PUBLIC HEALTH

Improving cancer screening programs

Evaluating diagnostic tests in learning screening programs could improve public health

By **Mette Kalager** *and* **Michael Bretthauer**

National cancer screening programs, such as mammography for breast cancer, are widely implemented to reduce cancer incidence and mortality in high-income countries. Their introduction is also being considered in low- and middle-income countries. For many cancer types, the benefits and harms of different screening tests and the intervals at which they should be implemented are unknown. Thus, randomized comparison testing is warranted. However, this is not possible because most people in high-income countries have already undergone screening or have refused screening and are not comparable (*1*). There is an ethical, medical, economic, and societal imperative for continuous evaluation of cancer screening programs to ensure that their benefits outweigh any harms. This may be achievable if the screening programs can become the arena for clinical testing through the implementation of learning screening programs.

Every year, ~46 million individuals (126,000 people each day) are offered cancer screening worldwide (*2*). The most commonly used screening tests include the Pap smear for cervical cancer; fecal immunochemical test (FIT), fecal occult blood test (FOBT), colonoscopy, or sigmoidoscopy for colorectal cancer; prostate-specific antigen (PSA) testing for prostate cancer; mammography for breast cancer; and computed tomography (CT) for lung cancer. The effects of such screening tests on cancer incidence and mortality need to be balanced against their harms, such as possible medical complications of screening procedures and subsequent treatment, and psychological distress. Cancer screening targets whole populations rather than individual patients, and each cancer occurs rarely. Thus, many people undergo testing without benefit. Recently, it has also become evident that some screening tests, such as PSA, mammography, and lung CT, lead to increased cancer incidence rates as a result of overdiagnosis—that is, the detection of cancers and cancer precursors that would not have progressed to symptoms or death in the absence of screening (*3*). Because overdiagnosed lesions cannot be distinguished from lesions that will progress, all patients are treated. Therefore, overdiagnosed patients only experience harms and do not gain any benefit from screening.

For many cancers, not only are the comparative benefits and harms of available screening tests unknown, there is also a lack of consensus regarding the appropriate choice of test interval and threshold for a positive diagnosis. Screening programs in different countries use different tests, intervals, and thresholds. For example, the UK National Health Service (NHS) offers a screening sigmoidoscopy at age 55 and FIT from age 60 in England, whereas in Scotland the NHS does not offer sigmoidoscopy and has a higher cutoff for FIT positivity than in England (*4*). No new prescription drug enters the market without rigorous testing in randomized controlled trials (RCTs); by contrast, many cancer screening tests and strategies have been introduced without such rigorous clinical trials. For the few tests that have been evaluated in RCTs, such as mammography and FOBT screening, the data are often outdated and there is controversy about whether the tests are beneficial in current screening programs. Because of the enormous improvements in treatment of many cancers, early detection through screening may not be as important to achieving improved cancer mortality rates today as it was 20 to 30 years ago when the screening trials were performed.

Future tests such as panels of genetic markers for prostate and breast cancer screening are expected to enter national screening programs soon (*5*). In the United States, a fecal DNA marker panel for colorectal cancer is recommended by some organizations on the basis of microsimulation modeling data (*6*) and not RCTs. When screening is widespread and everybody is exposed, traditional RCTs of screening tests outside of national screening programs are not possible because there is no control group for comparison.

A learning screening program continuously and systematically generates knowledge about what works—that is, which test is most effective to reduce mortality, has the optimal balance between benefits and harms, and is most likely to be acceptable in the population. Individuals in national screening programs are asked to be randomized to receive either a new screening test, interval, or threshold, or the standard option. Testing thus involves randomized comparisons of thousands or even tens of thousands of participants with clinically relevant end points, such as cancer incidence or mortality. After the testing phase is over, it will be possible to make valid estimates of benefits and harms. For example, overdiagnosis can be measured in terms of the difference between numbers of cancers detected in individuals randomized to one screening test versus those randomized to another. Then, the best test or method will be introduced to all. When a new test or method becomes available, the cycle begins again. This continuous cycle of testing, treatment, and evaluation ensures that the public receives optimal cancer screening (see the figure).

