Drug-Diagnostic Co-Development Concept Paper

Draft — Not for Implementation

Department of Health and Human Services (HHS)F
Food and Drug Administration (FDA)
April 2005

TABLE OF CONTENTS

Drug-Diagnostic Co-Development
Concept Paper

1. INTRODUCTION, BACKGROUND, AND SCOPE

This concept paper reflects preliminary Agency thoughts on how to prospectively co-develop a drug or biological therapy (*drugs*) and device test in a scientifically robust and efficient way. The thoughts and recommendations contained here are being put forward for discussion purposes only. The Agency intends to solicit initial input from the public on this concept paper, then develop a draft guidance for public comment according to the good guidance practices regulation (21 CFR 10.115).

1.1 Introduction

Drug/test combinations have the potential to provide many clinical benefits to patients including differential diagnosis of a disorder or identification of a patient subset, identification of potential responders to a specific drug, a way to target therapy, an approach to identifying individuals at risk for adverse events, an adjunct tool for monitoring responses to drugs, and a way to individualize therapy.

Co-development is an area of rapidly evolving technology and targeted therapy that may involve regulation of products across the FDA centers (e.g. the Center for Drug Evaluation and Research (CDER), the Center for Devices and Radiological Health (CDRH) and/or the Center for Biological Evaluation and Research (CBER), the Office of Combination Products (OCP)). This document contains general ideas on both process and scientific issues to be considered in the co-development of drugs in which a new diagnostic test may play a critical role in the clinical use of the drug.[1]

1.2 Background

The use of diagnostics to help select drug therapy is a well-established technique. As early as 1972, receptor hormones for estrogen were identified as valuable markers for selecting candidates for hormonal treatment in women with breast cancer (1). Since then, FDA has

[1] FDA has recently published two related guidances to help clarify options available for using new technologies in decision making in drug development and/or clinical use. In April 2003, CDRH published guidance for Multiplex Tests for Heritable DNA Markers, Mutations and Expression Patterns; Draft Guidance for Industry and FDA Reviewers (1). Comments have been received and the document is being revised. In March 2004, CDER published the final version of the guidance for industry *Pharmacogenomic Data Submissions*. See also the FDA Genomic Web page at http://www.fda.gov/cder/genomics/default.htm.

approved a small number of additional tests of this type, notably assays for progesterone receptor, Her 2 Neu DNA and protein, and for Epidermal Growth Factor Receptor protein. As a result of new technologies (most notably multiplex technologies) and as a result of increased information on the human genetic map and drug targets, interest in biomarkers based on pharmacogenomic information has been growing rapidly. These developments offer the opportunity for increased understanding of human biology, disease, and drug effects. Of particular interest is the ability of pharmacogenomic based tests to identify sources of inter-individual variability in drug response (both efficacy and toxicity) and that might be used to guide drug selection and target therapy to selected patient subsets.

In the field of pharmacogenomics — as is typical with a rapidly evolving science — the experimental results (e.g., biomarker validation data) have not always been well enough established scientifically to be suitable for regulatory decision making. To this end, the Agency has undertaken a series of public meetings (May 2002, November 2003, July 2004, and planned meeting on April 11) (2-9) to obtain input on relevant issues from the scientific community and interested stakeholders. In addition, an FDA docket was made available to provide an opportunity for public input on this topic (9).

The policies and processes outlined in this concept paper are intended to take the above factors into account and to assist in advancing the field of pharmacogenomics in a manner that will benefit both drug development programs and the public health.

1.3 Scope

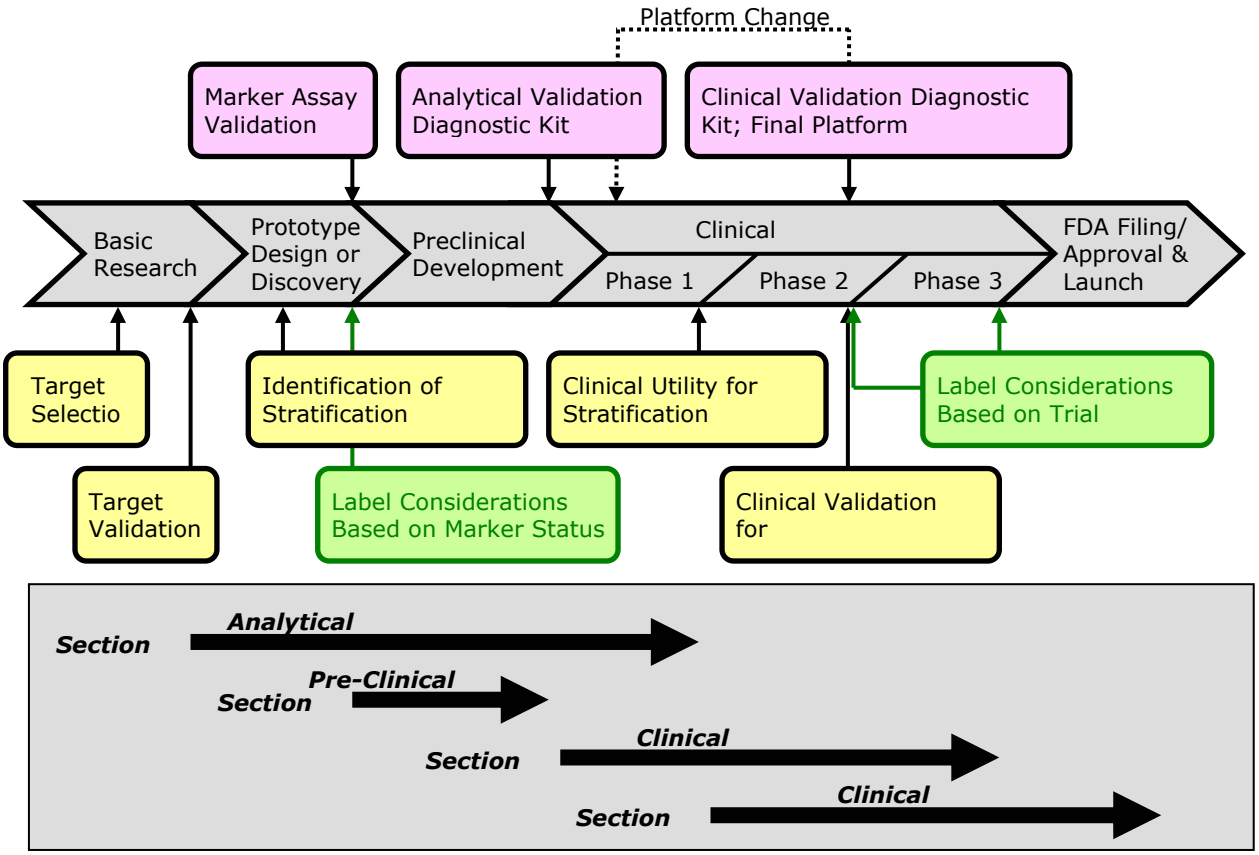
This document addresses issues related to the development of in vitro diagnostics (see glossary for definition of terms) for mandatory use in decision making about drug selection for patients in clinical practice. Both the test and the drug would be used in the clinical management of the patient. The diagnostic tests being considered in this context may be used to identify patients most likely to respond to a drug, patients most likely to fail to respond to a drug, and/or patients most likely to exhibit adverse events that might contraindicate drug administration. The tests may also be valuable as optional tests during the drug development process to assist in understanding mechanisms of a disease or in determining how to enrich or select patient populations to conducting more rapid and predictable clinical trials for new therapies.

This document addresses development of a single test in conjunction with a single drug. Figure 1, on the next page, depicts key steps during co-development. Particular attention should be paid to the status of a biomarker (i.e. exploratory or valid, detailed in the guidance *Pharmacogenomic Data Submissions*). The status of a biomarker can influence the stratification measures, clinical utility and validation, and, therefore, the label of the co-developed product. Figure 1 also puts in perspective sections 3 to 6 of this document along the development path of such products.

This document does not specifically address issues related to pharmacogenomic testing for the purposes of drug dosing determinations or monitoring of drugs, although it does contain principles that may be relevant to the development of these types of tests.

Figure 1. Drug-Device Co-Development Process: Key Steps During Development.

Of particular note are the label considerations based on the status of the marker used for stratification. Clinical validation of the marker has a direct influence on the clinical utility and therefore on the label of the co-developed product.



This paper does not cover other scenarios, such as the use of one test (e.g., CYP2D6 alleles) with multiple drugs or of several tests developed for serial or parallel use with a single drug. Furthermore, this paper does not address optional or exploratory tests that are not intended for further development or those that do not affect the results of clinical trials (e.g. those that are used in understanding mechanisms of disease).

The term *pharmacogenomics* is defined here as the use of a pharmacogenomic or pharmacogenetic test (see glossary for definitions) to be used in conjunction with drug therapy.

Among the important considerations discussed in this paper are:

- Review procedure issues: This section describes processes and procedures for submitting and reviewing a co-developed drug-test product.
- Analytical test validation: This section describes the in-vitro ability to accurately and reliably measure the analyte of interest, including analytical sensitivity and specificity, and focuses on the laboratory component of drug/test development.
- Clinical test validation: This section describes the ability of a test to detect or predict the associated disorder in patients and includes clinical sensitivity and specificity, and/or other performance attributes of testing biological samples.
- Clinical test utility: This section describes elements that should be considered when evaluating the patient risks and benefits in diagnosing or predicting efficacy or risk for an event (drug response, presence of a health condition).

2. REVIEW PROCEDURE ISSUES

2.1 Co-development and Intercenter Review Considerations

Co-developed products that would be used together may or may not be combination products as defined in 21 CFR 3.2(e).[2] FDA anticipates that many therapeutic drug and diagnostic test products will be marketed separately. For the purposes of this document, *co-development* refers to products that raise development issues that affect both the drug therapy and the diagnostic test, regardless of their regulatory status as a combination product or as a noncombination product. For example, when co-developed products are considered together, unique questions may arise that would not exist for either product alone. Scientific or technologic issues for one product alone may be minimal, but they may have substantial implications for the other product. Also, postapproval changes in one may affect the safety and effectiveness of the other. Subsequent sections of this document address some of these product development considerations.

