Developing an Evidentiary Criteria Framework for Safety Biomarker Qualification

April 14–15, 2016

The Food and Drug Administration's (FDA) Center for Drug Evaluation and Research (CDER), in co-sponsorship with the Foundation for the National Institutes of Health (FNIH) Biomarkers Consortium, convened this public workshop. The workshop was aimed at creating alignment among scientific stakeholders including FDA, NIH, the biopharmaceutical industry, academic researchers, and patient groups regarding a proposed framework for determining the levels of evidence required to qualify biomarkers for use in drug development, with an emphasis on biomarkers used in determinations of clinical safety assessments.

The workshop focused on (1) a proposed general evidentiary criteria framework to support biomarker qualification along with its specific application to different contexts of use (COUs) related to clinical safety, and (2) several specific case studies involving qualifying clinical markers of toxicity in different organ systems.

DAY 1

Welcome

David Wholley, M.Phil.
Director of Research Partnerships, FNIH, and Manager, Biomarkers Consortium

Mr. Wholley welcomed the participants and noted that the meeting was being webcast and recorded. He supplied some background on the purpose and structure of FNIH. The Biomarkers Consortium was founded by FNIH, FDA, the National Institutes of Health (NIH), and pharmaceutical companies for the purpose of developing biomarkers to support regulatory decision making in the context of precompetitive projects covering a broad range of disease. The Consortium has been involved in 22 projects covering many different diseases. Projects have ranged from incorporating modern clinical endpoints into antimicrobial trials to working out how novel biomarkers can be integrated in clinical trial designs.

The Consortium's executive committee started discussing a framework for evidentiary criteria about a year and a half ago. The discussion centered on the need for evidentiary criteria to reduce uncertainty in the qualification process. Related activity was taking place in Congress as part of the 21st Century Cures Act and in the Critical Path Institute (C-Path), a partner of the Biomarkers Consortium.

Last summer, a team comprising representatives of FNIH, the Pharmaceutical Research and Manufacturers of America (PhRMA), C-Path, pharmaceutical and biotechnology companies, and FDA began drafting the documents disseminated in advance of this meeting. In addition, sections on statistical and analytical considerations for validation of biomarkers were contributed from working groups that grew from the 2015 workshop co-sponsored by the FDA, C-Path, and the University of Maryland's Center for Excellence in Regulatory Science and Innovation. The framework document went

through many iterations within this core writing group. Prior to the meeting, the document was reviewed by multiple stakeholder groups to broaden the input.

Mr. Wholley credited Joseph Menetski, Ph.D., deputy director of FNIH, with helping with the development of case studies on qualification of safety biomarkers for hepatotoxicity, drug-induced kidney damage, and drug-induced vascular injury. The case studies were presented during the meeting.

Christopher Leptak, M.D., Ph.D.

Acting Director, Office of New Drugs Regulatory Science Program, and Co-Director, Biomarker Qualification Program, Office of Translational Sciences, CDER, FDA

Dr. Leptak said that this meeting was not an isolated effort; numerous meetings have been convened to investigate evidentiary standards or criteria to support integration of biomarkers into clinical trials. Various planning committees helped develop the meeting program, and several small groups focused on statistical and analytical aspects. Those efforts will likely lead to the publication of white papers, journal articles, and ultimately, FDA guidance(s).

One meeting, convened by the Brookings Institute, occurred in October 2015. It covered various challenges in the biomarker space and included discussions related to the development of the BEST (Biomarkers, EndpointS, and other Tools) Resource, data sharing, and regulatory aspects of biomarker acceptance and use. In December 2015, the National Biomarker Development Alliance conducted a meeting using case studies around four different diseases to explore the possibility of generalizing an evidentiary framework to support biomarker qualification for generalized COUs.

Dr. Leptak concluded by saying that a general framework for evidentiary criteria to qualify biomarkers for use in drug development would:

- Enhance the clarity, predictability, and harmonization of the biomarker qualification process
- Improve the quality of biomarker qualification submissions to FDA
- Support FDA in the development of relevant guidance documents for evidentiary criteria for biomarker qualification

Keynote and Charge to Participants

Janet Woodcock, M.D.

Director, CDER, FDA

What is the goal of biomarker qualification? Why are we working on qualifying biomarkers?
Dr. Woodcock explained that an explosion of knowledge has occurred in basic science, but the pace of innovation in therapy development has been relatively flat despite huge investments by the public and private sectors. A few years ago, FDA leaders published a paper showing that the return-on-investment and productivity of the biomedical research engine have diminished over time, with the exception of the introduction of biologics that led to an uptick in these measures.

Dr. Woodcock pointed out that drug developers face a huge amount of uncertainty at each step of the development process. Much of the uncertainty centers on questions about the validity of targets, concerns about off-target effects, and whether targeting provides benefit for patients. Uncertainty exists about measurement of clinical effects, good and bad. There is also uncertainty about regulatory benefit and risk. More than half of molecules that get to Phase 3 development still fail, primarily because they do not demonstrate the hoped-for effects, or they have unexpected toxicity. Some drug

candidates are not as good as expected and are abandoned for commercial development. Of those submitted to FDA, about three-fourths are approved, but much attrition occurs prior to that point.

Lagging progress can be attributed in part to a lack of evidence and a lack of innovation, but scientifically, the problem boils down to one thing: uncertainty. Dr. Woodcock said that no other industry faces this astounding degree of uncertainty and risk. Translational science is needed to reduce uncertainty. How can we measure and predict effects of drug candidates *early* in development?

We need reliable biomarkers that can reduce uncertainty and support decision making in the drug development process.

According to Dr. Woodcock, thousands of biomarker papers are published each year, but no company would gamble a multimillion-dollar development program, nor would a clinician make patient care decisions based on these papers. Neither would regulators trust these discoveries as a basis for regulatory decisions.

The question of whether a biomarker provides *new, actionable information* that can be relied upon for decision making in drug discovery, drug development, and health improvement cannot be answered by researchers working in the domain of basic biomedical science; their job is to publish papers. Nor is it the job of any one pharmaceutical company; their job is to develop molecules and get them into the market. The original belief was that single companies could develop all the science necessary to support development of innovative treatments, but the magnitude of the science required is more than one company can handle. Congressional staffers sometimes think it is FDA's job to develop biomarkers, but FDA is a regulatory agency—not a research entity—and, therefore, is not in a position to produce the scientific evidence needed.

For the past decade or so, biomarker science has been undertaken mainly by consortia of the willing: the FNIH-supported Biomarkers Consortium, FDA, C-Path, some academics and patient/disease advocates, and a few other entities. There are pockets of understanding about biomarkers in academia, but this scientific area is not well delineated for academics. Together, the Biomarkers Consortium, C-Path, and other groups have spearheaded several efforts to develop specific measures. Some of these biomarkers have succeeded, but others have failed in terms of predictive value. The Innovative Medicines Initiative in the EU has invested in translational research and is working through specific examples to try to develop a panel of biomarkers. All these efforts raise questions about methods and standards for evaluating biomarkers. How does one determine that a biomarker is reliable for its COU?

Some standards—analytical validation, for example—are well understood by some communities with particular expertise in diagnostics and so forth, Dr. Woodcock explained. Other standards, such as the actual performance of a biomarker, are poorly understood.

Dr. Woodcock provided some background on FDA's Biomarker Qualification Program (BQP); she noted that the EU and Japan have similar programs. Consortia or other entities can ask CDER to deem biomarkers as being fit for purpose. Recently, however, there has been a push to reinvigorate evidentiary criteria to support qualification.

Dr. Woodcock highlighted the BEST Resource, noting that it is a living document intended to harmonize the terminology around biomarkers. Its development represents a key step toward developing a

framework for evidentiary criteria for biomarker qualification. The resource was a joint effort of FDA and NIH.

The amount of evidence required for biomarker qualification is proportional to the potential benefit and risk associated with the biomarker's context of use. The case studies presented at the National Biomarker Development Alliance meeting provided an opportunity to generalize and think about evidentiary criteria. Now is the time to drill down to real-world examples to demonstrate what is needed.

Dr. Woodcock said that biomarker development, as a new scientific field, is worthy of the kind of recognition that goes to any sort of scientific endeavor. In other fields, such as aircraft manufacture, failure rates are much lower than they are for drug development. Aircraft manufacturers have translational information that they apply when building an aircraft. They have tolerance data for materials and computer models to predict the likelihood of achieving their objectives. They succeed most of the time. Biology is more complicated, and if we do not put in place an orderly approach to build a body of generalizable knowledge, the failure rate will continue to be too high and the risk excessive.

Dr. Woodcock recommended setting standards that everyone could use as they evaluate molecules. She encouraged the participants to provide their feedback. The documents distributed in advance of the meeting are excellent starting points. FDA has committed to issuing guidance, but it is just part of this effort, which is owned by the scientific community.

The pace of introducing innovation and developing new tools such as biomarkers can be slow, making it difficult to show impact of the BQP. However, if we do not establish evidentiary criteria and continue simply to have ad hoc use of biomarkers, failure rates will not improve and societal debates about drug development will continue. *We need to offer an alternative in the form of real translational science that can make a meaningful impact over the next several years.*

Update on FDA Biomarker Development and Qualification Process

Shashi Amur, Ph.D.
Scientific Lead, Biomarker Qualification Program, Office of Translational Sciences, CDER, FDA

Dr. Amur described how biomarkers have been used for many years across the spectrum of drug development ranging from basic research, to development of companion diagnostics, to use in clinical trials as surrogate endpoints. She highlighted the two main pathways for integrating biomarkers in drug development:[1]

1. Drug approval submissions (Investigational New Drug [IND] applications, New Drug Applications [NDAs], Biologics License Applications [BLAs])
2. Biomarker qualification

Dr. Amur presented the definitions of *qualification* and *COU* as given in the BEST Resource:

- **Qualification:** A conclusion, based on a formal regulatory process, that within the stated context of use, a medical product development tool can be relied upon to have a specific interpretation and application in medical product development and regulatory review.

[1] Amur et al. *Clin Pharm Thera.* 2015;98(1):34–46.

- **COU:** A statement that fully and clearly describes the way the medical product development tool is to be used and the medical product development-related purpose of the use.

