








Perspectives on statistical strategies for the regulatory biomarker qualification process

Suzanne B Hendrix^{*1} , Robin Mogg², Sue J Wang³ , Aloka Chakravarty⁴ , Klaus Romero⁵ , Samuel P Dickson¹ , John-Michael Sauer⁵  & Lisa M McShane⁶ 

¹Pentara Corporation, UT 84109, USA

²Bill & Melinda Gates Medical Research Institute, MA 02139, USA

³Office of Biostatistics, Office of Translational Sciences, Center for Drug Evaluation and Research, US Food & Drug Administration, MD 20993, USA

⁴Office of the Commissioner, US Food & Drug Administration, MD 20993, USA

⁵Translational and Safety Sciences Program, Critical Path Institute, AZ 85718, USA

⁶Biometric Research Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute, National Institutes of Health, MD 20892, USA

*Author for correspondence: shendrix@pentara.com

Qualification of a biomarker for use in a medical product development program requires a statistical strategy that aligns available evidence with the proposed context of use (COU), identifies any data gaps to be filled and plans any additional research required to support the qualification. Accumulating, interpreting and analyzing available data is outlined, step-by-step, illustrated by a qualified enrichment biomarker example and a safety biomarker in the process of qualification. The detailed steps aid requestors seeking qualification of biomarkers, allowing them to organize the available evidence and identify potential gaps. This provides a statistical perspective for assessing evidence that parallels clinical considerations and is intended to guide the overall evaluation of evidentiary criteria to support a specific biomarker COU.

First draft submitted: 19 August 2020; Accepted for publication: 19 March 2021; Published online: 26 May 2021

Keywords: analysis plan • context of use • diagnosis • enrichment • prognosis • qualification • risk/benefit • safety • statistical • validation

A biomarker is a defined characteristic that is measured as an indicator of normal biological or pathogenic processes, or biological response to an exposure or intervention, including therapeutic interventions [1]. Qualification is a conclusion, based on a formal regulatory process evaluating the scientific merit of the evidence provided, that a biomarker may be used by any person in drug, or medical product development for the qualified context of use (COU) [2]. An FDA guidance document provides detailed explanation of the qualification evidentiary framework [3] that is recommended as prerequisite reading for those unfamiliar with the concept. The evidentiary process for qualification has been outlined as a five-component process (Figure 1) [4]. Inherent in that process is accumulation, aggregation, summarization, scientific interpretation and statistical analysis of data to establish whether evidentiary criteria are met to support the utilization of a specific biomarker in a particular COU to facilitate medical product development. The focus of this paper is to provide further elaboration from a statistical perspective on the last component of this process, evidentiary criteria. Insights gained or limitations of data determined by statistical evaluation may feed back into the first four components, (need statement, COU, benefit, risk) potentially leading to refinement of the COU, deeper understanding of risk and benefits associated with use of the biomarker, or modification of the biomarker.

The August 2015 Center of Excellence in Regulatory Science and Innovation (CERSI) meeting ‘Evidentiary Considerations for Biomarker Qualification’ recognized the integral role of statistical evaluation in the biomarker qualification process. Therefore, a working group, Biomarker Qualification Statistics Working Group (BQSWG) was formed to focus on statistical principles and approaches relevant to biomarker qualification efforts [5]. Members of the working group include representatives from the US FDA; the NIH; the pharmaceutical industry; and Critical Path Institute’s (C-Path) Predictive Safety Testing Consortium (PSTC) and Critical Path for Alzheimer’s

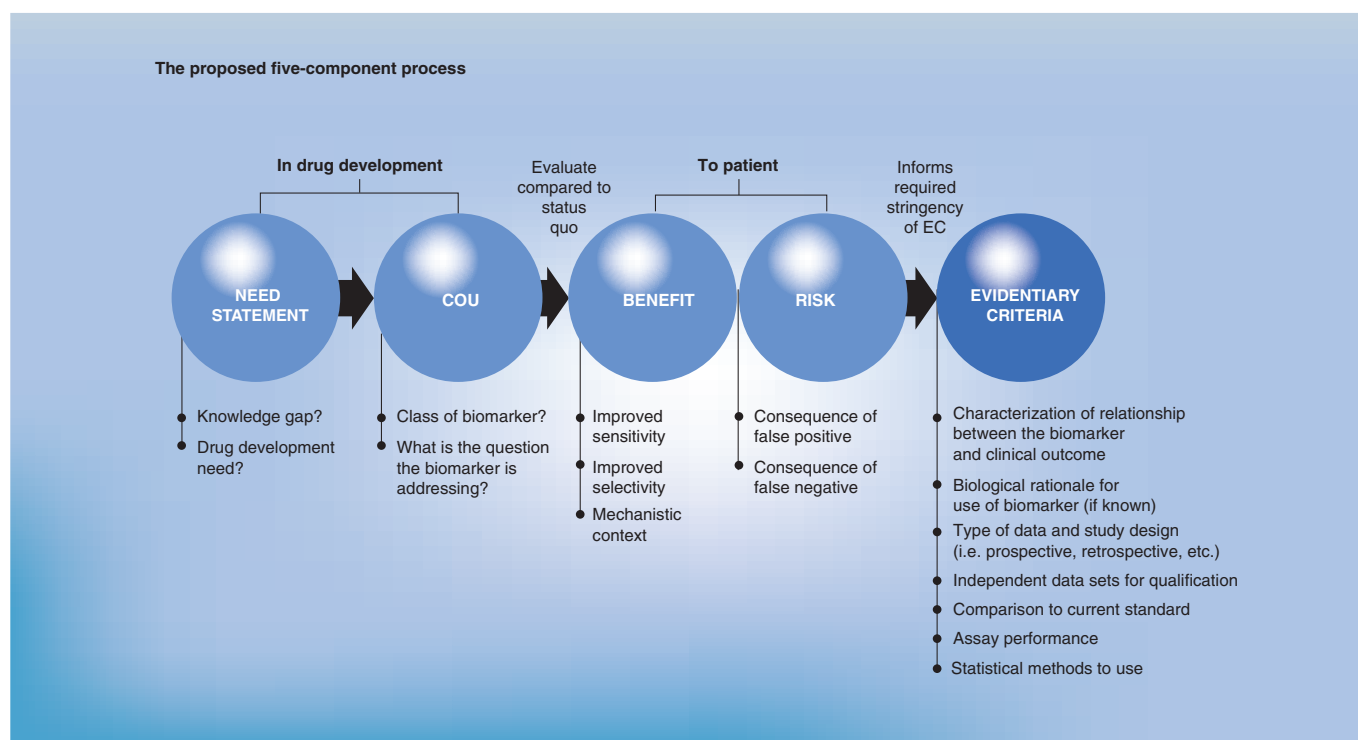


Figure 1. The proposed five-component biomarker qualification process.
 Reproduced with permission from [4].

Disease (CPAD) Consortium (formerly known as the Coalition Against Major Diseases). The goal of this working group was to develop a statistical ‘strategy’ that biomarker qualification teams could reference when defining design elements and considering analysis approaches to support qualification of safety biomarkers or prognostic enrichment biomarkers for specific COUs (draft guidance: ‘Enrichment Strategies for Clinical Trials to Support Approval of Human Drugs and Biological Products’). This statistical strategy is intended to serve as an adjunct to the broader biomarker qualification process and evidentiary criteria assessment map described therein. This paper does not address the validation of analytical procedures that ensure that the measurement method is reliable and accurate. This paper presents a statistical strategy proposal resulting from deliberations of the BQSWG, consideration of public discussion that took place at the FNIH-FDA Biomarker Qualification Workshop: Framework for Defining Evidentiary Criteria held in April 2016, as well as informal comments received from regulatory scientists. This paper does not constitute an official consensus or guidance document.

