

Screening tests

Introduction

You may wonder why we have a separate chapter on screening tests. After all, now that you have learned about sensitivity, specificity, likelihood ratios, Receiver Operating Characteristic curves and so forth, it seems like you should be well equipped to evaluate screening tests. However, whereas diagnostic tests are done on sick people to determine the cause of their symptoms, screening tests are generally done on healthy people with a low prior probability of disease. The problem of false positives and possible harms of unnecessary treatment looms larger. The questions of whether the patient benefits from being diagnosed and whether this benefit justifies the possible harms and costs of the test are more salient for screening. Finally, because decisions about screening are often made at the population level, political and other factors may be more influential. Thus, in this chapter, we focus explicitly on the question of whether doing the test improves health, not just whether it gives the right answer, and we pay particular attention to biases and nonmedical factors that can lead to excessive screening.¹

Definition and types of screening

Our favorite definition of screening is that suggested by Eddy (1991): “*the application of a test to detect a potential disease or condition in people with no known signs or symptoms of that disease or condition.*” The “test” being applied may be a laboratory test or x-ray, or it may be nothing more than a standard series of questions, as long as the goal is to detect a disease or condition of which the patient has no known symptoms.

¹ We do not wish to come across as complete screening nihilists. In fact, both of us have loved ones whose lives we believe may have been saved by screening. However, this is an area where we are concerned that enthusiasm has sometimes exceeded science, where there is a potential for harm, and where we see a growth industry that could consume ever greater resources with diminishing return. Hence our emphasis here tends to be on taking a critical approach to studies of screening, and on not overestimating the value of screening.

Table 6.1. Types of screening

	Unrecognized symptomatic disease	Presymptomatic disease	Risk factor
Examples	• Refractive errors in children • Depression • Iron deficiency	• Syphilis • Neonatal hypothyroidism • Many types of cancer • Glaucoma • Abdominal Aortic Aneurysm	• High blood pressure • High blood cholesterol
Number Labeled	Few	Few	Many
Number Treated	Few	Few	Many
Duration of Treatment	Varies, may be short	Varies, may be short or long	Usually long
Number Needed to Treat	Few	Few	Many
Ease of Showing Benefit	Often easy	Often difficult	Usually very difficult
Potential for Harm	False positives	• False positives • Pseudodisease • Labeling	• Risks from treatment, including delayed adverse effects • Labeling

This definition has two advantages over the definitions you will see elsewhere, which specify that screening involves “testing for asymptomatic disease.” First, “*no known symptoms*” is not quite the same as asymptomatic, because some people may have symptoms they do not recognize as such. Second, the Eddy definition includes testing not just for diseases, but for “*conditions*.” The goal of many screening tests is not to detect disease, but to detect risk factors – that is, to detect the condition of being at increased risk for one or more diseases.

Based on this definition, we can divide screening into three types:

- Screening for *unrecognized symptomatic disease*,
- Screening for *presymptomatic disease*, and
- Screening for *risk factors*.

The goals of these types of screening differ, thus the study designs, numbers of subjects, and amount of time needed to study them differ as well (Table 6.1). There is, however, some overlap between these categories. For example, glaucoma may be asymptomatic or cause unrecognized visual field loss, and osteoporosis might be considered a disease or just a risk factor for fractures.

Screening for unrecognized symptomatic disease is generally the most easily evaluated type of screening, because both the accuracy of the test and the benefits of early detection can be assessed in relatively short-term studies, often with modest sample sizes. Vision screening in children is a good example: children who have trouble with the eye chart are referred for further evaluation. If they are confirmed to have refractive errors, glasses are prescribed. No randomized trials are needed to tell that glasses will help the child see better, because the effect is immediate. Other examples

of this type of screening are screening for depression or iron deficiency anemia. When patients are already symptomatic, demonstrating a benefit from identifying and treating them does not require a long trial with many subjects.

Screening for presymptomatic disease is harder to assess. As is the case with unrecognized symptomatic disease, because the disease is already present at the time of screening, the accuracy of the screening test can be measured in the present, without a long follow-up period. But because the disease is initially asymptomatic, demonstrating benefits of treatment generally will require a follow-up study (often a randomized trial), to show that early diagnosis and treatment of disease reduces the frequency or severity of symptoms later. Examples include screening for cystic fibrosis, abdominal aortic aneurysms (AAAs), and breast cancer. On the other hand, if the natural history and pathophysiology of the disease are clear and the effects of treatment are dramatic (e.g., as with screening for syphilis or neonatal hypothyroidism), randomized trials of treatment may not be needed.

It is most difficult to decide whether to screen for risk factors for disease, because both the ability of the test to predict disease and the ability of treatment to prevent it generally must be assessed by using longitudinal studies, often with very large sample sizes. The first step is quantifying how well the measurement of the risk factor (e.g., a blood cholesterol measurement) predicts the risk of disease (heart attacks or strokes); the second step involves determining whether and by how much treatment lowers that risk. Because deaths or other serious events occur in only a small proportion of subjects (even for relatively common diseases like heart disease), this second step may require following many thousands of subjects for many years. An intermediate step, determining how well treatment lowers the level of the risk factor, is generally insufficient, because (as we will discuss in the next section) lowering the level of a risk factor may not lead to the expected lowering of the risk of disease (Guyatt et al. 2008). A dramatic example of this was the Cardiac Arrhythmia Suppression Trial, in which patients were screened for premature ventricular contractions (PVCs), a risk factor for sudden death after a heart attack. The PVCs were diminished by treatment with antiarrhythmic drugs, but unfortunately, this did not translate into fewer sudden deaths. In fact, the death rate was nearly 3 times higher in those treated, leading to an estimated 50,000 excess deaths in the U.S. (Moore 1995).

There is another important difference between screening for risk factors and screening for diseases. Most of what we learned about quantifying the accuracy of diagnostic tests in the previous chapters does not work well for risk factors. Sensitivity, specificity, and prior and posterior probability all refer to “prevalent disease”: disease that the person either already has or does not have. The time dimension for study designs for measuring these parameters is generally cross-sectional (measurements made at about the same time) as opposed to longitudinal (measurement of predictor followed by measurement of outcome).² On the other hand, for risk factor screening, we are generally trying to predict incidence (not

² As discussed in Chapter 5 under “Double Gold Standard Bias,” sometimes studies of the *accuracy* of cancer screening tests include a longitudinal component (for those who test negative). Studies of the *efficacy* of these tests have to be longitudinal.

prevalence) of disease. Thus, it is awkward to speak about sensitivity and specificity when quantifying the accuracy of risk factor screening tests, because the gold standard cannot be immediately applied. We will discuss this issue at greater length in the next chapter, which is about prognostic tests.

