Biomarker Qualification: Toward a Multiple Stakeholder Framework for Biomarker Development, Regulatory Acceptance, and Utilization

S Amur[1], L LaVange[2], I Zineh[3], S Buckman-Garner[1] and J Woodcock[4]

The discovery, development, and use of biomarkers for a variety of drug development purposes are areas of tremendous interest and need. Biomarkers can become accepted for use through submission of biomarker data during the drug approval process. Another emerging pathway for acceptance of biomarkers is via the biomarker qualification program developed by the Center for Drug Evaluation and Research (CDER, US Food and Drug Administration). Evidentiary standards are needed to develop and evaluate various types of biomarkers for their intended use and multiple stakeholders, including academia, industry, government, and consortia must work together to help develop this evidence. The article describes various types of biomarkers that can be useful in drug development and evidentiary considerations that are important for qualification. A path forward for coordinating efforts to identify and explore needed biomarkers is proposed for consideration.

Optimal drug therapy is predicated on selecting the most appropriate patient-specific pharmacological intervention (e.g., considering a specific pathology) at an individualized dose, and at the right time in the patient's disease process. Reliable diagnostic, prognostic, predictive, pharmacodynamic, and pharmacokinetic biomarkers are critical to assure correct patient selection, drug dosing, and monitoring for safety and efficacy of many therapies in clinical practice. Novel molecular and genetic markers are increasingly being used to guide treatment, although challenges exist in the validation and clinical uptake of newly discovered biomarkers.[1–4] These challenges range from logistical to cultural, and are often methodological or evidentiary.[5,6]

The discovery, validation, regulatory acceptance, qualification, and use of biomarkers adequate for a variety of drug development and regulatory decision-making purposes are areas of tremendous interest and need. In fact, many of the key activities intended to both "de-risk" and optimize drug discovery and development (e.g., target identification, target engagement, safety prediction and assessment, proof of concept, enrichment) are highly reliant on the availability of credible biomarkers (which do not always exist).

Over a decade ago, the US Food and Drug Administration (FDA) called for "critical path research. . .to develop new, publicly available scientific and technical tools—including. . .biomarkers. . .—that make the [drug] development process itself more efficient and effective and more likely to result in safe products that benefit patients."[7] In this seminal report, the agency identified the need to develop new biomarkers and surrogate endpoints, and promulgated a framework for development of promising biomarkers.

There are generally two pathways through which biomarkers can be accepted by the FDA for use in therapeutic product development. First, during the drug development process (under an investigational new drug [IND] application), a pharmaceutical developer may engage directly with FDA review staff to reach agreement on the use of a particular biomarker in a given drug's development program. These interactions are critically important if the biomarker is intended to either serve as a surrogate endpoint or be used as a criterion for restricting use in the population. While this pathway may be efficient for drug developers, it has inherent limitations. The confidential discussions between the FDA and drug sponsor are not subject to the broad input of

[1]Office of Translational Sciences, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, Maryland, USA; [2]Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, Maryland, USA; [3]Office of Clinical Pharmacology, Office of Translational Sciences, Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, Maryland, USA; [4]Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, Maryland, USA. Correspondence: S Amur (Shashi.Amur@fda.hhs.gov)
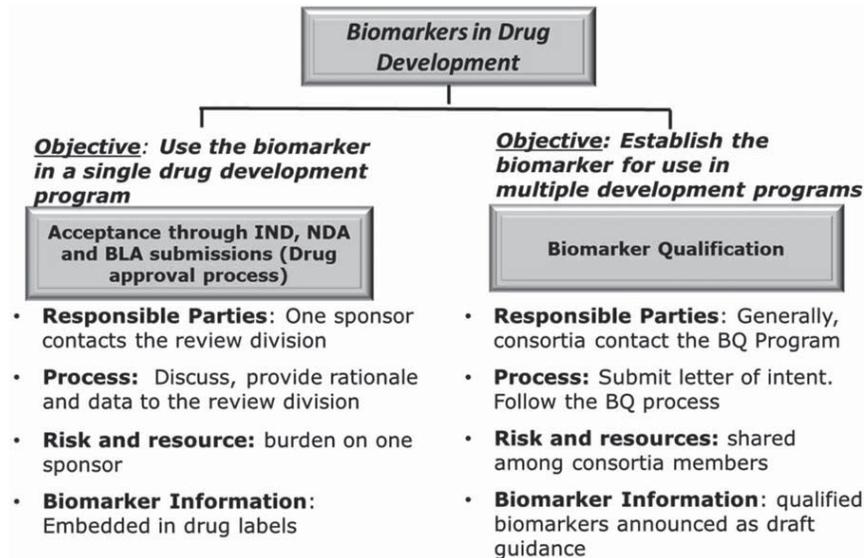
Figure 1 Pathways at FDA to integrate biomarkers into drug development.

the scientific community, and additionally, they may not contribute to ongoing agency policy development. In contrast, via a second fairly new mechanism, a pharmaceutical developer, patient- or disease-specific foundation, health research organization, or consortium may request regulatory "qualification" of a biomarker for a particular context of use through the FDA's Biomarker Qualification Program (BQP). The FDA defines qualification as "a conclusion that within the stated context of use, a biomarker can be relied upon to have a specific interpretation and application in drug development and regulatory review."[8] This mechanism is particularly advantageous for biomarkers with broader application across therapeutic areas and/or for which disparate data sources must be aggregated (e.g., through consortium efforts) to provide sufficient evidence of biomarker utility. Positive qualification decisions are publicly communicated to the drug development and research communities through regulatory guidance documents by the FDA's Center for Drug Evaluation and Research (CDER) (**Figure 1**). Publication is intended to facilitate widespread adoption and further evolution of the biomarker, and to contribute to increased understanding of how to develop evidence supporting biomarker use.

Biomarker qualification following this second pathway was established after the FDA's white paper publication.[7] After a pilot phase, the FDA formally established a biomarker qualification program to facilitate the regulatory–industry–academic interface on biomarker development, publishing a procedural guidance on CDER's Drug Development Tool Qualification programs.[8] The FDA also reached out to other regulators to establish international recognition of the qualification concept. Several biomarkers have been qualified through the FDA's program (**Table 1**)[9] as well as the European Medicines Agency's qualification program (**Table 2**).[10] The FDA Safety and Innovation Act (FDASIA), signed into law in July 2012, included a provision intended to advance the use of biomarkers in drug development and regulation. Consortia continue to be active in the area of bio-

marker discovery and development.[11–13] Despite these advances, ongoing challenges must be overcome to promote efficiencies in biomarker development, ensure rigor in the biomedical research enterprise *vis-à-vis* biomarker qualification, and create alignment on the evidentiary requirements for biomarkers intended to be qualified for regulatory purposes. Additionally, more clarity is needed as to when regulatory endorsement of biomarkers should follow the former or latter pathway as described above. Here, we present some of these high-level issues and opportunities and conclude with our recommendations for needed next steps.

TYPES OF BIOMARKERS
Our experience to date suggests a need for establishing a common lexicon for biomarkers and their uses. Standardizing these definitions could lead to efficiencies in defining specific use contexts in drug development and regulation. These, in turn, are expected to drive consensus development on evidentiary requirements for different biomarker uses. From a drug development perspective, the most commonly used biomarkers can be classified as diagnostic, prognostic, predictive, and response biomarkers (**Figure 2**).

Diagnostic biomarkers
Biomarkers that distinguish between patients with a particular disease and those who do not have the disease are commonly referred to as diagnostic biomarkers. Diagnostic biomarkers can be utilized to ensure that patients selected for a clinical study have the disease or the disease subset of interest.