The goal of the BQSWG was to describe a conceptual framework and strategy for development and evaluation of evidence to support aspects of biomarker qualification from a statistical perspective. The strategy represents a collection of steps that may be required for qualification and maps statistical methodologies that may be useful in the evaluation process to these steps. The statistical methodologies serve as examples; they are by no means exhaustive and alternate statistical methods or approaches could be proposed as the current thinking changes in this quickly evolving field. Further, it should be understood that statistical analyses alone are insufficient for a biomarker to achieve qualified status. The qualification process brings to bear medical, biological, analytical and scientific understanding and evidence. Statistical analyses are then applied to summarize and evaluate the strength of the evidence. Biomarker qualification is ultimately based on an overall assessment of risk versus benefit of utilization of a particular biomarker for a given COU in medical product development.

Two examples are used throughout the paper to illustrate key statistical concepts and steps in the qualification evidentiary process. Details of each example are provided in Supplementary Table 1. The first example is the qualification of total kidney volume (TKV) as a prognostic biomarker for clinical trial enrichment in autosomal dominant polycystic kidney disease (ADPKD). As an early indicator of ADPKD, increased TKV generally presents earlier than other signs and symptoms of renal failure, making it an attractive prognostic biomarker for clinical trial

enrichment. TKV was successfully qualified with FDA and the European Medicines Agency (EMA) by C-Path's Polycystic Kidney Disease Outcomes Consortium (PKDOC) as a prognostic biomarker for the clinically relevant end point represented by 30% decline in estimated glomerular filtration rate (eGFR) and will be referred to as the 'enrichment' biomarker example in this article [67–11]. The second example is an ongoing effort that is sponsored by the Foundations for the National Institutes of Health Biomarker Consortium (FNIH BC)/PSTC Kidney Safety Project with the goal of qualifying a panel of urine biomarkers for early detection of acute kidney injury resulting from exposure to an investigational nephrotoxicant agent in healthy participants in early phase medical product development studies and will be referred to as the 'safety' biomarker example in this article. The safety biomarker example has not been formally reviewed or endorsed by FDA or EMA; therefore the views and descriptions herein are considered from a requestor's perspective and have been presented publicly in various forums, including CERSI [5] and a C-Path webinar series [12]. A subset of the urine biomarkers in the safety example has been recently qualified by the FDA for monitoring safety of patient cohorts under a more limited COU that is not described herein [13].

Each biomarker qualification project may have unique aspects, but it is hoped that the principles and approaches illustrated through these examples provide a useful introduction and starting point for planning future biomarker qualification efforts.

Methods

The statistical approaches appropriate for addressing the key elements of the strategy will differ depending on the type of biomarker measurement(s) and the specific COU. For purposes of discussion, the biomarker qualification process is presented here in a linear fashion; however, the process may require iteration through some of the described steps to appropriately align evidence and refine the intended COU.

The overall statistical strategy is proposed as comprising the following steps:

- Step 1: Consider the drug or medical product development need and initial COU statement for the biomarker to formulate an overall strategy for data accession and statistical evaluation;
- Step 2: Collate existing evidence relating the biomarker to the key clinical outcome(s) in order to assess what information relevant to the intended COU is currently readily available and identify gaps;
- Step 3: Identify any additional data or specimens available for analysis that can be used to build on existing evidence and determine what prospective data may need to be generated to fill identified gaps;
- Step 4: Determine appropriate strategies for analyzing existing and newly generated data to support the intended COU by evaluation of the biomarker–clinical outcome relationship;
- Step 5: Statistically quantify risks and benefits associated with using the biomarker for the intended COU based on the totality of evidence, demonstrating the practical relevance of the biomarker;
- Step 6: Finalize the COU statement and summarize the supporting evidence.

A biomarker cannot be qualified without a reliable and accurate means of measurement. Implicit in the steps outlined above is that there is at least one method available to measure the biomarker. Qualification does not limit measurement of a biomarker to the assay specified in the qualification effort provided the new method has demonstrated similar performance characteristics [14]. As evidence is gathered during the qualification effort, the impact of the proposed biomarker measurement methods and their performance must be considered and may be a factor in the final qualification statement. Inability to identify assays with sufficiently good analytical performance in measuring a biomarker could preclude qualification of the biomarker [15]. Even for a qualified biomarker, there could be specific assays deemed to have analytical performance insufficient to support their use under the qualification context of use. As a qualification effort proceeds, it is important to continually identify factors that might influence biomarker performance so that these factors can be documented and variation in performance can be appropriately quantified.

Biomarker measurement methods supporting the final qualification should be analytically validated, including assessment of reliability, but detailed discussion of these requirements is beyond the scope of this discussion, which focuses on overall statistical strategy. Rather, researchers are directed to C-Path's 'Points to Consider Document: Scientific and Regulatory Considerations for the Analytical Validation of Assays Used in the Qualification of Biomarkers in Biological Matrices' for a comprehensive description of the analytical validation requirements pertinent to biomarker qualification [16].

Table 1. Diagram of a context of use statement with eight essential elements identified for the total kidney volume example.

COU statement for TKV as a prognostic enrichment biomarker (original statement text is split across the rows)	Step 1 element
This guidance provides qualification recommendations for the use of TKV, measured at baseline, as a prognostic enrichment biomarker...	Element 1: role of the biomarker
... defined as a confirmed 30% decline in the patient's estimated glomerular filtration rate (eGFR)	Element 5: specification of the outcome of interest
Baseline TKV can be used in combination with the patient's age and baseline eGFR...	Element 3: participant characteristics that may affect the biomarker/outcome relationship
... as an enrichment factor in ADPKD clinical trials to select ADPKD patients at high risk for a progressive decline in renal function	Element 4: development context for the drug or other medicinal products Element 7: decisions and actions based on the biomarker
Patients with ADPKD should be at least 12 years of age	Element 2: population
Various imaging modalities and post-processing methods are available to determine TKV. These modalities have different levels of precision. For patients with ADPKD at high risk for a confirmed 30% decline in their eGFR, TKV was qualified based on a collection of data from multiple study sites as well as on results from imaging modalities (i.e., magnetic resonance imaging [MRI], computed tomography [CT] or ultrasound [US]) and from analysis methodologies (i.e., stereology and ellipsoid calculations). TKV should be calculated from the left and right kidneys measured with a validated and standardized image acquisition and analysis protocol within the trial	Element 6: the measurement method and specific quantification of the biomarker (including timing of the biomarker measurement)
Proposed thresholds for decision making were included in the submitter's application, but not included in the FDA's TKV qualification of biomarker guidance	Element 8: thresholds for decision making on the biomarker
Data taken from [6].	

Step 1: Consider the drug or medical product development need & initial COU

Central to qualification of a biomarker(s) (e.g., a single biomarker or a panel comprising two or more biomarkers) is a COU statement indicating the medical product development use for the biomarker. An initial COU statement is typically formulated based on the characteristics of existing evidence, but it may undergo refinement as new data are generated or when extrapolations are supported by other scientific evidence and statistical analyses. Elements important in defining the COU, which drive what data are needed to support a qualification effort include: role of the biomarker in medical product development; targeted patient or participant population (e.g., healthy individuals; individuals with a disease or other medical condition or individuals receiving treatment for an existing medical condition; adults or children; etc.); patient or participant characteristics that may affect the relationship between the biomarker and the outcome (e.g., disease subtype; stage or severity; age or sex; risk of developing a disease); development context for the drug or other medical products; specification of the outcome of interest; specific quantification of the biomarker (including any processing to be performed on the raw measurements); decisions and actions based on the biomarker; and thresholds for decision making on the biomarker. Table 1 illustrates these eight elements for the COU of the TKV example. Each of these elements will be discussed in further detail below.

