Performance Measures 3

Performance measures reflect the link between cancer screening test results and cancer diagnoses. They provide no information about cause-specific mortality. Performance measures are used in the initial assessment of proposed cancer screening tests and also are used to monitor performance once cancer screening has disseminated. There are six key performance measures, with each interpretable as a probability (ranging from 0 to 1) or percentage (ranging from 0% to 100%).

Performance measures are calculated from the experience of individuals who have been screened. The cancer screening test result and whether cancer was present at the time of the screen need to be available for each individual to calculate performance measures.

3.1 The Building Blocks of Performance Measures

3.1.1 Cancer Screening Test Result

The cancer screening test result is classified as either positive or negative. A positive result indicates a suspicion of cancer and the need for diagnostic evaluation. A negative result indicates no suspicion of cancer and no need for diagnostic evaluation. The definition of a positive test result is not etched in stone; instead,

P. M. Marcus, *Assessment of Cancer Screening*,
https://doi.org/10.1007/978-3-030-94577-0_3

the medical community makes recommendations as to what constitutes a positive test. In practice, any abnormality deemed suspicious by the test interpreter is called positive, regardless of whether it meets the recommended definition of a positive test. For many cancer screening tests, particularly those than employ imaging, it is impossible for recommendations to include every finding or constellation of findings that creates a suspicion for cancer.

Recommendations are made after many factors are weighed, including the burden of positive tests and the gravity of missing a cancer. Medical communities may arrive at different recommendations. In the US, for example, a prostate-specific antigen (PSA) blood level of 4.0 ng/mL or higher is typically considered a positive test for prostate cancer, but in parts of Europe, a value of 3.0 ng/mL or higher is used.

At the extremes, there tends to be agreement as to whether a cancer screening test result should be classified as positive or negative. For example, a large spiculated lung mass observed on low dose computed tomography (LDCT) would be classified as positive for lung cancer, while a mammogram that shows only the anatomic structures of the breast would be called negative for breast cancer. The challenge comes when it is not obvious what a finding represents: a result that isn't exactly negative and isn't exactly positive. There is a move towards classifying these grey-zone findings as indeterminate and employing a less intense and usually non-invasive form of diagnostic evaluation. Some may disagree with use of the phrase diagnostic evaluation in the instance of indeterminates, as the recommended medical intervention is intended to watch for change in the abnormality rather than determine whether it is cancer. In that instance, the term monitoring can be used. For the purpose of calculating performance measures, I classify indeterminate cancer screening test results as positive. In my opinion, any cancer screening test that is not negative is positive, as it leaves uncertainty in the mind of the clinician and screenee.

Some biospecimen-based cancer screening tests return a numeric value or other quantitative measure. These values

correlate with the chance of the presence of cancer. PSA is one such test. A value greater than 4 ng/mL is usually considered a positive result in the United States, but active surveillance rather than biopsy is often recommended if the PSA is between 4 ng/mL and 10 ng/mL. A value of 10 ng/mL or greater, however, typically leads to imaging or biopsy. Other biospecimen-based cancer screening tests, such as cervical cytology, indicate whether abnormal cells are present. One form of cervical cancer screening, human papilloma virus (HPV) testing, indicates whether certain cancer-causing strains of HPV are present rather than indicating whether an abnormality suspicious for cancer is present.

Imaging-based cancer screening tests are used to determine if abnormalities are present. A cancer screening test will be called positive if an abnormality suspicious for cancer is revealed. These tests also can reveal abnormalities that are not suspicious for cancer and abnormalities whose significance with regard to cancer is unknown. Lung cancer screening with LDCT, for example, can lead to detection of non-calcified nodules (positive if above a certain size), calcified nodules (usually negative), or ground glass opacities (oftentimes of uncertain significance). Some imaging-based cancer screening tests also can lead to detection of abnormalities that represent or are suspicious for non-cancer conditions, called incidental findings or incidentalomas. For example, LDCT screening for lung cancer can lead to the detection of coronary artery calcification.

