Revisiting the Standard Blueprint for Biomarker Development
to Address Emerging Cancer Early Detection Technologies

Ruth Etzioni PhD\textsuperscript{1}, Roman Gulati MS\textsuperscript{1},
Christos Patriotis PhD\textsuperscript{2}, Carolyn Rutter PhD\textsuperscript{1}, Yingye Zheng, PhD\textsuperscript{1},
Sudhir Srivastava PhD\textsuperscript{2}, Ziding Feng PhD\textsuperscript{1}

Affiliations
\textsuperscript{1}Division of Public Health Sciences, Fred Hutchinson Cancer Center, Seattle, Washington, U.S.A.
\textsuperscript{2}Division of Cancer Prevention, National Cancer Institute, Bethesda, Maryland, U.S.A.

Correspondence
Ruth Etzioni, PhD, Division of Public Health Sciences, Fred Hutchinson Cancer Center, 1100 Fairview Ave N, M2-B230, Seattle, WA 98109-1024, USA. Tel: +1 206 6676561. Fax: +1 206 6677264. E-mail: retzioni@fredhutch.org.
Abstract

Novel liquid biopsy technologies are creating a watershed moment in cancer early detection. While evidence supporting population screening is nascent, a rush to market the new tests is prompting cancer early detection researchers to revisit the standard blueprint established by the Early Detection Research Network (EDRN) for evaluating novel screening biomarkers. In this commentary, we review the EDRN’s Phases of Biomarker Development (PBD) for rigorous evaluation of novel early detection biomarkers and discuss both hazards and opportunities involved in expedited evaluation. According to the PBD, for a biomarker-based test to be considered for population screening: (1) test sensitivity in a prospective screening setting must be adequate, (2) the shift to early curable stages must be meaningful, and (3) any stage shift must translate into clinically significant mortality benefit. In the past, determining mortality benefit has required lengthy randomized screening trials, but interest is growing in expedited trial designs with shorter-term endpoints. Whether and how best to use such endpoints in a manner that retains the rigor of the PBD remains to be determined. We discuss how computational disease modeling may be harnessed to learn about screening impact and meet the needs of the moment.
We are at a watershed moment in cancer early detection with novel liquid biopsy technologies creating the possibility to screen for multiple cancers with a simple blood test. Studies to date have shown that such multi-cancer early detection tests can identify cancers in persons already diagnosed,[1-3]. Industry messaging and public anticipation are driving a perception that widespread deployment of the tests is an urgent matter (e.g., [4]); indeed, several products are being marketed ahead of FDA approval while we await evidence of clinical utility, and a congressional bill has been introduced to allow Medicare coverage of the new tests once they have been approved and shown to be clinically effective.[5, 6] There is now growing concern among early detection researchers that the tests will become widely available before they have been properly vetted for benefit and harm.

In this commentary, we examine the road from demonstrating diagnostic performance of a new early detection biomarker to producing convincing evidence of its value for population screening. We explain why a measured, sequential approach has become standard and explore whether this can and should be adapted to meet the needs of the moment.

Cancer early detection has always been controversial because it involves intervening in a predominantly healthy population to benefit a relative few. A cancer screening test is not a simple matter; even if the test itself is not costly or harmful, it can lead to downstream interventions that are. In the last decade, research studies and the popular press focused far more on the harms of cancer screening than on its benefits. Newspaper articles warned about overdiagnosis and overtreatment,[7, 8] and research studies highlighted how the risk of unnecessary biopsies compounds dramatically under regular screening.[9]

We previously commented on the valuable lessons that the history of early detection holds for the field today.[10] The key lesson bears repeating: It is not just about the test’s diagnostic performance. Bridging preliminary evidence that a new screening test can detect cancer to establishing that the
population screening can materially reduce deaths in a sustainable fashion requires a whole sequence of pieces to fall into place.