Role of the biomarker

Biomarkers can have several possible roles in drug or medical product development. The BEST glossary [1] defines the most common of these roles (see Table 2). Some biomarkers could potentially satisfy additional criteria to qualify as a validated surrogate clinical trial end point, but the explanation of criteria for surrogacy is beyond the scope of this discussion and readers are referred elsewhere [17]. Other aspects of the biomarker's role include whether it will be used in isolation or in combination with other biomarkers or variables. The biomarker may be used as: a standalone in a setting where no conventional biomarker or accepted clinical variable exists, a standalone replacement biomarker in a situation where conventional biomarkers are not considered adequate or an add-on to conventional biomarkers or variables in settings where conventional biomarkers are considered acceptable but not optimal. Details on the role of the biomarker within the two examples are provided in Supplementary Table 2.

Population

The population in which the biomarker will be measured and applied to support a medical product development program should be clearly specified in the COU. Evidence supporting the intended use should include data gathered in that same population, if possible, or a reasonably similar or generalizable population.

Table 2. Roles for biomarkers in medical product development from BEST glossary.

Role	BEST glossary definition
Susceptibility/risk	A biomarker that indicates the potential for developing a disease or medical condition in an individual who does not currently have clinically apparent disease or the medical condition
Diagnostic	A biomarker used to detect or confirm presence of a disease or condition of interest or to identify individuals with a subtype of the disease
Monitoring	A biomarker measured repeatedly for assessing status of a disease or medical condition or for evidence of exposure to (or effect of) a medical product or an environmental agent
Prognostic	A biomarker used to identify likelihood of a clinical event, disease recurrence or progression in patients who have the disease or medical condition of interest
Predictive	A biomarker used to identify individuals who are more likely than similar individuals without the biomarker to experience a favorable or unfavorable effect from exposure to a medical product or an environmental agent
Pharmacodynamic/response	A biomarker used to show that a biological response has occurred in an individual who has been exposed to a medical product or an environmental agent
Safety	A biomarker measured before or after an exposure to a medical product or an environmental agent to indicate the likelihood, presence or extent of toxicity as an adverse effect

The patient population could be defined in a variety of ways depending on the specific role of the biomarker. For a safety biomarker, such as the kidney toxicity example, the patient population might include healthy participants or patients diagnosed with a disease for whom safety must be closely monitored due to participation in a clinical trial. The population of interest for a prognostic biomarker such as the TKV example might be patients who have been diagnosed with a specific disease and for whom participation in a clinical trial of a new therapy is being considered. The prognostic biomarker could be used to enrich the clinical trial population for patients most likely to benefit from the clinical trial intervention or most likely to progress in disease severity during the course of the clinical trial resulting in increased statistical power for the trial's treatment comparison or decreased timeframe for the trial.

Additional biomarkers with different roles have different population definitions. The population of relevance for a diagnostic biomarker might be individuals who are suspected of having a particular disease status or subtype. The population of relevance for a treatment response biomarker may be clinical trial participants receiving different interventions for the same disease. Details on the role of the biomarker within the two examples are provided in Supplementary Table 3.

Patient or participant covariates

Patient characteristics such as age, sex, race or co-existing medical conditions (e.g., immune disorders, obesity, diabetes, cancer) or concurrent treatments for other medical conditions may have an impact on the relationship between the biomarker and the clinical outcome of interest or on the analytical performance of the biomarker measurement method (e.g., imaging-based biomarkers might be more difficult to measure in children due to smaller anatomical features). Covariates should be collected to the extent possible in order to investigate their potential modifying effects on biomarker utility. Details on participant characteristics that may affect the biomarker – outcome relationship in the context of the two examples are provided in Supplementary Table 4.

Drug or other medical product development context

For biomarkers being qualified for use as safety biomarkers or prognostic or predictive enrichment biomarkers for development of a new drug, the drug's mechanism of action may be an additional critical factor in defining the biomarker's COU. Every drug works through one or more biological pathways to have an impact on a disease process, but could potentially also cause an unintended adverse effect on a patient. A particular biomarker may be relevant to only one or a few of these pathways. The biomarker is unlikely to be very informative for clinical benefit or adverse outcome related to a drug if it does not lie in one of the pathways relevant to the drug's mechanism of action unless it is a safety biomarker, which could be relevant to the off-target effects of many different mechanisms. Therefore, extrapolation from one class of drug or other medical product for which a biomarker has been qualified to a different class would require supporting evidence. Details on the development context for the two illustrative examples are provided in Supplementary Table 5.

Table 3. Statistical approaches for quantification of single or multiple biomarkers.

Example quantifications of a single biomarker		
Type of calculation	Examples of specific calculations	
Raw measurement	None	
Summary statistic over repeated measurements	Minimum or maximum Time until maximum or minimum Rate of change over time (slope) Area under the curve (AUC)	
Normalized or standardized measurement	Relative to laboratory variables or other biomarkers Relative to amount of biospecimen collected (urine/blood/cerebral spinal fluid/stools) Relative to age, education, sex or other clinical and demographic characteristics	
Measurement adjusted for baseline value	Relative (fold) change from baseline (post/pre) Absolute change from baseline (post-pre) Change from baseline ((post-pre)/pre) (%) Absolute measurement or change corrected for a baseline covariate	
Examples of statistical methods for developing quantifications involving combinations of multiple biomarkers or biomarkers with clinical variables		
Outcome type	Biomarker measure type	
Categorical	At least one quantitative	<ul style="list-style-type: none"> • Receiver operating characteristic (ROC) analyses based on scores from regression models; • Other classification methods (including deep learning or other machine learning)
Quantitative (including time to event)	All quantitative	<ul style="list-style-type: none"> • Regression (linear, logistic, ridge, LASSO, partial least squares, mixed effects); • Factor-based methods (principal components, tree-based modeling, other machine-learning approaches); • Cox regression (for time to event)
Categorical or quantitative	Mixed categorical and quantitative or all categorical	<ul style="list-style-type: none"> • Decision trees (which may include confirming one biomarker result with another); • Other <i>ad hoc</i> empirical approaches; • Deep-learning or other machine-learning methods

Specification of outcome of interest

The usefulness of a biomarker in medical product development usually relies on its association with one or more clinical outcomes (or ‘states’). Careful description of the clinical outcome and justification for the way it is measured are important components of the qualification process. Considering the example of a prognostic biomarker, clinical outcomes of interest might include disease-related events such as disease recurrence or progression, or symptom resolution. Example assessment methods include tumor progression measured by CT scan or cognitive ability as assessed by the mini mental state examination (MMSE). Additional important aspects might include time window of observation (e.g., organ failure within 30 days), definition of time zero for time-to-event outcomes, specific events included in composite outcomes (e.g., tumor progression or death for progression-free survival) and frequency of assessments. Dichotomous, ordinal and continuous outcomes can be useful and may be considered as appropriate outcomes.

Some outcomes are not straightforward to assess and may rely on a combination of clinical observations and subjective assessments; indirect measures of outcomes or states may be needed when direct assessment is not ethical or feasible. For example, definitive assessment of some neurological conditions would require brain biopsies or might only be possible upon postmortem examination. Due to the possibility that the relationship between the biomarker and clinical outcome of interest may be influenced by exactly how the outcome is defined and assessed, it is important that all of these aspects are carefully considered in qualification efforts. Details on specifying the outcome of interest for the two illustrative examples are provided in Supplementary Table 6.