Qualification is a voluntary process that allows biomarkers to be used in multiple drug development programs without a need for CDER to reconfirm the suitability of the biomarker's qualified COU. The BQP has the potential to advance public health by streamlining the drug development paradigm. Any individual or group can submit biomarkers for qualification; so far, research consortia have been the main entities to engage in qualification of biomarkers.

Regarding the COU, Dr. Amur clarified that it is a "how and why" statement that drives the level of evidence needed. The required level of evidence influences how long the biomarker qualification process takes.

The biomarker qualification process consists of three stages: (1) initiation, (2) consultation and advice, and (3) review. To date, 13 unique biomarkers have been qualified. A list of qualified biomarkers is available on the CDER website, which includes publicly available information about the biomarkers. A total of 27 projects are currently under way, some of which are listed on CDER's website. About 50% of submissions are safety biomarkers, 30% are response biomarkers, and the majority of the remaining 20% are patient compliance (monitoring) biomarkers.

Dr. Amur also clarified that a composite biomarker is a panel of biomarkers that can be qualified for a specific COU. For composite biomarkers, some important considerations apply, including the question of whether the single biomarkers comprising the composite have equal or unequal weightings.

Dr. Amur said that the main obstacles to biomarker qualification are related to the lack of a clear COU, insufficient supportive data, insufficient resources to support biomarker development and qualification, failure to identify a patient population, issues with sample storage, challenges in the aggregation of requisite data, and insufficient assay validity/reproducibility. The lack of clear evidentiary standards is a major challenge.

Submitters have provided some feedback about the BQP, indicating that the terminology is confusing, the qualification process takes too long, and that qualification can be a costly and time-intensive effort that draws upon limited resources to overcome what is seen as a very high evidentiary bar. The scientific community is also interested in having clearer timelines and deliverables. In addition, the review divisions of CDER have conflicting views on whether qualification is even needed. Why should qualification be pursued if biomarkers can achieve regulatory acceptance through IND/NDA/BLA submissions?

FDA is addressing these challenges in several ways:

- Working with NIH to develop the BEST Resource as a way to improve communication about biomarkers
- Streamlining the qualification process by issuing letters of support and permitting biomarker qualification on the basis of limited COUs
- Focusing more on communication with submitters and CDER staff
- Conducting surveys to identify scientific and therapeutic areas where biomarker development is needed
- Convening workshops on development of evidentiary standards

- Updating the FDA webpage with information and templates for submitters
- Disseminating information about the BQP via presentations and publications

Dr. Amur presented a timeline of BQP accomplishments. Activity spiked over a period extending from 2013 to the present, during which eight letters of support were issued in addition to the qualification of eight biomarkers and other biomarker projects in process.

Dr. Amur encouraged the meeting participants to send in any questions they have. Those who are interested in pursuing qualification of a biomarker should engage early and often with CDER staff.

Proposed Framework for Evidentiary Criteria for Biomarker Qualification

Christopher Leptak, M.D., Ph.D.
Acting Director, Office of New Drugs Regulatory Science Program, and Co-Director, Biomarker Qualification Program, Office of Translational Sciences, CDER, FDA

Background: Biomarker Terminology and the Framework

Dr. Leptak provided an overview of the process by which framework was developed and introduced the proposed framework, which consists of three elements:

1. The COU
2. Benefit and risks to the patient for the intended biomarker use
3. Consideration of the evidentiary criteria necessary to support qualification, based on the benefit and risk (i.e., the evidence map)

FDA has been regulating drugs for more than a century. Biomarkers, such as blood pressure, have always been part of FDA's purview. Over time, people developed confidence through community consensus that the biomarker (blood pressure, in this example) was clinically meaningful. The proposed evidentiary framework is intended to be a formalized way to delineate the types and levels of evidence necessary to develop confidence in a biomarker based on its benefits and risks.

Dr. Leptak said that the 21st Century Cures Act proposed the development of guidance documents outlining an evidentiary framework to support biomarker qualification. Dr. Woodcock has said that "it takes a village to raise a biomarker," meaning that no one entity, including FDA, can do it alone.

Dr. Leptak outlined how the BEST Resource—a glossary of terminology and examples of uses of biomarkers and endpoints in basic biomedical research, medical product development, and clinical care—was created by the FDA-NIH Biomarker Working Group. The Working Group examined guidance documents, regulations, and documents of the NIH extramural research program to gather definitions used by various communities and used these as a starting point. The hope is that a broad variety of clinical and scientific communities will use the terminology. The BEST Resource is publicly available through the National Library of Medicine website and will be updated over time with new definitions and examples.

According to the glossary, a biomarker is:

> *A defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes, or responses to an exposure or intervention, including therapeutic interventions. Molecular, histologic, radiographic, or physiologic characteristics are types*

of biomarkers. A biomarker is not an assessment of how an individual feels, functions, or survives.

Dr. Leptak pointed out that many elements of the 2001 consensus definition are encompassed in the definition in the BEST Resource. The term *biomarker* covers a spectrum ranging from serum proteins, changes in tumor size measured with imaging technologies, and algorithms for QT determination on electrocardiogram. The definition is consistent with longstanding goals and drug development processes. Regulatory acceptance focuses on how a biomarker is used in drug development (as opposed to clinical biomarkers used in treatment decisions).

The phrase *fit for purpose* refers to matching a biomarker to a drug development goal and a data-supported relationship. The BEST Resource categorizes biomarkers according to their function in drug development:

- susceptibility/risk biomarker
- diagnostic biomarker
- monitoring biomarker
- prognostic biomarker
- predictive biomarker
- pharmacodynamic/response biomarker
- safety biomarker

According to Dr. Leptak, the components of a successful biomarker development effort include a defined hypothesis and goals, data (e.g., scientific understanding of the topic, existing data and stored samples), resources (e.g., financial, staff, information technology expertise) available to the biomarker development team for additional data collection and analysis, strategies to mitigate challenges, relevant models or in silico options, and opportunities for collaboration (e.g., consortia, patient/disease advocacy groups, professional societies).

The COU defines the boundaries of the biomarker's known reliability and limitations. In other words, what questions can the biomarker answer? One topic of interest is the potential expansion of the COU with additional studies and data to support future qualifications for use in drug development or clinical trials.

In his concluding remarks, Dr. Leptak discussed the development of the proposed evidentiary framework and its three major concepts: COU, benefit/risk profile, and evidentiary criteria. He clarified that the risks have to do with accepting a bad biomarker or rejecting a good biomarker. The profile of benefit and risk could range from challenging to favorable.

Testing the Framework

John Wagner, M.D., Ph.D.
Senior Vice President and Head of Clinical and Translational Sciences, Takeda Pharmaceuticals

Dr. Wagner said that the case studies exemplify how the framework fits into the flow of evidentiary criteria. He recalled that the goals of the framework are threefold: clarity, harmonization, and predictability. *The vision is to create a predictable process that we, as a community, can apply to expand biomarkers into wider use.*

Dr. Wagner noted that confusion persists about the COU. FDA guidance gives granular detail on the COU, but it can be distilled into two elements: (1) What class of biomarker is proposed? (2) What information content would it provide? Stated otherwise, what is the biomarker's specific fit-for-purpose use? The intent of the biomarker should be elaborated before the COU is defined and should be re-addressed as new information becomes available.

The benefit/risk profile, given that the COU is related to the biomarker's value to drug development or clinical trials, is assessed from the patient's perspective. Risk relates to the potential harm if the biomarker's use is not aligned with expectations based on the COU.

The type and amount of data required for qualification depends on expected benefit and expected risk. The COU links the benefit/risk profile with the level of evidence needed for the biomarker.

With a favorable benefit/risk profile, a lower evidence level likely could suffice for qualification. For example, most biomarkers used for patient stratification could lead to a loss of resources but would not compromise patient safety if the biomarker does not perform as expected. In the setting of a clinical trial testing a targeted therapy, patient stratification would be more critical, however. A less favorable benefit/risk ratio likely would require a moderate level of evidence; an example would be a safety biomarker used in addition to traditional safety biomarkers. The degree of risk depends on the impact on decision making in drug development and the risk to patients enrolled. The highest level of evidence is required for surrogate endpoints, which can substitute for clinical endpoints in human studies; if the evidence is lacking or incorrect for a surrogate endpoint, regulators could make incorrect approval decisions, leading to ineffective drugs being marked or patients being denied access to effective therapy.

The evidence map in the framework was inspired by a framework proposed by Altar et al. in 2008.[2] It is a standardized visual representation of the level of evidence required for regulatory qualification, compared with the current state of the evidence. It focuses on the most important criteria needed for decision making, such as the assay and scientific understanding. Analytical validation of the assay for measuring the biomarker encompasses such measures as accuracy, precision, analytical sensitivity, analytical specificity, reportable range, reference interval, reproducibility, stability, and so forth.

Dr. Wagner asked the workshop participants to engage in the cross-stakeholder discussion in order to test the hypothesis that the proposed framework will advance biomarker qualification by enhancing clarity, predictability, and harmonization of the process. Will the framework improve the quality of biomarker submissions to CDER for qualification? Does the framework support FDA in the development of relevant guidance for evidentiary criteria in biomarker qualification?

Q&A Session

Q: Why are surrogate endpoints not listed as a type of biomarker in the BEST Resource?

A: Dr. Leptak said that the subset of biomarkers that could serve as surrogate endpoints is small; therefore, a separate category was not included. Most surrogate endpoints would be considered pharmacodynamic/response biomarkers. In some cases, surrogates would be intermediate markers.

[2] Altar CA, et al. A prototypical process for creating evidentiary standards for biomarkers and diagnostics. *Clin Pharmacol Ther.* 2008;83(2):368–71. Epub December 19, 2007. PMID: 18091762.

Q: The evidence map provides clear information about the COU. Is the goal of the framework to come up with COUs and levels of evidence? Or is it a continuous loop whereby the COU and evidentiary requirements are confounded? Why not provide more clarity by setting standards, as is done with drug/device submissions? Also, the number of potential COUs for safety biomarkers is probably finite. Why not come up with evidentiary criteria for each?

