Assessing the Clinical Utility of Diagnostics Used in Drug Therapy

J Woodcock

There is an ongoing debate over the evidentiary standards that should be applied for introduction of new diagnostics into routine clinical practice. Many call for evidence of “clinical utility,” i.e., a positive impact on patient outcomes. A diagnostic, when used with a medicine, has clinical utility if it improves the outcomes of drug therapy. Improved outcomes may be defined broadly, including benefits, harm reduction, and patient-reported outcomes. Much of the controversy centers around the methods of demonstrating clinical utility. For instance, are randomized prospective trials the only acceptable source of data? Practically speaking, many sources of evidence—mechanistic, pharmacologic, and observational—can contribute to a finding of clinical utility, depending on the circumstances. Clinical utility is highly indication specific, and achieving it is dependent on good analytical and diagnostic test performance.

Although the past several decades have been remarkable for the introduction of many new medical therapies, the upcoming years may well be known as the “age of diagnostics.” Aided by a host of new technologies, scientific discoveries from fields as disparate as physics and genomics are being exploited to yield diagnostic probes of unprecedented specificity and power. In the foreseeable future, clinicians may routinely image human biochemical processes, evaluate an individual’s genetic sequences, or catalog a tumor’s genomic aberrations. But these potential advances come at a time of intense debate over how to manage unsustainable increases in health-care costs and the role of new technologies in driving this trend. As a result, there is a corresponding debate about the evidence needed for any new diagnostic test to be introduced into clinical practice. Technology assessors/payers, entities responsible for establishing clinical guidelines, and other scientific bodies are calling for generation of this evidence, often referred to as the “clinical utility” of the diagnostic. However, exactly what constitutes clinical utility is not agreed upon. For example, diagnostic clinical utility clearly is not the same as the benefit/risk concept embodied in the safety and efficacy evidence generated for therapeutics, although some authors attempt to “force fit” diagnostics into the therapeutic evidence paradigm. Conversely, the fact that a diagnostic test accurately measures a biomarker is usually not adequate justification of its medical usefulness. Standard hierarchies of evidence, whether constructed for biomarkers or for “evidence-based medicine,” do not address certain issues specific to diagnostics. This article explores the concepts involved in determining the clinical utility of a diagnostic, with particular emphasis on diagnostics used in conjunction with medicines, either during drug development or in clinical practice.

Demonstration of the clinical utility of a diagnostic test is traditionally the culmination of a prolonged exploratory development process. The genesis of a new diagnostic begins with discovery of a candidate biomarker, perhaps a serum protein, a gene sequence, a gene expression profile, or a novel image. Historically, such candidate biomarkers have been discovered in academic laboratories and introduced to the biomedical community through scientific publications. These scientific papers are usually exploratory in nature and present evidence of an association between the biomarker, assayed in a test developed by the researcher, and a particular physiologic state, disease, or condition. Subsequently, if funding is available, further academic studies are published, expanding the findings. Collaborators may contribute patient samples so that larger series can be evaluated. Certain assays may become part of a diagnostic workup by specialists in the given field, who may request the original developers to perform the test as needed. As the volume of such requests increases, the assay may be licensed out to a commercial laboratory, which will offer it as a laboratory-developed test. Such laboratories typically conduct analytical validation of the assay, and perhaps some clinical validation; however, attempts at demonstration of clinical utility are uncommon. Rather, evidence accumulates through clinical use of the diagnostic. Similar

1Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, Maryland, USA. Correspondence: J Woodcock (janet.woodcock@fda.hhs.gov)

Received 13 August 2010; accepted 23 August 2010; advance online publication 27 October 2010. doi:10.1038/clpt.2010.230
development trajectories have occurred for novel imaging technologies (e.g., positron-emission tomography scanning). For some biomarkers, clear evidence of utility emerges from this process; for many others, controversy and uncertainty about their appropriate use may continue for decades despite their widespread use.

The development paradigm described above is becoming harder to justify in an environment increasingly focused on evidence-based medicine and cost-effectiveness. However, evidence generation is an expensive proposition. Sustainable business models for diagnostic development that include robust evaluations of clinical utility are hard to formulate. As a result, there is ongoing tension between the drive for scientific innovation through diagnostics and the need for more rigorous evaluation of their value prior to routine use in health care.

