Epidemiology series

Uses and abuses of screening tests

David A Grimes, Kenneth F Schulz

Screening tests are ubiquitous in contemporary practice, yet the principles of screening are widely misunderstood. Screening is the testing of apparently well people to find those at increased risk of having a disease or disorder. Although an earlier diagnosis generally has intuitive appeal, earlier might not always be better, or worth the cost. Four terms describe the validity of a screening test: sensitivity, specificity, and predictive value of positive and negative results. For tests with continuous variables—eg, blood glucose—sensitivity and specificity are inversely related; where the cutoff for abnormal is placed should indicate the clinical effect of wrong results. The prevalence of disease in a population affects screening test performance: in low-prevalence settings, even very good tests have poor predictive value positives. Hence, knowledge of the approximate prevalence of disease is a prerequisite to interpreting screening test results. Tests are often done in sequence, as is true for syphilis and HIV-1 infection. Lead-time and length biases distort the apparent value of screening programmes; randomised controlled trials are the only way to avoid these biases. Screening can improve health; strong indirect evidence links cervical cytology programmes to declines in cervical cancer mortality. However, inappropriate application or interpretation of screening tests can rob people of their perceived health, initiate harmful diagnostic testing, and squander health-care resources.

Screening is a double-edged sword, sometimes wielded clumsily by the well-intended. Although ubiquitous in contemporary medical practice, screening remains widely misunderstood and misused. Screening is defined as tests done among apparently well people to identify those at an increased risk of a disease or disorder. Those identified are sometimes then offered a subsequent diagnostic test or procedure, or, in some instances, a treatment or preventive medication.[1] Looking for additional illnesses in those with medical problems is termed case finding;[2,3] screening is limited to those apparently well.

Screening can improve health. For example, strong indirect evidence lends support to cytology screening for cervical cancer. Insufficient use of this screening method accounts for a large proportion of invasive cervical cancers in industrialised nations.[4] Other beneficial examples include screening for hypertension in adults; screening for hepatitis B virus antigen, HIV-1, and syphilis in pregnant women; routine urine culture in pregnant women at 12–16 weeks' gestation; and measurement of phenylalanine in newborns.[5] However, inappropriate screening harms healthy individuals and squanders precious resources. The nearly universal antenatal screening for gestational diabetes (a diagnosis in search of a disease)[6] in the USA[7] exemplifies the widespread confusion about the nature and aim of screening. Here, we review the purposes of screening, the selection of tests, measurement of validity, the effect of prevalence on test outcome, and several biases that can distort interpretation of tests.

Ethical implications

What are the potential harms of screening?

Screening differs from the traditional clinical use of tests in several important ways. Ordinarily, patients consult with clinicians about complaints or problems; this prompts testing to confirm or exclude a diagnosis.[8] Because the patient is in pain and requests our help, the risk and expense of tests are usually deemed acceptable by the patient. By contrast, screening engages apparently healthy individuals who are not seeking medical help (and who might prefer to be left alone). Alternatively, consumer-generated demand for screening, such as for osteoporosis and ovarian cancer, might lead to expensive programmes of no clear value.[5,9] Hence, the cost, injury, and stigmatisation related to screening are especially important (though often ignored in our zeal for earlier diagnosis); the medical and ethical standards of screening should be, correspondingly, higher than with diagnostic tests.[10] Bluntly put: every adverse outcome of screening is iatrogenic and entirely preventable.

Screening has a darker side that is often overlooked.[2] It can be inconvenient (the O'Sullivan screen for gestational diabetes), unpleasant (sigmoidoscopy or colonoscopy), and expensive (mammography). For example, a recent Markov model revealed that new screening tests for cervical cancer that are more sensitive than the Papanicolaou test (and thus touted as being better) will drive up the average cost of detecting an individual with cancer.[11] Paradoxically, these higher costs could make screening unattainable by poor women who are at highest risk.[4] The net effect might be more instances of cancer.

