# Principles of Epidemiology for Public Health
# EPID 160

## Natural history of disease / population screening

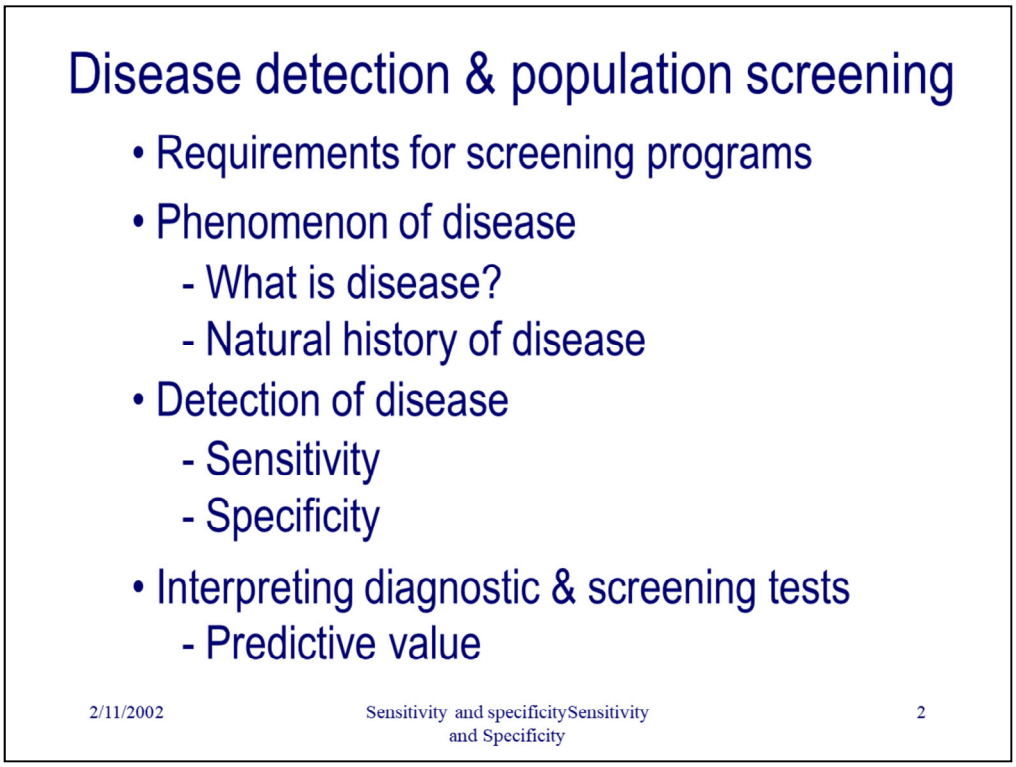
UNC School of Public Health
Department of Epidemiology
Summer 2002

Faculty: Victor J. Schoenbach, PhD
Lorraine K. Alexander, PhD

Developers: Carl M. Shy, PhD, Lorraine K. Alexander, PhD



This lecture is about the fundamental concept of natural history of disease, which will be presented in the context of population screening. The concept of natural history refers to the observation that disease is a process that evolves over time, rather than something that is absent one moment, present the next, and static thereafter. Population screening is a key strategy in what is called secondary prevention of disease. "Primary prevention" refers to the avoidance of disease by avoiding exposure to pathogens. "Secondary prevention" refers to the detection of disease in an early, more treatable stage. "Tertiary prevention" refers to the management of disease in ways that minimizes adverse consequences such as disability.




We will present the topic of disease detection through population screening under four headings – first, what do we mean by population screening and what are requirements for effective screening programs.

Second, what is disease, what do we mean by the natural history of disease, and how does the natural history of disease relate to disease detection and population screening.

Third, how do we evaluate the ability to detect disease, for example, the ability of a screening test or a diagnostic test to detect a condition of interest.

Finally we will consider the interpretation of diagnostic and screening tests from a population perspective, by looking at the concept of predictive value.

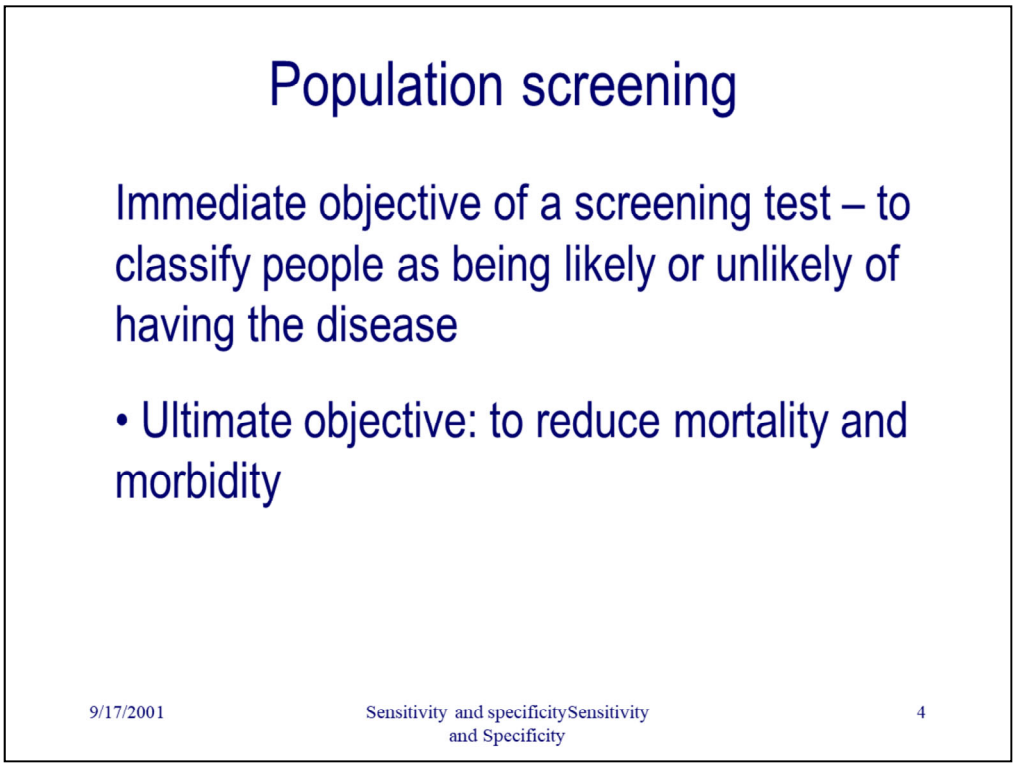
# Population screening

"application of a test to asymptomatic people to detect occult disease or a precursor state"
(Alan Morrison, *Screening in Chronic Disease*, 1985)



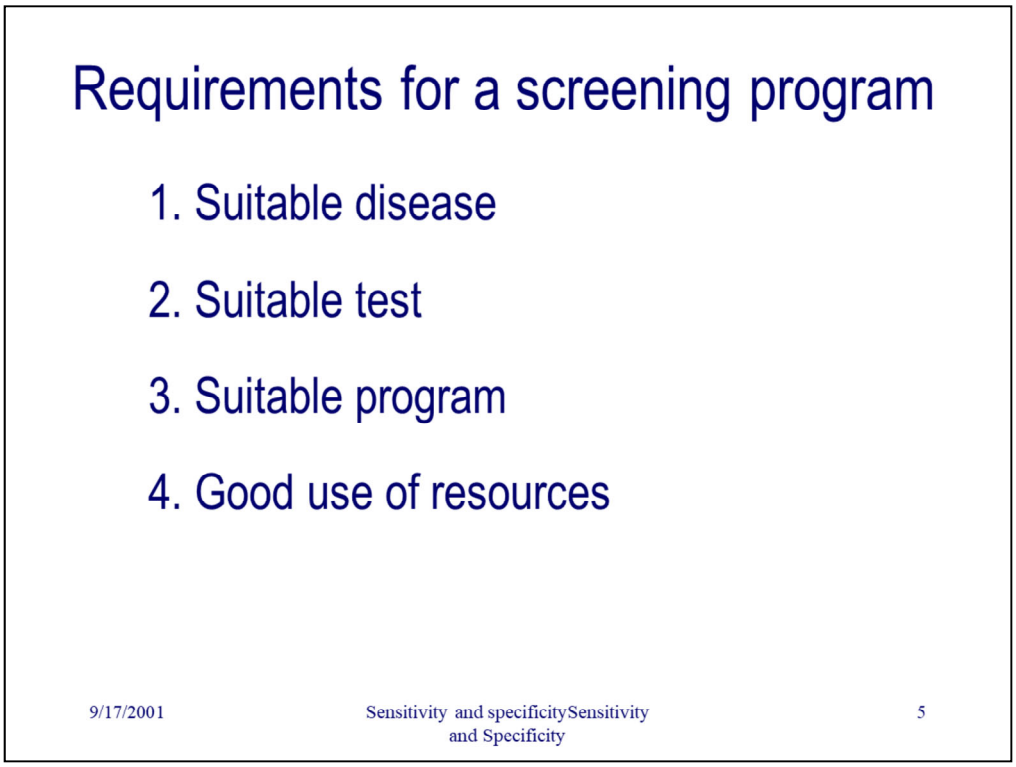
Alan Morrison defines population screening is "the application of a test to asymptomatic people to detect occult disease or a precursor state". Familiar examples of the use of population screening are cancer screening (e.g., Pap smears to detect cervical neoplasia, mammography and physical breast exam to detect breast cancer, PSA to detect early prostate cancer, and fecal occult blood testing to detect colorectal cancer or adenomas), screening programs for hypertension and diabetes to prevent complications, screening of newborns for phenylketonuria (PKU) to prevent mental retardation, screening for carotid arterial obstruction to prevent stroke.

Important points to note about the definition are that it refers to asymptomatic people and to "occult disease or a precursor state". A test to detect disease in a person who has symptoms is considered a diagnostic test. Even though the sametest procedure may be employed, there are important differences between applying a test to asymptomatic people in the population at large and applying the test to symptomatic people who have come to a health care provider. In general, even if the probability that someone with symptoms of a disease actually has that disease, that probability is many times greater than for an asymptomatic person in the general population. A test applied to asymptomatic people must have an extremely low risk of causing harm or serious discomfort. Furthermore, a patient who has come to a provider for evaluation is much more likely to proceed with the diagnostic workup and treatment than someone screened in the general population.




The objective of a diagnostic test is to determine whether someone with signs or symptoms of a disease actually has it. The objective of a screening test is to classify the person receiving it as being likely or being very unlikely to have the disease. If the classification is "likely", then a diagnostic evaluation will be conducted. If the classification is "unlikely", then the process stops there.

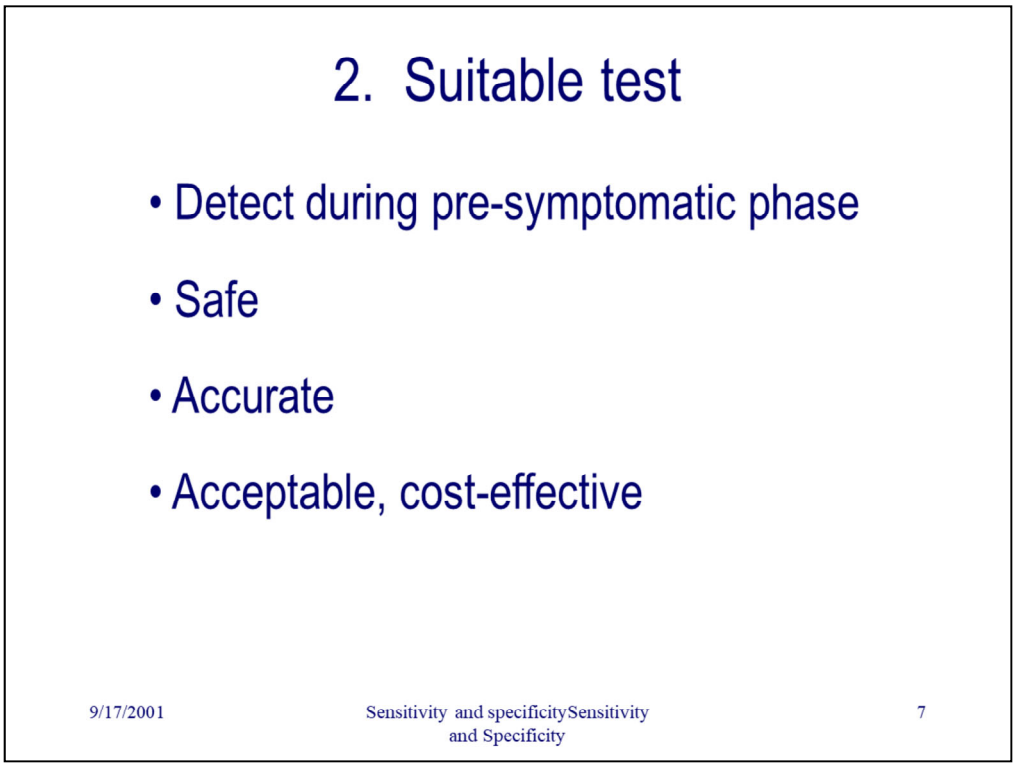
The overall objective, of course, is to reduce morbidity and mortality by detecting a condition in its early stages.




There are four requirements for an effective screening program, namely that the disease be suitable for early detection, that the test be suitable, that the program that provides the test must be suitable, and that the screening program and associated medical evaluation of people with positive tests represent a good use of healthcare resources.


To be a suitable target for a screening program, a disease must have serious consequences if it is untreated. Otherwise why invest the resources to detect it? In addition, the disease must be detectable before symptoms appear. Otherwise, what is the use of screening for it. And importantly, treatment that begins before the disease would be detected without screening must be superior to the outcomes of treatment initiated following clinical diagnosis. After all, if the outcome will not be better, why not just wait for symptoms to appear and bring the patient to a healthcare provider who can initiate treatment then?

As we shall see, it can be quite difficult to demonstratethat the last of these requirements – that outcomes be better as a result of initiating treatment early – is met.




The second requirement for an effective screening program is a suitable screening test. First, of course, the test must be capable of detecting the disease during its pre-symptomatic phase. A test that cannot detect the disease before it becomes symptomatic may be useful for diagnostic purposes, i.e., for establishing that someone suspected of having the disease actually has it. But such a test cannot be useful for population screening, since the whole point of screening is to detect people who would have no reason for being subjected to a diagnostic test. Once symptoms appear, then the person will be motivated (or can be educated to) seek evaluation by a healthcare provider.

In addition, while any medical test or procedure should be safe, there is an especially high premium on safety for a test that will be given to apparently healthy people in the general population. If the disease is rare, as is typically the case, hundreds, even thousands of people will have to be tested for each case identified. Even a 1% rate of serious side effects from the test could cause a great deal of harm in relation to the benefit gained.

Of course, the test must be accurate – it must be capable of doing its job of detecting the condition in its presymptomatic stage and must not misclassify too many people without the condition, which could cause needless worry and inconvenience, burden the healthcare system, and perhaps lead to harm from the diagnostic evaluation.

