CHAPTER FOUR

EPIDEMIOLOGIC TESTS

Epidemiology is the study of disease and its causative factors. Most commonly it involves studying a particular population of patients to determine the frequency of a disease and how it affects that population. Epidemiology also involves the assessment of various diagnostic tests and their clinical utility in evaluating and treating disease.

INCIDENCE, PREVALENCE, AND MORTALITY RATES

Epidemiologists have specific terms for disease occurrence which are commonly misused. The **incidence** of a disease is defined as the number of new cases of the disease per unit time divided by the population at risk for the disease at the beginning of the time period. The large size of some populations of interest can make accurate incidence measurements difficult and costly to obtain. The **prevalence** of a disease is defined as the number of individuals with the disease divided by the population at risk for the disease at a *specific point in time*.

$$\textbf{Incidence} = \frac{\textbf{number of new cases of disease over time}}{\textbf{number of patients at risk for disease at the beginning of the time period}}$$

$$\textbf{Prevalence} = \frac{\textbf{number of patients with the disease at a specific point in time}}{\textbf{number of patients at risk for disease at a specific point in time}}$$

Mortality rates are another aspect of epidemiology in which specific definitions are used. Mortality rates quantitate the incidence of death due to various causes in a particular population of interest and provide a standardized method by which to compare the frequency of death in different patient populations. The **crude annual mortality rate** is defined as the total number of deaths in a population at risk per year divided by the size of the population at risk at mid-year. The population size is measured at mid-year to establish an average size for the population as some patients will die early in the year while others will die late in the year. The crude annual mortality rate is used to measure mortality from all causes over the period of a year. A more specific rate is obtained from the **cause-specific annual mortality rate** in which a particular cause of death is of interest. It is defined as the number of deaths due to the cause of interest per year divided by the size of the population at risk measured at mid-year. Occasionally, we wish to investigate the death rate due to a particular cause by age; the **age-specific annual mortality rate** is defined as the number of deaths in a given age group at risk per year divided by the size of the population at risk measured at mid-year.

$$\textbf{Crude annual mortality rate} = \frac{\textbf{total deaths in population at risk per year}}{\textbf{total population at risk at mid-year}}$$

$$\textbf{Cause-specific annual mortality rate} = \frac{\textbf{total deaths from cause of interest per year}}{\textbf{total population at risk at mid-year}}$$

$$\textbf{Age-specific annual mortality rate} = \frac{\textbf{total deaths in a given age group per year}}{\textbf{total population at risk at mid-year}}$$

SURVIVAL ANALYSIS

Although mortality rates allow us to characterize the occurrence of death in a particular population per year, they do not provide us with information on the natural progression of the disease process. To obtain such information, special methods of analysis are required and are known by the terms **survival analysis**, **actuarial analysis,** or **life-table analysis**. These methods are used to follow a specific group of patients in order to determine the effect of time on the natural progression of the disease process of interest.

Survival analysis is based on probability theory. Recall from Chapter One that the probability of two independent events occurring together is given by the product of their respective probabilities. The probability that a patient survives (P_S) for two years is therefore given by the probability of surviving the first year (P_1)

multiplied by the probability of surviving the second year (P_2). The cumulative probability of a patient surviving for five years is given by:

$$P_S = P_1 \text{ x } P_2 \text{ x } P_3 \text{ x } P_4 \text{ x } P_5$$

Actuarial or **life-table analysis** and the **Kaplan-Meier product limit method** are common statistical methods for evaluating survival data. They each result in graphs which illustrate the survival rates of the patient population at various points in time**. Actuarial analysis** calculates the exact survival rate during specific time periods (see Chapter Twelve for an example). The actuarial method makes two assumptions: 1) that all deaths and "withdrawals" (patients lost to follow-up) occur on average at the midpoint of each time period, and 2) that the probability of survival for each time period is independent. Thus, the actuarial method is subject to significant bias if the occurrence of death does not occur evenly in the population or if a large number of patients are lost such that their outcome (death or survival) is not known.

The **Kaplan-Meier product limit method** does not look at survival rates during specific time periods, but rather estimates survival and recalculates this estimate each time a patient dies (see Chapter Twelve for an example). It is not subject to the potential bias introduced by withdrawals as is actuarial analysis. Both of these methods are easily performed by most computer statistics packages and are commonly seen in the medical literature. Statistical methods exist for comparing and detecting differences between survival curves. Such tests include the **Wilcoxon rank-sum test**, the **Kruskal-Wallis test,** and the **Mantel-Haenszel test**.