Prognostic biomarkers
Prognostic biomarkers provide information on the likely course of disease in an untreated individual. A prognostic biomarker informs about the aggressiveness of the disease and/or the expectation of how a particular patient would fare in the absence of therapeutic intervention. Typically, prognostic biomarkers identify patients who are probabilistically at either higher risk for

Table 1 Biomarkers qualified through the Biomarker Qualification Program at FDA

Response (Safety)	Nonclinical Urinary biomarkers: Albumin, β2- Microglobulin, Clusterin, Cystatin C, KIM-1, Total Protein, and Trefoil factor-3	For the detection of acute drug-induced nephrotoxicity along with traditional clinical chemistry markers and histopathology in toxicology studies; Kim-1, albumin, clusterin and trefoil factor-3 as biomarkers of drug-induced acute kidney tubular alterations and Total Protein, β2 Microglobulin and Cystatin C as biomarkers of acute drug-induced glomerular alterations/damage and/or impairment of kidney tubular re-absorption.	4/14/2008 Drug-induced Nephrotoxicity Biomarkers Reviews
Response (Safety)	Nonclinical Urinary biomarkers: Clusterin, and Renal Papillary Antigen (RPA-1)	For the detection of acute drug-induced nephrotoxicity along with traditional clinical chemistry markers and histopathology in toxicology studies for drugs for which there is previous preclinical evidence of drug induced nephrotoxicity or where it is likely given the experience with other members of the pharmacologic class: Urinary Clusterin: for use in the detection of acute drug induced renal tubule alterations in male rats, particularly when regeneration is present: Urinary RPA-1 for detecting acute drug-induced renal tubule alterations, particularly in the collecting duct.	9/22/2010 Drug-induced Nephrotoxicity Biomarkers Reviews
Response (Safety)	Nonclinical serum/plasma biomarkers: Cardiac troponins T (cTnT) and I (cTnI)	For detection of cardio toxicity in nonclinical safety assessment studies.	2/23/2012 Drug-induced Cardiotoxicity Biomarkers Reviews
Diagnostic (Patient selection)	Clinical serum/bronchoalveolar lavage fluid biomarker: Galactomannan	As a sole microbiological criterion to classify patients as having probable invasive aspergillosis for enrollment in clinical trials of patients with hematologic malignancies or recipients of hematopoietic stem cell transplant who also have clinical and radiologic features consistent with invasive fungal infection.	10/24/2014 Invasive Aspergillosis Biomarker Reviews

adverse disease-related events or a faster rate of decline in their health status.

Predictive biomarkers
Unlike prognostic biomarkers, predictive biomarkers are linked to treatment, as they provide a forecast of the potential for a patient to respond, in some identified manner (which may be favorable or unfavorable), to one or more specific treatments.

Response biomarkers
Response biomarkers are dynamic assessments that show a biological response has occurred in a patient after having received a therapeutic intervention.

Safety biomarkers. Biomarkers that can indicate adverse effects on biology in response to treatment may have important roles in safety assessment during drug development, especially if sensitive to early pathophysiological changes well in advance of overt toxicity. Predictors of future toxicity are also critically valuable.

Pharmacodynamic biomarkers. Most response biomarkers used to guide drug development are indicators of the intended activity of the drug. These pharmacodynamic biomarkers often precede clin-

ical outcome measures of drug effect and need not be indicative of clinically meaningful effects in and of themselves.

Efficacy-response biomarkers or surrogate endpoints. Efficacy-response biomarkers are a subset of response biomarkers that predict a specific disease-related clinical outcome and can serve as surrogates for a clinical efficacy endpoint.

It is important to recognize that an overlap in the categorization of the biomarkers is encountered often. For example, a prognostic biomarker can also be predictive. It is also important to recognize the contribution of interindividual variation to differences in disease susceptibility, disease progression, and response to drugs/biologics and that biomarkers can be used to identify susceptibility to exposure of a drug or biologic in individuals.

Of the above types of biomarkers, validated surrogate endpoints have been the most difficult to establish. A surrogate endpoint is "a biomarker that is intended to substitute for a clinical endpoint. A surrogate endpoint is expected to predict clinical benefit (or harm or lack of benefit or harm) based on epidemiologic, therapeutic, pathophysiologic or other scientific evidence."[14] Surrogate endpoints such as blood pressure or Hb A1C can, and frequently are, used as evidence of benefit for traditional drug approvals.

Table 2 Biomarkers qualified by European Medicines Agency (EMA)

Type of biomarker	Qualified biomarkers	Context of use	Supporting information
Response (Safety)	Nonclinical Urinary biomarkers: Albumin, β2- Microglobulin, Clusterin, Cystatin C, KIM-1, Total Protein, and Trefoil factor-3	The urinary kidney biomarkers (Kim-1, Albumin, Total Protein, β2 Microglobulin, Urinary Clusterin, Urinary Trefoil Factor 3and Urinary Cystatin C) are considered acceptable in the context of nonclinical drug development for the detection of acute drug-induced nephrotoxicity, either tubular or glomerular with associated tubular involvement.	http://www.ema.europa.eu/docs/en_GB/ document_library/Regulatory_and_ procedural_guideline/2009/10/ WC500004205.pdf
Response (Safety)	Nonclinical Urinary biomarkers: Clusterin, Renal Papillary Antigen (RPA-1)	These biomarkers can be included along with traditional clinical chemistry markers and histopathology in GLP toxicology studies which are used to support renal safety in clinical trials: Urinary Clusterin is a biomarker that may be used by Applicants to detect acute drug-induced renal tubule alterations, particularly when regeneration is present, in male rats. Urinary RPA-1: may be used to detect acute drug-induced renal tubular alterations, particularly in the collecting duct, in male rats.	http://www.ema.europa.eu/docs/en_GB/ document_library/Regulatory_and_ procedural_guideline/2010/11/ WC500099359.pdf
Prognostic (Patient selection)	Clinical cerebrospinal fluid (CSF) biomarkers: CSF Aβ1-42, total-tau [t-tau], and phosphorylated tau [p-tau]	The CSF biomarker signature based on a low Aβ1-42 and a high-tau (total-tau and phosphorylated tau) qualifies to identify mild cognitive impairment (MCI) patients as close as possible to the prodromal stage of Alzheimer's Disease (AD), who are at risk to evolve into AD-dementia.	http://www.ema.europa.eu/docs/en_GB/ document_library/Regulatory_and_ procedural_guideline/2011/02/ WC500102018.pdf
Prognostic (Patient selection)	Clinical biomarker: hippocampal volume	Low hippocampal volume, as measured by MRI and considered as a dichotomized variable (low volume or not), appears to help enriching recruitment into clinical trials aimed at studying drugs potentially slowing the progress/conversion to AD dementia of the included subjects.	http://www.ema.europa.eu/docs/en_GB/ document_library/Regulatory_and_ procedural_guideline/2011/10/ WC500116264.pdf
Diagnostic (Patient selection)	Clinical biomarker: Amyloid related positive/negative PET signal	Amyloid related positive/negative PET signal qualifies to identify patients with clinical diagnosis of predementia AD who are at increased risk to have an underlying AD neuropathology, for the purposes of enriching a clinical trial population.	http://www.ema.europa.eu/docs/en_GB/ document_library/Regulatory_and_ procedural_guideline/2012/04/ WC500125018.pdf

In a different regulatory pathway to market, "accelerated approval," a surrogate endpoint that is considered to be "reasonably likely, based on epidemiologic, therapeutic, pathophysiologic, or other evidence, to predict clinical benefit," is used in the development program, and confirmation of benefit is required postapproval. In a study that examined FDA approvals of therapeutic agents between 2005 and 2012, it was reported that surrogate endpoints are used in around 45.3% of all New Molecular Entity (NME) approvals, the majority of them for traditional approvals.[15]

BIOMARKER EVALUATION AND EVIDENTIARY CONSIDERATIONS
Biomarker qualification process
A biomarker may be used in any drug development program for the qualified context of use upon qualification.[9,16] The qualified biomarker can be included in IND or New Drug Application/ Biologic License Application (NDA/BLA) submissions without having to reconsider and reconfirm the acceptance of the biomarker as long as:

- There are no serious issues with the study design and data.
- The biomarker is not used outside the qualified context of use.
- There are no new and conflicting scientific facts that might limit the use of the biomarker.

