Chapter 7
Screening and Diagnostic Tests

Fen Liu

Key Points

- Screening is the process of using quick and simple tests to identify and separate persons who have an illness from apparently healthy people.
- The validity of a screening test is defined by its ability to correctly categorize subjects who do or do not have a disease into corresponding groups. The components of validity include sensitivity, specificity, Youden's index, and likelihood ratio. Reliability is an index that reflects the stability of the testing results. That includes agreement rate and Kappa statistic. The PPV is defined as the probability of the persons having the disease when the test is positive. The NPV is the percentage of the persons not having the disease when the test is negative. The position of the cutoff point for a screening test will determine the number of true positives, false positives, false negatives, and true negatives. For continuous measurement data of a screening test, the cutoff point is determined mostly by the ROC curve.
- Screening the high-risk population or performing multiple tests increased the validity of a screening test.
- Volunteer bias, lead-time bias, and length-time bias are three major sources of bias in screening test.

Screening is an effective strategy for early detection of diseases and is considered a secondary prevention program in public health; diagnostic tests are helpful in confirming diagnoses of diseases and can help the doctors determine the therapeutic plans for patients. Along with the progress in science and technology, novel

F. Liu (✉)
School of Public Health, Capital Medical University, Beijing, China
e-mail: liufen05@ccmu.edu.cn

C. Wang, F. Liu (eds.), *Textbook of Clinical Epidemiology*,
https://doi.org/10.1007/978-981-99-3622-9_7

screening and diagnostic tests are continuously put forward. Thus, the quality of screening and diagnostic tests is a critical issue. In this chapter, we will address the questions on how to assess the quality of various screening and diagnostic methods, in particular, the newly available ones, and how to make reasonable decisions on their application.

7.1 Design a Screening or Diagnostic Test

Although the purpose, observational subjects, and requirement of screening and diagnostic tests are different, the principle for the evaluation of these two types of tests is similar. Therefore, we take a screening test assessment as an example to discuss.

Screening is the process of using quick and simple tests to identify and separate persons who have an illness from apparently healthy people. To evaluate a new screening test, we need to compare the results of the test to that of a standard test, which is called the "gold standard" via using the blinding method.

7.1.1 Gold Standard (Reference Standard)

A "gold standard" method refers to the most reliable method to diagnose a disease, which is also referred to as standard diagnosis. Application of gold standard can distinguish whether the disease is truly present or not. The gold standard can be biopsy followed by pathological examination, surgical discovery, bacteria cultivation, autopsy, special examination, and imaging diagnosis; it also can be an integrated combination of several diagnostic criteria (such as Jones diagnosis standard, etc.). The outcomes of long-term clinical follow-up obtained by applying the affirming diagnostic methods were also used for the gold standard.

7.1.2 Study Subjects

The subjects of a screening test include the case group who has a specific disease and controls who do not have the disease. They should be representative of the target population. Therefore, the case group should include various types of the studied disease: mild, moderate, or severe; early, middle, or late stage; typical or atypical; with or without complication; treated or untreated, in order to make the result of the study more representative and applicable to the general population. In contrast, the control group should include individuals without the studied disease, but with other illnesses, particularly those that are not easily distinguishable from the studied disease. The testing of study subjects should be kept within the same research period through either continuous sampling or proportional sampling, rather than by the researchers' choice. Otherwise, a selection bias may be present, which influences the validity and reproducibility of the test.

7.1.3 Sample Size

The sample size is determined based on the following factors: sensitivity, specificity, permissible error, and alpha level. The formula for sample size calculation is as follows:

$$n = \frac{{Z_\alpha}^2 p(1-p)}{\delta^2} \tag{7.1}$$

n is the sample size of abnormal or normal subjects in the study.

Z_α is the Z value for normal distribution of cumulative probability, which is equal to α/2.

δ is admissible error, usually, it is set at a 0.05 ~ 0.10 level.

p is the estimation of sensitivity or specificity of the test. Sensitivity is used to calculate the sample size of the case group, while specificity is used for the control group. This formula requires the sensitivity or specificity approaching 50%. When the sensitivity or specificity ≤20% or ≥80%, the corrected formula is needed:

$$n = \left[\frac{57.3 Z_\alpha}{\sin^{-1}\left(\delta/\sqrt{p(1-p)}\right)}\right]^2 \tag{7.2}$$

7.2 Evaluation of a Screening Test

When evaluating a new screening test for a disease, the gold standard for the disease should be used simultaneously. The subjects will be divided into two groups based on the test results: case group and control group (non-disease group). The results of the gold standard and the screening test are then compared. The first step of this comparison is to generate a two-by-two table and calculate several indexes.