3.1.2 Cancer: Present or Not?

Cancer is either present or not present at the time of the cancer screening test, though only some cancers that are present can be detected through cancer screening. Recall from Chap. 2 (Fig. 2.1) that Phase A cancers are present but not detectable, while phase B cancers are present and have characteristics that should make them detectable. Knowing whether a cancer is present and detectable at the time of a cancer screening test is often not as simple as the four phase model implies, though. The most challenging

aspect is determining whether a negative screen that occurred prior to a symptom-detected cancer represents a true negative or a false negative, terms that are fairly self-explanatory and will be discussed later in this chapter. The following fictional scenarios represent quandaries that researchers face when trying to assess whether a Phase B cancer was present at the time of a negative screen:

Amanda had a lung cancer screening test and the result was negative. Three months later, she receives a symptom-prompted diagnosis of lung cancer. Was the cancer missed on screening, or was the cancer in Phase A at the time of the screening but moved through Phase B very quickly? Did the cancer exist at the time of the screen?

Arnie had a prostate cancer screening test and the result was positive. He received standard diagnostic evaluation and his clinician concluded that he did not have prostate cancer. Nine months later, he receives a symptom-prompted diagnosis of prostate cancer. Was the cancer in Phase B at the time of the screen but diagnostic evaluation failed in some way? Is the diagnosed cancer a new and fast growing abnormality. In other words, did the diagnosed cancer arise from an abnormality other than the one that prompted the positive result?

Astrid schedules her screening mammogram. Two days before the test, she finds a breast lump but does not tell anyone. Her mammogram is positive and diagnostic evaluation indicates that the lump she found is cancer. Astrid's cancer was present at the time of her mammogram, but should the test be considered a screening mammogram or a diagnostic mammogram?

The phrase interval cancer is used to describe cancers that occur between screening rounds and follow either a negative test or a resolved positive test. Resolved means that the conclusion of the diagnostic evaluation was that cancer was not present. Amanda's cancer and Arnie's cancer are interval cancers regardless of whether they were in Phase A or B at the time of the screen. If in Phase B, the previous screening test would be classified as a false negative.

It is clear that Astrid's cancer was present and in Phase C at the time of the screening test. The cancer could be classified as an interval cancer because it was symptomatic before the screen. Then again, it could be classified as screen detected because the screening test result was positive, even though it was beyond Phase B. Cancer screening tests can miss Phase C cancers, and that could have been Astrid's experience.

Most screen-detected cancers are in Phase B at the time of the cancer screening test. For simplicity's sake Phase C cancers that are detected as the result of cancer screening will be excluded for the remainder of this primer.

3.2 Calculating Cancer Screening Performance Measures

The six cancer screening performance measures are sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), false positive rate (FPR), and false negative rate (FNR). The receiver operating characteristic (ROC) curve is a graph of sensitivity versus FPR, which is equal to 1 minus specificity. The ROC curve demonstrates how those two values vary as the definition of a positive test changes. Its summary measure, area under the curve (AUC), is calculated so that ROC curves can be compared.

3.2.1 The Formulas

Table 3.1 presents the quantities that are needed to calculate performance measures. The four quantities in the center of the table are at the heart of performance measure calculations. They are true positive tests (a), false positive tests (b), false negative tests (c), and true negative tests (d). True positive tests are those positive tests that led to the diagnosis of a cancer, and true negative tests are those negatives tests that correctly indicated no suspicion

Table 3.1 The components of performance measure formulas

		Truth		
		Cancer present (in Phase B)	Cancer not present (includes Phase A)	Total
Screening test result	Positive	*a* true positives	*b* false positives	*a* + *b* all positives
	Negative	*c* false negatives	*d* true negatives	*c* + *d* all negatives
	Total	*a* + *c* cancers present	*b* + *d* cancers not present	*a* + *b* + *c* + *d* all screenees

Cancers in Phase C can be screen detected, but most screen-detected cancers are in Phase B. For simplicity's sake Phase C cancers are not included as cancers that are screen detected

of cancer. False positive tests are sometimes called false alarms; the test suggests something suspicious, but diagnostic evaluation reveals that cancer is not present. False negative tests are incorrectly negative: cancer is present and in Phase B, but the cancer screening test result is negative. Because Phase A cancers cannot be detected by cancer screening, they are considered not to be present when calculating performance measures.

The six performance measures are defined as follows. The formulas use the notation in Table 3.1.