The history of the introduction of new diagnostics into drug development is somewhat different. Drug regulators worldwide have generally been extremely reluctant to accept results from new diagnostics that are intended to be part of a drug's evidentiary base, unless the performance of the assay is very well understood. Generally, regulators have waited (in some cases, decades) until the biomarker's clinical utility is worked out by the medical community. This contrasts with drug discovery, in which cutting-edge science is rapidly incorporated into the processes. Because a strong evidentiary base is generally lacking for new biomarkers, there has been some stagnation in the toxicological and clinical evaluation phases of drug development. Recognizing this growing gap, particularly with new genomic tests, the Center for Drug Evaluation and Research at the US Food and Drug Administration (FDA) initiated the voluntary genomic data submission process in 2003. This process provided a nonregulatory “safe harbor” for industry and FDA staff to discuss the application of these new technologies to drug development. The voluntary genomic data submission process has since been broadened to allow discussions of multiple technologies. In 2004, the FDA launched the Critical Path Initiative, which was intended to help modernize the later phases of drug development. One key area of focus of the initiative is biomarker development. As an outcome of the voluntary genomic data submission process, the FDA recognized that there was no clear regulatory pathway for biomarkers to be evaluated and formally accepted by regulators as part of the preclinical or clinical drug development phases. The concept of “biomarker qualification” emerged from public discussion of these issues. “Biomarker qualification” can be conceived as a drug development version of “clinical utility” in health care: determination of fitness for use for a very specific application. It is hoped that biomarker qualification, like clinical utility, will be pursued in a more organized and rigorous fashion in the future; however, the issue of how this work will be supported in a sustainable manner also arises in this setting.

The following discusses in more detail the pathways and challenges of evidence development for diagnostics.

BIOMARKER DISCOVERY AND DEVELOPMENT

A biomarker has been defined as “a characteristic that is measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacological responses to a therapeutic intervention.” Diagnostic tests are intended to assay (in the case of in vitro diagnostics) or illustrate (e.g., in the case of electrocardiogram tracings or imaging technologies) the state of a biomarker. Therefore, in most cases, biomarker discovery and scientific evaluation are the initial steps in developing new diagnostics that are not simply refinements of preexisting assays. In some cases, the candidate biomarker fits into an existing body of scientific knowledge—for example, genetic tests for drug-metabolizing enzyme polymorphisms, known Mendelian disorders, or virus sequences. Under these circumstances, the nature of the association of the biomarker with various clinical states is already fairly well understood, and movement toward assay development can occur rapidly. In other cases, the candidate biomarker is generated empirically—for example, by correlation of gene expression patterns with a particular clinical event or state—and, ideally, much scientific work should be conducted before moving to the creation of a diagnostic test. Skipping too rapidly over this scientific phase of biomarker development is one of the causes of subsequent difficulties in diagnostic development. Currently, the pace of drug development (glacial though it may seem) may exceed the ability to generate and assess specific biomarkers relevant to the use of the therapeutic (e.g., for patient selection or prediction of therapeutic response). The majority of drugs currently under development are not accompanied by new biomarker assays.

ASSAY DEVELOPMENT

Once the biomarker(s) of interest are clearly identified, the transition from laboratory assay to clinical diagnostic must begin (Table 1). Usually, a significant amount of effort should be invested in optimizing parameters such as sample collection, processing, and storage; reagents; physical test configurations; test procedures; and data processing. New assays such as those utilizing gene expression or proteomic technologies that measure a very large number of analytes can pose particular challenges with respect to data management and test configuration. Once a robust assay has been developed, it must undergo analytical validation to evaluate its performance characteristics.

Analytical validation (evaluation of assay performance characteristics)

This evaluation generates evidence on how well the assay measures the biomarker(s) (accuracy, precision, bias, and interoperator variability) and how robustly it performs under various circumstances (such as range and interference by other substances). Often, diagnostic performance of new assays is compared with that of a reference standard; however, for certain novel biomarkers discussed here, this is usually not an option because there is no gold standard. Understanding assay performance is crucial to assessing subsequent evidence on clinical performance and utility.

Clinical validation (evaluation of diagnostic performance characteristics)

This evaluation generates evidence on the diagnostic accuracy and predictability of the assay—in other words, how well the
test correlates with the clinical condition of interest. Quite a few genomic diagnostics are performed using a preexisting test platform that has already undergone significant analytical evaluation; therefore, evaluating clinical performance may be the first major assessment effort. Diagnostic accuracy includes sensitivity and specificity. Assays that measure novel biomarkers may be compared against a “truth” standard such as well-established phenotype data in order to generate evidence on sensitivity and specificity. Such a gold standard may not be available for other novel markers, and estimations of diagnostic accuracy must be performed using various clinical outcomes (that usually constitute the association that was the impetus for developing the test). There is often some circularity inherent in evaluating diagnostic performance in this manner. During this evaluation, assay parameters, such as cutoff values, and weightings of various inputs, may be fine-tuned to optimize performance for the intended use. Often, receiver operating characteristic curves and likelihood ratios are generated to summarize diagnostic performance. It is not widely appreciated that sensitivity and specificity may be dependent on the tested population; this is true for genomic assays as well as for more traditional diagnostics. Optimization of diagnostic performance for the intended use is critical for efficient assessment of clinical utility: establishment of cutoff values and other test parameters have use-dependent influences on performance. Clinicians often prefer utilizing the predictive values (positive and negative) of a given assay. Such values are highly dependent on the population tested and should therefore be generated in populations that reflect, as much as possible, those encountered in clinical care. A common error is to evaluate diagnostic performance only in individuals already known to have or not have the condition of interest: performance in an unselected population may be completely different.