A second wave of injury can arise after the initial screening insult: false-positive results and true-positive results leading to dangerous interventions.[2] Although the stigma associated with correct labeling of people as ill might be acceptable, those incorrectly labeled as sick suffer as well. For example, labeling productive steelworkers as being hypertensive led to increased absenteeism and adoption of a sick role, independent of treatment.[12,13] More recently, women labeled as having gestational diabetes reported deterioration in their health and that of their infants over the 5 years after diagnosis.[14] By what right do clinicians rob people of their perceived health, and for what gain?[2]

Screening can also lead to harmful treatment. Treatment of hyperlipidaemia with clofibrate several decades ago provides a sobering example. Treatment of the cholesterol count (a risk factor, rather than an illness itself) inadvertently led to a 17% increase in mortality among middle-aged men given the drug.[2] This screening

Lancet 2002; **359:** 881–84

Family Health International, PO Box 13950, Research Triangle Park, NC 27709, USA (D A Grimes MD, K F Schulz PhD)

Correspondence to: Dr David A Grimes
(e-mail: dgrimes@fhi.org)

misadventure cost the lives of more than 5000 men in the USA alone.[2] Because of these mishaps, reviews of screening practices have recommended that clinicians be more selective.[5,15]

Criteria for screening

If a test is available, should it be used?

The availability of a screening test does not imply that it should be used. Indeed, before screening is done, the strategy must meet several stringent criteria. One checklist separates criteria in three parts: the disease, the policy, and the test.[1] The disease should be medically important and clearly defined, and its prevalence reasonably well known. The natural history should be known, and an effective intervention must exist. Concerning policy, the screening programme must be cost effective, facilities for diagnosis and treatment must be readily available, and the course of action after a positive result must be generally agreed on and acceptable to those screened. Finally, the test must do its job. It should be safe, have a reasonable cut-off level defined, and be both valid and reliable. The latter two terms, often used interchangeably, are distinct. Validity is the ability of a test to measure what it sets out to measure, usually differentiating between those with and without the disease. By contrast, reliability indicates repeatability. For example, a bathroom scale that consistently measures 2 kg heavier than a hospital scale (the gold standard) provides an invalid but highly reliable result.

Although an early diagnosis generally has intuitive appeal, earlier might not always be better. For example, what benefit would accrue (and at what cost) from early diagnosis of Alzheimer's disease, which to date has no effective treatment? Sackett and colleagues[2] have proposed a pragmatic checklist to help decide when (or if) seeking a diagnosis earlier than usual is worth the expense and bother. Does early diagnosis really benefit those screened, for example, in survival or quality of life? Can the clinician manage the additional time required to confirm the diagnosis and deal with those diagnosed before symptoms developed? Will those diagnosed earlier comply with the proposed treatment? Has the effectiveness of the screening strategy been established objectively?[5,15] Finally, are the cost, accuracy, and acceptability of the test clinically acceptable?

Assessment of test effectiveness

Is the test valid?

For over half a century,[16] four indices of test validity have been widely used: sensitivity, specificity, and predictive values of positive and negative. Although clinically useful (and far improved over clinical hunches), these terms are predicated on an assumption that is often clinically unrealistic—ie, that all people can be dichotomised as ill or well. (Indeed, one definition of an epidemiologist is a person who sees the entire world in a 2×2 table.) Often, those tested simply do not fit neatly into these designations: they might be possibly ill, early ill, probably well, or some other variant. Likelihood ratios, which incorporate varying (not just dichotomous) degrees of test results, can be used to refine clinicians' judgments about the probability of disease in a particular person.