- Sensitivity, sometimes abbreviated as *Se*, is the percentage of people with cancer who had a positive test; $a/(a + c)$.
- Specificity, sometimes abbreviated as *Sp*, is the percentage of people without cancer who had a negative test; $d/(b + d)$.
- PPV is the percentage of people with a positive test who had cancer; $a/(a + b)$.
- NPV is the percentage of people with a negative cancer screening test who did not have cancer; $d/(c + d)$.
- FPR is the percentage of people without cancer who had a positive test; $b/(b + d)$. FPR equals 1 *minus specificity*.

- FNR is not typically reported but will be defined here for completeness' sake. It is the percentage of people with cancer who had a negative cancer screening test; $c/(a + c)$. FNR equals 1 *minus sensitivity*.

Positivity and negativity rate usually are not referred to as performance measures, but it is important to present them nonetheless:

- Positivity rate is the percentage of people screened who have a positive test; $(a + b)/(a + b + c + d)$
- Negativity rate is the percentage of people screened who had a negative test; $(c + d)/(a + b + c + d)$.

Table 3.2 presents data from the Breast Cancer Surveillance Consortium (BCSC) Data Explorer, a public-access database of mammographic breast cancer screening experience from 1994 through 2009 [1]. These data are used in Table 3.3 to calculate the performance measures.

In the BCSC example, sensitivity and specificity are fairly high, as is often the case with cancer screening tests that are used in population-based cancer screening. The manner in which a positive test is defined generally drives sensitivity and specificity,

Table 3.2 Screening mammogram classification among women ages 50 to 59 at the time of screening. Breast Cancer Surveillance Consortium Data Explorer, 1994–2009

		Truth		
		Cancer present	Cancer not present	Total
Screening test result	Positive	7044 (true positives)	165,115 (false positives)	172,159
	Negative	1534 (false negatives)	1,623,399 (true negatives)	1,624,933
	Total	8578	1,788,514	1,797,092 (all screens)

Table 3.3 Performance measures, positivity rate, and negativity rate: formulas and calculations using data from Table 3.2

Performance measure	Formulas using Table 3.1 notation	Calculations using Table 3.2 data
Sensitivity	$a/(a + c)$ *true positives/cancers present*	7044/8578 82%
Specificity	$d/(b + d)$ *true negatives/cancer not present*	1,623,399/1,788,514 91%
PPV	$a/(a + b)$ *true positives/all positives*	7044/172,159 4%
NPV	$d/(c + d)$ *true negatives/all negatives*	1,623,399/1,624,933 >99%
FPR	$b/(b + d)$ *false positives/cancer not present* also equal to $1 -$ *specificity*	165,115/1,788,514 9%
FNR	$c/(a + c)$ *false negatives/cancers present* also equal to $1 -$ *sensitivity*	1534/8578 18%
Positivity rate	$(a + b)/(a + b + c + d)$ *all positives/all screened*	172,159/1,797,092 10%
Negativity rate	$(c + d)/(a + b + c + d)$ *all negatives/all screened*	1,624,933/1,797,092 90%

as do the capabilities and limitations of the cancer screening test itself. The definition of a positive screen is chosen so that most cancers are found (high sensitivity) and the absolute number of false positives is kept as low as possible (low FPR, or high specificity). NPV is high as well but PPV is very low.

3.2.2 The Relationship Between PPV, NPV, and Prevalence

PPV and NPV are driven by sensitivity and specificity, and they also are driven by the prevalence of disease. PPV and NPV can be calculated from sensitivity, sensitivity, and the prevalence of disease using the formulas in Box 3.1.

Box 3.1 Calculating PPV and NPV from sensitivity (Se), specificity (Sp), and prevalence

$$PPV = \left(Se\ x\ prevalence\right) / \left(Se\ x\ prevalence + \left(1 - Sp\right) x \left(1 - prevalence\right)\right)$$

$$NPV = \left(Sp\ x \left(1 - prevalence\right)\right) / \left(Sp\ x \left(1 - prevalence\right) + \left(1 - Se\right) x\ prevalence\right)$$

The Box 3.1 PPV formula indicates that PPV always will be low in the instance of a rare disease (low prevalence) because the numerator will be substantially smaller than the denominator. The Box 3.1 NPV formula indicates that NPV always will be high in the instance of a rare disease because the numerator and denominator will be nearly the same. Those statements are true because the quantity (*Se x prevalence*) will be close to zero when prevalence is low. Table 3.4 presents, using the BCSC sensitivity (82%) and specificity (91%), values of PPV and NPV for a range of prevalence values. The annual prevalence in the BCSC cohort is approximately 500 per 100,000 women. In Table 3.4, notice that PPV increases as prevalence increases, but it takes an implausible prevalence, 100 times that of the prevalence observed in the BCSC cohort (50,000 per 100,000 women), for PPV to rise to 90%. A prevalence of 50,000 per 100,000 women means that every other woman has breast cancer, something that is far from true for any cancer.