Some of the confusion and debate over diagnostic performance and clinical utility stems from the wide range of scientific contexts in which new diagnostics are introduced and the varying types of information that new biomarkers contribute. Diagnostic performance and clinical utility are not completely independent concepts. For example, an assay for polymorphisms of a drug-metabolizing enzyme might classify individuals into different subgroups with a certain accuracy (which might vary among subgroups), reflecting its analytical performance characteristics. If the intended use is dose adjustment, the diagnostic accuracy would relate to how well the specific genotype finding(s) predicts the parameter of interest (e.g., international normalized ratio (INR) and blood levels of a specific drug). This information might be extrapolated from previously acquired knowledge or obtained from retrospective samples. Both these methods suffer from various deficiencies, and, often, prospective testing would provide the most definitive answers. On the other hand, previous information might be adequate for the purpose. The “clinical utility” would be obtained from the benefits and harms of using the test to modify dosing. These outcome measures could be modeled using previously known dose-response relationships (for toxicity and benefit) in combination with the predictive values of the test or could be determined prospectively from randomized trials. In contrast, a test that, based on empirical evaluation, classifies patients into prognostic subgroups (e.g., cancer gene expression assays) may have little in the way of independent supporting scientific evidence and may therefore require confirmation of diagnostic accuracy in populations or data sets other than those used to generate evidence for the biomarker. However, once that is done, reliable prognostic information for certain cancers may be deemed valuable on the face of the evidence, without further demonstration of utility. A third example is genetic tests measuring the potential for drug resistance, for example, in HIV infection. In this case, a large amount of data exist regarding the mechanism of resistance and its relationship to virus gene sequences. Once diagnostic performance is established, clinical utility for monitoring drug therapy and for classifying patients as candidates for new therapies can be worked out in a straightforward manner.

Once diagnostic performance of an assay—i.e., a correlation with a particular state or condition—has been demonstrated, drug sponsors may have interest in using the test in drug development. One way to accomplish this is to “qualify” the biomarker for a particular use.

### USE OF NEW DIAGNOSTICS DURING DRUG DEVELOPMENT: BIOMARKER QUALIFICATION

After a new diagnostic shows reasonable performance in measuring something, the next question that arises is “what is it useful for?” In drug development programs, this question translates to: “can the evidence from the assay be used in regulatory filings and to support decision making?” Answers to such questions are highly use specific. The paradigm of “biomarker qualification” involves generating evidence on the fitness for use of a specific
biodarkk assay for contributing evidence about drug safety, efficacy, dosing, patient selection, and so on. Biomarker qualification establishes global, rather than product-specific, fitness for use; that is, it involves a finding that the information generated, for the specific use, is deemed reliable and will therefore be acceptable to regulators. In contrast, development of a biomarker assay for use with a specific drug usually involves evaluating its value for use in health care, i.e., its “clinical utility.”

The FDA has established a process through which biomarker assay sponsors, typically a consortium of interested parties, may submit information to the agency with a request to initiate the qualification process. If the FDA finds the data promising, it may agree to participate and provide advice on further evidence development. Frequently, the first steps taken by the consortium involve analytical and clinical validation of the assay. These can then be followed by evaluation of a specific use. Once complete, this evidence is submitted to the FDA (and often the European Medicines Agency and other regulators). The FDA will review the data; if fitness for use has been demonstrated, the agency will provide written notice of this finding. It is expected that the scientific data on which the qualification is based be made public. The FDA may wish to conduct an advisory committee meeting or other public discussions. The biomarkers that are currently furthest along in this process, new assays for drug-induced renal toxicity, have been qualified in this manner for animal toxicology studies, and clinical studies are under way to initiate qualification for use in drug development trials. These qualification studies generally seek to evaluate the diagnostic performance of the assays in the setting of drug-induced renal injury. If qualification is successful, and after this use of the assays is formally accepted by regulators, the assays could be used to monitor patients during trials of potentially nephrotoxic drugs. In many cases, such drugs would either be found not nephrotoxic or would be abandoned if they are found to cause renal toxicity. However, a nephrotoxic drug that otherwise had unique clinical value might be further developed, along with one or more of the assays to monitor renal function. In such a circumstance, if the drug were approved in conjunction with the assays, the “clinical utility” of the assays would have been demonstrated by a combination of evidence from the qualification studies and the specific drug development program (i.e., that use of the assays allowed the drug to be used with reasonable safety). This example illustrates why regulators are seeking to formalize the process of evidence development and evaluation for new biomarkers used in drug development. For example, individual drug sponsors are ordinarily not able to generate the quantity and variety of evidence needed to create confidence in results of new safety markers. But new safety tests are critically needed in many stages of clinical evaluation. Many stakeholders—clinicians, patients, researchers, and developers—would benefit if new, reliable diagnostics were to become available. Acceptance of such assays requires extensive scientific scrutiny and evaluation. Therefore, biomarker qualification is designed as a public process with broad scientific input and publication of evidence. Potentially, many assays undergoing qualification will ultimately be evaluated for clinical utility.