Importance of a critical approach to screening tests

Possible harms from screening

Although screening tests and resulting treatments, when properly selected and done, may have substantial benefits, there are also significant possible harms from screening. The potential to do harm is particularly great for risk factor-screening tests, because the number treated and duration of treatment may be much greater than for other screening tests. Some of the possible harms of screening apply to all persons screened, some only to those with specific test results, and others extend beyond those screened. These possible harms from screening, although perhaps generally underappreciated, are not conceptually difficult, so we will just list them with examples in Table 6.2, rather than discuss them at length.

Reasons for excessive screening

The possible costs and risks of screening are more than sufficient to justify a cautious approach. But there is another reason as well: awareness of the strong forces likely to lead to excessive screening. The main force may be the desire to do good – to do something that will help people live longer and healthier lives. But other forces tending to increase screening are worth considering as well (Table 6.3). Unlike the potential market for tests and treatments for symptomatic diseases, which is limited by the prevalence of the symptoms, the potential market for screening tests and resulting treatments has no such limits. The number of people at risk for each disease times the number of years for which they are at risk creates a vast potential market for screening tests, including the machines and personnel required to do them. The “patients” identified by screening become a similarly vast market for the drugs or other interventions intended to reduce their risk. In the case of disease screening, the market for treatments is limited by the number of people found to have the disease. The market for treatments for risk factors, in contrast, has no such limits, as there may be a measurable (or imagined) health benefit to treatment even at levels of the risk factor that are very prevalent in the population. Thus, we should not be surprised that companies selling products related to screening tests or to treating the diseases they are intended to diagnose or hospitals that have invested in these technologies should be very interested in moving the public toward more screening.

Pressure for increased screening does not arise solely from for-profit companies. For academic researchers like us, the greater the number of people who have, get, or worry about our disease of interest, the greater the importance of the research and the researcher, and the greater the opportunities for funding, collaborators, publications, and travel. Similarly, nonprofit organizations (like the American Liver

Table 6.2. Possible harms from screening

Group at risk or affected and type of harm	Examples
I. Costs and risks to those tested	
A. Everyone tested	
Time, cost of test	• CT scan for early lung cancer • Genetic testing for predisposition to breast and ovarian cancer
Pain, discomfort, anxiety, or embarrassment from the screening test or anticipation thereof	• Venipuncture • Digital rectal examination • Sigmoidoscopy • Mammography
Late adverse effects	• Cancer from radiation for mammography (Law and Faulkner 2001)
B. People with a negative test result	
Inappropriate reassurance leading to delay in diagnosis of target disease (false negative) or to unhealthy decisions with regard to other risk factors (false or true negative)	• Delay in evaluation of hearing loss in baby with falsely normal newborn hearing screen • Patients with normal cholesterol levels deciding they do not need to exercise or stop smoking
C. People with a positive test result	
Time, cost, pain, discomfort, anxiety, and complications of follow-up testing – generally much worse than costs and risks of initial tests	• Breast or prostate biopsies • Perforation from colonoscopy following fecal occult blood testing
Costs and risks of treatment for those testing positive; may exceed benefits, even in “true positives”	• Increased fractures when osteoporosis is treated with sodium fluoride (Riggs et al. 1990) • Increased mortality from use of clofibrate for high blood cholesterol (WHO 1980) • Increased mortality in patients with asymptomatic PVCs after myocardial infarction when treated with antiarrhythmic drugs (Epstein et al. 1993)
Unnecessary treatment of “pseudodisease”	• Prostatectomies, mastectomies, or lung resections for biopsy-proven cancer that would not have caused problems anyway
Loss of privacy or insurability	• Testing for hepatitis C, HIV, or syphilis
Labeling or other psychological distress; failure to be reassured after normal follow-up testing	• Increased absenteeism in steelworkers found to have hypertension (Haynes et al. 1978) • Self-restriction of activities following low bone density measurements in elderly women (Rubin and Cummings 1992) • Altered parent–infant relationship following false-positive newborn hypothyroidism screening (Fyro and Bodegard 1987) • Continued anxiety following false-positive mammograms (Barton et al. 2004)

Table 6.2 (*continued*)

Group at risk or affected and type of harm	Examples
II. Costs and risks to others	
Injuries to testing personnel	• Radiation, needle sticks, etc.
Harms to contacts, partners, family members	• False-positive or false-negative tests for sexually transmitted diseases • Finding of infant blood group inconsistent with supposed paternity
Time cost of patients and physicians informing themselves about tests the patient chooses not to have done	• Expensive screening tests being marketed directly to consumers (Lee and Brennan 2002)
Removal of resources from where they would do more good (Eddy 1997)	• Mammography for the wealthy in poor countries (Braveman and Tarimo 1994)

Foundation or the American Cancer Society) tend to favor screening tests for their disease or organ system. Aside from any medical benefits from screening, it has the potential to identify large groups of people likely to be interested in the work of the organization and to make donations. As discussed below, some of those most in favor of screening may believe that their lives were saved by screening tests.

Although managed care organizations might be expected to favor limiting screening (because, in most cases, it increases their costs), even they have reasons to encourage screening. Performance of screening tests, because it is popular with the public, easily measured, and little affected by complex, variable presentations of patients with symptoms, tends to be disproportionately weighted on quality “report cards” like the Health Plan Employer Data and Information Set (HEDIS 2009). In addition, health plans may find it advantageous to emphasize preventive care, if doing so attracts healthier patients.

Finally, the general public tends to be supportive of screening programs. Part of this is wishful thinking. We would like to believe that bad things happen for a reason, and that there are things we can do to prevent them (Marantz 1990). We also tend to be much more swayed by stories of individual patients (either those whose disease was detected early or those in whom it was found “too late”) than by boring statistics about risks, costs, and benefits (Newman 2003). Because, at least in the U.S., there is no clear connection between money spent on screening tests and money not being available to spend on other things, the public tends not to be swayed by arguments about cost efficacy (Daniels 1986; Mariner 1995; Eddy 1997). In fact, in the general public’s view of screening, even wrong answers are not necessarily a bad thing.

Schwartz et al. (2004) did a national telephone survey of attitudes about cancer screening in the U.S. They found that 38% of respondents had experienced at least one false-positive screening test. Although more than 40% of these subjects referred to that experience as “very scary” or the “scariest time of my life,” 98% were glad they had the screening test! As our gynecologist colleague George Sawaya (who

Table 6.3. Powerful non-medical forces that could lead to increased enthusiasm for screening

