Evidentiary Considerations for Integration of Biomarkers in Drug Development

A Symposium co-sponsored by

University of Maryland's Center of Excellence in Regulatory Science and Innovation (M-CERSI)

U. S. Food and Drug Administration (FDA)

Critical Path Institute (C-Path)

August 21, 2015

TABLE OF CONTENTS

Introduction

The Center for Drug Evaluation and Research's (CDER) Biomarker Qualification Program was established in 2005 by the U.S. Food and Drug Administration (FDA) to support work to develop biomarkers and provide a framework for scientific development and regulatory acceptance of biomarkers for use in drug development. Biomarkers have been defined as "characteristics that are objectively measured and evaluated as an indicator of normal biological process, pathogenic processes, or biological responses to a therapeutic intervention". Biomarkers are used across all stages of drug development, from understanding the molecular pathways underpinning a disease to determining the mechanism of action of a compound. Biomarkers are also used to assess safety in preclinical studies and clinical trials, to determine optimal drug dose, to stratify patients and select those most likely to respond to the treatment, and to track patient response to the treatment. In addition to the qualification of biomarkers, the role of the CDER Biomarker Qualification Program includes facilitating the integration of biomarkers into the regulatory review process, encouraging the identification of novel and emerging biomarkers and reaching out to stakeholders in industry and academia to foster biomarker development.

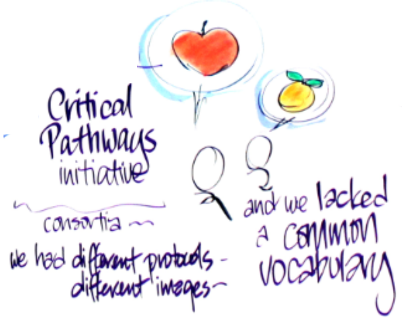
In the last eight years, there have been six qualification decisions by the FDA (encompassing thirteen specific biomarkers), providing the regulatory certainty to use these drug development tools (DDTs) in drug development programs. With the learning from the biomarker qualifications completed to date, and a recognition by all stakeholders of the need for greater clarity on evidentiary expectations within the process, the University of Maryland's Center of Excellence in Regulatory Science and Innovation (M-CERSI), the FDA and Critical Path Institute (C-Path) co-sponsored a symposium entitled "Evidentiary Considerations for Integration of Biomarkers in Drug Development," which was held at the University of Maryland School of Pharmacy on August 21, 2015. The first in a series, this symposium was designed to bring together biomarker qualification stakeholders from industry, academia, and regulatory agencies to begin a conversation which will ultimately lead to defining and codifying evidentiary considerations that will help drive the process of regulatory acceptance of biomarkers for use in drug development.

Key Note Address: Dr. Janet Woodcock

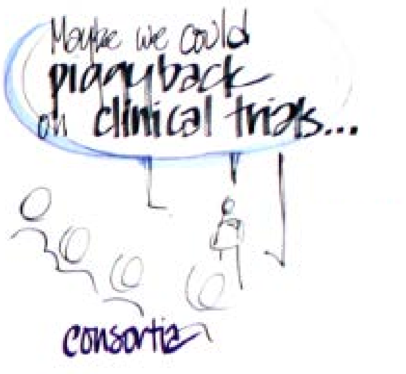
The key note speaker for the symposium was Dr. Janet Woodcock, Director of CDER at the FDA. During Dr. Woodcock's 20 year tenure at FDA, she has been involved in a number of new initiatives, including the Critical Path Initiative designed to move medical discoveries from lab to patient more efficiently. In her talk, Dr. Woodcock outlined where we are in the evolution of the biomarker qualification process, how we got to this point, and where we need to go from here.

The FDA has long accepted use of biomarkers for diagnosis, enrichment and monitoring of safety and efficacy, relying on the research community for supportive evidence. Historically, as the scientific community accepted the utility of a biomarker, the FDA also accepted it. On the other hand, certain failed examples and the lack of a clear process for scientific rigor around the development of biomarkers, led to guarded regulatory skepticism, slowed acceptance, and the sluggish use of biomarkers in drug development.

Dr. Woodcock pointed out the critical importance of partnerships to biomarker development. Single companies with limited data and a lack of scholarly vetting for their biomarker approaches may not have the resources to gain acceptance from the FDA. In addition, although academics can get funding to discover new biomarkers, they typically cannot get funding to develop robust performance data (e.g. analytical validation and clinical correlation studies) to support biomarker reliance for regulatory use. It clearly “takes a village”, and consortia were formed to bring together industry, academic and regulatory partners to develop biomarkers. These consortia began with the idea that they would first go after “low-hanging fruit” in the biomarker space.


Critical Pathways initiative
consortia
we had different protocols - different images -
and we lacked a common vocabulary

It soon became clear however, that there were issues moving even the “low hanging fruit” forward, due to the lack of standardized, validated assays for biomarker measurement and the use of different study protocols that prevented pooling of data. Other barriers included a lack of common vocabulary among clinicians and diagnostic scientists, and the lack of clear regulatory "goal posts" for qualification. In order to better define the evidence needed for biomarker qualification, the FDA developed the context of use (COU). The COU is meant to define what the biomarker will be used for and what decisions it will drive within a drug development program. The FDA can then use the COU to define the evidence needed to provide confidence in its use. However, Dr. Woodcock acknowledged that the evidentiary goal posts are not currently well understood by the scientific community or the FDA.



The next steps in the evolution of the biomarker qualification process should start with the basics, such as putting in place clear guidelines for assay validation, sample preparation, and standard practices. Additionally, it is essential to develop a taxonomy or common vocabulary, to enable mutual comprehension, and to understand the data needed to generate the evidentiary goal posts. Since these data may be derived from intervention trials and not just the natural history studies, it would be prudent to piggy back on the ongoing clinical trials to generate such evidence.