ASSESSING DIAGNOSTIC TESTS

Physicians utilize diagnostic tests each day to help "rule in" or "rule out" disease. We propose a diagnosis and use laboratory tests and procedures to confirm it. In doing so, we begin with a **pre-test likelihood of disease**. We have an impression from the patient's history, physical examination, and previous diagnostic tests of whether or not our proposed diagnosis is correct. We then perform the diagnostic test and modify our impressions based on the test results creating the **post-test likelihood of disease**. From the post-test likelihood we may either 1) confirm our diagnosis, 2) reject our diagnosis, or most commonly 3) strengthen our impression of the probability of disease, but not enough to either confirm or reject our hypothesized diagnosis.

Whether a diagnostic test is clinically useful in predicting disease (and therefore confirming our diagnosis) is dependent upon the number of patients it correctly identifies as having disease (**true positives** or **TP**), the number it correctly identifies as not having disease (**true negatives** or **TN**), the number it falsely identifies as having disease (**false positives** or **FP**) and the number it falsely identifies as not having disease (**false negatives** or **FN**). Based on these four outcomes, we can create a **2 x 2 contingency table** for any diagnostic test which summarizes its ability to accurately predict the presence of disease.

	Disease	**No Disease**
Test Positive	**True Positive**	**False Negative**
Test Negative	**False Positive**	**True Negative**

Figure 4-1: 2 x 2 contingency table

The clinical utility of a diagnostic test is most commonly quantitated using measurements such as **sensitivity, specificity, positive predictive value, negative predictive value, and accuracy**. The **sensitivity** of a test is defined as the proportion of diseased patients (true positives + false negatives) which the test correctly classifies as having the disease (true positives). Sensitivity can also be described as the "**true positive fraction**." A test with high sensitivity is important when we do not want to risk missing a disease if it is present. Such a test is therefore useful in screening for the presence of disease. Mathematically, the sum of a test's sensitivity (true positive fraction) and the proportion of patients with disease that it misses (the **false negative fraction**) must equal 1; thus, as the sensitivity of a test increases, the number of patients it misses (false negatives) must decrease.

$$\textbf{Sensitivity} = \textbf{TP/(TP+FN)}$$

The **specificity** of a test is defined as the proportion of non-diseased patients (true negatives + false positives) which the test correctly classifies as not having the disease (true negatives). Specificity can also be described as the "**true negative fraction**." A test with high specificity is important when false positive results would result in harm to patients physically or emotionally or require them to undergo unnecessary procedures or treatments. The specificity of a test is a measure of its ability to confirm the presence of disease.

Mathematically, the sum of a test's specificity (true negative fraction) and the proportion of patients that it falsely identifies as having the disease (the **false positive fraction**) must also equal 1; as the specificity of the test increases, the number of false positives must decrease.

Specificity = TN/(TN+FP)

The **positive predictive value** of a test is the proportion of positive tests (true positives + false positives) that correctly identify a patient with disease (true positives). The **negative predictive value** of a test is the proportion of negative tests (true negatives + false negatives) that correctly identify a patient without disease (true negatives).

Positive Predictive Value = TP/(TP+FP)

Negative Predictive Value = TN/(TN+FN)

The **accuracy** of a diagnostic test is the proportion of all test results, both positive and negative, which correctly identify the patient's disease status.

Accuracy = (TP+TN)/(TP+TN+FP+FN)

If we were to design the perfect diagnostic test, it would have a sensitivity of 1.0 (no false negatives) and a specificity of 1.0 (no false positives). In reality, few if any diagnostic tests possess these characteristics. As we have seen, the characteristics of a test will vary depending on how we define disease. If our disease of interest is a dichotomous or binary variable (i.e., live vs die, extubated vs reintubated), the presence of disease is fairly obvious. The vast majority of disease processes are not that straightforward however. In defining the presence of hypoxia, for example, should it be defined as an arterial oxygen tension (PaO_2) of < 50 torr?, < 60 torr?, < 70 torr? Depending on the **decision threshold** or "**cut-off point**" which we use to define it, the prevalence of our disease (in this case "hypoxia") will change significantly. If we are more stringent in our definition of hypoxia (a PaO_2 of < 50 torr for example), the sensitivity of our test will be high (there will be few false negatives as most patients with a PaO_2 < 50 torr will be hypoxic), but the specificity will be low (there will be many false positives as many patients with PaO_2's > 50 torr will also be hypoxic). Conversely, if we are less stringent in our definition of hypoxia (a PaO_2 of < 70 torr for example), the sensitivity of our test will be low (many patients are asymptomatic despite PaO_2's of < 70 torr), but the specificity will be high (hypoxic patients are unlikely to have a PaO_2 > 70 torr). The way in which we define our disease of interest can therefore significantly affect the characteristics of our test. In an attempt to improve our test's sensitivity, we must frequently accept a decrease in specificity, or vice versa. We must decide, based on the nature of the disease, which is worse: an increased false positive rate or an increased false negative rate and choose our decision threshold accordingly.