For simplicity, however, assume a population has been tested and assigned to the four mutually exclusive cells in figure 1. Sensitivity, sometimes termed the detection rate,[10] is the ability of a test to find those with the disease. All those with disease are in the left column. Hence, the sensitivity is simply those correctly identified by the test (a) divided by all those sick (a+c). Specificity denotes the

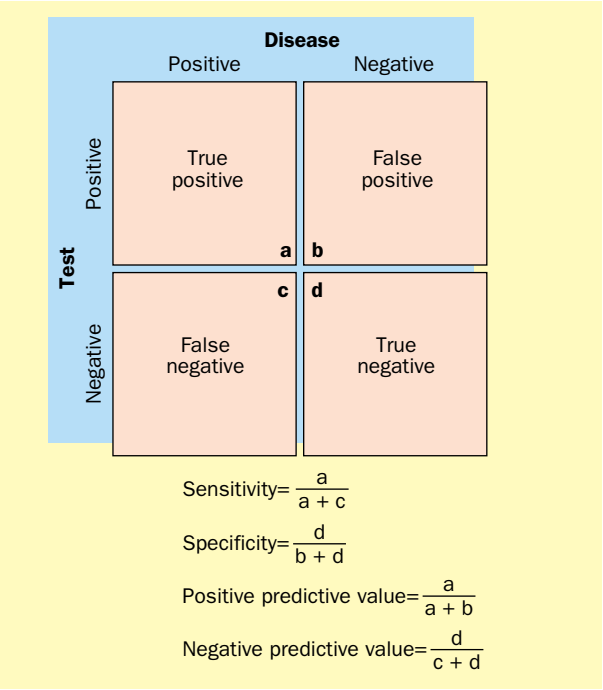

Figure 1: **Template for calculation of test validity**

ability of a test to identify those without the condition. Calculation of this proportion is trickier, however. By analogy to sensitivity, many assume (incorrectly) that the formula here is b/(b+d). However, the numerator for specificity is cell d (the true negatives), which is divided by all those healthy (b+d).

Although sensitivity and specificity are of interest to public-health policymakers, they are of little use to the clinician. Stated alternatively, sensitivity and specificity (population measures) look backward (at results gathered over time).[8] Clinicians have to interpret test results to those tested. Thus, what clinicians need to know are the predictive values of the test (individual measures, which look forward). To consider predictive values, one needs to shift the orientation in figure 1 by 90 degrees: predictive values work horizontally (rows), not vertically (columns). In the top row are all those with a positive test, but only those in cell a are sick. Thus, the predictive value positive is a/(a+b). The "odds of being affected given a positive result (OAPR)" is the ratio of true positives to false positives, or a to b.[10] For example, in figure 1, the OAPR is 75/5, or 17/1. This corresponds to a positive predictive value of 89%. Advocates of use of the OAPR note that these odds better describe test effectiveness than do probabilities (predictive values). In the bottom row of figure 1 are those with negative tests, but only those in cell d are free of disease. Hence, the predictive value negative is d/(c+d).

Learning (and promptly forgetting) these formulas was an annual ritual for many of us in our clinical training. If readers understand the definitions above and can recall the 2×2 table shell, then they can quickly figure out these formulas when needed. As a mnemonic, disease goes at the top of the table shell, since it is our top priority. By default, test goes on the left border.

Through the years, researchers have tried to simplify these four indices of test validity by condensing them into a single term.[8] However, none adequately depicts the important trade-offs between sensitivity and specificity that generally arise. An example is diagnostic accuracy, which is the proportion of correct results.[3] It is the sum of

the correctly identified ill and well divided by all those tested, or (a+d)/(a+b+c+d). Cells b and c are noise in the system. Another early attempt, Youden's J, is simply the predictive value positive plus the predictive value negative minus one.[17] The range of values extends from zero (for a coin toss with no predictive value) to 1·0, where predictive values of both positive and negative tests are perfect.

Trade-offs between sensitivity and specificity

Where should the cut-off for abnormal be?