A: Dr. Leptak referred to a 2004 document of C-Path as well as the 21st Century Cures Act, both of which use the word *standards.* For biomarker qualification, the decision was made to use the term *evidentiary considerations.* The hope is that standards may come about in the future, but given the broad range of biomarkers, coming up with standards will take a while. Some discussions have gone beyond "considerations," however. Standards can imply legal consequences. The goal of the BQP is to be transparent and consistent and to provide guidance. CDER does not want the process to come down to a series of checkboxes. The framework represents the sweet spot—evidentiary considerations. Additional modules could lead to future FDA guidances along with corresponding levels of evidence. Dr. Wagner recalled the trifold goals of harmonization, clarity, and increased predictability, and he noted that uncertainty is embedded in biology and science. We want to enhance the clarity of the process, but with general rules that balance a prescriptive approach with the flexibility needed for individual cases. The central question is, what is the right balance between being prescriptive and flexible?

Q: A participant referred to the COU description on the CDER BQP website, noting that it refers to the question of what decision would be made based on a biomarker result. Benefit/risk is a critical piece of the COU. Some biomarkers are merely informative, whereas others have clinical impact.

A: Dr. Leptak said that the framework went through more than a dozen versions. The planning committee compared benefit/risk and COU to decide which concept to describe. He said that there is overlap between the two concepts.

Q: Should the evidence level be based on the benefit/risk profile or on risk alone? In the examples put forward, the level of evidence could have been defined by just using risk. Consider that the level of evidence is based on the risk of getting the decision wrong, but one can take steps to mitigate the level of risk. The level of evidence could be lowered when justified by areas of unmet need or rare diseases.

A: Dr. Leptak said that this question has been raised before. Benefit/risk is part of the fundamental approach for biomarker qualification. In the device world, however, the focus is on risk classification, not benefit.

Q: A participant pointed out that benefit could be construed as improving drug development productivity or as helping patients directly.

A: Dr. Wagner agreed that good biomarkers can affect productivity of drug developers and that patients can benefit from innovative treatments that can result from improved productivity. Therefore, the two concepts of benefit connect at the "30,000-foot level." The writing committee extensively discussed a cost-benefit approach as a quantifiable way of assessing benefit and risk, but it is important to consider the patient perspective.

Q: FDA has gone through the process of biomarker qualification with six biomarkers. It seems that FDA is more often recognizing preclinical safety biomarkers with letters of support rather than qualification. A

letter of support opens the door for case-by-case use when the benefit/risk profile is appropriate. Preclinical safety biomarkers comprise 20% of submissions to FDA. A letter of support encourages development of promising biomarkers but signals a need for more data for qualification.

A: Dr. Leptak responded that the letter of support does not signal that FDA is moving away from qualifying preclinical biomarkers; it is a matter of research communities choosing a route leading to a letter of support rather than pursuing a qualification effort. Dr. Leptak underscored that qualification is a voluntary process. External parties can decide which route they want to follow given their resources.

Q: Some terms relevant to biomarkers do not appear in the BEST Resource. Examples include *probably valid* and *known valid* biomarkers.

A: Dr. Leptak said that the resource is intended to be a living document that can be modified as needed. He encouraged the participants to suggest changes.

Q: A participant from FDA suggested adding an arrow to figure 2 in the draft framework to indicate that the submitter should collect more evidence to support biomarker qualification.

A: Dr. Wagner thought that a revision of the diagram might be in order.

Statistical Design Considerations: Summary of the Statistics Working Group

Lisa McShane, Ph.D.
Chief, Biostatistics Branch, Biometric Research Program, Division of Cancer Treatment and Diagnosis, NIH/National Cancer Institute (NCI)

Dr. McShane said that the statistical group resulted from a suggestion to develop a statistical map for biomarker qualification teams to refer to when working on clinical biomarkers of safety and prognostic enrichment. She linked several important statistical considerations to key steps in the qualification process, which generally proceeds in a stepwise manner along with discussions with regulators:

1. **Construct an intended COU statement for the biomarker.** Statisticians should be part of a collaborative team working on the COU. The COU needs to articulate the purpose of the biomarker, type of population (e.g., animal vs. human, age, sex, race, comorbidities); clinical setting (healthy volunteers vs. population with a disease/medical condition; treated vs. untreated for disease/medical condition); biomarker measurement (a testing system [assay], single measurement vs. combination of multiple measurements of the same or different biomarkers or with standard clinical or pathological covariates).
2. **Collate existing evidence to determine initial support for the intended COU by examining relationships between the biomarker and relevant clinical outcome measures.** Dr. McShane cautioned about the pitfalls in retrospective prognostic studies that lack a rigorous prospective design and analysis plans. Many biomarker studies are conducted in highly heterogeneous patient populations, but sometimes that is not the mix represented in the intended COU. Also, it may be that different assays were used to measure the different biomarkers. She cited an example[3] evaluating the strength of prognostic associations and support for an assay based on six different subgroup analyses using different optimized biomarker cut-points. She recommended that qualification teams look for relevant differences in different studies when evaluating strength of evidence.

[3] *Clin Cancer Res.* 2013;19(23):6633–6643.

3. **Identify additional data and biospecimens available for analysis that could be used to build on existing evidence.** This step will help the team identify potential gaps, remaining data needs, and potential sources of additional data.
4. **Determine appropriate strategies for analyzing existing and newly generated data to evaluate the level of support for the intended COU.** The qualification team needs to plan ahead with a biostatistician to develop a game plan, which should include the following:
    a. Evaluate analytic variables and assess the analytical validity of assays.
    b. Assess the comparability of candidate assays used to generate existing data or for the planned-for generation of new data.
    c. Obtain standard clinical and pathologic variabilities for prespecified subgroup analyses and adjustments.
    d. Use internal validation strategies and/or designate samples to hold back to avoid over-optimism bias during biomarker quantification and biomarker-outcome model development. If possible, maintain strict separation of development work and final independent validation. It is best to have a truly independent data set for validation. She cautioned about validating a biomarker for predicting toxicity in a relatively healthy population if it might be applied in frailer subpopulations that might be more prone to toxicity. Also, changing the mix of patients could affect the risk of relying on the biomarker for a prognostic enrichment COU. Another important part of the process is to come up with a biomarker quantification method. It might be necessary to perform a covariate adjustment to get a biomarker that is appropriate for different patient groups. In some cases, one might be interested in measuring change from baseline; with prostatic surface antigen, it is the slope or doubling time that is important. One could take an "unsupervised" look to see how patients cluster into subgroups, or a supervised approach using scores and algorithms might be the best approach. Statistical methods can help qualification teams figure out composite biomarker scores and their relationship to the clinical outcome of interest. Dr. McShane emphasized the importance of building and testing models correctly. To avoid model overfitting (i.e., fitting to noise), statistical tricks (e.g., resampling methods) could allow a team to build a model based on a training set, run the quantification on a training data set, and then run it on a test data set that was not included in the training set. Re-substitution is the naïve method of testing model performance by plugging in the same data used to develop the model. Qualification teams need to include statisticians and computational scientists who have expertise in the nuances of modeling.
5. **Quantify risks and benefits associated with use of the biomarker in the intended COU based on the totality of the evidence.** Once a relationship between a biomarker and a clinical outcome of interest is established, several cut-points on a continuous biomarker may be considered for decision making.

Statistical Design Considerations: Panel Discussion and Audience Q&A

Lisa McShane, M.D., Discussion Leader

Panel Members:

Aloka Chakravarty, Ph.D., CDER, FDA
Klaus Romero, M.D., C-Path
Andrew Thomson, Ph.D., European Medicines Agency (EMA)
Sue Jane Wang, Ph.D., CDER, FDA

1. Elaborate on what you see as differences in evidence levels based on biomarker category.

   The panel members thought that the evidence level depends on the disease setting (unmet need, rare disease), the time scale for the biomarker (e.g., prognostic biomarkers depend on natural history of the disease), whether the submitter is intending to replace an existing biomarker or qualify a new one to use along with an existing biomarker, benefit assessment (patients or healthy volunteers), and the risk assessment (whether a toxicity is reversible or the ramifications of getting the qualification decision wrong). An early conversation with CDER staff about data requirements is critical.

2. What are the panel's thoughts about the challenges of collecting data for the analyses mentioned?

   Dr. Romero said that C-Path focuses on subject-level or experiment-level data, not meta-level data. Differences in published literature are sometimes difficult to discern, increasing the challenge of using the data from several studies to support biomarker qualification. Dr. Wang spoke of the importance of acknowledging the limitations associated with the data available and thinking carefully about additional data needed. Retrospective data might not be designed for the intended COU. One potential advantage of composite biomarkers is that the limitations of each separate biomarker may be overcome by considering them together. Dr. Thomson discussed the importance of transparency in clinical trials. There is potentially more individual-level data that could be accessed although they were never published originally. Dr. Romero said that by defining a draft COU and envisioning the analyses required, the qualification team might identify data gaps that would be important for experimentalists. He also said that letters of support provide an impetus for companies and sponsors to pursue the required data.

3. Regarding the recommendation for an independent validation data set, what are the limitations if such a data set is not available?

   Dr. Chakravarty said that an independent data set provides robustness that can never be matched by data splitting. Despite use of sophisticated statistical tools, the issue of cross-validation remains. Nevertheless, sometimes qualification teams have no choice other than to split a data set for validation. Dr. Thomson said that the limitations of data splitting can be mitigated to some extent if the data set is sufficiently large and if the investigators specified a priori that the data will be split. Dr. Romero said that study endpoints are also key: Hospitalization and death are clear endpoints, but many others are less clear. Dr. Wang spoke of her experience validating a biomarker using internal, confidential data.

4. Safety signals often occur at very low prevalence rates, making it difficult to use them as good evidence for qualification. What is an acceptable timeframe for safety data to use in qualification—during drug development only or also during the post-marketing period?

   Dr. Chakravarty said that aggregating results in a meta-analysis or in a combined look triggers many caveats. The choice of analytical method is driven by the number of events, thereby complicating the analyses for biomarker qualification (and for safety analysis generally). She recommended thinking about methods to consider a lack of events as being equivalent to a lack of evidence. Dr. Wang spoke about making a decision up front about low-prevalence safety biomarkers if a drug candidate seems very promising.

5. What are the key performance criteria (e.g., sensitivity, specificity, positive and negative predictive values) for a binary biomarker assay)?