Stakeholder	Reasons to favor screening ^a	Example
Companies selling tests or testing equipment	Sell more tests or testing machines	• Osteoporosis testing machines • Office cholesterol machines • Private companies marketing genetic tests or body scans
Companies selling products to treat the condition	Sell more product	• Schering–Plough has funded public awareness campaigns to encourage PSA and hepatitis C screening [they make Eulexin (flutamide) used to treat prostate cancer and Intron (interferon) used to treat hepatitis C]
Clinicians or hospitals who diagnose or treat the condition	More patients, procedures, income, importance	• Gynecologists tend to recommend more Pap smears and urologists more PSA testing than generalists • Thoracic surgeons or radiologists may favor CT screening for lung cancer
Politicians	• Appear sympathetic to those who have or are at risk of the condition • Be responsive to special interests or contributors	• U.S. Senate vote 98–0 overturning National Cancer Institute panel's recommendations that mammography decisions for 40- to 49-year-old women be individualized (Ernst 1997)
Nonprofit disease research and advocacy groups	• Increased importance of disease and hence of organization's work • More people with the disease or risk factor who become interested are active constituents and potential donors • Increase attractiveness for donations from industry	• American Liver Foundation Hepatitis C Screening promotion (paid for by Schering–Plough) • American Cancer Society recommendations for cancer screening often more aggressive than those of the U.S. Preventive Health Services Task Force
Academics who study the condition	• Increased importance, recognition, and funding for research for the condition • Accessible funding from industry	• Hypercholesterolemia, osteoporosis, and virtually everything else
Patients/the public	• Wishful thinking – wanting to believe bad things happen for a reason and that there are things we can do to prevent them • Individualistic perspective – lack of concern about costs if someone else is paying them	• Belief in and demand for PSA testing and mammography disproportionate to evidence of benefit • View that those (even elderly) not wishing to be screened are “irresponsible” (Schwartz et al. 2004)

^a Aside from the desire to help people, which is assumed to be a reason for all.

studies Pap smears) puts it, “the patients are so grateful when we come to the rescue and put out the fire that they forget that we were the ones who set it in the first place.”

We know of no similar survey that addresses how patients feel about false-negative results, but some may still be happy they had the test. Patients whose cancer is diagnosed at a late stage who did not get screened are likely to wonder if they could have been saved if they had been screened. Those who were screened and were (presumably falsely) negative will at least have the comfort of knowing it was not their fault and of not being blamed by their doctors, family, and friends (Marantz 1990). Another disturbing result of the survey by Schwartz et al. was that, even though (as of 2002) the U.S. Preventive Health Services Task Force felt that evidence was insufficient to recommend prostate cancer screening, more than 60% of respondents said that a 55-year-old man who did not have a routine PSA test was “irresponsible,” and more than a third said this for an 80-year old man! Thus, regardless of the efficacy of screening tests, they have become an obligation if one does not wish to be blamed for getting some illnesses.

Reasons for underscreening

We have emphasized many reasons to worry about excessive screening, but insufficient screening can occur as well. All of the potential problems that screening can cause (Table 6.2) are reasons why it might not be done even when a net benefit could be projected: it costs money, takes time, patients may fear discomfort or loss of privacy, etc. If screening leads to improved health but net increases in costs, managed care organizations could deliberately make it difficult to do the tests. Some hospitals may lack the confidence, competence, and capacity to deal with positive results. In order to make screening work, the systems for dealing with positive results and providing services to identified patients need to be in place.

Critical appraisal of studies of screening tests

The big picture

The general idea of a lot of screening (and diagnostic tests) is that, if you do the test, it will help you diagnose the disease, and if you diagnose the disease, you will improve outcome. If we want to know whether to do a test, we would really like to know whether people who get the test have a better outcome than people who do not (Fig. 6.1). Unfortunately, most studies do not address that question directly. Instead, studies either 1) correlate testing or test results with diagnosis or stage (e.g., studies that estimate diagnostic yield, sensitivity, specificity, Receiver Operating Characteristic curves, likelihood ratios, etc.) or 2) correlate diagnosis or stage with ultimate outcome. The latter studies are those susceptible to lead- or length-time biases, which we will discuss below.

For simplicity, assume that we are screening for presymptomatic disease, and in a subset of patients, the disease is fatal a predictable time period after symptoms develop (Fig. 6.2). The disease is detectable by screening some time after its biological

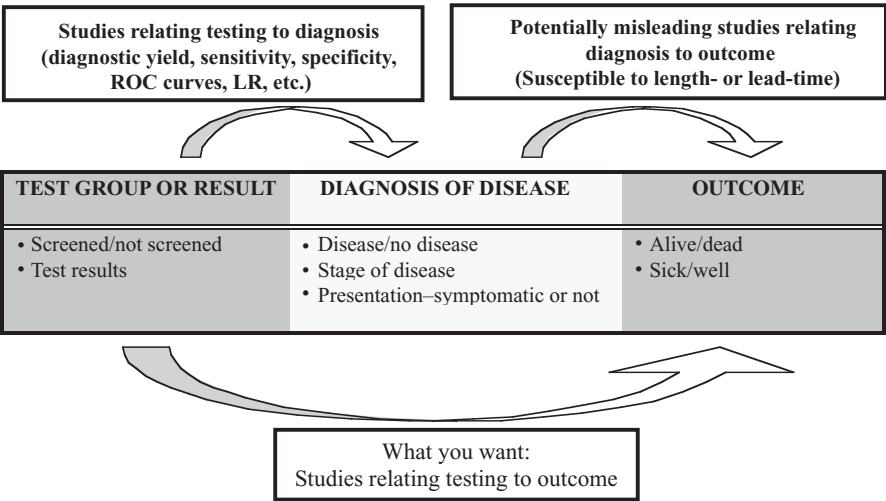


Figure 6.1 Predictor and outcome variables in studies of screening. The best studies bridge the gap and compare outcomes in those screened and not screened.

onset but before symptoms develop (Herman et al. 2002). The rationale for screening is that intervention during this detectable preclinical phase (“latent phase”) forestalls or prevents symptom onset and prolongs life.

The best way to assess screening is to randomize people into two groups: one that receives the screening test and one that doesn’t. As we will discuss in Chapter 9, randomization ensures against systematic differences between the two groups with respect to disease risk, health habits, and other factors that can affect the outcome of interest (e.g., life expectancy). Both the screened and unscreened groups will include

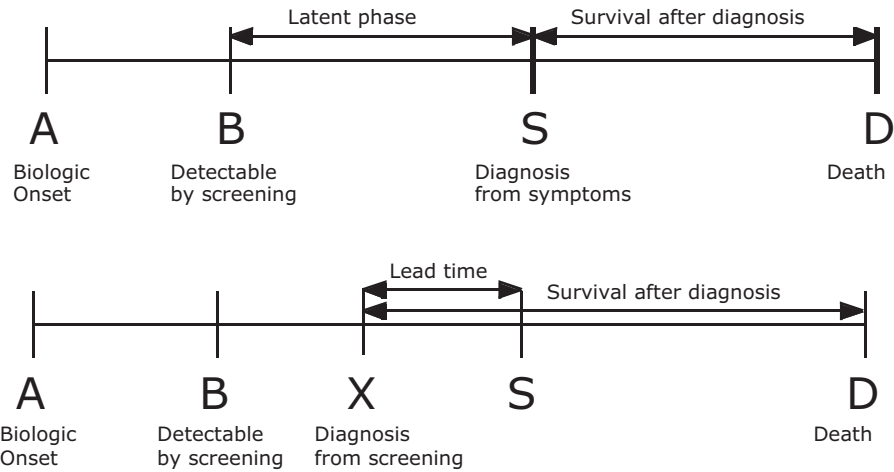


Figure 6.2 *Upper Panel:* Natural history of disease in people affected by lead-time bias. Disease progresses from biological onset (A) until it is detectable by screening (B), through the latent phase, until diagnosis from symptoms (S), through the clinical phase, until death (D). *Lower Panel:* Lead-time bias: early detection does not prolong overall survival, but does prolong the time from diagnosis to death. The period between diagnosis from screening (X) and when diagnosis would have occurred from symptoms (S) is the lead time.

