Adding Rigor to Biomarker Evaluations—EDRN Experience

Ziding Feng and Margaret S. Pepe

ABSTRACT

The cancer early-detection biomarker field was, compared with the therapeutic arena, in its infancy when the Early Detection Research Network (EDRN) was initiated in 2000. The EDRN has played a crucial role in changing the culture and the ways people conduct biomarker studies. The EDRN proposed biomarker developmental guidelines and biomarker pivotal trial study design standards, created biomarker reference sets and functioned as an unbiased broker for the field, implemented the most rigorous blinding policy in the biomarker field, developed an array of statistical and computational tools for early-detection biomarker evaluations, and developed a multidisciplinary team-science approach. We reviewed these contributions made by the EDRN and their impacts on maturing the field. Future challenges and opportunities in cancer early-detection biomarker translational research are discussed, particularly in strengthening biomarker discovery pipeline and conducting more efficient biomarker validation studies.

See all articles in this CEBP Focus section, “NCI Early Detection Research Network: Making Cancer Detection Possible.”

First 10 Years of the EDRN

Status of the field at 2000

When the EDRN was initiated in 2000, the field of cancer biomarker research, particularly for cancer early detection, was in its infancy compared with cancer therapeutic drug development in scientific rigor and sophistication. This is understandable because the non-invasive detection of cancer at an asymptomatic early stage is a very difficult problem: Lack of an enabling technology for detecting low cancer signal and the prohibitive cost of conducting prospective early-detection studies due to low incidence of cancer in an asymptomatic population. As a consequence, investigators in the field did not have much experience in designing and executing good early-detection studies. In contrast, the cancer therapeutic field had already had several decades of development in methodology for study design and evaluation, and accumulated rich experience from thousands of therapeutic trials, sponsored by drug companies or NIH-funded clinical trial groups. Researchers have a widely accepted guideline for moving a new drug target through the developmental pipeline, from phase 1 to 3 for FDA approval and phase 4 after market surveillance. A culture has been established and matured in conducting these trials. The drug evaluators and applied their favorite technology to search for biomarkers. Non-reproducible but highly acclaimed findings were rampant, but few successful translational products made it to clinical use.

The EDRN’s Contributions

Concept of “an arm’s-length away”

The first EDRN validation study was motivated by a high-profile article using MSA to detect bladder cancer (1) and an FDA registry trial was proposed by investigators to validate this test. There were debates within the EDRN on how to review the proposal and conduct this study. Clinical trialists and population scientists thought the validation study should be coordinated by an independent team and the original discovery group should be “arm’s-length away” whereas the discovery group thought they could carry out the multicenter validation study by themselves. This was due to a culture clash, and we were all inexperienced. As the first try-and-see case, this study had two coordinating centers, one at the discovery group and one at the EDRN coordinating center. Learning from this trial, the EDRN subsequently established the “an arm’s-length-away” policy: The discovery laboratory should be hands-off for validation, providing expertise but not driving the trial. Discovery laboratory biomarkers should be transferred to an EDRN Biomarker Reference Laboratory (BRL) to show that the test is transportable. The validation study should be coordinated by the EDRN Data Management and Coordinating Center (DMCC) that does not have vested interest in the test, to ensure an unbiased evaluation. That policy and the team science approach have been used for all subsequent validation studies and became an established culture.

Concept of five phases

Shortly after the EDRN was funded, a high-profile article was published (2) by an investigator outside of the EDRN on SELDI profiling to detect ovarian cancer, and consequently a clinical trial was planned to prospectively apply SELDI profiling to asymptomatic women to detect ovarian cancer. The EDRN, as the first and sole national consortium in this area, was under pressure to deliver something similar with high clinical impact. In response, EDRN developed a five-phase biomarker development guideline (3). The idea is to put the development of early-detection biomarkers to the same scientific rigor as therapeutic trials in terms of judging its...
readiness, providing a path toward clinical use, and specifying the objective and study characteristics of each phase (Fig. 1). This helps investigators by preventing premature jumps, and develops a road map to drive their biomarkers to clinical use. By applying the five-phase guideline to the SELDI profiling for ovarian cancer study, it was clear that they had simply completed a phase 1 study using case-control study design, and certainly were not ready for a phase 5 randomized clinical trial for morality benefit. An EDRN laboratory also published several articles using SELDI profiling for prostate cancer diagnosis (4), and these were also judged as only phase 1 studies. The EDRN team designed a phase 2 study to evaluate its performance in independent phase 2 study design. These studies demonstrated that SELDI profiling did not distinguish prostate cancer from non-cancer patients (5), and no further investigation was warranted.