   Dr. Wang recalled that FDA would not specify targets for sensitivity and specificity performance characteristics for a biomarker assay. False positive and false negative rates are important. For enrichment, the risk is borne by the sponsor. Because of an imperfect assay, the investigator might not enroll the right patients. But, if the biomarker is for enrichment based on prevalence, then the risk is shifted. Dr. Thomson said that, for safety biomarkers, the relative importance of performance criteria depends on whether one is developing something new or something better. Dr. McShane remarked on the distinction between population risk and individual risk. For enrichment biomarkers, the risk is more to the population, but for safety biomarkers, the risk is skewed toward the individual level.

6. Who should do the statistical analyses—bioinformaticians, computational scientists, biomedical statisticians?

   Dr. Chakravarty said that the team approach is always best. Statisticians by themselves could not elucidate the clinical implications of a particular threshold, but they should be an integral part of the planning and strategy for the analysis. Dr. Romero agreed, saying that clinical relevance has to be the absolute foundation, but the qualification team must include quantitative scientists and experimentalists.

Dr. Chakravarty provided some additional background on retrospective analyses, which can be useful but with several caveats: First, there must be a predetermined, independent scientific basis underpinning the biomarker hypothesis. Second, studies need to be well controlled and adequately powered. Third, prognostic factors not used in the stratification, but considered important, should be balanced across treatment arms within the subgroups to avoid systematic bias. Fourth, the collection of participants should be representative of the entire intention-to-treat population, not just a convenience sample enrolled at a given study site. (Ideally, the same biomarker assay should be used for all subjects.) Fifth, the analysis plan should account for the retrospective nature of the biomarker analysis. Data integrity would be in question if the analysis plan was set up after unblinding of the efficacy data and biomarker status (if known). Sixth, the analysis plan must control for multiplicity and the study-wide false positive rate.

Q&A Session

Q: Patient-level or experiment-level data are optimal, but how can statisticians use metadata to full advantage?

A: Dr. Romero cautioned against using metadata, as they can lead to erroneous conclusions and because the underlying studies rarely used the same methodology and populations. However, metadata could be used to develop a draft COU, formulate a potential analytical plan, and open a dialog with regulators about qualification. The regulators could help the qualification team focus on right data sets. He recommended changing incentive mechanisms to encourage more data sharing. Dr. Chakravarty added that some types of biomarkers require individual data on exposures.

Q: More guidance is needed on how to assemble combinations of biomarkers in a rational way.

A: Dr. Wang said that statisticians can help assemble panels of biomarkers. For qualification, the components of the panels are scrutinized. Simply putting biomarkers together through an algorithm is probably unwise. Rather, one must think about linking biomarkers with the mechanism of action. If a team uses a statistical method to assemble a panel of biomarkers, the benefit needs to be clear. Dr. Romero pointed out that the qualification guidance for total kidney volume for polycystic kidney disease makes it clear that the panel needs to be interpreted in conjunction with clinical factors, which take variability into account. He underscored the importance of capturing and measuring sources of variability for the clinical endpoints of interest.

Q: Two things seem to be slowing down the safety biomarker pipeline. First, statisticians are used to following the 2010 guidance on adaptive trial designs for drug approval to ensure that trials are done in a robust way. There is a hope that event rates will be higher with biomarkers, making them superior to clinical endpoints in a way. Can the trial *N* be decreased? Second, is it possible to look at event rates in a blinded way to preserve the integrity of the study? Is it possible to peek at the data in interim fashion without introducing excessive alpha error?

A: Dr. Wang said that, for adaptive trial designs, blinding is critical. Unblinding may be proposed in a prospective way, however. She also discouraged decreasing the study size in confirmatory studies. She offered an example of a blinded sample size re-estimation. The statistical method would increase or decrease the sample size in theory. In application, however, alpha needs to be controlled.
Dr. Chakravarty added that stopping a study early could entail a statistical penalty. Biomarker qualification already faces many challenges; one does not need to complicate the situation further. If a qualification team wants to build in an interim peek at the data, this needs to be prespecified: “You don’t want to service the plane in mid-air!”

Case Study 1: Biomarkers of Drug-Induced Kidney Injury (DIKI)

Frank Sistare, Ph.D.
Scientific Associate Vice President, Safety Assessment & Laboratory Animal Resources, Preclinical Development, Merck & Co., Inc.

Dr. Sistare presented the history of a biomarker qualification project for DIKI that started in 2006. The hypothesis was that new promising translational kidney safety biomarkers (albumin, β2-microglobulin, clusterin, cystatin C, [kidney injury molecule-1] KIM-1, total protein, and trefoil factor) detected in urine could:

1. Report injuries to different segments of the nephron
2. Respond earlier and with greater sensitivity than blood urea nitrogen (BUN) and serum creatinine (SCr)
3. Inform patient prognosis

4. Enable safe clinical drug development

Some of the biomarkers in the panel are functional biomarkers, some are injury response markers, and one is a leakage marker.

By 2008, evidence from 34 animal studies was available; FDA and EMA regulators supported qualification claims for the panel of biomarkers for nonclinical and limited clinical use. The qualification team expects to have a final qualification submission in 2017 based on two clinical studies currently in progress.

Dr. Sistare reviewed the context of use (clinical): The panel of qualified urinary kidney safety biomarkers may be used together with SCr and BUN as a more sensitive and/or earlier biomarker to monitor for renal tubular safety. The biomarkers inform an injury response and could enable or restrict planned dose escalations or drive decisions to interrupt or continue dosing. Specifically, the panel could be used to monitor clinical trial subjects in real time so that dosing of a subject could be interrupted or modified sooner or interrupt a dose cohort sooner.

Study subjects could be normal, healthy volunteers or patients with no concurrent kidney disease in early clinical trials of duration supported by animal studies conducted with the same test agent. Exclusions include urinary tract infections, urinary tumors, systemic inflammatory disorders, and elevated risk for developing kidney disease.

Dr. Sistare described a plan involving a creative step-down procedure to see if some of the biomarkers could be eliminated from the panel.

The gold standard for DIKI is histopathology. Inulin/iohexol clearance and biopsies are impractical. The biomarker assays are all commercially available, along with standards. Albuminuria and proteinuria are not new, but the others are novel biomarkers. The benefit of the biomarker panel is that it detects DIKI earlier than BUN and SCr. For example, the KIM-1 biomarker increases 10-fold before any increase in SCr is seen.[4]

Regarding the qualification path, these biomarkers are being used clinically in drug development. There is a growing evidence base to bring them to a higher level of qualification, making the case for interim qualification based on a limited COU to encourage development of more evidence and mitigate risk.

In terms of the benefit/risk assessment, Dr. Sistare said that it is difficult to calculate the risk of a false negative (i.e., BUN and SCr detect an injury but the biomarker panel does not). The risk of continuing to give someone a small dose of a drug is probably minimal although there could be a possibility of fibrotic disease that could compromise renal reserve over time. The risk of a false positive is mitigated because SCr and BUN serve as backstops.

A limited COU has been proposed as a composite measure of the biomarkers in a highly restricted clinical trial context. The final step would be an integrated evaluation of available evidentiary criteria levels against the proposed COU and benefit-risk assessment to support qualification of the DIKI panel for clinical use.

[4] *Am J Kidney Dis.* 2013;62(4):796–800.

Case Study 1: Panel Discussion and Audience Q&A

Frank Sistare, Ph.D., Discussion Leader

Panel Members:

Amanda Baker, Pharm.D., Ph.D., C-Path
Steve Hoffmann, M.S., FNIH
Paul Kimmel, M.D., NIDDK
Romaldas Mačiulaitis, M.D., Pharm.D., EMA
Irene Nunes, Ph.D., Merck & Co., Inc.
Aliza Thompson, M.D., CDER, FDA

1. Does the proposed framework provide a way to ensure that the value of the biomarker is articulated clearly? What is the biomarker's value to drug development and to patients? Will the effort of qualification justify the investment of resources? What is missing from the framework?

   Dr. Mačiulaitis thought that the framework was sufficient and that it allows the submitter to clarify the importance and value of the biomarker. He said that the framework helps everyone speak the same language. He spoke about three dimensions: benefit, risk, and the level of evidence.

   Dr. Kimmel said that the question about the value of the biomarker is very interesting, but it could only be answered in the future from a societal and commercial standpoint. By applying the biomarker, can drugs be developed more quickly and at a lower cost?

   Dr. Thompson said that regulators would want to see how the submitter assesses the evidence for each category given the proposed COU. The evidence map should be a starting point for a dialog with the regulator. The document needs to be a precise and succinct summary that includes a rationale for selecting a particular box on the evidence map. The COU dictates how much evidence is needed. Dr. Sistare envisioned the evidence map as a living document that would be updated as certain milestones are hit, publications come out, and data generated.

   Dr. Thompson said that the qualification guidance addresses the uncertainty about how the agency will interpret changes in the biomarker. Small companies often take on riskier drugs and are using biomarkers to address concerns about toxicity. They could benefit from guidance on using data for COUs.

2. There may be more potential benefit if no treatment exists for a particular indication. How do regulators evaluate evidence against that background?

   Dr. Thompson said the key issue is the evidentiary criteria, which are tied to the proposed COU. Then, the COU is tied to the risk of a false positive or negative. There is no point in engaging in a discussion with regulators if there is no unmet need that the biomarker is addressing.

3. What about the possibility of a limited COU that would go beyond the letter of support as an interim stage toward qualification?

   Dr. Mačiulaitis pointed out that pursuing a limited COU would still require 75% of the work required for full qualification.

Dr. Sistare spoke about the tension between the COU and research. For example, cisplatin is a single-use agent, so there was a discussion about limiting the COU to single-use agents, because the test cohort was treated with cisplatin. However, limiting the COU in this way would have rendered the qualification effort not worthwhile.

The group discussed causes of analytical interference, including proteinuria, urea and sodium concentrations, lipemia, and drugs for comorbidities. Some interferences can be negated by normalizing to SCr; in other cases, one could stratify on age, disease, or comorbidities.

Q&A Session

Q: A participant remarked on the risk of false negatives (i.e., overlooking DIKI over a period of time). If one observes a change in an analyte but the level returns to baseline, how can we know that the tubules are functioning normally? What is the effect of these little "hits" over time? The DIKI safety biomarkers will be helpful for answering these questions.