Measurement & quantification of the biomarker

A biomarker may potentially be measured or quantified in multiple ways or using various measurement techniques (Table 3A). For purposes of discussion, the biomarker quantification can be thought of as the key quantity relevant to the specified context of use. It is important to appreciate that the nature and strength of an association between a biomarker and a clinical outcome may be influenced by many aspects of the measurement approach, including

on what specimen or entity it is measured, by what method it is measured, timing of measurements, and how the measurement is quantified. Although initially it may be helpful to conceptualize the biomarker independently of the method of measurement, subsequent analyses must establish to what extent the measurement method is important for the reliable performance of the biomarker in the intended COU (see Step 2) [15].

A biomarker could be quantified in various ways. For example, a single biomarker may be expressed as an amount, concentration or an intensity. Repeated measures for a biomarker may be summarized across multiple time points or spatial locations, such as by using a change from baseline, slope or gradient, or by reducing to the minimum or maximum observed value in a defined time period or space. Raw measurements for biomarkers may be processed further (e.g., normalization or standardization using internal controls or external calibrators). Multiple measurements of the same or different biomarkers may be combined, or biomarkers may be combined with clinical variables, often referred to as covariates. The combination might produce a signature quantified by a composite score or classifier (Table 3B). Score or classifier development may require statistical parameter estimation to optimize performance, for example if techniques such as regression modeling are used. Machine-learning methods (including deep learning) could also be used to develop a classifier.

Some statistical methods that are helpful for deriving a quantification of a biomarker when it is a composite of different measures are shown in Table 3B. Examples include regression (e.g., linear, logistic, ridge, LASSO, principal components regression, partial least squares); receiver operating characteristic (ROC) curve analyses or other classification tools based on linear combinations of biomarker values; and factor-based methods (e.g., principal components analysis, tree-based modeling, and other machine-learning approaches). Other *ad hoc* empirical approaches may be considered to combine biomarkers of different measurement types (e.g., categorical vs continuous measurements) or on different scales. Many of these methods assess the composite biomarker's association with the clinical outcome at the same time that they develop the biomarker's quantification.

Some methods for biomarker quantification development explicitly attempt to identify the optimal combination of measurements in the context of the prediction of the outcome of interest. Details on the specific quantification of the biomarker for the two illustrative examples are provided in Supplementary Table 7.

Decisions & actions based on the biomarker

A qualified biomarker may be used to inform decisions during drug or medical product development, and those decisions may trigger certain actions. Every action has associated risks and benefits that could be experienced at an individual patient level and/or a population level (e.g., dose cohort). A biomarker that identifies patients thought most likely to respond to an experimental therapeutic agent (positive predictive) or least likely to respond (negative predictive) would be useful in the development process to define eligibility for clinical trials to appropriately balance benefits and risks for trial participants and most efficiently identify efficacy, if any. A safety biomarker that could provide early warning of an impending adverse effect of a drug or medical product on a clinical trial participant could avert serious or irreversible harms, for example by signaling that an action such as treatment discontinuation or dose reduction is needed to avoid or reduce the risk of organ damage. Details on the decisions and actions based on the biomarker for the two illustrative examples are provided in Supplementary Table 8.

Decision thresholds

Risks, benefits and biomarker values might each lie on a continuum. The balance of risks and benefits deemed acceptable may be context dependent, and thresholds on a biomarker used in medical product development may be adjustable to achieve the right balance. Typically, greater risks are accepted in studies of new therapies for patients with very advanced or life-threatening diseases in hope that urgently needed greater benefits might be realized. In such situations, thresholds of biomarkers potentially indicative of disease subsets most benefiting from the therapy might be relaxed to gather more evidence of the range of efficacy, whereas thresholds for biomarkers indicative of adverse effects might be set to tolerate some nonsevere or reversible adverse effects. These risk–benefit considerations are critical to specification of the qualification COU. Details on the decision thresholds in the context of the two illustrative examples are provided in Supplementary Table 9.

Step 2: Collate existing evidence (published & unpublished) relating the biomarker to the key clinical outcome(s) in order to assess what information relevant to the intended COU, is readily available & where there are gaps

For many candidate biomarkers, a large and often confusing body of evidence will have already been accumulated relating the biomarker to various biological characteristics or clinical outcomes. Biomarker measurement methods, patient populations, disease spectrum, treatment setting (e.g., drugs or other therapeutic interventions) and end points examined may vary substantially among existing studies. It is essential to carefully review this existing evidence to determine the level of comparability across these many characteristics, particularly to assess how variations in these characteristics across studies might affect key associations between the biomarker and clinical outcomes of interest.

A literature search or formal systematic review conducted according to a protocol may be helpful as a framework for summarizing a large body of literature. At a minimum, it is helpful to compile a table of the key study characteristics as described above, relevant study results, and important notes for each study to facilitate integration of evidence. A meta-analysis to produce summary estimates of measures of a biomarker's performance or its association with an outcome might be possible in some situations where the integrated studies are sufficiently comparable. The assembled information will serve as a foundation to guide additional data collection and analysis efforts (Step 3).

There are many important considerations in assessing the credibility and relevance of prior claims about a biomarker and the usefulness of existing data for qualification goals. Claims regarding presence or absence of associations between the biomarker and clinical outcomes made in prior studies should be critically examined to determine whether those claims are supported by appropriate statistical analyses, with particular attention to the possibility of false-positive or -negative findings. False-positive findings are readily generated from extensive data exploration that does not account for multiple testing. In contrast, small studies not designed to have sufficient statistical power to detect differences in biomarker values between groups or associations between the biomarker and the clinical end point(s) of interest, may result in statistically nonsignificant findings that might be inappropriately interpreted as negative [18]. A claim of a null effect (e.g., no difference, no association) requires statistical demonstration that the effect size is smaller than some meaningful value with high certainty, and this typically requires studies with large sample size such as those establishing bioequivalence of two drugs. Observational data need particular scrutiny to identify possible embedded biases. At the culmination of this evaluation process, the goal is to have assembled a useful summary of existing evidence from the literature and unpublished studies.

After review of the existing evidence for TKV, it was determined that literature evidence and existing data, when considered in totality, provided a sufficient evidence base for analysis to firmly establish the relationship between increased TKV and faster progression in ADPKD and to meet evidentiary criteria to qualify TKV as an enrichment biomarker. The literature review and assessment of available datasets for the kidney toxicity example indicated that prospectively collected data was required to support qualification. Details on the collation and evaluation of existing evidence relating the biomarker to key clinical outcomes in the context of the two illustrative examples are provided in Supplementary Table 10.

Step 3: Identify data or specimens available for analysis that can be used to build on existing evidence & determine what additional data may need to be generated to fill identified gaps

Initial review of existing data often uncovers gaps in the evidence needed to support an intended COU statement. Sometimes additional evidence can be obtained through access to raw, unprocessed data from prior studies that were not originally intended to assess biomarker associations. With full access to raw data, it may be possible to select data from specifically defined subsets of participants consistent with a preliminary COU and to reprocess the data in a uniform way to facilitate pooled data analyses across existing data sets or to apply alternate algorithms for comparison. For biospecimen-based laboratory biomarkers, availability of biospecimens may permit new analyses of biomarkers using alternative assay methods. New biomarker data generated from biospecimens may be used to evaluate the impact of particular assay methods on measured biomarker values and the strength of their associations with clinical outcomes.

Sometimes the only way to acquire the necessary data to support a biomarker for qualification is to launch a new, prospective clinical study. Prospective studies are usually time and resource intensive, so if they are necessary, careful planning is essential to ensure they will yield the proper evidence to permit definitive evaluation of the

Table 4. Examples of methods for initially establishing a relationship and later validating that relationship between a specified biomarker and outcome of interest.