Concept of blinding and randomization

Blinding and randomization are established practices in therapeutic trials. Many studies in the biomarker field did not use rigorous blinding, so bias arose from manipulating the specimens or data. It was a common mistake that a study would measure all cancer cases in one plate/day and health controls in another plate/day. The EDRN developed a rigorous blinding and randomization policy. All specimen aliquots were labeled in a blinded and randomized fashion so that one cannot tell whether two aliquots are from the same subject, as only the DMCC holds the key. Because the EDRN validation studies or reference sets have enough specimens to validate multiple biomarkers, the EDRN also has a policy that blinding will be maintained even after a laboratory has submitted the assay data to the DMCC, as long as there is still a specimen in the repository. The laboratory only receives the summary measures of its biomarker performance, and does not receive clinical data at the subject level. This policy is more strict than many biorepositories, and there is still internal debate whether this is too strict and does not help the laboratory to learn from any false calls.

The reasons for holding this strict blinding policy include that EDRN validation studies and reference sets are used as an unbiased broker for the field to validate biomarkers, not for discovery work. As some laboratories have leftover specimens, unblinding the data would unblind the remaining specimens, and the EDRN’s role as an unbiased broker would be compromised. For these reasons, the better strategy is to expand the reference sets that could allocate some specimens for discovery work and not to unblind the specimens reserved for validation.

In EDRN studies, we found in multiple incidences that compromised randomization led to bias. One EDRN-coordinated discovery study used specimens from another biorepository. The DMCC generated randomized specimen labels, but later we found an obvious bias in the data submitted from one of the participating laboratories, as the protein profiling had discriminating performance in almost all protein mass spectrum points. A discussion with the laboratory revealed that, in addition to the label the DMCC provided, there was another label on each aliquot and the laboratory technician used the numbers on that label for the mass spectrum run order. That label was provided by the repository and was not randomized. As a consequence, all cancer cases were profiled first, followed by all controls. This example illustrated that even when a laboratory is blinded, a failure in randomization can still have fatal consequence.

Concept of reference sets

Good biomarker discovery and validation studies need high-quality specimens. There are many biomarkers reported to have good performance, and conducting validation studies for all of them is not feasible. EDRN established 20 reference sets that are open to the whole field (6). Each reference set was designed with a specific clinical application in mind, and most of them were prospectively collected. Any laboratory can apply for access by submitting its preliminary data as evidence. If it is approved, the laboratory is provided a subset called a pre-validation set. If the data on this subset are promising, the laboratory receives the remaining set. This dual access approach allows a lower bar for the initial entry, and therefore more biomarkers can be tested. Every EDRN validation study collects more specimens than required for assaying the primary biomarker, and the remainder forms a new reference set. EDRN reference sets represent the most comprehensive biorepository in the cancer early detection and diagnosis arena, and its applicants are worldwide. EDRN really serves as an unbiased broker for the field.

Concept of study design standard for early detection biomarkers

To address the issue of many unreproducible biomarkers in the field, the EDRN proposed standards for biomarker study design, Prospective sample collection–Retrospective-Blinded Evaluation, or PRObE, standards (7). PRObE consists of four components: Clinical context, performance criteria, biomarker, and sample size/power (Fig. 2). It emphasizes that the clinical context should drive all other components for the study design. It provides a checklist for biomarker researchers when they design their studies. The standards are for biomarker pivotal trials, but its principles apply to biomarker discovery and early-stage validation as well. Because prospective sample collection is expensive, many biomarker discovery and early-stage validation studies could not be fully PRObE compliant. However, understanding the principles allows investigators to critically examine the potential biases due to the violation of the principles, take measures to address them, and not to overstate the study findings or push for premature clinical use.