Qualification can contribute to acceptance and application of the biomarker across multiple drug development programs. Having qualified biomarkers that can be utilized by many sponsors will aid in optimizing drug development and evaluation and can facilitate cross-study comparisons.
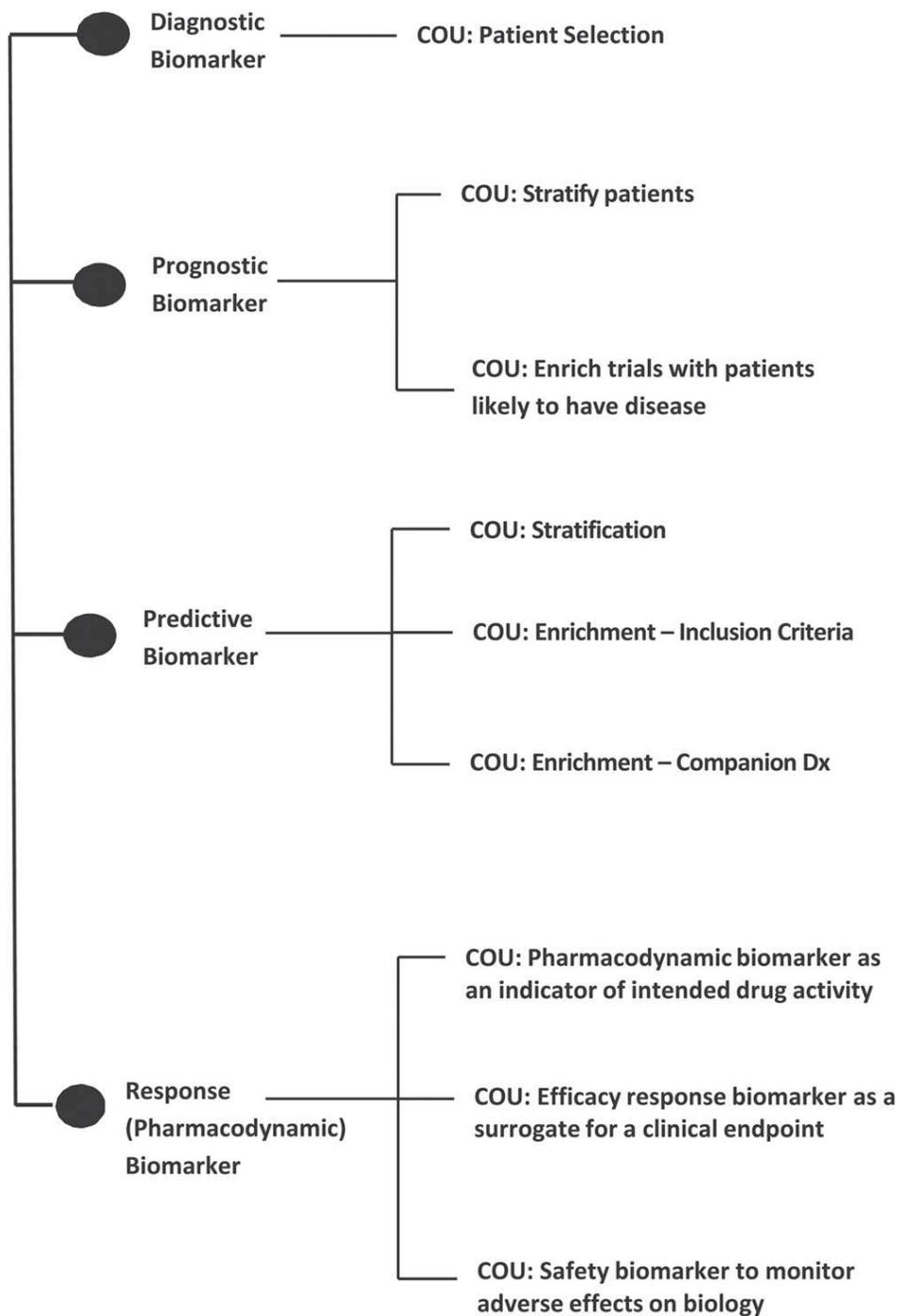
Figure 2 Biomarker types and potential contexts of use in drug development.

Qualification starts by defining the intended context of use (COU) of the biomarker in drug development and then determining the data sources and level of evidence required for that COU (**Figure 3**). Because the biomarker qualification process is iterative, the precise COU is often refined as biomarker development proceeds.

The CDER BQP works with submitters in a collaborative effort during biomarker development. In the *Initiation stage,*

after a letter of intent is submitted by the biomarker developer, the BQP provides guidance on clearly establishing the COU and identifying an analytically validated assay(s) to measure the biomarker(s) of interest (**Figure 4**). At this stage, a decision to accept the biomarker in the qualification program is made. For accepted proposals, a biomarker qualification review team (BQRT) consisting of individuals with appropriate scientific and regulatory expertise is assembled

Figure 3 The proposed context of use determines the level of evidence needed to support qualification.

to assess the data and interact with the submitter during the qualification process.

In the *Consultation and Advice* stage, preliminary data and analysis plans are reviewed and recommendations are provided to the submitter. This process is often iterative. Once the biomarker development is completed, a final qualification package is submitted and reviewed. This stage is referred to as the *review* stage.[8,17] A CDER qualification recommendation is announced as draft guidance in the Federal Register and posted on the FDA guidance webpage. The draft guidance is revised as needed and finalized after receiving public comment.

Evidentiary standards

The need for evidentiary standards to qualify biomarkers was identified in the FDA's Critical Path Initiative.[18,19] Clearly, different evidentiary standards are needed for the various types of biomarkers and their varying COUs. This concept of risk-based evidentiary requirements is shown in **Figure 5**. For example, prognostic biomarkers can be used to stratify patients during randomization and/or analysis or for enrichment of a specific subpopulation in clinical trials. Safety biomarkers may be used to supplement or substitute for certain safety tests, or they may be used to permit more aggressive dosing. The level of evidence needed to support qualification for stratification, enrichment, or various uses of safety biomarkers is not generally understood, although there is agreement on the need.