Outcome type	Biomarker type	Example analysis method
Quantitative	Quantitative	Linear, nonlinear or nonparametric regression methods
Quantitative	Categorical	T-test, ANOVA, or analogous nonparametric tests
Categorical (>2 categories)	Quantitative	Multinomial logistic regression Ordinal logistic regression
Categorical (binary)	Quantitative	Logistic regression ROC curve
Categorical	Categorical	Fisher's exact or chi-squared test
Time to event	Quantitative	Cox regression or parametric survival analysis modelling
Time to event	Categorical	Log rank test/Kaplan–Meier curves

qualification COU statement (see Step 4). Details on identifying available data or specimens to fill identified gaps in the context of the two illustrative examples are provided in Supplementary Table 11.

Step 4: Determine appropriate strategies for analyzing existing & newly generated data to support the intended COU by evaluation of the relationship between the biomarker & the clinical outcome

By the end of Step 3, sufficient evidence should be available to identify some candidate quantifications of the biomarker (or biomarker panel) that are likely to have associations with certain clinical outcome measures of interest. If a single biomarker quantification has already been determined, then Step 4 will be the validation step, and will only need to be performed once. If some refinement is still required to determine the final quantification, then iteration within step 4 may be required. By the time step 4 is completed, the biomarker is considered statistically validated. The words used to describe these components differ by field, and in quality management, the steps of 'verification, validation, and evaluation' correspond approximately to analysis of training, validation and test datasets. The first of these corresponds to defining the biomarker (parts of our step 1), the second one corresponds to refinement or further validation (early parts of step 4), and the third one refers to a final independent validation in a test set (final part of step 4).

A variety of clinical outcomes may be of interest on either an individual or population (e.g., dose cohort), including injury response, diagnosis, time to diagnosis of medical condition, clinical decline, disease progression or time to death. In Step 4, it is important to determine appropriate strategies for analyzing existing and newly generated data by establishing the relationship between the biomarker and the clinical outcome, demonstrating the relevance of that relationship, selecting thresholds and/or decision rules to assess performance in the specific COU, and validating the performance of the selected thresholds and decision rules.

Establish relationship

Several statistical approaches are available to establish a relationship between biomarker quantification and clinical outcome of interest (see examples in Table 4). Often this relationship is expressed in terms of a statistical model which can be confirmed later in a similar model with a new validation dataset. If a complex model were used to derive the biomarker, the final outcome of that model would be used as the independent variable for the validation model shown in Table 4. Because many biomarkers and outcome measures are inherently continuous in nature, it is preferable to maintain their continuous quantifications when first establishing these relationships and often later when confirming them. In some situations, it may be sufficient to establish this relationship at the population level (e.g., for an enrichment biomarker). For clinical decision-making purposes, the association between biomarker and outcome has to be examined at the individual patient level and is then typically summarized over the population. Quantitative biomarkers often need to be dichotomized or categorized when used for individual patient decisions, and performance of such thresholds must be evaluated as part of the qualification process. Several transformations of variables may be considered and may result in a better characterization of the relationship between the biomarker and the outcome. Initial iterations through this process could be used to further refine the biomarker, with the final iteration providing an independent validation of a specific well-defined biomarker.

Sometimes a 'gold standard' clinical outcome measure will not be available, in which case one or more accepted proxy measures might have to be used to represent the clinical outcome of interest. A proxy measure such as treatment exposure may be useful, particularly for a safety biomarker, if it is known that treatment is associated

with the outcome of interest. In a degenerative disease, time is often useful as a proxy for disease progression in a population without a clear gold standard, and biomarkers that maximally distinguish between patients with two different stages of disease or distinguishing change between treated and untreated at two time points may allow selection of a more progressive population. Other factors may also change over time, such as aging-related conditions, and these need to be considered when using this approach. Maximizing the cross-sectional separation between patients with and without a disease diagnosis may also be useful for selecting or assessing a biomarker, in which case the diagnosis is considered the gold standard. The combination of these two – maximizing the progression rate between patients with and without a disease diagnosis – can combine the progressive proxy and the diagnosis for a better gold standard. After establishing an initial association between the biomarker and clinical outcome, an iterative process may be required to select the biomarker quantification that associates most strongly with the clinical outcome.

One risk in this iterative approach is the possibility of generating spurious associations through multiple testing or overfitting of models. The concept of overfitting refers to situations where complex models may partially fit to noise in the data and hence not perform well when evaluated on independent data. Judicious use of the available data – for example, splitting it into training and validation sets or using resampling methods such as k-fold cross-validation or bootstrapping – can be useful approaches to minimize overfitting of models in earlier iterations of this step. Ideally, the last iteration of this step should use a completely independent dataset to rigorously validate a final, fully defined biomarker measure. The standards for full validation of the biomarker should reflect the context of the biomarker use and the risks and benefits.

Demonstrate relevance

Translation of an initially observed association between a biomarker and a clinical outcome into a biomarker-based tool that can be qualified for use in medical product development requires a careful characterization of the biomarker's performance in the specified COU performance is not adequately characterized by demonstration of a statistically significant association between the biomarker and clinical outcome. The association must be sufficiently strong that the biomarker can make meaningful clinical distinctions between subjects. For continuous biomarker quantifications, one or more thresholds are typically applied to the biomarker to define ranges of the biomarker values for which different clinical decisions or actions would be appropriate.

Consider the performance metrics of positive and negative predictive value for occurrence of an adverse clinical event, e.g., experiencing a toxicity, as an illustration of the simplest case of a binary biomarker quantification. Among those with a biomarker indicating a risk, do a sufficient proportion of participants actually experience a clinical event? For those with a biomarker indicating no risk, do a sufficient proportion fail to experience an adverse clinical event?

- Safety example: a single threshold below which participants are considered to have low risk of injury and above which the likelihood of injury is unacceptable;
- Enrichment example: several thresholds may be relevant in selecting participants for a trial of a new therapy. Each one should be shown with its corresponding summary statistic of the outcome of interest. Examples of summary statistics include event rate for a binary outcome, event-free proportion at a given time, or mean and standard deviation for a continuous outcome (e.g., blood pressure, or rate of kidney function decline).

Implicit in establishing a relationship between the biomarker and clinical outcome that has relevance for a specific COU is that there is at least one measurement method (e.g., laboratory assay for the biomarker) for which the relationship holds. If biomarker data generated by several different measurement methods are available, and the biomarker–outcome relationship is similar regardless of the measurement method or population, then it may sometimes be acceptable to pool data. A sensitivity analysis comparing the measurement approaches could be performed to quantify how much the practical performance of the biomarker may change depending on the assay method, and the conclusion may be that only some assay methods are acceptable. Similarly, a sensitivity analysis could be used to assess variability in performance across populations. If variation is observed across populations, it might be necessary to adjust biomarker values for characteristics that define different populations.

Appropriate ways to pool data will depend on the relationship between the biomarker quantifications generated by different measurement methods or how those measurements are affected by population characteristics. If results generated by two or more measurement methods are completely exchangeable, it may be reasonable to pool raw

biomarker data directly. In other situations, adjustments to the biomarker data including location and scale shifts, quantile normalization or calibration, may be helpful to transform biomarker data to be nearly exchangeable between measurement methods. Other adjustments to biomarker values might account for different distributions of biomarkers dependent on population characteristics. If comparability of biomarker values across measurement methods or populations cannot be achieved, then separate analysis and pooling of summary measures of association between biomarker and clinical outcome across studies using the different measurement methods can still provide supporting evidence. However, in this situation, biomarker qualification may require additional evidence specific to each proposed measurement method or population.

Select thresholds &/or decision rules & assess or validate their performance

Once a relationship between a biomarker and a clinical outcome of interest has been established, decision rules to be applied to the biomarker quantifications usually have to be more precisely defined and characterized (e.g., identifying particular thresholds for decision making in medical product development). This finer characterization of the biomarker test is necessary to establish that the test meets suitably rigorous performance standards necessary to qualify it as fit-for-purpose for the specific COU.