Second 10 Years of the EDRN

During the first 10 years, the EDRN established its culture of team science approach, developed a method of systematically and rigorously

<table>
<thead>
<tr>
<th>Preclinical exploratory</th>
<th>PHASE 1</th>
<th>Promising directions identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clinical assay and validation</td>
<td>PHASE 2</td>
<td>Clinical assay detects established disease</td>
</tr>
<tr>
<td>Retrospective longitudinal</td>
<td>PHASE 3</td>
<td>Biomarker detects preclinical disease and a “screen positive” rule is defined</td>
</tr>
<tr>
<td>Prospective screening</td>
<td>PHASE 4</td>
<td>Extent and characteristics of disease detected by the test and false-referral rate are identified</td>
</tr>
<tr>
<td>Cancer control</td>
<td>PHASE 5</td>
<td>Impact of screening on reducing the burden of disease on the population is quantified</td>
</tr>
</tbody>
</table>

Figure 1. Phases of biomarker development. Adapted from Pepe et al. J Natl Cancer Inst 2001;93:1054–61.
evaluating a biomarker, and built an infrastructure for biomarker validation studies. The EDRN organ collaborative groups are where the team science is in action. They identify lower-hanging fruits, establish a long-term goal, and map out strategies. The EDRN is also aware that, to accomplish its set mission, it is necessary to work with diagnostic industry and leverage on their resources, particularly in the development of clinical assays and pushing a biomarker through FDA approval. Most FDA-approved biomarkers since the EDRN’s inception have EDRN’s print on them, in that EDRN worked collaboratively with industry and helped the translational process toward FDA approval. Here, we illustrated the PCA3 case as an example.

PCA3 is a urine-based RNA biomarker for assessing prostate cancer risk. The EDRN genitourinary (GU) group has identified three clinical contexts as its long-term goals: Biomarkers to help men to answer the questions “Should I have a prostate biopsy?”, “What if the biopsy is negative?”, and “What if the biopsy is positive for prostate cancer?”. The first two clinical contexts are lower-hanging fruits because they are for predicting biopsy outcomes, and each year many men have biopsies so the studies are relatively easy to complete. EDRN also has the T2-ERG biomarker (8) in the pipeline, so the urine samples collected for PCA3 are also suitable for T2-ERG assay. The third clinical context is very important as it should predict whether a newly diagnosed prostate cancer is indolent or lethal. It requires many years of follow-up to get the prognosis outcome. To address the first two questions, the EDRN GU group collaborated with Gen-Probe. Gen-Probe provided assay kits to the EDRN BRL and performed quality checks on 10% of the study specimens. They also provided their preliminary study data for the EDRN to design the validation study. The EDRN investigators helped Gen-Probe design their FDA trial and participated in discussions with the FDA. The EDRN PCA3 trial was a success that demonstrated that PCA3 has value in assisting with prostate biopsy decisions (9). The FDA approval for PCA3 is also for the same clinical context.

The T2-ERG biomarker discovered by the EDRN investigators was licensed to Gen-Probe, and they developed a clinical assay. The EDRN GU group proposed a combination rule, developed by an independent cohort, to combine PCA3, T2-ERG, and PSA to address the first clinical context, that is, to reduce unnecessary initial biopsies but not...
miss high-grade prostate cancer. Because the PCA3 reference set was not unblinded after the PCA3 trial, this 3-biomarker combination rule could be readily validated using the PCA3 trial data without compromising its validity. Gem-Probe provided assay kits to the EDRN BRL and performed quality checks on 10% of the study samples. The study demonstrated that a simple "OR" combination rule that can be easily implemented in clinical practice can reduce unnecessary biopsy by 42% while maintaining 93% sensitivity for detecting high-grade prostate cancer (Gleason Score ≥ 7; ref. 10). More biomarkers are currently under validation using the PCA3 reference set samples to evaluate their clinical utilities for assisting biopsy decisions.

Sharpening Statistical and Computational Tools

EDRN promotes a team science approach where laboratory scientists, clinicians, and population scientists (epidemiologists, biostatisticians, bioinformaticians) all contribute and learn from each other. The EDRN, specifically in the first 3 cycles of its RFA, supported statistical methodology development research that is relevant to the EDRN mission. That vision certainly bore fruits. In addition to the five-phase and ProBE guidelines, biostatisticians at the DMCC have developed many statistical methods that were motivated by the needs of the EDRN and that also have broader applications in the biomarker field, several of which are described below.

