; the results are not expected for many years and the rapid adoption of gene expression technology into clinical practice has reportedly inhibited recruitment. MindACT is a large, multicenter, prospective RCT which initiated recruitment in early 2007 and plans to recruit 6000 node negative breast cancer patient within 3 years of the start date. Risk of relapse will be assessed by using both Adjuvant! Online and Mammaprint™ (Cardoso et al., 2008). If patients are considered low risk by both clinical and Mammaprint™ criteria, they receive adjuvant endocrine therapy alone. If they are considered high risk by both, they receive both endocrine and chemotherapy. If they have either clinical features of high risk but low Mammaprint™ profile, or vice verse, they are randomly assigned to use one or the other to make the decision to receive chemotherapy or not.

The TAILORx study, NCI’s Trial Assigning Individualized Options for Treatment, is designed to help address the challenge of integrating Oncotype DX™ into clinical practice. This study was designed to determine both the prognostic and predictive value of Oncotype DX™ based on a NCI organized clinical expert consensus design using a non-commercial mid-range category scores of 11-25 to be applied to approximately 11,000 randomized patients (Sparano and Paik, 2008). However, unlike the MindACT trial, TAILORx does not randomly assign patient decision making to use the test or not. Rather, eligible node negative, estrogen receptor rich breast cancer patients are all profiled with the test. Those with high risk results will receive endocrine therapy and
chemotherapy, and so it is assumed the test accurately identifies them as patients who will benefit. The study is designed to prospectively address the prognostic utility of the assay in those with low risk results, who will only receive endocrine therapy and be followed. Patients who are in an intermediate risk range, as determined by the test, will all receive endocrine therapy, but will be randomly assigned to chemotherapy or not. Thus, the trial already assumes the clinical value of the test at its higher range, and is prospectively testing its prognostic utility in the lower range (assuming optimal endocrine treatment) and predictive ability in the middle range. Furthermore, the risk cutoffs used for the test are not the same as are being used to designate high and low risk, a mismatch that could limit the utility of this design. This is not necessarily a design for prospective test evaluation that would be accepted for all tests, in that it assumes the value of the test in identifying good candidates for chemotherapy, but that is probable enough given the preceding results that this trial will likely be sufficient for approval or coverage purposes. The TAILORx trial will also follow patients for up to 20 years and multiple sample types will be collected for each patient in a central repository for the purposes of further test and algorithm development across the risk continuum of early stage breast cancer.

The use of non-randomized cohorts for the evaluation of predictive ability suffers from the same weaknesses as non-randomized assessments of treatment efficacy. If differently treated cohorts can be found whose prognostic characteristics, which include concomitant therapies, are plausibly equivalent, then the evidence of a large differential effect as a function of gene expression score might be deemed as sufficient for approval or coverage purposes. However, finding such cohorts can be challenging, as can convincing evaluators that they are prognostically equivalent. The measurement of outcomes other than clinical endpoints (e.g. cost, QOL) is also important for complete clinical utility assessment.

**Reporting**

1. **Reporting of developmental array experiments should follow MIAME guidelines** ([www.mged.org/Workgroups/MIAME/miame_2.0.html](http://www.mged.org/Workgroups/MIAME/miame_2.0.html)) (Box C).

2. **Reporting of all clinical studies validating a test should follow the ReMark guidelines** (McShane et al., 2006) (Box D).

3. **The STARD initiative has another set of reporting guidelines that apply in part to gene expression tests** (Bossuyt et al., 2003; Bossuyt et al., 2004) (Box E).

Beyond public repositories of information on gene expression profiling tests, there is a need for standardization in reporting of results on the diagnostic accuracy of these tests in the scientific literature. Bossuyt et al. has developed and published guidelines on how data and information on diagnostic tests should be reported. The objective/purpose of the STARD initiative is to improve the accuracy and completeness of reporting of studies on diagnostic accuracy to allow readers to assess the potential for bias in a
study and to evaluate the generalizability of its results (Bossuyt et al., 2003; Bossuyt et al., 2004).

4. Reports to clinicians and patients should include the absolute risks associated with the given risk score, based on the results from the independent validation studies, described in lay language, e.g., “The analysis of this test in a clinical population revealed low risk patients have a probability of 92% of metastasis free survival at 5 years.” (FDA, 2007a)
Box A.