many individuals who do not have the disease in question. If screening affects the life expectancy of these nondiseased individuals at all, it is likely to have a negative effect.³ Both groups will also include individuals with the disease. In the unscreened group, the disease will be diagnosed at symptom onset, but in the screened group, at least some cases of the disease may be diagnosed by screening. If screening genuinely allows interventions that forestall or prevent symptoms and prolong life, and if this effect exceeds the negative effect of screening on nondiseased individuals, the overall death rate should be lower and life expectancy should be longer in the screened group. So, the ideal study would be a randomized trial of screening versus no screening that compares the overall mortality between the two randomization groups. Although such a study may not be practical, keeping this ideal study design in mind can help you understand biases common in observational studies, to be discussed below.

Observational studies of screening tests

Observational studies of screening deviate in various ways from the ideal randomized trial of screening versus no screening. Some compare the outcome (such as death from prostate cancer) among persons who have been screened with those who have not been screened, but the assignment to the screened and unscreened groups is not random, and there are systematic differences between them. Others limit the comparison to those with the disease. The screened patients with the disease (even if it was missed on screening and diagnosed by symptoms) may be compared with the unscreened patients with the disease (all of whom were diagnosed by symptoms). Finally, those diagnosed by screening may be compared with those diagnosed by symptoms (whether or not they were ever screened). Observational studies are subject to several important biases that can make screening tests appear to be more beneficial than they are.

Volunteer bias

When assignment to the screening group is not random, comparisons between people who are and are not screened may be invalid because people who volunteer for screening are generally different from people who do not. The screened group may be at higher risk of poor outcome, if, for example, they volunteered for screening because of a symptom they did not disclose (people with symptoms are generally excluded from studies of screening tests). More typically, they may be at lower risk of poor outcome, because of healthier habits or better access to health care. For example, Otto et al. (2004) compared the number of deaths over a 5-year period of men who agreed to be in a randomized trial of prostate cancer screening with the number expected for that population and found that it was 13% lower. To address volunteer bias, investigators measure and attempt to control for factors that might be associated with both receiving the screening test and outcome (e.g., family history, education level, number of health maintenance visits, etc.), but the only way to

³ This is almost always the case, but a possible exception is the Multicenter Aneurysm Screening Study described in Problem 6–6.

eliminate the possibility of volunteer bias is to randomize the study subjects either to receive or not to receive the test.

Lead-time bias

Lead time is the apparent increase in survival obtained when a disease is detected before it would have become symptomatic and been detected clinically (Fig. 6.2). Lead-time bias affects the subset of the population destined to die of the disease whether or not they are screened. The trouble is, even if screening and/or treatment are completely ineffective, if you start counting years of survival from the date of diagnosis, moving the date of diagnosis earlier will make survival seem longer (Fig. 6.2). Lead-time bias is thus a problem when postdiagnosis survival is compared between persons whose disease was detected by screening and those whose disease was detected by development of symptoms. Lead-time bias cannot occur in a randomized trial of screening or a cohort study that compares the entire screened group with the entire unscreened group; survival time is counted from either randomization (in the trial) or inception (in the cohort), rather than from date of diagnosis.

Length-time bias

This bias gets its name from the fact that heterogeneity in the natural history of a disease can lead to subjects spending a variable **length of time** in the presymptomatic phase. A clearer name for it could be “different natural history bias.” Length-time bias may occur whether the study considers one-time screening or screening at regular intervals, but only when it compares survival time from diagnosis between those diagnosed by screening versus those diagnosed by symptoms.

When thinking about length-time bias, assume that the entire population is being screened for disease; there is no unscreened group. If the disease being screened for is heterogeneous (e.g., some tumors are indolent, whereas others rapidly metastasize and kill), the cases that are more slowly progressive (and have a longer latent phase) will be preferentially diagnosed by screening. In comparison to individuals whose disease is diagnosed by symptoms, those with disease diagnosed by screening have more indolent disease, and hence show improved survival. This is illustrated in Figure 6.3 in which three of the subjects have much more rapidly progressive diseases, as represented by the compressed progression time from disease onset (A) to detectability by screening (B) to development of symptoms (S) to death (D).

Because screening tests done at any particular point in time can only get the head start on detection if they catch the disease in its latent phase (between times B and S in Fig. 6.3), patients whose diseases spend a short time in that state are less likely to be identified by screening and more likely to present with symptoms. These patients will have a poorer prognosis due to the rapidly progressive nature of their disease. Thus, the basic problem is that, although detection by the screening test will be associated with a better prognosis, the causal inference is incorrect: both early detection and the improved prognosis are due to the better expected natural history of the disease (Fig. 6.4).

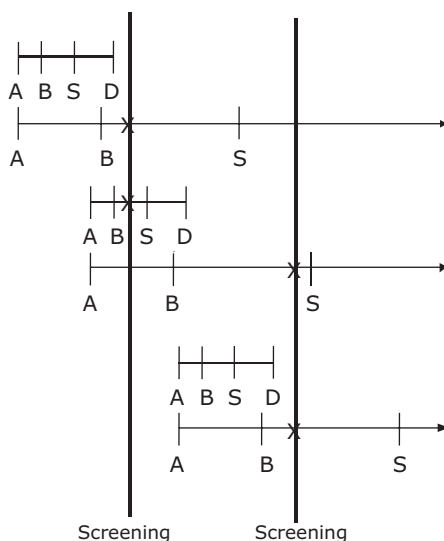


Figure 6.3 Length-time bias. If some cases of disease are rapidly progressive (indicated by short intervals between A, B, S, and D), they are less likely to be caught between B and S when a screening test is done, and hence more likely to present with symptoms. (A, biological onset; B, detectable by screening; X, detected by screening; S, symptom onset; D, death). In the figure, $3/6 = 50\%$ of the cases are rapidly progressive and have a bad prognosis. However, this is true of only $1/4 = 25\%$ of the cases detected by screening.

Length-time bias is only operative when disease is heterogeneous *and* survival from diagnosis is compared between persons whose disease was detected by screening and those whose disease was detected in other ways. Length-time bias will generally be accompanied by at least some lead-time bias. (In Fig. 6.3, the time between the vertical screening lines and points S is the contribution of lead-time bias.) However, the reverse is not always true: lead-time bias will occur even if the natural history of the disease is entirely homogeneous and there is no length-time bias.