Sequential design for efficient use of specimens

The paucity of high-quality specimens, especially prospectively collected specimens, has been the bottleneck for the cancer early-detection biomarker field due to the low incidence of cancer. The high FDR of biomarkers leads many highly acclaimed biomarkers to seek validation, but most of them will fail in a rigorous validation study. In order not to waste precious specimens, EDRN has established a review process and the policy of two-tiered access to reference sets described above, that is, an approved biomarker will be evaluated on a subset of the reference set and granted access to the remaining set if the results hold up. This approach is similar in spirit to the group sequential trials in therapeutic trials, and a similar challenge is how to incorporate this sequential decision feature in the study design and data analysis to ensure unbiased reporting of the performance and to maximize efficiency. If one simply combines data from two subsets and reports the biomarker performance, the performance measure is biased upward. For example, if the true sensitivity is 0.6 while we approve access to the remaining set if the observed sensitivity on the first subset is at least 0.7, then the observed sensitivity due to randomness could likely be, in this instance, between 0.5 and 0.7. Due to the interim analysis decision rule, we will only have the chance to see the data from the second set if the observed sensitivity from the first set is at least 0.7. Therefore, the final analysis has to take into account the interim decision rule otherwise the sensitivity from the combined data will be biased upward. The EDRN investigators developed an array of statistical methods for group sequential design (GSD) for binary outcome (sensitivity and specificity; ref. 11), sensitivity at the prespecified specificity (ROC(t); ref. 12), logistic regression (13, 14), the optimal ways to rotate and select initial subset according to available specimens (15). These tools facilitate better use of the reference sets and maintain rigor in study reports.

Strategy to select candidates

Biomarker discovery studies often face the challenge of selecting a small subset of promising candidates from among a large number of candidates for next-stage investigation. The most common selection criteria are fold change, t statistics or Wilcoxon statistics. Cancer screening often requires very high specificity. If this unique feature is not considered in candidate selection, the chosen biomarkers may not have the required specificity for cancer screening. The EDRN developed the biomarker selection and prioritization strategy using partial area under the ROC curve (pAUC; ref. 16). Biomarkers that have decent sensitivity in a very high specificity region will rank high by pAUC criterion. We also developed a method to combine biomarkers that maximizes pAUC (17).

The concept of high pAUC also has biological appeal when combining biomarkers. Cancers are heterogeneous with several histological subtypes and many more molecular subtypes. Their carcinogenic pathways are different, and the representative biomarker profiles are likely different. If a biomarker represents a unique molecular/pathway subtype not shared by other subtypes, consisting of, for example, 30% of tumors for that cancer site, then the diagnostic performance of this biomarker, if we can measure it perfectly, will have sensitivity 30% at 100% specificity, therefore a large pAUC in high-sensitivity region. Not only do we want to put high priority on selecting these kinds of biomarkers, we also want to combine such biomarkers using an "OR" rule, that is, the test is positive if either biomarker A or B (or C, etc.) is positive, where A, B, . . . are such highly specific biomarkers. We advocated identifying these biomarkers and then using an OR rule to form the decision rule (18). Several EDRN applications used this approach (10).

Often discovery laboratories could not find adequate sensitivity and specificity for all these types of specific biomarkers, and some biomarkers are indeed signaled across all or at least several cancer subtypes. In this case, a sensible approach extending the above idea is first to look for some biomarkers that are well motivated by their roles in the carcinogenic process and which, with high thresholds, could pick up some cancers without false-positive calls. Then, for the remaining cancers, a linear or other combination rule is developed to identify them with acceptable false-positive rate. This is a combination of "OR" rule and regression models. We have used this approach in a pancreatic cancer biomarker study (19). This approach deserves more investigation, such as how to set up the thresholds for each component of the OR rule and simultaneously determine the linear combination rule.

