Minireview

Biomarker definitions and their applications

Robert M Califf[1,2,3]

[1]School of Medicine, Duke University, Durham, NC 27710, USA; [2]Verily Life Sciences (Alphabet), South San Francisco, CA 94043, USA; [3]Department of Medicine, Stanford University, Stanford, CA 94305, USA
Corresponding author: Robert M Califf. Email: robert.califf@duke.edu

Impact statement

Biomarkers are critical to the rational development of medical diagnostics and therapeutics, but significant confusion persists regarding fundamental definitions and concepts involved in their use in research and clinical practice. Clarification of the definitions of different biomarker classes and a better understanding of their appropriate application could yield substantial benefits. Biomarker definitions recently established in a joint FDA-NIH resource place different classes of biomarkers in the context of their respective uses in patient care, clinical research, or therapeutic development. Complex composite biomarkers and digital biomarkers derived from sensors and mobile technologies, together with biomarker-driven predictive toxicology and systems pharmacology, are reshaping development of diagnostic and therapeutic technologies. An approach to biomarker development that prioritizes the quality and reproducibility of the science underlying biomarker development and incorporates collaborative regulatory science involving multiple disciplines will lead to rational, evidence-based biomarker development that keeps pace with scientific and clinical need.

Abstract

Biomarkers are critical to the rational development of medical therapeutics, but significant confusion persists regarding fundamental definitions and concepts involved in their use in research and clinical practice, particularly in the fields of chronic disease and nutrition. Clarification of the definitions of different biomarkers and a better understanding of their appropriate application could result in substantial benefits. This review examines biomarker definitions recently established by the U.S. Food and Drug Administration and the National Institutes of Health as part of their joint *Biomarkers, EndpointS, and other Tools (BEST)* resource. These definitions are placed in context of their respective uses in patient care, clinical research, or therapeutic development. We explore the distinctions between biomarkers and clinical outcome assessments and discuss the specific definitions and applications of diagnostic, monitoring, pharmacodynamic/response, predictive, prognostic, safety, and susceptibility/risk biomarkers. We also explore the implications of current biomarker development trends, including complex composite biomarkers and digital biomarkers derived from sensors and mobile technologies. Finally, we discuss the challenges and potential benefits of biomarker-driven predictive toxicology and systems pharmacology, the need to ensure quality and reproducibility of the science underlying biomarker development, and the importance of fostering collaboration across the entire ecosystem of medical product development.

Keywords: Biomarkers, cardiovascular, epidemiology, medicine, monitoring, pharmacology/toxicology

Experimental Biology and Medicine **2018; 243: 213–221. DOI: 10.1177/1535370217750088**

Introduction

Biomarkers are critical to the rational development of drugs and medical devices.[1] But despite their tremendous value, there is significant confusion about the fundamental definitions and concepts involved in their use in research and clinical practice. Further, the complexity of biomarkers has been identified as a limitation to understanding chronic disease and nutrition.[2]

Several years ago, this issue came to a head. At a joint leadership conference of the U.S. Food and Drug Administration (FDA) and the National Institutes of Health (NIH), it became apparent that leaders from each federal agency had differing impressions about the appropriate definitions of biomarkers in different contexts of use. A joint task force was therefore formed to forge common definitions and to make them publicly available through a continuously updated online document–the "Biomarkers, EndpointS, and other Tools" (BEST) resource.[3]

The importance of well-understood definitions and a shared understanding of how to apply them should not

be underestimated. Science has produced a surfeit of associations between biological measurements and models of disease at the subcellular, cellular, organ, biological system, and intact organism levels. This steadily increasing ability to make measurements in model systems, animals, and humans has led to an avalanche of potential biomarkers for states of disease and wellness, extending beyond pure research into medical product development, clinical practice, nutrition, and environmental policy development. But at the same time, the potential for much more acute biological measurement has been blunted by confusion about definitions that is slowing or even stalling progress toward development of useful diagnostic and therapeutic technologies.

The concept behind BEST is that improving our collective ability to match a biomarker with its appropriate purpose will enable greater speed, efficiency, and precision in the development of useful diagnostic and therapeutic technologies and strategies, as well as benefitting the development and implementation of public health policies. When scientific resources are devoted to developing a biomarker application that does not meet criteria for regulatory approval, reimbursement, or clinical use, the financial and human investments are wasted. Even in early translational research, mistaken concepts about future use can lead to an unfortunate diversion of funding and scientific effort toward biomarker development programs that are destined to yield inaccurate estimates of effects on animal or human health.

In this section, these definitions will be reviewed and placed into context. Examples from the field of cardiovascular disease will be used because of the author's specific experience in this field, although the concepts are applicable to all areas of human and veterinary medicine. The chapter does not go into detail about the validation process, which is covered in other sections. However, it is worth noting that the process of validation requires the specific and interdependent steps of analytical validation, qualification using an evidentiary assessment, and utilization (Figure 1).[2] These steps are specific to each condition of use for the biomarker.

Figure 1. Steps in the evaluation framework for biomarkers. Adapted from: Institute of Medicine. *Evaluation of biomarkers and surrogate endpoints in chronic disease.* Summary. Washington, D.C.: National Academies Press, 2010.