Current Determinants of Therapeutic Treatment

**National Institutes of Health (NIH) Consensus Development Criteria.**

In 2000, the NIH issued a consensus development conference statement on “Adjuvant Therapy for Breast Cancer.” This consensus panel statement was followed by an article by Eifel, et al. (2001) summarizing the statement. The statement addresses the following six areas: 1) factors used to select systemic adjuvant therapy; 2) For whom should adjuvant therapy be recommended; 3) For whom should adjuvant chemotheraphy be recommended; 4) For whom should post-mastectomy radiotherapy be recommended; 5) What side effects and quality-of-life issues factor into individual decision making about adjuvant therapy; 6) Promising new research directions for adjuvant therapy. The consensus panel made the following conclusions. Generally accepted prognostic and predictive factors include age, tumor size, lymph node status tumor histology and grade, mitotic rate, and hormone receptor status. Adjuvant hormonal therapy treatment decisions should be based on the presence of hormone receptors proteins in tumor tissues. Adjuvant polychemotherapy should be recommended for the majority of women with localized breast cancer. At the time of the statement there was little data available on the use of taxanes in the treatment of lymph node positive tumors. Additional evidence is available to support locoregional radiotherapy for high risk patients post-mastectomy.

**St. Gallen Guidelines.**

The 10th St. Gallen expert consensus meeting in march of 2007 refined it's target-oriented approach to adjuvant systemic therapy of early breast cancer. These guidelines focus on endocrine responsiveness combined with the presence or absence of human epidermal growth factor receptor-2 (HER2). There are three categories of endocrine responsiveness defined: highly endocrine responsive—tumor expresses high levels of both estrogen receptors (ER) and progesterone receptors (PgR) in the majority of cells; incompletely endocrine responsive—tumor expresses lower levels of ER or PgR, or lacks one of the hormone receptors; endocrine non-responsive—tumor has a complete absence of both ER and PgR. HER2 positivity is determined by either strong immunohistochemistry (IHC) staining or fluorescence in situ hybridization (FISH). A chart of current treatment recommendations can be found in Goldhirsch et al., 2007. Additional considerations of surgery and radiation therapy were also presented at the most recent conference, but were not subsequently added to the St. Gallen Guidelines.

**National Comprehensive Cancer Network (NCCN) Guideline.**

Compiled by a panel of twenty-six experts, the Breast Cancer Clinical Practice Guidelines present algorithms composed of decision trees and recommendations that address the clinical process from presentation and diagnosis to treatment and follow-up and are frequently updated to incorporate new evidence. Currently in version 2 for the year 2008, the guidelines include recommendations for a spectrum of disease including noninvasive and invasive breast cancer, as well as special considerations for Phyllodes tumor, Paget's disease, breast cancer during pregnancy, and inflammatory breast cancer. Breast cancers are classified into four groups for the purposes of treatment: 1) pure noninvasive carcinoma, 2) operable, local-regional invasive carcinoma, 3) inoperable, local-regional invasive carcinoma, and 4) metastatic or recurrent carcinoma. Therapeutic decisions depend on “tumor histology, clinical and pathologic characteristics of the primary tumor, axillary node status, tumor hormone receptor content, tumor HER2 status, presence or absence of detectable metastatic disease, patient comorbid conditions, patient age, and menopausal status.” The manuscript also emphasizes that patient preference is a major component of the decision making process as treatment options have not been optimized for any clinical situation and may offer equivalent survival rates.

**Adjuvant! Online.**

The purpose of this on-line tool is to help caregivers discuss risks and benefits of adjuvant therapy with patients who have the early stages of cancer. Adjuvant! Online developers wanted to address two problems: 1) patients not receiving enough information about their prognoses with and without adjuvant therapy; and 2) oncologists being uncertain about which quantitative estimates to apply to a patient. Details about how estimates of prognosis and estimates of the efficacy of can be found in Ravdin et al., 2001. There are two versions of Adjuvant! Online for breast cancer. The standard version is designed for decision making for patients with early breast cancer at the time of initial diagnosis and staging. The following variables are used to calculate survival (or relapse) at 10 years, deaths due to cancer in 10 years, and deaths due to other causes in 10 years: age, comorbidities, ER status, tumor grade, tumor size, and number of positive nodes. Additionally, effectiveness of adjuvant therapy options is
provided. The other version is designed for decision making for ER and/or PgR positive postmenopausal patients at the time of completing 5 years of adjuvant tamoxifen. The following variables are used to calculate no relapse after 5 years, relapse, and death from other causes in patients with no additional therapy after tamoxifen: age, comorbidities, ER status, tumor grade, tumor size, number of positive nodes, and 10-year initial risk.