Evidentiary standards can help define what the regulators require (i.e. what kind of evidence they want) to make a biomarker qualification decision. This is not the same process as making a marketing decision for a new drug, so evidentiary standards must be defined specifically for the purpose of biomarker qualification. The overarching framework of the evidentiary criteria that would match up with a biomarker will be related to the level of risk associated with dependence on the use of biomarker data to make decisions about human health. For example, a poorly performing biomarker used to enrich clinical trial populations carries the risk of increased noise, missed drug efficacy, failure of the trial and futile use of human subjects. A safety biomarker, particularly a stand-alone safety biomarker, carries an even higher risk of harm, so higher evidentiary standards are needed for safety biomarkers. Some intermediate steps are possible with safety biomarkers, but this is also a situation where more conversations are needed. C-Path was used as an example of a consortia-based approach that has taken on this high burden in safety biomarker qualification. It was also pointed out that surrogate endpoints for efficacy versus for accelerated approval have different legal standards and still need to be defined. Together, this illustrates the clear need for a biomarker classification system and the FDA is working with National Institutes of Health (NIH) to accomplish this. The results will need vetting and acceptance from the scientific community.

The FDA alone does not have all of the answers when it comes to the biomarker qualification process. In this era of molecular and precision medicine, all the qualification stakeholders are learning together. While there's has been much learning in eight years, now is the time to define evidentiary criteria for biomarker qualification and this will need to be done by the scientific community working together. While there is a considerable repository of knowledge, Dr. Woodcock suggested that biomarker qualification and the evidentiary criteria need to be made a field of scientific endeavor.

Session 1: Overview of Biomarkers in Drug Development

FDA's Efforts to Encourage Biomarker Development and Qualification

The FDA has two pathways to facilitate the use of biomarkers in drug development programs (Amur et al., 2015). The traditional path, where biomarkers are incorporated into an Investigational New Drug (IND), New Drug Application (NDA), or Biologics License Application (BLA) submissions, places the entire burden on the individual sponsor who is submitting the application, and acceptance of that biomarker pertains only to a single drug development program.

Biomarker qualification is an alternative pathway that allows a biomarker to be used for multiple drug development programs. In this case a submitter, often a consortium, would contact the Biomarker Qualification Program and follow the qualification process for a biomarker. Consortia are collaborative partnerships with members from academia, industry, patient advocacy groups and others, working together to collect, analyze and present data to support biomarker qualification, thus sharing the work and the risk among their members. While biomarker information is embedded in the drug labels and reviews in the case of individual use, qualified biomarkers are announced as draft guidance to the drug development community. The FDA has published several guidance documents related to biomarker qualification (FDA 2011; FDA 2012; FDA 2014). Qualification indicates that the FDA accepts the use of the biomarker in a drug development program within a stated COU.

A number of factors need to be considered in order to qualify a biomarker. First is the type of biomarker. Four specific types of biomarkers have been defined. Diagnostic biomarkers identify patients with a particular disease or a disease subset. Prognostic biomarkers indicate future clinical course with respect to a specified clinical outcome in the absence of therapeutic intervention. Predictive biomarkers identify patients likely to respond (favorably or unfavorably) to a specific treatment. Response biomarkers indicate that a biological response has occurred in a patient after having received a therapeutic intervention. In addition, there are three types of defined response biomarkers. Pharmacodynamic response biomarkers are indicators of the intended activity of the therapeutic and are not necessarily strong predictors of efficacy. Efficacy response biomarkers predict specific disease-related clinical outcome and could serve as primary clinical endpoints or surrogates for a clinical end point. Safety-related response biomarkers are indicators of potential adverse drug reactions and are likely to be specific for a type of drug toxicity and usually organ specific.

In addition to the type of biomarker, if data are available, the biological rationale for use of the biomarker is important to understand. Next, the COU, a comprehensive statement that fully and clearly describes the manner and purpose of use for the biomarker in drug development, should be considered. The relationships among the biomarker, the clinical outcomes, and the treatment (where applicable) need to be understood. The use of appropriate pre-specified statistical methods to demonstrate the hypothesized relationships for the COU are also required.

The type of data available to assess the strength of association of the biomarker with its proposed clinical outcome, whether retrospective or prospective, registry data, and/or randomized controlled trial (RCT) data, is a necessary consideration for qualification. In addition, analytically validated assay methods and an understanding of potential sources of variability in the measurement are key considerations for biomarker qualification. Reproducibility of data is a very important consideration and thus, it helps to have a dataset to evaluate the biomarker (test dataset) and a separate dataset (confirmatory dataset) to verify the findings. Finally, the strength of evidence supporting the biomarker is important to consider. The level of evidence that is necessary for biomarker qualification depends on the type of biomarker and its proposed COU.

The qualification process is comprised of three stages: Initiation, Consultation and Advice, and Review. On average, the biomarker qualification process takes approximately 2-3 years to complete, not including the time it takes for submitter to generate and analyze the data and the iterations in the process requiring submitter and FDA responses. To date, FDA has issued six qualification decisions including thirteen biomarkers – ten preclinical biomarkers and three clinical biomarkers. Three other qualification packages are in the review stage and an additional 24 have been submitted.

Recently, the FDA has initiated a new process called the Letter of Support (LOS) for biomarker development programs that do not yet have sufficient data for clinical qualification. The intention of the LOS is to express the FDA's support for the qualification effort, promote data collection and sharing, and stimulate additional studies that could eventually lead to qualification. To date, seven LOS have been issued.

Another way the FDA is encouraging biomarker qualification is through the use of a limited COU in order to expedite the integration of the biomarker in drug development and to potentially generate additional data that can help in qualifying the biomarker for an "expanded" context of use.

In addition, the FDA holds Critical Path Innovation Meetings (CPIM) on a case-by-case basis to enable CDER to meet with investigators from industry, academia, patient advocacy groups, and other governmental agencies to discuss general challenges in drug development. CPIM can also be used to discuss findings with exploratory biomarkers. A guidance document was recently issued to describe how one might request a CPIM (FDA 2015a).