Biomarkers, clinical outcome assessments, and endpoints

The basic definition of a biomarker is deceptively simple: "A defined characteristic that is measured as an indicator of normal biological processes, pathogenic processes or responses to an exposure or intervention."[3] This broad definition encompasses therapeutic interventions and can be derived from molecular, histologic, radiographic, or physiologic characteristics. For the sake of clarity, biomarkers should be distinct from direct measures of how a person feels, functions, or survives—a category of measure known as a clinical outcome assessment (COA). This difference between biomarkers and COAs is important, because COAs measure outcomes that are directly important to the patients and can be used to meet standards for regulatory approval of therapeutics, whereas biomarkers serve a variety of purposes, one of which is to link a measurement to a prediction of COAs. Only when a biomarker is validated can it serve as the primary basis for regulatory approval for marketing, except in circumstances where no effective therapy is available. In such situations, the biomarker may be used to support approval under one of several accelerated approval pathways[4] as deemed appropriate by FDA reviewers.

Biomarkers and COAs take on additional complexity—and corresponding need for scientific rigor—when used as endpoints in clinical studies. An *endpoint* is a precisely defined variable intended to reflect an outcome of interest that is analyzed using statistics to address a particular research question.[3] Although a biomarker or COA may be discussed in a more general sense, when either is used as an endpoint, a degree of rigor that includes multiple dimensions is required. What is the precise definition, and what are the steps that will be used to measure the endpoint of interest? When will the measurement(s) occur, and how will multiple measurements in the same individual be handled in the analysis? Thus, the investigation of a biomarker can posit a less specific construct for the general development of scientific and technological concepts, but clinical study endpoints must be precisely defined to yield reliable and reproducible results.

Biomarker definitions

A number of subtypes of biomarkers have been defined according to their putative applications. Importantly, a single biomarker may meet multiple criteria for different uses, but it is important to develop evidence for each definition. Thus, while definitions may overlap, they also have clear distinguishing features that specify particular uses.

Diagnostic biomarkers

A *diagnostic biomarker* detects or confirms the presence of a disease or condition of interest, or identifies an individual with a subtype of the disease.[3] As we move into the era of

precision medicine, this type of biomarker will evolve considerably. Such biomarkers may be used not only to identify people with a disease, but to redefine the classification of the disease. For example, the detection of cancer is moving rapidly toward a molecular and imaging-based classification rather than a largely organ-based classification scheme.

Given a diagnostic biomarker that can be measured with sufficient precision and reliability with a delineated context of use, the assessment of that biomarker remains complex. One goal is to define a method for validation that assures that the biomarker can be measured reliably, precisely, and repeatably at a low cost. All too often, assays are not validated, engendering misleading assumptions about the biomarker's value. The complexity of validation can be seen in the use of troponin, clearly an important biomarker for the diagnosis of acute myocardial infarction. The operating characteristics of the many assays for troponin vary considerably, especially at the lower limit threshold, where misclassification can lead to a major difference in medical care. Furthermore, while the advent of high-sensitivity troponin assays has opened many avenues for sophisticated diagnosis of small episodes of myocardial necrosis, it has created further confusion in the field. When small elevations of troponin occur at previously undetectable levels, the clinical consequences are unclear. We can expect that as measurement methods continue to improve, the understanding of the value of individual diagnostic biomarkers will likewise evolve.

If a diagnostic biomarker moves beyond a general application, such as advancing scientific concepts, to specific use in prospective research or clinical practice, close attention must be paid to the context of use. A diagnostic biomarker may be useful in one set of clinical circumstances but completely misleading in another context. For example: in low-prevalence diseases such as pancreatic or ovarian cancer for which a new diagnosis is psychologically devastating or would require invasive evaluation, a biomarker must have a very low false-positive rate. On the other hand, in screening for common diseases such as hypertension or hyperlipidemia for which repeated assessments can be done with little risk, higher false-positive rates are tolerable and the focus of concern may be on false-negative rates.

The use of receiver-operating characteristic curves has enabled a rational process of diagnostic biomarker evaluation to proceed.[5] A common problem, however, is the absence of a historical standard for defining the presence or absence of the disease or condition. Furthermore, decision thresholds and clinical utility are becoming important measures for assessing the value of biomarkers for clinical application. In the future, proof that a biomarker adds information about diagnosis may be necessary but not sufficient. Rather, the key question will be whether the additional information is substantial enough to lead to a change in clinical decision-making. Statistics for evaluating this issue, such as the net reclassification index, are evolving.[6] Researchers involved in early preclinical biomarker research would be well served to understand how the biomarker will eventually be evaluated, just as those doing early drug development should have the ultimate use in humans in view.[7]

Monitoring biomarkers

When a biomarker can be measured serially to assess the status of a disease or medical condition for evidence of exposure to a medical product or environmental agent, or to detect an effect of a medical product or biological agent, it is a *monitoring biomarker.* Monitoring is a broad concept, so there is overlap with other categories of biomarkers as described below.