**DETERMINATION OF CLINICAL UTILITY**

Evidence for clinical utility in health care addresses the question: “is the test worth doing?” How much value does it add, and how much harm will result from using it? Similar to biomarker qualification, determination of clinical utility is highly situational: “worth doing for what use?” Much of the current debate over demonstration of clinical utility is driven by differences of opinion about the weight that can be given to various types of evidence in different clinical scenarios. Arguably, evidentiary standards for clinical utility should depend on the circumstances of the test. Standards applied to tests intended for population screening or prenatal diagnosis might reasonably differ from those used for determining the risk of a rare, serious, drug-induced adverse event. Given the fact that resources for clinical trials are limited, it is rational to apply a risk-based approach to evidence generation. Positions on this issue in the biomedical community vary from “clinical utility should always be evaluated in prospective randomized trials” to “clinical utility can be demonstrated from first principles” (e.g., some images). To a great extent, the positions taken in this matter appear to be related to the party’s role—e.g., technology assessor, developer, provider, etc.

Issues related to the clinical utility of new diagnostics used with drug therapy vary depending on the point in the drug life-cycle at which the diagnostic is introduced. Diagnostics intended for use with marketed drugs might be considered as “retrofits”: they are intended to modify the longstanding use patterns of the drug by providing additional information. “Rescue” diagnostics are introduced before or around market approval: these diagnostics were not originally incorporated in the drug development plan but were subsequently added to improve the safety or effectiveness profile to support approval, utilization, or reimbursement decisions. Finally, and (currently) uncommonly, new diagnostics are incorporated prospectively during drug development in a process of drug–diagnostic co-development.

**DRUG–DIAGNOSTIC CO-DEVELOPMENT**

Many of the challenges of co-development of a drug and novel diagnostic appear to relate to proper timing. It has been difficult for developers to identify and analytically validate an assay (for example, to predict treatment response) in time to identify a population for efficacy trials. Much has been written about sequential trials, which first determine diagnostic performance of a test in conjunction with a drug, followed by a confirmatory stage. However, the actual difficulty at the moment appears more scientific in nature—identifying an assay that might predict treatment response in the first place. With regard to the clinical utility of the test in this situation, development programs that utilize test-determined eligible patients for trials, and do not evaluate the ineligible group, are obviously not designed to yield calculations of negative predictive value (this is true for enrichment designs based on classification into prognostic subgroups or treatment response). Therefore, it may not be clear what value, if any, the test added—would the outcomes have been different without the test? Such designs were used in the efficacy trial of
the breast cancer drug herceptin (in which only test-positive individuals were studied) and the recent JUPITER trial of rosuvastatin.\(^8\) In the herceptin trial, omission of the test-negative group and conclusions about the clinical utility of the assay were justified on the basis of mechanistic reasoning about the nature of the drug’s interaction with its target. The JUPITER trial studied patients who were not currently recommended for statin therapy, with elevated levels of a biomarker, “high-sensitivity C-reactive protein.” This marker (and its associated cutoff value) has previously been shown, in other studies, to have prognostic discrimination for risk of cardiac events. Higher-risk patients identified by this assay were randomized to drug or placebo, and a reduction of events was observed in the treated group. In this case, evidence of clinical utility of the diagnostic (identifying patients worth treating with a statin) would be assembled from the previous demonstration of risk stratification and the drug trial. However, not all members of the biomedical community were convinced that the performance of the marker, including justification of the cutoff, had been adequately explored.\(^9\) Similar considerations apply to trials of drug interventions in cancer that enroll patients with presumed poor prognosis based on a genetic “classifier.” The trials will demonstrate whether or not the particular population benefits but not whether the test added any value. That information would have to come from other sources: mechanistic reasoning, previous retrospective studies, or other trials.