A prototypical framework for developing evidentiary standards for biomarkers was developed by Pharmaceutical Research and Manufacturers of America (PhRMA) and discussed in a workshop with the FDA and academia.[20] Tolerability of the risk introduced by use of the biomarker was identified as the key factor in determining the weight of evidence. Thus, it was proposed that the "specific purpose and context of use" would determine tolerability of risk and, thereby, the type and level of evidence needed to qualify any biomarker. This concept is similar to the "COU" employed by the BQP. With the goal of developing clear evidentiary standards, the FDA's Center for Food Safety and Applied Nutrition (CFSAN) in conjunction with the Center for Drug Evaluation and Research (CDER) reached out to the Institute of Medicine (IOM) in 2008 for advice on biomarker and surrogate endpoint evaluation in chronic disease. The IOM report recommended a three-part framework comprised of 1) Analytical validation—Can a biomarker be measured accurately? 2) Qualification—Is the biomarker associated with the
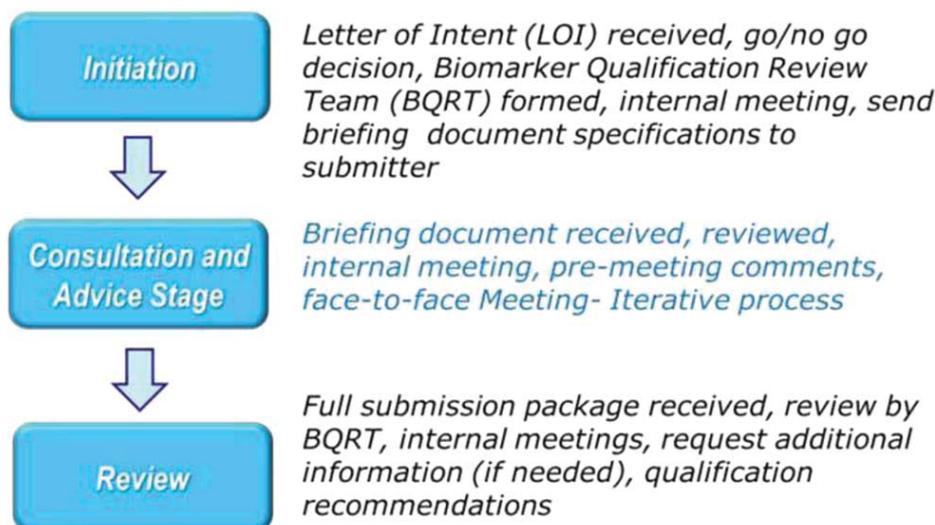

Biomarker Qualification Process

Initiation — Letter of Intent (LOI) received, go/no go decision, Biomarker Qualification Review Team (BQRT) formed, internal meeting, send briefing document specifications to submitter

Consultation and Advice Stage — Briefing document received, reviewed, internal meeting, pre-meeting comments, face-to-face Meeting- Iterative process

Review — Full submission package received, review by BQRT, internal meetings, request additional information (if needed), qualification recommendations

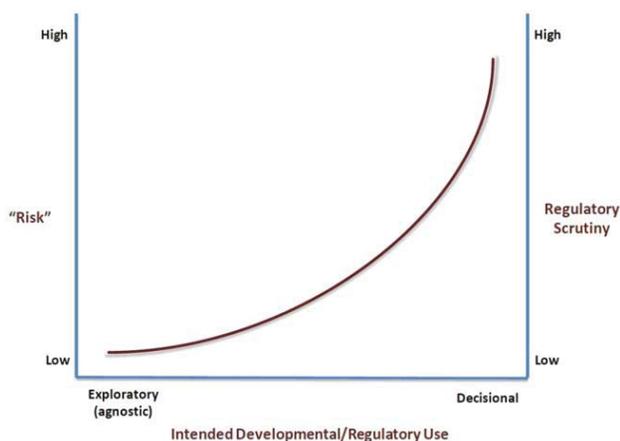**Figure 4** The biomarker qualification process.

Figure 5 Evidentiary requirements and level of engagement of regulators will likely be a function of the intended use for the biomarker in drug development and regulatory decision making (x-axis) and "risk" (y-axis; the impact of the decision to be made based on the biomarker results [e.g., restricting enrollment in clinical trials, shunting patients away from standard of care therapy, clinical decision making at the patient level]). The higher the risk or the more likely the results will be used to make definitive regulatory decisions (e.g., accelerated approval, full approval, inform labeling), the higher the likely regulatory scrutiny.

clinical endpoint of concern? 3) Utilization— What is the specific COU for biomarker qualification?[21] Of note, the use of the term "qualification" in this report differs from its use by CDER in the "qualification for regulatory use" process.

The IOM report recommended that the "FDA should use the same degree of scientific rigor for evaluation of biomarkers across regulatory areas..." i.e., for food health claims as well as medical products. The group "recognized that the decision to use a biomarker for a regulatory purpose is a subjective one" and that "many biomarkers work perfectly well in particular contexts while failing in other contexts."[21] However, the IOM report did not provide the actual elements of the decision framework or the criteria by which regulatory decisions should be made.

Many factors are involved in evaluating a biomarker for qualification for regulatory purposes. One important exercise is an assessment of the confidence in whether the biomarker adequately predicts what is intended to predict vs. the risk if it does not. For example, the benefit of using a surrogate endpoint can be significant, allowing for an earlier and often easier determination of a drug's efficacy. If the surrogate endpoint is not adequately predictive, however, its use may result in approval of a drug that is not effective. Thus, the level of evidence needed to support the confidence in a surrogate endpoint is likely to be high. Currently, evidentiary standards are not defined for all types of biomarkers for various COUs. The development of evidentiary standards necessitates the participation of multiple stakeholders including scientists from academia, the pharmaceutical industry, and government to help aggregate the requisite information.

Data submission considerations
The FDA has learned from the submitted biomarker qualification packages and has begun to develop a framework for the types

of evidence needed for qualification. Such evidentiary considerations are informed by factors including, but not limited to:

- The use of the biomarker (e.g., prognosis, surrogate endpoint, safety monitoring).
- The COU of the biomarker for use in drug development (e.g., predictive biomarker for stratification vs. surrogate for use in lieu of a clinical endpoint).
- Biological rationale for use of the biomarker (if available) including a comprehensive understanding of the causal pathway of the disease process, and how the biomarker is positioned in the disease pathway.
- Characterizations of the various relationships among the biomarker, the clinical outcomes, and the treatment (where applicable) required for the proposed COU.
- Assay considerations (analytically validated method and understanding of potential sources of variability in the measurement). These include reliability, reproducibility, sensitivity, and specificity considerations.
- Type of data available to assess the strength of association of the biomarker with its proposed clinical outcome: retrospective or prospective, registry data, and/or randomized controlled trial (RCT) data.
- Reproducibility of data (need for test dataset and confirmatory dataset). This factor is especially important when published data are selected since only 20–25% data have been reported to be reproducible.[5] Also, there seems to be a consensus that reproducibility is a concern in science.[22,23] The lack of reproducibility may be related to small sample sizes or low statistical power in the studies.[24]
- Use of appropriate, prespecified statistical methods to demonstrate the hypothesized relationships for the COU.
- Strength of evidence: the level of evidence depends on the type of biomarker and its COU.

Qualifying pharmacodynamic biomarkers
Generally speaking, pharmacodynamic biomarkers are used in early drug development and their use in this setting has little associated patient or societal risk. Therefore, such biomarkers are often used in drug development without regulatory qualification. Nevertheless, pharmacodynamic biomarkers may evolve over time to become surrogate endpoints; so appropriate rigor and standardization in their definition and collection can help advance a particular disease area. In other words, while the use of pharmacodynamic measures for exploratory purposes is common in drug development, quality experimental designs and conduct could be quite enabling in building a knowledge base for use in other contexts over time.