Monitoring biomarkers have important applications in clinical care. When blood pressure is treated or low-density lipoprotein (LDL) cholesterol-lowering drugs are used, blood pressure or LDL cholesterol levels are monitored. Similarly, when HIV infection is treated, CD4 counts are monitored. But while the general concept of monitoring for clinical purposes is intuitive, arriving at a more refined understanding of what changes in the biomarker should signal a particular change in clinical course and decision-making (e.g. more testing or intervention) is complex and often less precise than is desirable.

For example, target measurements for hemoglobin (Hb) A1C,[8] blood pressure,[9] and LDL cholesterol[10] remain controversial despite these being among our most well-studied and accepted biomarkers. Similarly, we often lack sufficient empirical confirmation of the most helpful interval between measurements or the duration of the clinical course during which measurements should be made. Many biomarkers routinely used in clinical practice have very imprecise operating characteristics, so that they are used in a clinical "gestalt" along with the phrase "clinical judgment is needed." Yet the specifics of clinical parameters that should go into a good clinical judgment are unspecified.

When medical products are developed, changes in biomarkers are routinely used to make decisions about whether key thresholds have been reached, allowing developers to conclude that the therapy affected a biological target enough to merit continued development of the product. Most initial biomarkers used for this purpose measure effect on the assumed target of the intervention, so that changes in the biomarker indicate target engagement and related activity. As discussed below, the ability to measure off-target effects on biological systems will increasingly bring panels of biomarkers and systems measurement into play to evaluate intermediate findings in medical product development.

Monitoring biomarkers are also important in ensuring the safety of human research participants. For example, the safety threshold for drugs with possible liver toxicity is monitored through serial measurement of liver function tests, and cardiovascular events are measured through the use of serial troponins.

Monitoring biomarkers are also useful for measuring pharmacodynamic effects, to detect early evidence of a therapeutic response, and to detect complications of a disease or therapy. International normalized ratio (INR) is a classical pharmacodynamic measure used to titrate the dose of warfarin anticoagulation. Similarly, when blood pressure is treated, a reduction in the measure of blood pressure provides evidence that the therapy is working.

One of the more interesting aspects of monitoring biomarkers is the almost unalterable belief held by many researchers and clinicians that changes in biomarker measurements give the best measure of the likely outcome for a patient or population. However, in many circumstances the actual measure, not the change, is the best predictor of outcome, even if the change is the best way to monitor whether the therapy itself is having an effect. For example, an angiotensin-converting enzyme (ACE) inhibitor may cause an elevation of serum creatinine and/or potassium, and this provides a measure of drug effect. However, the risk to the patient or research participant is primarily determined by the actual creatinine or potassium level, not the change in levels.

Pharmacodynamic/response biomarkers

When the level of a biomarker changes in response to exposure to a medical product or an environmental agent, it can be called a *pharmacodynamic/response biomarker*. This type of biomarker is extraordinarily useful both in clinical practice and early therapeutic development. If one is treating hypertension or diabetes and no reduction in blood pressure or glucose occurs with a therapy, there is good reason to eschew that intervention and pursue another. Similarly, a candidate drug for a condition that does not alter the key parameter of that biomarker in phase 1 trials would hardly be worth pursuing. A special circumstance is phase 1 studies of normal individuals. It would be unexpected for a disease-related biomarker to show a major change (for example, blood pressure) in persons with normal baseline values. In this circumstance, the main focus is on developing preliminary evidence that the drug will be safe to use in individuals with the target disease. For many drugs, dosing is determined by measured change in a pharmacodynamic/response biomarker when a therapy is given.

However, the interpretation of pharmacodynamics/response biomarkers is not always simple or straightforward. In the case of ACE inhibitors, the initial view was that acute titration of dose in the intensive care unit could guide dosing in heart failure patients. And indeed, it was possible to see major differences in the responsiveness of different patients to the same dose. But unfortunately these acute responses did not adequately predict long-term responses. It is therefore critically important to validate that the measured change in the pharmacodynamics/response biomarker provides a reliable signal for the expected therapeutic response.

Another complex problem arises when easily measureable biomarkers do not reflect true pharmacodynamic responses. With intravenous fibrinolytic agents, serum pharmacokinetics do not reflect the activity of the agent in the thrombus. Similarly, amiodarone is heavily deposited in fat and therefore has a much longer duration of activity than simple measurement of serum levels would predict.

Predictive biomarkers

A *predictive biomarker* is defined by the finding that the presence or change in the biomarker predicts an individual or group of individuals more likely to experience a favorable or unfavorable effect from the exposure to a medical product or environmental agent.[3] Proving that a biomarker is useful for this purpose requires a rigorous approach to clinical studies. Ideally, patients with or without the biomarker are randomized to one of two or more treatments (or a placebo comparator) and differences in outcome as function of treatment are significantly related to the difference in presence, absence, or level of the biomarker. Proof of a reliable predictive biomarker thus represents a "high hurdle" to clear.

Predictive biomarkers are important for *enrichment strategies*[11,12] in the design and conduct of clinical trials. Especially in the pre-registration phase of development, focusing enrollment on participants with elevated levels of a predictive biomarker enables a clearer signal that the treatment actually has an effect by enrolling people in whom the treatment is likely to "work." Using predictive biomarkers for enrichment is a more targeted approach than using prognostic biomarkers, which can be used to increase event rates but not to select specific patients who are more likely to respond or not respond to therapy.

