Overview of Screening: Where We Are and Where We May Be Headed

Russell Harris*

* Correspondence to Dr. Russell Harris, Cecil G. Sheps Center for Health Services Research, 725 Martin Luther King, Campus Box 7590, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599-7590 (e-mail: russell_harris@med.unc.edu).

Accepted for publication February 23, 2011.

This volume of Epidemiologic Reviews continues a discussion about screening within the evidence community that has been going on for many years. From various perspectives, the authors of these reviews consider the benefits and harms of screening for multiple conditions; the balance between benefits and harms (and costs) is often not clear. With few exceptions, the contribution of screening to improving the health of the public is small, yet it has become a popular and growing form of prevention. It may be that we are learning that the magnitude of benefit from screening is less than we hoped, and the harms may be greater than we thought. Perhaps we should not think of screening as our primary prevention strategy but rather use screening to make a real, but limited contribution to population health. We might target screening to smaller subpopulations with the highest potential benefit and the lowest potential harm. The payoff for population health could be greater if we shifted some resources we now devote to screening to developing, testing, and implementing alternative approaches to preventing the important threats to population health. There needs to be a wider discussion about these issues with the public.

This volume: Discussions about evidence

The articles in this volume of Epidemiologic Reviews deal with the scientific complexity of screening. These are discussions among clinicians, academics, program leaders, and investigators—the “evidence community.” The discussions here review various types of evidence about several important examples of current screening controversies: screening

Abbreviations: HIV, human immunodeficiency virus; PPH, predictor of poor health; USPSTF, U.S. Preventive Services Task Force.

A recurring cartoon frame appears from time to time in The New Yorker magazine. In the frame is an older man with a long beard who sees a sign marked “Truth” pointing in one direction or another. There are a number of variations of this cartoon, but one version particularly appeals to me. The older man is walking on a treadmill—going fast but not gaining on “Truth.”

The idea of screening was first raised more than 150 years ago (1); screening programs, research, and debate have been going on since at least the 1950s. One might ask, with all of this effort, whether we are making any progress in finding the “truth” about screening. What have we learned about how to view the contribution of screening to individual and population health? Is it the once and future answer to many of our health problems, or is it a “half-way technology” (2) to sustain us for a while until we find a better approach?

The role of screening in individual and population health has grown dramatically over the past 50 years. Development of new and more sensitive tests, discoveries about the origins and precursors of disease, and new treatment approaches have led us to screen for more and more conditions, for larger and larger populations, and with lower and lower thresholds for abnormality. We have developed a strong interest in “risk stratification”—a form of screening—to find subgroups to target for more aggressive testing, follow-up, and treatment. When we are not making headway in reducing the burden of suffering from conditions such as human immunodeficiency virus (HIV) infection or diabetes or lung cancer, we often develop a screening strategy as a way to advance the health of the public. In many ways, screening has become our primary and our default prevention strategy.

THIS VOLUME: DISCUSSIONS ABOUT EVIDENCE

The articles in this volume of Epidemiologic Reviews deal with the scientific complexity of screening. These are discussions among clinicians, academics, program leaders, and investigators—the “evidence community.”
for colorectal cancer, breast cancer, prostate cancer, diabetes, HIV, and genetic conditions. Some of the articles deal with general screening issues that cross multiple conditions: the problems of predisease, of overdiagnosis, of risk prediction, of how to measure mortality benefit from screening in randomized controlled trials, of assessing the evidence for implementing screening programs, and of how best to implement an effective screening program to reach the greatest number of eligible people. Ideally, the outcome of these discussions would be agreement on the magnitude of benefits and the magnitude of harms of screening, with further agreement about the balance between the 2. From this agreement would come advice to decision makers (whether decisions about individuals or populations) about when and how to implement screening programs. Given the complexity of the evidence about screening, reaching consensus on what the advice should be is often challenging.

EXAMPLES OF THE DISCUSSIONS ABOUT EVIDENCE: HIV AND DIABETES

To illustrate these discussions about the benefits and harms of screening, it is helpful to consider the 2 domains of evidence: certainty and magnitude. In the discussion of HIV screening described by Chou (3), the Centers for Disease Control and Prevention is more certain than the U.S. Preventive Services Task Force (USPSTF) that screening reduces risky sexual behaviors and transmission of HIV. The disagreement over certainty regarding this important issue is partly due to a more critical appraisal of a meta-analysis by the USPSTF. Chou raises an important issue that may help explain the Centers for Disease Control and Prevention’s lower bar for certainty of evidence. It may well be that, because the Centers for Disease Control and Prevention sees itself as dealing with an urgent public health problem (i.e., the unchanging incidence of HIV infection), it is making a “best guess” recommendation given the existing evidence. There may be a feeling of needing to act, even if the degree of certainty that the action will be effective is not high.