**Professional society guidance**

**ASCO**

American Society of Clinical Oncology (ASCO) Tumor Marker Guidelines Committee. ASCO fist convened this panel in 1995, and it has systematically reviewed the clinical utility of putative markers for breast and colon cancer over the last one and one-half decades. The Committee also established a level of evidence scale for tumor markers that is widely cited for determination of whether a marker should or should not be used in clinic (Hayes et al. 1996). The most recently updated guidelines were published in 2006 for gastrointestinal (Locker et al. 2006) and breast cancers (Harris 2007). In the latter, the Committee recommended routine use of ER, HER2, the 21 gene recurrence score, and UPA/PAI-1.

**ASCO and College of American Pathologists (CAP) ad hoc Committee to Evaluate HER2.**

Following the reports of the remarkable benefits of adjuvant trastuzumab in HER2 positive breast cancer patients, ASCO and CAP partnered to jointly assess clinical and technical criteria for HER2 testing. This panel’s recommendations have been highly cited and are generally felt to have resulted in widespread improvement in HER2 testing, based in part on new proficiency testing required to achieve CAP accreditation (Wolff et al. 2007).
Box B.

Medical Device Classification Regulations.

**Class II Devices:**

Devices for which general controls alone are insufficient to assure safety and effectiveness, and existing methods are available to provide such assurances; these devices are exempt from premarket notification.

Class II medical devices intended for use as a gene expression profiling test system for breast cancer diagnosis are regulated under the Code of Federal Regulations, 21CFR866.6040 and require a Premarket Notification 510(k): The FDA “Guidance for Industry and FDA Staff: Class II Special Controls Guidance Document: Gene Expression Profiling test System for Breast Cancer Prognosis” provides guidelines for test kit manufacturers regarding preparation on premarket notifications and labeling. Specific and detailed information on how to gain marketing clearance from the FDA can be found on the FDA "Device Advice" web page, (http://www.fda.gov/cdrh/devadvice/3122.html), and from the FDA Guidance for Industry and FDA Staff: Interactive Review for Medical Device Submissions.

**Class III devices:**

Devices that support or sustain human life, are of substantial importance in preventing impairment of human health, or which present a potential, unreasonable risk of illness or injury. Due to the level of risk associated with Class III devices, FDA has determined that general and special controls alone are insufficient to assure the safety and effectiveness of class III devices.

Class III medical devices require a Premarket Approval (PMA): the FDA process of scientific and regulatory review to evaluate the safety and effectiveness of Class III medical devices. Detailed information on how to obtain a PMA can be found on the FDA "Device Advice" web page, (http://www.fda.gov/cdrh/devadvice/3122.html), and from the FDA Guidance for Industry and FDA Staff: Interactive Review for Medical Device Submissions.
Box C.

Minimum Information About a Microarray Experiment - MIAME 2.0
(www.mged.org/Workgroups/MIAME/miame_2.0.html)

The six following elements must be provided to support microarray based publications.

The raw data for each hybridization.

The raw data are defined as data files produced by the microarray image analysis software, such as CEL files for Affymetrix or GPR files for GenePix. These files should be provided in the native formats and should match their respective array designs.

The final processed data for the set of hybridizations in the experiment (study)

The processed data is defined as the normalized and/or summarized data on which the conclusions in the related publication are based. For instance, these can be MAS5 or RMA normalized data matrices for Affymetrix data. In gene expression experiments the final processed data is typically a matrix of genes and experimental conditions characterizing the expression of each gene under each condition. The identifiers used in these processed data files should match the array annotation or locations on the arrays.