The same thinking underlies much of the current consensus about treatment choice in clinical practice. Antihypertensive medications are prescribed for patients with elevated blood pressure; blood transfusion is used in people with anemia measured by low Hb levels; acute reperfusion is indicated in patients with ST-segment elevation on an electrocardiogram—all of these are examples of biomarkers that differentially select patients likely to respond to therapy. Similarly, populations at increased risk due to high levels of predictive biomarkers are identified as needing additional intervention in population health strategies. For example, patients with high levels of HbA1C have the most to gain from aggressive therapies to treat diabetes. In addition, a major growth area in predictive biomarkers is the development of genetic and genomic markers for precision medicine, as in the case of cancer patients with HER2 receptor positive assays who are more likely to respond to treatment with herceptin.

The biomarker-guided use of LDL cholesterol-lowering drugs offers an excellent example of the complexity of these issues. LDL cholesterol is clearly a *susceptibility/risk biomarker* and a *prognostic biomarker*. Patients with elevated LDL cholesterol are at increased risk both of developing atherosclerosis and of experiencing an event such as death, stroke, or myocardial infarction once they have diagnosed disease. Statins, the selective cholesterol absorption inhibitor ezetimibe, and PCSK9 inhibitors all lower LDL cholesterol levels and reduce mortality and critical clinical events such as stroke. However, in multiple clinical trials cumulatively enrolling more than 100,000 patients, the relative effect on event reduction is similar across all levels of LDL cholesterol, including levels well within the normal range.[13] Therefore, in clinical trials, the event reduction is a function of the overall relative risk reduction and the absolute risk of an event, which is determined not only by LDL cholesterol levels, but also by multiple factors, including age, smoking status, diabetes, and blood pressure. Environmental exposures have similar characteristics. Individuals and subpopulations may have particular risks

associated with specific biomarkers such that preventive measures are most likely to be useful in people with elevated levels of those biomarkers.

Prognostic biomarkers

A *prognostic biomarker* is used to identify the likelihood of a clinical event, disease recurrence, or disease progression in patients with a disease or medical condition of interest. Although this distinction is not uniformly accepted, the BEST working groups concluded that *prognostic biomarkers* should be differentiated from *susceptibility/risk* biomarkers, which deal with association with the transition from healthy state to disease. Furthermore, they are distinguished from *predictive biomarkers*, which identify factors associated with the effect of intervention or exposure.

In clinical trials, prognostic biomarkers are routinely used to set trial entry and exclusion criteria to identify higher-risk populations. The key issue is that the statistical power of a trial is determined by the number of events rather than the sample size. When trials are enriched in this manner, the event rates are increased; if the treatment is effective, the differences in outcomes as a function of treatment are magnified quantitatively but not qualitatively. In addition, prognostic biomarkers are especially important for predicting the risk of an event or poor outcome in an individual. This information is key to decisions about length of stay in hospital and/or in intensive care units. Yet another major use of prognostic biomarkers is for resource allocation in population health: by stratifying the risk for both negative clinical and financial outcomes, a healthcare organization can distinguish which patients could benefit from more intensive evaluation while allowing others to avoid unnecessary additional diagnostic tests or medical interventions.

Safety

A *safety biomarker* is measured before or after an exposure to a medical intervention or environmental agent to indicate the likelihood, presence, or extent of a toxicity as an adverse event. For many therapies, monitoring for hepatic, renal, or cardiovascular toxicity is critical to assuring that a given therapy can be safely sustained.

Safety biomarkers are useful for identifying patients who are experiencing adverse effects from a treatment. When antiarrhythmic drugs are prescribed, prolongation of the QT interval on the electrocardiogram is used as a safety biomarker because it predicts the risk of developing the lethal arrhythmia torsades de pointes and can be used to identify patients in need of countermeasures for effective therapy. Similarly, safety biomarkers can be used to monitor a population for exposure to an environmental risk or to monitor a population after an exposure.

An interesting aspect of developing safety biomarkers is the balance that should be sought between safety and the potential benefits of therapy. Returning to the example of QT interval monitoring: the effect such monitoring has had on drug development has been a topic of frequent discussion and controversy. It is possible that a drug whose benefits outweighed its risks has been missed because development was stopped when QT interval prolongation was detected. The Cardiac Safety Research Consortium, which includes representatives from the FDA, industry, and academia, is working on strategies for establishing an optimal balance between the ability to measure risk through early biomarker detection with the potential for benefit.[14]

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 is classified as a *susceptibility/risk* biomarker. The concept is similar to prognostic biomarkers, except that the key issue is the association with the development of a disease rather than prognosis after one already has the diagnosis. These types of biomarkers are foundational for the conduct of epidemiological studies about risk of disease.