Qualifying safety biomarkers
The rigor needed to use a new safety biomarker for drug development or regulatory decision-making depends on the COU. Safety biomarkers that are used to supplement traditional safety monitoring are basically exploratory and can generally be introduced into trials without qualification. In contrast, safety biomarkers that are intended to provide an additional safety margin, and to be relied

upon for patient safety assurance, of necessity have a higher evidentiary bar. Most existing biomarkers used for patient safety monitoring in trials have been in clinical use for a decade or more, and were adopted for drug development use after long clinical experience. This very long lag period does not meet the needs of modern drug development, and thus there is great interest in new safety monitoring capabilities. The FDA has qualified several urinary biomarkers for drug-induced renal toxicity in animals. These biomarkers are now undergoing the clinical evaluation needed for use in drug development. These biomarker programs were initiated, appropriately, because existing monitoring for drug-induced renal toxicity in humans is insensitive and unlikely to adequately protect patients. As a result, promising drugs with equivocal animal signals of drug-induced renal toxicity were unlikely to be taken into the clinic. However, the direct result of focusing this unmet need is that evidence development is likely to be difficult, since no gold standard comparison exists, and the stakes for reliance on the safety biomarkers are high. Qualification of new safety biomarkers for regulatory use is a new and evolving field.

Qualifying prognostic and predictive biomarker qualification: statistical considerations

In our experience, ostensibly prognostic and predictive biomarkers are among the most commonly submitted types of biomarkers to the CDER BQP and deserve some additional consideration here. The choice of statistical methods for biomarker qualification will be a function of the context of use for a particular biomarker.[25] While specific guidelines for statistical analyses to support qualification for the different COUs have not been developed or agreed upon by all interested parties, there are a few basic principles that have been used by the CDER to date to guide the statistical evaluation of biomarkers. In general, the statistical rigor required for qualifying a prognostic biomarker should be less than that required for a predictive biomarker, with qualification as a surrogate endpoint being the most demanding. Within each of these three broad categories of use, there are gradations of rigor expected, each depending on other aspects of the qualification process, such as the availability of additional biomarkers serving the same purpose, past experience with and/or validation of the biomarker's use in similar contexts, etc. In addition to the choice of statistical methods, adequacy of data sources for characterizing the relationship between the biomarker and other variables is an important factor to consider in determining the feasibility of the qualification process; in some cases, qualification may not be feasible until additional data or other types of data (e.g., from RCTs) are available. Descriptions of the available data sources and details of the statistical analyses proposed to support qualification should be included in the statistical analysis plan and submitted to the agency for review. Following are general guidelines for determining the statistical methods appropriate for qualifying a prognostic or predictive biomarker.

Prognostic biomarkers.

- Assumptions required for qualification include consensus about the clinical outcome(s) considered essential for characterizing the disease course (i.e., outcomes that describe the way

a patient feels, functions, or survives and upon which regulatory decisions can be based) and some knowledge about the nature of the relationship between the biomarker and the clinical outcome. The first step in qualification is then to show that a relationship exists and to estimate the strength of that relationship. For example, if past experience tells us that the relationship is approximately linear, with increases in the biomarker signifying increases in the clinical outcome by uniform amounts, then a statistically significantly positive Pearson's correlation coefficient will usually suffice to characterize the relationship. More complex relationships may require modeling to establish the prognostic ability of the biomarker with respect to clinical events that are hallmarks of the disease.

- Once the relationship between the biomarker and clinical outcome has been established and reasonably well characterized, specifics about the intended use come into play. Prognostic biomarkers may be used to stratify patients into homogeneous subgroups at the time of randomization and/or analysis (randomization stratification factors should always be used as stratification factors during analysis, but not vice versa) in order to improve the power to detect treatment effects in clinical trials. The "value" of the stratification will depend on the level of homogeneity within each stratum (and, therefore, the heterogeneity across strata). Required for stratification is the ability to define discrete values of the biomarker for use in grouping patients into homogeneous strata, if the original measurement scale is an interval scale. Alternatively, an internally scaled biomarker may be used for covariate adjustment in analysis of covariance (ANCOVA) or other model-based analyses to achieve the same objective, namely, increased power, without the need for categorization. For use in stratification or covariate adjustment, no additional analyses beyond analyses showing a sufficiently strong relationship between the biomarker and outcome are typically required for qualification, due to the fact that the patient population targeted for enrollment in the trial is not being affected by use of the biomarker.

- Use of a prognostic biomarker to enrich a study population by enrolling a specific number or proportion of patients within a specified range of values for the biomarker will usually require additional evidence for qualification. In the extreme, a prognostic biomarker may be used to enroll only those patients in the target range. In either case (enriching the population or limiting enrollment altogether), the "value" of the enrichment strategy may be an important factor in the evaluation of the biomarker for qualification. One method for demonstrating value is to consider the increase in power (or decrease in sample size) the strategy affords. Providing such evidence will usually require simulating various clinical trial scenarios corresponding to reasonable assumptions about the design factors in addition to incorporating the parameter estimates characterizing the biomarker–outcome relationship. As with stratification, discrete values or cutpoints are usually designated in order for a biomarker to be used to modify enrollment criteria, although the analysis conducted for qualification can examine several choices for that designation. The strength of the biomarker–outcome relationship will determine the value provided by the

enrichment strategy, and the decision as to how much added value is enough for qualification will be subjective and will likely depend on other factors, e.g., the availability of other enrichment factors, the difficulty enrolling or completing a trial, etc.

- Note that it is not necessary to have data from randomized clinical trials available for use in establishing a biomarker as prognostic. Observational studies that include biomarker measurements at baseline and long-term follow-up for clinical outcome assessment may suffice, provided the population is consistent with future clinical trial populations and the completeness and quality of the data collection and outcome ascertainment is deemed sufficiently high.

Predictive biomarkers.

- As defined above, a predictive biomarker is measured prior to therapy and identifies patients susceptible to a particular drug effect, whether benefit or risk. The most straightforward way to establish that a biomarker is predictive is to demonstrate a statistically significant interaction between treatment and biomarker status in the context of the analysis used to demonstrate an overall treatment effect. Ideally, the biomarker hypothesis would have a scientific basis, be prespecified as part of an RCT, and be prospectively evaluated on samples from the full study population with appropriate Type I error control. In practice, most predictive biomarker evaluations are conducted retrospectively, and some caveats apply. Caution is needed if the samples available to determine biomarker status are from a convenience sample, particularly if the treatment effect in the sample differs from that in the overall study population, and if the marker positive and marker negative subgroups differ on important prognostic variables. Power for interaction tests is often low, particularly when not anticipating the need for such a test, requiring subjective judgment to be used in determining whether a clinically meaningful difference exists between the treatment effects observed within each biomarker-defined subgroup. Interpretation of retrospective subgroup analyses is complex and may require confirmation. Depending on what is known about a biomarker *a priori*, it may be difficult to enroll adequate numbers of marker-negative patients in a study designed to establish its predictive capability. These and other considerations are discussed more fully in the context of oncology trials, an area where predictive biomarkers have proven very useful in advancing precision medicine, and regulatory experience is available to guide the qualification process.[26]
- Note that predictive biomarkers need not be prognostic—the differential treatment effects are attributable to the ability of the biomarker to separate patients more likely to receive benefit or experience risk from the treatment and may not be predictive of the clinical course of the disease for a given patient.