The essential sample annotation, including experimental factors and their values

Experimental factors (conditions) and their values are the most essential information about the samples used in the experiment. The experimental factors are the key experimental variables in the experiment, for instance "time" in time series experiments, "dose" in dose response experiments, "compound" in compound treatment experiments, or "disease state" (normal or otherwise) in disease studies. The same experiment may have several experimental factors, for example, compound, dose and time may all be experimental factors in a dose response experiment in which several compounds are used to treat samples over a time course. In addition to experimental factor values, additional sample information that is required to interpret the experiment must be given, for instance, the organism and organism part from which the sample has been taken.

The experiment design including sample data relationships

The purpose of the experimental design description is simply to specify the essential relationships between different biomaterials, such as samples and arrays, and the data files which are produced in each hybridization. In a simple one channel one sample - one array experiment, this may be a table listing all samples and the respective raw data files. If relevant, it is important to show which hybridizations in the experiment are replicates, and which are technical and which are biological replicates. More generally, the experimental design can be described as a graph where nodes represent biomaterials (e.g., samples or their sources) and data objects (e.g., files), and edges or arrows show their relationships. MAGE-TAB provides a simple format to encode such graphs.

Sufficient annotation of the array design

Essential array design information is the reporter (probe) sequence information and/or the database accession numbers that characterize a sequence. For synthetic oligonucleotides the precise DNA sequence must be given. For commercial or other standard array platforms this information is typically provided by the array vendors or manufacturers.

Essential experimental and data processing protocols

The essential laboratory and data processing protocols are usually described in the journal methods section. It is sufficient to simply reference the standard experimental or data processing protocols, such as MAS5 or RMA. However, if a protocol depends on parameters that can be varied, their values should be provided. If novel or non-standard data processing protocols are used, these should be described in sufficient detail to allow the user to understand what exactly has been done in the experiment and how the data have been analyzed to reach the conclusions of the study.
Box D.

REMARK

REporting recommendations for tumor MARKer prognostic studies (REMARK) (McShane et al., 2006)

Introduction
1. State the marker examined, the study objectives, and any pre-specified hypotheses.

Materials and methods
Patients
2. Describe the characteristics (e.g., disease stage or co-morbidities) of the study patients, including their source and inclusion and exclusion criteria.
3. Describe treatments received and how chosen (e.g., randomized or rule-based).

Specimen characteristics
4. Describe type of biological material used (including control samples) and methods of preservation and storage.

Assay methods
5. Specify the assay method used and provide (or reference) a detailed protocol, including specific reagents or kits used, quality control procedures, reproducibility assessments, quantitation methods, and scoring and reporting protocols. Specify whether and how assays were performed blinded to the study endpoint.

Study design
6. State the method of case selection, including whether prospective or retrospective and whether stratification or matching (e.g., by stage of disease or age) was used. Specify the time period from which cases were taken, the end of the follow-up period, and the median follow-up time.
7. Precisely define all clinical endpoints examined.
8. List all candidate variables initially examined or considered for inclusion in models.
9. Give rationale for sample size; if the study was designed to detect a specified effect size, give the target power and effect size.

Statistical analysis methods
10. Specify all statistical methods, including details of any variable selection procedures and other model-building issues, how model assumptions were verified, and how missing data were handled.
11. Clarify how marker values were handled in the analyses; if relevant, describe methods used for cutpoint determination.

Results
Data
12. Describe the flow of patients through the study, including the number of patients included in each stage of the analysis (a diagram may be helpful) and reasons for dropout. Specifically, both overall and for each subgroup extensively examined report the numbers of patients and the number of events.
13. Report distributions of basic demographic characteristics (at least age and sex), standard (disease-specific) prognostic variables, and tumor marker, including numbers of missing values.

Analysis and presentation
14. Show the relation of the marker to standard prognostic variables.
15. Present univariate analyses showing the relation between the marker and outcome, with the estimated effect (e.g., hazard ratio and survival probability). Preferably provide similar analyses for all other variables being analyzed. For the effect of a tumor marker on a time-to-event outcome, a Kaplan–Meier plot is recommended.
16. For key multivariable analyses, report estimated effects (e.g., hazard ratio) with confidence intervals for the marker and, at least for the final model, all other variables in the model.
17. Among reported results, provide estimated effects with confidence intervals from an analysis in which the marker and standard prognostic variables are included, regardless of their statistical significance.
18. If done, report results of further investigations, such as checking assumptions, sensitivity analyses, and internal validation.