Prognostic versus predictive biomarkers

The distinction between *prognostic* and *predictive* biomarkers is critically important when assessing likely disease outcomes with treatment. Prognostic biomarkers are associated with differential disease outcomes, but predictive biomarkers discriminate those who will respond or not respond to therapy. For example: ST-segment deviation on the electrocardiogram is a prognostic biomarker, but the direction of the ST-segment change is a crucial predictive biomarker and ST-segment elevation predicts response to fibrinolytic therapy, whereas ST-segment depression predicts a lack of response to therapy. The issue is easiest to visualize in the context of an "all-or-nothing" response scenario in which the treatment effect is clearly different depending on the level of the biomarker. However, in many cases, the response is graded (a spectrum of responses), probabilistic (the treatment is effective in most, but *more or less* effective in those with the biomarker), or both.

Surrogates

The single most common and serious error in the evaluation of biomarkers is the assumption that a correlation between the measured level of a biomarker and a clinical outcome means that the biomarker constitutes a valid surrogate. In fact, for a biomarker to qualify as a surrogate, the biomarker must not only be correlated with the outcome, but the change in the biomarker must "explain" the change in the clinical outcome. The term "explains" invokes statistical inference, which can only be made with confidence if the observation is made in multiple therapies that all change the biomarker. This high bar means that the overwhelming majority of biomarkers are not valid surrogates; further, even when a surrogate is validated, that validation only pertains to a specific context of use.

The classic work of Fleming and DeMets[15] and Prentice[16] clearly delineates the reasons that "a correlation does not a surrogate make" (Figure 2). Biological pathways and therapeutic effects are multifaceted and redundant.

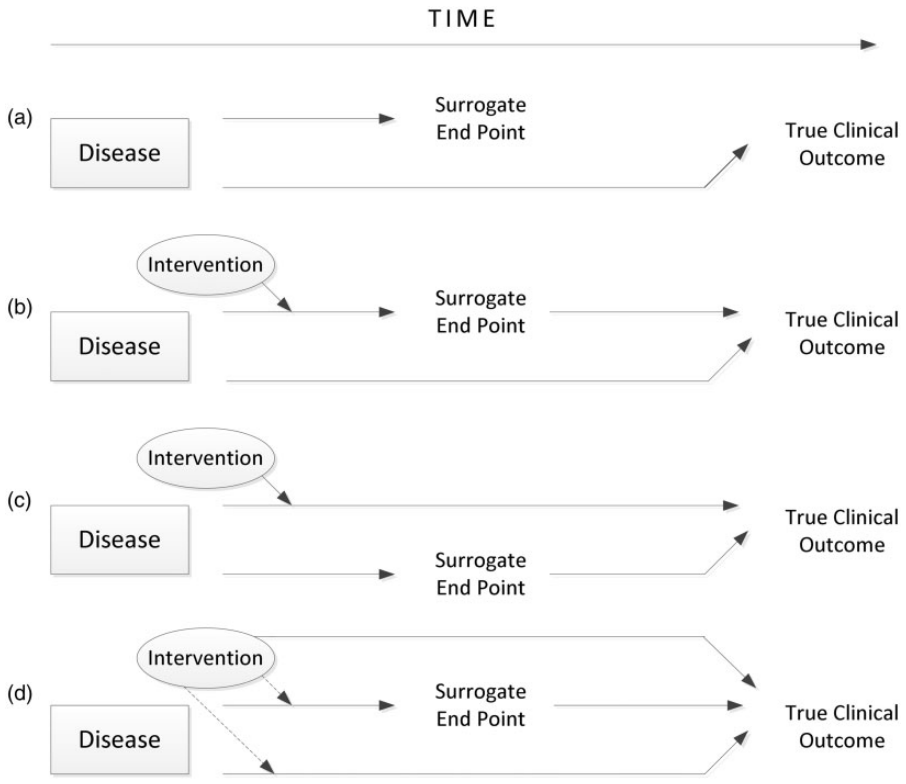

Figure 2. Reasons for failure of surrogate endpoints. (a) In this situation, the disease affects the putative surrogate endpoint and the true clinical outcome via different mechanisms, so that any correlation between the two is not causal. (b) The intervention affects the putative surrogate endpoint, which has some impact on the true clinical outcome. Unfortunately, the disease affects the true clinical outcome by other mechanisms, which make the change in the putative surrogate an unreliable measure of change in the true clinical outcome. (c)The intervention affects the putative surrogate endpoint through mechanisms independent of its effect on the true clinical outcome. Thus, the change in the surrogate endpoint is not a reliable measure of the change in the true clinical outcome. (d) All of the above issues are in play. Adapted from: Fleming TR, DeMets DL. Surrogate end points in clinical trials: are we being misled? *Ann Intern Med* 1996;**25**:605–13.

This means that a therapy can change an outcome without affecting the putative surrogate, it can change the putative surrogate without changing the clinical outcome, or it can change both to a variable degree. Among many excellent examples: high-density lipoprotein (HDL) cholesterol is a notably excellent prognostic and susceptibility biomarker, but when employed as a surrogate, it has failed multiple times across many classes of drugs. People with low levels of HDL are susceptible to developing atherosclerosis and thus are more likely to have poor outcomes, but drugs that raise levels of HDL cholesterol have had either no effect or detrimental effects on clinical outcomes.