---

[2] Under 21 CFR 3.2 (e), a combination product is defined to include:

(1) A product comprised of two or more regulated components, i.e., drug/device, biologic/device, drug/biologic, or drug/device/biologic, that are physically, chemically, or otherwise combined or mixed and produced as a single entity;
(2) Two or more separate products packaged together in a single package or as a unit and comprised of drug and device products, device and biological products, or biological and drug products;
(3) A drug, device, or biological product packaged separately that according to its investigational plan or proposed labeling is intended for use only with an approved individually specified drug, device, or biological product where both are required to achieve the intended use, indication, or effect and where upon approval of the proposed product the labeling of the approved product would need to be changed, e.g., to reflect a change in intended use, dosage form, strength, route of administration, or significant change in dose; or
(4) Any investigational drug, device, or biological product packaged separately that according to its proposed labeling is for use only with another individually specified investigational drug, device, or biological product where both are required to achieve the intended use, indication, or effect.

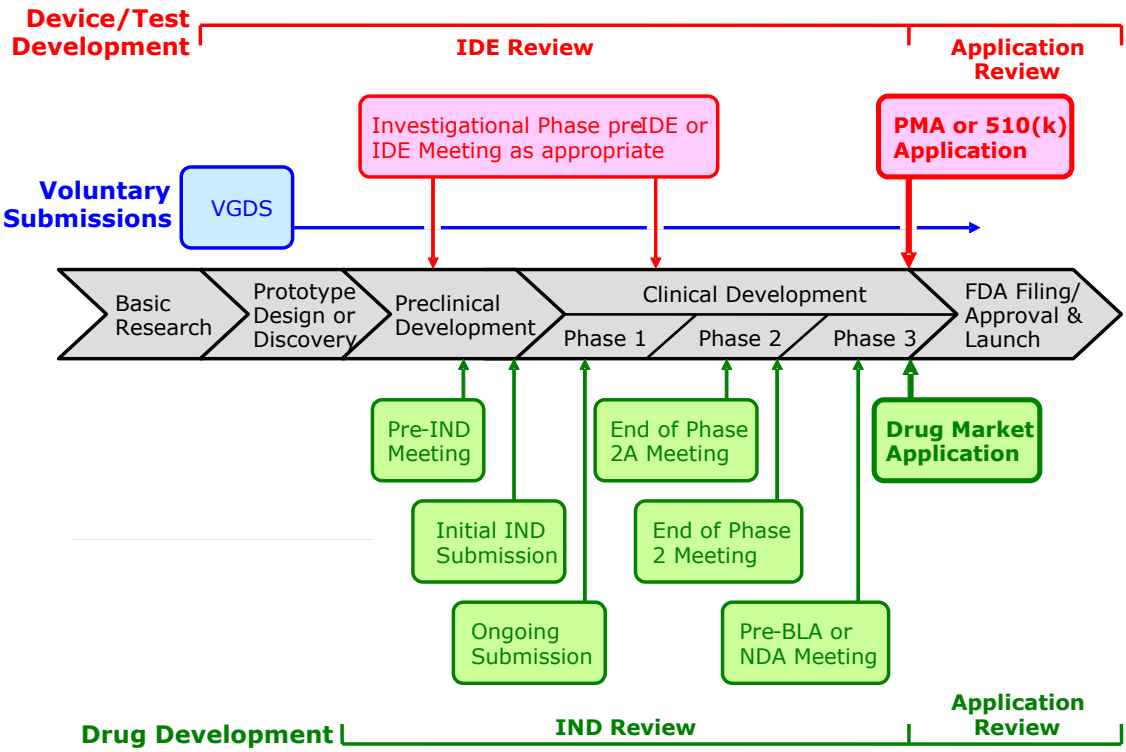
2.2 Procedures

The parallel development of a drug and a diagnostic is a relatively new aspect of drug development and calls for careful coordination. Figure 2 illustrates approximate time points during the drug development process for a noncombination product at which formal industry-FDA interactions normally take place. For additional information on combination products and the intercenter review process, see the Office of Combination Products Website at http://www.fda.gov/oc/combination.

Voluntary submissions (i.e. Voluntary Genomic Data Submissions, VGDS), a new approach introduced in the guidance for industry *Pharmacogenomic Data Submissions*, can be used throughout this development process to present and discuss data with the Agency that are not used for regulatory decision making, but could have an effect on the overall development strategy. Such voluntarily submitted data will not be used for regulatory decision making by the FDA and is not included in the evaluation of an IND, IDE, or market application.

The co-development pathway for the in vitro diagnostic should be determined early in development. FDA recommends that sponsors seek discussions with FDA that involve the reviewing centers, the OCP, and the manufacturers of both the diagnostic and the therapeutic drug, as appropriate. These pre-IND/IDE processes are outlined in existing FDA guidance documents.

Figure 2. Drug-Device Co-Development Process: Formal Industry-FDA Interactions
(Noncombination Product Example)



In preliminary discussions with FDA about new submissions with or without use of the pre-IDE process, sponsors may want to consider questions such as the following:

(1) When and how should the diagnostic test be validated analytically and clinically and what constitutes validation in the context of proposed use?

(2) What additional information is needed for information previously submitted under a VGDS if a VGDS becomes a required submission?

(3) What analytical and feasibility test data on the diagnostic are recommended before beginning clinical studies and when should such data be obtained?

(4) What analytical and clinical data are needed to support prespecified retrospective development and validation of a diagnostic test?

(5) What analytical and clinical attributes of diagnostic tests can be validated in one protocol and what characteristics will need separate protocols?

(6) What is the most appropriate regulatory pathway for co-development? Is the product apt to be a combination product or noncombination product?

(7) If it is not a combination product, is sequential or simultaneous approval most appropriate?

Biomarkers for drug selection, with the exception of the estrogen receptor test, were not addressed in the classification of in vitro diagnostic devices promulgated by FDA during the late 1970s and early 1980s. As a result, few if any appropriate predicates exist (see glossary) for use for this class of diagnostic devices. FDA would expect many of these products — in particular those with high risk profiles — to be processed as class III products subject to premarket review under the premarket application approval (PMA) process (http://www.fda.gov/cdrh/pmapage.html).

Additional discussion of the number of investigational and marketing applications for combination products goes beyond the scope of this paper (10). FDA intends to develop guidance on this topic.

3. ANALYTICAL TEST VALIDATION

3.1 General Recommendations to Support Premarket Review

The following general recommendations are for analytical studies to support premarket review of the analytical quality of commercially distributed test kits.

A major hurdle for the co-development of a diagnostic test with a drug is the importance of obtaining and securing adequate specimens from patients in the clinical trials that can be used as evidence of drug efficacy and/or safety. When possible, we recommend that a diagnostic test for subsequent pivotal efficacy and/or safety studies be developed and analytically validated early in the drug development process to allow clinical test validation and clinical test utility

determination during the late stage clinical trials. Study design should take into account statistical considerations for both the drug and the diagnostic. Clinical trial specimens should be banked in optimal storage conditions to enable subsequent test development and/or retrospective hypothesis generation or confirmation of test performance.

3.2 Device Description

It is recommended that careful characterization of device platforms for all relevant design elements be included in all test development programs. The test system's methodology for detecting the analytes of interest should be described in detail with design elements relevant to optimization of the test system characterized appropriately. For additional information see Addendum A.

If the test kit includes reagents for sample preparation, there should be a description of the methodology and specimen preparation.

Illustrations or photographs of nonstandard equipment or methods can be helpful in understanding novel methodologies and any approaches to risk management.

3.3 Analytical Studies

Analytical validation studies are recommended to evaluate the following performance characteristics of the assay, where applicable, for each analyte claimed in the clinical use statement. A complete description of each study should be provided, including protocol and results, to adequately interpret the study outcomes. For additional information, see Addendum B. Some important considerations in analytical validation are listed below. These are not intended to be prescriptive, rather to give an overview of the types of information that evolve from analytical studies of test validation.

(1) Studies to show that test performance can be applied to expected clinical use as a diagnostic with acceptable accuracy, precision, specificity and sensitivity: A demonstration of the device's ability to accurately and reproducibly detect the analyte(s) of interest at levels that challenge the analyte concentration specifications of the device should be provided. (See number 3 below).

(2) Sample requirements: All relevant criteria and information on sampling collection, processing, handling and storage should be clearly outlined.

(3) Analyte concentration specifications: It is recommended that, when appropriate, a range of analyte concentrations that are measurable, detectable, or testable be established for the assay.

(4) Cut-off: It is recommended that there be a clear rationale to support an analytical characterization of cut-off(s) value(s).

(5) Controls and calibrators: All external and process controls and calibrators should be clearly described and performance defined.

(6) Precision (Repeatability/Reproducibility): All relevant sources of imprecision should be identified and performance characteristics described.

(7) Analytical specificity (interference and cross reactivity studies): Cross-reactive and interfering substances should be identified and their effect on performance characterized.

(8) Assay conditions: The reaction conditions (e.g. hybridization, thermocycling conditions), concentration of reactants, and control of nonspecific activity should be clearly stated and verified.

(9) Sample carryover: The potential for sample carryover and instructions in labeling for preventing carryover should be provided.

(10) Limiting factors of the device should be described, such as when the device does not measure all possible analyte variations, or when the range of variations is unknown.

3.4 Software and Instrumentation

3.4.1 Data processing

If the device includes software, there should be specific information about the software in the test submission. It is recommended that computational methods be developed and verified using the CDRH software development and validation guidance documents that are available at http://www.fda.gov/search/databases.html. Evidence should be provided that the software has met all necessary verification tests. If applicable, computational concerns that are addressed by the software should be described, such as probe saturation level, background correction, and normalization.