Discussion
19. Interpret the results in the context of the pre-specified hypotheses and other relevant studies; include a discussion of limitations of the study.
20. Discuss implications for future research and clinical value.
Box E.

The STARD Criteria

1. Is this a study of diagnostic test (gene expression profiling test) accuracy?
2. Research question:
   a. What are the study aims?
      i. estimation of diagnostic accuracy
      ii. comparison of diagnostic accuracy between tests
      iii. Comparing diagnostic accuracy across patient populations (participant groups)
3. Who are the study participants?
   a. Population characteristics
   b. inclusion/exclusion criteria
   c. setting and location of data collection (primary care, tertiary care)
   d. is the diagnostic test being used for screening, or confirmation of diagnosis, guidance on treatment
4. Patient recruitment
   How were the eligible participants identified?
   a. consecutive recruitment, randomized recruitment, cohort from a previous study, etc.
   b. have patients been diagnosed with the target disease
   c. are all participants evaluated by the gene expression profiling test and an index test (Adjuvant!, St. Gallen criteria, IHC)
5. Patient sampling?
6. Data Collection
   a. was data collection planned before the gene expression profiling test and the reference standard were performed?
   b. was data collection performed after patients underwent the gene expression profiling test and the reference standard was performed?
      i. retrospective data collection may better reflect routine practice
      ii. retrospective data collection may not identify all eligible participants
      iii. retrospective studies may not provide high-quality data
7. Reference standard description
   a. How/why was this standard chosen?
8. Technical description/specifications of materials and methods.
   a. How were methods and execution of both the gene expression profiling text and the reference standard conducted (or cite references)?
   b. variations in execution of tests will/can be sources of diagnostic variation in accuracy
9. What is the definition and rationale for the results of the gene expression profiling test and the reference standard?
   a. include cut-offs for results
10. Who is executing and evaluating the tests?
   a. include number and expertise
   b. variability in the way tests are processed and read will impact the diagnostic accuracy.
11. Blinding?
12. How is diagnostic accuracy measured?
13. What methods are used to calculate test reproducibility?
   a. factors that influence reproducibility: observer variability or imprecision, analytic method variability.
14. When was study done?
   a. include beginning and ending dates of recruitment
15. What are the clinical and demographic characteristics of the study population?
16. How many participants qualifying for the study did not participate?
   a. may want to track using a flow diagram.
   b. diagnostic accuracy will be biased of the result of the gene expression profiling test influences the decision to order a reference standard
17. What is the time interval between the application of the gene expression profiling test and the reference standard?
   a. these tests should be performed at the same time.
18. What is the distribution of the severity of the disease in question?
   a. also include severity of non-target conditions
   b. comorbid conditions may produce false-negative or positive results
19. Have the gene expression profiling test been cross-tabulated with the reference standard?
20. Were any adverse events from performing the gene expression profiling test or reference standard reported?
21. Are measures of diagnostic accuracy reported as measures of statistical uncertainty?
   a. how well do the test results correspond with the target condition (presence or absence) as established by the reference standard?
22. How were indeterminate results, missing responses, and outliers handled?
   a. include reasons why a test’s results may have been uninterpretable—such as technical test failure, comorbid conditions, treatment of comorbid conditions.
   b. intermediate test results should be reported—they may have diagnostic value.23.
23. Are the estimates of variability of diagnostic accuracy between subgroups, test readers, etc., reported?
24. What are the estimates of test reproducibility?
   a. example, is there inter-observer variability?
25. What is the clinical applicability of the study findings?
   a. results of the study may not apply to decision making in problem/question of interest.
   b. point out differences of study and other settings
Gene Expression Profile Tests for Early Stage Breast Cancer

Bibliography


FDA. Ensuring the Safety of Marketed Medical Devices. CDHR’s Medical Device Postmarket Safety Program. 2006a.


FDA. Guidance for Industry and FDA Staff Commercially Distributed Analyte Specific Reagents (ASRs): Frequently Asked Questions. 2007b.


 Locker GY, Hamilton S, Harris J et al. ASCO 1996 update of recommendations for the use of tumor markers...