As shown in Table 7.1, in cell *a*, the disease of interest is present, and the screening test result is positive, a true-positive result. In cell *d,* the disease is absent, and the screening test result is negative, a true-negative result. In both *a* and *d* cells, the screening test result agrees with the actual status of the disease. Cell *b* represents individuals without the disease who have a positive screening test result. Since these test results incorrectly suggest that the disease is present, they are considered to be false positives. Subjects in cell *c* have the disease but have negative screening test

Table 7.1 Comparison of the results of a screening test with the gold standard

Screening test	Gold standard		Total
	Patients	Controls	
Positive	True positive (*a*)	False positive (*b*)	$a + b$
Negative	False negative (*c*)	True negative (*d*)	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d$

results. These results are designated false negatives because they incorrectly suggest that the disease is absent.

7.2.1 Validity of a Screening Test

The validity of a screening test is defined by its ability to correctly categorize subjects who do or do not have a disease into corresponding groups. The components of validity include sensitivity, specificity, Youden's index, and likelihood ratio.

7.2.1.1 Sensitivity

The sensitivity of a screening test is defined as the proportion of persons with the disease in the screened population who are identified as ill by the test. Sensitivity is calculated as follows:

$$Sensitivity(Sen) = \frac{a}{a+c} \times 100\% \tag{7.3}$$

If someone with the disease is incorrectly called "negative," it is a false-negative result. The false-negative rate is complementary to sensitivity.

7.2.1.2 Specificity

Specificity of a test is defined as the proportion of disease-free people who are so identified by the screening test. Specificity is calculated as follows:

$$Specificity(Spe) = \frac{d}{d+b} \times 100\% \tag{7.4}$$

If some people without a disease are incorrectly called "positive," it is a false-positive result. The rate of false-positive is complementary to the specificity.

7.2.1.3 Youden's Index

Youden's index (*YI*) is also called the accuracy index, which is frequently used to evaluate the overall performance of a test. The formula of the Youden's index is:

$$YI = Sen + Spe - 1 \tag{7.5}$$

It ranges from 0 to 1. The greater the index is, the better the validity.

7.2.1.4 Likelihood Ratio

The likelihood ratio (*LR*) reflects the validity of screening test; it is an integrative index that can reflect the sensitivity and specificity altogether, i.e., the ratio of true-positive or false-negative rates in disease group to the false-positive or true-negative rates in the group without the disease. Using the results of the screening tests, we can calculate all the *LR* of the tests, which thus reflect the overall validity of a screening test.

The positive likelihood ratio of a screening test is the ratio of true-positive rate to false-positive rate, and negative likelihood ratio is a ratio of false-negative rate to true-negative rate. The computation formulas for positive likelihood and negative likelihood ratios are as follows:

$$\mathrm{LR}^{+} = \frac{\mathrm{a}/(a+c)}{b/(b+d)} = \frac{Sen}{1 - S\mathrm{pe}} \tag{7.6}$$

$$\mathrm{LR}^{-} = \frac{c/(a+c)}{d/(\mathrm{b}+d)} = \frac{1 - Sen}{Spe} \tag{7.7}$$

The likelihood ratio is more stable than sensitivity and specificity, and it is less influenced by prevalence.

There is an example that would be helpful in understanding the calculation of these indices.

Example Suppose, we perform a diabetes screening test in a cohort of 1000 people, of whom 20 are diabetic patients and 980 are not. A test is available that can yield either positive or negative results. We want to use this test to distinguish subjects who have diabetes from those who do not. The results are shown in Table 7.2. How do we evaluate the validity of the screening test?

These results showed that of the study population, 90% were positive in the screening test, but the remaining 10% were not diagnosed. Among the individuals without diabetes, 95% tested negative with the screening, and 5% were misdiagnosed in the screening.