Finland, Norway, and Poland have started to apply learning screening programs (*7*–*9*). The Finnish colorectal cancer screening program was designed as a randomized comparison of FOBT versus no screening, using the entire population of 60- to 69-year-olds in Finland (*8*). The program did not find a mortality benefit of screening versus no screening. FOBT screening was stopped, and half the Finnish population was prevented from undergoing an ineffective screening test. This year, Finland is starting the learning screening program cycle again to test FIT screening. The Norwegian cervical cancer screening program currently invites women with even-numbered dates of birth to be tested for human papilloma virus (HPV, a cause of cervical cancer) with Pap smears, whereas women born on odd-numbered dates receive the standard Pap smear. Women with negative tests in the two groups are offered new testing every 5 and 3 years, respectively, and compared for cancer incidence (*9*). The Polish colorectal cancer screening program has started to randomize individuals to immediate or delayed colonoscopy screening, or to no screening, and will assess cancer incidence and mortality (*7*).

Learning screening program activities that could be established immediately include assessment of colorectal cancer, cervical cancer, and breast cancer screening

Clinical Effectiveness Group, Institute of Health and Society, University of Oslo, Oslo, Norway, and Department of Transplantation Medicine, Oslo University Hospital, Oslo, Norway. Email: mette.kalager@medisin.uio.no

tests. The NHS colorectal cancer screening program in England offers FIT screening with a positivity threshold of 20 μg of hemoglobin per gram of feces (μg Hb/g) for further diagnostic assessment with colonoscopy, whereas NHS Scotland offers FIT screening with a positivity threshold of 80 μg Hb/g (*4*). It is unknown whether colorectal cancer incidence and mortality rates differ with these thresholds and whether cases of early colorectal cancer are being missed in Scotland (*10*). A learning screening program could randomize individuals to the two thresholds. After 10 years of follow-up, the data from the program will establish the optimum thresholds for reduction of colorectal cancer incidence and mortality, as well as for harms and burdens (e.g., number of false positives, burden of further testing by colonoscopy, and side effects from further testing such as colon perforations or hemorrhage), and could be established and implemented to all. Then, the program could continue and, for example, women could be randomized to colonoscopy versus FIT, because there is evidence from RCTs that women have less benefit from FIT and sigmoidoscopy screening to reduce colorectal cancer risk (*11*).

The cervical cancer screening program in Norway could be expanded to a true learning screening program (with continuous cycles of testing) by randomizing women with HPV-negative cytology to a 10-year screening interval in the next testing cycle. Furthermore, HPV-vaccinated women could be randomized to Pap smear screening perhaps every 10 years versus every 15 years, because it is unknown how frequently HPV-vaccinated women should be screened, if at all (*12*). In a learning screening program, it is possible to find out.

Although mammography screening resulted in reduced breast cancer mortality in RCTs carried out 30 years ago, recent improvements in therapy and care may have reduced this effect (*13*). The balance of benefits and harms of cancer screening is delicate and may shift with better quality of clinical service and improved cancer treatment and diagnostics (*13*). In a learning screening program, women could be randomized to mammography screening that is currently offered (every 1 to 3 years) or to screening, for example, every 5 years. If screening every 5 years is as effective in reducing mortality and also reduces overdiagnosis, then screening every 1 to 3 years could be phased out by randomizing women to screening every 5 years or to longer intervals, or even to no screening.

There are important challenges to consider for learning screening programs: All individuals in these programs must be informed about the randomized comparisons. Information needs to include descriptions of the uncertainty of current tests and strategies, and estimates of new tests and strategies that are tested. Information needs to be easily accessible without overselling and promoting the new test. Eligible individuals should be given the opportunity to consent to testing and evaluation. Opt-out may be considered to facilitate the need for testing large numbers of individuals. The opportunity to consent also means that individuals who do not want to participate (that is, do not provide consent) need an alternative screening option, the standard screening test. The screening program will hence be similar to any RCT with a pre-randomization consent. Independent ethical oversight must also be provided, as occurs with clinical trials for therapeutics. Screening has been strongly advocated and many individuals overestimate the effect of screening. Hence, an independent review should focus on the individual's ability to give informed consent and may require evidence that the individuals are informed properly.