The USPSTF, on the other hand, likely feels no such urgency; its allegiance is to the evidence, and it seeks to make “pretty darned sure” rather than “best guess” recommendations. The USPSTF is trying to minimize the number of positive recommendations for interventions that, later on, are found to be ineffective in improving population health. The USPSTF is acting on a “first, do no harm” principle. As a result, the USPSTF recommendation is targeted, whereas the Centers for Disease Control and Prevention recommendation is universal screening for people aged 13–64 years.

One may wonder, however, what effect either recommendation—or screening in general—will have on the burden of suffering of populations from HIV–acquired immunodeficiency syndrome (AIDS). A recent modeling study found that even greatly expanded screening for HIV would have only a minor effect on the HIV epidemic (4).

There are similarities with the discussion about screening for diabetes: the options again are a “best guess,” universal approach versus a “pretty darned sure,” targeted approach.

In this case, although the evidence about benefits and harms is only indirect, there is enough evidence to make a reasonable estimate of magnitude. The root of the diabetes screening disagreement comes down to differences in estimating the magnitude (rather than the certainty) of the benefits and harms, given our less-than-perfect evidence.

Although not explicitly discussed here, the American Diabetes Association recommends universal screening for people aged 45 years or older and for many people (e.g., overweight and African-American or physically inactive people) younger than age 45 years (5). The USPSTF, on the other hand, recommends targeting screening to people with hypertension (6). Echouffo-Tcheugui et al. (7) discuss the evidence gaps but conclude that there are likely benefits in reducing cardiovascular disease. A recent report at an international meeting from an important trial of intensive versus conventional treatment for people with screen-detected diabetes found no benefit in the primary composite outcome, although there was a small absolute reduction in nonfatal myocardial infarction (8). In absolute terms, the magnitude of health benefits appears to be small (9).

Echouffo-Tcheugui et al. (7) read the evidence as showing that the harms of diabetes screening are either small or nonexistent. Another reading leads to a different conclusion. The meta-analyses of tight glycemic control have found increased rates of hypoglycemia (9), a problem with potentially lasting effects (10). A short-term report from the same recently completed trial noted above (8) of intensive treatment for people with screen-detected diabetes found small increases in several negative psychological states (11). A second study of people with screen-detected diabetes found that psychological distress increased with time from diagnosis (12); a third large US cohort study found an increased risk of depression for people with diabetes compared with similar controls without diabetes (13). There is a case for estimating the absolute magnitude of harms from diabetes screening as being higher than has been appreciated.

Modeling studies of the cost-effectiveness of screening for diabetes have varied in their estimates of the magnitude of benefits and harms. Several studies have estimated both to be small in absolute health terms (14, 15). In this situation, changing the estimate of benefits or harms/costs by only a small degree can result in a much larger change in the cost-effectiveness ratio. While the cost-effectiveness studies differ in their findings about the cost-effectiveness ratio, several agree that screening for diabetes is an exercise in small effects—likely small benefits with (arguably) small harms. As in the case of HIV screening, it is unlikely that screening will have a major effect on the burden of suffering from diabetes.

THE CONTRIBUTION OF MODELING AND COST-EFFECTIVENESS ANALYSIS

As shown by the cases of HIV and diabetes screening, modeling and cost-effectiveness analyses are increasingly used to help with screening decisions, in situations of both low evidence certainty (for “best guess” analyses) and higher evidence certainty (for “pretty darned sure”
analyses). Lansdorp-Vogelaar et al. (16) review cost-effectiveness analyses of colorectal cancer screening, finding large variation in model assumptions about such important issues as the natural history of colonic adenomas and thus very different results about the effect of screening on colorectal cancer mortality. Although we have more evidence about the effects of colorectal cancer screening than for many other health problems, certainty is still low for critical variables. In addition to gaps in the certainty of evidence about benefits, most cost-effectiveness analyses have difficulty modeling the harms of screening, partly because of less evidence about harms. Another reason, however, is that benefits and harms are usually expressed in different metrics and are thus difficult to combine into a single measure of “net benefit” (17). Given these problems, we agree with Lansdorp-Vogelaar et al. that decision makers should use modeling as a tool for making decisions rather than as the sole “decider.” Modeling can sometimes extend and clarify evidence, but it should not be considered evidence in itself.