The reason it is so important to get this concept right is because surrogates substitute for clinical outcomes and thus can be used to draw inferences about whether a treatment is clinically beneficial. The FDA has the statutory authority to approve medical products for marketing based on validated biomarkers, but the amount of work required to validate a biomarker is substantial. For each biomarker and endpoint, this means that multiple clinical trials that measure both the outcome and the biomarker must be done to demonstrate that the relationship between the change in biomarker and the change in outcome is generalizable across therapies.

The future

The application of these definitions would require substantial discipline even if the underlying scientific fields were static. However, we are currently witnessing tremendous developments in systems biology. At the same time, continuous progress in our capacity to store, collate, and compute massive amounts of information is fundamentally changing our understanding of both biology and clinical outcomes. Taken together, these developments augur a period of explosive growth and rapid change in the field of biomarkers that will occur in tandem with a blossoming in the fields of clinical pharmacology and toxicology. Some examples of critical trends are given below.

Complex biomarkers

The field of biomarkers has been built on critical measures with profound associations with disease that can be understood in a straightforward paradigm. For instance: LDL cholesterol is associated with the risk of cardiovascular disease and lower LDL cholesterol is better; higher systolic blood pressure is associated with stroke and lower systolic blood pressure is better. However, biological systems are complex and multidimensional. As increasingly sophisticated biological models are developed, it is clear that evaluating one biomarker in the absence of an understanding of others can lead to erroneous conclusions. In addition, measurement of complex, composite biomarkers may enable better predictions because multiple biomarkers each play a small role in the summative outcome of interest.

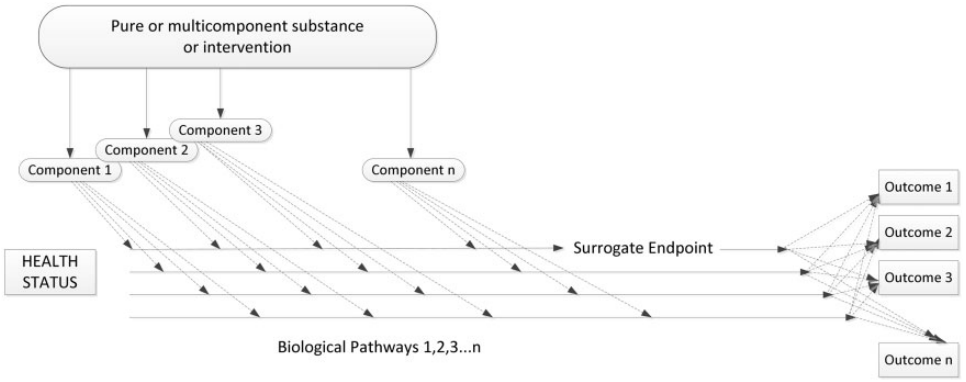

Figure 3. Multiple components, biological pathways, and outcomes all contribute to the complexity of using biomarkers and surrogate endpoints in the context of chronic disease. Adapted from: Institute of Medicine. *Evaluation of biomarkers and surrogate endpoints in chronic disease*. Summary. Washington, D.C.: National Academies Press, 2010.

The effort required to understand a single biomarker becomes many times more complex when the interrelationships of multiple biomarkers are considered. Fortunately, changes in computing and measurements are making such an approach increasingly feasible. The result of ongoing investigations such as Verily/Alphabet's Project Baseline[17] and the NIH's All of Us[18] will produce a vast array of complex biological data as well as context for how these data relates to more traditional clinical outcomes such as survival, major clinical events, and quality of life. Figure 3 provides a visual representation of why these relationships are so complex and intertwined.[2]

Digital biomarkers

One rapidly developing frontier is the field of digital biomarkers.[19] Sensors and personal devices now enable rapid and continuous assimilation of information about a person that provides insight into complex measures such as psychological state, exercise level, cognitive abilities, eating patterns, motion, and tremor. Because these data are in large part derived from new sources including smartphones and wearable electronic devices and facilitated by novel technologies that allow for the streaming and storage of complex data, standards for evaluating these biomarkers are just now developing. Although the Clinical Trials Transformation Initiative has recently published recommendations on standards for quality in the field,[20] a great deal of additional study is needed to link digital phenotypes and endpoints to traditional outcome measures. For example, the 6-min walk test has become a standard method for assessing exercise tolerance, and seated resting systolic blood pressure has become the standard measure for blood pressure assessment. But the relationship between the patient's activity status and measurements derived from wearable accelerometers, including ones embedded in wristwatches or cellphones, is a work in progress,[21] while sensors and smartphone apps for blood pressure measurement are likewise undergoing evolution.[22] Dealing with missing data, outlier values, and reduction of massive volumes of data into measures that can inform decisions will entail considerable work.

Ultimately, it is likely that digital biomarkers will open up entirely new measures of phenomena that are already used in practice. For example, it may be that total activity over the course of the day or some composite of peak activity and continuous activity would be a better measure to predict onset of new diseases (*risk/susceptibility biomarker*), prognosis for those who already have a disease (*prognostic biomarker)* or response to treatment (*response biomarker*). Similarly, it is likely that when very frequent blood pressure measurements are possible, derivative measures from the array of blood pressures and activities will be a better indicator of response to therapy for hypertension than seated resting blood pressure measurement.