First, the promising performance of a screening test in known cases and non-cases must lead to adequate discrimination in the prospectively screened (intended-use) population. In practice, degradation of screening test sensitivity in the prospective setting is expected because the case mix will be skewed towards cases that are earlier in their natural history and may include a fraction that are clinically insignificant.[11, 12] Moreover, in this setting, sensitivity will depend on what happens after a positive test; if accurate confirmation testing is not readily accessible, the sensitivity of the entire screening episode will be further degraded.[13]

Second, screening in the intended-use population must lead to material movement of the timing of diagnosis to an earlier, more treatable point in disease natural history. Since we most often conceptualize early versus late diagnosis in terms of clinical stage of disease, we generally refer to this condition as screening producing an adequate “stage shift”. But this can only happen if the cancer provides adequate opportunity – via a long enough early-stage duration – to be detected at this point. Thus, the natural history of the target cancer will factor critically into the ability of a screening program to change the fate of the cancers detected. Learning a cancer’s natural history, however, requires incidence data with and without screening, which is not available for most types of cancer. Whether a candidate test can achieve adequate stage shift is virtually impossible to establish for these cancers without prospective evaluation.

Third, any stage shift produced by screening should translate into an adequate and sustainable reduction in cancer mortality. Whether this will be the case depends on the cancer – the expected mortality reduction from a given stage shift is highly variable across cancers – and on the implementation of the screening program, including any subsequent diagnostic and treatment interventions. In practice, access to confirmation testing can be heterogeneous, and changes in available
treatments can impact the effectiveness of screening. Thus, a number of factors determine whether a promising stage shift will result in a clinically significant and sustainable mortality reduction.

Establishing that a new test meets these three requirements while also controlling adverse outcomes such as unnecessary biopsies and overdiagnosis involves a corresponding sequence of studies that typically take many years to complete. In 2001, the Early Detection Research Network of the National Cancer Institute established the Phases of Biomarker Development (PBD) to codify these studies and specify criteria for progression from one phase to the next.[14, 15]

The PBD blueprint begins with discovery (Phase 1) and assessment of the discriminative performance in known cancer cases and non-cases (Phase 2). It progresses to evaluation of pre-diagnosis performance using stored serum samples (Phase 3) and ascertainment of the change in incidence and stage induced by biomarker-based screening in prospective cohort studies (Phase 4). Phase 4 studies may also inform about screening test performance in the prospective setting, but simple empirical estimates of sensitivity from such studies are frequently overly optimistic.[16, 17] The final phase is the randomized screening trial with disease-specific mortality as a primary endpoint (Phase 5).

Screening trials not only avoid selection bias due to random assignment of screening; they also examine the collective contributions of screening, diagnostic confirmation testing, and treatment to delivering mortality benefit. The demonstration of significant mortality benefit in a randomized screening trial has become established as a condition for acceptability of a new test; the US Preventive Services Task Force (USPSTF) generally requires such evidence as a prerequisite to recommending population screening.

To a certain extent, the PBD sequence of studies is being followed with the new cancer tests. Several retrospective (Phase 2) and prospective (Phase 4) studies of test performance have been or are currently being conducted.[11, 18] Phase 3 studies are rare because of the specimen volume required by the new multi-cancer tests (e.g., 20 mL of blood for one product).[19] A recent Phase 3 study presented findings regarding detection rates up to three years prior to clinical diagnosis based on specimens from a
population cohort study which drew 36 mL of blood per participant.\cite{20} At this time, very few Phase 5 trials are in process, and test developers are making the case that trials with disease-specific mortality as primary endpoint are too lengthy, costly, and complex.\cite{4, 21}

Randomized screening trials are indeed lengthy, costly, and complex, particularly in the average-risk population. The rarity of disease-specific mortality in this population means that such trials must enroll a very large number of participants; further, the time required to observe enough of these events in an initially asymptomatic cohort means that long follow-up is generally needed. Given their expense and duration, randomized trials can generally only evaluate one or two screening strategies. Further, because diagnostic and treatment practices can continue to evolve, the trial results may be less relevant to contemporary practice when they finally become available. Given the sheer number of liquid biopsy-based tests currently under development, it is unrealistic to conduct a screening trial for each one. To this end, there is growing interest among cancer researchers in expediting screening test evaluation in a manner that gains efficiency but retains the intent of the PBD to rigorously assess the mortality benefit that may be reasonably expected from a novel cancer screening biomarker.\cite{22}