A variety of statistical criteria are commonly used to establish presence of associations and effects. Examples include two-sided p-values <0.05, likelihood ratios above 5 or below 0.2, and AUCs above 0.8. Although these criteria are helpful guidelines, the specific criteria for decision making may differ by specific context and should be carefully considered.

It is important to understand how this step goes beyond the mere confirmation of a statistical association between a biomarker quantification and a clinical outcome. As an illustration, consider a continuous biomarker and its association with a binary outcome (e.g., toxicity event). One way to establish an initial association would be to construct ROC curves displaying the sensitivity and 1-specificity pairs for detection of the binary outcome corresponding to the range of possible thresholds applied to the continuous biomarker values. If the area under that ROC curve is statistically greater than the null value of 0.5, one could say that an association has been established between the biomarker and the binary outcome. However, establishing that a biomarker is fit for purpose for use as a tool in medical product development, would require that an appropriate decision point exists that has acceptable risk benefit for the specific intended use. This requires more stringent standards be met compared with validating sensitivity at a specified threshold or simply confirming an association between biomarker and outcome. The demands of qualification process must be intimately tied to the COU and are best evaluated in collaboration with regulatory authorities.

Performance of the final precisely defined biomarker test may require validation on a new independent dataset. For biomarkers already in use for other COUs, sometimes it might be possible to transfer existing decision rules or thresholds if they achieve an acceptable balance of risks and benefits for the new COU; otherwise, new rules or thresholds may need to be derived. A range of thresholds may be considered. For instance, a strict cutoff for a prognostic marker might be desired to select a very severe disease population. However, a less stringent cutoff might be appropriate for selecting a broader population with severe and moderate disease. It may be appropriate to characterize or calibrate the biomarker by defining several cutoffs across the range of the biomarker and then evaluating the performance for each of those cutoffs. The performance of the biomarker could be evaluated against a gold standard of performance.

Selecting at least one appropriate threshold might involve combining categories of discrete biomarker values or applying thresholds to continuous biomarker quantifications to define ranges of biomarker values to indicate certain clinical decisions or actions (e.g., eligible for a trial, reduce dose, halt therapy). The qualification needs to support the decisions intended to be made with that specific biomarker and threshold.

When biomarker thresholds or decision rules are selected using a specific dataset, it is important to remember that the observed performance on that same data will be better than that expected if those thresholds are applied to completely independent data. This results from the same phenomenon mentioned earlier in the context of overfitting multivariable models. For this reason, thresholds or decision rules that are selected based on available data typically require validation on independent data or by using techniques such as k-fold cross-validation or bootstrapping.

Similar to Table 4, which referred to establishing preliminary biomarker-clinical outcome associations, the method of analysis for definitive clinical performance assessment will depend on whether the clinical outcome variable is binary, time-to-event or continuous. Table 5 presents statistical analysis approaches that may be helpful

Table 5. Analyses to assess clinical performance of biomarker thresholds and decision rules.

Type of outcome	Analysis [†] – performance assessment at specific thresholds
Binary (e.g. organ failure)	<ul style="list-style-type: none"> • Proportion with event at different thresholds (corresponds to different points on the ROC curve); • Sensitivity/specificity; • Positive predictive value/negative predictive value
Time to event outcome	<ul style="list-style-type: none"> • Cox regression – outcome is time to ‘gold standard’ event, covariates are included, specific decision points are assessed for clinical performance; • Kaplan–Meier curves – simple, but do not accommodate covariates; • Many different parametric curves; • Cure models, not assuming proportional hazards
Quantitative/continuous outcome	<ul style="list-style-type: none"> • Mean (SD) decline in biomarker positive group compared with negative group (and total group) with CIs; • Proportion above a critical threshold

[†] Specific analyses may depend on the proposed COU.

in performance assessment, but these analyses must then be supported by an assessment of clinical relevance to establish the utility of specific decision points for clinical decision making.

It is important to appreciate that performance of selected thresholds or decision rules, may depend on population characteristics and specific patient covariates. For example, sensitivity and specificity of biomarkers to detect disease often depend on population characteristics such as disease spectrum or individual-level covariates such as age, gender, disease stage or severity, co-morbidities or medication use. Influence of these factors can be addressed either by demonstration of biomarker performance separately by subgroups or by use of biomarker decision tools (e.g., models or risk scores) that incorporate these factors in addition to the biomarker values. If this relationship is complex, it may make sense to use a computer supported look up table or simulation tool instead of a printed table.

Qualification requires that at least one threshold or decision rule be shown fit for purpose in the specified COU. Thresholds or decision rules used to support a particular qualification are not automatically considered acceptable in all contexts. If several thresholds or decision rules are shown to be relevant in the qualification documentation, then sponsors have more flexibility in the selection to meet their specific needs. If sponsors wish to select thresholds or decision rules different than those in a qualified biomarker then they should discuss with the presiding FDA clinical division to determine whether their choice is suitable for a specific development program or deemed similar enough to one in the initial qualification to be acceptable. A specific threshold or decision rule may require additional supporting evidence if it is not already included in the qualification, particularly if it is out of the range of those previously evaluated.

As described previously, overfitting can occur when the model mean-squared error (MSE) is evaluated using the same data that was used to fit the statistical model. One way to address this issue is by validating the model fit using an independent set of data (i.e., a set of data that was not used to fit the statistical model). Prospective data can be expensive to obtain, and it can be challenging to obtain an independent set of data that has been drawn from the same underlying population. Another approach to model validation is by using k-fold cross-validation, described in the table below, but validation on an independent dataset when available is considered a more objective reflection of model performance and stronger evidence if supported.

Details on determining the appropriate analytic strategy to support the intended COU, including evaluation of the relationship between the biomarker and clinical outcome, demonstrating the relevance of that relationship, selecting thresholds and decision rules and validating such rules are provided in Table 6 (also Supplementary Table 12) in the context of the two illustrative examples.

Step 5: Quantify risks & benefits associated with using the biomarker in the intended COU based on the totality of evidence

Evaluation of risk–benefit balance should be based on the totality of evidence and may extend beyond examination of a single clinical end point or single biomarker. As an illustration, early phase clinical trials of new anti-cancer therapies usually enroll patients with advanced disease who have progressed on standard therapies. While those investigational therapies might have substantial toxicities, biomarkers used to enrich the study population for patients most likely to benefit from the medical product or used to monitor for early signs of toxicity can help tip the balance of potential risks and benefits toward an overall more favorable outcome.

Table 6. Determining the appropriate analytic strategy to support the intended context of use for the two illustrative examples.