Q: Another individual recommended using the term *historical comparator* instead of *gold standard.*

Case Study 2: Glutamate Dehydrogenase (GLDH) as a Biomarker of Drug-Induced Liver Injury (DILI)

Jiri Aubrecht, Ph.D., Pharm.D.
Senior Director, Safety Biomarkers, Pfizer

Dr. Aubrecht outlined some of the challenges in detecting DILI. The liver has large reserve; the conventional biomarker-based DILI diagnostic paradigm detects liver injury only after substantial (sometimes irreversible) damage has occurred. The gold standard, alanine aminotransferase (ALT), is sensitive but not specific enough as it is also present in muscle. Transient ALT increases in clinical trials are common. Patients with underlying muscle disease (e.g., rhabdomyolysis, myositis, hereditary dystrophies) have elevated ALT levels, which exclude them from participating in trials of agents with the risk of DILI and makes diagnosis of liver injury problematic. Bilirubin is sufficiently specific, but it lacks sensitivity.

Therefore, an unmet clinical need exists in clinical care and in research. A new liver-specific biomarker strategy is needed to provide context for ALT increases in clinical trials and clinical practice. GLDH has been shown in preclinical species to be a liver-specific biomarker of liver toxicity. Limited clinical data are published on GLDH.

The COU reads thus: Elevated serum GLDH is a measure of hepatocellular liver injury that can be used in target populations including healthy subjects and patients with elevated serum ALT from suspected extrahepatic sources to evaluate liver specificity of observed increases in ALT. GLDH will be used in conjunction with standard measures of hepatotoxicity, including bilirubin.

Insofar as the benefit/risk analysis is concerned, Dr. Aubrecht postulated that application of GLDH adds no risk because the biomarker would be used in conjunction with the current diagnostic paradigm for DILI and will not eliminate or supersede current safety guards.

Toxicity studies in humans are not feasible, and access to DILI cases is limited. For the qualification effort, the team is monitoring biomarker performance in human disease that approximates DILI and

monitoring cases of acetaminophen hepatotoxicity. Randox (UK) manufactures a kit for measuring GLDH; it is validated under Clinical Laboratory Improvement Amendments (CLIA). The qualification team evaluated the analytical validity of the assay, tested sample stability, and evaluated sources of interference.

The objectives for clinical qualification of GLDH were the following:

1. Establish a reference range for GLDH
2. Establish GLDH as a sensitive biomarker of liver injury at least as good as ALT
3. Determine GLDH specificity for liver injury in nonclinical studies and clinical studies to assess the ability of GLDH to differentiate between muscle and liver injury and to detect liver injury

Regarding objective 2, the qualification team tested 32 model agents with a wide range of mechanisms of action in rat models. GLDH significantly outperformed ALT with an overall increase in the area under the curve for the receiver operating characteristics. In addition, the team conducted a very comprehensive evaluation of GLDH as biomarker of liver injury in humans. The results were used to come up with cutoff levels of GLDH to indicate potential DILI and correlations of those results with ALT levels. Regarding the specificity of GLDH, the study team found that GLDH is not affected in subjects with Duchenne muscular dystrophy.

Dr. Aubrecht said that the evidence assessment map in the proposed framework was very helpful in terms of the strength of the evidence for qualification. GLDH met the very high standard for the assay, for biological performance, and for the scientific understanding. GLDH improves specificity while retaining the sensitivity of ALT for detecting liver injury. In terms of the data and samples proposed to establish qualification, the team has hundreds of subjects contributing data through banked samples. The high level of evidence supports a very strong benefit/risk analysis.

Case Study 2: Panel Discussion and Audience Q&A

Jiri Aubrecht, Ph.D., Pharm.D., Discussion Leader

Panel Members:

Elizabeth Hausner, D.V.M., FDA
Christopher Leptak, M.D., Ph.D., FDA
Joseph Menetski, Ph.D., FNIH
John-Michael Sauer, Ph.D., C-Path
Shelli Schomaker, Pfizer
Paul Watkins, M.D., University of North Carolina

The panel members thought that the case dovetailed with the framework. Dr. Aubrecht considers the framework as a guide—or a philosophy—for thinking about how to get the evidence required for qualification.

Dr. Sauer appreciated the strategy for GLDH because it occupies a niche, a very specific COU. The team did not try to "boil the ocean"; rather, they used leftover samples and easily obtainable samples for the qualification and worked step-by-step to increase the value of the biomarker.

Dr. Leptak favored use of the term *historical comparator* in this case because the so-called gold standard is not very good. This qualification effort aims to improve upon the historical comparator in stepwise fashion.

Dr. Watkins thought that adding GLDH as a routine liver test might not be justified. The rapid half-life of GLDH means that if there is a delay after liver injury, levels may return to normal, giving the wrong picture. Also, some drugs cause injury that resolves sooner or later. His point was that biomarkers are challenging. It will take more than a couple of studies with a few academic partners to support use of GLDH. FDA can create a safe haven where companies can collaborate and share data to validate biomarkers. Dr. Aubrecht said that large collaborations require substantial funding. The EU uses fund matching with contributions from drug companies to support such efforts.

Dr. Hausner emphasized the importance of the submissions including a thorough examination and synthesis of the literature to show acceptance of a biomarker by the community over time. Having this background gives regulators more confidence and can reduce the number of new studies that are needed. She pointed out that at least one qualification was done just by literature review.

Dr. Menetski discussed the potential use of a limited COU to encourage parties to use the biomarker. The COU can be thought of like a drug label. One possibility is to extend the consortium-like approach to continue the qualification process because no single company has the wherewithal to do it. Others agreed about the importance of leveraging data being generated around proprietary compounds. Valuable data are being lost. Discussions between FDA and C-Path are focusing on repurposing data and leveraging them for biomarker development.

The partnership with the Pfizer team was very important in bringing in the preclinical data on GLDH. Dr. Aubrecht said that Pfizer having access to the expertise and data of the PSTC was very valuable.

Q&A Session

Q: If the impact of biomarker qualification can be proven by making drug development more efficient and by benefiting public health, most likely, funds will become available.

A: Dr. Aubrecht thought that if GLDH is used in hospitals for patients with rhabdomyolysis or for children with Duchenne muscular dystrophy, this would be a tangible public health benefit.

Q: With regard to sharing proprietary data, companies might hesitate to do so because of potential negative impact on their development programs.

A: Companies are not afraid of generating exploratory data; confirmatory data are more of a problem, because some action must be taken.

Q: One participant asked about creating a data repository and using the data to better define the utility of biomarkers and extend their COUs. This would require data owners to be willing to contribute de-identified data. A pilot study could be conducted to see if sufficient data could be aggregated in this way. Another person said that FDA knows that sponsors are measuring certain parameters, but that information is confidential. Would it be possible for FDA to send letters to sponsors to inform them about biomarker qualification activities of C-Path? Yet another participant recalled that many years ago, FDA and representatives of drug companies convened a workshop to draft FDA's safe harbor policy. At that time, companies were being encouraged to submit their pharmacogenomic data to FDA. Companies have a great deal of safety-related data, but FDA needs the authority to take the initiative and encourage industry to collaborate. How could we set up an infrastructure for data sharing?

A: An attendee said that it is challenging for FDA to curate data internally due to data standards. However, FDA could reach out to sponsors to encourage them to donate their data, and the agency could establish a safe harbor data repository.

Panel Discussion and Audience Q&A: Understanding the General Framework

David Wholley, M.Phil., Discussion Leader

Panel Members:

Martha Brumfield, Ph.D., FDA
Gary Kelloff, M.D., NCI
Christopher Leptak, M.D., Ph.D., CDER, FDA
Rajesh Ranganathan, Ph.D., PhRMA
Thorsten Vetter, M.D., EMA
John Wagner, M.D., Ph.D., Takeda

1. How does the EMA view biomarker qualification?

   Dr. Vetter said that the elements proposed in the framework would be very helpful in formulating the evidence map. He also said that the framework fits in with EMA's thinking about biomarker qualification. If an applicant filled in all the topics mentioned in the evidence map, it would fuel a very productive conversation about how to advance the project from an agency point of view.

   He has some reservations about the diagram used for benefit/risk calculations and thought it could be improved. In terms of assessment and evidentiary requirements, there are no substantive differences between FDA and EMA. Both entities look for solid evidence of benefit and risk.

   EMA has a novel, procedural-based methodology platform for providing feedback for biomarkers and drug development. There is a very interactive planning phase with limited funding periods for consortia. When development is further along, EMA has the option of providing qualification advice and follow-up. The process is flexible and open, and it attracts fees.

   Dr. Vetter recommended streamlining communications between FDA and EMA to address evidence concerns of both agencies. Regarding the evidentiary criteria framework, Dr. Vetter recommended a more robust discussion of assay performance and methodology—areas that have proven to be quite challenging, often very late in the development process. It is also important to make sure both sides agree about confirmatory evidence required.

2. What about decision thresholds based on maturity of biomarkers to prevent the application of immature biomarkers across different drug programs?

   Dr. Leptak said that the framework includes loops for modifying the COU, but perhaps a loop is needed to indicate that more evidence needs to be collected. Dr. Wagner responded that the process should not become an infinite loop; however, this is science and there is a need for a feedback loop.

With regard to the need for FDA guidance, Dr. Wagner said that the field is at an inflection point. Everyone is ready for the clarity that evidentiary criteria will provide. Mr. Wholley recalled discussions about whether the framework should be a cookbook or whether it should be viewed philosophically. The writing team opted for a middle course. The intent was to help qualification teams lay out the evidence. Recipes are prescriptive and reduce flexibility. In contrast, qualification needs to be a flexible process.

Dr. Wagner remarked that the DIKI case study showed not only how a biomarker is being qualified, but also how it was used by the company for decision making. There is interplay based on confidence in a qualified biomarker.

3. Does the framework have value for academia?

   Dr. Kelloff said the evidence document is very helpful, but academics are concerned about work overload between their academic duties, research activities, and clinical practices. Their incentives consist of publications and grants. This work is costly and depends on convenience samples collected by busy surgeons. Also, assays are often unproven and not broadly available. The reward structure is inadequate. We need to change the incentives to engage academia.

   Greater investment is needed along with partnerships among stakeholders, diagnostic developers, advocates, academics, and industry. Academics are indispensable to this process, but we need to equip them to do the work right.