The FDA is also working to improve communication around the biomarker qualification process by enhanced interaction with submitters. This includes beginning COU discussions very early on in the process. There are also ongoing efforts to enhance interactions with consortia through organizations such as National Center for Advancing Translational Sciences (NCATS), Foundation for the National Institutes of Health (FNIH), and C-Path, and European partners such as the Innovative Medicines Initiative (IMI). Presentations, publications and an FDA webpage containing information for submitters, as well as information about previously qualified biomarkers, are also part of the communication strategy. The FDA has recently completed an external survey via a Federal Register (FR) posting (FDA 2015b) intended to identify potential biomarkers for qualification and describing contexts of use, to address areas important to drug development. A summary of comments was recently posted on FDA's webpage (FDA 2015c). In addition, an internal survey is currently ongoing within FDA. The FDA is also initiating additional collaborative workshops aimed at developing evidentiary standards, including the M-CERSI symposium summarized here.

Evidentiary Considerations for Biomarkers: Statistical Considerations

The statistical considerations for determining evidentiary standards for biomarker qualification are integrally tied to the role the biomarker plays in drug development whether it be stratification, prevention, screening or diagnosis of disease; prognosis; prediction of risk or therapy-related risk management; or therapy monitoring or surveillance (IOM 2010).

Statistical considerations prior to the design of the statistical analysis plan include the relation of the biomarker to its COU, including acceptable, analytically validated measurement methods, the operating characteristics of the assay, the relationship of the biomarkers' components to each other for composite measures (CM), and patient characteristics or covariates that have an effect on biomarker expression.

The use of a "learn and confirm" paradigm is one approach that can be applied to a qualification program. This approach uses exploratory analyses in the learning stage with expression, definition of threshold and relevant covariates. In the confirmatory phase, consideration of sample size and derivation of an implementable analysis plan are important.

The design of the statistical analysis plan should consider statistical methods for appropriately identifying multiple predictors. An understanding of biomarker levels and how they would be measured is also important. Specific considerations that should be addressed include the baseline biomarker level, the change relative to baseline, and whether change will be measured as a difference from baseline or a relative change. If intra-subject variability is greater than the inter-subject variability, then a linear change from baseline is a better measure. If inter-subject variability is greater than the intra-subject variability, then a relative change from baseline may be preferred. The threshold for meaningful change is important to establish and may involve relevant single or multiple time points. The threshold for meaningful change should also consider replicability and possible multiplicity issues. The model selected to investigate the threshold should be parsimonious to avoid over-fitting.

Reference standards are critically important to the design of any qualification program. If there is no established gold standard biomarker, all available information should be used. If a flawed gold standard or a “pseudo-gold standard” is used as reference, the new biomarker may lack sensitivity and its estimation may be biased. The extent of this bias depends on the correlation between the new gold standard and the pseudo-gold standard biomarker. In some cases, the use of adjudication committees may help to mitigate this bias.

Analysis issues are also important to consider. The use of cross validation can be a powerful tool if all model building steps are included and the model does not require external variable selection or outcome evaluation. In addition, whether the model uses a single-fold or k-fold validation approach needs to be clearly described. Furthermore, if a credible validation comes from a separate trial, then lack of outcome knowledge is preferable.

Interim analysis can be used in both learning and confirming phases. Early interim analysis can assess initial performance, modify biomarker thresholds or be used for sample size re-estimation. Later interim analysis can assess performance improvements based on modifications, or stop the trial based on futility. However, the objective of the interim analysis and its effects on Type I error must be clearly defined.

The analysis plan should include pre-specified hypotheses of interest, possible multiplicity adjustments, procedures to handle missing data, and plans for secondary comparison. The analysis plan should also avoid inflation of Type-I error.

Key elements for a retrospective analysis include acceptable, well controlled studies with a large enough sample size to ensure adequate power. Furthermore, the biomarker of interest must be evaluated in the intention to treat population. It is important that assays utilized in the study are well-characterized, with acceptable analytical performance and the same assay(s) used in all the studies. The integrity of the analysis plan is questionable if it occurs after the efficacy data have been unblinded and the biomarker status is known. The analysis plan should control multiplicity and the study-wise Type I error. It is important to emphasize that retrospective evaluation is not to be used to salvage a negative study.

If using a prospective-retrospective design, the biomarker hypothesis is prospectively specified prior to diagnostic assay testing. Samples are collected prior to treatment initiation and may be stored for later use. The biomarker classification is then conducted using a validated assay to characterize the biomarker for the proposed COU. The clinical outcome data may have already been (partially) collected, unblinded, and analyzed. However, if the prior analyses did not include biomarker data, the biomarker analysis might be considered as "prospectively" performed with a "retrospective classifier analysis".

In summary, meticulous planning prior to undertaking the biomarker qualification project is critical. It is also critical to account for mid-course modifications to the project. Collaboration across stakeholders is also important, including planning for multiregional factors which can be especially challenging. And finally, early engagement of regulators is a step that can improve the chances of overall biomarker qualification success.

Assay validation and reproducibility considerations for biomarkers used in drug development

The use of biomarkers in target validation, early compound screening, pharmacodynamic assays, patient selection and as surrogate endpoints has increased the demands on biomarker assay performance. For preclinical, *in vitro* experiments or assays to measure biomarkers in animal models, research-grade assays and possibly molecular imaging are commonly utilized. Biomarker assays in Phase 0-I trials must meet minimal analytic performance standards, while in later Phase II-III trials assays must demonstrate more robust analytic performance.

An assay must be validated to establish that it is "fit for purpose" for the role that it is intended to play in the drug development process. Specifically, analytic validation must address whether the assay reliably measures what it is intended to measure and whether the assay can be performed on the types of specimens available. The analytic performance of the assay must meet acceptable standards to establish that it is appropriate for the biomarker's COU. In situations where the biomarker is analyzed across multiple laboratories, harmonization of the assay's conduct and performance across all sites is critical.

Specimen and pre-analytic factors that must be considered include patient physiologic factors and state at specimen collection, specimen collection method, processing and storage, specimen quality screening, minimum required amount of specimen, and the feasibility of collecting the necessary specimens in a clinical trial setting.

Critical performance characteristics to be established in analytic performance evaluation include precision and reproducibility, linearity, bias and accuracy, analytic sensitivity, analytic specificity and sample stability. Evidence must be provided to support a clinical cut-off (if applicable). The evaluation should recognize that the trade-off between sensitivity and specificity will depend on risks associated with false positives and false negatives. "Optimized" cut-offs produce overly optimistic accuracy results and require validation with independent data. Cut-offs may not be transportable for assays that lack reproducibility between laboratories or scorers.