Regulatory considerations for qualification of surrogate endpoints

While surrogate endpoints offer many potential advantages, the evidentiary bar for their development and acceptance is relatively high, given the consequences of clinical and regulatory decision-making based on their use, and their checkered history. Surrogate endpoints can reduce the cost and time required to develop new therapeutics; expedite patient access to promising treatments, allow for timely assessment; and stimulate researchers and developers who are searching for novel therapies.[21,27] Use of surrogate endpoints may provide the only practical way to evaluate therapies intended to change the course of slow, progressive diseases or to prevent expression of a disease that has a long latency period.[28]

Regardless of these advantages, diseases are heterogeneous and complex and it may not be possible to find one surrogate endpoint for each disease.[30] Additionally, the lack of a "gold standard" clinical outcome measurement hobbles surrogate endpoint development for many conditions.[28]

The history of surrogate endpoint development has been one of tremendous successes and failures.[31] Failures include the use of antiarrhythmic drugs for suppression of postmyocardial infarction ventricular premature beats, for which the outcome trial CAST demonstrated a significant increase in mortality, and an attempt to raise high-density lipoprotein (HDL) cholesterol to prevent cardiovascular events. There have been multiple such failures over the last three decades, along with compelling successes. There are many "failure modes" for a potential surrogate endpoint. A surrogate endpoint may overestimate or underestimate the effect of a therapeutic on clinical outcomes for a variety of reasons, including, but not limited to: 1) The biomarker is not in the causal pathway of the disease; 2) multiple pathways may influence the outcome; 3) a therapeutic agent affects a biomarker favorably and/or affects the clinical outcome independently; 4) the therapeutic has off-target effects.[21,32]

The statistical community has put forth various proposals for "validation" of surrogate endpoints, but evidentiary standards for acceptance and qualification of surrogates for both traditional and accelerated approval need further development. These criteria are likely to include both mechanistic and statistical criteria (see statistical considerations section for details). The FDA has utilized numerous new surrogate endpoints over the past several decades, particularly in serious and life-threatening diseases (both traditional and accelerated approvals). However, the criteria for use of surrogate endpoints for chronic progressive diseases and for prevention indications need to be clarified.

In order to highlight the complex, multifactorial, and often idiosyncratic nature of regulatory acceptance of surrogate endpoints, we highlight two examples.

Hemoglobin A1C (Hb A1c). The use of Hb A1c for monitoring the degree of control of glucose metabolism in diabetic patients was proposed in 1976.[33] A meta-analysis of 23 publications in 1999 confirmed a strong association between Hb A1C and mortality in type 2 diabetes mellitus.[34] In 2006, the American Diabetes Association (ADA), World Health Organization (WHO), and American College of Endocrinologists (ACE) published recommendations that Hb A1c level is considered the "gold standard" in assessment of metabolic control and that the treatment of both type 1 and type 2 diabetes mellitus should be aimed at the

specific level of Hb A1c described.[35] The FDA accepted Hb A1c as the primary efficacy endpoint for approval of drugs in the mid-1990s and published a guidance on the use of Hb A1c as the surrogate endpoint to support an indication of glycemic control,[36] and improvement in glucose control captured using Hb A1c is an accepted biomarker for full approval. It took about two decades after a publication that reported the utility of Hb A1c for monitoring glucose metabolism in diabetes patients to establish Hb A1c as a surrogate endpoint.

HIV viral load (HIV RNA). A report of correlation of viremia measurement and its relation to disease progression was published in 1991[37] and an article showing prediction of progress to AIDS using serum HIV-1 RNA and CD4+ count was published in 1995.[38] In 1996, the AIDS Clinical Trials Group (ACTG) 116B/117 Study Team and the ACTG Virology Committee Resistance and HIV-l RNA Working Groups published that baseline HIV-1 RNA level was an independent predictor of disease progression,[39] and in 1997, the FDA convened an advisory committee meeting to consider the use of changes in HIV RNA levels as endpoints in clinical trials supporting traditional approval of antiretrovirals. This was followed by a publication in 1999 by the FDA that concluded that HIV RNA "appears to be a rigorous benchmark for assessing the efficacy of antiretroviral regimens."[40] A draft guidance was issued by the FDA in 2002 that allowed accelerated approvals to be based on shorter-term reductions in HIV RNA (e.g., 24 weeks) by a drug, while traditional approvals to be based on trials that show a drug's contribution toward durability of HIV RNA suppression (e.g., for at least 48 weeks).[41] It took about 7 years to establish HIV viral load as a surrogate endpoint. Before HIV viral load was accepted as a surrogate, risk of death among patients with AIDS was the clinical endpoint used for full approval; accelerated approval was granted based on drug-induced changes of CD4+ T-cell counts for didanosine and zalcitabine. HIV viral load as a surrogate biomarker offers several advantages including ease of measurement, earlier evaluation of drug activity, and rapid identification of loss of response[40] for both accelerated as well as full approval of antiretrovirals.

Statistical considerations in qualification of surrogate endpoints

Biomarkers that indicate response to treatment and also predict the clinical outcome at a future time have tremendous potential to facilitate the drug development process by providing earlier readout on a drug's activity than the longer-term or more difficult to assess clinical outcomes, as described earlier. These biomarkers are not treatment-specific but, rather, disease-specific, and they often play an important regulatory role by serving as the basis for accelerated approval in cases of unmet medical need. As such, the statistical rigor required to establish a biomarker as surrogate is quite high and has been the subject of much discussion in the literature. The least controversial but often unattainable criterion for establishing surrogacy is the Prentice criterion.[42] This criterion can be assessed in a single clinical trial and is satisfied if the treatment effect with respect to the clinical outcome is wholly explained by the treatment effect with respect to the biomarker. For example, assume an ANCOVA model is the appropriate context for assessing the treatment effect with respect to the clinical outcome, and significant treatment effects have been demonstrated with respect to both the biomarker and the clinical outcome. If the biomarker is then included in the ANCOVA as a covariate, not only is it a significant predictor of the clinical outcome, but the treatment effect is no longer significant, after controlling for the biomarker. That is, the biomarker has assumed all of the "significance" of the treatment in predicting the clinical outcome.

The Prentice criterion has proven to be difficult to satisfy in practice, due primarily to difficulties with the last step, i.e., proving that the treatment effect is null once the biomarker has been accounted for. This finding is often biologically unlikely, for reasons discussed above. Other methods based on single trials have been proposed, including methods based on likelihood theory[43,44] and on causal inference,[45,46] each with certain advantages and disadvantages. Validation methods based on multiple clinical trials were introduced around 2000[47,48] in order to provide replication of the observed correlations between the biomarker and clinical outcomes and reduce variability in estimating the relationships.

A practical multiple-trial approach for establishing surrogacy that the agency has relied on in evaluating efficacy response biomarkers is to demonstrate that treatment effects with respect to the biomarker are correlated or predictive of treatment effects with respect to the clinical outcome.[49] This approach usually proceeds in two steps. First, individual changes in the biomarker (or biomarker response, if measured as an event) attributable to treatment are shown to be correlated with or predictive of individual changes in the clinical outcome (or clinical response). Correlational analyses of patient-level outcomes are usually adequate for normally distributed biomarker and outcome variables, but response or time-to-event variables require modeling at this step. Second, a meta-analysis of RCTs that include measurements on both the biomarker and the clinical outcome demonstrate a significant relationship between treatment effects with respect to the two variables. This second step is usually more difficult than the first and will depend on both the number of RCTs available and the strength of the treatment effect in each. Consequently, it is possible to qualify a biomarker as satisfying only the first step. While not achieving full surrogacy status, the biomarker may still be useful for giving an early readout of promising therapies. For use as the basis of full regulatory approval, however, both steps are usually required, at a minimum. It should be noted that no absolute standards for surrogacy have been established by the FDA, nor has general agreement been reached among academic and other researchers on a single criterion for this purpose.