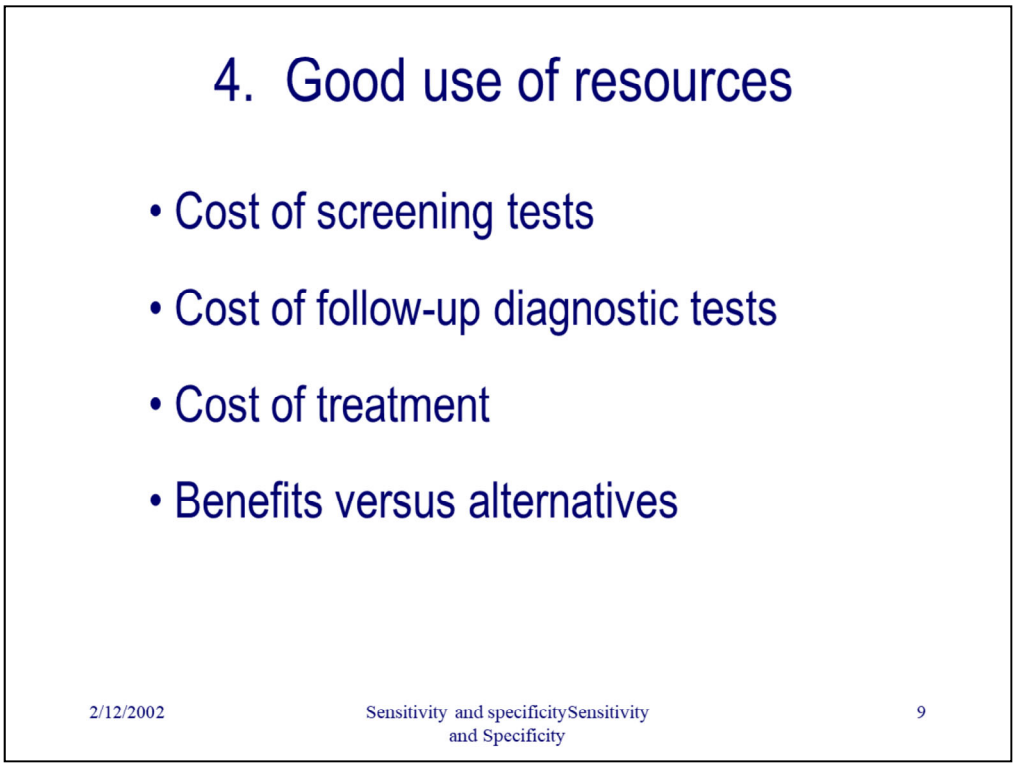
Finally, a screening test must be acceptable to patient and provider and cost-effective in its application.


A suitable disease and a suitable test provide the essential basis for an effective screening program. But to achieve good results, screening tests must be applied in the context of a suitable program. The program must reach the appropriate target population, i.e., people at greatest risk of having the condition being screened for. For example, people who come for screening may greatly overrepresent the "worried well" – people who are at low risk but are very concerned to protect their health. While the screening program may provide them with reassurance, false positive tests will increase worry and may do harm. Moreover, the objective of reducing morbidity and mortality from the disease cannot be met unless groups at greatest risk are screened.

Quality control procedures must be adequate, since population screening is "mass production". A test may be highly accurate in a low-volume, highly controlled laboratory but be much less accurate in a high-volume, commercial one. Inaccuracy in reading test results can easily destroy much of the value of the program.

The program must ensure that people with positive screening tests receive appropriate follow-up. A poorly conducted program may simply label people as at risk (e.g., hypertensive) without producing any benefit unless follow-up is adequate.

And as in any large-scale program, lack of efficiency will drive costs up, straining resources.




Even if all of the foregoing requirements are met, a screening program may still not represent a good use of resources. A full evaluation requires taking account of the cost of administering and interpreting the screening tests themselves, but also the costs of the follow-up diagnostic work-ups for people who receive positive tests – including those with falsely-positve tests. Treatment costs must also be considered, since screening may detect cases that would not otherwise come to the attention of the healthcare system (e.g., because the person might have died of something else). As always, costs and benefits must be weighed against alternative strategies for controlling the disease of concern as well as against other health objectives.

## Phenomenon of disease: what is disease?

Difficult to define, e.g.:

"a type of internal state which is either an impairment of normal functional ability–that is, a reduction of one or more functional abilities below typical efficiency–or a limitation on functional ability caused by environmental agents"

(C. Boorse, *What is disease*? In: Humber M, Almeder RF, eds. *Biomedical ethics reviews*. Humana Press, Totowo NJ, 1997, 7-8 (quoted in Temple *et al*., 2001)



Before proceeding further with our inquiry into population screening, it will be valuable to develop a more sophisticated understanding of the phenomenon of disease. Although most of us speak as if we know what disease is, the term is notoriously difficult to define. For example, the above definition refers to "normal" and "typical efficiency", which begs the question since we then need a definition for "normal".

Conditions that rapidly lead to death rarely arouse debate about whether or not they qualify as "disease". But many non-fatal conditions are less easily classified. Two current examples are attention deficit hyperactivity disorder (ADHD) and chronic fatigue syndrome. Both of these syndromes (collections of symptoms) are regarded by many as resulting from a "disease", yet in the absence of regularly demonstrable biological dysfunction or abnormality how does one differentiate these syndromes from other variation in "normal" experience?

## Phenomenon of disease: what is disease?

Difficult to define, e.g.:

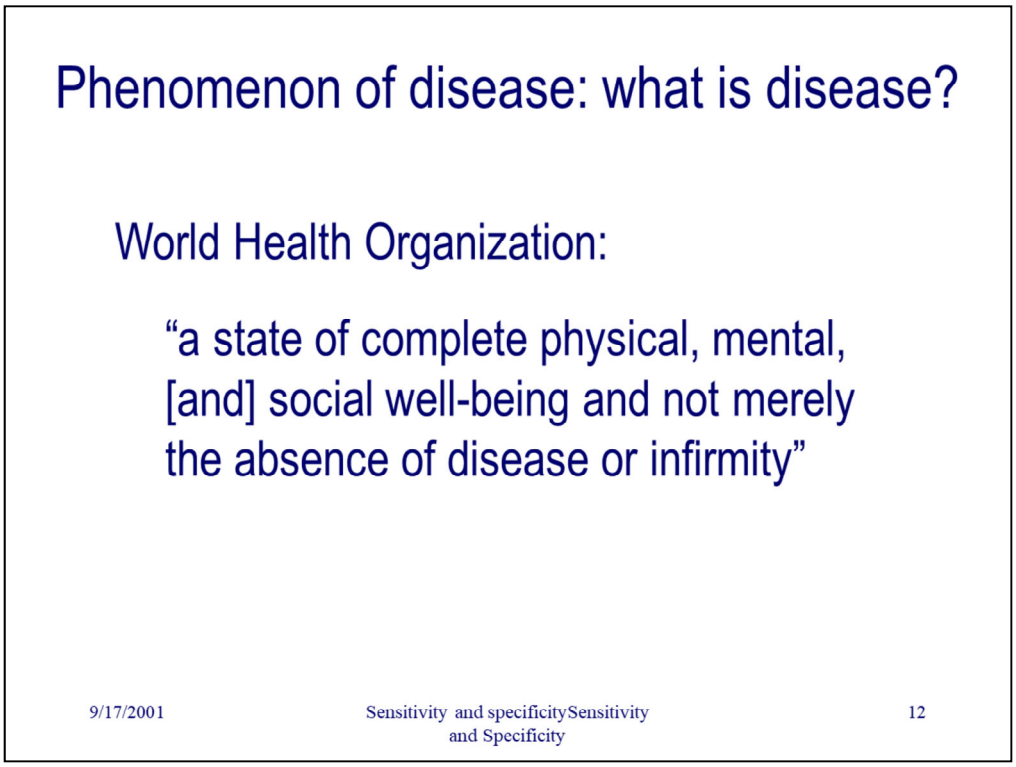
"a *state* that places individuals at *increased risk of adverse consequences*"

(Temple LKF *et al*., Defining disease in the genomics era. *Science* 3 Aug 2001;293:807-808)



An attempt to avoid the word "normal" is also problematic. There are many states that place individuals "at increased risk of adverse consequences" that we would probably feel uncomfortable classifying as diseases. What about, for example, riding in a car without a seatbelt? Or being a police officer. Or working in a convenience store? Or living in a high-crime neighborhood? Or perhaps even being a graduate student?

Perhaps these are not "states". However, intoxication is a state – is it meaningful to call it a "disease"?



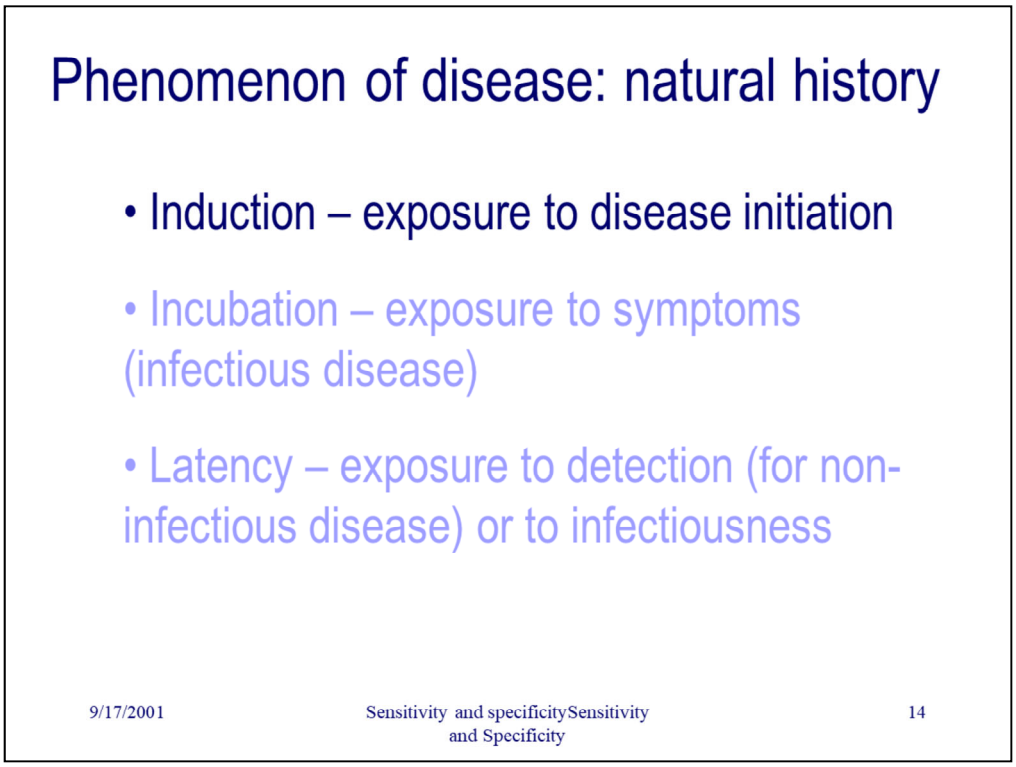

The World Health Organization's definition of health provides a lofty and inspiring vision. But this definition, too, is difficult to operationalize.

So we are left in the somewhat unsatisfactory situation – or at least unsatisfying situation – of not having an adequate definition of disease, which is, after all, the object of much of epidemiology.


Nevertheless, most of us think that we know disease when we see it, so we may as well move on to the key concept of natural history.

The concept of natural history arises from the realization that disease, whatever it is, is a process that unfolds over time. Natural history refers to the evolution of the disease process over time, from the earliest pathological developments to their resolution. In principle, the "natural" in "natural history" refers to the evolution of a disease in the absence of intervention or treatment, though the meaning is not always restricted in this way.

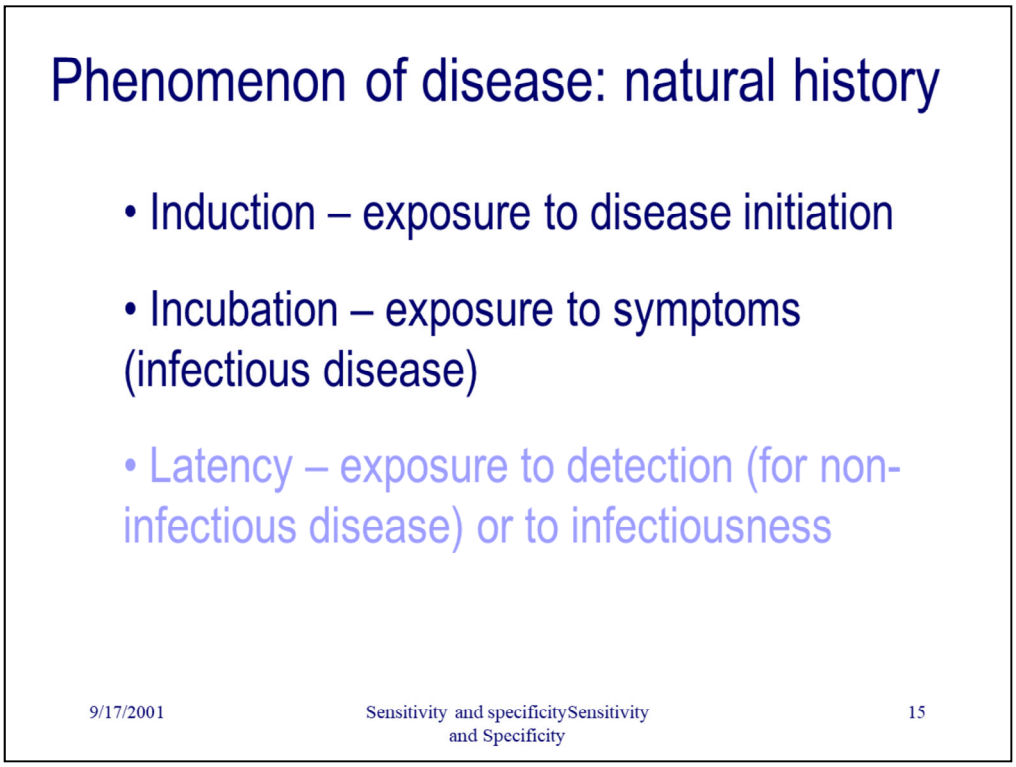
By extension, natural history can be extended to include developments and experiences that are not "pathological" as such but which can be regarded as forerunners of pathological change. For example, cigarette smoking could be viewed as part of the natural history of lung cancer as it most often occurs. Early onset of sexual intercourse with numerous partners could be viewed as part of the natural history of cervical cancer.



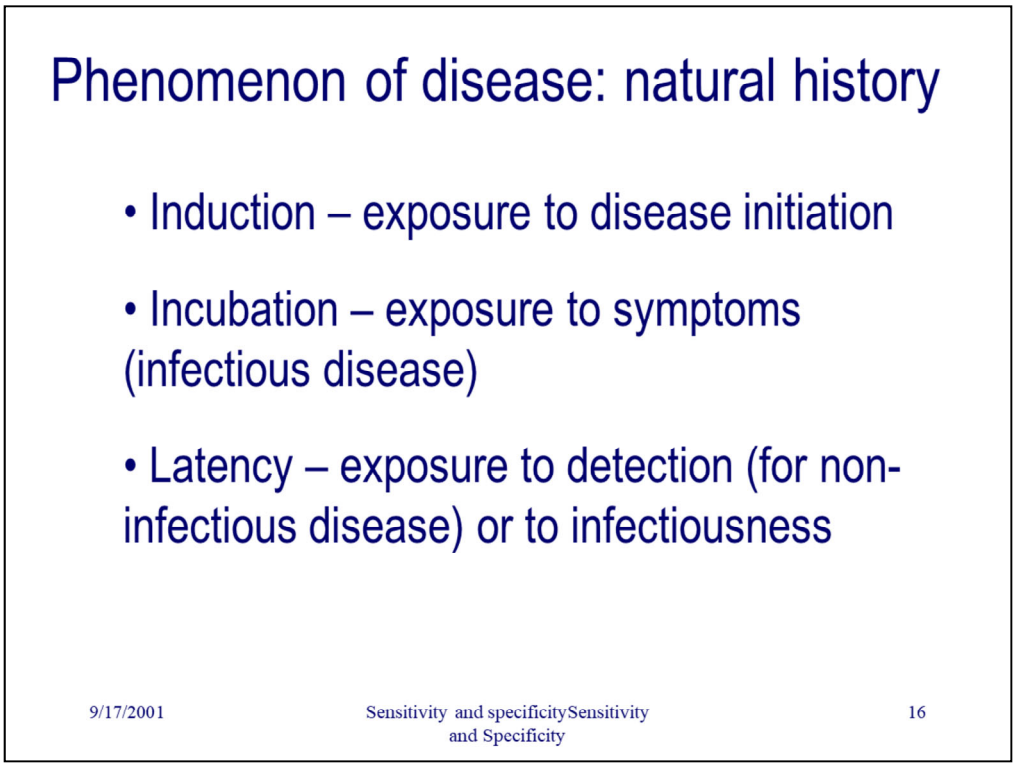

Several important landmarks in natural history are the initiation of the disease itself (however we are defining it), the onset of symptoms, the time period when a communicable disease can be transmitted, and the point where a disease can be and/or is usually detected.