Finally, as long as a study (randomized trial or cohort study) compares the entire screened group with the entire unscreened group (between-groups comparison), lead-time and length-time bias are not issues.

Stage migration bias

Newer, more sensitive diagnostic tests can lead to the diagnosis of disease at an earlier or milder stage, and also to patients being classified as being in a higher stage of disease than would have been known previously (Fig. 6.5). For example, a more sensitive bone scan might lead to some patients being classified as having stage IV

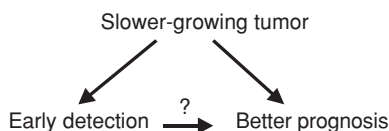


Figure 6.4 A noncausal relationship between early detection and a better prognosis is the cause of length-time bias.

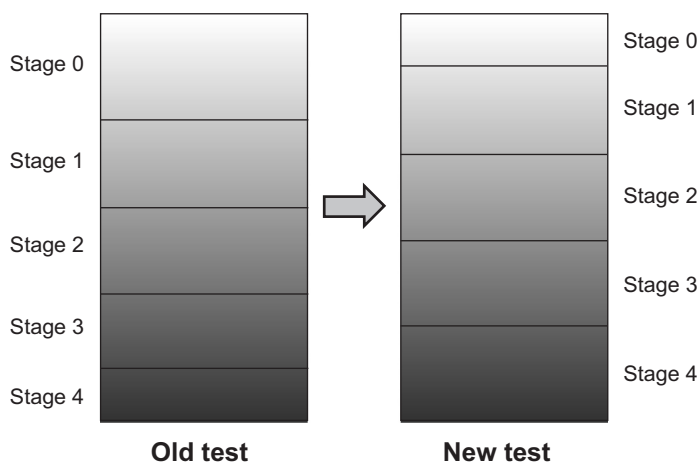


Figure 6.5 Stage migration bias. Newer, more sensitive tests lead to less severe disease and a better prognosis at each stage.

prostate cancer, when previously they would have been thought to be in a less advanced stage. These patients likely have a longer life expectancy than those with the more significant bone metastases detectable by a less sensitive scan. The end result is that stage-specific survival (e.g., survival of patients with stage IV disease before and after the new test) will appear to improve with the more sensitive test, even if no one lives longer. The survival of those at lower stages is improved by having the patients with a worse-than-average stage-specific prognosis leave their stage and be classified in a higher stage. Survival at higher stages is increased because of the entry of subjects from lower stages with better-than-average, stage-specific prognosis for their new stage. However, if a change in the distribution to more advanced stages is the cause of the improvement in stage-specific survival, overall survival will be the same (Feinstein et al. 1985). If a study reports stage-specific improvement in survival with a new screening test, comparing overall survival between screened and unscreened groups is a good way to check for stage migration bias.

Stage migration bias can also occur in the absence of changes in diagnostic testing, simply because of changes in the diagnostic criteria for different stages over time. This was demonstrated for breast cancer, when changes in classification of lymph node involvement between the 5th and 6th editions of the American Joint Committee on Cancer staging system dramatically altered stage-specific survival (Olivotto et al. 2003; Woodward et al. 2003).

Pseudodisease

In Chapter 5 on biases in studies of test accuracy, we described double gold standard bias, in which some patients could be designated as D+ on surgical pathology but as D– on clinical follow-up if they have either transient or dormant disease. For patients like this, if a positive index test leads to biopsy but a negative index test leads to clinical follow-up, the index test will always appear to give the right answer. Here, we are not worried about overestimating the accuracy of an index test, but rather,

overestimating the effectiveness of a screening program. In this context, the problem is the possibility of detecting “pseudodisease” – that is, disease that never would have affected the patient had it not been diagnosed. (This is also called overdiagnosis.) It is difficult to identify pseudodisease in an individual patient, because it requires completely ignoring the diagnosis. (If you treat pseudodisease, the treatment will always appear to be curative, and you won’t realize the patient had pseudodisease rather than real disease!) In some ways, pseudodisease is an extreme type of stage migration bias. Patients who were not previously diagnosed as having the disease are now counted as having it. Although the incidence of the disease goes up, the prognosis of those who have it improves.

We would like to believe that pathologists can look at a biopsy and reliably distinguish benign from malignant tissue. However, there is abundant evidence that this is not always the case – some tumors that microscopically are diagnosed as breast, prostate, and even lung cancers do not behave as cancerous (Welch 2004).

Lack of understanding of pseudodisease, including the lack of people who know they have had it, is a real problem, because most of us understand the world through stories (Newman 2003). Patients whose pseudodisease has been “cured” become strong proponents of screening and treatment and can tell a powerful and easily understood story about their experience. On the other hand, there aren’t people who can tell a compelling story of pseudodisease – men who can say, “I had a completely unnecessary prostatectomy,” or women who say, “I had a completely unnecessary mastectomy,” even though we know statistically that many such people exist.

The existence of pseudo-lung cancer was strongly suggested by the results of the Mayo Lung Study, a randomized trial of chest x-rays and sputum cytology to screen for lung cancer among 9,211 male cigarette smokers (Marcus et al. 2000). Because it was a randomized trial and screening should lead to early detection of lung cancer but not affect its cumulative incidence (over a sufficiently long follow-up period), the number of new lung cancers in the 2 groups should have been the same, with (if screening worked) more tumors at lower stages in the screened group and more tumors at higher stages in the control group. In fact, however, after a median follow-up of 20.5 years, there was still a highly significant 29% *increase* in the cumulative incidence of lung cancer in the screened group. There was an excess of tumors at an early, resectable stage, but no decrement in late-stage tumors. The screened group therefore had more lung “cancer” resections, but no overall decrease in lung cancer deaths. In fact, there was a trend ($P = 0.09$) toward an increase in deaths attributed to lung cancer in the screened group; Marcus et al. 2000.

Pseudodisease has been divided into two types (Black and Welch 1997). Type I pseudodisease is related to length-time bias. Just as some individuals can have particularly aggressive forms of the disease, others can have particularly indolent forms, which are detectable on screening but would never cause the patient any symptoms. Type II pseudodisease occurs even if the natural history of the disease is homogeneous. Some people with preclinical disease will die from another cause before the disease becomes symptomatic. This is why screening an octogenarian for cancer rarely makes sense.

Randomized trials of screening tests

We have said that the best way to determine whether a test is of benefit is to perform a randomized trial in which subjects are randomized to be tested or not. These have been done only for a few major screening tests, like mammography and stool occult blood testing. A drawback to randomized trials is that they may need to be very large and of long duration. Aside from the fact that the target diseases may be quite uncommon, the sample size has to be increased even further to make up for the bias toward the null (finding no effect) that occurs as a result of crossover between groups: some subjects randomized to screening will decline it, and some randomized to usual care will get screened anyway. One controversy that comes up particularly for randomized trials of screening tests is the choice of outcome variables.