The prevalence of the disease can also affect the characteristics of a test. If the disease of interest is rare, unless the test is very accurate in predicting disease, the test will have a high false positive rate. For example, a test with a sensitivity of 0.95 (false negative rate of 5%) and specificity of 0.95 (false positive rate of 5%) will be a very good test if the disease prevalence is high, but will be a poor test if the prevalence is low and the disease is rare. If, for example, the disease prevalence is only 1% of the population, out of 10,000 patients only 100 will truly have the disease. However, 590 of the 9,900 patients without the disease will have a positive test (9,900 x false positive rate of 5%) and will be incorrectly predicted to have the disease. With regard to this test, therefore, despite a high sensitivity and high specificity, the overall accuracy of the test in correctly identifying the disease is very low (only 1%).

	Disease	No Disease	
Test Positive	95	495	590
Test Negative	5	9405	9410
	100	9900	10000

Sensitivity = 95/(95+5) = 0.95
Specificity = 9405/(9405+495) = 0.95
Positive Predictive Value = 95/(95+495) = 0.16
Negative Predictive Value = 9405/(9405+5) = 0.99
Accuracy =(95+5)/(95+5+495+9405) = 0.01

This is an extreme case, but it illustrates one of the potential problems with using sensitivity and specificity to assess test utility. Sensitivity, specificity, and accuracy are the "gold standard" methods for evaluating diagnostic tests in the medical literature. The effect of decision thresholds and disease prevalence, however, should always be kept in mind when interpreting data analyzed with these tests.

RECEIVER OPERATING CHARACTERISTIC (ROC) CURVES

A statistical method which avoids the problems associated with the use of sensitivity and specificity is **receiver operating characteristic (ROC) curve analysis**. It is being used more and more frequently in the medical literature. As we have seen, in order to improve a test's sensitivity we must frequently accept a decrease in specificity due to the way in which we define disease. One way to demonstrate this relationship is to construct the ROC curve for the test. This curve plots sensitivity (the true positive fraction) against 1 - specificity (the false positive fraction).

To create an ROC curve, we must first determine all possible decision thresholds for the test and calculate the sensitivity and specificity of the test at each point. After plotting the sensitivity and 1 - specificity for each decision threshold, we can then choose the sensitivity that maximizes the specificity and identify the decision threshold for that point. Once the curve is plotted, we can also calculate the area under the ROC curve and use that as a measure of the test's usefulness. Since the "perfect" diagnostic test has a sensitivity of 1.0 and a specificity of 1.0, the "perfect" ROC curve has an area under the curve of 1.0. In comparing two diagnostic tests, the test with the largest area under the ROC curve will have the fewest false positives and false negatives. By comparing the area under each test's ROC curve and determining whether they are statistically different, one diagnostic test can be compared with another irrespective of the decision thresholds utilized for each test.

As an example, consider a study in which 98 patients underwent weaning from mechanical ventilation using negative inspiratory force (NIF) measurements as the determing factor for extubation. In this study, 81 patients remained extubated, while 17 required reintubation. We will define our "disease" as successful extubation from mechanical ventilation. To begin an ROC analysis, we must first choose a number of decision thresholds which are clinically of value. NIF is usually measured from 0 to 65 cm H_2O with an NIF of > 30 cm H_2O traditionally being used to identify patients who are ready for extubation. Using a range of decision thresholds (X) between 0 and 65 cm H_2O, therefore, each patient can be classified as having had one of four possible outcomes:

1) The patient was successfully extubated and had an NIF of X at extubation
2) The patient was subsequently reintubated and had an NIF of less than X
3) The patient was subsequently reintubated and had an NIF of X
4) The patient was successfully extubated and had an NIF of greater than X

For each decision threshold, we evaluate each of the 98 patients to determine which of the four possible outcomes they fall into. A table of our data might appear like that in Figure 4-2.

NEGATIVE INSPIRATORY FORCE (cm H_20)

Decision Threshold (X)	0	5	10	15	20	25	**30**	35	40	45	50	55	60	65
Extubated at X	0	1	0	1	3	13	**19**	11	11	3	5	3	11	0
Reintubated at < X	0	0	0	0	1	2	**4**	9	9	14	14	16	16	17
Reintubated at X	0	0	0	1	1	2	**5**	0	5	0	2	0	1	0
Extubated at > X	81	81	80	80	79	76	**63**	44	33	22	19	14	11	0

Figure 4-2: ROC analysis: extubation outcome stratified by decision threshold

For an NIF decision threshold of 30 cm H_2O (the traditional threshold or “cut-off” value for NIF), for example, 19 patients were successfully extubated with an NIF of 30 and 63 patients were successfully extubated with an NIF of > 30. Meanwhile, 5 patients were initially extubated with an NIF of 30, but required reintubation, and 4 patients were initially extubated with a NIF of < 30, but also required reintubation.