Use longitudinal trajectory information

The idea of using longitudinal biomarker information to improve the early detection of cancer is natural and biologically sound. Each person has his/her biomarker trajectory with random variation over time. If a cancer biomarker has a relatively small variation over time within a subject as compared with the variation across subjects, then when serial biomarker measurements accumulate, each subject has his/her trajectory functioned as his/her own reference control value. When a cancer-induced change point occurs, this change is compared to the subject’s reference trajectory and is easy to identify because the between-subject variation is filtered out. Though conceptually simple, the implementation is not easy. For example, without a sufficient number of serial measurements, a jump due to noise could be misjudged as a cancer signal. A more natural way is to combine information of the biomarker distributions from the population when there are few serial measurements within a subject, and gradually increase the weight of the information from an individual trajectory when the serial measurements accumulate. EDRN investigators have been pioneers in this area both in developing appropriate methodology (20, 21) and in applying this to cancer early
detection studies (22, 23). For a single biomarker trajectory, EDRN investigators developed two approaches, a full Bayesian change point model (FBM; ref. 20) and a parametric empirical Bayesian model (PEB; ref. 21). FBM is a full Bayesian change point model that incorporates all information accumulated to date, both within and across individuals, to provide a posterior probability that a change point indeed happened. The threshold of this posterior probability can be set at the required specificity, often between 90% and 99%, to make a call. The PEB does not use cancer case data when it makes a decision rule. It simply calculates the posterior mean and variance for each individual biomarker trajectory, after a transformation to improve normality and standardization. That mean and variance estimate is a weighted combination of this subject’s mean and variance and the mean and variance for the study population. Therefore, when there are few serial measurements, more weight is put on the study population mean and variance when computing the PEB estimate of the subject mean and variance, that is, the estimate is shrinking towards the population mean. As serial measurements accumulate, more weight is put on the subject mean and variance. The threshold is chosen using the α% one-sided tail probability of standard normal distribution where α is the targeted false-positive rate. FBM has more power when the modeling assumptions (e.g., prior distributions used) are reasonable but is less robust when the assumptions are not reasonable. PEB has the advantage of fewer required assumptions and does not need use of data from cancer cases (less training optimism). However, it could be less powerful if the signal arises at a slow but steady rate, because each time the PEB estimate of the mean could be slowly increasing but not enough to make a call. It is designed for settings where the tumor characteristic jump is adequately large and the trajectory arising is fast. In our investigations, we did not observe big performance differences between the two methods. Both methods were originally developed for a single biomarker trajectory. EDRN investigators extended the FBM to multiple biomarkers (24), and investigation is currently ongoing for multivariate PEB. The developed methods are motivated from and are used by EDRN studies. For example, the EDRN Hepatocellular carcinoma Early Detection Strategy cohort study will use trajectories of AFP, DCP, and AFP-L3 to improve the early detection of liver cancer in cirrhosis populations.

Predictiveness curve

It is crucial to understand biomarker behavior in risk stratification, sensitivity, specificity, PPV, and NPV for the full spectrum of the potential cutoffs to be used. With that knowledge, we can give a biomarker the best chance to have clinical utility, whether for public health policy decision-making or for a specific patient making care decisions. The EDRN investigators proposed a predictiveness curve (PC; refs. 25, 26) that could visually illustrate all of the above features and also developed a statistical inferential procedure (hypothesis testing and confidence interval) to evaluate PC. A PC plots absolute cancer risk against the percentile of the biomarker distribution (Fig. 3). Sensitivity and specificity can be plotted on the same figure not as ROC curves but as two curves vary across percentile of the biomarker. PC adds two important pieces of information to a figure over ROC curve: the absolute risk and the biomarker distributions in the population. PC provides more information for public health policy decision makers to evaluate the impact of an intervention policy. Before making a clinical decision, a patient may want to know the absolute risk he/she has related to the test results, in addition to the true- and false-positive rates. The PC has gained attention in the biostatistical field and has begun to see some applications in clinical articles (ref. 27; Fig. 3).

Looking into the Future: Challenges and Opportunities

The EDRN during the past 20 years has made significant impacts on cancer biomarker early detection and the biomarker field in general. It has established the guidelines, the standards, and the team science
culture that are necessary to raise the rigor of biomarker evaluation to that compatible with therapeutic trials. Its achievement is also confirmed by the fact that most of the FDA approved biomarkers in the past 15 years for cancer risk assessment and diagnosis have some association with the EDRN. This is remarkable given that the EDRN’s mission is not to effect commercialization by itself, but to accelerate the process. The commercialization step must be completed by the diagnostics industry. When a field is new and must be built from scratch, it is absolutely necessary to constantly have critical self-examination, reinvent as necessary, not be afraid of taking corrective actions, and respect and learn from other disciplines and work together. Sponsor’s vision is crucial. It is uncommon for an NIH translational research consortium to fund statistical methodology development for 3 cycles of funding, but the EDRN did. Consequently, the EDRN became an early detection and diagnostic statistical powerhouse, responsible for a large portion of the methods developed in this area since 2000.