One approach that has been gaining attention with strong support from industry is to use short-term outcomes in screening trials for the new tests. A prime candidate is the reduction in the incidence of late-stage disease, which has been proposed as a surrogate or provisional endpoint for mortality benefit.\cite{4, 23} At this time there is no clear consensus for how this would be operationalized; for example, would conclusions about screening benefit be made based on this endpoint or would the stage shift be used to predict the reduction in disease-specific mortality? We strongly recommend against the first option. We have previously shown that a given reduction in late-stage incidence does not imply the same reduction in disease mortality across cancers; a seemingly dramatic reduction in late-stage incidence would be expected to produce a very modest mortality benefit for some cancers.\cite{24} In short, we do not know what might constitute a clinically significant reduction in late-stage incidence. And given
the challenges of de-Implementing cancer screening programs once they have been initiated, basing screening decisions on a seemingly favorable reduction that does not lead to a reasonable mortality benefit could be suboptimal for both clinical and policy purposes.

The second option – using the mortality reduction predicted by late-stage incidence rates an endpoint – may be more reasonable. Indeed, studies have previously proposed using this endpoint in breast and colorectal cancer screening trials; results suggested advantages over the mortality endpoint both in terms of timeliness and statistical power.[25, 26] However, this will still require further investigation into different ways to predict mortality benefit given a stage shift and their validity across cancers. For example, we have shown that accounting for prognostic subtype when substituting early-stage for late-stage survival among cases shifted from a late- to an early stage by screening may affect the predicted mortality reduction.[27]

Are there other approaches that might be harnessed to expedite evaluation of novel screening tests and produce evidence to support well-founded population screening decisions?

Real-world data has been suggested as an evidence source that might be useful as the tests disseminate in the clinic and produce data on their use and consequences in practice.[28] However, there are likely to be many challenges to the development of reliable evidence from real-world data beyond the primary challenge of ensuring that key variables are collected in a consistent manner in the clinical setting. First, selection bias in real-world data is a well-understood problem that cannot always be remedied. This will almost certainly be a major issue in the evaluation of benefits of novel early detection tests given their costs and other barriers to their use as well as to accurate confirmation testing. Long-term outcomes will take years to accumulate, and short-term outcomes will be subject to the same concerns as those pertaining to surrogate endpoints for screening trials. And data sharing issues will need to be addressed or alternatives to data sharing adopted to permit analysis of large, representative datasets. If real-world data can be made available, they will be most useful for
assessments of short-term outcomes: patterns and predictors of novel test utilization, access issues with respect to both screening and confirmation testing, and diagnostic performance in the intended-use setting.

Computational disease modeling is an established approach to learn from and extrapolate beyond the empirical results of screening studies. A computational model for studying cancer screening is a mathematical representation of the events in disease progression that drive screening outcomes, such as disease onset, metastasis, diagnosis, and survival. Once the rates of the relevant events have been estimated, the model permits projecting virtually (e.g., by simulation) the impact of screening and treatment on key clinical outcomes, such as late-stage incidence and mortality. Models have been used to expand the range of screening strategies beyond those studied in actual trials via simulated trials that examine a range of screening ages and intervals, biopsy referral criteria, and strategies tailored to disease risk.[29-36] In a sense, modeling has become an informal sixth phase of the PBD and has been relied on by the USPSTF and other national guidelines panels to inform their policies for lung, breast, cervical, and colorectal cancer screening.[37] The most recent USPSTF guidelines for both lung and colorectal cancer screening were directly informed by modeling. In the case of lung cancer, the modeling studies showed that changing the eligibility criteria from 30 to 20 pack-year smoking histories dramatically increased the fraction of the population eligible for screening and the life-years saved while only modestly increasing harms.[36] In the case of colorectal cancer, the modeling studies showed that strategies that started screening at age 45 increased life-years saved and yielded fewer colorectal cancer cases and deaths than similar strategies that started screening at age 50 or 55.[38] The new USPSTF lung cancer screening guidelines expand eligibility to 20 or more pack-year smoking histories[39] and the new colorectal cancer screening guidelines recommend beginning screening at age 45.[40]

While computational modeling cannot replace every screening trial, a calibrated model of the disease process – one that replicates observed results when simulating existing studies – is a powerful
tool. Indeed, a calibrated model permits rigorous and transparent projections that may increase efficiencies and even eliminate the need for some trials. Here we consider two ways in which modeling could be used to potentially expedite the evaluation of novel tests, but there are likely to be many more.