Table 7.2 The results of a screening test and the gold standard test for diabetes

	Gold standard		
Results of screening	Have the disease	Don't have the disease	Total
Positive	18	49	67
Negative	2	931	933
Total	20	980	1000

Sensitivity = (18/20) × 100% = 90%
Specificity = (931/980) × 100% = 95%
False-negative rate = (2/20) × 100% = 10%, or 1 – 90% = 10%
False-positive rate = (49/980) × 100% = 5%, or 1 – 95% = 5%
Youden's index = 0.90 + 0.95 – 1 = 0.85
$LR^{+} = 0.90/0.05 = 18.00$
$LR^{-} = 0.10/0.95 = 0.11$

7.2.2 Evaluation of the Reliability of a Test

Reliability or repeatability is an index that reflects the stability of the testing results, i.e., if the results are replicable when the test is repeated. In a study, almost all variations of measured data stem from the observer's variation (intra-observer and inter-observer variation), measuring instruments, reagents variation, and research object's biological variation (intra-subject variations), etc.

7.2.2.1 Coefficient of Variation

For a continuous variable, the variations of data are commonly measured with standard deviation (*SD*) and coefficient of variation. The coefficient of variation (*CV*) is obtained by dividing the *SD* by mean (percentage).

$$CV = \left(\frac{SD}{\overline{X}}\right) \times 100\% \tag{7.8}$$

7.2.2.2 Agreement Rate and Kappa Statistic

Agreement (consistency) rate is also called accuracy rate, which is defined as the proportion of the combined true positive and true negative number of the total population evaluated by a screening test, i.e., the percentage of the results of a screening test that is in accordance with those of the gold standard method. Below is the formula for calculating accuracy rate:

$$\text{Agreement rate} = [(a + d)/(a + b + c + d)] \times 100\% \tag{7.9}$$

Table 7.3 Kappa value judgment standard

Kappa value	Consistency strength
<0	Poor
0 ~ 0.2	Weak
0.21 ~ 0.40	Light
0.41–0.60	Moderate
0.61 ~ 0.80	High
0.81 ~ 1.00	Strong

For counted variable, the observation coincidence rate or kappa statistic is used to determine data reliability (repeatability or precision).

This is the calculation of kappa:

$$Kappa = \frac{\begin{pmatrix} Percent \\ agreement \\ observed \end{pmatrix} - \begin{pmatrix} Percent \\ agreement \\ expected \\ by \\ chance \\ alone \end{pmatrix}}{100\% - \begin{pmatrix} Percent \\ agreement \\ expected \\ by \\ chance \\ alone \end{pmatrix}} \tag{7.10}$$

Kappa is an index that judges consistency in levels between different observers. Landis and Koch suggested that kappa greater than 0.75 represents an excellent agreement beyond chance, while a kappa less than 0.40 shows poor agreement, and a kappa of 0.40 to 0.75 represents intermediate to good agreement (Table 7.3). Testing for the statistical significance of kappa, please refer to the relevant book.

7.2.3 Predictive Value

Sensitivity and specificity are indicators of the accuracy of a test, which can be considered the characteristics of a screening or diagnostic test itself. However, the predictive value is affected by both the sensitivity and specificity of the test and the prevalence of the disease in the population to be tested. There are positive predictive value (*PPV* or *PV*+) and negative predictive value *(NPV* or *PV*–).

The *PPV* is defined as the probability of the persons having the disease when the test is positive. The *PPV* is calculated as follows:

$$PPV = \frac{a}{a+b} \times 100\% \tag{7.11}$$

The *NPV* is the percentage of the persons not having the disease when the test is negative.

$$NPV = \frac{d}{c+d} \times 100\% \tag{7.12}$$

Take the data in Table 7.2 as an example again for the calculation of predictive values:

$$PPV = \frac{18}{18+49} \times 100\% = 26.87\%$$

$$NPV = \frac{931}{2+931} \times 100\% = 99.79\%$$

The *PPV* of 26.87% means that 67 individuals are positive in screening, but among them, the number of real patients is 18, accounting for 26.87% of the total positive results. The *NPV* of 99.79% indicates that 933 persons have negative test results, and among them, the number of individuals "not having the disease" is 931, accounting for 99.79% of the total negative results.