The ideal test would perfectly discriminate between those with and without the disorder. The distributions of test results for the two groups would not overlap. More commonly in human biology, test values for those with and without a disease overlap, sometimes widely.[18] Where one puts the cut-off defining normal versus abnormal determines the sensitivity and specificity. For any continuous outcome measurement—for example, blood pressure, intraocular pressure, or blood glucose—the sensitivity and specificity of a test will be inversely related. Figure 2 shows that placing the cut-off for abnormal blood glucose at point X produces perfect sensitivity; this low cut-off identifies all those with diabetes. However, the trade-off is poor specificity: those in the part of the healthy distribution in pink and purple are incorrectly identified as having abnormal values. Placing the cut-off higher at point Z yields the opposite result: all those healthy are correctly identified (perfect specificity), but the cost here is missing a proportion of ill individuals (portion of the diabetic distribution in purple and blue). Placing the cut-off at point Y is a compromise, mislabeling some healthy people and some people with diabetes.

Where the cut-off should be depends on the implications of the test, and receiver-operator characteristic curves are useful in making this decision.[19] For example, screening for phenylketonuria in newborns places a premium on sensitivity rather than on specificity; the cost of missing a case is high, and effective treatment exists. The downside is a large number of false-positive tests, which cause anguish and further testing. By contrast, screening for breast cancer should favour specificity over sensitivity, since further assessment of those tested positive entails costly and invasive biopsies.[20]

Prevalence and predictive values

Can test results be trusted?

A badly understood feature of screening is the potent effect of disease prevalence on predictive values.

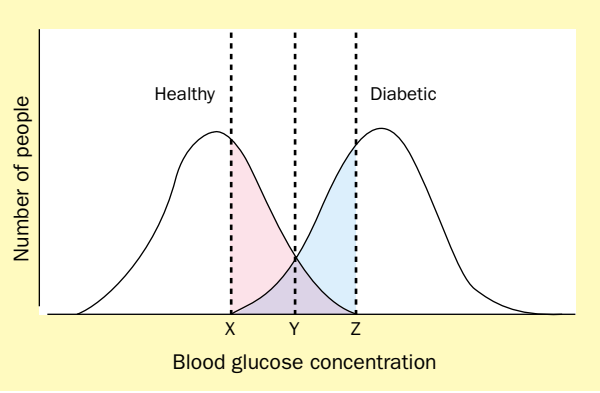

Figure 2: **Hypothetical distribution of blood glucose concentrations in people with and without diabetes**
Setting cut-off for abnormal at X yields perfect sensitivity at the expense of specificity. Setting cut-off at Z results in perfect specificity at the cost of lower sensitivity. Cut-off Y is a compromise.

Clinicians must know the approximate prevalence of the condition of interest in the population being tested; if not, reasonable interpretation is impossible. Consider a new PCR test for chlamydia, with a sensitivity of 0·98 and specificity of 0·97 (a superb test). As shown in the left panel of figure 3, a doctor uses the test in a municipal sexually transmitted disease clinic, where the prevalence of *Chlamydia trachomatis* is 30%. In this high-prevalence setting, the predictive value of a positive test is high, 93%—ie, 93% of those with a positive test actually have the infection.

Impressed with the new test, the doctor now takes it to her private practice in the suburbs, which has a clientele that is mostly older than age 35 years (figure 3, right panel). Here, the prevalence of chlamydial infection is only 3%. Now the same excellent test has a predictive positive value of only 0·50. When the results of the test are positive, what should the doctor tell the patient, and what, in turn, should the patient tell her husband? Here, flipping a coin has the same predictive positive value (and is considerably cheaper and simpler than searching for bits of DNA). This message is important, yet not widely understood: when used in low-prevalence settings, even excellent tests have poor predictive positive value. The reverse is true for negative predictive values, which are nearly perfect in figure 3. Although failing to diagnose sexually transmitted diseases can have important health implications, incorrectly labeling people as infected can wreck marriages and damage lives.

Tests in combination

Should a follow-up test be done?

Clinicians rarely use tests in isolation. Few tests have high sensitivity and specificity, so a common approach is to do tests in sequence. In the instance of syphilis, a sensitive (but not specific) reagin test is the initial screen. Those who test positive then get a second, more specific test, a diagnostic treponemal test. Only those who test positive on both receive the diagnosis. This strategy generally increases the specificity compared with a single test and limits the use of the more expensive treponemal test.[20] Testing for HIV-1 is an analogous two-step procedure.