Epidemiology is still working on standardizing its terminology. The above terms are used by a major epidemiology textbook (Rothman and Greenland, *Modern epidemiology*) to represent these four stages:

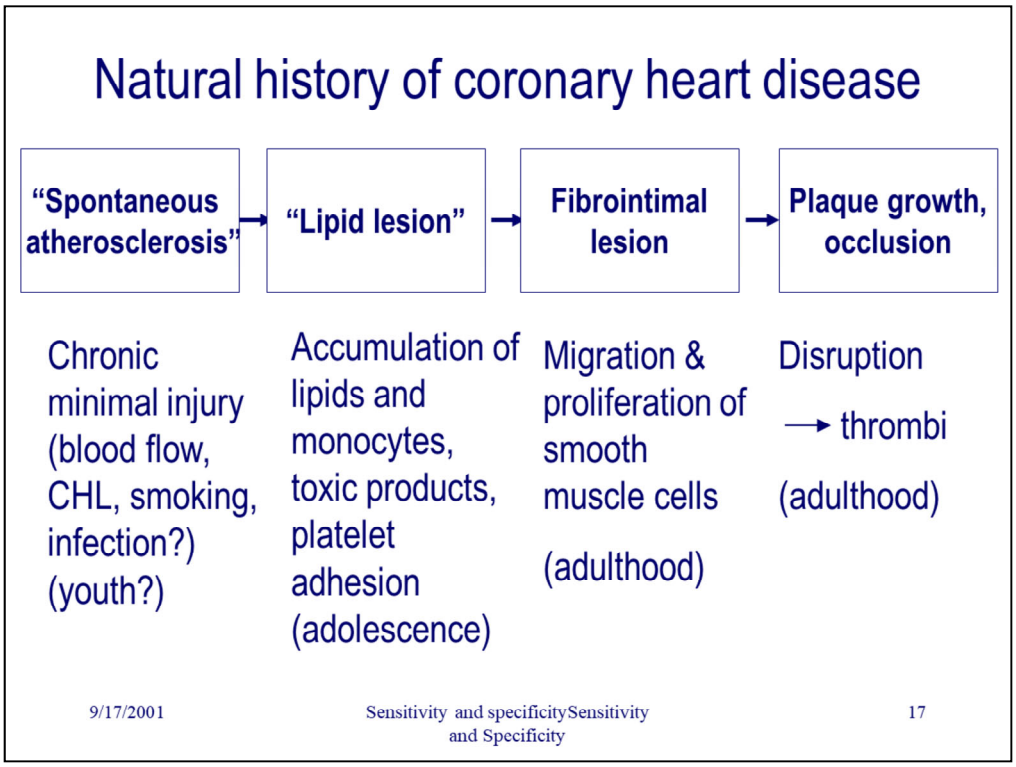
Rothman applies the term induction period to the period of time between exposure to a causal agent and the initiation of the disease. Since many diseases, particularly those whose etiology involves a combination of factors, Rothman considers each causal agent as having its own induction period. The beginning of actual disease, even if it can be defined, is often unobservable. Nevertheless, the concept of the induction period is a useful one.



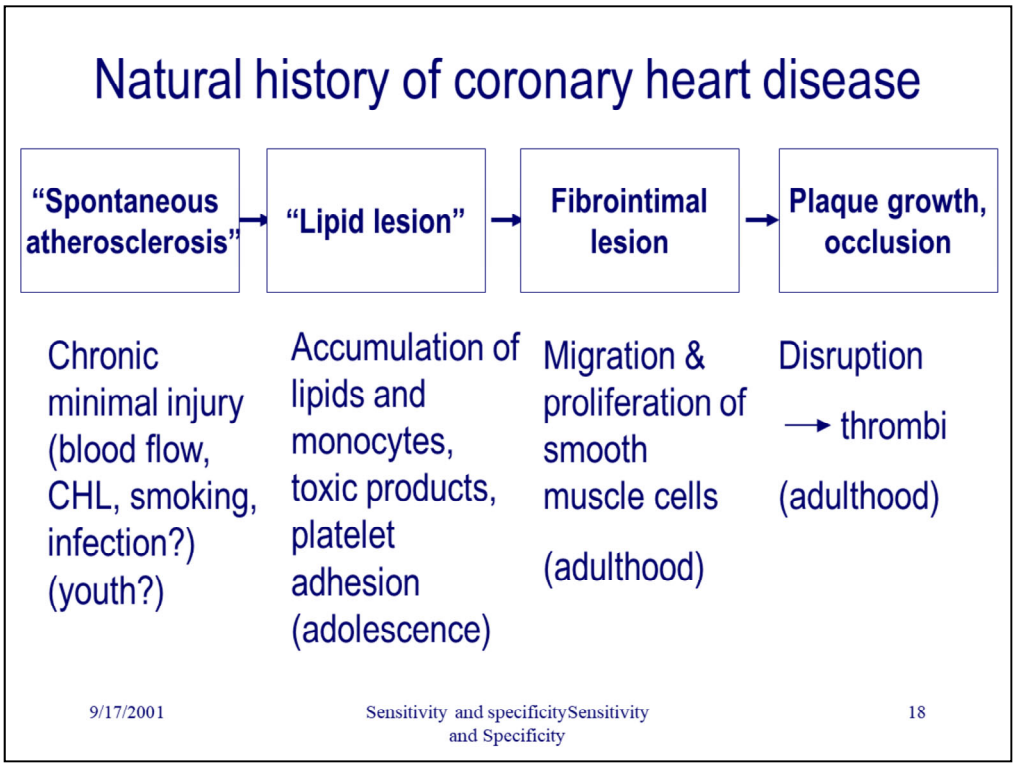

The incubation period, a term that originated to describe infectious diseases but is sometimes applied more broadly, refers to the period of time during which an infectious organism multiplies in the body to the point where symptoms result, either from the actions of the microorganism or from the immune response to it.



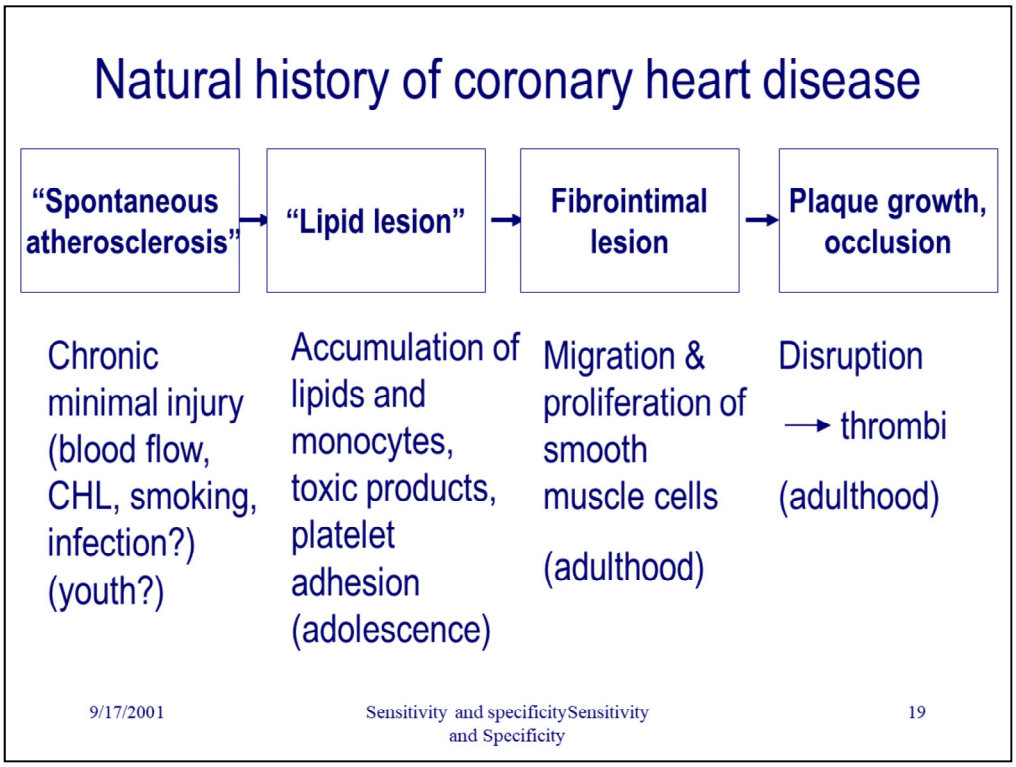

The latency period, or simply latency, refers to the period between the initiation of disease and its detection. Detection usually follows the appearance of symptoms. However, if screening for the disease is common, latency may be regarded as ending at the time of usual detection through screening. Note that for infectious diseases, latency refers to the period before which the disease can be transmitted, a very important event in the control of communicable diseases.




We illustrate the concept of natural history with the example of coronary heart disease. Most heart disease in the industrialized world arises through a lengthy process that spans decades. The process, referred to as "atherosclerosis", begins with chronic, minimal injuries to the arterial endothelium – the lining of the inner wall of the blood vessels that carry blood from the heart. Atherosclerosis can occur in many locations in the vascular system. Coronary atherosclerosis involves the coronary arteries, that supply the heart muscle with oxygen and nutrients. Carotid atherosclerosis involves the carotid arteries that supply blood to the brain.



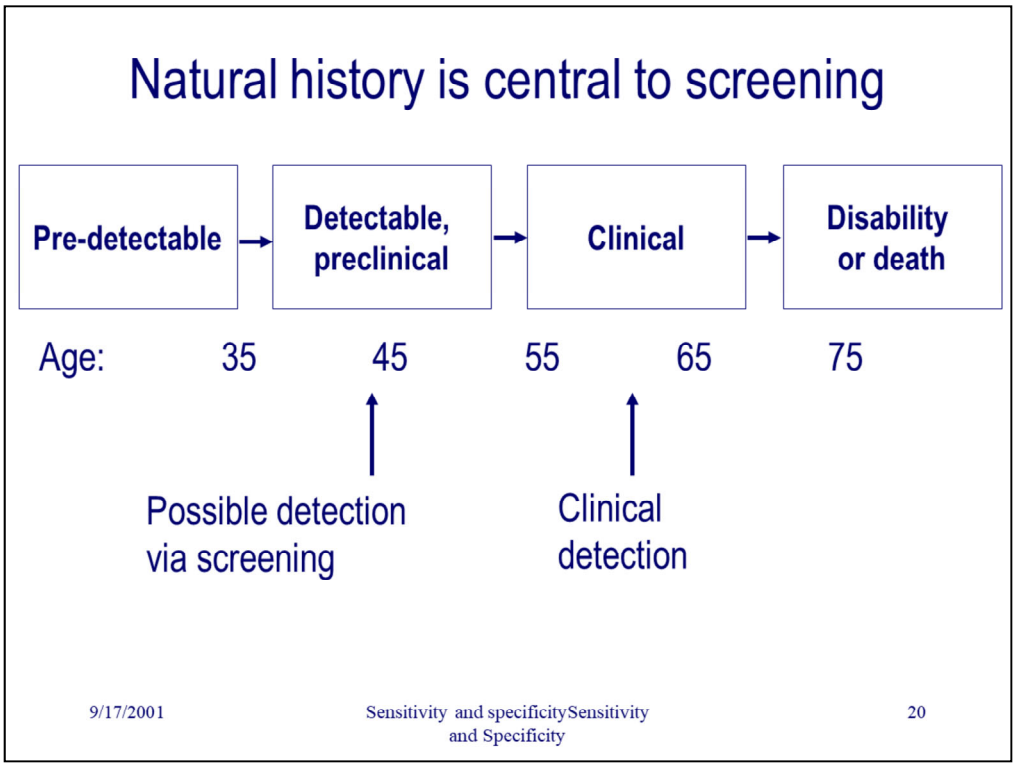

The initial injury to the arterial lining results naturally from disturbances in the flow of the blood at locations where the arteries branch. The "wear and tear" from blood flow can be exacerbated by "insults" such as elevated serum cholesterol (CHL), tobacco smoke constituents, and possibly infections. These Type I injuries probably begin in youth and are not harmful in themselves. However, places where the smooth endothelium has been compromised in this way accumulate blood fats ("lipids") and macrophages (blood cells that make up an important component of the body's immune system). The macrophages release toxic products, which are helpful when used to kill invading microorganisms or to destroy infected cells. In the arterial endothelium, however, these toxic products cause a worsening of the damage by attracting other blood cells called platelets (involved in forming blood clots), which then adhere to the endothelium producing. The macrophages, platelets, and the endothelium itself secrete growth factors that attract smooth muscle cells and cause them to proliferate and to take in lipids from the blood, producing a "fibrointimal lesion" or "lipid lesion" ("lesion" means injury).



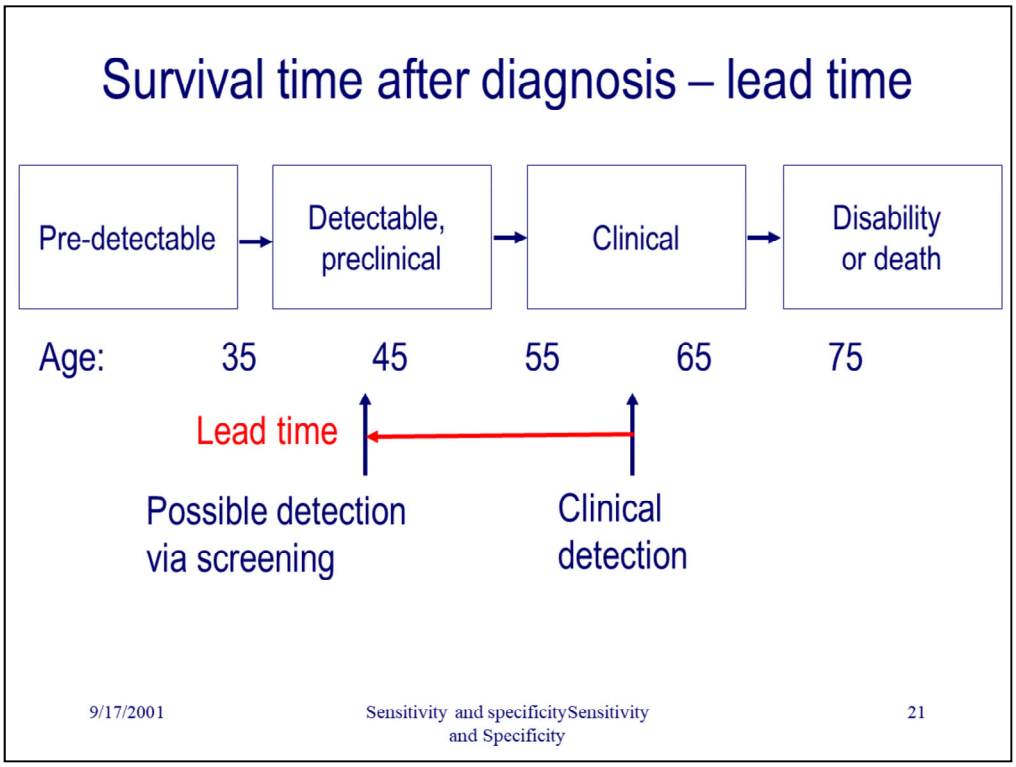

The lipid lesions can grow to partially obstruct the flow of blood. Worse, disruption of a lipid lesion ("Type III damage") leads to thrombus formation - clotting. The clot can block the remaining portion of the artery, shutting off the flow of blood through it. Or, a portion of the thrombus may become dislodged and travel through the blood vessels to become lodged in a smaller vessel, blocking it off. The complete obstruction of blood flow produces a myocardial infarction (MI, or "heart attack") or an occlusive stroke (blockage of blood to an area of the brain).

These events are generally accompanied by symptoms, though it is possible to have a "silent" MI or "silent stroke". If the blockage is not cleared quickly, then heart or brain tissue is damaged and disability or even death ensues.