# Appendix A

## Advisory Group Members

<table>
<thead>
<tr>
<th>Stakeholder Group</th>
<th>Name</th>
<th>Organizational Affiliation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product industry</td>
<td>P. Terry</td>
<td>Genomic Health</td>
</tr>
<tr>
<td>Product industry</td>
<td>Xiao-Jun Ma</td>
<td>AviaraDx (H/I test)</td>
</tr>
<tr>
<td>Product industry</td>
<td>Laura van’t Veer</td>
<td>Agendia (MammaPrint™)</td>
</tr>
<tr>
<td>Product industry</td>
<td>Steve Teutsch</td>
<td>Merck</td>
</tr>
<tr>
<td>Former Payer</td>
<td>Bruce Quinn</td>
<td>Formerly w/ National Heritage Insurance Company</td>
</tr>
<tr>
<td>Payer</td>
<td>Gerald Peden</td>
<td>Independence Blue Cross</td>
</tr>
<tr>
<td>Payer</td>
<td>Mitchell Burken</td>
<td>Mid-Atlantic Medicare Medical Director Trailblazer Health Enterprises</td>
</tr>
<tr>
<td>Payer</td>
<td>Derek van Amerongen</td>
<td>Humana of Ohio IOM Committee on Technologies for the Early Detection of Breast Cancer</td>
</tr>
<tr>
<td>Clinician</td>
<td>Laura Esserman</td>
<td>University of California, San Francisco</td>
</tr>
<tr>
<td>Clinician-Researcher</td>
<td>Daniel Hayes</td>
<td>University of Michigan, Clinical Director, Breast Oncology Program SW Oncology Group TAILORx Trial Lead ASCO Representative</td>
</tr>
<tr>
<td>Clinician-Oncologist</td>
<td>Eva Thomas</td>
<td>Kaiser Permanente NCR Hematology/Oncology</td>
</tr>
<tr>
<td>Patient Advocacy Organization Rep</td>
<td>Musa Mayer</td>
<td>AdvancedBC.org</td>
</tr>
<tr>
<td>Patient Advocacy Organization Rep</td>
<td>Karen Sepucha</td>
<td>Foundation for Informed Medical Decision-Making</td>
</tr>
<tr>
<td>Patient Advocacy Organization Rep</td>
<td>Sharon Terry</td>
<td>Genetic Alliance</td>
</tr>
<tr>
<td>Patient Advocate</td>
<td>Helen Schiff</td>
<td>National Coalition on Breast Cancer</td>
</tr>
<tr>
<td>Epidemiologist</td>
<td>Laurie Habel</td>
<td>Kaiser Permanente DOR Oncotype DX Study PI</td>
</tr>
<tr>
<td>Methodologist</td>
<td>David L Veenstra</td>
<td>Pharmaceutical Outcomes Research and Policy Program; Institute for Public Health Genetics University of Washington</td>
</tr>
<tr>
<td>Methodologist</td>
<td>Maggie Piper</td>
<td>Blue Cross Blue Shield Association TEC EGAPP</td>
</tr>
<tr>
<td>Methodologist</td>
<td>Steve Pearson</td>
<td>Institute for Clinical and Economic Review</td>
</tr>
<tr>
<td>Observer</td>
<td>Kristin Brinner</td>
<td>Health and Human Services, Office of Personalized Health</td>
</tr>
<tr>
<td>Observer</td>
<td>Sue Levine</td>
<td>Hayes, Inc.</td>
</tr>
<tr>
<td>Observer</td>
<td>Linda Bradley</td>
<td>Centers for Disease Control National Office of PH Genomics</td>
</tr>
<tr>
<td>Observer</td>
<td>Elizabeth Mansfield</td>
<td>US Food and Drug Administration</td>
</tr>
<tr>
<td>Observer</td>
<td>Gurvaneet Randhawa</td>
<td>Agency for Healthcare Research and Quality</td>
</tr>
<tr>
<td>Observer</td>
<td>Jeff Roche</td>
<td>Centers for Medicare and Medicaid Services</td>
</tr>
</tbody>
</table>
Appendix B

Current Technical Documents
Updated November 17, 2008

FDA Guidance for Industry:

#229

This guidance is intended to be used as a companion to the guidance Pharmacogenomic Data Submissions (March 2005). It reflects experience gained since the issuance of that guidance with voluntary genomic data submissions as well as with review by the FDA of numerous protocols and data submitted under investigational new drug (IND) applications, new drug applications (NDAs), and biologics license applications (BLAs). This document details specific methodological issues (analytic validity) for developers when considering submitting gene expression data from microarrays. Additional information beyond that requested in this report may be required for microarray data supporting the clearance or approval of a diagnostic device.