3.4.2 Validation of instrumentation

If the device can be used on a generic platform (e.g., a generic thermocycler), specifications should be provided in the labeling so that the user may select an instrument that is suitable for their purposes. If the device includes proprietary instrumentation, whether manufactured by the sponsor or by another company, specific information about the instrument(s) should be included in the submission. It is recommended that the following general attributes be addressed in validating instrumentation:

- A characterization of the instruments used in the device: We recommend that information on how the instrument assigns values to or interprets assay variables (e.g., feature location, size, concentration, volume, drying of small samples, effect of small volume reactions) be included along with its impact on test results.
- An explanation for how the instrument is calibrated and what materials are used during calibration.
- Uncertainties should be included that describe and quantify potential sources of error in results introduced by hardware components (e.g., scanners).

If a particular instrument is specified (by manufacturer or brand), there should be assurance that any changes made to the instrument (by the sponsor or the manufacturer) are tracked throughout analtyical development. If changes introduce new or different assay performance issues, the sponsor should be responsible for validation of the device under the changed conditions.

3.5 Analytical Validation of Changes to a Device in Late Stages of Development

In some cases, the device configuration used during certain drug trials for efficacy and safety may not be ideal for commercial use in clinical practice. Major changes to a device platform can be validated using an independent prospective clinical data set, or by testing retrospectively banked specimens from the original studies. The stability and validity of using banked samples should be documented by demonstrating that the original assay results can be repeated at the time when the new assay results are obtained from the specimens. It is also recommended that the FDA review the validation protocol for the new or modified assay prior to beginning new clinical studies.

For smaller, or more defined modifications to device configuration, analytical studies alone may suffice to validate these changes. For example, if the specimen or sample storage conditions used during development and validation of the device (e.g., rapid freezing of samples) turn out to be impractical in a clinical setting (i.e., settings where only refrigeration is available), analytical validation of new storage conditions for patient specimens and processed samples may be acceptable**.**

3.6 Analytical Considerations for Specific Types of Diagnostic Products

If a multi-analyte diagnostic test (e.g., a gene expression array) is used, the degree of analytical validation will depend on the number of features or readouts represented on the test. If the feature number is relatively low (e.g., 2 to 10), each feature can be validated (depending on the system). However, it is infeasible to verify each feature in a test containing, for example, 100,000 features. In that case, typical measures (e.g., accuracy, precision, analytical specificity and analytical sensitivity) of the assay may be studied using the system as a whole to prove the validity of the diagnostic test.

In many cases, particularly in the cases of patient stratification (e.g., for drug efficacy improvement), it is anticipated that relatively simple diagnostic tests measuring just a few analytes simultaneously, derived from probing the patient population with highly multiplexed assays, can be used. Statistical considerations in deriving a small number of biomarkers from a large amount of parallel multiplexed data should be properly addressed. A new test with fewer biomarkers developed for diagnostic purposes (i.e., patient stratification) should be properly validated, ideally in clinical trials that enrolled patients with the intended indication. When validating a gene or expression *pattern,* instead of a set of individual biomarkers, a rigorous statistical approach should be used to determine the number of samples, and the methodology used for validation. It is recommended that the validation strategy be discussed in advance with FDA.

3.7 Resources for Software Submissions

FDA has published guidances on general principles of software validation, such as content of premarket submissions for software contained in medical devices and off-the-shelf software use in medical devices. In addition, the American National Standards Institute (ANSI)/Institute of Electrical and Electronics Engineers, Inc. (IEEE) has developed 21 standards describing software design/validation requirements that may be of interest to drug-test co-developers.[3]

4. PRECLINICAL PILOT FEASIBILITY STUDIES

4.1 Introduction

After a new diagnostic test has been analytically characterized, additional studies should be performed to determine clinical validation. Optimally, these studies will be performed based on information known from analytical studies and based on pilot studies or careful analysis to determine relevant populations to be studied to establish clinical test performance and target cut-off points in biological specimens.

Ideally, a new diagnostic intended to inform the use of a new drug will be studied in parallel with early drug development (phase 1 or 2 trials) and diagnostic development will then have led to prespecification of all key analytical and clinical validation aspects for the subsequent (late phase 2 and phase 3) clinical studies. These include the intended population and selection of diagnostic cut-off points for the biomarker intended to delineate test positives, test negatives, and, when appropriate, equivocal zones of decision making.

4.2 Prespecification of Assay Cutoffs

The cutoff that defines test positive and test negative results should be selected prior to performing the pivotal clinical drug/diagnostic study or studies that provide evidence of adequate clinical test validation and clinical utility. Estimates of performance can be severely biased when test cut-offs are chosen post-hoc to optimize test performance. This is because if the cut-off of the in vitro test is chosen post-hoc using a point to maximize clinical accuracy or to maximize sensitivity for a given minimum specificity or vice versa, the cut-off becomes a random variable and uncertainty related to the cut-off should be accounted for in the statistical analysis (e.g., confidence intervals).

This method of establishing clinical test validation can lead to overestimates of test quality measures. Cross-validation, bootstrapping, or other statistical techniques are available to obtain unbiased estimates of performance in such situations. However, estimates based on these techniques may not be as clear or convincing as performance based on independent validation of the cut-off points of interest. For example, the clinical trial for trastuzumab revealed that the 2+ category, which was previously chosen to be a positive test category, was really an indeterminate

[3] See http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfStandards/search.cfm.

test category. This conclusion was reached because there weren't enough patients in the drug clinical trial with a 2+ test result to arrive at a statistically significant determination for the appropriate cut-off**.**

4.3 Multi-Dimensional Examination of Setting of the Cutoff

It is important to examine the potential test cut-off in detail and to capture all relevant contributing information. For example, for an immunohistochemistry test in oncology, cut-off points may be defined in terms of the following:

- Cancer tissue percent of the specimen
- The type of cellular staining, e.g., membrane, cytoplasmic, nuclear
- The intensity of staining (0. 1+, 2+, 3+)
- The staining pattern (e.g., homo-, heterogeneous or focal staining; membrane staining, complete or partial/incomplete)
- The presence of leading edge staining and background staining intensity

4.4 Use of Receiver-Operating Characteristic (ROC) Curves to Aid in Setting the Cutoff Values for Diagnostic Tests

A ROC curve is a plot of sensitivity (true positivity) vs. 1-specificity (false positivity) for all cutoff points for any test. The ROC graph of all potential cutoffs can aid one in choosing an optimal cutoff for the intended use of the diagnostic.

The cutoff can be chosen to produce balance between true positive and false positive results, to emphasize true positives (when drug use is informed by the test for avoiding highly toxic adverse events), or to emphasize false positives (when drug use is informed by the test to ensure patients likely to respond are appropriately selected).

The area under the curve (AUC) ranges between 0 and 1, with better tests having larger areas. Generally, a test is informative if its AUC is greater than 50 percent. A useful guide for ROC curve analysis is the Clinical Laboratory Standards Institute (CLSI) Document titled "Assessment of the Clinical Accuracy of Laboratory Tests Using Receiver Operating Characteristic Curves (ROC) Plots; Approve Guidelines (1995) – GP10-A. A number of articles have been written on this methodology (11-14)

4.5 Identification of Indeterminate or Gray Zones

Cutoff values can be chosen to avoid or to include the presence of an indeterminate or gray zone for decision making based on the test. Two types of data can be considered in making this decision: clinical and analytical.

4.5.1 Clinical Factors

The ability of a test to discriminate between positive and negative results at a given cut-off point will depend in part on the strength of the clinical outcome signal being studied in patients with and without the drug response being sought. If there are strong and distinct differences in that signal between, for example, test-positive and test-negative patients, separation of drug responsive and nonresponsive populations is likely to be significant. There is probably little value to an indeterminate or gray zone.

However, if there appears to be a significant amount of overlap between outcomes in stratified patient groups, an indeterminate or gray zone may be of critical value in ensuring test results are properly interpreted and provide meaningful patient results.

4.5.2 Analytical Factors

The ability to discriminate between positive and negative analytical results at a given cut-off point will depend in part on the precision of the analyte signal being studied. With decreased precision (particularly for values near the designated cut-off point), the likelihood increases that a test determination will misclassify a patient.

In selecting cut-offs for clinical studies, attention should be paid to the precision determined during the analytical phase of method evaluation. Of note, between-laboratory differences may be significant for some tests and should be taken into account in decision making about whether and how wide to make any indeterminate or gray zone.

An example of the value for in-depth analysis of the cutoff and gray zone is the 2+ result of the immunohistochemistry Her-2/neu tests. Reproducibility studies revealed that readers had a difficult time separating 2+ from 1+ and 3+ results. The clinical trial confirmed that fewer persons with 2+ results were having positive drug outcomes than persons with clear 3+ results, and, as a result, 2+ results were re-categorized as representing indeterminate rather than positive results. To address uncertainty of values in this gray zone, a recommendation in the clinical practice was introduced to have all 2+ results evaluated by re-assay with another test method.

4.6 Clinical Test Validation

When a new diagnostic is being considered for use in selecting patients to receive or to avoid a particular drug therapy (i.e., drug/test co-developed product) or to stratify patients in some other way, two distinct, but related, issues should be addressed.

The first is the ability of the test to select (or deselect) patients with the biomarkers (analyte(s)) of interest. This is clinical test validation — use of a test to detect or predict the associated disorder of interest in biological specimens from the target patient groups. This should be considered the domain of clinical test validation.

The second is the ability of the test to result in patient selection that will improve the benefit/risk of the drug in the selected and nonselected groups. This would occur when the test identifies patients with a higher likelihood of benefit/risk or those at higher risk of an adverse effect, or

potentially both. This is considered to be the domain of clinical test utility (the risks and benefits to the patient associated with use of the test.)

Because these properties are separate, but related, studies should be conducted to ensure that there is evidence to support both the use of the test analytically in patients and the use of the drug in test-positive and test-negative subgroups.