The cancer early-detection biomarker field is still relatively new in terms of team science and systematic coordinated approach to evaluate biomarkers for their clinical utilities. The EDRN has made important contributions to bring maturity to the field, particularly in the area of biomarker validations. To speed progression and fulfill the NIH mission to translate biomarkers from bench to clinics, there are several challenges the EDRN needs to address in the future. The challenges described below are not exhaustive and only reflect our personal opinions based on our 20 years of EDRN experience.

**Strengthen discovery pipeline**

Biomarker discovery remains a weak link in the whole translational process. The lack of a robust pipeline with biomarkers ready for definitive validation is an undisputable fact. There are many reasons for this. First, the biomarker discovery process does not have a framework or standards for optimal discovery. A brilliant biological insight, a new technology innovation, or pure luck could lead to great new biomarker. However, lacking a framework, guidelines, and standards makes rapid progress difficult. This is the nature of the discovery game that we cannot alter. Second, some technologies are still not adequate for measuring low cancer signals at a tumor’s early stage, particularly for measuring proteins. Technological progress may change this in the near future. Third, the difficulty of and prohibitive cost to secure high-quality specimens collected in the right clinical context, for example, at asymptomatic stage for early detection, make it impossible to use optimal specimens for all discovery work. This is where the EDRN can make a difference. Using the wrong specimen has many ramifications that bias discovery. Many acute reactant signals present at the symptomatic stage but not at the early stage, so using specimens collected at the symptomatic stage may lead a discovery lab to select these biomarkers and later see that signal disappear in validation studies. Retrospectively collected samples, that is, after knowing disease status, often have wired-in biases due to the systematic differences between cases and controls based on how they are selected and the settings and the timing of the blood draw and processing. This is in contrast to a prospective early detection study where the disease status is unknown at the time of specimen collection, so all future cancer cases and controls selected go through sample collection and processing standard–operating procedures (SOP) the same way, therefore avoiding wired-in bias. This crucial point has not been adequately appreciated by many biomarker researchers. For example, many biomarker researchers think this kind of bias will be eliminated by using samples from early-stage cancer at the time of diagnosis. Although this approach has merit, it is still possible that aggressive cancer with rapid development may be more likely to be diagnosed at an advanced stage. Biomarker discovery laboratories do not have resources nor expertise to design and conduct large prospective studies. We propose several ways to strengthen the discovery pipeline.

First, biomarker discovery laboratories should be part of multidisciplinary team comprising biological, clinical, and population science expertise. That team with inputs from clinicians will identify unmet clinical needs. Biomarker discovery laboratories provide insights on which biomarker candidates are promising and the challenges in moving them through the translational pipeline. The team develops a road map toward biomarker translational path to meet these needs. Population scientists help design a series of studies to triage and evaluate these candidates. This team approach will give biomarker discovery labs a better chance at developing high-quality biomarkers for discovery and early triage investigations.

Second, a team approach is more powerful when requesting access to large cohorts and biorepositories that have appropriate samples for early detection. Large cohorts should change the culture that only grants access to their samples for validation studies. To strengthen the biomarker discovery link, large cohorts should step in and help discovery work. Rigorously coordinated biomarker discovery teamwork will ensure the best use of these precious specimens.

Third, the team approach should ensure binding and randomization when possible during discovery process. In our experience many discovery laboratory personnel do not like to be blinded or for assay orders to be randomized, considering it an unnecessary headache at the discovery stage. We have empirical evidence that binding the assay laboratory personnel to disease status and randomizing the assay order for case and control samples on the plate/day will reduce the bias and variation, leading to more reproducible findings.

Fourth, the field needs to address the “death valley” in the translational process, that is, that discovery laboratories have neither resources nor incentives to develop clinical grade assays when the diagnostics industry does not want to invest resources to do so until they are convinced by performance data. However, research assays have larger analytical variation compared with clinical grade assays, therefore reducing the performance of a biomarker when measured on research assays. This catch-22 situation is very common in our initial validation studies using research assays. NIH has a social responsibility to fill this gap, and as the largest NIH-funded enterprise in the early-detection area, the EDRN has a responsibility to help find a solution. It is conceivable that a government–academic–industry partnership can be formed to coordinate efforts to prioritize biomarkers for clinical assay development and provide resources and incentives to execute these priorities.