Predictive value is affected by the prevalence of a disease in a specific population, or by the pretest probability of the presence of a disease in an individual. We can use the formula derived from Bayesian theorem of conditional probability to show the relationships of predictive value, sensitivity, specificity, and prevalence.

$$PPV = \frac{Sensitivity \times Prevalence}{Sensitivity \times Prevalence + (1 - Specificity) \times (1 - Prevalence)} \tag{7.13}$$

$$NPV = \frac{Specificity \times (1 - Prevalence)}{(1 - Sensitivity) \times Prevalence + Specificity \times (1 - Prevalence)} \tag{7.14}$$

The more sensitive a test is, the higher will be its negative predictive value (the more confident clinicians can be that a negative test result rules out the disease being sought). Conversely, the more specific the test is, the better will be its positive predictive value (the more confident clinicians can be that a positive test confirms or rules in the diagnosis being sought). Because predictive value is also influenced by prevalence, it is not independent of the setting in which the test is used.

As the numbers in Table 7.4 show, positive results even for a very specific test, when applied to patients with a low likelihood of having the disease, will be largely false positives. Similarly, negative results, even for a very sensitive test, when applied to patients with a high chance of having the disease, are likely to be false negatives. In summary, the interpretation of a positive or negative result of a

Table 7.4 The screening results of diabetes in populations with different values of sensitivity, specificity, and prevalence

				Gold standard				
Prevalence (%)	Sensitivity (%)	Specificity (%)	Screening results	Patients	Non-patients	Total	PPV (%)	NPV (%)
50	50	50	+	250	250	500	50	50
			–	250	250	500		
			Total	500	500	1000		
20	50	50	+	100	400	500	20	80
			–	100	400	500		
			Total	200	800	1000		
20	90	50	+	180	400	580	31	95
			–	20	400	420		
			Total	200	800	1000		
20	50	90	+	100	80	180	56	88
			–	100	720	820		
			Total	200	800	1000		

screening or diagnostic test is dependent on the setting in which the test is carried out, in particular, the estimated prevalence of the disease in the target population.

7.2.4 Determination of Cutoff Point for a Screening Test

Ideally, the sensitivity and specificity of a screening test both should be 100%. In practice, when we plot the value of a screening test for a disease group and non-disease group on the same graph, the distribution often overlaps, the test does not separate normal from diseased with 100% accuracy. Figure 7.1 is the schematic graph showing the distributions of test results for patients with and without the disease. The area of overlap indicates where the test cannot distinguish normal and abnormal. We need to determine a balance by an arbitrary cutoff point (indicated by A and B) between normal and disease. The position of the cutoff point will determine the number of true positives, false positives, false negatives, and true negatives. If we want to increase sensitivity and include all true positives, we can use A as a cutoff point, but by doing this, we increase the number of false positives, which means decreased specificity. Likewise, if we want to increase specificity by using B as a cutoff point, it will lead to decreased sensitivity.

We can also use the blood sugar data in Table 7.5 as an example to illustrate how changes in the cutoff point will affect the sensitivity and specificity of a screening test.

To make decisions on the appropriate cutoff point for a screening test, the following principles need to be taken into consideration. For a proven serious disease that can be cured if diagnosed early, a high sensitivity may be suggested. If a false-positive result would detrimentally affect a patient both mentally and physically, such as cancers, which may put a patient at risk of surgery and chemotherapy, a test with high specificity would be required. If both the sensitivity and specificity are important, the junction point of curves might be used as the cutoff point.

Fig. 7.1 Blood sugar level distribution in normal people and diabetes patients

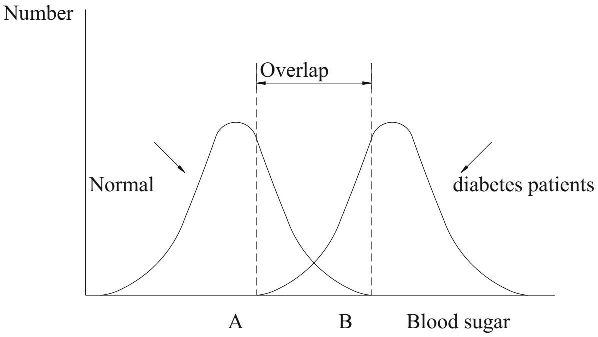

Table 7.5 The effects of cut-off points of 2 h after-meal blood sugar on sensitivity and specificity of the screening test