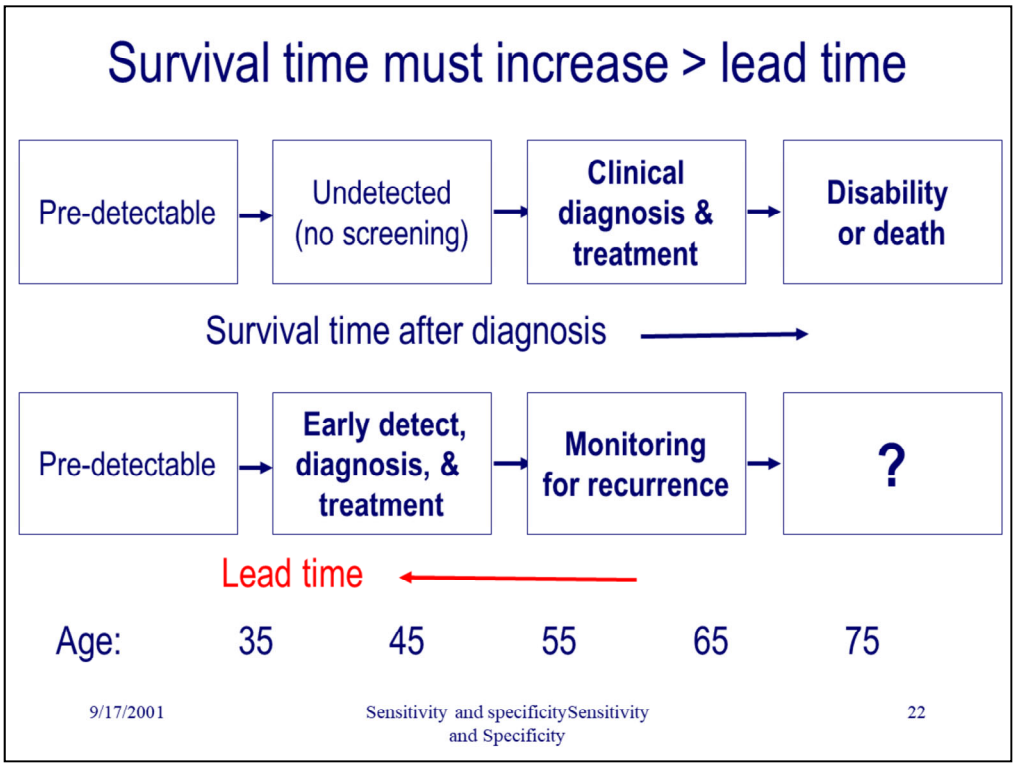
Partial occlusion of a coronary artery can lead to a syndrome called "angina pectoris" – pain in the chest brought on by a relative insufficiency of oxygen to the heart muscle. In "typical" angina, the pain comes on during strenuous exertion, where the heart muscle needs more oxygen than the compromised blood vessel can provide. In "unstable angina", the pain can come without an obvious cause, and the danger of a complete blockage occurring is considerably higher.



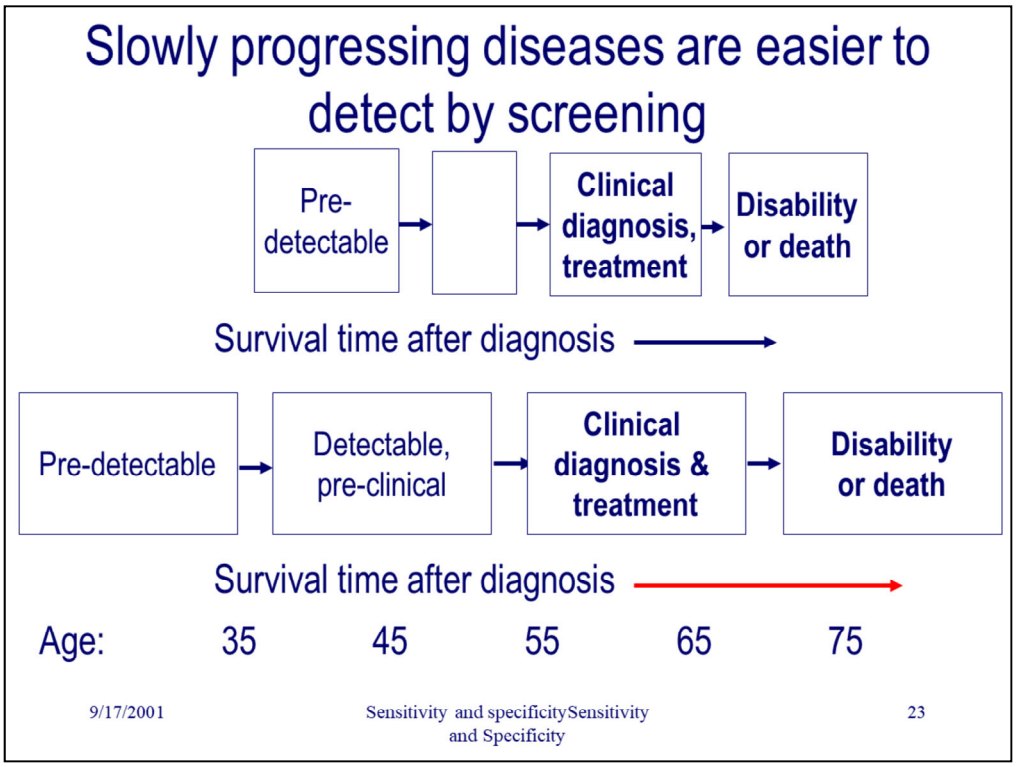

The concept of natural history of disease is central to population screening. To be a good candidate for screening, a disease needs to have a lengthy "detectable, preclinical" stage that can serve as the target for the screening test. Before the disease is detectable, screening is futile. After the disease becomes symptomatic, screening is superfluous. So the opportunity to reduce morbidity and mortality through screening depends upon their being a detectable, preclinical stage that lasts long enough to be detected by screening on some reasonable schedule.




A successful screening program will advance the time of detection from the point at which symptoms appear (clinical detection) to some earlier point in the natural history of the condition. The time by which detection is advanced is sometimes called the lead time. The lead time provides the opportunity for treatment to gain the upper hand over the disease. The lead time also provides the opportunity for screening to appear to be effective when in fact it produces no benefit. This phenomenon, called "lead time bias", is illustrated on the following slide.

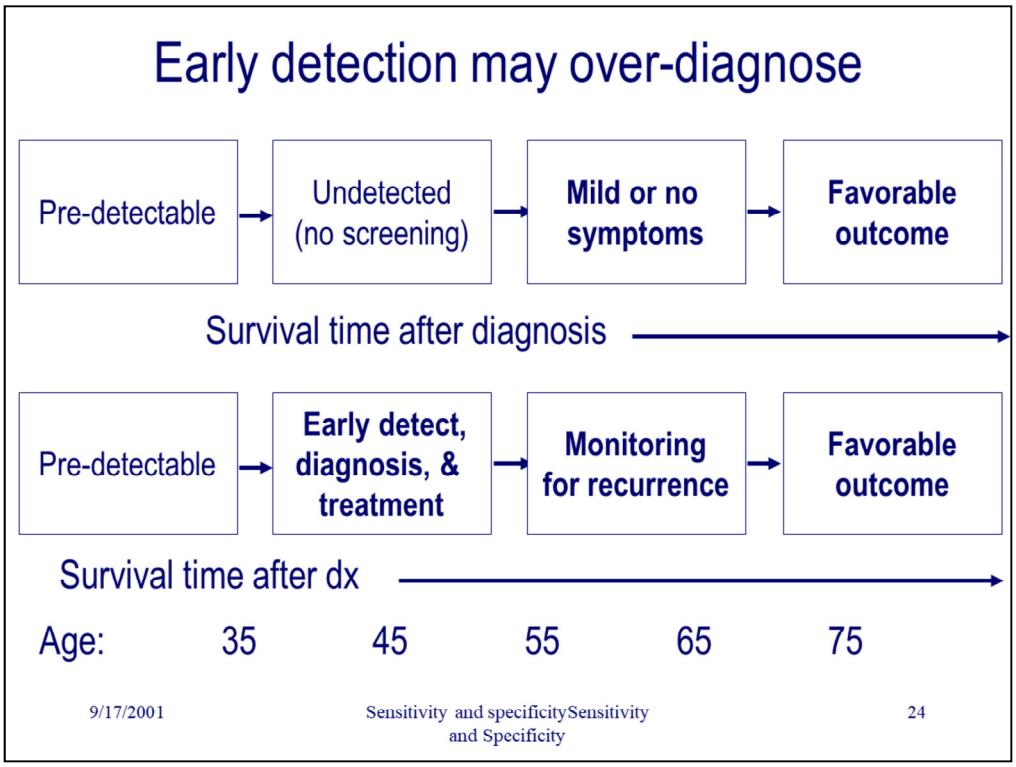



Lead time bias occurs in a comparison of survival time after diagnosis that fails to take account of the lead time. For example, suppose that the usual age at diagnosis is 60 years and cases live on average 10 years following diagnosis, so that the average age at death is 70 years. Suppose that due to the screening program, the disease is now being detected 15 years earlier, at age 45 years. Even if people now survive for 25 years following diagnosis, they are will die at age 70. So a screening program with a treatment that does not alter the natural history of the disease compared to what would happen with treatment following clinical diagnosis will nevertheless appear to lengthen survival time following diagnosis. Such a program has in fact merely lengthened the time during which the person is considered a patient. So comparisons of length of survival following diagnosis need to adjust for lead time resulting from earlier diagnosis in order to avoid being affected by lead time bias.



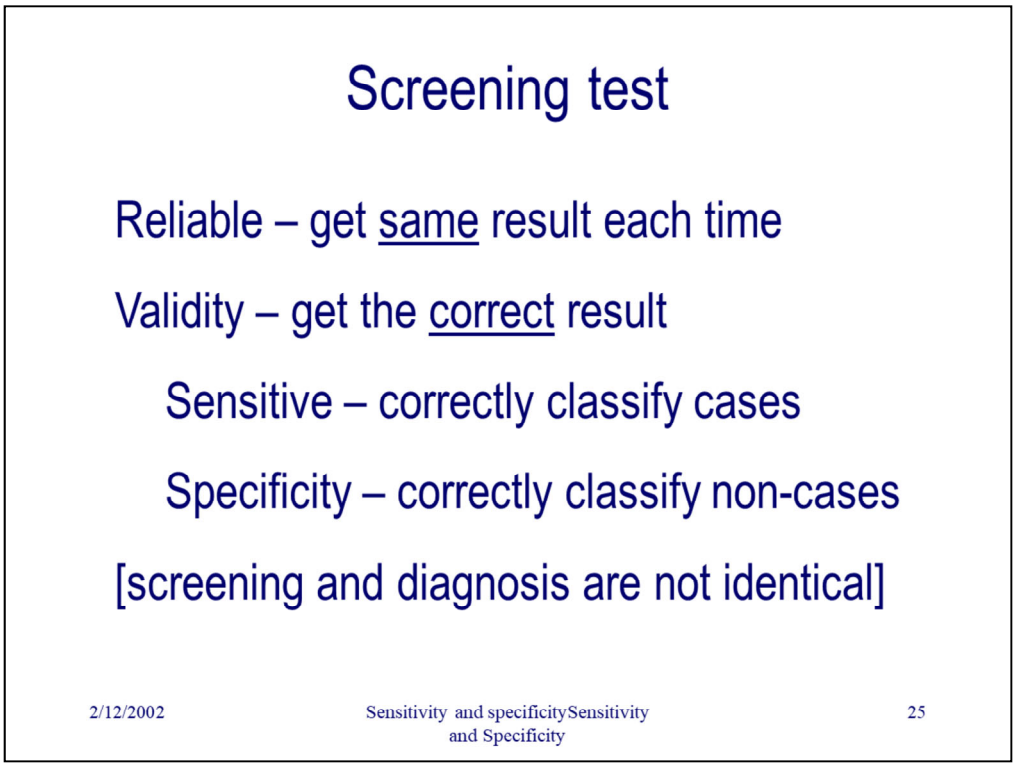

Another type of bias that can complicate the interpretation of survival data from a screening program is "length bias". For example, one apparently logical way of assessing the effectiveness of screening is to compare survival time following diagnosis for cases detected during regular screening with cases that are detected during the interval between screens ("interval cases"). The idea is that the cases detected by screening have been earlier, allowing the treatment to be more efficacious. However, if there is some heterogeneity in the length of the natural history of the disease – i.e., if some cases progress more rapidly than other cases – then the comparison of outcomes between screen-detected cases and interval cases can be misleading. The reason for the problem is that interval cases are more likely to be rapidly growing, aggressive disease which may be more lethal regardless of the circumstances. Had this aggressive disease been detected by screening, it woud still have a poor prognosis. But rapidly progressing disease is less likely to be detected by screening, because its detectable, preclinical stage is shorter.

If a group of horses are running toward a finish line, and several of my colleagues and I jump on some of them as they run by, you will observe that the ones we are riding will take longer to reach the finish line than the rider-less horses. One interpretation is that having a rider slows the horse down. Another possibility, though, is that we were more likely to jump onto slower horses. In this way, screening preferentially detects more slowly progressing disease, and such disease may be easier to treat even if detected when it becomes symptomatic.




Heterogeneity of disease can lead to yet another problem in interpreting survival data from a screening program. Knowledge about the prognosis of various conditions has been gained from observing clinical cases – people who have developed symptoms and then been treated. With screening, the possibility exists of detecting conditions that might never actually cause disease but happen to look like an early form of a condition that does cause disease.

For example, prostate cancer is a very common condition among older men. About 4% of 50-year-old men have asymptomatic prostate cancer, a figure that rises to about 10% at age 60 and about 20% at age 70. These prevalence estimates are based on autopsy studies, since most of the prostate cancers will never produce clinical disease during the man's lifetime. However, Prostate Specific Antigen (PSA) screening will detect a certain number of these prostate cancers, whose prognosis was excellent without treatment and will presumably be excellent with treatment – (except for the side effects of treatment. Thus, treatment of PSA-detected prostate cancer will appear to be more efficacious than treatment of clinical prostate cancer at least in part because many of the PSA-detected cases would never become clinical disease.




Now let's turn our attention to the screening test, which of course must be accurate. There are two major dimensions to accuracy. One dimension is reliability. A reliable test will give us the same result each time that we use it, assuming that the conditions are the same. However, a test can be reliable (consistent) but still give an incorrect result. So the other key dimension is validity. A test is valid if it gives the correct result.

One of the reasons we differentiate between reliability and validity is that in practice we may not know what the correct result is. In such a case, the best we may be able to do is to assess whether the test at least gives consistent results. If the results are inconsistent, for example, when the test is repeated after a short interval, then it must be giving the wrong result fairly often.

Although we could assess the validity of a test by how often it correctly classifies a person as having or not having a condition, in epidemiology we examine each of these classifications – of cases and of noncases – separately. For one, the consequences of misclassification are often different for the two situations. Also, the cases are usually so much rarer that a test that classifies everyone as a "non-case" would nevertheless appear to be "highly accurate".

The ability to correctly classify cases – people with the condition – is called sensitivity. The ability to correctly classify noncases – people without the condition – is called specificity. Note that sensitivity and specificity are both probabilities of correct classification, i.e., they are both desirable properties for a test.