Table 3.4 PPV and NPV by prevalence of disease (sensitivity of 82% and specificity of 91%)

Prevalence	PPV	NPV
250 per 100,000	2.2%	>99%
500 per 100,000	4.3%	>99%
1000 per 100,000	8.2%	>99%
50,000 per 100,00	90.8%	>99%

Data are fictional

Table 3.5 PPV as a function of sensitivity and specificity (disease prevalence of 500 per 100,000)

	Sensitivity		
	90%	95%	99%
Specificity			
90%	4.3%	4.6%	4.7%
95%	8.3%	8.7%	9.0%
99%	31.1%	32.3%	33.2%

Data are fictional

Those who are new to assessment of cancer screening often are amazed that PPV is so low for cancer screening tests even when sensitivity and specificity are high. Table 3.5 demonstrates, for a typical cancer prevalence of 500 per 100,000, how changes in sensitivity and specificity affect PPV. Notice that even at a sensitivity and specificity of 99%, values that are yet to be achieved for cancer screening modalities, PPV is only 33%. The data in Table 3.5 demonstrate that it is virtually impossible for PPV to rise above 10% given typical prevalence, sensitivity, and specificity associated with today's cancer screening tests.

3.2.3 The Implications of Low PPV

A low PPV indicates that most positive cancer screening tests are false alarms. A PPV of 4% means that 96% of positive tests do not lead to a cancer diagnosis. In the BCSC data (Table 3.2), there are 7000 true positives but 165,000 false positives. There is disagreement as to whether false positives should be classified as a harm of cancer screening. One point of view is that any test, including the diagnostic evaluation tests that accompany a false positive, is a test worth having if it rules out cancer. The other point of view is that false positives are a harm of cancer screening as they cause patients to worry unnecessarily and to receive unneeded medical tests and procedures, some of which can be risky.

3.2.4 Can PPV Be Improved?

As was demonstrated in Box 3.1, PPV depends on three quantities: prevalence, sensitivity, and specificity. Disease prevalence is, for all intents and purposes, not modifiable (and definitely not in the short term), and while we do have some control over sensitivity and specificity, their upper bounds are determined, realistically, by the abilities of the cancer screening tests. So PPV will remain low. And cancer screening will continue to generate many more false than true positive tests.

Recall that the intent of cancer screening is not to diagnose; rather it is to identify individuals who need additional medical attention to determine if they have cancer or to rule that out. A cancer screening test with a value of 100% for sensitivity, specificity, PPV, and NPV would be possible if a cancer screening test had perfect discriminatory ability, which is contrary to the goal of cancer screening. We could guarantee 100% sensitivity by assigning a positive test result to every screening test, but in that instance, PPV will still be low: it will equal the prevalence of the cancer. We could guarantee 100% specificity by assigning a negative test result to every cancer screening test, but in that instance, no cancers would be screen detected.

3.3 ROC Curves and AUC

An ROC curve demonstrates the trade-off between detecting more cancers and increasing the FPR. The curve is formed by graphing the sensitivity and FPR for different definitions of positivity. Usually an established screening cohort with information on cancer diagnoses and specifics of what was observed on the cancer screening test (rather than only a positive/negative test result classification) is used and scenarios are created. Prostate cancer screening provides a straightforward example. A PSA of 4 ng/mL or greater is the usual definition of a positive prostate cancer

screening test in the US, but what would have happened if the cut-off was 3 ng/mL or 5 ng/mL, say? How many additional cancers would be detected with the lower cut-off, and how many additional cancers would be missed with the higher cut-off? The FPR would increase with the lower cut-off and decrease with the higher cut-off, but by how much?