In summary, expectations for assay analytic performance in the COU for a specific drug development program or to support formal qualification of a novel biomarker need to be more clearly defined and communicated to the various stakeholders. The codification of such expectations should be a fundamental objective of defining evidentiary considerations for biomarker qualification.

Session 2: Evidentiary Considerations for Clinical Safety Biomarkers

Mechanisms of Drug Toxicity & Relevance to Pharmaceutical Development

Drug toxicity is one of the major reasons for termination of drug candidates in development. The different types of drug toxicities include on-target toxicity (e.g. mechanism-based; same receptor, wrong tissue such as the statins), hypersensitivity and immunological reactions (e.g. penicillins), off-target pharmacology (e.g. terfenadine and hERG channel effects), bioactivation to reactive intermediates (e.g. acetaminophen) and idiosyncratic toxicities. The key to better enabling drug development is to move assessment of toxicological issues to earlier stages of compound development using *in vitro* strategies and improved, translatable biomarkers for drug-induced tissue injury. Safety biomarkers offer the opportunity to better understand the relevance of preclinical safety signals and allow for a more informed discussion during the clinical phase of drug development.

A Case Study – Clinical Safety Biomarkers – Including Methodological Considerations

Translational safety biomarkers are intended to increase the ability to monitor potential drug-induced tissue injury. The objective of translational safety biomarker qualification is to demonstrate the predictive accuracy of the biomarker to detect tissue injury in humans. In nonclinical studies, the performance of the novel biomarker can be anchored to histopathological changes (i.e. the gold standard biomarker in nonclinical toxicology studies), as well as to standard biomarker performance. In clinical studies, demonstration of the predictive accuracy of the biomarker to tissue injury cannot be directly determined, as histopathology is rarely evaluated in clinical studies. In clinical studies, it must be proved that the novel biomarker outperforms the standard biomarker (i.e. gold standard for safety in clinical studies) where one exists. Scientific and regulatory expectations, or the evidentiary standards needed for biomarker qualification, increase as the COU broadens and as risk increases with failure of the biomarker. Two hypothetical examples of evidentiary standards associated with safety biomarkers are given below.

The first hypothetical case is for a safety biomarker for drug-induced pancreatic injury with a broad COU based on two prospective purpose-designed clinical trials with supporting nonclinical data. The clinical diagnosis of drug induced pancreatic injury remains a challenge due to the lack of specific symptoms. Amylase and lipase are the gold standard biomarkers for pancreatic injury. Amylase concentrations greater than three times the upper reference limit indicate injury. Lipase activity parallel the increased amylase activity. However, many conditions that might present with similar clinical symptoms are also associated with increased amylase and lipase activities. In addition, amylase and lipase tests are among the more poorly standardized tests in laboratory medicine.

The hypothetical COU for two novel biomarkers of drug-induced organ specific injury, Protein RA1609 and Protein RT2864, in healthy volunteers and patients with normal pancreatic function is as follows:

> The qualified pancreatic safety biomarkers are proposed to be used together with monitoring of conventional pancreas biomarkers (e.g., serum amylase and lipase), in early clinical drug development research to support conclusions as to whether a drug is likely or unlikely to have caused a mild injury response in the pancreas at the tested dose and duration.

The nonclinical data supporting use of these novel biomarkers in a clinical trial design include (1) demonstration that the biomarkers are responsive to pancreatic injury in an animal toxicology study, (2) evidence of mild pancreatic injury that is expected either not to be human relevant or to have a satisfactory safety margin over the targeted clinical therapeutic exposure, and (3) prior evidence in an animal toxicology study that pancreatic injury can be safely monitored.

Implementation of the novel safety biomarkers in a clinical trial designed to evaluate the safety of a drug may include qualified biomarkers (Serum RA1609 and RT2864) with the following conditions. The biomarkers will be used to make decisions in real time such that an individual patient or an entire dose-cohort of subjects may be triggered to stop or to pause dose escalation of a drug when a pre-specified biomarker threshold is exceeded. The change in biomarker serum concentrations, as defined by change from baseline, will enable the conclusion that a mild pancreatic injury response to a drug candidate was likely or not likely to have occurred in response to a drug in individual subjects. The biomarkers are intended to complement the use of the standard biomarkers, including lipase and amylase, and should be evaluated in conjunction with standardly used safety monitoring.

A generic decision tree for safety biomarker implementation would include a step where the response of the novel biomarkers is correlated with organ injury. If there is no correlation, then the novel biomarkers are not appropriate for use in this COU. If there is a correlation, then the novel biomarkers can be used in early clinical studies, in conjunction with standardly used safety monitoring, to evaluate organ injury including a determination of whether or not the biomarkers respond as described in the COU. If the biomarkers respond as described in the COU, then it can be concluded that drug-induced organ injury is occurring in an individual. If the biomarker does not respond as described in the COU, then it can be concluded that drug-induced organ injury is absent in an individual.

The predictive accuracy of a safety biomarker can be assessed in nonclinical studies using multiple studies with multiple organ specific toxins primarily in the rodent with limited studies in canine and nonhuman primate. Biomarker response is then correlated to pathology and the performance compared to other biomarkers. The mechanism of the response and the relevance to toxicity can also be determined. There should be a consistent response across mechanistically different compounds, a similar response across sex, strain, and species, the presence of a dose response, and a temporal relationship to the magnitude of response. Understanding the biomarker response to toxicities in other tissues or to pharmacologic effects without toxicity in the target organ, will define the specificity of the biomarker response to the observed toxicity.

In clinical studies, two prospective studies in patients with currently used medications that have the potential to cause organ specific injury should be evaluated. The predictivity of the novel biomarker is then compared to standard biomarkers using a formal adjudication procedure and a predefined statistical evaluation. The risk of novel safety biomarkers that lack predictive accuracy is the safety of individuals in clinical trials and must be a consideration in the level of evidentiary standards required. In this case, the risk is mitigated by the fact that the novel biomarkers will be used in conjunction with the gold standards.