# Reliability

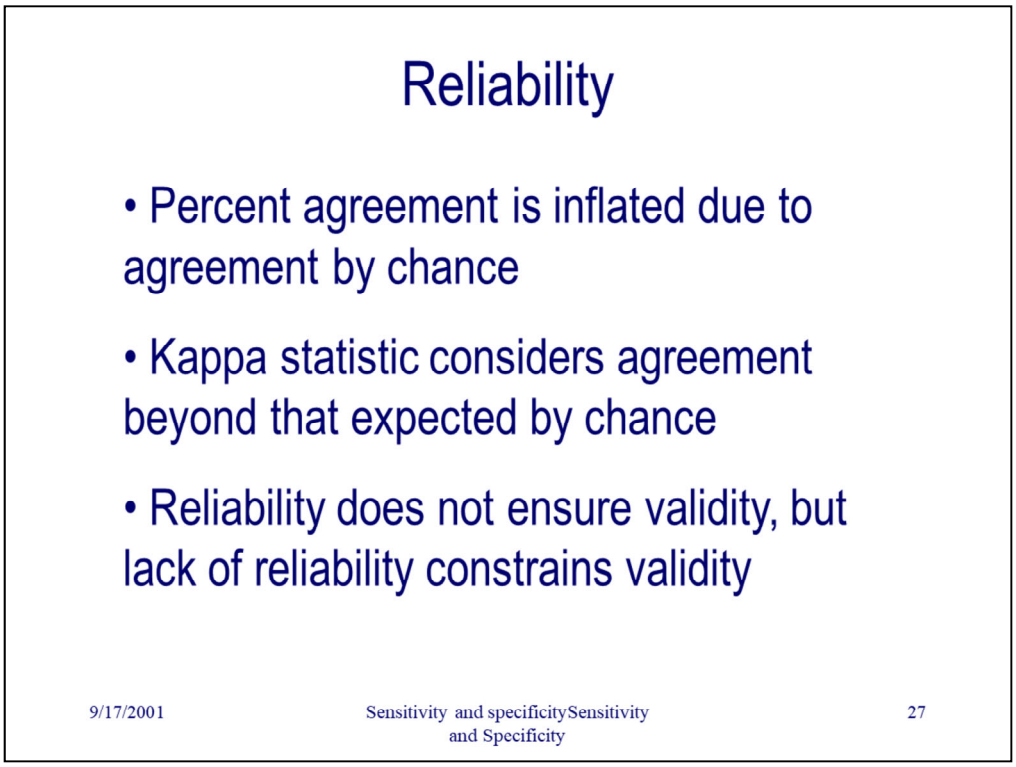
Repeatability – get same result

- Each time
- From each instrument
- From each rater

If don't know correct result, then can only examine reliability.



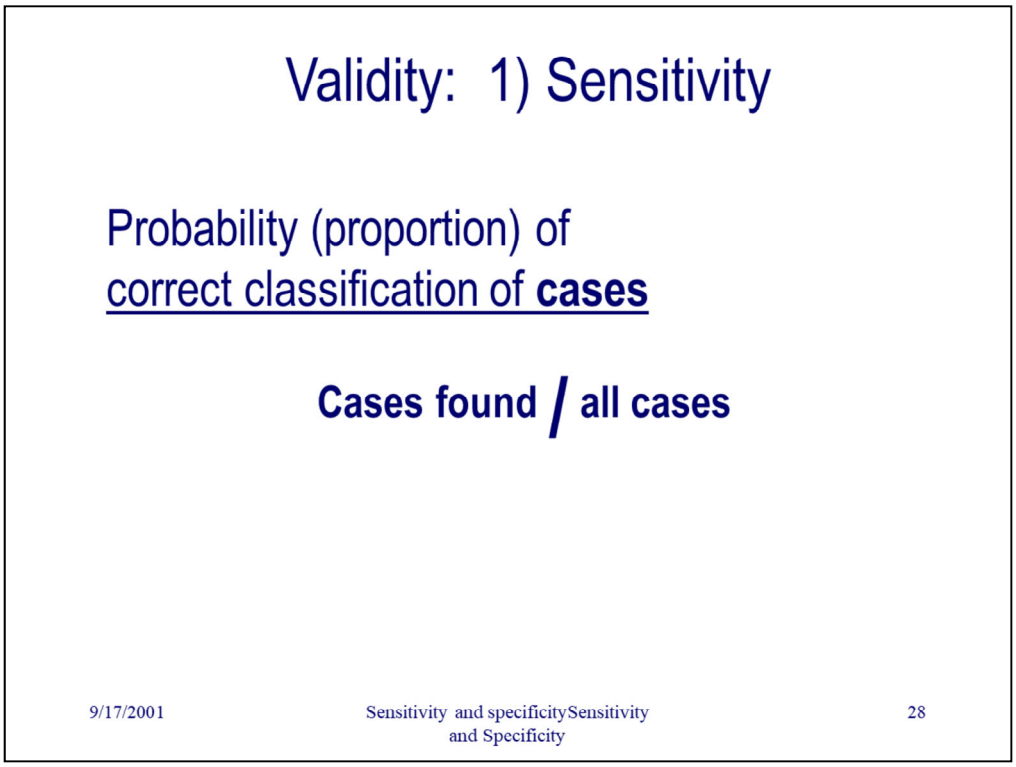
Reliability can be thought of as "repeatability" – a reliable test gives us the same result each time. We can examine reliability over time ("test-retest reliability"), reliability across instruments we believe to be equivalent (e.g., two forms of the same psychological scale), or reliability across examiners ("inter-rater reliability"). If we do not have an authoritative measure (a "gold standard") that can tell us what the correct result is, then we can still assess reliability, but not validity.



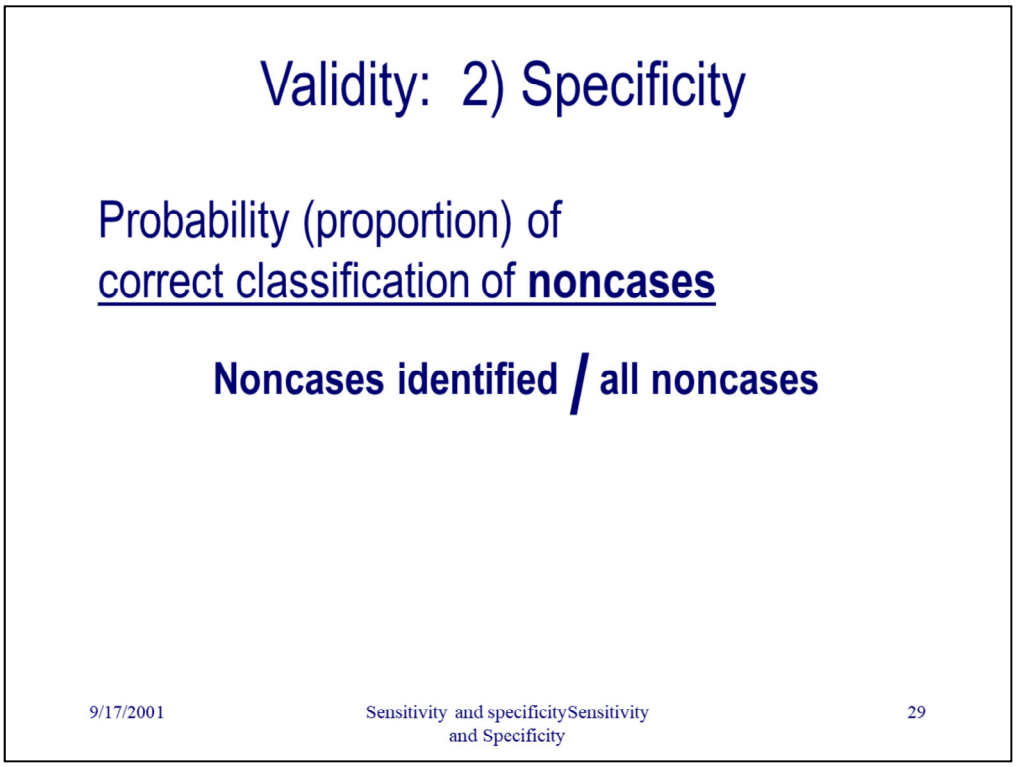

When we assess agreement between two instruments or raters, we should keep in mind that some agreement can be expected by chance alone. Just as one can get the correct answer on a multiple choice test by flipping a coin, so two independent examiners can come up with the same classification by chance. For example, suppose you decide to answer a true-false question by flipping a coin. As a budding epidemiologist, you may want to check the reliability of your coin. If you flip the coin a second time, you will find that on average it agrees with itself, on average, 50% of the time. So if two raters agree 50% of the time, we should not necessarily be impressed.

The kappa statistic is a measure of agreement that adjusts for the amount of agreement expected by chance, so it is often preferred to simple percent agreement (see the *Evolving Text* for more on measuring agreement and kappa).

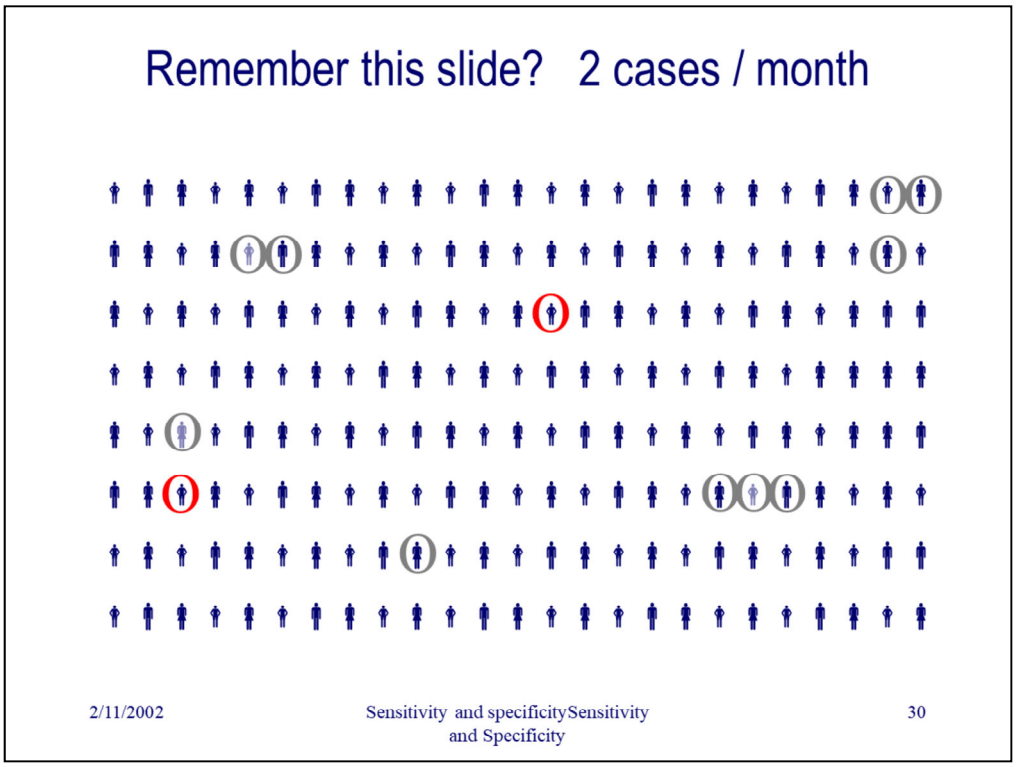
A test that is unreliable cannot be valid. The converse is not true, however. Even if test-retest agreement is very high (for example, 100%), the test could simply be consistently incorrect. However, a test that is unreliable cannot be valid.



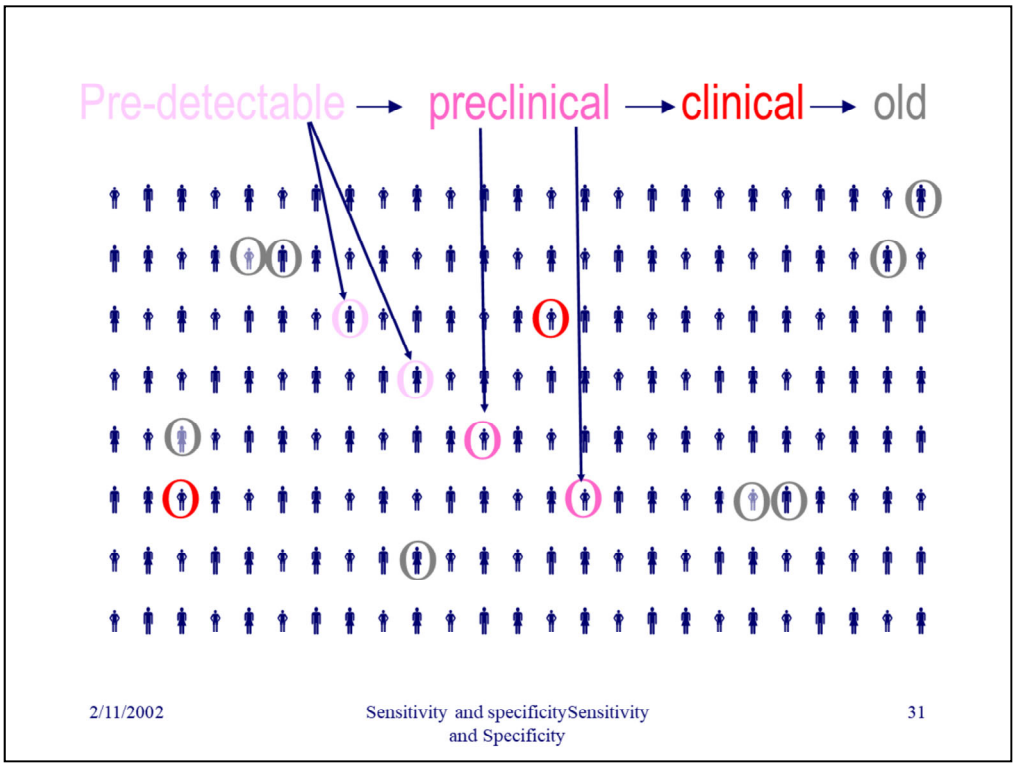

Sensitivity is defined as the probability that the test will correctly classify a case. Given a population of people with a condition, sensitivity gives the proportion whom we expect to be correctly identified as cases.



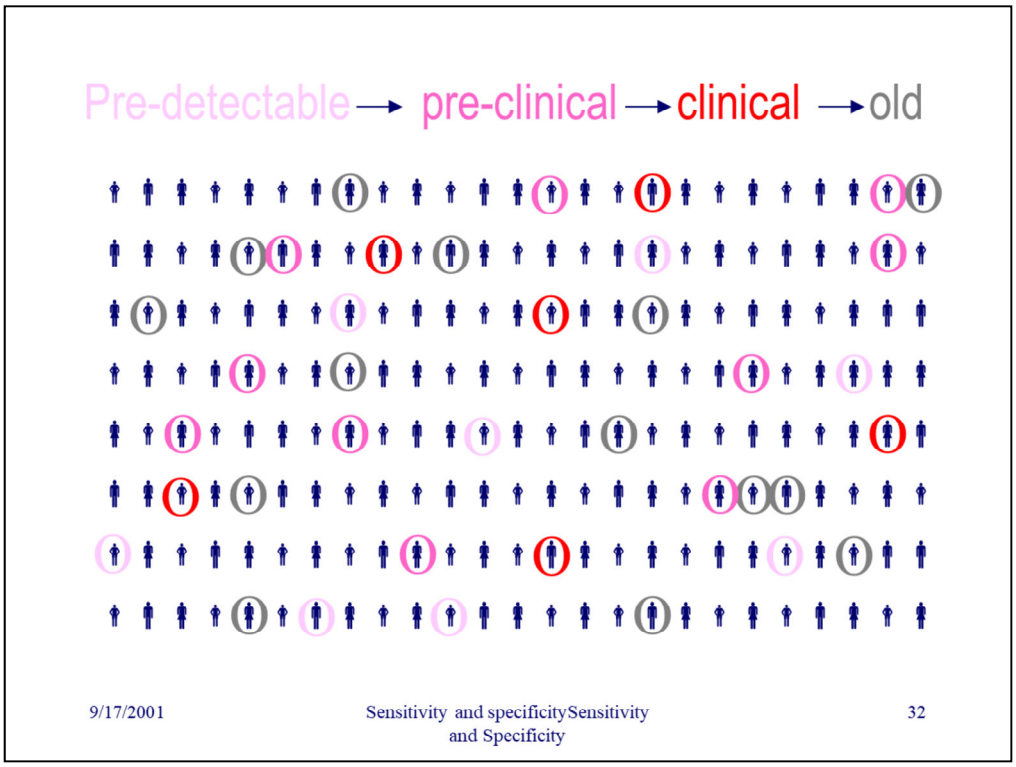

Specificity is defined as the probability that the test will correctly classify a noncase. Given a population of people without the condition, specificity gives the proportion whom we expect to be correctly identified as noncases.




Do you remember this slide? We saw it in the lecture on incidence and prevalence. The slide showed a population of 200 people in which 10 people had a condition and two more people (circled in red on the slide) just developed it during the month.