THE PROBLEM OF NONADHERENCE

One important screening problem that Lansdorp-Vogelaar et al. (16) discuss is how to model adherence to screening. If a model assumes 100% adherence, it likely overestimates benefits for a population and underestimates the cost of any interventions to increase adherence. Nonadherence is an important factor that reduces the population health effect of a screening program. If, for example, adherence for a fecal test for colorectal cancer is much higher than adherence for an invasive test such as colonoscopy, the fecal test may achieve a higher population benefit than colonoscopy just because of greater adherence.

Von Wagner et al. (18) point out the strong, graded association between cancer screening participation and socioeconomic status. Rather than accept lower socioeconomic status as an adequate explanation, however, these authors seek to develop a framework for understanding how lower socioeconomic status leads to reduced screening participation. They choose 3 corollaries of socioeconomic status—stressors and resources for change, educational opportunities and experience, and illness experiences—and consider how each may affect screening participation.

Although much nonadherence is associated with such characteristics of potential participants as socioeconomic status, other factors can be traced to organization of the health care system. Levin (19) discusses the possibilities of “organized screening programs” in improving adherence to colorectal cancer screening. As the author demonstrates, these programs—which include such features as an explicit policy, a defined target population, management and health care teams, quality assurance structures, and ongoing measurement of disease occurrence in the population—have generally increased adherence to colorectal cancer screening.

Even though nonadherence due to personal disadvantage or to poor health care organization is clearly problematic, it is also important for screening advocates to remember that screening is a decision, not a mandate. Even the most ben-

A PROBLEM INHERENT IN SCREENING: OVERDIAGNOSIS AND THE HETEROGENEITY OF SCREENING ABNORMALITIES

Nonadherence is not the only factor that limits the net health benefits of screening for a population. Another factor is the heterogeneity of the abnormalities found by screening. As Viera (20) points out, disease and normality are not dichotomous but exist on a continuum separated by various degrees of increasing risk to health. Screening necessarily detects more than just “high-risk” abnormalities destined to progress to adverse health outcomes; it also detects moderate or low-/no-risk “abnormalities” that would never have progressed to important health conditions. More sensitive screening strategies may lower our threshold for labeling the findings of screening tests as abnormal, leading us to create whole new categories of people with “predisease,” a heterogeneous mixture of high- and low-/no-risk people. The label of “abnormal” often implies to patients and clinicians that all of these people should be “treated” in some way, many of whom are thusly overdiagnosed and overtreated. For most conditions, our understanding of the health risk continuum is limited. Thus, as Rose (21) pointed out many years ago, a variable but usually large number of people who suffer an adverse health event could not have been placed in a risk category beforehand that, in an absolute sense, we would consider “high.”

Harris et al. (22) suggest that we redefine the target of screening, replacing categorical thinking such as “disease” and “predisease” with an umbrella concept, “predictor of poor health” (PPH). Each PPH is characterized by the probability it signifies for the future adverse health event we are trying to prevent. PPHs are to be considered along a continuum: a strong PPH is associated with a high absolute probability of developing an adverse health outcome in a defined period of time, whereas a weak PPH signals a low probability, with all degrees of PPHs in between. Screening should seek to detect people with strong PPHs for further evaluation, while people with weak PPHs can be reassured that they do not need further evaluation and perhaps not even further screening. The fact that there are relatively few examples of strong PPHs (and many examples of PPHs with an uncertain probability of a future adverse health event) limits the potential for screening to reduce a population’s burden of suffering from a condition without causing harm from overdiagnosis.

Using this framework, we can think of overdiagnosis as resulting from treating a person with a weak PPH the same way we would treat a person with a strong PPH. That is, when people have findings on screening that signal a low probability of developing the adverse health outcome (i.e., a weak PPH), they often still receive further evaluation, sometimes being diagnosed with a condition that would
be unlikely to ever cause them important health problems. Because we often screen for PPHs of uncertain strength, we may not know the strength of the PPH we have found and end up treating everyone as if they had a strong PPH. Many of these people are treated unnecessarily.

Interestingly, PPHs can be made stronger by raising (rather than lowering) thresholds for defining abnormality on screening tests. PPHs can also be made stronger by targeting screening to higher prevalence populations rather than expanding screening to broader, more universal populations. Although raising test thresholds and targeting screening limit the effect of screening on the burden of suffering in a population, they do allow a better, more efficient trade-off between potential benefits and harms/costs. Targeted strategies that focus on strong PPHs amount to a “low hanging fruit” approach to screening.