   Dr. Ranganathan, speaking from the perspective of PhRMA, saw great value in this consortium-driven process. The framework reduces uncertainty, helping entities know what to expect when they embark on a qualification project. He recommended revamping the evidence table and clarifying the evidence level needed to drive decision making.

4. With regard to the benefit/risk chart, specifically how does one determine where a biomarker is on the x-axis?

   The benefit of a biomarker derives from how it enhances decision making, and that should be clear from the COU.

   Mr. Wholley clarified that benefit/risk is not a ratio. Some participants thought that separate graphs should be used for benefit and risk to make this clear.

   Dr. Leptak clarified that if the benefit/risk profile is unfavorable, then the required level of evidence is higher. The grid shifts based on the COU.

Q&A Session

Q: What about decision thresholds? Nothing in the framework would prevent people from taking action on the basis of premature biomarkers. Where is the protection or understanding that an immature biomarker will not be applied prematurely across different drug programs?

A: Dr. Leptak said that the framework includes loops for modifying the COU, but perhaps a loop is needed to indicate that more evidence needs to be collected.

Q: Clarity on the evidentiary standards would help submitters understand the sort of data they need for the COU. How can submitters plan effectively if qualification involves infinite iterative loops?

A: Dr. Wagner responded that the process should not become an infinite loop; however, this is science, and there is a need for a feedback loop.

Q: Would it be possible to write down the evidentiary standards for a few potential safety COUs, framed by a benefit/risk assessment and expected evidentiary standards?

A: Dr. Leptak thought that guidance could be issued along these lines. It would be easier to define limited COUs in guidance than in the framework. Dr. Brumfield thought that issuing a guidance with such limited experience, based on only six qualified biomarkers, would be premature. Dr. Wagner disagreed, saying that the field is at an inflection point. Everyone is ready for the clarity that evidentiary criteria will provide.

Mr. Wholley recalled discussions about whether the framework should be a cookbook or whether it should be viewed philosophically. The writing team opted for a middle course. The intent was to help qualification teams lay out the evidence. Recipes are prescriptive and reduce flexibility. Qualification needs to be a flexible process.

Q: Several participants raised questions about the benefit/risk chart, specifically how one determines where a biomarker is on the x-axis. The benefit of a biomarker derives from how it enhances decision making, and that should be clear in the COU.

A: Mr. Wholley clarified that benefit/risk is not a ratio. Some participants thought that separate graphs should be used for benefit and risk.

Q: One participant referred to the evidence map as a snapshot of the currently available information, but it is not clear where one needs to be in terms of the evidence. Is the goal to get to the high standard? Is the minimum ever acceptable? Is the map submitted during the initiation stage?

A: Dr. Leptak clarified that if the benefit/risk profile is unfavorable, then the required level of evidence is higher. The grid shifts based on the COU.

Q: Some drug development is very complicated (e.g., gene therapy, precision medicine, silencing RNAs, and antibodies). It would be helpful to learn from FDA, C-Path, and drug companies how to use upcoming technologies in their decision making.

A: Dr. Wagner remarked that the DIKI case study showed not only how a biomarker is being qualified but also how it was used by the company for decision making. There is interplay based on confidence in a qualified biomarker.

Case Study 3: Biomarkers of Drug-Induced Vascular Injury (DIVI)

Tanja Zabka, D.V.M.
Pathologist/Senior Scientist, Genentech

According to Dr. Zabka, the dose-effect relationship forms the basis for the philosophical approach for all toxicity evaluations. However, there are some unique considerations for vascular injury biomarkers. For one thing, no clinical gold standard exists for measuring vascular injury in the clinic; a need exists to fill this gap to enable more efficient and safer drug development.

The lack of a gold standard led to the unique task of qualifying biomarkers in this setting. Dr. Zabka outlined the challenges involved: the lack of a clinical pathology readout, the fact that one or multiple organs may be involved, and the limitation of using only data on human vasculitides (acute flare or chronic) in qualification because of the lack of an animal model. Current biomarkers are nonspecific inflammation biomarkers (e.g., C-reactive protein, complement) or organ injury, if relevant. Functional assessment of local effects is challenging.

The goal was to qualify sensitive and specific translational biomarkers with a greater tendency toward false positive results (rather than false negative results).

The qualification team started by aligning with similar efforts by other groups working on DIVI biomarkers. The aim was to use morphology and biomarker biology relative to biomarker performance and panel profile. The team also sought a flexible approach for statistical analysis.

The proposed COU focused on a panel of qualified vascular injury safety biomarkers that could be used in conjunction with the totality of information in healthy volunteers with no concurrent vascular disease to monitor for vascular safety in early clinical trials. Biomarkers were needed to report vascular injury, tissue response, and repair processes with greater specificity and sensitivity than conventional biomarkers. The biomarkers would be used to inform dose continuation/escalation/cessation; establish a safe dose for subsequent trials; and enable safer progression to higher clinical exposures relative to preclinical no observed adverse event levels (NOAEL).

The intended population comprised healthy human subjects without underlying vascular disease or inflammatory conditions.

With regard to the state of the evidence, Dr. Zabka said the team started with a long panel of biomarkers and then used univariate analysis to determine which clustering biomarkers could be safely eliminated. None of the biomarkers was strong enough on its own, but as a composite, they were very powerful.

For the benefit/risk profile, Dr. Zabka noted that the lack of a gold standard was compelling; the potential benefit was very high. The risk was that false negatives would incorrectly give researchers confidence to continue treatment or escalate dosing. The risk posed by a false positive was low because dose escalation would stop prematurely. To mitigate risk, the team used conventional nonspecific biomarkers of inflammation and organ injury. The qualification plan captured the breadth of morphology and multiple mechanisms of DIVI for small and large molecules.

In terms of positioning on the evidence map, Dr. Zabka thought that the evidence level was minimal for all parameters. The DIVI biomarker would be low risk for patients but could be high risk for the sponsor. Potential benefit derives from enhanced ability to safely translate from preclinical to clinical study. Risk is low, assuming a low false negative rate. Working in a safe harbor could help the team quickly collect the data needed to meet evidentiary requirements.

Case Study 3: Panel Discussion and Audience Q&A

Tanja Zabka, D.V.M., Discussion Leader

Panel Members:

Christopher Leptak, M.D., Ph.D., CDER, FDA
Romaldas Mačiulaitis, M.D., Pharm.D., EMA
Hobart Rogers, Pharm.D., Ph.D., FDA
John-Michael Sauer, Ph.D., C-Path
James Weaver, Ph.D., FDA

1. Regarding the relationship between morphologic outcomes and biomarkers, is this approach an adequate path forward? What about using patients with human diseases for qualification?

   Dr. Weaver reiterated that no gold standard exists for DIVI. One could look at the mechanism of DIVI using small molecules. Another approach would ignore the mechanism and focus on finding biomarkers that can signal injury. Dr. Sauer said that the difficulty lies in translational studies because animals and humans have different proteins involved in vascular injury and repair. No method exists for monitoring DIVI; hence, the team is applying a creative approach that involves accepting a level of risk but taking steps to mitigate that risk.

   Animal models have led to failures in the past, according to Dr. Rogers, but there is nothing else to go on from a regulatory context. He suggested that the team think about a limited COU based on small molecules known to have strong effects. Dr. Zabka said that growing confidence in animal models might help regulators define evidentiary goal posts. Dr. Mačiulaitis said that in many cases, animal models do not exist and even proof-of-concept studies are being done in humans. Ethics committees seem to be accepting of this.

   Dr. Mačiulaitis suggested setting some priorities and focusing on the main risk factors. Dr. Zabka said that inflammation is nonspecific, but there are many markers available for assessing it. The team is interested in reducing the number of endpoints.

   Dr. Sauer noted that the lack of a gold standard means no safety net if the biomarker fails. Dr. Mačiulaitis said that it should be simple to prove that the biomarker is better than nothing.

   Dr. Leptak observed that no monitoring is done currently for any vascular effects. What is the likelihood that companies would be willing to study this because of potential negative effects on their development programs? Companies might be concerned that development programs might be stopped by regulators.

   Dr. Weaver pointed out that the absence of a path into the clinic probably has stopped 100 development programs in the past 10 years, so having an enabling technology would have a

significant public health benefit. Advanced in vitro models could add more data around risk and benefit.

Dr. Weaver spoke about using the available, imperfect models. There is no known case of DIVI of this pathological type to serve as a positive control. We can only look at available human conditions with vascular injury to see what tracks with animal models with the biomarkers. Dr. Zabka said that in preclinical models, one can examine morphology to detect vascular injury, but this is not possible in humans. We need more specific and sensitive markers to monitor vascular injury in the clinic; that way, we can dose to efficacy prior to onset of injury.

Referring to the evidence grid, Dr. Sauer said it is most helpful in demonstrating to submitters where they are and what they need to collect as evidence in order to meet minimum evidentiary requirements. If submitters have met the minimum standard, they should be ready to tell regulators how they met it. Each table should be different based on the minimum standard for each biomarker for the COU.

Regarding FDA's concern about false negatives, C-reactive protein might be an appropriate comparator for some patient populations. Industry would be primarily concerned about false positives and the need to distinguish DIVI from gastrointestinal injury, for example. Dr. Leptak thought some of these considerations would not be necessary for an initial, limited COU.

Dr. Zabka said that the qualification team does not want to go to the effort of qualifying a bunch of inflammatory markers. She recommended looking at the evidentiary standards to make sure we are moving in the right direction.

2. How much credence does the panel give to DIVI pathology as a way to move forward, based on clinical experience with a number of these small molecules?

   Dr. Rogers suggested having a panel of biomarkers for comparing a healthy volunteer cohort and a treatment cohort.

   Dr. Mačiulaitis said that no one knows the truth about DIVI risk in humans. Published data could be very helpful, but the quality of the evidence from published data is important to consider. Regulators look for data from prospectively designed studies with a defined hypothesis. Are the studies randomized? Are the investigators replicating findings in the same or different cohorts?

   A clinical investigator said that if there were a clear preclinical signal from a biomarker in one or more species, he would be willing to use the biomarker, assuming that it is validated. If he observed a positive signal using the biomarker in humans, he would pause the study, especially if approaching the therapeutic dose. On the flip side, if the biomarker signal is flat in the first-in-human study, is that believable? Some human studies would help shore up the utility of the biomarker.