NEXT STEPS

Scientifically accepted biomarkers offer the promise of reducing the length, cost, and uncertainty of drug development and potentially unlocking targets for precision medicine. They do so by providing rapid, reliable information on such key metrics as: 1) whether a patient is susceptible to or already has a disease;

2) whether a patient is likely to respond to a treatment or develop side effects; 3) how far a disease has progressed; 4) whether a potential treatment targets the correct disease pathway (proof-of-concept); and 5) whether a treatment has worked. Among the most important barriers to the development and/or qualification of biomarkers are: i) lack of standardized methods for measuring new biomarker(s) and resulting absence of reliable evidence about their performance (that is, how well they predict what they are thought to predict); ii) lack of generally accepted evidentiary standards for qualifying new biomarkers for a large range of diseases and purposes; iii) inadequate prioritization and coordination of the limited public and private resources available to identify and qualify biomarkers in areas of unmet need; and iv) inadequate scientific information on the causes, biochemical pathways, and natural histories of certain diseases, making identification of disease-specific biomarkers in those diseases difficult or impossible. While the FDA has an important role to play in qualifying potential biomarkers for regulatory use, it does not have all the requisite expertise, resources, or—in the case of inadequate scientific research—the mission, to address these key barriers to biomarker development.

As the 2012 President's Council of Advisors on Science and Technology (PCAST) report[50] recognizes, advancing new drug development tools like biomarkers cannot be the sole responsibility of the FDA, the NIH, or industry alone. Instead, it requires a coordinated partnership involving the larger non-federal community, including industry, academic researchers, patient and consumer groups, physicians, and insurance companies. The report specifically suggested that the external community "develop recommendations for the conceptual framework and scientific standards that should be applied to the qualification process." Sadly, very few examples of this type of robust coordinated partnership exist to date. There are currently a handful of public–private partnerships and consortia working to develop biomarkers in specific areas. However, additional actions are needed to direct these efforts to the highest priority needs, provide infrastructure and resources to facilitate aggregation and curation of biomarker data, assure that consortia are working collaboratively to share critical information, establish clear evidentiary standards for qualifying the biomarkers that are developed, and accelerate the process for qualifying biomarkers, including those that can be used as surrogate endpoints.

While the ultimate decisions regarding qualification of proposed biomarkers currently rest with the FDA, the process could be accelerated if diverse experts and stakeholders came together to identify and prioritize needs, gather relevant scientific information, and develop community consensus in an open and transparent process (**Figure 6**).

Specifically, the following needs seem to be overarching:

- Greater understanding and articulation of levels of evidence needed to support various contexts of use.
- Enhanced data sharing and collaborative efforts among consortia.
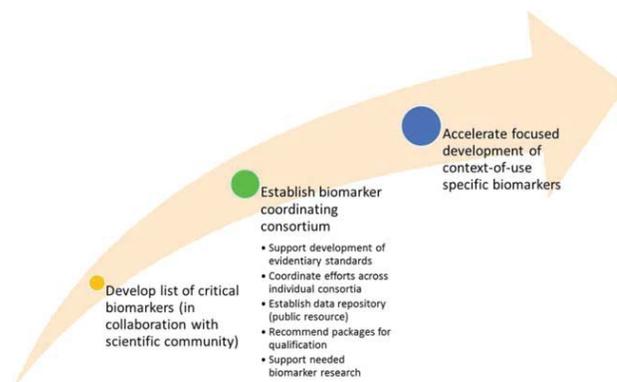- Development of qualification packages that support a defined COU.

Figure 6 Toward a multiple stakeholder framework for biomarker development.

- Data and specimen repositories which can support expanded contexts of use for biomarkers once additional data is aggregated.
- More communication about the value and progress made by consortia efforts.
- Greater clarity around levels of evidence for qualification of different types of biomarkers—this takes the entire scientific community—not just the FDA.

The level of evidence required for a biomarker should be calibrated to how it will be used. For example, very strong evidence, i.e., multiple replicate studies, might be necessary to rely on a biomarker for public health decisions that affect millions of Americans, whereas some tradeoffs that have less certainty may be acceptable to satisfy an unmet need. We need to have the help of industry, government entities, and academia to help us determine what levels of evidence befit different types of biomarkers, based on their context of use. To this end, we have asked the scientific community to begin hosting public workshops so the experts may help us build this evidentiary framework.

To support this effort, on February 13, 2015, the FDA issued a Federal Register Notice to identify potential biomarkers for qualification and describe contexts of use to address areas important to drug development. It is hoped that feedback from this request will help fuel broader discussions and prioritization efforts. Similarly, PhRMA conducted a survey of industry to obtain their feedback as well, and publication of those results is anticipated.

In an effort to enhance the attention on the efforts of consortia who have worked to aggregate data on prospective biomarkers that may be suitable for qualification, the FDA has developed a Letter of Support concept. This mechanism provides public attention to focus on specific biomarkers that need additional evidence to support qualification for a particular context of use. While a Letter of Support is distinct from qualification, it is hoped that it will bring greater visibility to potential drug development tools for which the scientific community can help provide additional evidence to support future qualification.

Due to the number of competing efforts underway to support both preclinical and clinical biomarker development, coupled with the resources needed to adequately and appropriately

provide confirmatory evidence to support qualification, better identification of potentially meaningful targets to support coordination among consortia is critical. Remarkably, the Predictive Safety Testing Consortium (PSTC), Safer and Faster Evidence-Based Translation (SAFE-T) Consortium, and the Biomarkers Consortium have been pioneers in demonstrating how this can be done in the setting of kidney injury biomarkers. PSTC developed some of the requisite evidence to support preclinical biomarkers that are also being evaluated in clinical trials conducted through SAFE-T and the Biomarkers Consortium. However, this is only one example, and many more are needed.

As alluded to in the 2012 PCAST report, an overarching "uber consortium" could serve a variety of purposes, including to:

a. Coordinate existing partnerships and consortia so that they effectively direct their efforts toward development and qualification of the priority biomarkers identified by the FDA and the scientific community;
b. Develop and maintain the infrastructure for aggregation and curation of relevant biomarker data to expand qualification of priority biomarkers (e.g., develop data and/or sample repositories);
c. Conduct substantive reviews and make recommendations to the FDA on the sufficiency of data packages developed by industry and public–private partnerships to support qualification of new biomarkers; and
d. Support biomedical research that is necessary as the basis for development of new biomarkers.

We remain optimistic that with focused, coordinated attention and prioritization of putative biomarkers for development, coupled with greater clarity regarding the level of evidence needed to support qualification, attention to reproducibility of studies, and data quality, that great strides can be taken to help streamline medical product development. The FDA called for this over a decade ago in the call to action for the Critical Path Initiative. While progress has been made, we still have more to achieve and it must be done collaboratively with government, academia, and industry at the table together to advance the needed science.

CONFLICT OF INTEREST
The authors report no conflicts of interest.