Overdiagnosis is increasingly recognized as one of the most important harms of screening. Although most agree that overdiagnosis is widespread, there is disagreement about its magnitude in terms of both frequency and its negative effect on health. De Gelder et al. (23) consider the frequency of overdiagnosis for breast cancer screening. They point out the wide variation in the estimates of breast cancer overdiagnosis, ranging from about 1% to as much as 52%. The contribution of de Gelder et al. is to help us understand these different estimates so we can weigh the magnitude of this harm against the benefit. Although de Gelder et al.’s model does not exactly predict the results of the Netherlands Screening Program, especially in the “steady-state” period, it does demonstrate that some of the differences in the estimates come from studies using different denominators and different time periods. Probably the estimate most relevant to screening is the proportion of screen-detected cancers during the screening period that are overdiagnosed. De Gelder et al. estimate this proportion as about 20% (extension phase), which is not far from Welch and Black’s (24) estimate of 24% based on the Malmö study. To many, this will be an impressive frequency for a harm that results in such unnecessary treatment as surgery, radiation therapy, and hormonal treatment.

**REDUCING OVERDIAGNOSIS BY TARGETING SCREENING**

Buijsse et al. (25) attempt to move us toward a more targeted approach to controlling the adverse health effects of diabetes. By reviewing risk assessment tools for identifying individuals at increased risk of developing diabetes, they help us understand the challenges of targeting prevention interventions, whether a screening or lifestyle strategy. Risk assessment (or “risk stratification”) can be thought of as the first of a possible 2-step screening process. Step 2 (which may involve blood or more invasive tests) is targeted to people whose results are positive in step 1 (risk assessment, usually on the basis of easily available information). For diabetes, Buijsse et al. found that a minority of the risk assessment tools had been validated in external populations and that discrimination was limited outside of the derivation population. There were multiple methodological concerns, and further concerns about feasibility of application. When easily available data are used, the present tools are certainly not strong PPHs, especially if we consider the PPH as predicting not just diabetes incidence but instead diabetes-related adverse health events. Still, we agree that the idea of improving risk assessment tools to target preventive interventions is a good one; tools will hopefully improve with more research.

**THE CONTRIBUTION OF GENETICS TO SCREENING**

One frequently discussed potential approach to improving risk assessment (and thus targeting step 2 of the screening process) is genetic screening. As discussed by Burke et al. (26), genetic screening can be either a single- or a double-step process. When used for newborn screening or for screening during pregnancy, the intent is to detect in a single step PPHs that strongly predict adverse health outcomes. Another potential use of genetic screening is to improve risk assessment—to stratify a population into subgroups more or less susceptible to an adverse health problem. People with greater susceptibility could then undergo further testing or surveillance, while those with lesser susceptibility could be reassured. To date, there have been few clinically useful applications of genetic tests to assess risk and thus target further screening or surveillance.

Some of the potential harms of genetic screening discussed by Burke et al. (26) are common to other screening situations: false-positive screening tests, ambiguous test results, and incidental findings. Larger numbers of false-positive and ambiguous screening test results weaken PPHs; a positive screening test is less predictive of an adverse health outcome. This is also true for incidental findings that must be evaluated but do not lead to treatment that improves health outcomes. The incidental finding itself becomes a weak PPH. The potential harms of false-positive and ambiguous screening tests, and incidental findings, have been insufficiently appreciated. We need more rigorous study to determine the effects of these types of abnormal screening tests on people’s health.

**THE MAGNITUDE OF BENEFIT FROM SCREENING TRIALS**

Our growing appreciation of the problem of overdiagnosis has come from cancer screening. For example, several estimates of prostate cancer overdiagnosis from screening have been in the 25% to nearly 50% range (27, 28). This large magnitude of overdiagnosis must be weighed against the magnitude of benefit from screening.

How to measure the magnitude of benefit is Hanley’s topic (29), who shows the complexity of estimating the steady-state screening benefit from cancer screening trials. The author makes the theoretical point that one must consider the benefit in terms of time from randomization, noting that there is typically less cancer mortality reduction in the initial time period, followed by a greater reduction a few years after randomization. After the trial, when screening has stopped, there is again less mortality reduction as the
effect of screening diminishes. Using this concept, he makes brief calculations that suggest larger mortality reductions for prostate, breast, and colorectal cancer screening than have been reported. The issue may be even more complex, however, because it is also necessary to take into account the effects of competing mortality, of overdiagnosis, and of differences in lethality of cancers in younger versus older people. There are further concerns about bias within the trials. The effects of screening are likely not constant over time, so there may be no single "steady-state" screening effect. Most helpful, of course, would not be calculations but rather long-term follow-up of randomized controlled trials of screening, analyzed, as Hanley proposes, by time since randomization and time after screening stops.