In our discussion of incidence and prevalence, we necessarily focused on cases that were detected. However, we are now concerned about pre-clinical disease as well, so here is the diagram with four additional case – two circled in very light pink to represent pre-detectable cases and two others circled in pink to represent preclinical, but detectable cases. It is these detectable preclinical cases that are the object of screening




Here is another hypothetical population displayed in a diagram using the same color scheme: light pink circles for pre-detectable cases, pink circles for detectable, pre-clinical cases, red circles for new clinical cases, and gray circles for older cases. Let's count how many of each type there are.

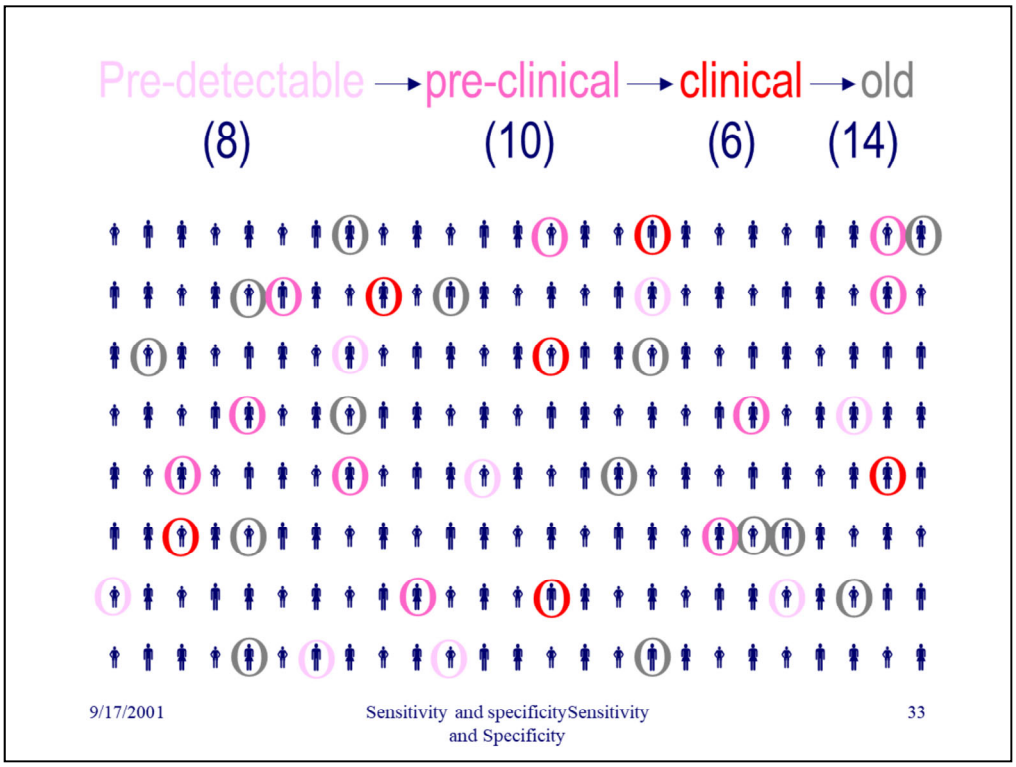
Pre-detectable:

Pre-clinical:

Clinical:

Old cases:

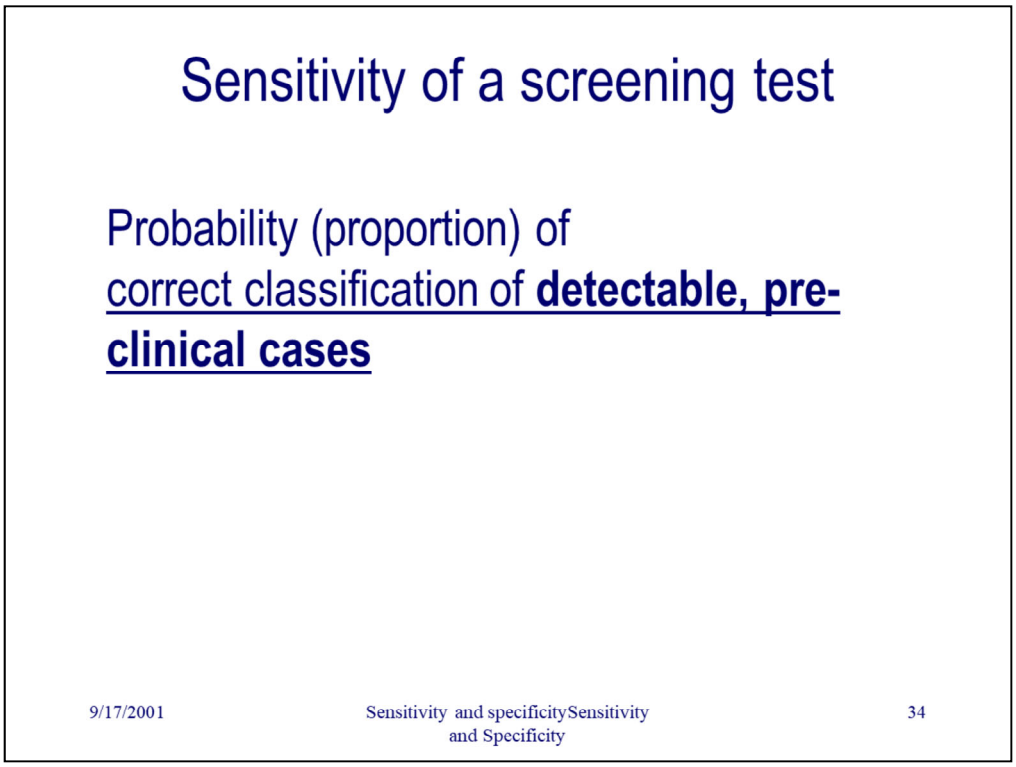
The pre-detectable cases are a bit difficult to see, as we would expect for cases that are pre-detectable!



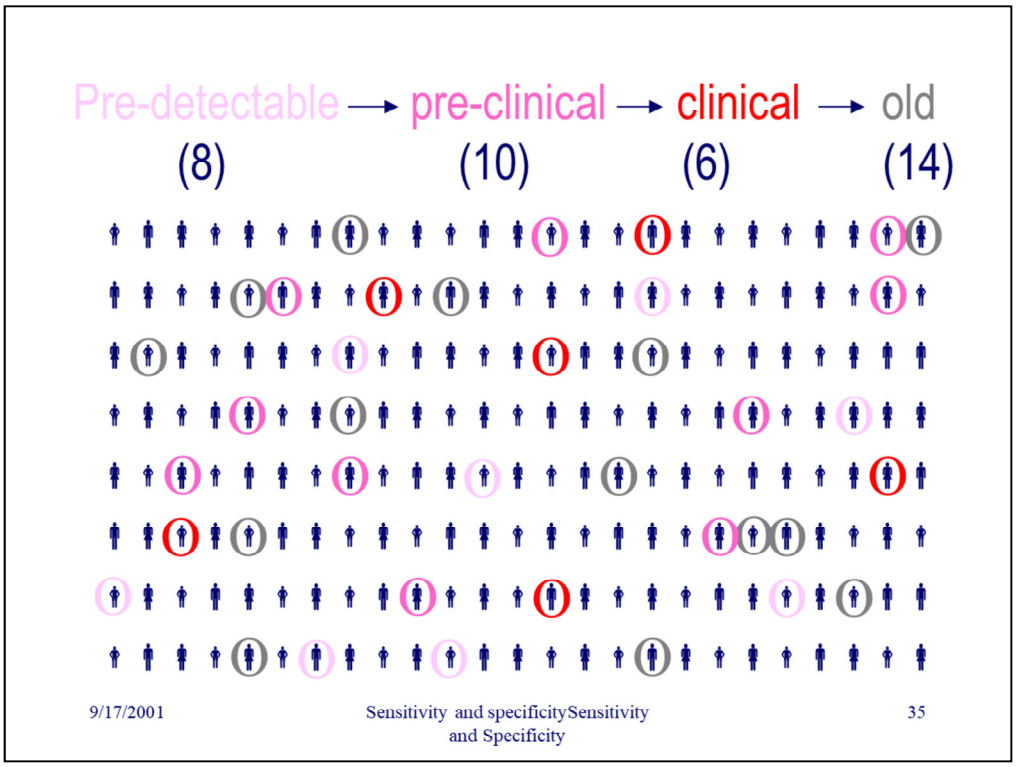

Here is my count – 8 pre-detectable, 10 pre-clinical, 6 clinical, and 14 old cases.

So, before we move on to estimate sensitivity and specificity, what is the prevalence of the condition in this population?

That depends, of course, on what we mean by "the condition". The prevalence of clinical disease is (6 + 14) / 200 = 10%. The prevalence of clinical and detectable pre-clinical diseasetogether is (10 + 6 + 14) / 200 = 15%. The prevalence of detectable, pre-clinical disease by itself is 10 / 180 = 5.5%. This prevalence is the most relevant one for planning or evaluating a screening program, since it is the detectable, pre-clinical cases that are the target of the screening test, and there would be no point in screening people whom we already knew to have the disease.

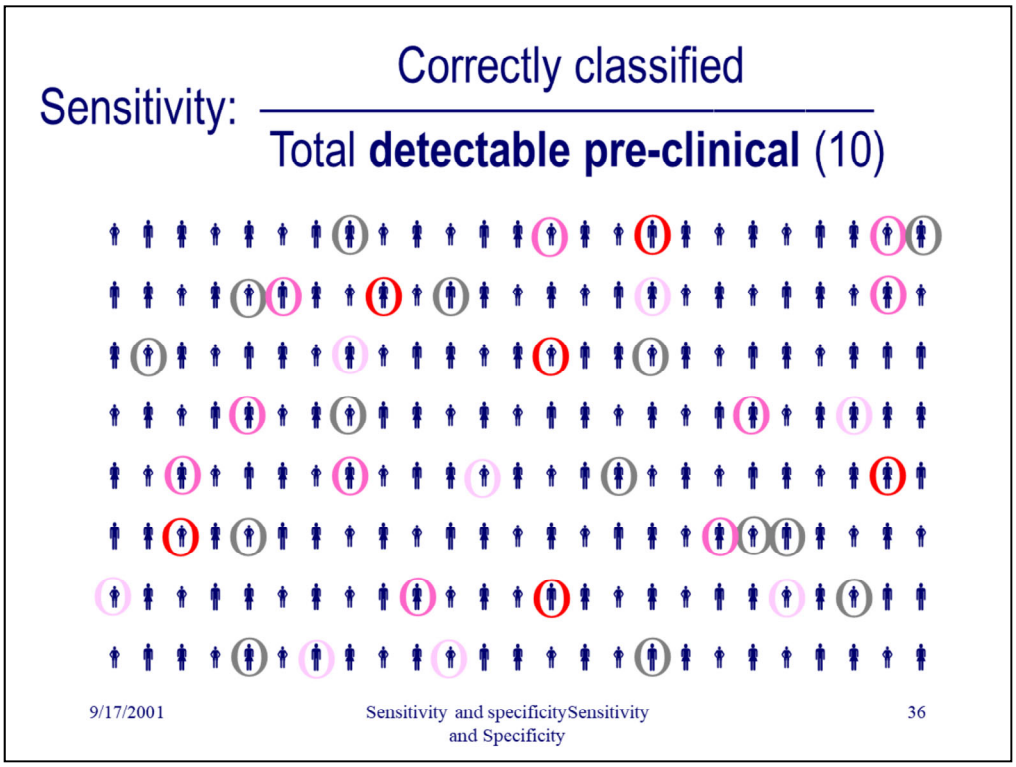



So for a screening test we can write a more precise definition of sensitivity as the probability of correctly classifying someone who has a detectable, pre-clinical case.

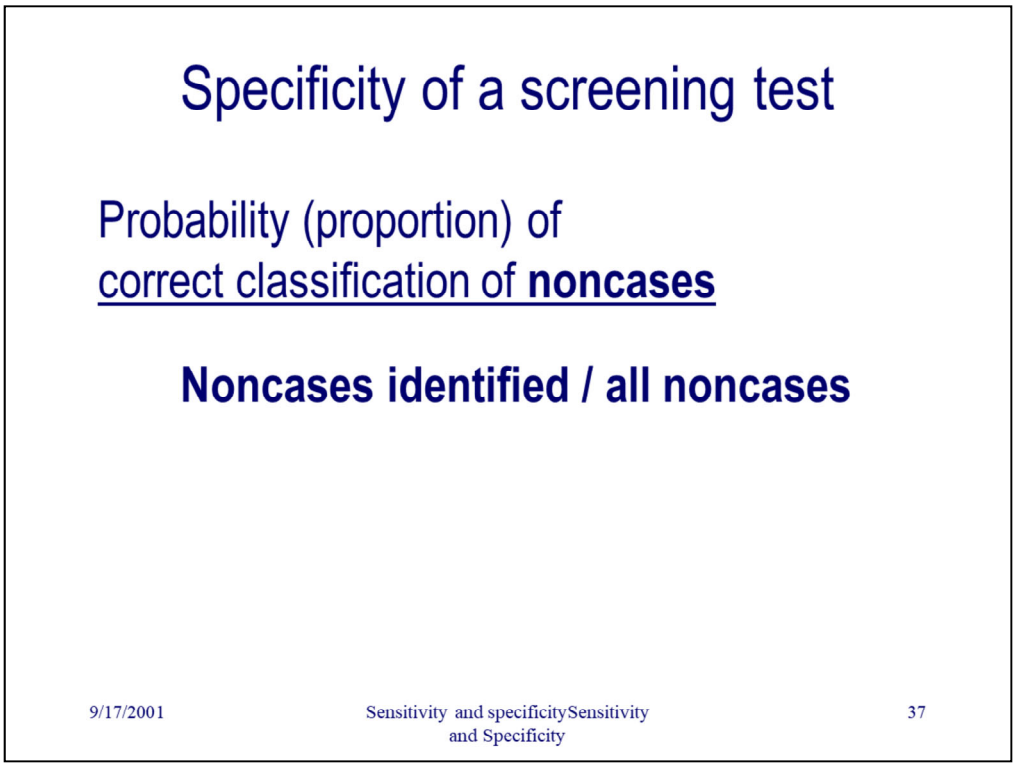



So what would the denominator of a sensitivity estimate be for this population?

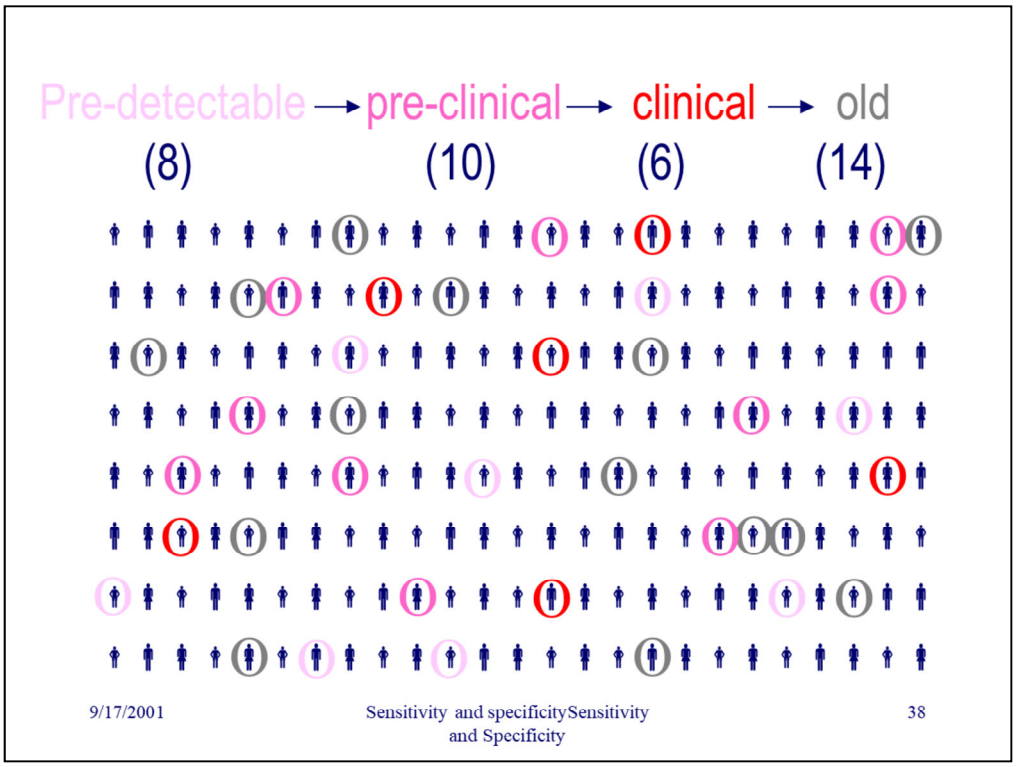
As noted, sensitivity concerns the correct classification of detectable, pre-clinical cases, so the denominator would be 10.