5. GENERAL APPROACHES TO DEFINE CLINICAL TEST VALIDATION

Clinical test validation of a new diagnostic for use in selecting drug therapy or avoiding drug therapy should be characterized by studying the test in relation to the intended clinical outcome in patient subgroups with and without the analyte of interest.

Endpoints used in a clinical trial to evaluate treatment efficacy or safety should be the same endpoints used to indicate the clinical utility of a tested biomarker and should provide information on the clinical impact of an analytical test result. For example, HER 2 testing is not used for the purpose of detecting the presence of HER 2 per se in biological samples (analytical validity), but to identify patients likely to respond to treatment with trastuzumab (clinical validity) to ensure that patients receive optimum treatment choices (clinical utility). The clinical utility of Her 2 measurements refer to the effect that the measurements have on efficacy and/or safety (i.e., benefit/risk) of drug use.

For simplicity of discussion, the clinical efficacy and safety endpoints discussed will be limited to categorical endpoints although clinical outcomes are often continuous. For example, survival time could be categorized in such a way that patients surviving longer than a target duration (e.g., over 6 months) as compared to those that do not (e.g., less than 6 months) are considered to have positive and negative treatment outcomes, respectively. Conversely, for safety, continuous variables may be dichotomized. For instance, hepatotoxicity may be described by a certain level of ALT elevation (3 times the upper limit of normalcy) so that above a threshold value is considered a significant adverse event and below a threshold value is not considered significant. For efficacy endpoints, subjects with a positive treatment outcome are referred to as *responders*, and those with negative treatment outcomes are referred to as *nonresponders*. A relevant efficacy biomarker is one that is good at predicting a priori what is considered to be the beneficial response in subjects and to differentiate responders from nonresponders.

For safety questions, subjects experiencing an adverse event or meeting a predetermined criterion for a safety event are referred to as *cases* and those that do not as *controls*. A relevant safety biomarker is one that is good at predicting patients becoming either cases (i.e., high risk of developing an AE) or controls (i.e., low or no risk of developing an adverse event).

Although clinical accuracy, clinical sensitivity (positive test results in patients with the condition of interest) and clinical specificity (negative test results in patients without the condition of interest) all provide valuable information to analyze the value of a diagnostic test — and these

values should be reported — other metrics are available to provide additional insight into the clinical usefulness of the test in individual patients.

Clinical test validation can also be evaluated by the predictive value of a positive or negative test result. The positive predictive value (PPV) of a test is the likelihood that a patient with a positive test will have the clinical condition of interest (in this case, a defined beneficial or adverse response to a drug). It is a measure of the probability of being a responder or a case (i.e., having an adverse event) in test positive patients.

The negative predictive value (NPV) of a negative test result is the probability that a patient with a negative test will not have the clinical condition of interest (a beneficial response or adverse response to drug). It is a measure of being either a nonresponder or a control (a patient without an adverse event) if the test is negative.

Because prescribers and patients are usually interested in the probability of the patient being a responder or at risk for an adverse event, the clinical usefulness of a test is generally better measured by positive and negative predictive values than by sensitivity and specificity alone. Although predictive values are dependent on the sensitivity and specificity of the test being used, they are also dependent on change in prevalence of the condition being tested for.

This means that positive and negative predictive values of a test should be determined in patient populations similar to the patient populations for the indication. If predictive values are estimated from patient populations that have been enriched (e.g., through selective enrollment in a study), they may not be representative of values likely to be found in unselected patient populations in clinical practice (i.e., the results would not be generalizable). Since enrichment strategies for clinical trial response is acceptable and not uncommon, especially in proof-of-efficacy-concept studies during drug development, consideration should be given in drug-test co-development programs to how to generalize the results from enrichment studies to the target population for the drug and test.

To provide more useful information in test labeling and to avoid confounding by prevalence on diagnostic test sensitivity and specificity, additional metrics (e.g., positive and negative diagnostic likelihood ratios or LR) have been suggested to increase the ability to distinguish patient subtypes. For example, a positive likelihood ratio compares the likelihood of a test positive result in a population with the outcome of interest (e.g., being a responder or case) compared to another population without the outcome of interest. (15-17). For additional information see Addendum C.

5.1 Statistical Considerations in Drug-Test Co-Development

Results for clinical sensitivity and specificity of a new diagnostic test for use in patient selection for drug therapy should be generated with sufficient numbers of patients, whenever possible, to allow calculation of confidence intervals that are precise enough, or as a measure of uncertainty, to be clinically relevant to the therapeutic question being considered. The imprecision expressed by confidence intervals is to a large extent affected by the square root of the sample size. Thus,

in some cases, such as predicting which patient will be at risk for an adverse event, it may not be possible to obtain tight confidence intervals if the numbers of cases are relatively small. Also, it may not be possible to power clinical studies and specify beforehand the number of cases of toxicity to achieve a fixed confidence interval target since the number of cases to be expected is an unknown quantity. The value that a test contributes to decision making is usually based on clinical as well as statistical interpretation and depends on the question being asked and the clinical performance of the test.

Global assessments of clinical performance can be made by comparing ROC curves, likelihood ratios (at a set cut-off) of positivity and negativity in responders to nonresponders, or using overall testing efficiency (sometimes called clinical accuracy) at a set cut-off. The latter value can be calculated by adding the true positive and negative results together and dividing by all results.

6. CLINICAL UTILITY

A definitive clinical study for a drug used in conjunction with a predictive biomarker would be one that allows for assessment of a drug's safety and efficacy (i.e., risk/benefit), as well as for verification of the clinical utility of the biomarker in guiding the drug's use including appropriate patient selection. Ideally, analytical and feasibility studies performed during early drug development (phase 1 or 2 trials along with diagnostic test development described in Sections 4 and 5 of this document) will have already led the sponsor to a diagnostic test of potential value in designing pivotal clinical studies, defining subject inclusion and exclusion criteria based on the diagnostic test, and selecting drug doses. At this point, the sponsor should have identified preliminary cut-off points (and, if applicable, targeted equivocal zones) for further study as necessary.

If these performance parameters and other aspects of analytical and clinical test validation are not established at the point where phase 3 clinical utility studies are being commenced, acceptable documentation of clinical utility may not be possible within these studies. Rather, in such cases, the phase 3 clinical trials of the drug should be aimed at exploring clinical performance of the test and identifying appropriate cut-offs. To confirm clinical performance, including clinical utility, additional clinical studies may be called for to avoid post-hoc specification of the diagnostic cut-off points. If changes are made to a test during the clinical validation process that result in major analytical changes, the ability to use and pool data from differing time periods or different sites may be compromised and may therefore undermine the evaluation of the clinical utility process.

6.1 Coordinating Drug and Diagnostic Studies

6.1.1 Study Objective and Timing

The objective of joint drug-diagnostic studies is to ensure that the results of diagnostic testing in the target population have been properly linked to the expected response, that is, safety and/or

efficacy of treatment using the drug at specified doses. Some examples of goals that may be pursued in these co-development studies are as follows:

- Identifying patients who are good candidates for a therapy and therefore will have a greater chance of having a favorable efficacy (i.e., responders).
- Identifying patients who are likely to develop adverse outcomes with a therapy (i.e., experience toxicity following drug administration) and are therefore not good candidates for a given drug treatment.

These objectives can be translated into diagnostic, clinical, and statistical hypotheses in designing a prospective study that assesses both drug response, as well as the quality of the diagnostic in guiding therapy. In each case, the analytical and clinical endpoints should be carefully chosen and intended use patients carefully selected (preferably in as *naturalistic* a manner as possible so to better reflect actual clinical use). The clinical results should be recorded in a blinded manner and then analyzed in relation to test results.

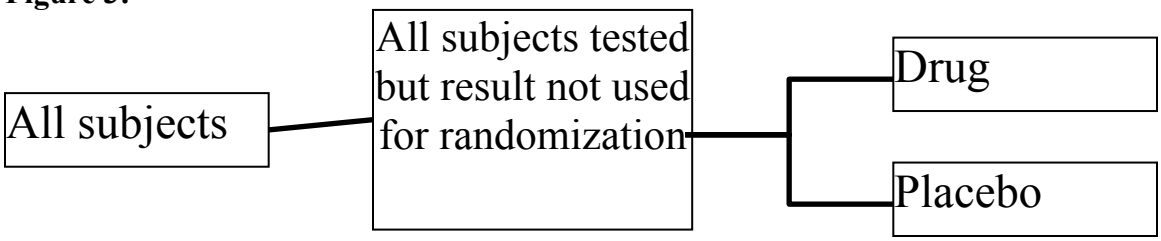
The optimal time to perform studies of a new biomarker as potential diagnostic tests to be used in informing the use of a new drug or to obtain samples for a future biomarker diagnostic test study (if a new biomarker diagnostic test has not yet been developed and clinically validated) is at the time of conducting adequate and well-controlled clinical trials for that drug in phase 3 of drug development. This timing offers a unique opportunity to study a population that represents the intended use population in a controlled manner (e.g., with assignment to drug and placebo, and results of the diagnostic test blinded). The results obtained from well-controlled trials provide information on the predictive results of the diagnostic test as they relate to drug response (safety and/or efficacy), as well as on any differential in the drug effect in diagnostic test positive and test negative patients, and between drug and placebo.

6.1.2 Clinical Trial Design Considerations

Careful attention to experimental clinical design can help minimize bias and assure that the results of the trials address the primary study hypothesis. There is considerable flexibility in drug-test clinical trial designs, and there are several design features that should be considered in a joint drug-diagnostic study.