Total mortality versus cause-specific mortality

There are good arguments for saying that, in order to know that a screening test is beneficial, we need to see a decrease in total mortality in a group randomized to screening as opposed to just a decrease in cause-specific mortality.⁴ Whereas mortality is easy to ascertain objectively, cause-specific mortality is subject to judgment and might be influenced by the screening test. It also may be difficult or impossible to know whether some deaths occurred as late effects of the treatment or of the screening test itself. This is a particular problem with large, population-based studies where death certificates are used. Although blinding those assigning cause of death to treatment group assignment will reduce that problem, it will not eliminate it, because screening produces information and events that become part of the patient's medical history.

Black et al. (2002) describe two major biases that result from using cause-specific rather than overall mortality as the outcome. "Sticky diagnosis bias" refers to the likelihood that, once a disease (particularly cancer) is diagnosed, deaths are more likely to be attributed to it. For example, sometimes patients die of unclear causes. If they previously had a cancer diagnosed by screening, their death would be more likely to be attributed to that cancer. The diagnosis of cancer "sticks" to the patient. This is a bias that will make a comparison of cause-specific mortality look *worse* for screening. Those in the screened group will tend to have higher cause-specific mortality attributed to the cancer they were screened for, even if they die of other conditions.

On the other hand, another possibility is what Black et al. call "slippery linkage bias." This occurs when the linkage between deaths due to screening, follow-up, or treatment "slips," so that deaths that may have occurred as a result of screening are not counted in the cause-specific mortality for the disease. This can occur from late complications from the screening test itself or from complications of treatment. For example, if a patient in a randomized trial of fecal occult blood testing to screen for colon cancer eventually dies after a series of complications that began with a

⁴ This discussion is focused on screenings whose goal is to prolong life.

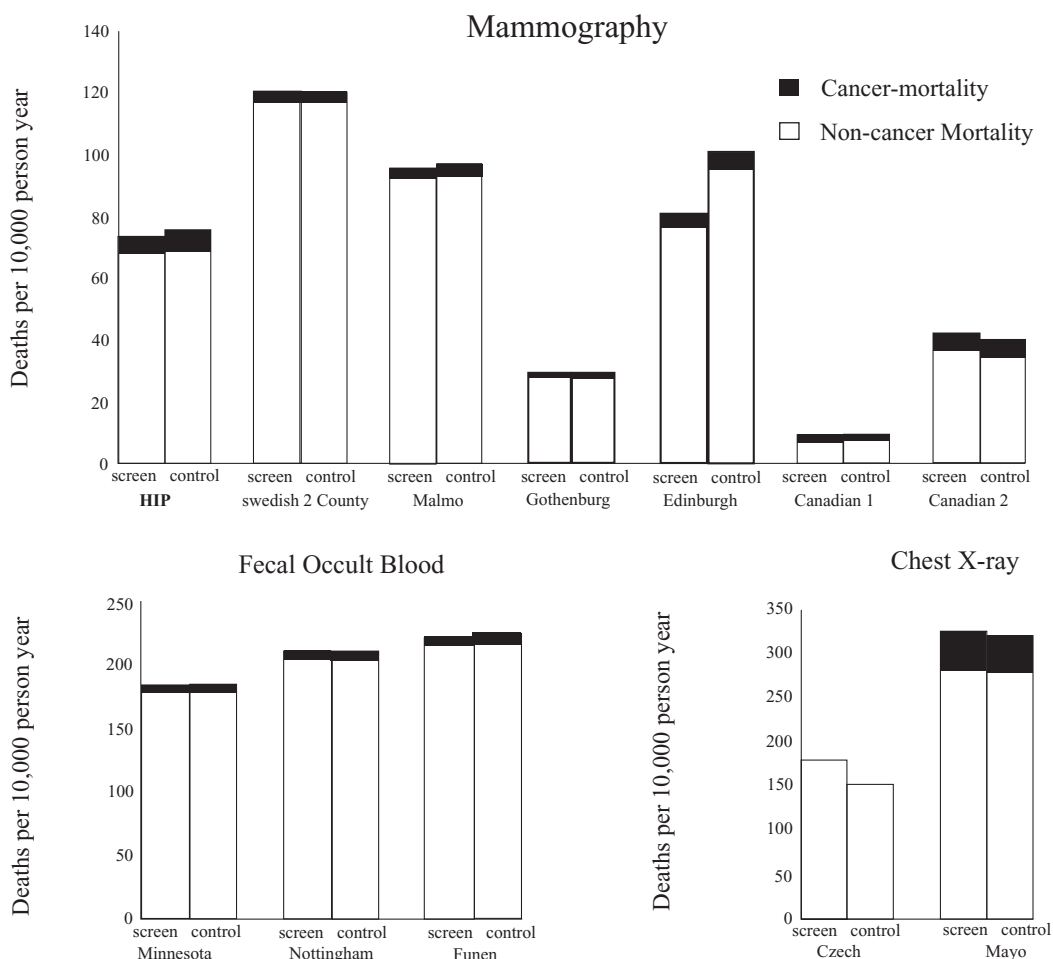


Figure 6.6 Cancer and noncancer mortality in randomized trials of cancer screening. From Black et al. (2002).

colonic perforation during colonoscopy for a false-positive fecal occult blood test, the death would not be counted as a colon cancer death, although it was caused by screening for colon cancer. Similarly, there is good evidence from randomized trials that radiation therapy for breast cancer is associated with a late increase in coronary heart disease death rates (Early Breast Cancer Trialists' Collaborative Group 2000). These deaths may occur with greater frequency in screened women, who are more likely to receive radiation, but will be difficult or impossible to link to screening.

There really is only one problem with using total mortality as an endpoint in screening trials, but it is a big one: deaths from causes unrelated to screening or the target condition will generally swamp deaths affected by screening, making it virtually impossible to identify beneficial (or harmful) effects. This is illustrated graphically in Figure 6.6. When only a few percent of deaths are likely to be due to the target condition, it is difficult to detect any effect on total mortality.

Biases that make screening tests look worse

We have focused on biases that tend to make tests look better than they really are.⁵ This is because, at least historically, people doing studies of tests have often been advocates of the tests, so these were the biases to be most concerned about. But as more people (like us) who are skeptical about tests write articles about them, we should consider biases that can make tests look *worse* in a study than they might be in practice:

1. **Inadequate power:** It is easy to fail to find any benefit of a test if your sample size is too small or duration of follow-up too short. Another way to reduce your power is to look at total mortality, rather than cause-specific mortality.
2. **People performing the test are unskilled:** If the test takes some skill, by studying it in a setting where it is not done well, you can find that it doesn't work.
3. **People who test positive are not properly followed up or treated:** For example, if one wanted to show that fecal occult blood testing was worthless, one could study it in a setting where many patients were not followed up or where those who were followed up were not well treated.