ROC curves provide useful comparisons, though it is necessary to make assumptions when using the scenarios. We must assume that the experience that follows the cancer screening test is the same regardless of the positivity definition employed. For example, we must assume that any cancer diagnosed through cancer screening ultimately would be symptom-detected (no overdiagnosis), and we also must assume that the intensity of diagnostic evaluation is the same regardless of the positivity definition employed. An ROC curve is built by selecting a finite number of positivity definitions, graphing the sensitivities and FPRs that would have resulted from those positivity definitions, and connecting the dots either in a linear fashion or by way of smoothing.

Examples of cancer screening test ROC curves can be found in the biomedical literature [2–4]. For illustrative purposes, the BCSC data presented in Table 3.2 were used to lay the foundation for a fictional ROC curve.

3.3.1 ROC Curves

Figure 3.1 presents our fictional ROC curve. Sensitivity is plotted along the y-axis and the FPR is plotted along the x-axis. The ROC curve rises steeply as sensitivity moves away from zero, indicating a large gain in sensitivity with only small increases in FPR. All ROC curves have a turning point, a point at which the incremental ability to improve sensitivity becomes increasingly more expensive in terms of FPR.

All ROC curves include the points [0,0] and [1,1]; it is the path the curve takes from [0,0] to [1,1] that varies. [0,0] represents the unrealistic situation in which all test are negative, which results in a sensitivity of zero and an FPR of 0. [1,1] represents the

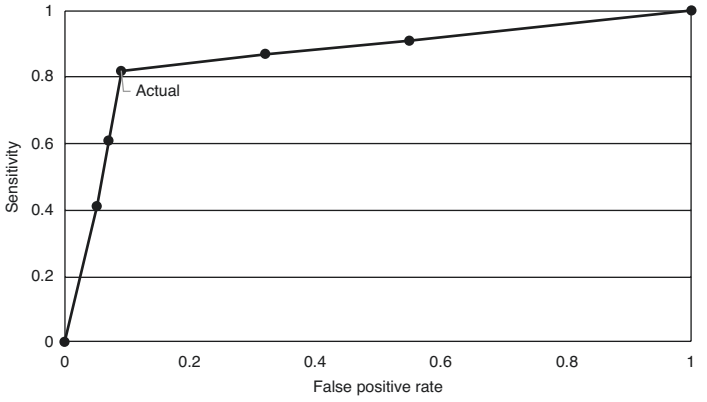

Fig. 3.1 ROC curve

unrealistic situation in which all tests are positive, which results in a sensitivity of 1 and an FPR of 1.

ROC curves can be created for cancer screening tests that return continuous measures, such as PSA, by selecting and varying the value that defines positivity. They also can be used for tests that return categorical classifications, such as the BI-RADS classification for breast abnormalities [5], by collapsing the categories into only two: positive and negative. Let's say that a cancer screening test returns a value of 1, 2, 3, 4, or 5. To create the ROC curve, a positive test result could be defined as a value of 2 or greater, a value of 3 or greater, or a value of 4 or greater. Sensitivity and FPR would then be calculated for each of the three scenarios to create the ROC curve.

The ROC curve (Fig. 3.1) was created using a small number of data points for ease of calculation and presentation. The points were developed by modifying the BCSC values of sensitivity (82%) and FPR (9%) (Table 3.3): the number of false positives and false negatives were varied by percentages rather than examining test findings and reclassifying according to new positivity definitions. The actual point and the derived points are presented in Table 3.6.

Table 3.6 Values of sensitivity and FPR used to calculate the ROC curve in Fig. 3.1

Sensitivity	FPR	Veracity
0.41	0.05	Fictional
0.62	0.07	Fictional
0.82	0.09	Actual
0.87	0.32	Fictional
0.91	0.55	Fictional