Step 4: determine appropriate analytic strategy to support intended COU	
Enrichment example (total kidney volume)	Safety example (kidney toxicity)
<p>Together with the nonlinear mixed effects model for TKV dynamics, the requestor used multivariate Cox regression to investigate the relationship between the covariates and TKV with the outcome of time to 30% worsening of eGFR. Three predictors – age, baseline eGFR and log-transformed baseline TKV – were each associated with the time to 30% decline in eGFR. The ROC curves at the 1-year and 5-year time points resulted in area under the curve of 0.75 and 0.70 at years 1 and 5, respectively, in the model that includes age, baseline eGFR, log (baseline TKV), and all two-way interactions</p> <p>Performance of the enrichment was assessed for groups less than 40 years of age vs at least 40 years of age, and eGFR of at least 50 ml/min vs less than 50 ml/min for those with TKV <1 l and those with TKV of at least 1 l. The results differed within these groups, demonstrating that quantitative terms in a model may need to be assessed at specific decision points to support research and clinical decision making. Cox models demonstrated the added value of including TKV (along with age and eGFR) in the final biomarker algorithm, as compared with using age and eGFR without including TKV. The FDA used the AIC to compare the best fit models with and without TKV, and observed a substantial improvement of 86 in AIC for the model with log (TKV) over the model without TKV, supporting qualification of the use of age, eGFR, and baseline TKV for enriching patient populations for clinical trials</p> <p>Multiple thresholds, representing different levels of enrichment, were identified as useful for clinical decision-making. An electronic simulation tool was provided in relation to the final biomarker qualification submission. In addition, the FDA assessed the findings regarding TKV's added clinical utility on a confirmed 30% decline in eGFR by analyzing an independent dataset during the statistical review</p> <p>The following excerpt from the Qualification Executive Summary section of the FDA's Biomarker Qualification Review for Total Kidney Volume describes the processes of cross-validation and validation, which were done by FDA for the enrichment example: <i>"Including TKV in the model provides a seemingly modest improvement over the best model using age and eGFR alone in terms of predictive performance on event risk based on estimation of a time to event concordance measure that is free of study-specific censoring distributions. This was observed internally (by k-fold cross validation) using the submitter's dataset. FDA performed an independent validation using a separate dataset available internally that further supported the predictive performance and the draft qualification of the prognostic biomarker."</i></p> <p>Further details about the FDA's assessment of the model are provided in the FDA's Biomarker Qualification Review for Total Kidney Volume</p>	<p>The selected reference end point was change in serum creatinine concentrations, which was an imperfect and insensitive indicator of the unverifiable definitive clinical end point of interest (kidney tubular injury). The relationship between urine biomarkers and the gold-standard of kidney histopathology in rodents had previously been established. Many of these biomarkers were expected to be translatable and relevant to humans and were tested in a learning dataset of normal healthy volunteers and mesothelioma patients exposed to cisplatin, a known nephrotoxicant. The majority of the statistical methods discussed above were applied to these learning datasets in an exploratory fashion, and in the end, thresholds were established for individual biomarkers using ROC analyses that best separated the populations of normal healthy volunteers from patients exposed to cisplatin who had evidence of AKI based on serum creatinine</p> <p>The biomarkers that showed trends of interest were further evaluated as candidates for clinical validation, where final selection was based on statistical results and other factors such as practical implications (e.g., assay cost) and biological understanding of relevance to toxicity (e.g., anatomical association)</p> <p>A simple biomarker algorithm was developed to identify patients exposed to a nephrotoxicant, where the algorithm specified a positive result when two or more biomarkers exceeded their individual normal range thresholds, which was established from the normal healthy volunteers in the learning dataset. This decision rule showed acceptable specificity and improved sensitivity of identifying exposure to a nephrotoxicant relative to the standard marker serum creatinine</p> <p>The biomarker panel algorithm was then pre-specified for use in two prospective studies of patients exposed to known nephrotoxicants and controls. In the prospective studies, pre-specified hypotheses will be tested to assess the question of whether the biomarker panel can detect exposure to known nephrotoxicants with improved sensitivity and acceptable levels of specificity relative to the conventional measure of serum creatinine. The prospectively planned statistical analyses to support these hypotheses were detailed in a statistical analysis plan that was reviewed and accepted by the FDA and EMA</p>

Data taken from [19,20].

Performance of selected thresholds or decision rules must be assessed to demonstrate acceptable risk–benefit balance when using the biomarker in the prescribed manner. For illustration of risk–benefit balance, consider a single biomarker measured as a continuous value at time of diagnosis that is prognostic for a clinical outcome. There are many possible thresholds that could be applied to that biomarker if it were used for enrichment of a clinical trial population, but the most appropriate choice of threshold would be dependent on the COU. If the clinical trial involved a randomization of patients to evaluate a more aggressive therapy compared with a standard therapy, enrichment based on the biomarker might be of interest to identify patients most likely to have poor outcome on standard therapy and most likely to benefit if the new therapy is more effective. The biomarker threshold would be set to identify a particularly high-risk subgroup of patients. In contrast, if the goal of the trial was to determine whether a medical product with less toxic side effects (or perhaps a lower dose) would result in similar survival, the biomarker threshold might be set to identify a subgroup of patients with a particularly favorable outcome on standard therapy.

The risk–benefit assessment must take into consideration a variety of factors including feasibility of performing the biomarker assays in a timely fashion so as not to unacceptably slow enrollment times. The overarching goal of biomarker qualification is to make medical product development faster, more efficient and less costly. Details on the qualitative considerations of risks and benefits associated with use of the biomarker are provided in Supplementary Table 13 in the context of the two illustrative examples.

Step 6: finalize the COU statement & summarize the supporting evidence

The COU statement and the level of evidence necessary to support it will differ depending on whether the biomarker is intended to stand alone, replace an existing biomarker, or be used in conjunction with an existing biomarker. The biomarker qualification process is meant to be an iterative and collaborative process of proposing, and refining the supporting data and documentation, with a goal to qualify the biomarker. The summary and current status of the qualification process is given below in Supplementary Table 14 for the enrichment and safety biomarker examples.

Discussion & conclusion

The BEST glossary provides a basis for the categorization of biomarker applications, but such categories do not, by themselves, constitute fully defined COU statements. The COU statement requires the definition of a target population for use, and a specific description of how the biomarker will be used in medical product development (for which types of trials, and for which types of decisions). Once a COU has been identified, biomarker qualification requires several different types of statistical analyses to support the biomarker for that specified COU and these are detailed in a statistical analysis plan submitted at the letter of intent phase of the submission. The conceptual framework described here for the development and evaluation of evidence to support the use of a biomarker in medical product development can be used as a reference for teams to consider when defining design elements and statistical analysis considerations to support biomarker qualification efforts. Although the proposed framework involves a stepwise process and was presented in linear fashion, in practice it may require iteration to appropriately align evidence and refine the intended COU.

The framework provided here applies to clinical safety biomarkers as well as biomarkers for prognostic enrichment for various COUs as illustrated by the kidney safety biomarker panel and total kidney volume examples. With appropriate adaptations, the framework can be extended to qualification of biomarkers for other uses in medical product development. Clearly, the statistical aspects of the qualification process must be considered in combination with other clinical and biological factors. Discussion of the broader aspects of qualification can be found elsewhere [3,21]. The examples described in this article represent biomarker qualification decisions by the FDA before the 21st Century Cures Act legislation [22]. As such, it is important to keep in mind that these examples should not be viewed as precedent for ongoing or future biomarker qualification reviews under the new procedures associated with this legislation.

Because the biomarker qualification process entails assembly of evidence to support use of a biomarker in diverse clinical development programs, the evidentiary standards for qualification will generally be higher than the standards for validating a biomarker in the context of an individual clinical development program as explained in FDA guidance [20]. Yet, standards for qualification do not necessarily meet those for routine clinical use of a biomarker-based test as explained on the CDER biomarker qualification website [21] emphasizing that “*qualification of a biomarker does not imply that a specific test device used in the qualification process for the biomarker has been reviewed by FDA and cleared or approved for use in patient care.*” Sponsors may choose which biomarker development pathway is most suitable for their needs including the Fit for Purpose Initiative which “*provides a pathway for regulatory acceptance of dynamic tools in medical product development programs*” [22]. Groups initiating a biomarker qualification process are strongly urged to communicate early with FDA and EMA and work collaboratively with the regulatory authorities to achieve their goals most efficiently.