Blood sugar (mg dL^{-1})	Sensitivity (%)	Specificity (%)
80	100.0	1.2
90	98.6	7.3
100	97.1	25.3
110	92.9	48.4
120	88.6	68.2
130	81.4	82.4
140	74.3	91.2
150	64.3	96.1
160	55.7	98.6
170	52.9	99.6
180	50.0	99.8
190	44.3	99.8

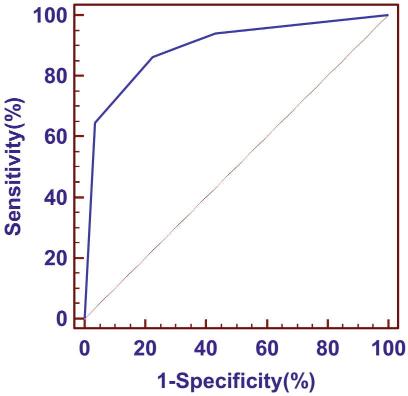

Fig. 7.2 The ROC curve of blood sugar in the diabetes diagnosis

7.2.4.1 ROC Curve

For continuous measurement data of a screening test, the cutoff point is determined mostly by the receiver operator characteristic (ROC) curve. ROC curve is a graphical plot of true positive rate (sensitivity, *Y*-axis) against the false negative rate (1 – specificity, *X*-axis) for different cutoff point. A ROC curve could reflect the relationship between the sensitivity and the specificity of a test (Fig. 7.2). By convention, the point nearest to the top-left corner of the ROC curve is set for optimal cutoff point.

As shown in Fig. 7.2 and Table 7.5, when sensitivity is 88% and specificity is 68%, the sum of the false positive and false negative rates is the minimum. Accordingly, the blood sugar level of 120 mg dL^{-1} can be set as the optimal cutoff point for diabetes screening in this population.

7.2.4.2 The Area under ROC Curve

ROC curves can also be used to compare clinical values of two or more screening tests, thus helping clinicians choose the best screening test. The area under the ROC curve is a measure of the test's accuracy. The larger the area under the ROC curve, the better the diagnostic test. The maximum value for the area under the ROC curve is 1, which indicates a perfect test; an area of 0.5, on the other hand, represents a worthless test.

We can use statistical software, such as MedCalc, SPSS, and SAS, to compute the area under the ROC curve and compare the areas under ROC curve between two or more screening tests (for details, please refer to related statistics books).

7.3 Improving the Efficiency of Screening and Diagnostic Tests

In order to increase the sensitivity and specificity of a screening test, several methods can be used, such as screening high-risk population or performing multiple tests.

7.3.1 Selecting Population with a High Prevalence

The predictive value of a test is influenced by the sensitivity, specificity, and prevalence of a disease. When sensitivity and specificity are constant, it is influenced mainly by the prevalence rate. Since morbidity has larger influence on the positive predictive value, the latter would have very low value if a screening test is carried out in a population with a low prevalence rate of the disease to be tested. However, if a high-risk population is screened, the positive predictive value can be significantly increased.

7.3.2 Use of Multiple Tests

A method combining two or more tests is called multiple tests. In general, multiple tests can be carried out in two ways, simultaneous testing and sequential testing.

7.3.2.1 Simultaneous Testing

In simultaneous testing (parallel tests), the sample is evaluated with more than one screening test simultaneously; a positive result of any test is considered evidence for

Table 7.6 The results of simultaneous and sequential testing

Multiple tests	Test results		Diagnosis
	Test A	Test B	
Simultaneous testing	+	–	+
	–	+	+
	+	+	+
	–	–	–
Sequential testing	+	+	+
	+	–	–
	–	+	–
	–	–	–

Table 7.7 An example of screening results using multiple tests (%)

Screening methods	Sensitivity	Specificity	PPV	NPV
Test A	80	60	33	92
Test B	90	90	69	97
Simultaneous testing (A and B)	98	54	35	99
Sequential testing (A and B)	72	96	82	93

the target disease. Simultaneous testing can improve sensitivity and negative predictive value, but lower the specificity and positive predictive value (Table 7.6).

7.3.2.2 Sequential Testing

Sequential testing (serial testing) means multiple screening tests are used in series, the individual is considered to be positive if all the test results are positive but is stopped when the previous test result is negative. Sequential testing increases specificity and positive predictive value but decreases sensitivity and negative predictive value.