# 334
Reports on the Status of Postmarketing Study Commitments. FDA Guidance for Industry: February 2006

This report contains procedures for submitting a postmarketing study commitment status reports for an approved human drug or licensed biological product. The FDA can require post-marketing reporting from a manufacturer if there are concerns about clinical safety.

FDA Guidance for Industry and FDA Staff:

#290

“The Secretary of Health and Human Services has delegated to FDA the authority to determine whether particular tests are "simple" and have "an insignificant risk of an erroneous result" under CLIA and thus eligible for waiver categorization (69 FR 22849, April 29, 2004). The Centers for Medicare & and Medicaid Services (CMS) is responsible for oversight of clinical laboratories, which includes issuing waiver certificates. CLIA requires that clinical laboratories obtain a certificate before accepting materials derived from the human body for laboratory tests. 42 U.S.C. § 263a(b). Laboratories that perform only tests that are "simple" and that have an "insignificant risk of an erroneous result" may obtain a certificate of waiver. 42 U.S.C. § 263a(d)(2)."

This document outlines approaches recommended by the FDA to demonstrate that a device is simple and poses insignificant risk due to erroneous results. Other approaches to demonstrate this are allowable.

# 289
Class II Special Controls Guidance Document: Gene Expression Profiling Test System for Breast Cancer Prognosis. FDA Guidance for Industry and FDA Staff. May 9, 2007

This document was developed as a guidance to support the classification of gene expression profiling test systems for breast cancer prognosis into class II (special controls). It provides recommendations to manufacturers regarding preparation of premarket notifications and labeling for a gene expression profiling test system for breast cancer prognosis; applicable to RNA expression assays such as RT-PCR and microarrays. It is issued in conjunction with a Federal Register notice announcing the classification of gene expression profiling test systems for breast cancer prognosis. This guidance document identifies the classification regulation and product code for gene expression profiling test system for breast cancer prognosis.
A gene expression profiling test system for breast cancer prognosis is a test that may require instrumentation for clinical multiplex test systems. Instrumentation for clinical multiplex test systems is regulated under 21 CFR 862.2570. Guidance for such instrumentation is available in the FDA Guidance for Industry and FDA Staff, "Class II Special Controls Guidance Document: Instrumentation for Clinical Multiplex Test Systems."

#228

This guidance document describes the roles of both FDA and industry (applicants) in an interactive review process for 510(k)s, original PMAs, PMA supplements, original BLAs, and BLA supplements.

The FDA can collect user fees and use these to establish the framework for a more aggressive set of performance goals in the review of medical device submissions. A formalized interactive review process to encourage and facilitate communication between FDA staff and industry during the review of specific medical device premarket submissions to help accomplish the following has been established to: improve the interaction between the FDA review staff and the; prevent unnecessary delays in the completion of the review; try to ensure that FDA’s concerns are clearly communicated to the applicant during the review process; minimize the number of review cycles; minimize the number of review questions conveyed through formal requests to applicants for additional information; and ensure timely responses from applicants.

#381
Class II Special Controls Guidance Document: Instrumentation for Clinical Multiplex Test Systems. FDA Guidance for Industry and FDA staff. March 10, 2005

This guidance document was developed as a special controls guidance to support the classification of instrumentation for clinical multiplex test systems into class II (special controls). This type of device is intended to measure and sort multiple signals generated by an assay from a clinical sample.