   Dr. Zabka said that the goalpost seems to keep moving. Regulators want submitters to design studies with adequate statistical power and sufficient data, but there needs to be a point when submitters know the race is done. Otherwise, people will give up on biomarker qualification. Qualification teams are interested in knowing where the goalpost is for benefit/risk and for evidence, filling the evidence gaps, and getting the biomarker into the clinic.

Dr. Mačiulaitis said that if the data are not sufficient for the proposed COU, then a limited COU should be articulated. If the evidence is not robust enough to support qualification, then the regulator may issue a letter of support. Any of these options is beneficial to the developer. How can the framework cover this spectrum?

Dr. Weaver recalled that the original idea behind biomarker qualification was to start small with a minimal data set and then expand the COU as more evidence accumulates.

Q&A Session

Q: Regarding the benefit/risk diagram, a participant suggested that one should not make assumptions about the false negative rate. The false negative rate should be addressed more robustly in order to convince regulators that patients will not be placed at increased risk.

A: *Benefit/risk* is shorthand; it should not be construed as a ratio. Benefit and risk need to be assessed independently. Several participants thought that the shorthand implies that the submitter should combine the two for the qualification process.

Q: Should the evidence map be modified to depict both the current state of the evidence and the aspirational state?

A: The evidence map might shift depending on where the submitter is in the qualification process. Early in the process, the diagram would show where more evidence is needed. Dr. Zabka said that the goalpost seems to keep moving. Regulators want submitters to design studies with adequate statistical power and sufficient data, but there needs to be a point when submitters know the race is done. Otherwise, people will give up on biomarker qualification. Qualification teams are interested in knowing where the goalpost is for benefit/risk and for evidence, filling the evidence gaps, and getting the biomarker into the clinic.

Q: What is the role of the limited COU?

A: The COU is a spectrum. Many biomarker developers have trouble understanding how changing the COU affects data collection. A limited COU is narrow in scope (e.g., population, use in drug development). It can be very restrictive based on the data available to support the COU. In one setting, the submitter may have data to support a limited COU; but if the COU changes, the data might be insufficient. A limited COU is not the same as a letter of support, but it could evolve into qualification. From a framework perspective, we have ideas in a broad sense about gradations of COUs that are mapped to the framework for the data set. A limited COU is a qualification pathway.

Dr. Mačiulaitis added that if the data are not sufficient for the proposed COU, then a limited COU should be articulated. If the evidence is not robust enough to support qualification, then the regulator may issue a letter of support. Any of these options is beneficial to the developer.

Q: How can the framework cover this spectrum of COUs?

A: Dr. Weaver recalled that the original idea behind biomarker qualification was to start small with a minimal data set, and then the submitter could expand the COU as more evidence accumulates. A participant said that a limited COU is not the endgame; rather, it is a milestone on the way to a goal that

justifies the investment and has an impact on society. A limited COU allows the biomarker to be used, resulting in collection of more evidence.

A participant remarked that biomarkers are used all the time in developing drugs and diagnostics. If the COU is so narrow that it applies only to one sort of therapeutic, then qualification might not be the best route. In this case, the developer should seek regulatory acceptance for a single program. Qualification should be for biomarkers to be used in multiple development programs.

Q: A participant spoke about the need for clarity and having an actionable set of correspondences between COUs and evidentiary standards; otherwise, the goalposts will keep moving. Published guidance could include an example of an incremental set of COUs and associated evidentiary standards.

Analytical Validation of Biomarker Assays

Steve Piccoli, Ph.D.

Research Fellow in Immunochemistry and Biomarkers, Bristol-Myers Squibb

In his presentation, Dr. Piccoli discussed the *Bioanalytical Method Validation Draft Guidance*, fit-for-purpose validation, extant validation comparisons, and the work done for the DIKI safety biomarker validation. The draft guidance was issued in 2013, followed by a series of collaborative workshops for various stakeholders.[5,6]

Often when people talk about validation of an analytic test, they are referring to fit-for-purpose validation for use in support of an IND, NDA, or BLA. He clarified that this sort of validation does not apply to clinical applications. The focus of Dr. Piccoli's presentation was on analytic validity, which drives clinical validity.

The assumptions behind assay validation for biomarkers include the following:

- The validation expectations for biomarker assays are the same as those outlined for pharmacokinetic or toxicokinetic assays.
- Assays used in the qualification or implementation of biomarkers in drug development are evaluated such that the performance characteristics align with the COU and the benefit/risk profile.
- Assays used in the qualification or implementation of biomarkers in drug development are not sufficient as de facto substitutes for an in vitro diagnostic device or for approval or clearance by the CDRH.
- Qualified biomarkers are suitable for use in drug development but are not directly acceptable for clinical practice as regulated by CLIA.

In the guidance document, the information on biomarker validation for submissions is very sparse and focused solely on biofluid-based biomarkers.

Dr. Piccoli reviewed the fundamental parameters (e.g., accuracy, precision, selectivity, sensitivity) for analytic validation of biomarker assays. Validation involves specific laboratory investigations to ensure that the performance characteristics of a method are suitable and reliable for the intended analytical

[5] Lowes S, Ackermann BL. AAPS and US FDA Crystal City VI workshop on bioanalytical method validation for biomarkers. Bioanalysis. February 2016;8(3):163–7. Epub January 22, 2016. PMID: 26795584.

[6] Arnold, King, et al. In press.

applications. For pivotal studies that require regulatory action, the bioanalytical methods should be fully validated. For exploratory methods used for the sponsor's internal decision making, a lower level of validation may be sufficient.

The key points to remember are these:

- The guidance pertains only to measurement of biomarkers in biological matrices.
- The fit-for-purpose validation approach should be used.
- Biomarkers for regulatory action require full validation.
- Biomarker data in support of early drug development may be fit-for-purpose validation.

What is fit-for-purpose validation? Accuracy, precision, selectivity, range, reproducibility, and stability of a biomarker assay are important characteristics that define the method. Some of these parameters might not be necessary to ensure that biomarker assays are safe and reliable for their intended use. The fit-for-purpose paradigm is a highly interactive process starting with exploratory method qualification. There are checkpoints based on predetermined criteria.

Dr. Piccoli also discussed the concept of total analytic error. This parameter includes all sources of error including systematic error and random error. This information is necessary in order to set the error bars for clinical decision making. The cutoffs should minimize the "gray zone." Compared with pharmacokinetic assays, biomarker assays have heightened requirements in terms of trueness, accuracy, parallelism, and precision to ensure clinical accuracy.

Preanalytical variables are factors that occur during the time between when a biospecimen is removed from storage and when it is analyzed. These factors affect the integrity of biospecimens and, ultimately, the results of analyses. Frequently, if a problem arises, analytical integrity is questioned when actually it is preanalytical variables affecting the assay. Other sources of variability include, but are not limited to, biological variability, environmental variability, and collection variability.

The fit-for-purpose validation must be sufficiently robust to support eventual clinical validation. Per the draft guidance, analytical validation should include demonstrations of (1) selectivity; (2) accuracy, precision, and recovery; (3) the calibration curve; (4) sensitivity; (5) reproducibility; and (6) stability of analyte in spiked samples. The criteria for fit-for-purpose validation of biomarkers vary according to extent of validation (e.g., discovery/exploratory, partial validation, full validation).

With regard to the clinical validation of the kidney safety biomarker for a limited COU study, Dr. Piccoli said that this was a collaborative effort, intended to extend support for the translational utility of a composite of five urinary kidney safety biomarkers. Each biomarker was qualified for use in rat studies. Dr. Piccoli described efforts to provide support for clinical validation of three of the biomarkers.

The qualified biomarkers were to be used together with conventional kidney biomarker monitoring (e.g., SCr, BUN) in early clinical drug development research under an IND or a Clinical Trial Application). Dr. Piccoli explained the methods used for assessing precision, functional sensitivity, dilutional linearity, sample stability, and other parameters. The clinical submission is in final review, and a decision is expected shortly. The output is a multiplex analysis—a composite measure—rather than results of the singlicate markers. The biomarkers are measured independently but interpreted together as a panel.

To ensure valid and reliable clinical conclusions, the level of analytical rigor and quantity of generated data must be based mainly on the biomarker-specific COU and the implications of the concomitant benefit/risk profile. Therefore, assay performance must be derived directly from the COU.

Analytical Validation Considerations: Panel Discussion and Audience Q&A

Steven Piccoli, Ph.D., Bristol-Myers Squibb, Discussion Leader

Panel Members:

Shashi Amur, Ph.D., FDA, CDER
Robert Becker, M.D., Ph.D., FDA, CDRH
Steve Gutman, M.D., M.B.A., Myraqa/Illumina
Gary Kelloff, M.D., NCI
Dan Krainak, Ph.D., FDA, CDRH
Meena Subramanyam, Ph.D., Biogen, Inc.

1. The documents identify characteristics that define assay performance. Is this a complete list?

   Dr. Gutman said that more detail could be added. For example, sample stability should include more specifications. Dr. Subramanyam said that sample handling (e.g., collection tubes, time in transit, shipping method) are all very critical for biomarkers.

   Dr. Krainak said it is helpful if the submitter has thought about how the biomarker might be used in the future. What are the important characteristics for the given COU? He prefers to provide some guidance for future users knowing that they might not use the same exact equipment in the future.

2. When should test parameters be locked in? What if the team wants to make changes to enhance performance?

   Dr. Kelloff recommended locking in parameters early. Otherwise, the team might have to start over or do a comparison study if something changes.

   Dr. Krainak recommended standardizing procedures for acquisition, analysis, and interpretation. From his experience as a member of a Biomarker Qualification Review Team, he suggested defining the performance requirements rather than the methods. Dr. Amur suggested starting with another validated assay and then doing a bridging study. Going from a functional assay to an imaging assay would be challenging, but just changing a reagent should not be a major problem. Validation involves many parameters, in the exploratory phase, a team might focus on fitness for purpose rather than a full validation until it is sure that the assay will be used for decision making. If the COU changes, Dr. Subramanyam pointed out that the team might have to revalidate the assay based on medications or infections in study population.