**Conduct better and more efficient validation studies**

The EDRN does not have the resources to conduct many large prospective early-detection trials due to the low incidence of cancer even for high-risk populations. It would save resources and time if FDA registry trials could be done from one large prospective study for multiple biomarker tests for a single cancer site, or for multiple cancer sites. The PROBE design standards provide principles to do so. To implement these principles for multiple biomarker validations for a single cancer site, we must address four challenges. First, retrospective evaluation of biomarkers on prospectively collected samples requires demonstration that specimen storage does not have appreciable impact on biomarker performance. It is important to plan studies that address this concern early on because it takes time to address storage effect, for example, to show biomarker measures will not
change over a 5-year period would require that biomarkers be measured at sample collection and again 5 years later. Second, maintaining blinding for multiple biomarker studies over time using the same repository to ensure that early results do not influence the later trials. The biomarker holders after the biomarker validation study is completed often want to unblind the samples so they can learn and improve. That desire conflicts with the rigorous blinding requirements for later validation trials. Third, we should ensure the study population fits the population of intended clinical use. This is particularly challenging for pan-cancer endpoints because the screening populations may differ, for example, patients with cirrhosis for HCC or heavy smokers for lung cancer early detection. Finally, we must reduce the sample collection burden. Many biomarkers require a short time window, for example, 4-hours from blood draw to freeze. That makes it hard for broad implementation of prospective studies. If better collection kits can be developed that stabilize biomarkers under room temperature, that would greatly facilitate blood collection. Some argue that only those biomarkers that are very robust to specimen handling and freezethaw cycles are worth studying. The field has yet to reach a consensus on that, as one may argue that once a biomarker is shown to perform well, a better collection kit can be developed for it later. We need rigorous initial specimen collection SOPs to give this kind of biomarker a chance to perform.

New enabling statistical and computational tools

As new imaging modalities are emerging and their costs dropping, there is great need for statistical and computational tools to extract useful features from imaging and combine them with body fluid–based biomarkers to help clinical decision making. Images are high-dimensional in nature, so machine learning and artificial intelligence (AI) are useful tools. One challenge is the training sample size because unlike many commercial uses of AI, cancer early-detection studies usually do not have large training sample sizes that are required by machine learning and AI. Another challenge is the study design. After imaging is done, some interventions, such as biopsies, will be taken based on imaging findings. That could lead to verification bias, that is, cancer outcomes from patients with negative screening images are not ascertained until later when clinical symptoms have developed. Imaging when combined with blood collection also makes it hard to predict the lead time a blood-based biomarker can detect before clinical diagnosis because the clinical onset time is missing when imaging detects the lesion. Imaging also could detect a lesion that will never kill a patient if left undetected, that is, overdiagnosis and subsequent overtreatment. Statisticians will have great challenges and opportunities to develop innovative study designs that address these challenges when biomarkers and imaging are incorporated into one trial.

Another promising area is integrating high-dimensional data of different types. In the biomarker discovery stage, there are analytical tools to integrate pathway information to help people understand the relationships between biomarkers. Many biomarker discovery people hope statisticians/informaticians can take the complicated data and implement machine learning and AI to find a model that greatly improves their current single biomarker test or a small biomarker panel. We think the integration of various omics data and other phenotype data (imaging, epidemiologic, etc.) is likely more useful for providing insights to select and prioritize biomarkers, not as a black box for early-detection decision making. The final clinical decision rule is still likely to be based on a small number of predictors. Therefore, biologists should work more closely with quantitative scientists to extract useful information from massive data, not just simply hand over the data and expect them to solve your problems.

Authors’ Disclosures

No disclosures were reported.

Acknowledgments

The work was supported by the National Institutes of Health grant, U24 CA086386 (to Z. Feng).

Received October 25, 2020, revised October 25, 2020; accepted October 26, 2020, published first November 10, 2020.

References

18. Feng Z. Classification versus association models: should the same methods apply? Scan J Clin Lab Invest 2010;70:53–8.
Adding Rigor to Biomarker Evaluations—EDRN Experience

Ziding Feng and Margaret S. Pepe