Alternatively, tests can be done in tandem (parallel or simultaneous testing).[3,21] For example, two different tests might both have poor sensitivity, but one might be better at picking up early disease, whereas the other is better at

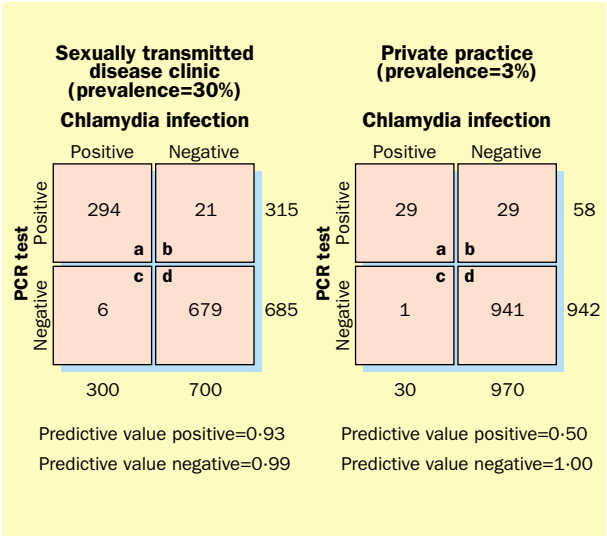

Figure 3: **Predictive values of a PCR test for *Chlamydia trachomatis* in high-prevalence and low-prevalence settings**

identifying late disease. A positive result from either test would then lead to diagnostic assessment. This approach results in higher sensitivity than would arise with either test used alone.

Benefit or bias?

Does a screening programme really improve health?

Even worthless screening tests seem to have benefit.[2] This cruel irony underlies many inappropriate screening programmes used today. Two common pitfalls lead to the conclusion that screening improves health; one is an artifact and the other a reflection of biology.

Lead-time bias

Lead-time bias refers to a spurious increase in longevity associated with screening. For example, assume that mammography screening leads to cancer detection 2 years earlier than would have ordinarily occurred, yet the screening does not prolong life. On average, women with breast cancer detected through screening live 2 years longer than those with cancers diagnosed through traditional means. This gain in longevity is apparent and not real: this hypothetical screening allows women to live 2 years longer with the knowledge that they have cancer, but does not prolong survival, an example of zero-time shift.[2]

Length bias

Length bias is more subtle than lead-time bias: the longevity association is real, but indirect. Assume that community-based mammography screening is done at 10-year intervals. Women whose breast cancers were detected through screening live 5 years longer on average from cancer initiation to death than those whose cancers were detected through usual means. That screening is associated with longer survival implies clear benefit. However, in this hypothetical example, this benefit indicates the inherent variability in cancer growth rates and not a benefit of screening. Women with indolent, slow-growing cancers are more likely to live long enough to be identified in decennial screening. Conversely, those with rapidly progressing tumours are less likely to survive until screening.

The only way to avoid these pervasive biases is to do randomised controlled trials and then to assess age-specific mortality rates for those screened versus those not screened.[10] Moreover, the trials must be done well. The quality of published trials of mammography screening has raised serious questions about the utility of this massive and hugely expensive enterprise.[22–24]

Conclusion

Screening can promote or impair health, depending on its application. Unlike a diagnostic test, a screening test is done in apparently healthy people, which raises unique ethical concerns. Sensitivity and specificity tend to be inversely related, and choice of the cut-off point for abnormal should indicate the implications of incorrect results. Even very good tests have poor predictive value positive when applied to low-prevalence populations. Lead-time and length bias exaggerate the apparent benefit of screening programmes, underscoring the need for rigorous assessment in randomised controlled trials before use of screening programmes.