3. What happens if one changes the format (platform) of the assay? What must be done to revalidate the assay? Also, if an assay has been validated in one laboratory, is it considered to be validated in all laboratories?

   One suggestion was to build a migration pathway that could reduce the work necessary. Bridging studies could also work. The goal is to match the original method. One could reserve

some of the original samples to test with the new method or prepare spiked samples for bridging studies.

Dr. Amur clarified that when CDER qualifies a biomarker, it is not qualifying the assay. The assay needs to work well; that is why CDER asks for validation. The assay is described in the draft guidance for the biomarker. If someone wants to use an alternative assay, the review division should be consulted for advice.

When a team qualifies a biomarker, it uses a specific cutoff using a specific assay. If others use alternative assays, how can the biomarker result be correctly interpreted?

Dr. Amur said that CDER staff consult with their CDRH colleagues who provide advice. CDER often relies on bridging assays. Dr. Subramanyam said that with a new assay, it might be necessary to set a new cutoff, but the submitter needs to demonstrate via a bridging test that the same COU is supported. Dr. Becker recommended looking at key performance characteristics and checking the new assay carefully to identify any new vulnerabilities.

4. Should the bar for validation be the same for all types of biomarkers (e.g., proteins, inorganic compounds, microRNAs)?

   The bar is set on the basis of the risk of the assay for the COU. Dr. Amur added that for RNA, sample preservation is key. Other validation parameters (e.g., accuracy, precision) would be the same.

   Regarding interfering substances, the list could get very long if the submitter has to take into account how the biomarker would ultimately be used in the clinic. Dr. Piccoli suggested listing the drugs that would most commonly be encountered; it is not possible to test everything in advance. The majority of patients in the target population might be on certain drugs.

Q&A Session

Q: What about adding a pH stabilizer, such as boric acid, to urine samples? Also, should it be a first-morning urine sample or a random sample?

A: Dr. Piccoli said that many components vary widely in urine. The guidance covers handling, stabilization, and other processing steps. The qualification team tested and controlled for many factors and demonstrated the range of acceptability for the assay. Dr. Becker added that reference materials and methods are very important when elucidating performance characteristics. Biological variability (e.g., diurnal variation), use of whole blood vs. serum, anticoagulant selection, and even the brand of collection tube can affect analytes.

Q: Would it be acceptable to prospectively specify some variables rather than expending a great deal of money to study them?

A: Dr. Krainak said that specification is acceptable, but future users will probably have to examine the qualification documentation to see if a variable was explored or specified.

Q: It sounds as if there is an expectation that submitters will have to enable future uses of the qualified biomarker. We do not have a single guidance that pulls all this information together.

A: Dr. Piccoli said that he hopes that one outcome from this meeting would be clarification of expectations from FDA in the form of a guidance, which could benefit the whole community.

Panel Discussion and Audience Q&A — Defining Evidentiary Criteria for Safety Biomarker Qualification: Testing the Framework and Refining the Model

John Wagner, M.D., Ph.D., Discussion Leader

Panel Members:

Gary Kelloff, M.D., NCI
Christopher Leptak, M.D., Ph.D., CDER, FDA
John-Michael Sauer, Ph.D., C-Path
Frank Sistare, Ph.D., Merck & Co., Inc.
Thorsten Vetter, M.D., EMA
David Wholley, M.Phil., FNIH

General Framework

The important question is whether the framework is right or not. During the course of the meeting, many opinions were heard about the best order for the various elements of the framework. The group voted by show of hands on four different options in terms of the order of the thought process.

1: COU — Benefit/Risk — Evidence (about 8 votes)
2: Benefit — COU — Risk — Evidence (about 17 votes)
3: COU — Evidence — Benefit/Risk (about 20 votes)
4: COU — Risk-Mitigating Circumstances (i.e., Benefit) — Evidence (about 10 votes)

Dr. Leptak contrasted the lumping and splitting paradigms for assessing benefit and risk. He talked about considering benefit and risk in terms of the COU.

Mr. Wholley said that the impetus behind the development of evidentiary criteria was to reduce uncertainty and minimize iterative back-and-forth discussions with regulators. He further clarified that benefit and risk is defined from the patient's perspective. However, no one expects that every interaction between submitter and regulator will follow this path. It is not envisioned as a rigid process. It will flow back and forth differently. Some reviewers will look at risk first and then consider benefit and then whether the COU needs revision. We need to incorporate the patient perspective, hence the inclusion of benefit.

COU

Dr. Sauer suggested focusing COUs more narrowly. Dr. Leptak thought it would be a good idea to add language to the framework to clarify that expanding the breadth of a COU most likely would require more evidence.

Dr. Leptak said that simply understanding what would be needed to use a biomarker in a clinical trial is important.

The panel members discussed the possibility of defining some COUs for safety in the hepatic space and identifying some potential sources of evidence. The next step would be to propose evidentiary standards, at least to a first approximation.

Benefit/Risk Profile

Dr. Wagner said that no participants seemed satisfied with the "rising sun" diagram in the proposed framework depicting benefit/risk profile and evidence level. Dr. Wagner advocated using both graphics and text to explain the components of benefit and risk and ways to mitigate risk. He also recommended asking risk experts to weigh in.

Dr. Sistare spoke of the benefit of a biomarker being based on the unmet need it is intended to fulfill. He also said that the risk of a biomarker does not translate to the IND space in the same way as for qualification.

Dr. Wagner said that, for biomarker qualification, when one speaks of risk, the question is straightforward: What is the risk of getting a wrong answer? The biomarker developer and the regulator need to understand that risk.

Evidence

Dr. Wagner said that the feedback heard during the workshop indicated some dissatisfaction with the evidence map. The consensus seemed to be that it is salvageable, however. He noted that there is some confusion about whether the grid represents the current state, the aspirational state, or both. The grid depicts the expectations of the biomarker's analytical performance and other requirements for a given COU.

Dr. Sauer thought that two tables would be needed in order to compare the current state with the minimum standards required. Dr. Leptak agreed that both the current state and the aspirational state should be included in order to home in on the elements that need further discussion and/or development.

According to Dr. Sistare, the intention behind the evidence map was to establish in the submitter's mind what is needed in terms of ensuring that a biomarker assay is fit for purpose and based on the scientific understanding of the biomarker. The question is twofold: Where do we stand today? What is the minimum standard for qualification compared to the current state? The qualification team must achieve the minimal status for each element of the COU.

Dr. Leptak said that the messaging around the evidence map will be very important. The evidence map should depict the range of variability for any given element, taking into account the full knowledge of the field.

Q&A Session

Q: How does the benefit/risk profile affect the level of evidence required?

A: A participant emphasized the concept that the benefit/risk profile influences the level of evidence needed and suggested the benefit/risk assessment could be used to rank COUs and set evidentiary standards. It would be rational to rank COUs based on seriousness of risk and triviality of benefit. If there are few consequences for untrue results (as would be the case for biomarkers for Parkinson's disease, for example), then the evidentiary requirement would be minimal.

Q: What are some alternatives to the "rising sun" diagram depicting the relationship between benefit/risk and evidence levels?

A: Alternatives suggested included a diagram based on quadrants, using text (not graphics) to explain how the benefit/risk profile influences the level of evidence required, using both graphics and text to explain the components of benefit and risk and ways to mitigate risk, and applying a quasi-mathematical thought process to assess risk and benefit. Benefit and risk need to reflect the patient perspective.

Summary of Next Steps — Toward a Guidance for Safety Biomarkers

Christopher Leptak, M.D., Ph.D.
Acting Director, Office of New Drugs Regulatory Science Program, and Co-Director, Biomarker Qualification Program, Office of Translational Sciences, CDER, FDA

Dr. Leptak summarized the highlights of the workshop and outlined some action items:

1. Use the terminology set forth in the BEST Resource.
2. When pursuing biomarker qualification, talk to FDA staff early and often.
3. Consider issuing a set of safety COUs along with specific starting evidence levels.
4. Address stakeholders' concerns about the iterative nature of biomarker qualification to ensure that the process does not become a series of never-ending loops.
5. Improve formatting of conversations between submitter and FDA during the advice and consultation step.
6. Find ways to encourage broader access to data sets.
7. Convene an analytical validation workshop and consider writing a paper on the topic.

Dr. Leptak said that the workshop was an excellent opportunity to engage with the external community: pharmaceutical companies, patient representatives, academia, and clinicians. This is a unique and emerging field. Each person has his or her own perspective, but it takes a community to understand this field and move it forward.

Action Items and Wrap-Up

David Wholley, M.Phil.
Director of Research Partnerships, FNIH, and Manager, Biomarkers Consortium

Mr. Wholley asked the participants for their thoughts on structuring the framework and guidance for evidentiary criteria and other suggestions to advance biomarker qualification. He asked for input to help the writing committee. The planning committee will continue to meet and work for a few months to further refine the documents and make them available to the public.

Several concrete action items for FDA were suggested by one participant:

1. Set up "safe harbor" databases to aggregate precompetitive data on acute DIKI and DILI. CDER can take action on this.
2. Develop "mother" guidance and "baby" guidances. The overarching guidance for the conceptual framework could address such questions as: How do we quantify benefit and risk? Why do we need this level and quality of evidence? Possible "baby" modules could include assay validation, statistical considerations, analytics, and particular evidentiary frameworks (e.g., for safety biomarkers).
3. Explore the possibility of convening a workshop on analytical validation, being mindful of the boundaries around the Center for Devices and Radiological Health (CDRH) and clinical validation.

Another person remarked that the topic of data sharing was discussed at a Predictive Safety Testing Consortium (PSTC) workshop at FDA. Sharing precompetitive data from animal and human studies is of paramount importance. A safe harbor database could help advance new biomarkers and other areas of science. Additionally, the IQ Consortium is talking about data sharing for biomarkers. C-Path is committed to the PSTC and other groups to make sure analytical and statistical conferences occur.

A participant recommended taking actions for reducing the research burden to stimulate progress in rare diseases. Surrogate endpoints are of particular importance. Some participants discussed requirements for postmarketing data collection to show clinical benefit if surrogate efficacy endpoints are used in IND trials.

Mr. Wholley adjourned the meeting after thanking the members of the writing committee, authors of the case studies, meeting organizers, FNIH staff, panelists, presenters, and attendees.