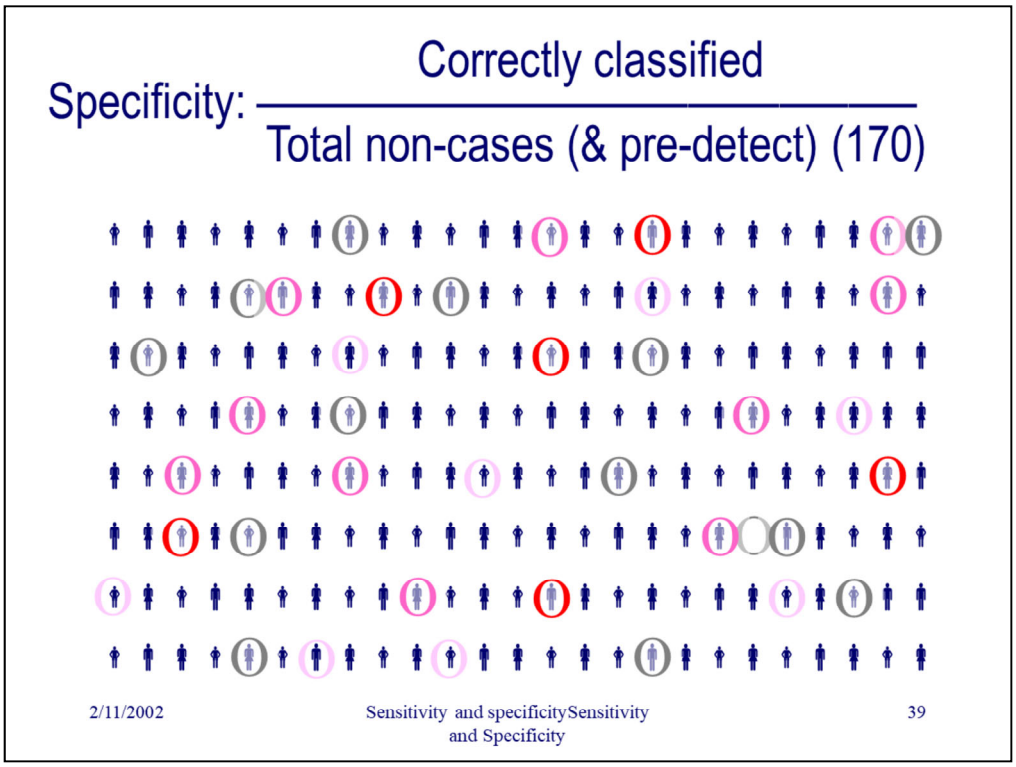

The sensitivity of a screening test that we applied to this population would be the number of correctly classified cases from among the 10 people with detectable, pre-clinical disease.



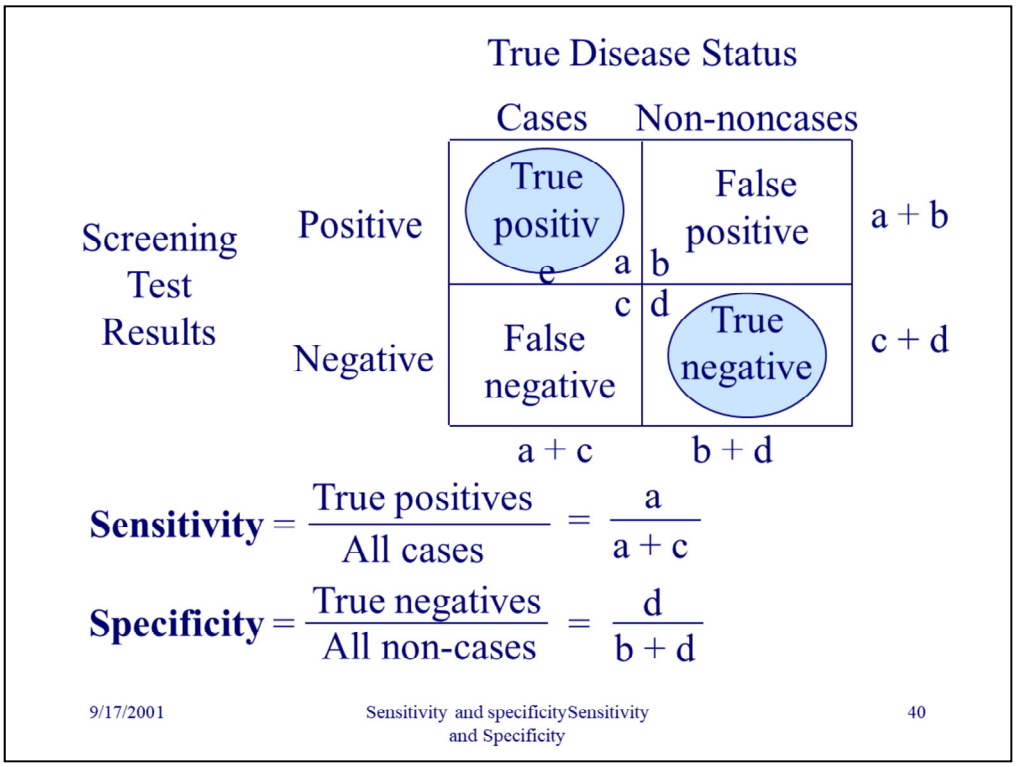

Now, consider the other aspect of validity, correct classification of noncases. Specificity is the probability of correctly classifying someone without the condition, so who would fall into that category?




The denominator for specificity would be the 170 people who have neither clinical disease (new and old) nor detectable, pre-clinical disease. Since this diagram shows us the 8 people with pre-detectable disease, we may feel somewhat uncomfortable including these 8 people as non-cases for our specificity calculation. However, these people are by definition not detectable, so we would never know how many of them there were anyway.



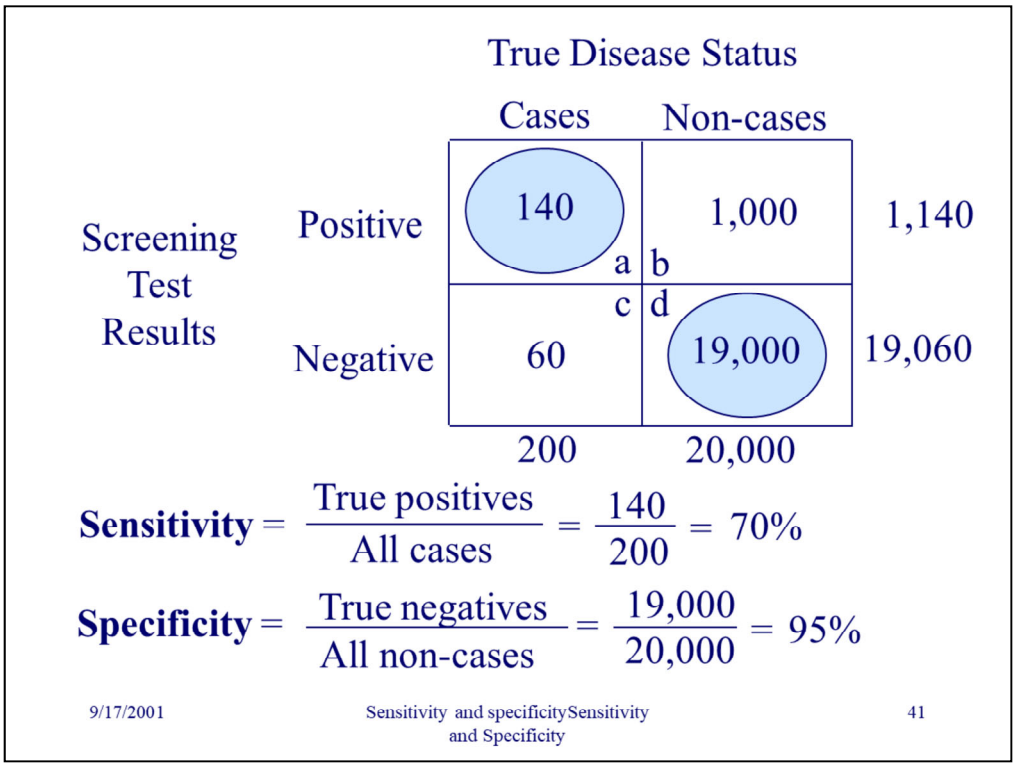

So our specificity measure would be the number of correctly classified non-cases divided by the 170 people with neither clinical nor detectable, pre-clinical disease (here, the people with detectable disease are grayed out so that it's easier to see the population from which specificity is calculated).




Data for estimating sensitivity and specificity are typically displayed in a 2 x 2 table that classifies people according to their disease status and test results. The above table has the True disease status along one dimension, with a column for cases and a column for non-cases, and the Test results on the other dimension, with a row for people who tested positive and a row for people who tested negative. In the top left-hand corner – the "a" cell – are the people who have the disease and whose test came up positive. They are "true positives", cases who were correctly classified. In the lower right-hand corner – the "d" cell – are the people who do not have the disease and whose test came up negative. They are "true negatives", non-cases who were correctly classified. The other two cells, b and c, contain people who were misclassified. Non-cases given who nevertheless received a positive test are often called "false positives", and cases who received a negative test are often called "false negatives", but these terms are not always employed with these meanings.

If cell "c" is in the lower left-hand corner of the table, then the left-hand column – the cases – has a total of (a + c) people, and we we can write the formula for sensitivity as a / (a+c): the number of cases correctly classified divided by the total number of cases.

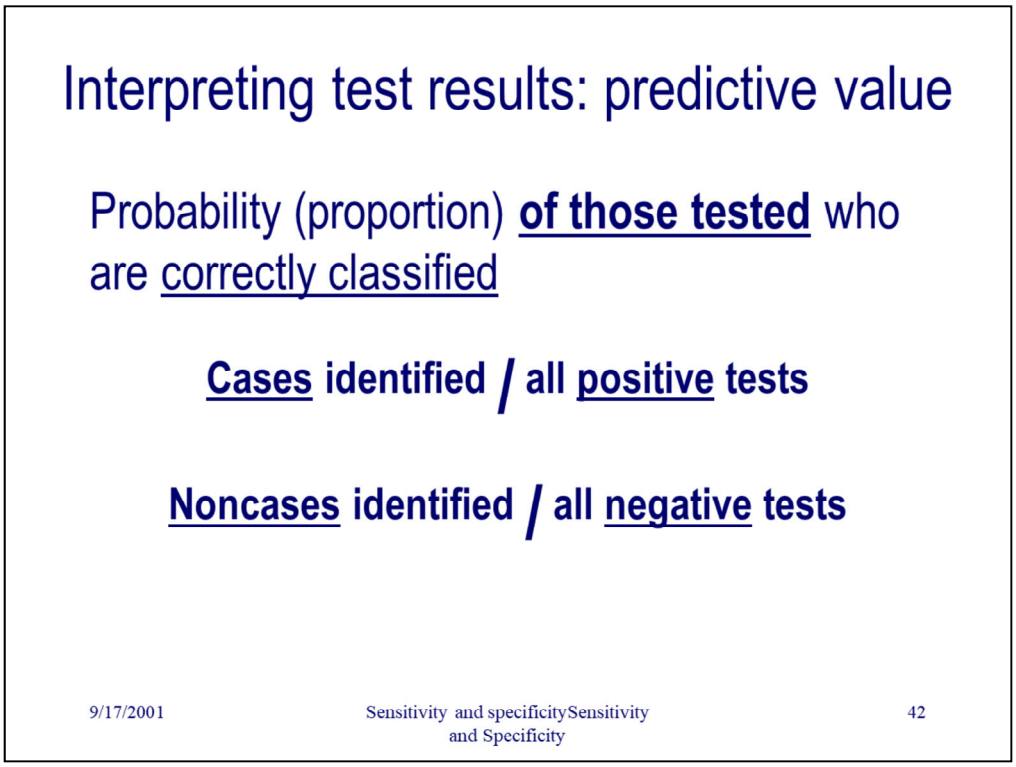
Similarly, the formula for specificity is d / (b+d): the number of correctly classified non-cases divided by the total number of non-cases.




If a population has a total of 200 cases, and the test correctly identifies 140 of them as cases, then a = 140, a+c = 200, and the sensitivity is: a / (a+c) = 140 / 200 = 70%

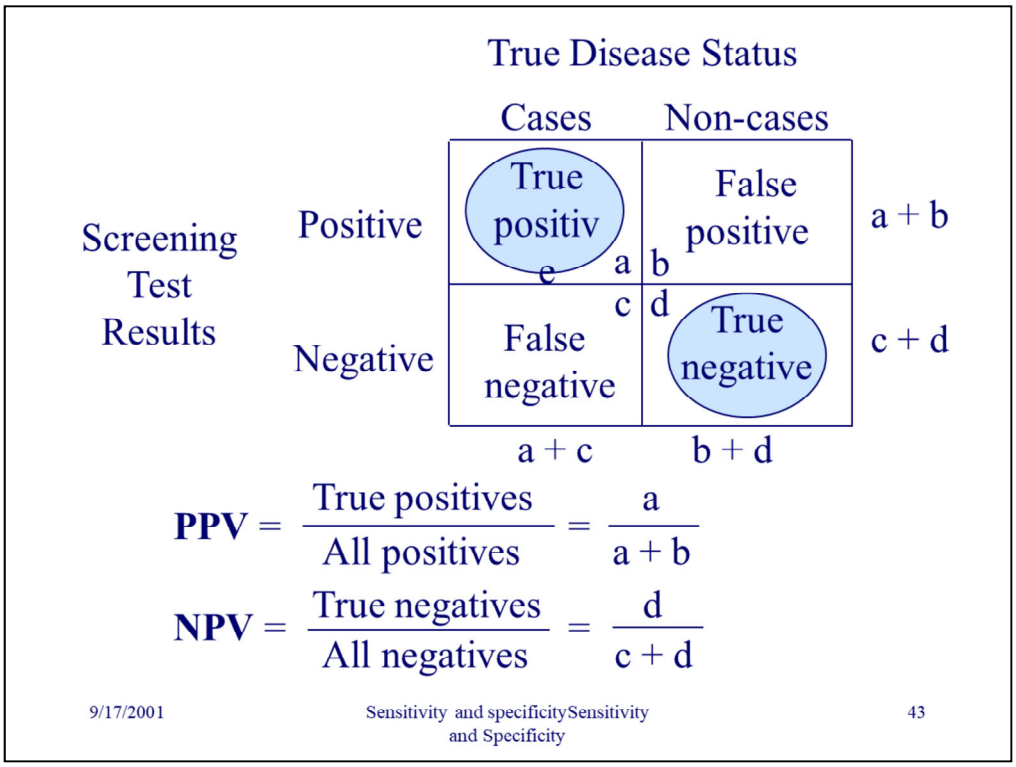
If there are 20,000 people without the disease, and the test correctly classifies 19,000 of them as non-cases, then d = 19,000, b+d = 20,000, and the specificity is: b / (b+d) = 19,000 / 20,000 = 95%.

As is often the case for a rare disease, even with what seems like a high specificity (95%), the number of false positives can easily exceed the number of true positives. This observation brings us to the concept of predictive value.