Acknowledgments

We thank Willard Cates and David L Sackett for their helpful comments on an earlier version of this report. Much of this material stems from our 15 years of teaching the Berlex Foundation Faculty Development Course.

References

1 Cuckle HS, Wald NJ. Principles of screening. In: Antenatal and neonatal screening. Oxford: Oxford University Press, **1984:** 1–22.
2 Sackett DL, Haynes RB, Guyatt GH, Tugwell P. Clinical epidemiology: a basic science for clinical medicine, 2nd edn. Boston: Little, Brown and Company, 1991.
3 Lang TA, Secic M. How to report statistics in medicine. Philadelphia: American College of Physicians, 1997.
4 Sawaya GF, Grimes DA. New technologies in cervical cytology screening: a word of caution. *Obstet Gynecol* 1999; **94:** 307–10.
5 US Preventive Services Task Force. Guide to clinical preventive services, 2nd edn. Baltimore: Williams and Wilkins, 1996.
6 Enkin M, Keirse MJNC, Neilson J, et al (eds). A guide to effective care in pregnancy and childbirth, 3rd edn. Oxford: Oxford University Press, 2000.
7 Gabbe S, Hill L, Schmidt L, Schulkin J. Management of diabetes by obstetrician–gynecologists. *Obstet Gynecol* 1998; **91:** 643–47.
8 Feinstein AR. Clinical biostatistics XXXI: on the sensitivity, specificity, and discrimination of diagnostic tests. *Clin Pharmacol Ther* 1975; **17:** 104–16.
9 NIH Consensus Development Panel on Ovarian Cancer. Ovarian cancer: screening, treatment, and follow–up. *JAMA* 1995; **273:** 491–97.
10 Wald N, Cuckle H. Reporting the assessment of screening and diagnostic tests. *Br J Obstet Gynaecol* 1989; **96:** 389–96.
11 Myers ER, McCrory DC, Subramanian S, et al. Setting the target for a better cervical screening test: characteristics of a cost-effective test for cervical neoplasia screening. *Obstet Gynecol* 2000; **96:** 645–52.
12 Haynes RB, Sackett DL, Taylor DW, Gibson ES, Johnson AL. Increased absenteeism from work after detection and labeling of hypertensive patients. *N Engl J Med* 1978; **299:** 741–44.
13 Taylor DW, Haynes RB, Sackett DL, Gibson ES. Longterm follow-up of absenteeism among working men following the detection and treatment of their hypertension. *Clin Invest Med* 1981; **4:** 173–77.
14 Feig DS, Chen E, Naylor CD. Self-perceived health status of women three to five years after the diagnosis of gestational diabetes: a survey of cases and matched controls. *Am J Obstet Gynecol* 1998; **178:** 386–93.
15 Canadian Task Force on the Periodic Health Examination. The Canadian guide to clinical preventive care. Ottowa: Minister of Supply and Services Canada, 1994.
16 Yerushalmy J. Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. *Pub Health Rep* 1947; **62:** 1432–49.
17 Youden WJ. Index for rating diagnostic tests. *Cancer* 1950; **3:** 32–35.
18 Griffith CS, Grimes DA. The validity of the postcoital test. *Am J Obstet Gynecol* 1990; **162:** 615–20.
19 Begg CB. Biases in the assessment of diagnostic tests. *Stat Med* 1987; **6:** 411–23.
20 Hennekens CH, Buring JE. Epidemiology in medicine. Boston: Little, Brown and Company, 1987.
21 Riegelman RK, Hirsch RP. Studying a study and testing a test, 2nd edn. Boston: Little, Brown and Company, 1989.
22 Gotzsche PC, Olsen O. Is screening for breast cancer with mammography justifiable? *Lancet* 2000; **355:** 129–34.
23 Olsen O, Gotzsche PC. Cochrane review on screening for breast cancer with mammography. *Lancet* 2001; **358:** 1340–42.
24 Horton R. Screening mammography: an overview revisited. *Lancet* 2001; **358:** 1284–85.