Having established the number of patients for each outcome at the various decision thresholds, we can then calculate the sensitivity (true positive fraction) and 1-specificity (false positive fraction) of the test at each decision threshold. Since our “disease” is successful extubation, we calculate sensitivity and specificity as follows:

Sensitivity = TP/(TP + FN)
= Extubated at > X / All extubated
= Extubated at > X / 81

Specificity = TN/(TN + FP)
= Reintubated at X / All reintubated
= Reintubated at X / 17

This results in the following table of sensitivity, specificity, and 1 - specificity for each of the decision thresholds from 0 to 65 cm H_2O:

NEGATIVE INSPIRATORY FORCE (cm H_20)

Decision Threshold (X)	0	5	10	15	20	25	30	35	40	45	50	55	60	65
Sensitivity	0	0	0	0	.06	.11	.24	.53	.53	.82	.82	.94	.94	1
Specificity	0	0	.99	.99	.98	.94	.78	.54	.41	.27	.23	.17	.14	0
1 - specificity	0	0	.01	.01	.02	.06	.22	.46	.59	.73	.77	.83	.86	1

Figure 4-3: ROC analysis: Sensitivity, specificity and 1 - specificity at each decision threshold

Using these numbers, the sensitivity (true positive fraction) is then plotted against 1 - specificity (false positive fraction) to generate the ROC curve for the test.

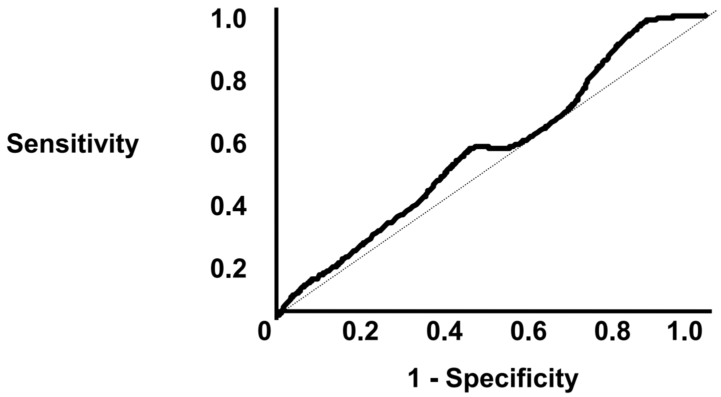

Figure 4-4: Receiver operating characteristic curve for NIF in predicting successful extubation

A useful diagnostic test is generally considered to be one with an area under the ROC curve above 0.50 (noted by the dashed line on Figure 4-4). The ROC curve for NIF is above this line connecting 0,0 with 1,1 and thus has an area above 0.50. The actual area under the NIF curve is 0.54 (see references 4-7 for details on calculating the area under ROC curves). We can also deduce from the curve that for this study the NIF which maximizes sensitivity and specificity is 35 cm H_2O and not the traditional value of 30 cm H_2O. The area under the NIF curve can be compared to that of other diagnostic tests to determine which test is a better predictor of successful extubation. Such calculations, although involved, are readily performed and the reader is referred to the articles by Hanley (1983) and Beck (1986) for further details and examples.

SUGGESTED READING

1. Sox HC. Probability theory in the use of diagnostic tests. Ann Intern Med 1986;104:60-66.
2. Doubilet PM. Statistical techniques for medical decision making: applications to diagnostic radiology. AJR 1988;150:745-750.
3. Pauker SG, Kassirer JP. Decision Analysis. NEJM 1987;316:250-258.
4. Metz CE. Basic principles of ROC analysis. Semin Nuc Med 1978;8:283-298.
5. Beck JR, Shultz EK. The use of relative operating characteristic (ROC) curves in test performance evaluation. Arch Pathol Lab Med 1986;110:13-20.
6. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves dervied from the same cases. Radiology 1983;148:839-843.
7. Hunink MG, Richardson DK, Doubilet PM, Begg CB. Testing for fetal pulmonary maturity: ROC analysis involving covariates, verification bias, and combination testing. Med Decis Making 1990;10:201-211.
8. Griner PF, Mayewski RJ, Mushlin AI, Greenland P. Selection and interpretation of diagnostic tests and procedures. Ann Intern Med 1981;94(Part 2):553-600.
9. Fletcher RH, Fletcher SW, Wagner EH. Clinical epidemiology - the essentials. Baltimore: Williams and Wilkins, 1982: 41-58.
10. Wassertheil-Smoller S. Biostatistics and epidemiology: a primer for health professionals. New York: Springer-Verlag, 1990:60-74.
11. Dawson-Saunders B, Trapp RG. Basic and clinical biostatistics (2nd Ed). Norwalk, CT: Appleton and Lange, 1994:188-209, 232-248.
12. O’Brien PC, Shampo MA. Statistics for clinicians: 11. Survivorship studies. Mayo Clin Proc 1981; 56:709-11.