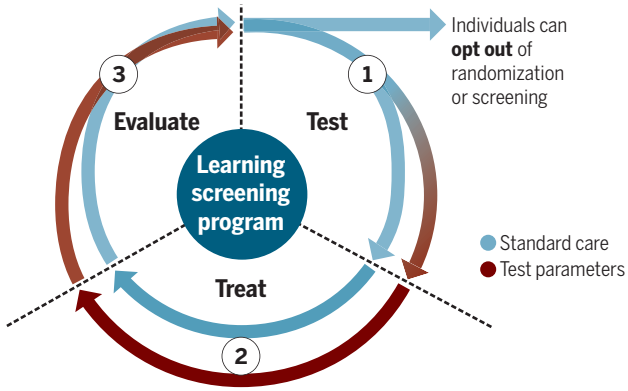
Randomized testing in learning screening programs requires the application of different trial methodologies depending on the settings and feasibility. For example, cluster-randomized trials may be applicable where individual randomization is deemed difficult to implement. Stepped-wedge cluster-randomized trials, where new interventions are implemented in a randomized order between clusters, may facilitate the implementation of learning screening programs, because this design combines the desired implementation of new tests and strategies with the advantage of randomized comparison. Platform clinical trials that combine multiple new strategies with a joint standard-of-care group may also be considered.

Learning screening programs need new funding mechanisms because they are operating between public health and research. Ideally, funding should be through public health programs rather than research budgets. Randomization does not increase costs in already functioning, conventional screening programs. Learning screening programs may also be applicable to other areas of public health and clinical practice such as diagnostic disease assessment; for example, radiology and endoscopy techniques could be compared for inflammatory bowel disease diagnosis.

The traditional divide between clinical trials and disease detection has become an obstacle to evidence-based screening programs. The establishment of learning screening programs may provide high-quality evidence for screening strategies with the most favorable benefit-to-harm ratio. ■

Learning screening programs

Patients enrolled in national screening programs agree to be randomized into testing arms that assess different tests, intervals, or thresholds to identify the optimal screening test that improves population mortality without overdiagnosing patients for additional treatment.



1 Volunteers undergoing national screening are randomized into **different treatment arms**

2 Individuals receive **different screening tests**, at different intervals or with different thresholds for positivity

3 Treatment arms are **compared** and the optimal screening test is selected as the standard test, which can be compared to additional variations in the future

GRAPHIC: N. CARY/*SCIENCE*

REFERENCES AND NOTES

1. R. J. Connor, P. C. Prorok, D. L. Weed, *J. Clin. Epidemiol.* **44**, 1215 (1991).
2. World Population Review, *Western Countries* 2019; http://worldpopulationreview.com/countries/western-countries/.
3. M. Bretthauer, M. Kalager, *Br. J. Surg.* **100**, 55 (2013).
4. U. K. Bowel Cancer, *Bowel Cancer Screening*; www.bowelcanceruk.org.uk/about-bowel-cancer/screening/.
5. J. D. Cohen *et al.*, *Science* **359**, eaar3247 (2018).
6. U.S. Preventive Services Task Force, *JAMA* **315**, 2564 (2016).
7. M. F. Kaminski *et al.*, *Endoscopy* **47**, 1144 (2015).
8. J. Pitkäniemi *et al.*, *BMJ Open Gastroenterol.* **2**, e000034 (2015).
9. Cancer Registry of Norway, *HPV i primærscreening*; www.kreftregisteret.no/en/screening/Cervical-Cancer-Screening-Programme/Helsepersonell/screeningstrategi-og-nasjonale-retningslinjer/hpv-i-primarscreening/.
10. S. Moss *et al.*, *Gut* **66**, 1631 (2017).
11. Ø. Holme *et al.*, *BMJ* **356**, i6673 (2017).
12. P. G. Rossi *et al.*, *Prev. Med.* **98**, 21 (2017).
13. H. G. Welch, P. C. Prorok, A. J. O'Malley, B. S. Kramer, *N. Engl. J. Med.* **375**, 1438 (2016).

ACKNOWLEDGMENTS

Supported by Norwegian Research Council grant 250256 and Norwegian Cancer Society grants 6741288 and 190345.

10.1126/science.aay3156

Improving cancer screening programs

Mette Kalager and Michael Bretthauer

Science **367** (6474), 143-144.
DOI: 10.1126/science.aay3156