The second hypothetical COU for serum Protein RA1609, Protein RT2864, and Trypsinogen-3 is as follows:

> A Composite Measure (CM) of serum Protein RA1609, Protein RT2864, and Trypsinogen-3 is a qualified safety biomarker of pancreatic injury response for use in normal healthy volunteer trials supporting early drug development.

The nonclinical data supporting use of these novel biomarkers in a clinical trial design includes (1) demonstration of the biomarkers' responsiveness to pancreatic injury in an animal toxicology study, (2) evidence of mild pancreatic injury that is expected either not to be human relevant or to have a satisfactory safety margin over the targeted clinical therapeutic exposure, and (3) evidence in an animal toxicology study that pancreatic injury can be safely monitored.

Implementation of the novel safety biomarkers in a clinical trial designed to evaluate the safety of a drug may include qualified biomarkers (Serum RA1609, RT2864, and Trypsinogen-3) with the following conditions. The CM is defined as a measure of serum RA1609, RT2864, and Trypsinogen-3 expressed as fold change from baseline. The group average CM is qualified for study sponsors to determine if there is an increased likelihood of a pancreatic injury response for a dose of an investigational drug in a dose cohort when benchmarked to results provided herein for normal healthy volunteers (NHVs). The CM is not qualified for individual subject safety monitoring. The biomarkers are intended to complement the use of the standard biomarkers, including lipase and amylase, and should be evaluated in conjunction with standardly used safety monitoring.

A generic decision tree for safety biomarker implementation would include a step where the response of the novel biomarkers is correlated with organ injury. If there is no correlation, then the novel biomarkers are not appropriate for use in this COU. In this case, if there is a correlation, the three novel biomarkers will be used as a measure in the single ascending dose (SAD) first-in-human NHV study to evaluate average CM for each dose group in conjunction with standardly used safety monitoring. If the biomarker data are greater than the CM threshold, the dose is potentially unsafe. A decision to investigate this dose further should be considered in the context of other clinical data. If the biomarker data are less than the CM threshold, the drug doses can continue to be investigated in the next NHV trial (i.e. multiple ascending dose (MAD) study), assuming other clinical data are reassuring with no evidence of a safety signal. The three novel biomarkers should be measured in the NHV trial to evaluate the average CM for each dose group.

In this COU for a CM, in addition to the considerations listed above for the first hypothetical COU, it is necessary to demonstrate that a CM of novel biomarkers can differentiate cohorts of healthy subjects experiencing drug-induced pancreatic injury from cohorts not experiencing injury. This requires one study in healthy subjects to define the variability associated with the biomarkers and one study in Crohn's disease patients treated with azathioprine known to cause pancreatic injury.

In summary, two approaches to qualification of translational safety biomarkers were presented and some of the scientific expectations for these hypothetical projects were delineated. However, the expectations must be aligned and codified in this area, as well as in other areas including nonclinical and clinical data expectations for (translational) qualification of clinical safety biomarkers, biomarker assay validation and performance expectations, expectations around clinical data generation and most importantly the statistical methodology expectations for confirmatory data analysis.

Statistical Considerations for Clinical Safety Biomarkers

Statistical considerations are delineated for biomarker qualification based on the two hypothetical COU examples given (i.e., expanded and limited) for safety biomarkers in the previous section.

The supportive studies for the first expanded COU example include two prospective case/control studies in patients using medications that have the potential to cause pancreatic injury (azathioprine in Crohn's disease patients or mesalazine in ulcerative colitis patients with normal pancreas function). These studies are designed to show greater diagnostic predictivity of the novel biomarkers compared to amylase and lipase with a formal adjudication procedure and a predefined statistical evaluation.

A “learn and confirm” approach is used to ensure that ample learning is completed prior to initiating two long and costly prospective studies. Paramount to designing the prospective studies is a clearly defined COU, leading to clearly defined objectives, so that study results will support specific conclusions which in this case are related to greater diagnostic predictivity of pancreatic injury response as defined by the biomarkers. At the confirmatory stage, with two prospective studies, the biomarkers are already identified and the biomarker measure defined (i.e., dynamic change from baseline instead of single time point concentration), although the COU description does not clearly describe exactly how the two biomarkers will be used together (i.e., as individual biomarkers or a combination).

A statistical evaluation of the two prospective studies is predefined in which the study results must support the defined COU. Clear hypotheses regarding how biomarkers are to be considered for use (relevant null and alternative) must be established. For the first example COU the hypothesis might be:

> Using biomarkers + conventional markers relative to conventional markers alone will improve the sensitivity (and/or specificity) to identify patients treated (or not treated) with medications known to potentially cause pancreatic injury.

In all cases, individual analyses are clearly stated to support each hypothesis. In this example, the lower bound of the 95% confidence interval (CI) on difference > 0 will support greater diagnostic predictivity. However, it is still required to define exactly how to identify patients as having potential injury response. Multiple possibilities exist, including a signal in any one biomarker, a signal in two of three biomarkers, a signal in all biomarkers, or a signal in a measure that combines and reduces three biomarker measures into one CM. Also important is an understanding of what the defined signal is predictive of (e.g., injury, exposure, or perhaps only outside the variation of NHV).

Another critical aspect in identifying appropriate hypotheses and evidentiary standards is determining whether a standard biomarker is a pseudo standard or a true gold standard. A true gold standard, such as histopathology, may be unavailable, too invasive, or too expensive. However, if a true gold standard exists, the new biomarker performance can be readily assessed through standard methods, such as receiver operating characteristic (ROC) analysis, to show “comparability” to the gold standard. However, a pseudo-gold standard is often what is available, and is inadequate by itself (i.e., amylase/lipase in pancreatic injury lacks specificity). Comparing the new biomarker using a pseudo-gold standard as a reference is unlikely to show improvement without accompanying preclinical data. In this case, using exposure to drug as the reference is an option to show performance improvement of the biomarkers.

Type I vs Type II error can be used to assess the risk in biomarker qualification. In this safety context, a Type I error results in qualification of a biomarker that does not predict toxicity. Alternatively, a Type II error fails to qualify a biomarker that does predict toxicity. Which error is worse, depends on the intended use of the biomarker and current standard of practice. For example, if the intended use of a new biomarker is to expand the testing of a new drug when conventional biomarkers alone are considered inadequate or too risky, then a Type I error must be avoided and it must be clear that the new biomarker predicts toxicity in order to ensure patient safety. On the other hand, if the intended use of the biomarker is to improve sensitivity in a context where the conventional biomarkers alone are considered adequate, a Type II error may be more critical, as one would not want to reject new biomarkers that do help predict toxicity in this setting.