**DISCUSSING SCREENING WITH THE PUBLIC**

When one considers all of the complexities discussed here—and many we have not discussed—it is not surprising that the evidence community often disagrees about the certainty and magnitude found in the evidence about screening. Fletcher’s (30) provocative and wide-ranging perspective gives us hope that we are making some progress, although we are also discovering even more complexity over time. The author points out that, in addition to the discussions highlighted here, there is another discussion we should be having more than we are—a discussion about screening with the public. The public’s enthusiasm for screening is well documented (31). For example, although Burke et al. (26) caution us not to undertake widespread genome-wide genetic screening without more research and “careful deliberation,” recent studies show that many members of the public are interested in genetic testing even if the results are sometimes wrong and ambiguous, even if there is no resultant treatment, and even if they have to pay for the tests themselves (32, 33). A news article reporting on these studies refers to those who would go slowly on direct-to-consumer advertising of genetic screening tests as “paternalists” (34). It is not clear whether a wider and clear-eyed discussion with the public about the magnitude of benefits and harms of screening (as well as the costs) would affect public demand for screening, but this discussion seems long overdue.

**WHERE WE ARE AND WHERE WE MAY BE HEADED**

Overall, then, the present truths about screening are that it is sometimes effective in improving health outcomes, that it is also associated with important harms, and that the balance between benefits and harms (and costs) is often not clear. With few exceptions, its contribution to improving the health of the public is small, yet it has become a popular and growing form of prevention. It may be that we are learning that the magnitude of benefit from screening is less than we hoped, and the harms may be greater than we thought.

Perhaps we should not think of screening as our primary prevention strategy but rather use screening to make a real, but limited contribution to population health for a few conditions. Perhaps we should generally target screening to smaller subpopulations with the highest potential benefit and the lowest potential harm, using higher thresholds to define an abnormality. Our tendency to expand the population eligible for screening, to lower the thresholds for a positive test, and to develop more sensitive screening tests may well be causing more harm than benefit. Such factors as the biologic heterogeneity of abnormalities found on screening tests and the problems of nonadherence, false-positive and ambiguous test results, incidental findings, the existence of few strong PPHs, and overdiagnosis make screening more a halfway technology (2) than a once and future answer. One wonders whether the payoff for population health would be greater if we shifted some of the resources we now devote to screening to developing, testing, and implementing alternative approaches to prevent the important threats to population health.

Fletcher’s (30) thoughtful perspective on breast cancer screening is consistent with this vision of the future of screening. Among other insights, the author points out that the “truth” about screening may change over time. If treatment of clinically detected abnormalities improves, if the public is more vigilant about noticing earlier symptoms, then screening may become less and less important to our work of improving the health of the public.

**A COUNTERVIEW**

A counterview is suggested by the genetic screening enthusiasts described by Burke et al. (26). In this view, universal genetic screening holds great promise for highly accurate risk stratification that would allow efficient second-step testing. If true, this strategy could indeed reduce the population burden of suffering from multiple conditions. In this view, screening could conceivably make a major contribution to population health. Those who work in this framework often see the benefits of screening as going beyond reducing the burden of population suffering to also include providing information alone (even if sometimes inaccurate) as an important benefit (32).

I will admit to being more convinced by the first view than the second, but, if given new evidence, I also agree with John Maynard Keynes (34): “When the facts change, I change my mind. What do you do, sir?” More evidence is surely needed; we will need to wait to see whether “the facts” change.

**THE OLD MAN AND THE “TRUTH” SIGN**

There is another variation of the cartoon frame from The New Yorker with the old bearded man and the “Truth” sign. This frame has a second “Truth” sign. While the first sign is simple and plain, the second is a large, marketing sign with colors and glittering lights. The signs point in opposite directions and the old man is caught between the 2, wondering which way to go. Screening has an intuitive appeal much like a large marketing sign. In our discussions among ourselves and in our discussions with the public, our responsibility is not to be distracted by the glittering lights, the intuitive
appeal, but to continue to seek the several “truths” about the complex world of screening.

ACKNOWLEDGMENTS

Author affiliations: Cecil G. Sheps Center for Health Services Research, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina; Department of Medicine, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina; and Public Health Leadership Program, School of Public Health, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina.

Conflict of interest: none declared.

REFERENCES