Despite these challenges in determining the predictive value of a co-developed diagnostic, other problems—such as identifying and rapidly developing a relevant assay—will probably dominate drug–diagnostic co-development for the foreseeable future. If existing diagnostics can be utilized, evidence of analytical validation should be thoroughly evaluated prior to conducting drug efficacy trials. If the evidence for diagnostic performance for the purpose at hand—however generated—is not strong, consideration should be given to conducting a stratified or sequential trial in order to first obtain this information and then demonstrate clinical utility.

“RESCUE” DIAGNOSTICS

Drug development is notorious for late surprises—lower-than-hoped-for efficacy and unexpected toxicities. As a result, the late search for diagnostic tests that improve either efficacy or safety is a recurring pattern—a more scientifically driven version of the retrospective subgroup analysis. A recent example involves the monoclonal antibodies cetuximab and panitumumab,\(^10,11\) which are intended for treatment of advanced colon cancer. These antibodies target the epidermal growth factor receptor and were approved in the United States on the basis of efficacy trial data. However, the European Medicines Agency review pointed out that a significant proportion of exposed patients failed to respond and asked the developers to analyze genetic predictors of response. Retrospective analysis for mutations of kRAS showed that patients with wild-type kRAS had a much higher likelihood of response, and those with mutations had a negligible response rate. Mechanistic considerations support these results. US regulators accepted the safety findings
and showed that use of the marker significantly diminished improve the use-limiting safety profile of a drug. Abacavir is associated with hypersensitivity reactions that can be life threatening on re-challenge. The drug developers conducted an extensive investigation of the reactions, culminating in a randomized clinical trial. The program identified a genetic risk marker and showed that use of the marker significantly diminished the safety problem. Clinical utility of the diagnostic for this use was confirmed, and prescribers could use the drug with more confidence. Some observers thought that establishing a strong association between a positive test and an increased risk would have been sufficient evidence of clinical utility in this case. They proposed that clinical implementation of the test, followed by post-test documentation of reduction in events, would provide sufficient evidence of the test’s contribution. In some clinics, this policy was followed (given that the assay was available for other purposes). However, the clinical trial demonstrated convincingly that screening with the risk marker actually decreased reactions, provided evidence of the magnitude of risk reduction, and arguably improved clinical uptake. This case illustrates some of the areas of contention around clinical utility. Depending on the clinical situation, some observers are more willing to accept extrapolations from existing data than are others. From a pragmatic viewpoint, if pharmacogenomic science is successful, the capacity to conduct outcome trials could become a rate-limiting step if prospective randomized trials were to be required before clinical implementation. Strong association with risk may be acceptable in many cases, especially if the biological bases for these reactions can be elucidated. As more knowledge emerges about the causes of serious drug-related adverse reactions, use of assays for very specific risk factors may become more common in drug development, so that such assays could be incorporated prospectively in trials.

**DIAGNOSTICS USED FOR “RETROFIT”**

Many marketed drugs have been available for decades and are off-patent. Their safety or dosing problems are well known, and there is less pharmacological information available about them than about the more recently developed drugs. The use of such drugs may be improved by incorporation of modern diagnostic tests. However, there is often no sponsor interested in pursuing this approach. Additionally, bringing about changes in prescribing patterns, once established, can be very difficult. Nevertheless, academics and regulators have been assessing a variety of diagnostic tests intended to improve dosing or address a specific safety problem. In general, use of diagnostics intended to identify individuals at higher risk of a rare, serious adverse event (e.g., carbamazepine and Stevens–Johnson syndrome) has been accepted with little concern. The rarity of the outcome precludes prospective testing, and the life-threatening nature of the event might raise ethical issues about randomizing individuals, even if such a trial were feasible. There appears to be acceptance that there is clinical utility in reducing the at-risk population, even if the magnitude of risk reduction is poorly understood. Importantly, the incidence of such serious reactions is usually several orders of magnitude lower than abacavir hypersensitivity. Therefore, many factors may contribute to setting the evidentiary standard needed for clinical adoption of safety tests.