Obtaining agreement around the statistical analysis plan requires a predefined statistical evaluation. In addition to what is already described above, other Important considerations for the analysis plan include utilizing appropriate methods to combine data from multiple studies (e.g. pooling, meta-analysis), how to handle missing data (e.g. ignore/remove, last observation carried forward [LOCF], imputation) and specifying important sensitivity analyses. Statistical strategies that may improve the efficiency of the qualification process may also be considered, including adaptive design approaches, and/or techniques such as cross-validation to identify important biomarker subsets when assessing a panel of biomarkers. Both of these strategies attempt to incorporate and de-risk additional learning while assessing data from confirmatory trials.

Two supportive studies were used for the second, more limited, example COU. The first was a longitudinal study in healthy subjects at two visits to define and characterize a CM of the two biomarkers in NHVs. The second study, in patients with known pancreatic injury, was used to show an association of the derived CM with known injury. The COU prescribed utilizing the CM to predict the evidence that a cohort has CM measures that substantially deviate from what would be expected in a cohort of NHVs.

Some potential limitations of the learning-phase data include that it may only confidently be use to predict deviation from NHV, that it may contain multiple time points for exposed patients and limited time points for NHV, that the signal may be much larger using the maximum across all time points instead of using a CM derived at each time point, and that the observed association may not have a causal relationship to injury. Given the limitations of the data, for a single timepoint thresholds can be derived for the CM that suggest deviation from healthy subjects using bootstrap resampling (any test or metric that relies on random sampling with replacement). To derive thresholds for the maximum signal across multiple time points, an extensive modeling and simulation exercise was performed based on some data-driven assumptions.

Within any COU, the right biomarker measure must be established, whether it be raw concentration, normalized concentration, or change from baseline (absolute or fold-change). It is also important to establish normal ranges, which can be estimated using a variety of methods including “robust” (Horn et al., 1998), non-parametric bootstrap, or assumptions of normality.

In summary, defining universal evidentiary standards for safety biomarker qualification is difficult given the significant diversity in potential COUs. Appropriate evidentiary standards will rely on core statistical principles. While some standards may mimic traditional evidentiary standards associated with drug development (e.g. clear hypotheses, analyses, multiplicity, missing data, etc.), some may not (e.g. settings in safety qualification where Type II error may be important, integrating more than one study for final analysis, etc.) Key considerations for defining evidentiary standards for safety biomarkers, beyond statistics, include collaboration (e.g. consortia), regulatory interactions and patience.

Session 3: Evidentiary Considerations for Biomarker-Based Enrichment of Clinical Study Populations to Increase Efficacy or Safety of Drugs

Biomarker-based enrichment of clinical study populations

Biomarkers can be used for enrichment of clinical study populations by measuring the biomarker at screening/baseline with the results defining trial participant eligibility. The biomarker might be measured one time at screening or more times during the trial. Once qualified, a biomarker can be used for development and evaluation of therapeutics interventions, according to the COU. Biomarkers may support clinical trials with a less appropriate endpoint, an endpoint that requires too long a period of time to study, and where the endpoint reflects serious disease progression. The biomarker may or may not ultimately become part of a disease population diagnostic test regimen, or be implemented in medical practice and these possibilities should be considered early in the biomarker qualification process.

The evidence for biomarker qualification may emerge over time from multiple clinical trials. For molecular biomarkers, if appropriate samples have been banked, and the analyte is stable, carefully planned retrospective analyses may speed qualification. The evidence for biomarker qualification may also emerge from positive correlations between the biomarker and the disease process and outcome from a range of studies, the ability to measure the pathological/physiological process based upon advances in the measurement of a biomarker. Qualification can also be achieved, in the case of prognostic biomarker, through increased understanding of the importance of a pathological/physiological condition for a patient subgroup. The key understanding is the relationship between the biomarker and the disease, and its longitudinal progression.

Careful definition of the COU for the specific biomarker is critical and is the foundation of biomarker qualification. In the case of the use of samples/data from already completed trials, determining whether those trials were designed in a manner that supports the specific COU is important. The level of predictive accuracy indicates potential utility and is dependent on the COU. The availability of tools to measure the biomarker and harmonization of the process of biomarker measurement across sites is critical and should be evaluated prior to any samples being tested. Another consideration is whether this biomarker will be used to identify the proper patients for an approved pharmaceutical agent (i.e., companion diagnostic) and this may inform how the assays for measurement of the biomarker are developed. In that case the biomarker will have to undergo its own regulatory approval process.

Neuroimaging enrichment biomarkers for CNS diseases

Hippocampal volume (HV) in Alzheimer's disease (AD) is a case study of a neuroimaging enrichment biomarker. Atrophy of the hippocampus as measured by structural MRI is an early and progressive feature of AD. Neuroimaging changes are key to mediating the memory deficits that represent a cardinal feature of AD. A framework of biomarker development and regulatory qualification begins with the rationale for use of the biomarker and proceeds to assessing the predictive accuracy of the biomarker for an appropriate stage of disease. In the particular case of HV, the rationale is that smaller HVs in patients with mild cognitive impairment (MCI), a pre-de-dementia stage of AD, are associated with more rapid clinical decline and progression to dementia. Clinical trials of MCI patients are in need of enrichment biomarkers for subject selection, because this group is very heterogeneous and cohorts defined on clinical criteria alone show a high degree of variability in the rate of progression. The use of a qualitative evidence map as a checklist for evaluating biological plausibility of the biomarker to linkage to clinical outcomes was discussed. Replication of the predictive accuracy of the biomarker is also important using independent datasets, yet the availability of data poses a significant challenge. An additional important consideration is the standardization of methods for accuracy and precision; for HV measurement, the use of standardized image analysis algorithms and centralized analysis is recommended.