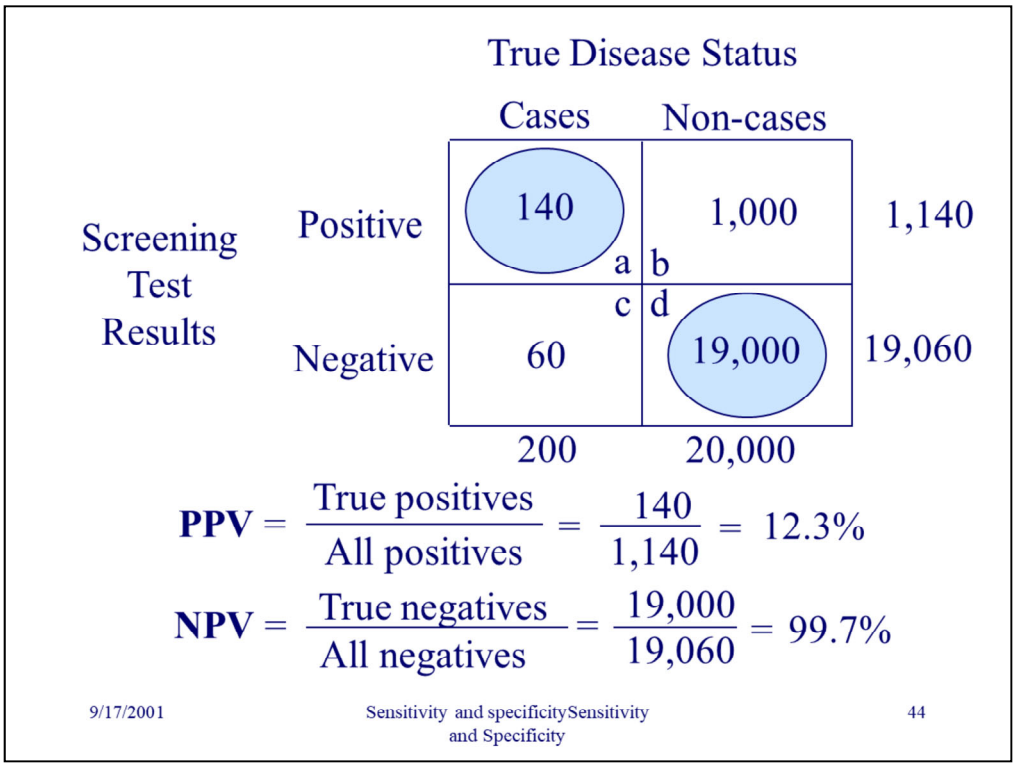

Sensitivity and specificity tell us what happens to cases and non-cases, respectively. However, appropriate interpretation of the results of a test – both screening tests and diagnostic tests – makes use of another concept that is very important for both the epidemiologic and the clinical perspectives, predictive value.

Predictive value is also a probability of correct classification, but here the universe for the probabilities is based on the way people have been classified by the test. There are two predictive values – predictive value of a positive test and predictive value of a negative test. Predictive value tells us the probability that the test was correct. This is obviously a key question for the clinician (and the patient), since we generally do not know whether someone is a case or not, but we do know whether the person tests positive or negative.




The table for examining predictive value is the same as that for sensitivity and specificity. Instead of using the total numbers of cases and non-cases, though, predictive value involves the total number of people with a positive test and the total number with a negative test. Positive predictive value, abbreviated PPV or PV+, is the proportion of all people with positive tests who truly have the condition – a / (a+b) in the above table.

Negative predictive value (NPV or NP-) is the proportion of all people with negative tests who truly do not have the condition – d / (c+d) in the above table.




Using the same numbers as in our example for calculating sensitivity and specificity, we find that the predictive value of a positive test (PPV) is only 140 / 1,140 = 12.3%. The predictive value of a negative test (NPV) is 19,000 / 19,060 = 99.7%.

Although the NPV is very high, that is not such an impressive result in this population, since the prevalence of the condition is only 200 / 20,200, which is not quite 1%. That means that if we select a person at random from the population, there is a 1% probability that the person will be a case. The probability that a person who tests positive actually is a case is 12.3%, so the test raises increases the probability substantially. On the other hand, the probability that a person randomly selected from the population does not actually have the condition is already 99%, so the further information that the person tested negative cannot provide that much more information.

However, the PPV of 12.3% poses a dilemma. Of the 1,140 people who tested positive, the vast majority – 87.7% – are falsely positive. They do not have the disease. If every person who gets a positive test must undergo an expensive and possibly painful diagnostic work-up, then 7 people who do not have the disease and therefore will not derive any benefit will have to have the work-up for every person who does have the disease and may benefit. This scenario is a common dilemma in population screening for a rare disease.

## Positive predictive value, Sensitivity, specificity, and prevalence

| Prevalence (%) | PV+ (%) | Se (%) | Sp (%) |
|---|---|---|---|
| 0.1 | 1.8 | 90 | 95 |
| 1.0 | 15.4 | 90 | 95 |
| 5.0 | 48.6 | 90 | 95 |
| 50.0 | 94.7 | 90 | 95 |

6/13/2002 Sensitivity and specificitySensitivity and Specificity 45

The above table illustrates the relation among positive predictive value (PPV), sensitivity, specificity, and prevalence of the condition. Note that sensitivity and specificity are being regarded as properties of the test, unaffected – in principle – by the rarity of the condition. In contrast, prevalence is a property of the population in which the test is being used, and PPV shows the result of applying a test with a given sensitivity and specificity to a population with a given prevalence.

For a sensitivity held constant at 90% and specificity held constant at 95%, PPV is only 1.8% for a disease with a prevalence of 1 in 1,000, but rises to almost 50% when the prevalence is 5%. This table illustrates the difference between using a test for screening and for diagnosis. Using the test in the general population, where the disease is rare (say, less than 1%), will result in a predictive value below 15% – most people who test positive will not have the condition. In contrast, people with symptoms are much more likely to have the condition. If the prevalence among them is above 5%, then the proportion of false positive tests is greatly reduced.

The challenge in population screening is to try to target a population at sufficiently high risk that the false positive rate (1 – specificity) is acceptable and yet a sufficient proportion of the cases are included.

A point often not mentioned in introductory presentations is that while sensitivity and specificity are in principle properties of the test, in practice a test is not a fixed entity. Various factors can affect the sensitivity and specificity of a test when it is actually implemented, since there are often human factors involved in interpreting test results, equipment may require frequent calibration, etc.

## Example: Mammography screening of unselected women

Disease status

| | Cancer | No cancer | Total |
|---|---|---|---|
| Positive | 132 | 985 | 1,117 |
| Negative | 47 | 62,295 | 62,342 |
| Total | 179 | 63,280 | 63,459 |

Prevalence = 0.3% (179 / 63,459)

Se = 73.7% Sp = 98.4% PV+ = 11.8% PV– = 99.9%

Source: Shapiro S et al., Periodic Screening for Breast Cancer



The above table shows data from the classic randomized trial of breast cancer screening conducted by Shapiro and colleagues in New York City's Health Insurance Plan (HIP). The table is taken from Gordis' textbook.

The prevalence of breast cancer is less than 1%, and even with specificity of greater than 98% only 11.8% of the women with a positive mammogram actually had breast cancer (PPV=11.8%).

Note that in a real study, only people with a positive test undergo a diagnostic work-up. So we can learn the numbers of true positives and false positives, which permit us to calculate predictive value. Estimating sensitivity and specificity is more problematic, since we do not know how many cases were missed, i.e., the number of false negatives. We can estimate that number based on the number of people who develop the disease soon after being screened, but you can see why that method leaves something to be desired. Without knowing the false negatives we also do not know the exact number of true negatives. For this reasons, Gordis' textbook says that sensitivity and specificity cannot be estimated from the results of a screening program. However, if we have a good estimate of the prevalence of the disease or at least know that it is very low, then we can often make a good estimate of the specificity based on the number of false positives and the estimated number of cases (and therefore of non-cases) based on disease prevalence. Note that a small inaccuracy in the denominator of the specificity will not alter the numerical result nearly as much as will a small inaccuracy in the denominator of the sensitivity estimate.

## Effect of Prevalence on Positive Predictive Value

Sensitivity = 93%, Specificity = 92%

| | | Surgical biopsy ("gold standard") Cancer | No cancer | Prev. |
|---|---|---|---|---|
| Without palpable mass in breast | | | | |
| Fine needle aspiration | Positive | 14 | 8 | 13% |
| | Negative | 1 | 91 | PV+ = 64% |
| With palpable mass in breast | | | | |
| Fine needle aspiration | Positive | 113 | 15 | 38% |
| | Negative | 8 | 181 | PV+ = 88% |

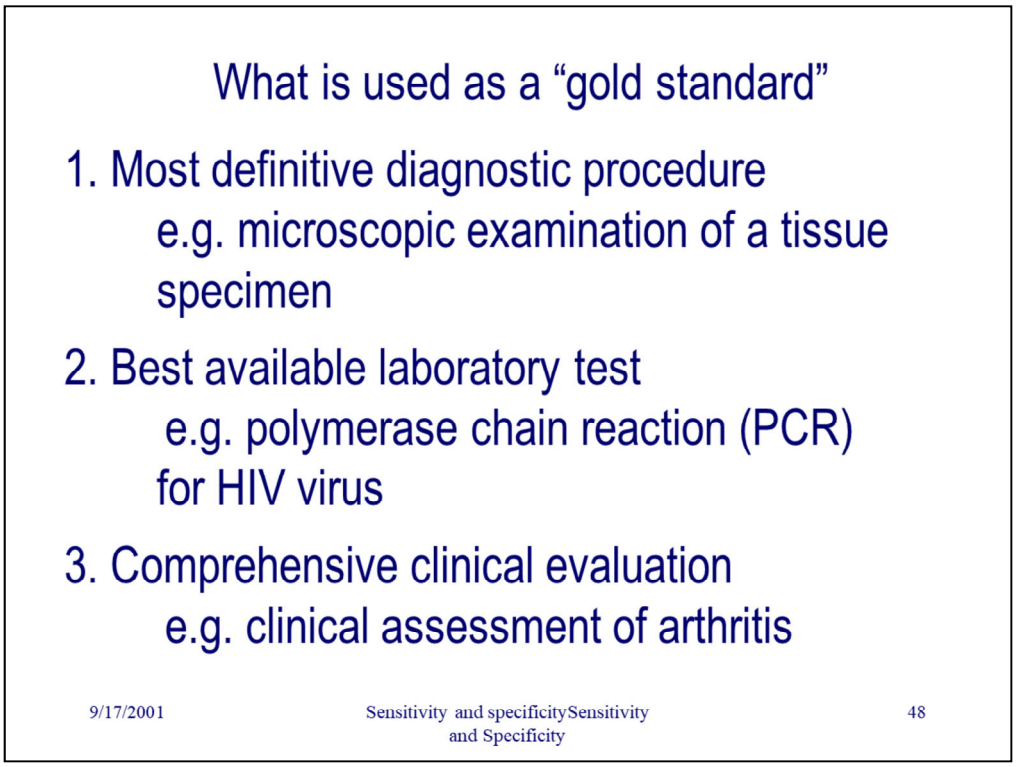
9/17/2001 Sensitivity and specificitySensitivity and Specificity 47

To provide a taste of clinical epidemiology, for those who would like it, here are possible results from the use of final needle aspiration to evaluate a suspected breast cancer. The compares the results from fine needle aspiration with the determination of cancer based on a surgical biopsy, whose results are regarded as definitive. The upper portion of the table shows the data for women with a positive mammogram but no palpable breast mass. 13% of these women actually have breast cancer. The PV+ is 64%.

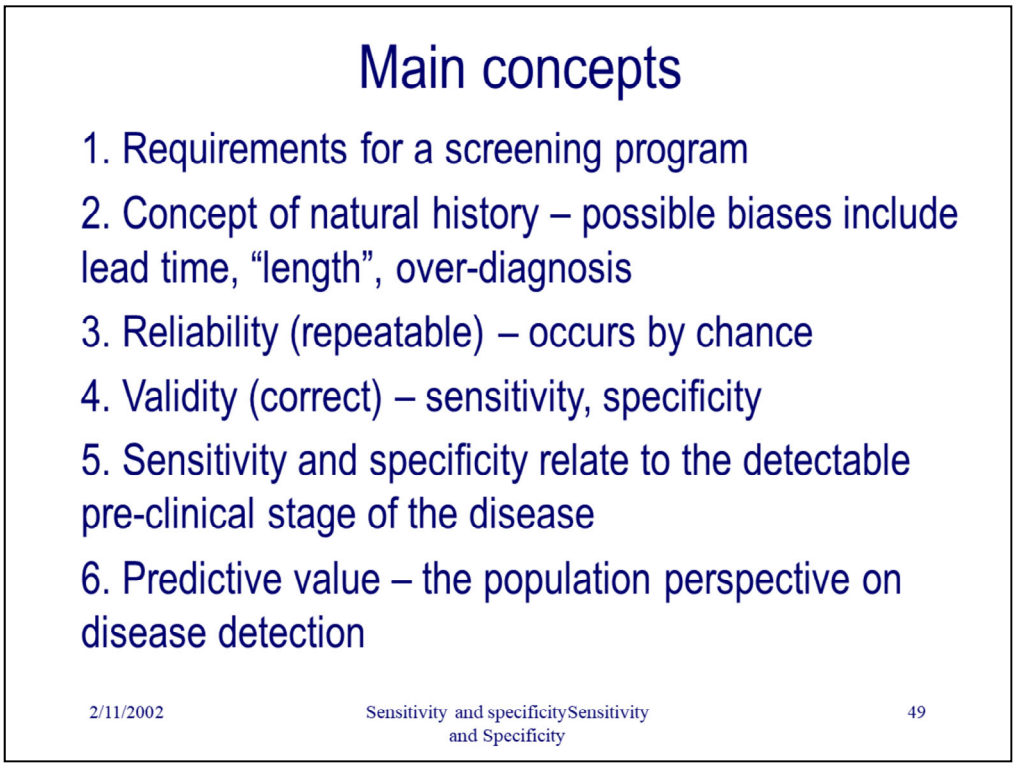
The lower portion of the table shows women with a palpable breast mass. Here, 38% of the women have breast cancer, and the PV+ is considerably greater, at 88%. The prevalence is referred to as the "prior probability" or "pretest probability", since it represents the probability that a randomly selected women has cancer. The PV+ is referred to as the "posttest probability" or "posterior probability", since it represents the probability that a woman with a positive test has cancer. The higher the ratio of the posterior to the prior probability, the more informative the test.

In clinical medicine, the informativeness of a diagnostic test is quantified by its likelihood ratio. The likelihood ratio of a positive test is ratio of the sensitivity of the test to its false positive rate (1 – specificity). It turns out that this ratio equals the ratio of the posterior odds to the prior odds (recall that the odds are simply the ratio of the probability to its inverse). The greater the likelihood ratio of a positive test, the more likely it is that a positive test indicates the presence of the condition. (For a detailed exposition, see David L. Sackett, R. Brian Haynes, Gordon H. Guyatt, Peter Tugwell. *Clinical epidemiology: a basic science for clinical*

*medicine*. 2nd ed. Boston, Little, Brown, 1991.)




As noted, calculation of sensitivity and specificity, and therefore predictive value as well, requires a way to authoritatively determine who has and does not have the condition of interest. This "gold standard" is typically the most definitive diagnostic procedure (for example, the definitive diagnosis of cancer is generally based on microscopic examination of a tissue specimen), the best avaialble laboratory test (for example, a polymerase chain reaction (PCR) test for the actual virus, as opposed to a test for antibody to the virus), or a comprehensive clinical evaluation, where there is no definitive laboratory test. For example, the best diagnosis for arthritis might be obtained through an examination.




There is a lot of detail in this lecture, so it may be helpful to review briefly the main points.

1. What are the requirements for an effective screening program – suitability of the disease, the test, and the program (and an appropriate use of resources).

2. Some understanding of the concept of natural history – disease is a process – and ways in which we can be misled if we fail to take account of the way in which early detection interacts with natural history and disease heterogeneity (lead time, rate of progression, and over-diagnosis)

3. The two dimensions of accuracy – reliability (repeatability), and the fact that some agreement occurs by chance alone

4. Validity, the other dimension, which requires that we know the "truth", and its two components – sensitivity and specificity

5. Sensitivity and specificity when used for population screening relate to the detectable, pre-clinical stage of the disease, rather than to clinical disease

6. Predictive value provides essential population and clinical perspectives for interpreting the results of screening and diagnostic tests