First, as we have already discussed, a model that has been calibrated to stage-specific incidence in a trial could be harnessed to predict the corresponding mortality reduction. A model that has been calibrated to incidence and mortality patterns could go further, predicting these outcomes beyond the trial duration. We previously used modeling to predict the long-term mortality benefit and ratio of overdiagnoses to lives saved under prostate cancer screening based on results obtained under limited follow-up from the European Randomized Study of Screening for Prostate Cancer.[41] While long-term empirical data do not always exist to validate such projections, their availability in two colorectal cancer screening studies permitted verification that their long-term results matched those projected by models.[42, 43] This application of models not only bridges from the trial to the policy setting, which necessarily requires quantifying outcomes over a long-term (ideally lifetime) horizon, but also opens the door to potentially shortening trial durations via judicious blending of empirical and model-based results. Prediction of the mortality reduction given the observed late-stage incidence in screen versus control groups, validated in one setting (e.g., annual testing) and used to anticipate effects in another setting (e.g., biennial testing), would be an example of such a blended model.

Second, a model of screening could, in principle, build from long-term studies of an older test to project outcomes of screening using a test with different performance characteristics. Modeling studies (e.g., [38]) have projected outcomes of newer stool-based tests for colorectal cancer by superimposing these tests, given their sensitivity, on existing models of colorectal cancer natural history that were calibrated to adenoma prevalence data and colorectal cancer incidence rates in the US.[44] This application of models could reduce the need for trials of novel tests when their performance in the
prospective screening has been well estimated, and calibrated models of screening for the tests’ target
cancers are available.

Naturally, modeling is subject to challenges and limitations; models require extensive, high-quality
data for adequate calibration, can be difficult to estimate even when such data are available,[45] and
often make unverifiable assumptions about natural history, screening performance, or screening
benefit. But the science of modeling has advanced over the last several decades, and methods for
mitigating these limitations have been developed and continue to evolve.[46] Notable advances include
metrics for independently developed models to examine validity of unobservable quantities (such as
sojourn times),[47, 48] efficient algorithms to calibrate models with potentially many parameters to
multiple data targets,[49] and methods to propagate uncertainty in model inputs to uncertainty in policy
preferences based on model outputs.[50] As data from studies of novel screening tests become
available, this will facilitate rigorous development of models for cancers with unknown natural histories
and provide new opportunities to validate existing natural history models.

In conclusion, marketing pitches for the new tests argue that lack of screening for many cancers has
created an urgency to deploy novel tests as quickly as possible. But the real urgency is driven by
concerns that the tests may be released to an unsuspecting public before researchers are able to
establish that they will do more good than harm. The early detection research community must now
address the conflict between the established blueprint of the PBD and the pressure to accelerate
evaluation of novel screening tests. Will we continue to require a significant reduction in disease
mortality from a randomized study to greenlight a new test? Or are we willing to expand the scope of
what constitutes adequate evidence to recommend such a test? If the latter course is followed, then
building a rigorous program of objective analytical and modeling studies to learn from and extend the
results of PBD studies would be a worthwhile investment.
Funding

This work was supported by the Rosalie and Harold Rea Brown Endowed Chair (to R.E.) and the National Cancer Institute (grant numbers R35CA274442 to R.E., R50CA221836 to R.G., and U24CA086368 to the EDRN DMCC).

Acknowledgments

The funder had no role in the design of the study; the collection, analysis, or interpretation of the data; or the writing of the manuscript and decision to submit it for publication.

Disclosures

The authors have no conflicts of interest to declare.

Data availability

No new data were generated or analyzed in this manuscript.

References