1. Pirmohamed, M. Acceptance of biomarker-based tests for application in clinical practice: criteria and obstacles. *Clin. Pharmacol. Ther.* **88**, 862–866 (2010).
2. Zineh, I. & Huang, S.M. Biomarkers in drug development and regulation: a paradigm for clinical implementation of personalized medicine. *Biomark. Med.* **5**, 705–713 (2011).
3. Fuzeny A.K., Levin, J., Chan, M.M. & Chan, D.W. Translation of proteomic biomarkers into FDA approved cancer diagnostics: issues and challenges. *Clin. Proteomics* **10**, 13 (2013).
4. Salter, H. & Holland, R. Biomarkers: refining diagnosis and expediting drug development — reality, aspiration and the role of open innovation. *J. Intern. Med.* **276**, 215–228 (2014).
5. Ioannidis, J.P. Why most published research findings are false. *PLoS Med.* **2**, 696–701 (2005).
6. Begley, C.J. & Ioannidis, J.P. Reproducibility in science: improving the standard for basic and preclinical research. *Circ. Res.* **116**, 116–126 (2015).
7. FDA. Innovation or stagnation: challenges and opportunity on the critical path to new medical products. <http://www.fda.gov/ScienceResearch/SpecialTopics/CriticalPathInitiative/CriticalPathOpportunitiesReports/ucm077262.htm>. (2004).
8. Guidance for Industry and FDA Staff Qualification Process for Drug Development Tools. <http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM230597.pdf>. (January 2014).
9. FDA-Biomarker Qualification webpage. <http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm284076.htm>.
10. EMA-Qualification of novel methodologies for medicine development webpage. <http://www.ema.europa.eu/ema/index.jsp?curl=pages/regulation/document_listing/document_listing_000319.jsp&mid=WC0b01ac0580022bb0>.
11. Parekh, A., Buckman-Garner, S. & McCune, S. Catalyzing the critical path initiative: FDA's progress in drug development activities. *Clin. Pharmacol. Ther.* **97**, 221–233 (2015).
12. Wholley, D. The Biomarkers Consortium. *Nat. Rev. Drug. Discov.* **13**, 791–792 (2014).
13. Stephenson, D. & Sauer, J-M. The Predictive Safety Testing Consortium and the Coalition Against Major Diseases. *Nat. Rev. Drug. Discov.* **13**, 793–794 (2014).
14. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin. Pharmacol. Ther.* **69**, 89–95 (2001).
15. Downing, N.S., Aminawung, J.A., Shah, N.D., Krumholz, H.M. & Ross, J.S. Clinical trial evidence supporting FDA approval of novel therapeutic agents, 2005–2012. *JAMA* **311**, 368–377 (2014).
16. FDA. DDT webpage. <http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/default.htm>.
17. FDA-CDER-MAPP-Drug Development Tool Qualification Programs. <http://www.fda.gov/downloads/AboutFDA/CentersOffices/OfficeofMedicalProductsandTobacco/CDER/ManualofPoliciesProcedures/UCM407254.pdf>.
18. Critical Path Opportunities List. <http://www.fda.gov/downloads/ScienceResearch/SpecialTopics/CriticalPathInitiative/CriticalPathOpportunitiesReports/UCM077258.pdf>.
19. Critical Path Opportunities Report. <http://www.fda.gov/downloads/ScienceResearch/SpecialTopics/CriticalPathInitiative/CriticalPathOpportunitiesReports/UCM077254.pdf>.
20. Altar, C.A. *et al*. A prototypical process for creating evidentiary standards for biomarkers and diagnostics. *Clin. Pharmacol. Ther.* **83**, 368–371 (2008).
21. IOM Report on Evaluation of Biomarkers and Surrogate Endpoints in Chronic Disease. <http://iom.edu/Reports/2010/Evaluation-of-Biomarkers-and-Surrogate-Endpoints-in-Chronic-Disease.aspx>.
22. Prinz, F., Schlange, T. & Asadullah, K. Believe it or not: how much can we rely on published data on potential drug targets? *Nat. Rev. Drug Discov.* **10**, 712–713 (2011).
23. Collins, F. & Tabak, L.A. NIH plans to enhance reproducibility. *Nature* **505**, 612–613 (2014).
24. Button, K.S. *et al*. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).
25. Chakravarty, A.G. Regulatory issues in use of biomarkers in drug development. In *Clinical and Statistical Considerations in Personalized Medicine* (eds. Carini, C., Sandeep, M. & Chang, M.) 313–331 (Chapman & Hall-CRC Press, Boca Raton, FL, 2014).
26. Chakravarty, A.G., Rothmann, M. & Sridhara, R. Regulatory issues in use of biomarkers in oncology trials. *Stat. Biopharm. Res.* **3**, 569–576 (2012).
27. Lathia, C.D. *et al*. The value, qualification, and regulatory use of surrogate end points in drug development. *Clin. Pharmacol. Ther.* **86**, 32–43 (2009).
28. Woodcock, J. Expecting the unexpected: the problematic nature of surrogate endpoints. The National Biomarker Development Alliance (NBDA) workshop VI held on December 1, 2014 in Washington D.C.

29. Katz, R. Biomarkers and surrogate biomarkers: An FDA perspective. *J. Am. Soc. Exp. Neurother*. **1**, 189–195 (2004).

30. Colburn, W.A. Optimizing the use of biomarkers, surrogate endpoints, and clinical endpoints for more efficient drug development. *J. Clin. Pharmacol*. **40**, 1419–1427 (2000).

31. Ciani, O. *et al*. Validation of surrogate endpoints in advanced solid tumors: systematic review of statistical methods, results, and implications for policy makers. *Int. J. Technol. Assess. Health Care* **30**, 312–324 (2014).

32. Fleming, T.R. & DeMets, D.L. Surrogate end points in clinical trials: are we being misled? *Ann. Intern. Med*. **125**, 605–613 (1996).

33. Koenig, R.J. *et al*. Correlation of glucose regulation and hemoglobin Alc in diabetes mellitus. *N. Engl. J. Med*. **295**, 417–420 (1976).

34. Groeneveld, Y. *et al*. Relationship between blood glucose level and mortality in type 2 diabetes mellitus: a systematic review. *Diabet. Med*. **16**, 2–13 (1999).

35. American Diabetes Association. Standards of Medical Care in Diabetes 2006. *Diabetes Care* **29**, S4–S42 (2006).

36. FDA. Hb A1C guidance. <http://www.fda.gov/downloads/Drugs/Guidances/ucm071624.pdf>.

37. Semple, M. *et al*. Direct measurement of viremia in patients infected with HIV -I and its relationship to disease progression and Zidovudine therapy. *J. Med. Virol*. **35**, 38–45 (1991).

38. Loveday, C. & Hill, A. Prediction of progression to AIDS with serum HIV-I RNA and CD4 count. *Lancet* **345**,790–791 (1995).

39. Coombs, R.W. *et al*. Association of plasma human immunodeficiency virus type 1 RNA level with risk of clinical progression in patients with advanced infection. *J. Infect. Dis*. **174**, 704–712 (1996).

40. Murray, J.S., Elashoff, M.R. & Iacono-Connors, L.C. The use of plasma HIV RNA as a study endpoint in efficacy trials of antiretroviral drugs. *AIDS* **13**, 797–804 (1999).

41. FDA HIV Guidance. <http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm070968.pdf>.

42. Prentice, R.L. Surrogate endpoints in clinical trials: definition and operational criteria. *Stat. Med*. **8**, 431–440 (1989).

43. Alonza, A. Prentice's approach and the meta-analytic paradigm: a reflection on the role of statistics in the evaluation of surrogate endpoints. *Biometrics* **60**, 724–728 (2004).

44. Alonza, A. & Molenberghs, G. Surrogate marker evaluation from an information theory perspective. *Biometrics* **63**, 180–186 (2007).

45. Frangakis, C. & Rubin, D. Principle stratification and causal inference. *Biometrics* **58**, 21–29 (2002).

46. Li, Y., Taylor, J.M.G. & Elliot, M.R. A Bayesian approach to surrogacy assessment using principal stratification in clinical trials. *Biometrics* **66**, 523–531 (2010).

47. Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D. & Geys, H. Statistical validation of surrogate endpoints: problems and proposals. *Drug Inf. J*. **34**, 447–454 (2000).

48. Gail, M.H., Pfeiffer, R., van Houwelingen, H.C. & Carroll, R.J. On meta-analytic assessment of surrogate outcomes. *Biostatistics* **1**, 231–246 (2000).

49. Burzykowski, T. & Buyse, M. Surrogate threshold effect: an alternative measure for meta-analytic surrogate endpoint validation. *Pharm. Stat*. **5**, 173–186 (2006).

50. PCAST report. <http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast_future_research_enterprise_20121130.pdf>.