Future perspective

Some of the steps outlined here may be helpful for validation of individual biomarkers or development and validation of complex biomarkers. As new biomarkers are developed and as they become more commonly used, the biomarker qualification process may be relevant for improving the efficiency of clinical programs that use these biomarkers. As more biomarkers are qualified, it will be important to assess the real-world performance of these biomarkers. It will be of interest to track how often they are used as medical product development tools and whether they have achieved the goal of making medical product development faster, more efficient and less costly.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: www.futuremedicine.com/doi/suppl/10.2217/bmm-2020-0523

Disclaimer

This article reflects the personal views of the FDA and NIH authors and should not be construed to represent the views, policies, or guidance of the US FDA or the NIH. Views expressed in written materials or publications and by speakers and moderators do not necessarily reflect the official policies of the Department of Health and Human Services; nor does any mention of trade names, commercial practices or organization imply endorsement by the United States Government.

Acknowledgments

The authors would like to thank A Brownlee and A Porter for medical writing assistance. The formation of the Statistics Working Group on Biomarker Qualification Statistical Considerations was facilitated by ShaAvhree Buckman-Garner immediately following the M-CERCI workshop in 2015.

Financial & competing interests disclosure

Funding for the TKV qualification example was made possible, in part, by the FDA through grant (U18 FD 005320). The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

Writing assistance was provided by Allison Brownlee at Pentara and Amy Porter at CPath.

Open access

This work is licensed under the Attribution-NonCommercial-NoDerivatives 4.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>

Executive summary

- Qualification of a biomarker for a specific context of use requires a statistical strategy to align available evidence with the proposed context of use, identify any gaps to be filled and plan any additional research required to support the qualification. Accumulating, interpreting and analyzing available data is outlined in the following steps.
 - Step 1: consider the drug or medical product development need and initial COU statement for the biomarker to formulate an overall strategy for data accession and statistical evaluation;
 - Step 2: collate existing evidence relating the biomarker to the key clinical outcome(s) in order to assess what information relevant to the intended COU is currently readily available and identify gaps;
 - Step 3: identify any additional data or specimens available for analysis that can be used to build on existing evidence and determine what prospective data may need to be generated to fill identified gaps;
 - Step 4: determine appropriate strategies for analyzing existing and newly generated data to support the intended COU by evaluation of the biomarker–clinical outcome relationship;
 - Step 5: statistically quantify risks and benefits associated with using the biomarker for the intended COU based on the totality of evidence, demonstrating the practical relevance of the biomarker;
 - Step 6: finalize the COU statement and summarize the supporting evidence.
- The specific level of evidence required differs depending on the COU and the risk and benefit of the proposed use.
- The biomarker qualification process is meant to be an iterative and collaborative process involving the statistical steps outlined above integrated with clinical considerations. These steps aid researchers in organizing and evaluating available evidence to support the qualification of a biomarker.

References

1. FDA-NIH Biomarker Working Group. BEST (biomarkers, endpoints, and other tools) resource (2016). <http://www.ncbi.nlm.nih.gov/books/NBK338448/>
2. US Food and Drug Administration. Qualifying a biomarker through the biomarker qualification program (2018). <https://www.fda.gov/drugs/cder-biomarker-qualification-program/qualifying-biomarker-through-biomarker-qualification-program#2>
3. US Food and Drug Administration. Qualification process for drug development tools guidance for industry and FDA staff. *Draft Guidance* (2019). <https://www.fda.gov/media/133511/download>
4. Leptak C, Menetski JP, Wagner JA *et al.* What evidence do we need for biomarker qualification? *Sci. Transl. Med.* 9(417), eaal4599 (2017).
5. Critical Path Institute. Evidentiary considerations for integration of biomarkers in drug development symposium 2 (2015). <https://c-path.org/evidentiary-considerations-for-integration-of-biomarkers-in-drug-development-symposium-2/>

6. Lawrence J, Wang S-J, Hung HMJ. Regulatory experience of qualifying total kidney volume as a prognostic biomarker for clinical trial enrichment: statistics in biopharmaceutical research. *Stat. Biopharmaceut. Res.* 11(3), 239–251 (2019).
7. Perrone RD, Mouksassi M-S, Romero K *et al.* Total kidney volume is a prognostic biomarker of renal function decline and progression to end-stage renal disease in patients with autosomal dominant polycystic kidney disease. *Kidney Int. Rep.* 2(3), 442–450 (2017).
8. Perrone RD, Mouksassi M-S, Romero K *et al.* A drug development tool for trial enrichment in patients with autosomal dominant polycystic kidney disease. *Kidney Int. Rep.* 2(3), 451–460 (2017).
9. Schutz C, Boulware DR, Huppler-Hullsiek K *et al.* Acute kidney injury and urinary biomarkers in human immunodeficiency virus-associated cryptococcal meningitis. *Open Forum Infect. Dis.* 4(3), ofx127 (2017).
10. Schaub JA, Parikh CR. Biomarkers of acute kidney injury and associations with short- and long-term outcomes. *F1000Res.* 986, 5 (2016).
11. Parikh CR, Moledina DG, Coca SG, Thiessen-Philbrook HR, Garg AX. Application of new acute kidney injury biomarkers in human randomized controlled trials. *Kidney Int.* 89(6), 1372–1379 (2016).
12. Critical Path Institute. Approaches to statistical analysis for biomarker qualification (2014). <https://c-path.org/approaches-to-statistical-analysis-for-biomarker-qualification-defining-the-components-to-drug-development-tool-ddt-qualification/>
13. US Food and Drug Administration. US Food and Drug Administration C for DE and FDA Qualification of Composite Measure of 6 urinary biomarkers to aid in detection of kidney tubular injury in Phase 1 trials of healthy volunteers (2018). <https://www.fda.gov/media/115690/download>
14. US Food and Drug Administration. CDER biomarker qualification program (2018). <https://www.fda.gov/drugs/drug-development-tool-ddt-qualification-programs/cder-biomarker-qualification-program>
15. Ewen JB, Sweeney JA, Potter WZ. Conceptual, regulatory and strategic imperatives in the early days of EEG-based biomarker validation for neurodevelopmental disabilities. *Front. Integr. Neurosci.* 13, 45 (2019).
16. Biomarker Assay Collaborative Evidentiary Considerations, Writing Group, Critical Path Institute (C-Path). Points to consider document: scientific and regulatory considerations for the analytical validation of assays used in the qualification of biomarkers in biological matrices (2019). <https://c-path.org/wp-content/uploads/2019/06/EvidConsid-WhitePaper-AnalyticalSectionV20190621.pdf>
17. Buyse M, Molenberghs G, Paoletti X *et al.* Statistical evaluation of surrogate endpoints with examples from cancer clinical trials. *Biom J.* 58(1), 104–132 (2016).
18. Institute of Medicine (US) Committee on Qualification of Biomarkers and Surrogate Endpoints in Chronic Disease. *Evaluation of Biomarkers and Surrogate Endpoints in Chronic Disease*. Micheel CM, Ball JR (Eds). National Academies Press, DC, USA (2010).
19. U.S. Food and Drug Administration. Biomarker qualification review for total kidney volume (2018). <https://www.fda.gov/media/93143/download>
20. Vlasakova K, Erdos Z, Troth SP *et al.* Evaluation of the relative performance of 12 urinary biomarkers for renal safety across 22 rat sensitivity and specificity studies. *Toxicol. Sci.* 138(1), 3–20 (2014).
21. Cagney DN, Sul J, Huang RY, Ligon KL, Wen PY, Alexander BM. The FDA NIH Biomarkers, EndpointS, and other Tools (BEST) resource in neuro-oncology. *Neuro-Onc.* 20(9), 1162–1172 (2018).
22. U.S. Food and Drug Administration. Drug development tool qualification process: transparency provisions (2019). <https://www.fda.gov/drugs/drug-development-tool-ddt-qualification-programs/drug-development-tool-qualification-process-transparency-provisions>