Predictive toxicology and systems pharmacology

For all the reasons described in this section, individual biomarkers cannot be considered the primary goal of biological discovery or therapeutic development. In particular, understanding the effect of an intervention or exposure will develop directly as a function of understanding its complex ramifications in biological systems.[23] For the most part, early evaluation of therapies has involved a relatively static set of assays. Yet it is well known that extrapolations from animal models to human biology have often been unreliable.[24] However, dramatic and continuing reductions in the cost of measurement of genetic, genomic, and integrated biological measures[25] and the expansion of computing and analytical power are increasingly conferring the ability to look beyond the specific mechanism of action of a technology. A tremendous amount of validation and confirmation of increasingly complex models will be needed, but ultimately this work should enable much more effective prediction of an intervention's impact on integrative biology and clinical outcomes.

Quality and reproducibility of the underlying science

The benefit of using a biomarker for a specific purpose is directly related to the quality of the research supporting it. All too often, the basic research underlying assessment of a biomarker for a specific context of use cannot be reproduced. This lack of reproducibility recently presented a major problem in the regulatory evaluation of a new

treatment for Duchenne muscular dystrophy when assessment of the key biomarker was not done in rigorous, blinded fashion.[26] New policies implemented at the NIH are improving commitment to rigorous methodology, transparency, and reproducibility.

The importance of working together across the ecosystem

If we are to make needed progress in the proper application of these definitions so that medical product development and environmental policy can improve, academics, industry, and trial sponsors must expand their horizons to encompass new methods and approaches. Despite the best efforts of all involved, there is a tendency to continue to use the same methods over time in regulated studies because of a shared level of comfort with the use of well-worn measures. The discipline of regulatory science is the common ground for all elements of the ecosystem to come together to advance the field.[27] By continuing to evolve in our current thinking about biomarkers, endpoints, and other tools in medical product development, we will accelerate our understanding of biological science and improve the efficiency and pace at which effective technologies are developed for the prevention, diagnosis, and treatment of disease.

Summary

Biomarkers are critical to the fabric of discovery science, medical product development, and healthcare for the individual and population. Recent and ongoing explosive growth in measurement, computation, and analysis are producing rapid change in the field. The NIH and FDA have worked together to create a set of definitions that should guide researchers in developing needed evidence and practitioners in the application of biomarkers in health care, while other organizations such as the Clinical Trials Transformation Initiative and the Foundation for the National Institutes of Health Biomarkers Consortium are following suit in extending this work. An approach to biomarker development that incorporates collaborative regulatory science involving multiple disciplines is needed to ensure that rational, evidence-based biomarker development keeps pace with scientific and clinical need.

DECLARATION OF CONFLICTING INTERESTS

Dr. Califf was the Commissioner of Food and Drugs, US Food and Drug Administration from February 2016 to January 2017. He currently receives consulting payments from Merck and is employed as a scientific advisor by Verily Life Sciences (Alphabet).

FUNDING

The following statements relate to relationships which ended in February 2015 when Dr. Califf was appointed to the FDA as Deputy Commissioner for Medical Products and Tobacco. The current disclosures of note are those listed in the above declaration of conflicting interests. Califf received research grant funding from the Patient-Centered Outcomes Research Institute, the National Institutes of Health, the US Food and Drug Administration, Amylin, and Eli Lilly and Company; research grants and consulting payments from Bristol-Myers Squibb, Janssen Research and Development, Merck, and Novartis; consulting payments from Amgen, Bayer Healthcare, BMEB Services, Genentech, GlaxoSmithKline, Heart.org – Daiichi Sankyo, Kowa, Les Laboratoires Servier, Medscape/Heart.org, Regado, and Roche; he also held equity in N30 Pharma and Portola.