The Coalition Against Major Disease (CAMD) has successfully achieved biomarker qualification of low baseline HV as an enrichment biomarker with EMA. The consortium performed a systematic survey of the published literature which indicated strong evidence for the use of low baseline HV as an enrichment biomarker in MCI trials and found that baseline HV predictive accuracy is consistent multiple different image analysis algorithms. Test-retest reliability is high for HV and a crucial consideration for biomarker qualification. Operational considerations and practical implications for clinical trials were also discussed with decreased cost and reduced sample size as benefits of using an enrichment biomarker.

In summary, key evidentiary questions to be addressed by a putative biomarker include heterogeneity of the clinically-defined target population; strength of supporting data and robustness of findings across different studies, cohorts, and geographies; test-retest reliability of the methodologies; sensitivity of the methodologies to technical variations; and operational considerations including time and cost. HV, currently in the Advice and Consultation phase of qualification with the FDA, provides a case study of a neuroimaging enrichment biomarker for prognostic use. Finally, biomarker qualification has the potential to improve the chances of a successful trial, reduce the number of subjects exposed to an experimental treatment that may have side effects, and reduce both time and cost of trials.

Prognostic Biomarker Qualification: Case Study: Autosomal dominant polycystic kidney disease (ADPKD) and total kidney volume (TKV)

Polycystic kidney disease (PKD) is the 4th leading cause of end-stage renal disease (ESRD) with over three million patients worldwide. There is no specific race or gender affected. Cysts are found in the kidneys, liver, spleen and brain and the disease begins in utero. However, patients with autosomal dominant polycystic kidney disease (ADPKD) are typically asymptomatic for decades and do not have progressive loss of kidney function until they are in their third or fourth decade of life. PKD patients suffer renal complications prior to loss of kidney function with over 50% having at least one complication by age 30. Clinical complications typically also happen in adulthood such that greater than 95% of patients demonstrate renal cysts by ultrasound by age 30, hypertension occurs in 60% of patients by age 30, greater than 50% of patients will have had an episode of gross hematuria by age 40, and proteinuria (low grade) occurs in approximately 25% of patients with important prognostic implications. Importantly, all these characteristics have now been shown to mediate their risk through total kidney volume (TKV).

Increased TKV in PKD is due almost exclusively to cyst burden with a highly variable rate of kidney growth with high inter-individual variability. Increase in TKV precedes a decline in kidney function by decades in PKD, and increases in TKV strongly predict future loss of kidney function in this disease. Therefore, the PKD Outcomes Consortium (PKDOC) was formed by the Polycystic Kidney Disease Foundation and C-Path in order to qualify TKV as a prognostic biomarker for use in clinical trials evaluating patients with ADPKD.

The first research objective of the PKDOC was to determine the predictive accuracy of baseline TKV, baseline estimated glomerular filtration rate (eGFR), baseline age and other prognostic factors (e.g. sex, PKD genotype, race) in estimating the risk of worsening of eGFR and ESRD, in order to support the regulatory qualification of TKV as a prognostic biomarker for use in clinical trials. The second research objective of the PKDOC was to develop a joint model to simultaneously assess longitudinal TKV measurements and the probability of disease outcomes, and use the joint model as a DDT for trial enrichment strategies.

A decision tree was developed for using TKV as a prognostic biomarker for use in clinical trials. Use of TKV to select appropriate patients for clinical trials is expected to positively impact clinical therapeutics development by decreasing the number of patients needed to test medications, shortening the duration of the study by measuring an outcome that is easily and accurately measured, reducing clinical trial costs, reducing exposure to potential drug toxicities, and improving the success rate of clinical drug development.

The HALT PKD Study illustrates the value of image stratification of ADPKD. Class severity was shown to associate with greater rates of TKV increase and eGFR decline. Changes in TKV and eGFR were shown to be negatively correlated. The treatment effect of low BP increases with class severity. In the patients with the most severe disease (class D-E), low BP was shown to associate with slower eGFR decline after month four and overall. Restriction of enrollment to class D-E patients would have detected a stronger low BP effect on TKV growth and eGFR decline, with a much lower number of patients (187 versus 551). These results stress the importance of optimal patient selection to reduce the cost and the chance of a Type II error. Using this information and TKV as a prognostic biomarker, interventional trials can be designed focusing on those patients most likely to progress to renal failure.

Statistical Considerations for Biomarker Qualification for Biomarker-Based Enrichment in Clinical Studies

Enrichment biomarkers can be used for diagnosis or to inform the definition of inclusion/ exclusion criteria for a clinical trial. The latter prognostic application may be used to separate groups or to enrich an already diagnosed population. Distinct from these applications, predictive biomarkers are used to forecast a treatment effect.

Statistical principles applied to use of enrichment biomarkers must include sources of variation, misclassification, sensitivity, specificity, predictive value and disease prevalence. Sources of variation include within patient variability (day to day), measurement error associated with instruments, calibrations, reading or administration errors, the experience level of the person taking measurements, and subject experience with measurement (learning effects). In addition, between-subject variability (covariates) is important to consider. But between-subject sources of variation can be reduced.

Ignoring important covariates and random error result in misclassification. The misclassification rate depends on disease prevalence. Few tests are inherently dichotomous. Continuous traits are used to categorize individuals. This may result in substantial variation of the same diagnostic test in different populations. This variation also depends on measurement error. The misclassification rate depends on the ratio of between to within patient variability and prevalence. In early AD, within-patient variability is larger, resulting in more misclassification.

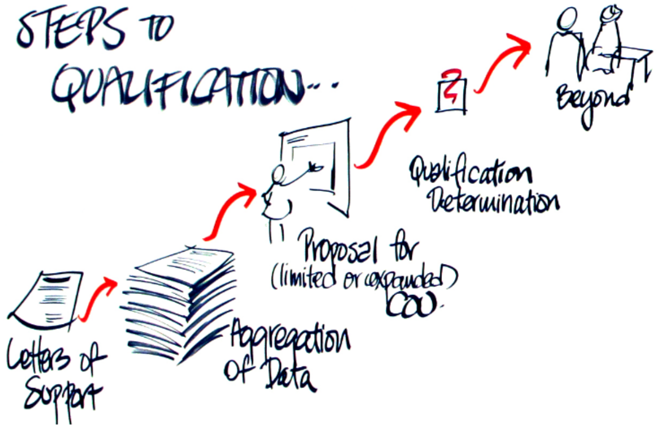
Sensitivity, specificity and predictive value must be calculated against a “gold standard”. In prodromal AD, the “gold standard” is future diagnosis with AD. Other standards include amyloid imaging, future clinical decline, and post-mortem plaque load. The level of evidence required depends on assessment of risks and benefits. The predictive value of a test varies with prevalence.