Warfarin is another drug, marketed for many years, that causes a significant burden of adverse events including mortality. Some of the harm is related to incorrect dosing, and it is well known that genetic factors influence the individual dose requirement.14 There is a long-standing controversy about the evidence needed to support the use of genetic information for setting the initial dose level. The current standard of care involves an empirical starting dose followed by careful and frequent monitoring of the pharmacodynamic INR, with dose titration to a stable INR. Proponents of genetic testing point out that its use allows patients to be started at doses closer to their ultimate stable dose: opponents reply that these putative benefits are mainly derived from retrospective studies and that the potential harm from starting at test-assigned doses is not well understood. A randomized trial to evaluate these issues has been initiated, although the surrogate end point of stable INR rather than benefits and harm will be used, given the resource constraints.15 Such a trial must, by ethical necessity, compare the results of genetic test-directed dosing vs. those of the highest attainable standard of care. This raises the issue of comparative effectiveness. Multiple studies have shown that, outside of trials and selected centers, individuals rarely receive INR monitoring that resembles this standard (because prescribers fail to order the monitoring, or individuals fail to comply, or both). Comparative effectiveness studies evaluate the clinical utility of various interventions or approaches to therapy in real-world settings. A comparative effectiveness study of genetic test-directed dosing vs. real-world care would have to utilize cluster randomization or a similar mechanism to evaluate the performance of each in actual practice settings. It is highly likely that the use of genetic testing would have better relative performance in such settings because the lower frequency of monitoring would reduce its effectiveness. Additionally, the concept of clinical utility of diagnostics (and other interventions) is often held to include patient-reported outcomes such as quality of life. If genetic testing decreases the amount of titration (and consequently minimizes the number of venipunctures, time off work, etc.) needed, the psychosocial and economic impact on patients could be significant. This impact is not usually evaluated in trials, although it is a major reason for real-world noncompliance with monitoring. These impacts could not be assessed very easily in a cluster randomized trial and could not usually be evaluated at all in a typical randomized trial in which all participants receive the same intensity of care. All these considerations illustrate the complexities of developing evidentiary standards for clinical utility of diagnostic tests for use with drug therapy.
Finally, emerging scientific knowledge (that fits in none of the above categories) may offer explanations of observed variability in drug response that may have serious consequences for individual patients (e.g., clopidogrel). Efforts to link that knowledge to drug outcomes in an expeditious manner will continue to raise the same sorts of controversies as the instances above. The resolution of some of these controversies may help guide the development of overall policy approaches to evidence development for drug–diagnostic pairs.

**DIAGNOSTICS AS SUROGATE END POINTS**

There is great interest in using diagnostics as surrogate end points in clinical trials during drug development. Clinical end points are usually defined as direct measures of how an individual “feels, functions, or survives,” whereas surrogate end points are biomarker assays of some kind (e.g., imaging, laboratory assays, electrocardiograms, and blood pressure) that are supposed to correlate with the clinical end point(s) of interest. Clinical end points are ordinarily used to determine drug efficacy in registration trials. Surrogate end points are desirable because they can often be evaluated much more quickly than many clinical outcome measures, and they are almost always easier to quantify. Such surrogates are most useful in cases in which the clinical outcomes take a long time to occur. Preventive interventions in particular (for example in chronic diseases) may take many years to result in a clinical effect; in some such cases, pre-approval trials using clinical end points may not be feasible. The FDA and other regulatory agencies routinely accept certain surrogate end points, such as blood pressure reductions, in clinical trials for drug approval. These are known as validated surrogate end points. The FDA also has a mechanism known as accelerated approval. Under this regulatory program, the Agency can accept less than “validated” surrogate end points—ones that are “reasonably likely” to correlate with clinical benefit—in serious or life-threatening illnesses that lack acceptable alternative therapies. Many drugs for HIV (surrogate = HIV-1 RNA copy number, a laboratory test) and cancer (surrogate = tumor shrinkage or progression-free survival, primarily as determined through imaging) have been approved through this route. Drugs approved under accelerated approval must be further studied after approval with clinical end points to verify the effect. Most diagnostic assays that are evaluated as surrogate end points already have acknowledged clinical utility for some purpose in health care. However, the evaluation of a diagnostic as a surrogate end point is similar to biomarker qualification—it involves assessment of fitness for use for a specific purpose in drug development rather than a use in health care.