REFERENCES

1. Robb MA, McInnes PM, Califf RM. Biomarkers and surrogate endpoints: developing common terminology and definitions. *JAMA* 2016;**315**:1107–8
2. Institute of Medicine. *Evaluation of biomarkers and surrogate endpoints in chronic disease*. Washington, D.C.: National Academies Press, 2010, www.nationalacademies.org/hmd/Reports/2010/Evaluation-of-Biomarkers-and-Surrogate-Endpoints-in-Chronic-Disease.aspx (accessed 22 September 2017)
3. FDA-NIH Biomarker Working Group. BEST (Biomarkers, EndpointS, and other Tools) Resource. Silver Spring (MD): Food and Drug Administration (US); Bethesda (MD): National Institutes of Health (US), www.ncbi.nlm.nih.gov/books/NBK326791/ (2016, accessed 22 September 2017)
4. U.S. Food and Drug Administration. Fast track, breakthrough therapy, accelerated approval, priority review. Updated September 14, 2015, www.fda.gov/forpatients/approvals/fast/ucm20041766.htm (accessed 27 September 2017)
5. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 2007;**115**:928–35
6. Pencina MJ, Demler OV. Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvement for normal variables and nested models. *Stat Med* 2012;**31**:101–13
7. Coller BS, Califf RM. Traversing the valley of death: a guide to assessing prospects for translational success. *Sci Transl Med* 2009;**1**:10cm9
8. The Action to Control Cardiovascular Risk in Diabetes Study Group, Gerstein HC, Miller ME, Byington RP, Goff DC Jr, Bigger JT, Buse JB, Cushman WC, Genuth S, Ismail-Beigi F, Grimm RH Jr, Probstfield JL, Simons-Morton DG, Friedewald WT. Effects of intensive glucose lowering in type 2 diabetes. *N Engl J Med* 2008;**358**:2545–59
9. The SPRINT Research Group, Wright JT Jr, Williamson JD, Whelton PK, Snyder JK, Sink KM, Rocco MV, Reboussin DM, Rahman M, Oparil S, Lewis CE, Kimmel PL, Johnson KC, Goff DC Jr, Fine LJ, Cutler JA, Cushman WC, Cheung AK, Ambrosius WT. A randomized trial of intensive versus standard blood-pressure control. *N Engl J Med* 2015;**373**:2103–16
10. Writing C, Lloyd-Jones DM, Morris PB, Ballantyne CM, Birtcher KK, Daly DD, Jr, DePalma SM, Minissian MB, Orringer CE, Smith SC. Jr., 2017 Focused update of the 2016 ACC expert consensus decision pathway on the role of non-statin therapies for LDL-cholesterol lowering in the management of atherosclerotic cardiovascular disease risk: a report of the American College Of Cardiology task force on expert consensus decision pathways. *J Am Coll Cardiol* 2017;**70**: 1785–1822
11. Antman EM, Loscalzo J. Precision medicine in cardiology. *Nat Rev Cardiol* 2016;**13**:591–602
12. US Food and Drug Administration. Draft guidance for industry: enrichment strategies for clinical trials to support approval of human drugs and biological products. December 2012, www.fda.gov/downloads/drugs/guidancecomplianceregulatoryinformation/guidances/ucm332181.pdf (accessed 27 September 2017)
13. Collins R, Reith C, Emberson J, Armitage J, Baigent C, Blackwell L, Blumenthal R, Danesh J, Smith GD, DeMets D, Evans S, Law M, MacMahon S, Martin S, Neal B, Poulter N, Preiss D, Ridker P, Roberts I, Rodgers A, Sandercock P, Schulz K, Sever P, Simes J,

Smeeth L, Wald N, Yusuf S, Peto R. Interpretation of the evidence for the efficacy and safety of statin therapy. *Lancet* 2016;**388**:2532–61
14. Sager PT, Gintant G, Turner JR, Pettit S, Stockbridge N. Rechanneling the cardiac proarrhythmia safety paradigm: a meeting report from the Cardiac Safety Research Consortium. *Am Heart J* 2014;**167**:292–300
15. Fleming TR, DeMets DL. Surrogate endpoints in clinical trials: are we being misled? *Ann Intern Med* 1996;**125**:605–13
16. Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. *Stat Med* 1989;**8**:431–40
17. Verily Life Sciences. Project baseline, www.projectbaseline.com/ (accessed 27 September 2017)
18. National Institutes of Health. National Institutes of Health All of Us Research Project, https://allofus.nih.gov/ (accessed 27 September 2017)
19. Insel T. Digital phenotyping: technology for a new science of behavior. *JAMA* 2017;**318**:1215–6
20. Clinical Trials Transformation Initiative. CTTI Recommendations: developing novel endpoints generated by mobile technology for use in clinical trials, www.ctti-clinicaltrials.org/files/novelendpoints-recs.pdf (accessed 27 September 2017)
21. Yap J, Lim FY, Gao F, Teo LL, Lam CS, Yeo KK. Correlation of the New York Heart Association classification and the 6-minute walk distance: a systematic review. *Clin Cardiol* 2015;**38**:621–8
22. Plante TB, Urrea B, MacFarlane ZT, Blumenthal RS, Miller ER, 3rd, Appel LJ, Martin SS. Validation of the instant blood pressure smartphone app. *JAMA Intern Med* 2016;**176**:700–2
23. Antman E, Weiss S, Loscalzo J. Systems pharmacology, pharmacogenetics, and clinical trial design in network medicine. *Wiley Interdiscip Rev Syst Biol Med* 2012;**4**:367–83
24. Engber D. The mouse trap. Slate.com, www.slate.com/articles/health_and_science/the_mouse_trap.html (accessed 27 September 2017)
25. National Human Genome Research Institute. The cost of sequencing a human genome. July 6, 2016, www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/ (accessed 27 September 2017)
26. US Food and Drug Administration. Center for Drug Evaluation and Research. Summary Review. Scientific dispute regarding accelerated approval of Sarepta Therapeutics' eteplirsen (NDA 206488) – Commissioner's Decision. September 16, 2016, www.accessdata.fda.gov/drugsatfda_docs/nda/2016/206488_summary%20review_Redacted.pdf (accessed 27 September 2017)
27. US Food and Drug Administration website. Advancing regulatory science. Updated August 9, 2017, www.fda.gov/scienceresearch/specialtopics/regulatoryscience/default.htm (accessed 9 December 2017).