Back to the big picture

So what should we do to avoid recommending screening tests that might do harm, while not taking a completely nihilistic stance? First, every effort must be made to perform studies that answer the main question of whether screening leads to better outcomes among patients. Because the ideal study design (randomized trial with total mortality as the outcome) is rarely feasible, keep several criteria in mind when considering the alternatives. First, studies should attempt to capture morbidity and mortality due to the screening test itself. Second, we should recognize that the need to examine total mortality varies with the screening test and the intervention. For fecal occult blood screening, for example, where the test involves no exposure to radiation and the treatment is primarily surgical, we have fewer concerns about late adverse effects than with mammography. Treatment resulting from mammography may involve radiation and/or systemic treatment with hormone analogs or chemotherapeutic agents that may have significant effects on causes of death other than breast cancer that may not be apparent for years. Finally, large, relatively simple, randomized trials and, when possible, much lower cost observational alternatives, like natural experiments, are desirable to address specific concerns about increases in mortality from causes other than the disease being screened for. Randomized trials and their alternatives will be discussed in Chapters 9 and 10.

Summary of key points

1. The purpose of screening tests is to identify unrecognized symptomatic disease, presymptomatic disease, or risk factors for disease.

⁵ Except Sticky Diagnosis Bias, which makes the screening test look worse in terms of cause specific mortality.

2. A critical approach to screening tests is important because screening tests can cause harm and because there are many forces and biases that tend to favor screening.
3. The most definitive way to assess screening tests is with randomized trials that have total mortality as the outcome, but these are seldom feasible, necessitating care when interpreting observational studies and trials focused on cause-specific mortality.

References

- Barton, M., D. Morley, et al. (2004). "Decreasing women's anxieties after abnormal mammograms: a controlled trial." *J Natl Cancer Inst* **96**: 529–38.
- Black, W. C., D. A. Haggstrom, et al. (2002). "All-cause mortality in randomized trials of cancer screening." *J Natl Cancer Inst* **94**(3): 167–73.
- Black, W. C., and H. G. Welch (1997). "Screening for disease." *AJR Am J Roentgenol* **168**(1): 3–11.
- Braveman, P., and E. Tarimo (1994). *Screening in Primary Health Care: Setting Priorities with Limited Resources*. Geneva, World Health Organization.
- Daniels, N. (1986). "Why saying no to patients in the United States is so hard. Cost containment, justice, and provider autonomy." *N Engl J Med* **314**(21): 1380–3.
- Early Breast Cancer Trialists' Collaborative Group (2000). "Favourable and unfavourable effects on long-term survival of radiotherapy for early breast cancer: an overview of the randomised trials." *Lancet* **355**(9217): 1757–70.
- Eddy, D. (1991). *Common Screening Tests*. Philadelphia, PA, American College of Physicians.
- Eddy, D. M. (1997). "Breast cancer screening in women younger than 50 years of age: what's next?" *Ann Intern Med* **127**(11): 1035–6.
- Epstein, A. E., A. P. Hallstrom, et al. (1993). "Mortality following ventricular arrhythmia suppression by encainide, flecainide, and moricizine after myocardial infarction. The original design concept of the Cardiac Arrhythmia Suppression Trial (CAST)." *JAMA* **270**(20): 2451–5.
- Ernster, V. L. (1997). "Mammography screening for women aged 40 through 49—a guidelines saga and a clarion call for informed decision making." *Am J Public Health* **87**(7): 1103–6.
- Feinstein, A. R., D. M. Sosin, et al. (1985). "The Will Rogers phenomenon. Stage migration and new diagnostic techniques as a source of misleading statistics for survival in cancer." *N Engl J Med* **312**(25): 1604–8.
- Fyro, K., and G. Bodegard (1987). "Four-year follow-up of psychological reactions to false positive screening tests for congenital hypothyroidism." *Acta Paediatr Scand* **76**(1): 107–14.
- Guyatt, G., D. Rennie, et al. (2008). *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice, 2nd Edition*. New York, NY, McGraw Hill Medical, pp. 113–143.
- Haynes, R. B., D. L. Sackett, et al. (1978). "Increased absenteeism from work after detection and labeling of hypertensive patients." *N Engl J Med* **299**(14): 741–4.
- HEDIS (2009). "The Health Plan Employer Data and Information Set (HEDIS®)." Available from: <http://www.ncqa.org/tabid/784/Default.aspx>. Accessed 10/3/08.
- Herman, C. R., H. K. Gill, et al. (2002). "Screening for preclinical disease: test and disease characteristics." *AJR Am J Roentgenol* **179**(4): 825–31.
- Law, J., and K. Faulkner (2001). "Cancers detected and induced, and associated risk and benefit, in a breast screening programme." *Br J Radiol* **74**(888): 1121–7.

- Lee, T., and T. Brennan (2002). "Direct-to-consumer marketing of high-technology screening tests." *N Engl J Med* **346**(7): 529–531.
- Marantz, P. R. (1990). "Blaming the victim: the negative consequence of preventive medicine." *Am J Public Health* **80**(10): 1186–7.
- Marcus, P. M., E. J. Bergstralh, et al. (2000). "Lung cancer mortality in the Mayo Lung Project: impact of extended follow-up." *J Natl Cancer Inst* **92**(16): 1308–16.
- Mariner, W. K. (1995). "Rationing health care and the need for credible scarcity: why Americans can't say no." *Am J Public Health* **85**(10): 1439–45.
- Moore, T. J. (1995). *Deadly Medicine: Why Tens of Thousands of Heart Patients Died in America's Worst Drug Disaster*. New York, Simon & Schuster.
- Newman, T. B. (2003). "The power of stories over statistics." *Br Med J* **327**(7429): 1424–7.
- Olivotto, I. A., P. T. Truong, et al. (2003). "Staging reclassification affects breast cancer survival." *J Clin Oncol* **21**(23): 4467–8.
- Otto, S. J., F. H. Schroder, et al. (2004). "Low all-cause mortality in the volunteer-based Rotterdam section of the European randomised study of screening for prostate cancer: self-selection bias?" *J Med Screen* **11**(2): 89–92.
- Riggs, B. L., S. F. Hodgson, et al. (1990). "Effect of fluoride treatment on the fracture rate in postmenopausal women with osteoporosis." *N Engl J Med* **322**(12): 802–9.
- Rubin, S. M., and S. R. Cummings (1992). "Results of bone densitometry affect women's decisions about taking measures to prevent fractures." *Ann Intern Med* **116**(12 Pt 1): 990–5.
- Schwartz, L. M., S. Woloshin, et al. (2004). "Enthusiasm for cancer screening in the United States." *JAMA* **291**(1): 71–8.
- Welch, H. G. (2004). *Should I Be Tested for Cancer? Maybe Not, and Here's Why*. Berkeley, CA, University of California Press.
- WHO (1980). "W.H.O. cooperative trial on primary prevention of ischaemic heart disease using clofibrate to lower serum cholesterol: mortality follow-up. Report of the Committee of Principal Investigators." *Lancet* **2**(8191): 379–85.
- Woodward, W. A., E. A. Strom, et al. (2003). "Changes in the 2003 American Joint Committee on Cancer staging for breast cancer dramatically affect stage-specific survival." *J Clin Oncol* **21**(17): 3244–8.