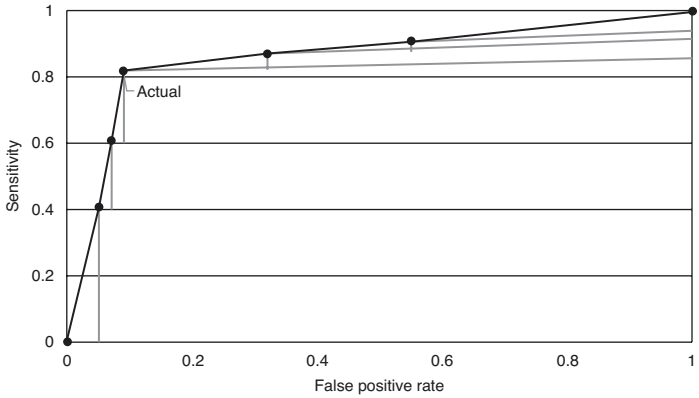

Fig. 3.2 Calculating AUC by partitioning the space under the ROC curve into rectangles and triangles

3.3.2 Calculating AUC

ROC curves can be summarized and compared by calculating the area underneath them. That area, the AUC, is circumscribed by the curve itself, the x-axis, and a right sided y-axis, and can be calculated using simple formulas for area or, if desired, integral calculus. The AUC for the ROC curve in Fig. 3.1 is 0.87 and was calculated by dividing the area into 5 rectangles and 6 triangles and summing those areas (Fig. 3.2). Many ROC curves presented in the literature are smoothed, however. Smoothing involves advanced mathematics, which is beyond the scope of this primer. Smoothing an ROC curve should change the AUC only slightly.

AUC ranges from 0.5 to 1.0. An AUC of 0.5 represents a cancer screening test with no discriminatory ability, meaning that the result does not depend on whether cancer is present. The cancer screening test is, in effect, no better than flipping a (fair) coin to assign the result. An AUC of 1.0 indicates perfect discriminatory ability: the point [1,0] defines the curve. In that instance, sensitivity is 1 and the FPR is 0. The points [0,0] and [1,1] are not viable scenarios in cancer screening, but they create standard anchors for the curve so that AUCs can be calculated and compared.

3.4 Performance Measures: Evidence or Not?

Performance measures are useful for describing the discriminatory ability of cancer screening tests and for comparing one cancer screening test to another. But they measure the ability of cancer screening to lead to detection of cancer, not the ability of cancer screening to reduce cause-specific mortality. Chapter 5 explains that improvement in cancer detection does not guarantee a reduction in cause-specific mortality.

Performance measures are rarely considered sufficient evidence to implement cancer screening for the first time. However, a new cancer screening test, one that is similar to an established test known to reduce cause-specific mortality, often disseminates into practice if its performance measures are superior to those of the older test. Examples include the change from film mammography to digital mammography (breast cancer) [3] and the change from guaiac FOBT to immunochemical FOBT, also known as FIT (colorectal cancer) [6]. Adoption of new cancer screening tests based on comparison of performance measures with that of past tests is discussed in more detail in Chap. 8.

References

1. BCSC data explorer [Internet]. Seattle: Breast Cancer Surveillance Consortium; 2011 – [cited 2019 Oct 20]. Available from: http://tools.bcsc-scc.org/dataexplorer/

2. Thompson IM, Ankerst DP, Chi C, Lucia MS, Goodman PJ, Crowley JJ, Parnes HL, Coltman CA. Operating characteristics of prostate-specific antigen in men with an initial PSA level of 3.0 ng/mL or lower. JAMA. 2005;294(1):66–70.
3. Pisano ED, Gatsonis C, Hendrick E, Yaffe M, Baum JK, Acharyya S, Conant EF, Fajardo LF, Bassett L, D'Orsi D, Jong R, Rebner M, for the Digital Mammographic Imaging Screening Trial (DMIST) Investigators Group. Diagnostic performance of digital versus film mammography for breast-cancer screening. New Engl J Med. 2005;353(17):1773–83.
4. Pinsky PF, Gierada DS, Nath H, Kazerooni EA, Amarosa J. ROC curves for low-dose CT in the national lung screening trial. J Med Screen. 2013;20(3):165–8.
5. ACR BI-RADS atlas 5th ed. [Internet]. Reston: American College of Radiology; 2013 – [cited 2019 Oct 20]. Available from: https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-Rads
6. Allison JE, Sakoda LC, Levin TR, Tucker JP, Tekawa IS, Cuff T, Pauly MP, Shlager L, Palitz AM, Zhao WK, Schwartz JS, Ransohoff DF, Selby JV. Screening for colorectal neoplasms with new fecal occult blood tests: update on performance characteristics. J Natl Cancer Inst. 2007;99(19):1462–70.