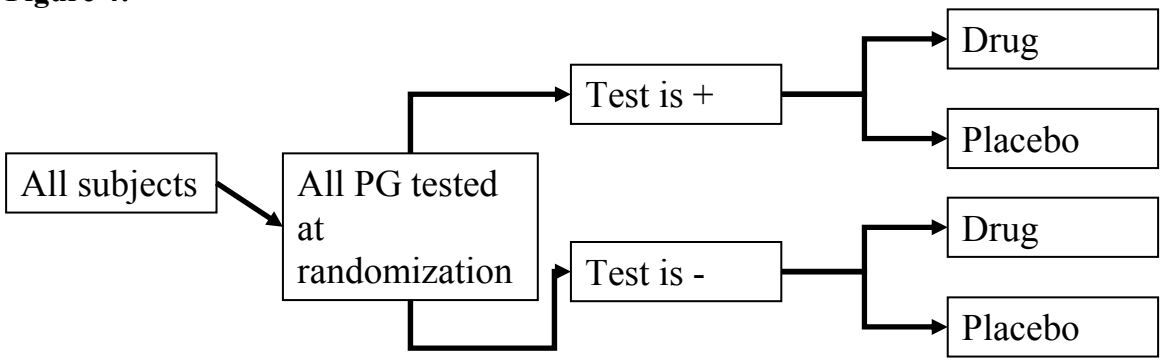
To explore the value of a diagnostic test within a drug clinical trial, the usual simple two-arm randomization comparing a treatment and a control may be employed, with the results from the diagnostic test or biomarker that is being investigated used as a prespecified stratification factor in the post-hoc statistical analysis. This would potentially allow for identification of a *treatment by diagnostic test result* interaction. One reason such a design may be adopted would be if the results of the testing will not be readily available at the clinical sites for informing randomization. A graphic (Figure 3) depicting this design follows:

Figure 3:



Alternatively, in Figure 4, randomization within differing strata by diagnostic test result (e.g., positive or negative subgroup) may be favored, particularly in circumstances where the test results are readily available at all clinical sites. Randomization ensures a balance in patient allocation between the treatment and the control for both the diagnostic test positive and test negative subsets.

Figure 4:



6.2 Issues and to Consider in Selecting Study Populations

In some cases, sponsors may wish to use enriched study populations to evaluate the likelihood of response to a drug treatment, such as in a proof of concept trial in early phase 2 of drug development. In these cases, careful explanation and justification of the enrichment technique used (diagnostic test, demographic information, other) should be provided. Consideration should be given to how enrichment will relate to the ultimate claims made for the drug being evaluated. That is, are the results generalizable, and will drug use be restricted to patients matching the enriched population studied and/or will there be efforts to justify use in different or broader patient populations.

Many of the important considerations that must be taken into account in designing clinical programs in which data from a test-defined subset of patients will be analyzed are quite familiar. Some of these considerations include:

- The clinical utility of the test (i.e., the strength of the association between the test results and a particular treatment response, whether beneficial or toxic, and the size of the difference between treatment response between tested and untested groups)
- Whether patients are readily identifiable in a clinical practice setting (i.e., would the test serving as the basis for enrichment be subsequently readily available in practice)? The prevalence of the marker being used to identify patients for treatment or for exclusion for treatment
- The intended use of the diagnostic in relation to the drug (i.e., will it be used for selecting patients for treatment, for identifying patients who should not receive treatment, and/or for making dosing decisions in test-defined subsets)
- In some cases, mechanistic and/or specific clinical data to support the hypothesis that a diagnostic test predicts enhancement of efficacy or safety in a tested population when compared to an unselected population may exist. In this case, the clinical development program for the drug should be designed to define the response in both patients with prior, known test status (both test positive and test negative) and in unselected patients. This is important to help establish the clinical validation and utility of the test. It is also important to establish an overall risk-benefit ratio for the treatment when used in the general population, since either efficacy or safety may differ (or both) in the test-positive and test-negative populations.

For some drugs when the indication is serious and life-threatening (e.g., drugs used in cancer), there is a reasonable likelihood that their use would occur in a wider population than the test-targeted population since clinical outcomes most likely will not be *all or none*. The wider populations would likely consist of untested patients or patients tested but without the expected result to guide therapy. In such cases, during co-development, studies should be conducted in which testing is done in an appropriate mix of test positive, test negative and/or untested patient populations, if possible, to be able to estimate clinical validation parameters and the overall benefit/risk of the drug in the general population of patients as well as the various subsets of patients. The sample size for these studies should be discussed with the appropriate review division for the specific therapeutic area.

The amount and extent of clinical trial data to verify the clinical utility of a test will differ, depending on the prior knowledge of the pathophysiology involved, and the mechanistic understanding of the way that the drug therapy exerts its pharmacological effect in relationship to the test, the magnitude of difference in clinical outcome between tested and untested patient groups, and the amount of previous relevant clinical data. In terms of confirming the value of a test in informing drug use, the evidentiary considerations are very similar to those of any other clinical hypothesis, and normally data from two or more adequate and well-controlled trials would be collected to confirm clinical effectiveness, thereby establishing the clinical utility of the test. Although prospective data are preferred, in cases where the analyte is stable and where collection bias (including spectrum bias, verification bias, and sampling bias) can be carefully characterized and addressed, prospectively designed retrospective clinical utility studies may be

possible. The design of these studies should be discussed in advance with the relative review divisions.

It will often be the case that a test is first used clinically during phase 3 trials, even if the trials are not specifically designed to be enriched based on the test status or otherwise designed to formally test a hypothesis related to these test results. In cases where the testing is done as an ancillary part of the trial (i.e., not incorporated into the trial design or primary outcomes), resulting associations between test results and clinical outcomes would usually be considered exploratory and therefore these results would be more appropriate for assessing clinical test performance or generating hypothesis about clinical utility rather than confirming clinical performance or utility.
For instance, if a clinical trial showed an overall marginal effect on the primary endpoint, but testing done retrospectively showed that there was an apparent greater response related to a particular test result (e.g., the subset of test-positive patients showed a larger, "statistically significant" response and those with a negative status showed little or no benefit), this observation may be confirmed in another clinical trial and the design and size of that trial should be discussed with the appropriate review division. It will depend on information above discussed above.

Optimally, further confirmatory testing would be performed in prospective trials. In some cases if the pathophysiological status of disease is well known, drug and diagnostic mechanisms well elucidated, and all of the effect comes from a defined subset, alternative retrospective validation methods may be considered.

In some cases, consideration could be given to banking samples for the purposes of retrospective analyses for associations between events of interest (including safety outcomes) and testing. The approach to these associations and analysis should be specified in advance and not after the study is completed. This technique may be of particular value in trials that are expected to explore doses at or near the toxicity threshold. In such instances, it would be important to establish that the target analyte is stable, that it is not subject to performance changes (particularly changes in interference) as a result of specimen storage, and that samples were obtained and banked without selection bias. Banking samples for later assay may lead to privacy and ethical concerns and regulatory requirements for samples. Sponsors should be sensitive to these requirements and seek informed consent and IRB approval as new sample banks are established. If global informed consent can be obtained, future and unpredicted studies or evaluation of links between test results and various clinical outcomes may be possible without re-consenting patients.

6.3 Data Collection and Data Standards

Uniform data standards for drug-diagnostic studies are being sought through many on-going academic, industry and government efforts. The data elements, data structure, terminology and content can be much more complicated when studies are designed to factor in independent drug and diagnostic effects. It is recommended that these issues be addressed in joint submissions

either as part of the IND submitted for review by CDER or CBER and/or as part of a pre-IDE or IDE submitted for review by CBER or CDRH.

In situations where a test is co-developed with a drug treatment, the ability of FDA reviewers to be able to audit clinical data and to link the results of treatment outcome to a subject's test result during the review process should be considered in designing the trial and archiving samples. In instances when testing may lead to privacy issues (e.g., genetic testing), test results may be masked to protect privacy, as long as it does not disallow test-outcome associations as described in the informed consent document. There should be processes in place to protect privacy, and these should be addressed in the informed consent and local IRB approval. In addition, it may be important to link review or audit of cases under study and use coded samples (single or double-coded), rather than fully anonymized samples, depending on data requirements for a particular study.

6.4 Verification of Clinical Test Utility — Statistical Considerations

Whether samples are collected and assayed prospectively, or collected prospectively, banked, and then analyzed retrospectively, every effort should be made to verify the clinical hypothesis being claimed with a study that is independent from the analytical and clinical study(ies) on which the diagnostic test was initially developed. That is, the analytical characterization (e.g., accuracy, sensitivity, cut-points etc) of a diagnostic test should be based on a dataset that is independent from and prior to the prospective or retrospective samples on which it is to be clinically verified. Otherwise, objective validation of clinical utility may not be possible. Preferably the clinical studies used to develop the diagnostic and the clinical utility study should have the same objectives (e.g., defined clinical outcomes, specified patient population).

Post-hoc characterization of a test based on the clinical utility data can be very misleading unless it is prespecified. For example, consider a multiplex diagnostic marker whose features and feature cut-point values are defined based on the clinical validation samples. This post-hoc characterization of the test marker can often identify a subgroup that appears to be associated with drug response or drug toxicity, but may actually be due to chance. A chance association is particularly likely when the number of features that could have been selected for inclusion as part of the multiplex marker is large since the chance of a spurious association increases with the number of features. The chance of a spurious association also increases when the selected features are combined post hoc to maximize test performance. An additional prospective study is ordinarily used to confirm the clinical validation of test utility defined post hoc. However, FDA could alternatively consider retrospective validation of the test utility if statistical techniques being using are robust, particularly in cases where the mechanism of action is understood, the strength of association is high, and replicate testing of independent collection of samples is possible.