Much controversy and confusion surround the use of surrogate end points in drug approval. There are no formal regulatory processes for evaluation or adoption of new surrogate end points, although various criteria have been proposed in the literature (see ref. 19, Appendix A, for a comprehensive review). Some of these criteria have called for any new surrogate end point to reflect “all” the clinical outcomes related to the drug intervention, so that there would be no loss of information as compared with the use of clinical end points. In fact, the use of a surrogate end point is an explicit trade-off between faster drug access (or any drug access at all) and certainty about the drug’s effects. It is unrealistic to expect a biomarker assay to predict all possible outcomes from use of a drug, or even all the beneficial outcomes. Generally, when qualifying a biomarker assay as a surrogate for drug efficacy, the mathematical relationships between the candidate surrogate end point and an accepted clinical outcome measure in the disease are explored. It is possible, using natural history data related to the biomarker and to clinical outcomes in combination with data from successful interventional trials in the disease, to correlate changes in the biomarker with both disease progression measures and improvements in outcomes resulting from therapy. Ordinarily, trial data from several different interventions (e.g., different drugs and different drug classes) are needed to assure generalizability. This evaluation forms the evidentiary basis for accepting a surrogate end point for efficacy. It is conceptually similar to diagnostic performance evaluation. There have been many calls for surrogate end points to predict all clinical efficacy outcomes—in other words, to have perfect predicitive performance—but evaluations of currently accepted surrogate end points show that they do not correlate completely with either disease progression or responses to interventions. This shortcoming has not prevented them from having an important impact—for example, the use of HIV-1 RNA copy number in stimulating the development of antiretroviral therapy that allows management of HIV as a chronic disease. Acceptance of a surrogate end point represents a judgment that the benefits of more rapid and efficient drug development in a specific disease outweigh the risks of using a surrogate end point. There are two risks associated with this judgment: the truncated safety evaluation might not adequately evaluate drug risks and the effects on the surrogate brought about by the drug might not correlate with clinical benefit. This benefit/risk calculation is very specific to each disease and to each drug development program. For example, the use of blood pressure measurements in trials of new antihypertensive drugs is well established. There is general confidence based on decades of experience that lowering blood pressure into certain ranges, using a wide variety of drugs from different classes, will improve long-term outcomes. Trials of new antihypertensive drugs can be of a duration and size to permit an extensive safety evaluation. Therefore, the risks of surrogate use are very low, which is appropriate in a field in which there are many acceptable therapies. In contrast, the use of a surrogate end point for a rare, serious disease for which there is sparse natural history data and no accepted therapy involves taking many more risks; however, these risks may be acceptable for patients who lack any type of treatment option. Weighing the benefits vs. the risks of surrogate end point use in a specific disease is conceptually similar to evaluating the clinical utility of a diagnostic; however, in the case of a surrogate, the outcomes often need to be modeled rather than based on data. The following points should be explicitly taken into account when framing discussions about surrogate end points:

1. There are inherent limitations in the safety evaluation when using a surrogate end point to evaluate drug efficacy.
Methods to mitigate the following limitations should be carefully explored:

(a) A biomarker that reflects the disease process and changes with interventions in the disease process cannot be also expected to predict off-target effects. Therefore, the biomarker assay will not predict most drug toxicities.

(b) The use of a biomarker that has a greater sensitivity to change than the clinical outcome measure will require fewer subjects in the clinical trial(s). This is one factor that makes surrogate end points attractive. However, as a result of lower numbers of patients exposed to the drug, rare or less common side effects may not be detected, information about common effects will be diminished, and several-fold increases of an adverse event over a background rate may be missed.

(c) A major benefit of using a surrogate end point is that trials can be much shorter in duration. However, this is also a major liability. Toxicity problems that worsen over time, or that accumulate slowly, may be missed.

(d) Those experienced in clinical drug development will recognize (a–c) above as the usual situation, simply exacerbated by “downsizing” the clinical evaluation. In fact, all drug development involves trade-offs between certainty and feasibility. In assessing the use of a surrogate end point in a development program, ways to mitigate the loss of safety information should be explored (e.g., by conducting parallel safety trials and by initiating an early safety cohort).

2. Great care should be taken to define exactly what clinical outcome the surrogate end point for efficacy represents, as well as what time point in the disease the surrogate and clinical outcome reflect. Surrogate end points are primarily used in chronic diseases because clinical outcomes can readily be measured in acute illness. Chronic diseases often have complex, multidimensional clinical outcomes that may evolve over time.

(a) Much of the thinking about the use of surrogate end points for efficacy seems to reflect past experience in cardiovascular medicine, in which survival is a key end point. In fact, survival has been referred to as the “ultimate end point” for efficacy. However, in most chronic illnesses, although survival is important, it is not the major disease manifestation being targeted. In fact, survival is usually a safety outcome (does use of the drug result in excess deaths?). Lumping safety and efficacy outcomes together in deliberations about surrogate use has led to some of the confusion and controversy in this field. Whereas an efficacy surrogate might be expected to reflect improved survival in a disease characterized by significantly increased mortality, the surrogate would be unlikely to predict increased mortality caused by drug toxicity. The use of a surrogate will usually result in less information pertaining to survival, given the trial designs employed: this is one of the risks that must be weighed. In addition, it is by no means clear that survival is the “ultimate end point” in many diseases. Multiple patient-preference studies have shown that many people would trade duration of life for quality of life: for example, they would avoid a therapy that causes increased suffering even if it improves their chances for longer survival.21,22 Therefore, even studies that measure the “ultimate end point” may miss the point, from the perspective of the patient.