FDA DRAFT Guidance

#271
In Vitro Diagnostic Multivariate Index Assays. FDA Draft Guidance for Industry, Clinical Laboratories, and FDA Staff. September 7, 2006

This guidance addresses the definition and regulatory status of a class of In Vitro Diagnostic Devices referred to as In Vitro Diagnostic Multivariate Index Assays (IVDMIAs). Regulatory classifications are driven by intended use(s) and device risk. There has been some confusion about IVDMIAs associated with the Analyte Specific Reagent (ASR) rule which classifies and regulates ASRs that move in commerce. The rule does not extend to tests developed in-house by clinical laboratories using commercially available ASRs and used exclusively by that laboratory or ASRs created in-house and used exclusively by that laboratory for in-house testing. For purposes of this guidance, IVDMIAs are test systems that employ data, derived in part from one or more in vitro assays, and an algorithm that usually, but not necessarily, runs on software to generate a result that diagnoses a disease or condition or is used in the cure, mitigation, treatment, or prevention of disease. The guidance also addresses premarket pathways and postmarket requirements with respect to IVDMIAs.

#226

This guidance document is written to help the manufacturer, sponsor, applicant, investigator and the IVD device industry in general in the development of IVD studies, particularly those exempt from most of the requirements of the IDE regulation and to provide a broad regulatory framework pertaining to the development phase of IVD devices. The document is intended to facilitate the movement of new IVD technology from the investigational stage to the marketing stage. The Center for Devices and Radiological Health (CDRH) and the Center for Biologics Evaluation and Research (CBER) each have regulatory responsibilities for IVD devices; information included in this document applies to Class I, II, and III IVD devices regulated by either Center.
Other Government Documents and Recommendations

# 236

The goal of this advisory committee was to undertake the development of a comprehensive map of the steps needed for evidence development and oversight for genetic and genomic tests, with improvement of health quality as the primary goal. This report focuses on the oversight of genetic testing and the application of genetic information in patient care and management.

#382
The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) initiative: methods of the EGAPP working group.

The Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Initiative, was established by the National Office of Public Health Genomics at the CDC. It supports the development and implementation of an evidence-based process for evaluating genetic tests and other genomic applications for clinical and public health practice in the United States. A EGAPP Working Group (EWG), a multidisciplinary expert panel selects topics, oversees the systematic review of evidence, and makes recommendations based on that evidence. This article describes the EGAPP processes and details the specific methods and approaches used by the EWG. Key objectives of the EWG are to develop a transparent, publicly accountable process, minimize conflicts of interest, optimize existing evidence review methods to address the challenges presented by complex and rapidly emerging genomic applications, and provide clear linkage between the scientific evidence and the subsequently developed EGAPP recommendation statements. This document does provide a table of the hierarchies of clinical trial and observational study design for the assessment of these tests but does not attempt to determine what sort of information is minimally sufficient to draw conclusions on the clinical effectiveness and utility of a test.

This article is an overview of the topics EGAPP has looked at or will look at. It discusses who is doing the reviews (such as ours) and how they oversee them. Very little outside of what we know about the process already is in this document. There is some wording about looking into “less expensive” and less time-consuming ways to conduct these reviews. The main end-users for this article are meant to be physicians and additional partnerships will need to be created to develop evidentiary standards and to build additional evidence review capacity.

# 374

This report discusses the CDRH medical device postmarket safety framework and the approaches used to monitor and address adverse events and risks associated with the use of medical devices that are currently available in the market. Primary goals of the postmarket programs are: (1) to be aware of and have access to any surveillance data 2) Establish partnerships and alliances throughout; (3) maintain an on-site enforcement inspection and assessment presence throughout the medical device manufacturing; (4) communication of significant medical device risk in a timely and appropriate manner; (5) build postmarket learning into premarket device assessment; (6) Identify and communicate examples of excellence and best practice; (7) Build and manage information and knowledge systems that support regulatory and public health responsibilities; (8) continued training of staff who will be skilled and knowledgeable with future medical device issues and priorities.

Other Documents

#372
The Clinical and Laboratory Standards Institute (CLSI) is a global, nonprofit, standards-developing organization that promotes the development and use of voluntary consensus standards and guidelines within the health care community (www.clsi.org). Their guideline MM17-A, “Verification and Validation of Multiplex Nucleic Acid Assays; Approved Guideline,” provides broad guidance on evaluating and establishing a test’s (applicable to gene expression profiling tests) analytic validity. The guideline is especially useful in providing definitions of terms and suggesting a validation pathway for developers, although it does not include procedures to ascertain health benefit.