Depending on the primary endpoint(s), the sample size used in drug-diagnostic studies can be estimated in a similar manner to that used in usual nongenomic study endpoints in clinical trials. Other factors that should be considered include recruitment numbers, based on the marker prevalence in specific sub-populations, and the magnitude of expected drug effects in subsets

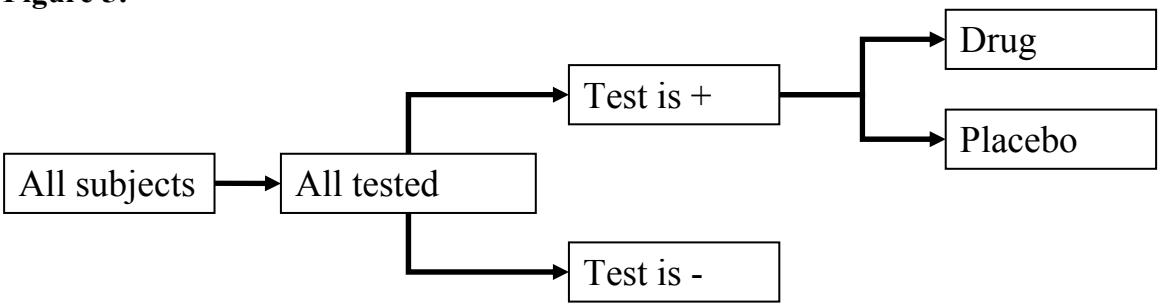
defined by testing or other stratification factors. Sample size should be calculated to address the drug effect (relative to a placebo or active control) in a subset of patients for which the biomarker diagnostic test facilitates treatment efficacy or safety.

6.5 Comments on Drug Efficacy and Safety Studies

Biomarker testing may be used as an aid in the drug development process by providing insight into differences in response among the patient populations being studied. Although valid biomarkers may be used to provide valuable insights into clinical dose-response, these may or may not be found to be of importance in clinical selection of patients to receive or avoid particular drugs. For instance, if a biomarker were developed which showed that asthma patients with a particular marker status have a greater response to an inhaled beta agonist than those without, this may be useful information, but may not impact on the choice of therapy if all patients had a reasonable, albeit different, response to treatment. In these cases, performance of the biomarker for purposes of understanding use of the drug are important but can be subsumed in the general review of the therapeutic and may not require independent credentialing of the assay as a diagnostic test for expected clinical use of the drug. This principle depends on the magnitude of efficacy differences in the target population and the relative benefit/risk is patient subsets defined by the biomarker test result.

However, in some cases diagnostic testing may prove to be so integral in the use of the new drug that testing will be considered a prerequisite to use. In cases when multi-site testing is expected in practice, new diagnostics will mean premarket review by CBER or CDRH. Clinical efficacy and safety of drug use may be linked and labeled with ***prerequisite*** test use in mind. In such cases, depending on the knowledge base at the time of the clinical trials, the trials may not even study patients with a particular test result. For instance, if by the time of the clinical trials there is reasonable certainty that drug response will only occur in biomarker positive patients (perhaps because the drug's mechanism works by means of that biomarker pathway), the design in Figure 3 could be considered.

Figure 5:



As stated above, drugs approved on the basis of such restricted testing would be integrally linked with availability and use of the biomarker test at the time of approval.

Although these types of study designs can be informative, they introduce difficulties in the evaluation of both the drug and diagnostic effects. In the example above, the drug effect relative to control can only be estimated in test positive patients. Descriptions of test sensitivity and specificity will not be possible using the above type of design without drug and placebo data in test negative subsets, and careful planning should be applied to determining how performance of both test and drug will be characterized.[4]

[4] See Begg and Greenes, 1983, *Biometrics*, 39, 207-215.

REFERENCES

1. McGuire L: Current status of estrogen receptors in breast cancer. Cancer. 1975; 36(2):638-44.
2. Lesko LJ, Salerno RA: Pharmacogenomic data: FDA voluntary and required submission guidance. Pharmacogenomics. 2004; 5(5):503-5.
3. Lesko LJ, Salerno RA: Pharmacogenomics in Drug Development and Regulatory Decision-making: the Genomic Data Submission (GDS) Proposal. Pharmacogenomics. 2004; 5(1):25-30.
4. Lesko LJ, Salerno RA, Spear BB, Anderson DC, Anderson T, Brazell C, et. al.: Pharmacogenetics and pharmacogenomics in drug development and regulatory decision making: report of the first FDA-PWG-PhRMA-DruSafe Workshop. J Clin Pharmacol. 2003; 43(4):342-58.
5. Leighton, JK, DeGeorge J, Jacobson-Kram D, MacGregor J, Mendrick D, Worobec A.: Pharmacogenomic data submissions to the FDA: non-clinical case studies, Pharmacogenomics. 2004; 5(5):507-511.
6. Ruaño G, Collins JM, Dorner AJ, Wang S-J, Guerciolini R, Huang S-M., Pharmacogenomic data submissions to the FDA: clinical pharmacology case studies. Pharmacogenomics. 2004; 5(5):513-517.
7. Trepicchio WL, Williams GA, Essayan D, Hall ST, Harty LC, Shaw PM, et. Al. Pharacogenomic data submissions to the FDA: clinical case studies. Pharmacogenomics. 2004; 5(5):519—525.
8. Lesko L, Woodcock J: Translation of pharmacogenomics and pharamacogenetics: a regulatory perspective, Nat Rev Drug Discov. 2004;3:763-770.
9. Federal Register notice for accepting comments for the July 24, 2004 workshop (Nancy, do you have the address? Shiew-MEi)
10. Manual of Standard Operating Procedures and Policies General Information – Review, Intercent Consultative/Collaborative Review Process Version 4 Date: June 18, 2004 (http://www.fda.gov/oc/ombudsman/intercentersop.pdf)
11. Obuchowski NA, Lieber ML, Wians FH, ROC curves in clinical chemistry: uses, misuses, and possible solutions. Clin Chem. 2004; 50(7):1118-25.
12: Zweig MH, Campbell G. Receiver-operating characteristic) ROC plots: a fundamental evaluation tool in clinical medicine. Clin Chem. 1993; 39(4):561-77.
13. Begg CB, Advances in statistical methodology for diagnostic medicine in the 1980's. Stat Med 1991; 10(12):1887-95).
14. Beck JR, Shultz EK, The use of relative operating characteristic (ROC) curves in test performance evaluation. Arch Pathol Lab Med. 1986110(1):13-20.
15. McNeil BJ, Hanle JA, Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. Med Decis making. 1984; 4(2):137-50.

16. Radack K, Rouan G, Hedges J. The Likelihood Ratio: an improved measure for reporting and evaluating diagnostic test results, Arch Pathol Lab Med, 1986; 110:689-693.

17. Weissler Am, Bailey KR. A critique on contemporary reporting of likelihood ratios in test power analysis, Mayo Clin Proc, 2004; 79:1317-18. 1. In Vitro Diagnostic Devices follow 21 CFR 809.10(b).

18. Guidance for industry: Guidance for Submission of Immunohistochemistry Applications to the FDA contains a model package insert and a check list for immunohistochemistry tests.

19. Guidance for industry: Clinical Studies Section of Labeling for Prescription Drugs and Biologics—Content and Format (http://www.fda.gov/cder/guidance.htm)

20. Guidance for industry: Content and Format of the Adverse Reactions Section of Labeling for Human Prescription Drugs and Biologics (http://www.fda.gov/cder/guidance.htm)

21. Guidance for industry: Labeling for Human Prescription Drug and Biological Products—Implementing the New Content and Format Requirements (under development)

GLOSSARY OF TERMS

Analytical Sensitivity – *In Quantitative Testing,* the change in response of a measuring system or instrument divided by the corresponding change in the stimulus; *In Qualitative Testing,* the test method's ability to obtain positive results in concordance with positive results obtained by the reference method.

Analytical Specificity – *In Quantitative Testing,* the ability of an analytical method to determine only the component it purports to measure or the extent to which the assay responds only to all subsets of a specified analyte and not to other substances present in the sample; In *Qualitative or Semiquantitative testing,* the method's ability to obtain negative results in concordance with negative results obtained by the reference method.

Analytical Validation – The in-vitro ability to accurately and reliably measure the analyte of interest. This aspect focuses on the laboratory component.

Clinical Performance – describes sensitivity and specificity, and other performance attributes of testing biological samples.

Clinical Sensitivity –The proportion of patients with a well-defined clinical condition (i.e. response to drug) whose test values are positive or exceed a defined decision limit (i.e., a positive result and identification of the patients who have a disease);

Clinical Specificity – The proportion of subjects who do not have a specified clinical condition (i.e. non-response to drug) whose test results are negative or within the defined decision limit

Clinical Utility – The elements that need to be considered when evaluating the risks and benefits in diagnosing or predicting risk for an event (drug response, presence or risk of a health condition.)

Clinical Validation – The process of determining the ability of a test to detect or predict the associated disorder (phenotype); this includes assessment of clinical sensitivity, clinical specificity, and/or other attributes of testing biological samples

In Vitro Diagnostic Device (IVD) – an "in vitro" reagent and any component part or accessory which is intended for use in the diagnosis of disease or other conditions, in man or other animals. (Section 201(h) of the Federal, Food, Drug, and Cosmetic Act) or "those reagents, instruments, and systems intended for use in the diagnosis of disease or other conditions, including a determination of the state of health, in order to cure, mitigate, treat, or prevent disease or its sequelae…Such products are intended for use in the collection, preparation and examination of specimens taken form the human body." (21 CFR 809.3.)

Negative Predictive Value (NPV) – The likelihood that a patient with a negative test will not have the clinical condition of interest (response to drug).

Pharmacogenetic test – An assay intended to study interindividual variations in DNA sequence related to drug absorption and disposition (pharmacokinetics) or drug action (pharmacodynamics), including polymorphic variation in the genes that encode the functions of transporters, metabolizing enzymes, receptors, and other proteins.

Pharmacogenomic test – An assay intended to study interindividual variations in whole-genome or candidate gene, single-nucleotide polymorphism (SNP) maps, haplotype markers, or alterations in gene expression or inactivation that may be correlated with pharmacological function and therapeutic response. In some cases, the *pattern or profile of change* is the relevant biomarker, rather than changes in individual markers.

Positive Predictive Value (PPV) – The likelihood that a patient with a positive test has the clinical condition of interest (response to drug).

Precision – closeness of agreement between independent test results obtained under stipulated conditions

Predicate Device – The legally marketed device(s) to which equivalence is drawn by FDA in reviewing class II premarket notification submissions.