Take the hypothetical example in Table 7.7 as an example, in which a population is screened for hepatocellular carcinoma using ultrasonography and serum alpha-fetoprotein (AFP) level. If two tests with 80% and 90% sensitivity, respectively, were used simultaneously, the sensitivity of the simultaneous testing will be increased up to 98%. However, there is a loss of specificity (decreased to 70%) compared to each test alone. In sequential testing, there is a gain in specificity (increased up to 96%), but a loss in sensitivity (down to 72%).

From the results above, we can summarize the regular pattern of sensitivity and specificity in different multiple tests. How to make the decision to choose either simultaneous or sequential testing is based on the actual situation.

7.4 Potential Bias in Screening Tests

There are three major sources of bias, which are specified to each screening test.

7.4.1 Volunteer Bias

The characteristics may be different between people who attend a screening and those who do not, especially when those factors are directly related to the survival of patients. Individuals with a higher risk of a disease are more likely to voluntarily join a screening program, as they might be more health-conscious and with higher compliance tend to have a better prognosis. For example, women with a significant family history of breast cancer are more likely to join a mammography program than those without it. This tendency is reflected by a higher rate of diagnosis in a series of screening tests than what is truly reflective of the population. Likewise, the screened people tend to have a larger percentage of adverse clinical outcomes than it would be in the general population.

The most effective way to avoid volunteer bias is to recruit a pool of volunteers and then assign them randomly to receive screening or not to receive it.

7.4.2 Lead-Time Bias

Lead time refers to the duration from early detection of disease (usually by screening) to the presentation of clinical symptoms and thus being diagnosed in the standard way. Especially for chronic diseases, the cases of which progress slowly, therefore patients with those diseases are more likely to be detected by screening and likely to have increased survival time than unscreened cases. In fact, the screening has no effect on the outcome of the disease; it only resulted in an earlier diagnosis of the disease when compared to traditional diagnostic methods. To illustrate the lead-time bias, we take a cancer screening test (shown below) as an example. As shown in the illustration, the tumor is detected at different ages with or without the screening test, but the patients die at the same age (Fig. 7.3), indicating that the overall survival of patients is not altered by the screening test.

So, unless we have some idea of the actual lead-time, perhaps from previous studies, we should not use survival time after diagnosis to evaluate a screening program. Instead, we should consider the effects on longer-term age-specific morbidity or mortality rates of the disease. The survival rates are therefore less likely to reflect the true benefits of early treatment better.

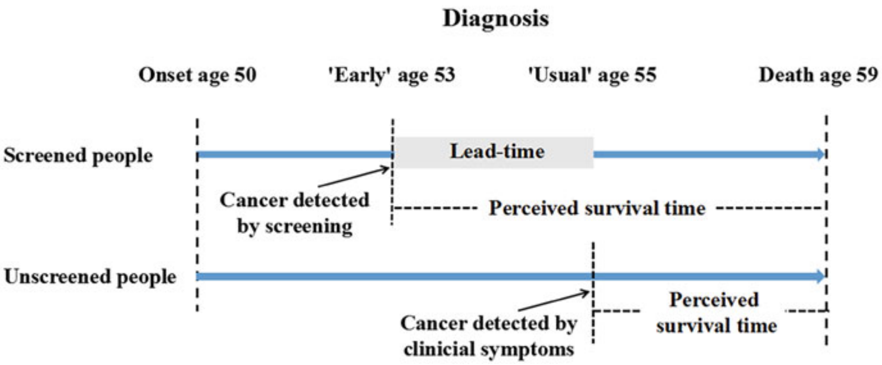

Fig. 7.3 An illustration of lead-time associated with screening and cancer development process

7.4.3 Length-Time Bias

Many screening programs are implemented to detect cancers. Doctors and researchers hypothesize that tumors with low growth rates have better outcomes than more aggressive types. However, it is found that screening is more likely to detect slower-growing, less deadly tumors due to their longer preclinical stages. In other words, patients with concealed, less fatal cancers may not know the fact before their death from other diseases, if without screening. This example results in the “length-time” bias associated with screening tests, which gives the appearance that screening can benefit patients and prolong their life span, when, in fact, the test selectively detects those diseases that progress slowly, thus allowing the patient to live longer.