Chapter 6 Problems: screening tests

1. For each of the following study descriptions, name and briefly explain the bias most likely to account for the results.
 - a) A study on early treatment of lupus-related kidney disease (nephritis) compared patients who had a kidney biopsy early in their clinical course with patients biopsied late in their course. The study measured time to renal failure from the time of the biopsy and found that those biopsied earlier had a longer time to renal failure.
 - b) One way to screen for *colon* cancer is to have patients collect a small amount of stool on a Hemoccult® card that can be chemically tested for the presence of blood. A study of fecal occult blood screening finds a dose-response between the number of Hemoccult cards returned and decreased risk of *lung* cancer death.

- c) A new policy requires all asthmatics to have a $p\text{CO}_2$ measured in the emergency department, with automatic admission to the ICU rather than the ward if the $p\text{CO}_2$ exceeds 45 mm Hg. (A $p\text{CO}_2 > 45$ mm Hg is an indication of increased severity.) Death rates from asthma in both the ICU and on the ward decline.
2. A 2006 paper in the *New England Journal of Medicine* (Henschke et al. 2006) reported 88% estimated 10-year lung cancer-specific survival among 412 patients with pathologically proven stage I lung cancer detected by CT screening. They contrasted this with about 5% survival among lung cancer patients in general. Which of the following statements about this finding are true? [True/False for (a–d); include explanation for (e).]
- Because survival was counted beginning at the time of diagnosis, some improvement in survival would be expected due to *lead-time bias*, even if there was no advantage to early detection.
 - “Sticky diagnosis bias” could be an explanation for these findings in favor of screening.
 - Because tumors identified on screening tests tend to be slower-growing and have a more benign prognosis than tumors that present with symptoms, *length-time bias* could contribute to these favorable results.
 - Overdiagnosis* of lung cancer (pseudodisease) in these participants is unlikely because all diagnoses were confirmed pathologically.
 - These results show that CT screening reduces lung cancer mortality by 80% or more. Explain your answer.
3. The Multicenter Aneurysm Screening Study (Ashton et al. 2002) was a randomized trial of the effectiveness of ultrasound screening for Abdominal Aortic Aneurysm (AAA) in reducing aneurysm-related mortality. Men aged 65–74 were randomized either to be invited to receive a screening abdominal ultrasound scan or not. Aneurysm-related and overall mortality in the two randomization groups are reported below:

Multicenter Aneurysm Screening Study

	N	AAA-Related Deaths	%	Total Deaths	%
Invited	33,839	65	0.19%	3,750	11.08%
Not Invited	33,961	113	0.33%	3,855	11.35%
Total	67,800	178		7,605	

- Does screening appear to be effective in reducing aneurysm-related deaths?
- You can see that, in those invited for screening, there were 48 fewer AAA deaths ($113 - 65$) and 105 fewer total deaths ($3855 - 3750$). Thus, there were $105 - 48 = 57$ fewer *non*-AAA deaths in those invited for screening. Which of the following do you think are the most likely explanations for this: Volunteer or Selection Bias; Lead-Time Bias; Length-Time Bias; Stage Migration Bias;

Misclassification of Outcome; Misclassification of Exposure; Cointerventions; or Chance?

- c) The authors also did a *within groups* analysis in the invited group only, comparing those who did and did not get the ultrasound scan. Results are summarized below, same format as before:

Multicenter Aneurysm Screening Study – Invited Group Only

	N	AAA Death	%	Total Death	%
Scanned	27,147	43	0.16%	2,590	9.54%
Not Scanned	6,692	22	0.33%	1,160	17.33%
Total	33,839	65		3,750	

The total mortality rate in the invited patients who were scanned (9.54%) was 45% lower than that of the invited patients who were not scanned (17.33%). Again, which of the following explanations are most likely responsible for this difference: Volunteer or Selection Bias; Lead-Time Bias; Length-Time Bias; Stage Migration Bias; Misclassification of Outcome; Misclassification of Exposure; Cointerventions; or Chance?

- d) This was a randomized trial, so the safest way to analyze the data is by group assignment – an “Intention to Treat” analysis. Nonetheless, it is sometimes of interest to compare groups according to how they were actually treated – an “As Treated” analysis. Do you believe the “As Treated” comparison of AAA deaths (not total deaths) between the scanned and not scanned patients within the Invited group is biased? Why or why not?
4. Torres et al. (1994) studied a population of 86 children who had been diagnosed and treated for posterior fossa medulloblastoma (a brain cancer). After initial treatment, these 86 children were screened for recurrence with a brain scan every 6 months, or scanned sooner if they developed symptoms or signs suggestive of recurrence. There were 23 children with recurrences: 4 were detected on interval screening and 19 presented with symptoms or signs of recurrence between surveillance scans. All 23 recurrences resulted in death. In the group of 4 recurrences that were detected by a regular screening scan (not prompted by signs or symptoms), the median survival was 20 months. In the group of 19 recurrences that presented with symptoms or signs, the median survival was only 4 months ($P = 0.03$). In the Letters to the Editor about this article, there was some debate about whether lead-time and/or length-time bias could explain this survival difference.
- a) Could lead-time bias explain the entire survival difference? If not, how much of the survival difference could be explained by lead-time bias?
- b) Could length-time bias explain the survival difference?
5. Mastroiacovo et al. (1992) studied the all-cause mortality of children with Down syndrome (DS) in Italy. As expected, they found that the strongest predictor of death was congenital heart disease (CHD). They noted that DS patients with CHD in northern Italy had greater survival than those with CHD in southern Italy. Also, DS patients without CHD in northern Italy had greater survival than

those without CHD in southern Italy. The authors suspect that medical care for the children in the South might not be as good. In the discussion, they state:

The insufficient resources for pediatric care available in the South could explain the low proportion of CHD diagnosed among DS infants there (10.6% as compared with 21.7% in the North).

Is it possible that the overall survival for DS patients (combining patients with and without CHD) in southern Italy could be just as high as in northern Italy? Explain.

References

- Ashton, H. A., M. J. Buxton, et al. (2002). "The Multicentre Aneurysm Screening Study (MASS) into the effect of abdominal aortic aneurysm screening on mortality in men: a randomised controlled trial." *Lancet* **360**(9345): 1531–9.
- Henschke, C. I., D. F. Yankelevitz, et al. (2006). "Survival of patients with stage I lung cancer detected on CT screening." *N Engl J Med* **355**(17): 1763–71.
- Mastroiacovo, P., R. Bertollini, et al. (1992). "Survival of children with Down syndrome in Italy." *Am J Med Genet* **42**(2): 208–12.
- Torres, C. F., S. Rebsamen, et al. (1994). "Surveillance scanning of children with medulloblastoma" *N Engl J Med* **330**(13): 892–5.