Pre-IDE – an informal submission to FDA for review of a protocol prior to initiating a formal study. These are reviewed with a 60 day time line and are offered to sponsors at no cost. FDA will frequently hold teleconference with sponsors or meet with them to discuss findings and issues identified during the pre-IDE review.

Probable valid biomarker – A biomarker that is measured in an analytical test system with well-established performance characteristics and for which there is a scientific framework or body of evidence that appears to elucidate the physiologic, toxicologic, pharmacologic, or clinical significance of the test results. A probable valid biomarker may not have reached the status of a known valid marker because, for example, of any one of the following reasons: The data elucidating its significance may have been generated within a single company and may not be available for public scientific scrutiny.

Reproducibility (of results of measurements) – closeness of the agreement between the results of measurements of the same measure and carried out under changed conditions of measurement

Valid biomarker – A biomarker that is measured in an analytical test system with well-established performance characteristics and for which there is widespread agreement in the medical or scientific community about the physiologic, toxicologic, pharmacologic, or clinical significance of the results

ADDENDUM A: DEVICE DESCRIPTION – EXAMPLES OF ELEMENTS TO BE DESCRIBED

- Analyte definition: single vs. multiple analytes
- Overall design of the test, including quality control of feature identity and placement, where applicable
- Platform (e.g., quantitative RT-PCR, arrayed elements, flow cytometry, etc.)
- Assay components such as buffers, enzymes, fluorescent dyes, chemiluminescent reagents, other signaling and signal amplification reagents, instruments, software, etc.
- Internal controls and external controls used
- Sequence or identity of probes, primers, antibodies, or other capture elements
- Composition and spatial layout of arrays or other spatially fixed platforms
- Methods used in attaching the probe material to a solid surface, if applicable
- Hybridization conditions, washing procedures and drying conditions (e.g., temperature, length of time), and variations thereof, where applicable
- Specificity of probes for analytes of interest
- Range of measurable, detectable or testable analyte concentrations, as appropriate (see Addendum B, 3)
- Stability and reproducibility of the platform when used for its intended use
- For multiplexed tests in which the target molecules will contact a number of different probes, the potential for specific and non-specific probe cross-hybridization (design and functional testing)
- For multiplexed tests in which many probes are handled during the manufacturing process, the potential for probe cross-contamination during manufacturing
- Analysis software for the interpretation of multiplex test results and key parameters or settings used in the analysis.

ADDENDUM B: STUDY DESIGN – EXAMPLES OF ISSUES TO BE CONSIDERED

A complete description of each study would include protocol and results. This will enable an adequate interpretation of study results.

1. General analytical performance considerations

Studies should show that test performance can be applied to expected clinical use. You should demonstrate the device's ability to accurately and reproducibly detect the analyte(s) of interest at levels that challenge the analyte concentration specifications of the device (see number 3 below).

2. Sample requirements

If you intend to provide reagents for specimen processing you should demonstrate that the chosen sample preparation method consistently provides quality samples that yield reproducible test results for each specimen type included in your intended use. If you do not intend to provide sample preparation reagents in your kits, you should provide specifications for assessing the quality of the assay input sample, so that the user can validate its own sample preparation method and reagents. You should provide justification for these specifications in the submission. You should also determine the approximate amount of specimen to be collected from a patient to generate the required input of processed sample. We also recommend careful characterization of sample stability (i.e. validation of storage and handling recommendations).

When fresh samples for rare biomarkers, patterns, or other analytes are scarce, we will consider the use of archived or retrospective samples. Although natural samples are preferred, we will also consider artificially prepared materials, provided that you mimic natural matrices to the greatest degree possible. In particular, when using cloned or amplified material, the copy number tested should approximate that found in a natural sample. In addition, you should demonstrate the copy number range that is detectable (or testable, as appropriate) in the sample.

3. Analyte concentration specifications

We recommend that you establish a range of analyte concentrations that are measurable, detectable or testable by your assay as appropriate for your intended use. Analyte concentrations at the edge of these ranges should be included in subsequent tests of analytical performance to adequately stress the system. For quantitative or continuous-scale assays we recommend that a measuring (linear) range be established. For qualitative and discrete-scale tests we recommend that you determine the 95% limit of detection (LOD) of analyte, as well as the possibility of a saturation limit at high levels of analyte. For more guidance, see "Protocols for Determination of Limits of Detection and Limits of Quantitation;" Approved Guideline, NCCLS, EP17-A and "Evaluation of the Linearity of Quantitative Analytical Methods;" Approved Guideline, NCCLS, EP-6A. For high density platforms in which it would not be feasible to establish these parameters for each analyte detected we recommend establishing these parameters for a subset of representative analytes. The feasibility of establishing these parameters for each analyte should

be determined by the sponsor and if not feasible, the reasons should be justified in the submission.

Determining a "testable" range of analyte concentrations would be appropriate for devices such as genetic tests, where "detectable" levels of nucleic acid may not be sufficient to discriminate between alleles. We recommend you determine the minimum amount of input nucleic acid needed to obtain a correct genotype 95% of the time. The possibility of a saturation limit at high levels of nucleic acid should also be explored.

4. Cut-off

We recommend you provide the following to support an analytical characterization of your cut-off(s), if applicable:

- Study design and analytical data to support the established cut-off.
- Rationale for the units, cut-off and/or categories of the results.
- A description of specimen preparation including analyte levels, matrix, and how levels were established.
- A definition of "equivocal zone" if applicable.
- Statistical methods used (e.g., Receiver Operator Characteristic Analysis).

5. Controls and calibrators

For external controls and calibrators, describe analyte levels, matrix, method of preparation, value assignment and validation. Indicate how the recommended calibration and control testing frequency were established. If external controls or calibrators are not provided, indicate commercial availability or method of preparation. For internal controls, describe reactions and functions monitored. For different technologies, these controls may differ, but the user should be able to determine if critical reactions have proceeded properly. Controls should contain analyte concentrations that adequately stress the system (i.e. for quantitative tests, control concentrations should span the measuring range of the assay).

We recommend that you instruct the user on calibration of systems where it can aid in generation and interpretation of results. Depending on the technology, calibration may or may not be critical for proper use of the device.

6. Precision (Repeatability/Reproducibility)

Perform studies to determine estimates of total variability for each specimen type. For information on precision studies we recommend that you consult "Evaluation of Precision Performance of Clinical Chemistry Devices;" Approved Guideline-2nd Edition, National Committee for Clinical Laboratory Standards (NCCLS), Document EP5-A2 and "User Protocol for Evaluation of Qualitative Test Performance;" Approved Guideline, NCCLS, EP12-A at

http://www.nccls.org/. Include as appropriate repeatability (same day, site, operator, instrument and lot) and reproducibility (between runs, days, sites, operators, instruments and lots) studies. Precision panel test samples should contain analyte levels that adequately stress the system. You should carry out reproducibility at three or more sites, with different operators with skill levels appropriate to "real world" use, and preferably using different lots of the device. It is also preferred that testing be performed over several weeks and at different times of the day to maximize detection of potential sources of variability. The protocol should include evaluation of sample preparation reagents provided with the kit. If sample preparation reagents are not included in the test kit, each site should use and validate their own specimen processing procedures and demonstrate that the resulting sample meets manufacturer-supplied specifications.

7. Analytical specificity (interference and cross reactivity studies)

Potential inhibitors present in patient specimens may not be efficiently removed by sample preparation procedures and may even interfere with sample preparation itself. We recommend that you examine potential interfering substances commonly present in the indicated patient specimens for their effects on sample preparation and assay performance. Test samples should contain analyte levels that adequately stress the system. For more information on interference studies we recommend that you consult "Interference Testing in Clinical Chemistry;" Approved Guideline, NCCLS, Document EP-7A.

For both cross-reactive and interfering substances tested you should include the following:

- The concentrations at which these substances were present in the samples
- Sample description and preparation including matrix and analyte level
- The number of replicates tested for each substance
- How interference and cross reactivity were defined in relation to the results obtained for the reference positive and negative control samples
- A description of the degree of interference or cross-reactivity observed

8. Assay conditions

As applicable, you should verify reaction conditions (e.g. hybridization, thermocycling conditions), concentration of reactants, and control of non-specific activity. In the case of multiplex tests, you should examine and describe optimization of multiple simultaneous target detection. When thermocycling is used, you should verify optimization, specificity, and robustness of amplification.

9. Potential for sample carryover

We recommend that you assess the potential for sample carryover and provide instructions in your labeling for preventing carryover.

10. Limiting factors of the device

You should describe any known limitations of the device, such as when the device does not measure all possible analyte variations, or when the range of variations is not known.

ADDENDUM C: DETERMINING IF A DIAGNOSTIC TEST IS INFORMATIVE

The first step in interpreting diagnostic test results is determining if a test is informative. A test is clinically useful only if it provides information to discriminate between patients with and without the condition or interest (e.g., response or adverse event). Examples of standard diagnostic test performance metrics are clinical sensitivity and specificity.

Clinical Sensitivity and Specificity

For efficacy determinations, clinical sensitivity is the proportion of patients with a well-defined condition (e.g., response to drug) whose test values are positive. Most commonly, it is expressed as 100 times the number of true positive tests divided by the true positive plus false negative tests. Sensitivity is the rate of pick-up of the responders when using the test. In an ideal world, clinical sensitivity is 100%. In the real world, sensitivity is likely to be less than 100% since false negatives are to be expected.

In contrast, clinical specificity for efficacy is the proportion of patients who do not have the specified clinical condition (e.g., non-response to drug) whose test values are negative. Most commonly, it is expressed as 100 times the number of true negative tests divided by the number of true negative plus false positive tests. Specificity is the rate at which a test can exclude the possibility of being a responder. In an ideal world, again not likely, clinical specificity is 100%. Specificity is likely to be less than 100% due to inevitable false positives.

While sensitivity and specificity provide valuable information on the diagnostic test of genomic biomarker and we would anticipate these values being reported, other metrics are available to provide additional insight in the usefulness of the test.