(b) Often, a single clinically important outcome measure is selected for use in efficacy trials, and this outcome will be correlated with the candidate surrogate. This outcome may not encompass all the important disease domains. In many chronic diseases, a number of clinical outcomes are not routinely measured, even in drug clinical trials (e.g., patient-reported quality of life). The use of a surrogate end point further diminishes the available information on how the intervention truly affects patients.

(c) Finally, there must be clarity on what clinical outcomes are represented by the surrogate. For example, in efficacy trials in type 2 diabetes, a laboratory test for glycated hemoglobin (HbA1c) is used to evaluate ongoing glucose control. Type 2 diabetes is a chronic, symptomatic disorder with long-term consequences. Glucose control (at various cutoff levels of HbA1c) correlates with symptoms and also with long-term microvascular complications.23,24 However, improved glycemic control has not been shown to decrease the macrovascular complications of type 2 diabetes. Cumulative measures of glycemic control also will not provide an insight into most drug safety problems. Therefore, the use of HbA1c applies to the drug’s effect on control of blood sugar and, by implication, microvascular complications. There have been concerns raised about the adverse cardiovascular effects of various drugs used in the treatment of type 2 diabetes; the FDA has addressed this aspect by requiring specific cardiovascular safety studies for new agents.25 Various authors have linked some of these controversies to the use of HbA1c as a measure of glycemic control. Clinical outcome measures for type 2 diabetes include acute and subacute symptoms of hyperglycemia, medical complications of hyperglycemia (e.g., infections), and long-term microvascular complications. Therefore, a surrogate end point linked to these clinical outcomes would not be expected to encompass beneficial or adverse cardiovascular outcomes.

No simple standard can be applied to a decision to use a surrogate end point to establish drug efficacy. In each disease state, the potential benefits and risks of surrogate use must be weighed. The FDA has formally acknowledged this in its regulation on accelerated approval, which states that “unvalidated” surrogates may be accepted only in cases of serious or life-threatening diseases that lack alternative therapy. That is, as with conclusions about clinical utility, the standards for surrogate use depend on the facts of the particular case at hand. The
acceptance of a surrogate should involve formal evaluation—not simply an evaluation of the mathematical relationship of the surrogate to the outcomes but also of the benefits and risks of accepting decreased levels of certainty in a particular disease state. Such benefits and risks will emerge over time, as treatments become available, as knowledge about the disease grows, or as information about performance of the surrogate end point becomes available. For this reason, the utility of surrogate end points may change over time and should therefore be re-evaluated at intervals.

The Institute of Medicine has recently released a study report entitled “Evaluation of Biomarkers and Surrogate Endpoints in Chronic Disease.” This study was requested by the Center for Food Science and Applied Nutrition of the FDA as part of an effort to establish standards for the use of biomarkers in health claims for foods. The study committee recommended a framework for evaluation of biomarkers (that differs slightly from that discussed in this article). They also recommended that the FDA should use the same “degree of scientific rigor” for biomarkers across all regulatory areas and that the FDA should convene expert panels to evaluate biomarkers and biomarker assays. The study report contains a fairly comprehensive review and discussion of the relevant literature.

**SUMMARY**

There are varied, and strongly held, opinions about the types of evidence needed to demonstrate clinical utility and to qualify biomarker assays for use in drug development and as surrogate end points. The basic differences relate to the extent to which information (mechanistic, pharmacologic, observational, etc.) other than findings from randomized controlled trials can be used to support the use of the diagnostic. A closer look at the issue reveals that, under certain circumstances, almost everyone can accept lesser evidence, and so the real challenge is where to draw the line in a specific case.

Ideally, relevant diagnostics will be evaluated during drug development and be available for use in efficacy trials. Because of technical challenges, this is not yet generally the case. There will continue to be instances of “rescue” diagnostics. Additionally, the use of certain older drugs may be improved by incorporating diagnostics—the “retrofit.” Developing the evidence to support such uses will continue to be challenging.

Biomarker qualification is a promising new route for establishing novel diagnostics for use in drug development. The pathway for establishing new surrogate end points is challenging and not well defined.

The controversy over clinical utility of diagnostics in drug therapy is a reflection of the underlying progress in understanding the basis of variability in human responses to interventions; in this regard, it is good news. Most scientific progress is ushered in by disputes and disagreements; hopefully, these will not cause us to lose sight of the promise of safer, more effective drugs in the near future.