In summary, biomarker qualification requires estimation of and reduction in sources of variability. Composites, repeated measurements and covariates may reduce variability.

Prevalence must be a consideration in the process. Biomarker validation depends on the risk/benefit of classification within the specified COU.

Session 4: Round Table Discussion (CERSI/FDA)

The final discussion of the day proved to be the most informative around the direction that evidentiary considerations would proceed in the near term. In that session, a coherent, strategic approach to delineating the steps towards obtaining biomarker qualification was defined and captured by the graphic illustrator. The first early and preliminary outcome along the process of clinical biomarker qualification is an optional Letter of Support (LOS), which as described above, indicates the potential value of a biomarker and encourages use of the biomarker. The next step is aggregation of additional data needed for qualification with specific questions concerning what data are needed and how data should be aggregated. In addition, at this step, the need for data sharing, the necessary data quality, standards and reproducibility are important considerations. A logical next outcome is to seek qualification with either a limited or expanded COU. Here the considerations are around the assay and the statistical analysis plan including assay methods and performance characteristics, sample handling and analyte stability. Qualification determination based on the submission is the next step in the process, with the risk/benefit associated with use of the biomarker driving the discussion. Considerations beyond qualification include whether or not the biomarker has implications for clinical practice or drug labeling and how to leverage data from INDs to aid in development of the biomarker.


STEPS TO QUALIFICATION...
Letters of Support
Aggregation of Data
Proposal for (limited or expanded) COU.
Qualification Determination
Beyond

Beyond the basic strategy around biomarker qualification, filling other gaps that would better enable the qualification process were discussed. These were quickly defined as “enablers of biomarker development” and include data standards, data quality, data reproducibility, statistical considerations, considerations for assay validation including imaging applicable across multiple methodologies (if necessary), and establishing cut points for biomarker implementation. Of course, a major unanswered question is how to disseminate current and best thinking around these gaps. To this end, there was a discussion around the need to conduct future workshops on these topics.

The round table discussion concluded with an overview of these upcoming meetings designed to continue the discussion started at this workshop. In October, 2015, the Brookings Institute will host a meeting where the goal is to inform ongoing policy efforts, both within FDA and the broader scientific and policy communities that are seeking to improve the development, qualification, and use of biomarkers in drug development. The main objectives of the upcoming meeting will be (1) to discuss the common lexicon for biomarker development that is currently being developed by FDA and NIH, (2) to use case studies to explore biomarker characteristics (including COU) that can inform whether and under what circumstances a biomarker should be targeted for qualification, and (3) to develop an initial set of strategies that can help to ensure better cross-sector collaboration and communication in the area of biomarker development and qualification, including strategies related to standardization, aggregation, and dissemination of biomarker data. This small and invitation-only workshop is being convened under a cooperative agreement with FDA, and will include representatives from across the FDA, as well as from NIH, academia, industry, professional associations, and patient and disease advocacy groups. Furthermore, the FNIH is considering hosting a series of workshops focused on specific case examples proposed for qualification and their associated evidentiary considerations. Initial discussions indicate that their first workshop will be in April of 2016. During the closing session, several statisticians from industry, academia and FDA agreed to form a small working team to map key elements of the COU cases discussed against critical statistical considerations. It is expected that this group can report out at one of the upcoming workshops.

There has been much learning in the eight years since the Biomarker Qualification Program was initiated by the FDA. Now is the time for the stakeholders in industry, academia and regulatory agencies to come together to better define the evidentiary standards necessary for biomarker qualification based on this wealth of experience.

References

Amur S, LaVange L, Zineh I, Buckman-Garner S, Woodcock J. **2015**. Biomarker Qualification: Toward a Multiple Stakeholder Framework for Biomarker Development, Regulatory Acceptance, and Utilization. *Clin Pharmacol Ther* 98(1):34-46.

Food and Drug Administration. **2011**. Guidance for Industry: Use of histology in biomarker qualification studies. Available at: http://www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm285297.pdf Accessed on 5 October 2015.

Food and Drug Administration. **2012**. Guidance for Industry: Enrichment strategies for clinical trials to support approval of human drug and biological products. Available at: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM332181.pdf Accessed on 5 October 2015.

Food and Drug Administration. **2014**. Guidance for Industry and FDA Staff: Qualification Process for Drug Development Tools. Available at: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM230597.pdf Accessed on 5 October 2015.

Food and Drug Administration. **2015a**. Critical Path Innovation Meetings: Guidance for Industry. Available at:

http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM417627.pdf Accessed on 5 October 2015.

Food and Drug Administration. **2015b**.FDA survey to identify potential biomarkers for qualification. Available at: http://www.fda.gov/Drugs/DevelopmentApprovalProcess/DrugDevelopmentToolsQualificationProgram/ucm459144.htm Accessed on 5 October 2015.

Food and Drug Administration. **2015c**. Identifying Potential Biomarkers for Qualification and Describing Contexts of Use to Address Areas Important to Drug Development. Available at: https://www.federalregister.gov/articles/2015/02/13/2015-02976/identifying-potential-biomarkers-for-qualification-and-describing-contexts-of-use-to-address-areas Accessed on 5 October 2015.

Horn PS, Pesce AJ, Copeland BE. (1998) A robust approach to reference interval estimation and evaluation. *Clin Chem*. 44(3):622-31.

Institute of Medicine. 2010. Evaluation of Biomarkers and Surrogate Endpoints in Chronic Disease. Available at: http://iom.nationalacademies.org/Reports/2010/Evaluation-of-Biomarkers-and-Surrogate-Endpoints-in-Chronic-Disease.aspx Accessed on 5 October 2015.