Predictive Values (positive and negative)

Clinical performance of a test can also be evaluated by the predictive value of a positive or negative test result. The positive predictive value (PPV) of a test is the proportion of patients with a positive test result that have the clinical condition of interest. (i.e., for efficacy response to drug; for safety, adverse event to therapy). Specifically, it is 100 times the number of true positives divided by the number of true test positives plus false tests positives. It is a measure of the probability of being a responder or a case (adverse event) in test positive patients.

The negative predictive value (NPV) is the proportion of patients with a negative test result that do not have the clinical condition of interest. Specifically, it is 100 times the number of true negatives divided by the true test negatives plus false test negatives. It is a measure of being either a nonresponder or a control patient without an adverse event if the test is negative

Because prescribers and patients are usually interested in the probability of the patient being a responder or at risk for an adverse event, the clinical usefulness of a test is generally better

measured by positive and negative predictive values rather than by sensitivity and specificity alone.

In a perfect test, the PPV and the NPV would each have a value of 100%. The lower the value (the nearer to zero), the less useful is the test. Note that in a rare situation, e.g., a rare adverse event, even a screening test with a very high sensitivity may result in a low PPV. The reason for this is that while positive and negative predictive values are determined by test sensitivity and specificity, they are also influenced by the likelihood of a response or a case (prevalence).

Impact of Prevalence on Predictive Values

As prevalence increases, the predictive value of a positive test result increases and predictive value of a negative test result decreases, assuming that the test's sensitivity and specificity are fixed. In contrast, as prevalence decreases, the predictive value of a negative test increases and the predictive value of a positive test gets smaller. This means that attempts made at defining the positive and negative predictive value of a test should be done in populations approximating the patient population of intended use. If the population studied is different from that in which the test will be used, differences in disease spectrum and in interfering conditions may cause differences in performance which will make extrapolation between the two populations inappropriate. In addition study populations should be sized to assure reasonable confidence intervals for parameters of interest. If testing for predictive value is done in a patient population that has been otherwise enriched (through other testing or clinical selection methods), the predictive values determined in such a population may not be representative of those likely to be found in an unselected, clinical practice setting and therefore are not generalizable.

Likelihood Ratios

To provide more useful information in test labeling and to avoid confounding by prevalence on diagnostic test performance, additional metrics (e.g., positive and negative diagnostic likelihood ratios or LR) have been suggested, although at the current time, these metrics are less commonly used than PPV or NPV.

The diagnostic likelihood ratio is the likelihood that a test result would be expected in a patient with a condition (an efficacious or adverse response to a drug) compared to the likelihood that the same test result would be expected in a patient without that condition or response.

From this definition, the LR of a positive test result (+LR, or PLR) for the condition is sensitivity (true positive) divided by [100% minus specificity] (false positive).

Two definitions can be used for the LR of a negative test (–LR or NLR) – one to establish LR for no condition (non-response) for a negative result (–LRn) and one to establish LR for the condition for a negative results (–LRd). These represent the same information content but are inverse ratios with –LRn being calculated as specificity (true negative) divided by [100% minus sensitivity] (false negative) and with –LRd being calculated as 1-sensitivity (false negative) divided by specificity (true negative) . While –LRd is most commonly used in the laboratory

and statistical literature, some clinicians recommend the use of –LRn [(16, 17); Mayo Clin Proc. 2004;79:1317-1318].

Positive and negative likelihood ratios are attractive because they do not depend on the prevalence of the condition of interest and yet provide predictive value information. One can show that a positive LR of 5 for being a responder means that the odds of the drug-induced condition in a test positive patient are increased 5 times relative to the pre-test odds. A negative LR of 1/5 for –LRd means that the odds of the condition in a test negative patient are 5 times less likely relative to the pre-test odds. Note that if the positive or negative LR equals one, the odds are unchanged by the test result, indicating the test is uninformative.

Odds Ratio

Another measure of performance of a diagnostic test is the ratio of the odds in test positive patients to the odds in test negative patients of having the condition of interest (being a responder or a case).

By definition, the odds of having the condition given a positive test result and a negative test result, respectively, are PPV / (100 - PPV) and (100 -NPV)/NPV. Thus, the odds ratio combines these parameters as follows: PPV*NPV / [(100 -PPV)*(100 -NPV)].

The odds ratio is commonly used in studies that are enriched with patients with the condition of interest. Since enrichment distorts the underlying prevalence, PPV and NPV are not generalizable to the intended use population. However, the odds ratio combines the +LR and –LR into a single value and allows for an unbiased performance estimate in enriched studies. An odds ratio of 1 indicates that the test is non-informative (see below "Determining When a Test is Informative").

In short, the pairs (sensitivity, specificity), (PPV, NPV), and (PLR, NLR), taken together with the odds ratio, are very useful measures for assessing overall the clinical outcome and accuracy of screening tests.

How Prevalence Affects Predictive Values, Likelihood Ratios, and Odds Ratios — Examples

For a test with any given sensitivity or specificity, the actual interpretation of either positive or negative results in a patient can be markedly impacted by the prevalence of the condition. Prevalence refers to the pre-test likelihood of the endpoint of interest. In this case that endpoint is the response or adverse event caused by the drug selected.

Two hypothetical examples follow as they might apply to drug effectiveness. Similar modeling would be applicable to drug safety as well.

Suppose a new drug is found to produce a therapeutic response in 10% of treated patients. The prevalence of response is thus 10%. Suppose a new diagnostic test is identified that predicts response to the new drug with sensitivity of 90% and specificity 90%.

For a sample 100 patients, the expectation is that 10 will be drug responders and 9 of the 10 test will be identified with a positive test. However, 90 patients are non-responders and 81 of the 90 will be test negative (Table 1).

Note that the test is quite informative because sensitivity plus specificity is 180%, closer to the ideal of 200% than the uninformative state of 100%.

Although test sensitivity and specificity are high in this scenario, prevalence is low (10% = 10/100). As a result the predictive value of a positive test (PPV) is only 50% (9/18). The predictive value of a negative test(NPV) is 99% (81/82). Thus, the test is good at identifying nonresponders. It will also increase the yield of responders but certainly not guarantee a good response rate.

The increase in the yield of responders can be more precisely defined using the positive likelihood ratio (PLR). The PLR is the number of true positive test results divided by false positive test results (sensitivity/ 100-specificity) or 90/10 which equals 9.

The increase in the yield of nonresponders can also be more precisely defined using the negative likelihood ration (NLR). The NLR is the number of false negative test results divided by true negatives (100-sensitivity)/specificity or 10/90 or .11.

Suppose the same test is used in a setting where the prevalence of a therapeutic response in treated patients is increased from 10% to 80% (Table 2). Assuming the sensitivity and specificity of testing is unchanged (not always a safe assumption because of the introduction of spectrum bias in disease states), results will be very different.

With unchanged sensitivity and specificity, because prevalence is higher, PPV rises to 97% (72/74). NPV, however, falls to 69% (18/26). While positive results are more likely to predict response; negative results are less likely to predict failure to respond.

Table 1. 2x2 Table of Expected Results for 100 Patients Studied: Prevalence 10%, Sensitivity 90%, Specificity 90%.

	Untested population	Test positive Patients	Test negative Patients
Responders	10	9	1
Nonresponders	90	9	81

Table 2. 2x2 Table of Expected Results for 100 Patients Studied: Prevalence 80%, Sensitivity 90%, Specificity 90%.

	Untested populations	Test positive Patients	Test negative Patients
Responders	80	72	8
Nonresponders	20	2	18

While PPV and NPV are much different in the two examples due to the change in prevalence, the increase in the likelihood of response and of nonresponse are the same. The PLR is still 9, indicating that the likelihood of response when a patient tests positive is still increased 9 times over the pretest likelihood. The NLR is still .11, indicating the likelihood of nonresponse when a patient tests negative is still decreased to .11 the pretest likelihood.

Determining When a Test Is Informative

A test is informative only if its sensitivity plus its specificity is greater than 100%. For tests with a combined sum of more than 100%, the strength of the test should be considered in terms of both numerical and clinical impact of the combined numbers. Obviously, the closer the sum comes to 200% (sensitivity and specificity each of 100%), the better the test performs. However, values between 100% and 200% that are considered clinically meaningful would depend on clinical rather than mathematical considerations.

Performance measures other than sensitivity and specificity can also be used to determine if a test is informative. A test is informative only if one of the following equivalent statements is true: (1) sensitivity plus specificity is greater than 100%, (2) PPV plus NPV is greater than 100%, (3) +LR or -LRn is greater than 1, or (4) the odds ratio is greater than 1.

To illustrate failure of a test to impact diagnosis, suppose a new drug is found to produce a therapeutic response in 10% of treated patients. Suppose a new diagnostic test is found that putatively predicts response to the new drug. Assume its sensitivity is 80% and its specificity is 20%. Because the sum is 100%, the test is uninformative. To elaborate, for a sample of 100 patients, the expectation is that 20 are responders (16 of the 20 with positive test results) and 80 are non-responders (64 of the 80 with positive test results). (Table 3). Because the proportion testing positive is the same (80%) in both responders and non-responders, the test is uninformative, that is, fails to have any impact in diagnosing response or non-response.

This example also illustrates how the other performance measures discussed indicate that the test is uninformative. For 100 patients, the expectation is that PPV is 20% (16/ (16+64)) and NPV is 80% (16/20) (Table 1). Because the sum is 100%, the test is uninformative. With regard to likelihood ratios, +LR is sensitivity / (100 - specificity) = 80% / 80% = 1 and -LR is either (100 - sensitivity) / specificity or specificity/ (100-sensitivity) = 20% / 20% = 1. Because both are 1, they indicate that the test is uninformative. Finally, the odds ratio is PLR / NLR = 1, again indicating that the test is uninformative.

Table 3. 2x2 Table of Expected Results for 100 Patients Studied: Prevalence 20%, Sensitivity 80%, Specificity 20%.

	Untested population	Test positive Patients	Test negative patients
Responders	20	16	4
Nonresponders	80	64	16