Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Determining the Reliability and Validity and Interpretation of a Measure in the Study Populations

There are many reasons for using a molecular test in an epidemiologic study. The test might be used to confirm or exclude disease, assess disease severity, or identify the precise location of disease. It might characterize how the host responds to disease or how microbes interact with each other or with the host. The test might be used to distinguish between disease types, identify microbes, or differentiate among microbes of the same species. The test might detect a known or suspected marker of disease prognosis, a known or suspected marker of exposure that modifies disease risk, or one that is known or suspected to interact with a defined exposure disease relationship. Alternatively, the measure might detect a marker in search of an association. New technologies have identified genetic, transcript, and protein variants whose association with disease is unknown. Regardless of the reason for measurement, a test is only useful if it is reliable and valid, and interpreted appropriately. The validity and reliability of a test result depend on everything from whether the specimen was collected correctly to whether the results were recorded accurately. Once a test is selected and the validity and reliability determined in the hands of the research team, the continued validity and reliability is assured by quality control and quality assurance procedures (discussed in Chapter 10).

The extent that a test result reflects the true value, that is, it is valid, depends on minimizing two major classes of error: systematic error (also known as bias) and random error (Figure 8.1). Systematic error is an error that occurs in one direction, for example, a scale that shows the weight as two grams higher than the true value. Though a systematic error can be corrected post hoc if detected, it is best avoided by regularly calibrating instruments. Depending on the direction of the bias, a systematic error can lead to the overestimation or underestimation of the frequency of exposure or disease. However, the primary concern is if the systematic error occurs differentially between cases and controls or exposed and unexposed individuals. This difference can lead to an erroneous association. For example, if values are always higher for cases, it will appear that there is an association with being a case, even if the effect is due solely to systematic error in measurement. A well-designed and well-implemented study protocol will avoid systematic errors by (1) setting inclusionary and exclusionary criteria that result in unbiased selection and follow-up of study participants; (2) making collection, storage, and processing of specimens from all participants as similar as possible; and

FIGURE 8.1 Random and systematic errors (bias). Source: Adapted, with permission, from Atkins1 (2000).
arranging laboratory procedures so that any effects of storage or testing equally impact specimens from cases and controls or exposed and unexposed participants.

The reliability of a test, that is, the extent that the same test yields the same value or very close value on repeated testing, is affected by the extent of random error. Random error increases variation and decreases the ability to detect a difference between groups – if a difference exists. Random errors can be minimized, but not avoided entirely, by minimizing technical variability in procedures, using uniform reagents and instruments, and training and periodically retraining personnel in the study protocol. Ideally, each specimen will be collected, handled, processed, and tested in exactly the same way from each study participant. If we could attain this ideal, all observed differences between participants would be attributable to true biological variation – given a test with perfect reliability and validity. After ensuring that the protocol does not introduce systematic error, our goal is to develop a study protocol and laboratory procedures that minimize random error.

This chapter describes the steps towards developing an optimal study protocol for a valid and reliable test result, and identifying any issues of interpretation of the selected measure for the study population (Table 8.1).

### 8.1 IDENTIFY ALL DATA HANDLING AND PROCESSING STEPS, FROM SPECIMEN COLLECTION TO RECORDING DATA IN A DATABASE

Identifying all data handling and processing steps is essential for establishing a study protocol that minimizes error. This listing is also required for establishing quality control and quality assurance procedures intended to minimize errors throughout the conduct of the study (see Chapter 10). The listing should be as exhaustive as possible (see Table 8.2 for an example from a study of group B Streptococcus).

The amount of observed variation for a specific test depends on how sensitive the test results are to all the steps up to and through laboratory testing. Variation also is affected by the type of specimen, the quality of the specimen, the test itself, and how the results are to be interpreted. Results from tests of liquid specimens can be quantitative, because the test can be referred back to a sample of known volume. The amount of bacteria isolated from urine is reported as the number of colony-forming units per milliliter urine. Tests of solid specimens may give only qualitative results. The amount of bacteria isolated from a throat swab is reported as heavy, moderate, little, or no growth. It is possible to process solid specimens or that collected on swabs by dissolving, or grinding, to test a known amount. Depending on processing, results may be reported qualitatively or semiquantitatively. Although many tests give results per volume, others give values relative to some standard. For example, we can use real-time polymerase chain reaction (PCR) to estimate the total bacterial load and then the proportion of that load due to a specific species, enabling semiquantitation even if the initial specimen is solid or a secretion collected on a swab. Even if the results are quantitative the interpretation may not be. The level of discrimination detected by the measuring tool may

### TABLE 8.1 Determining an Optimal Study Protocol

- Identify all data handling and processing steps, from specimen collection to recording data in a database
- Assess the potential for error at each step, and the error tolerance
- Determine the reliability of the selected measure across a range of values
- Determine the validity of the selected measure
- Determine the intralaboratory and interlaboratory reliability
- Determine the appropriate interpretation of measurement
not reflect a true biological difference. There may be a wide range of values corresponding to disease or exposure so that the actual interpretation is dichotomous (diseased/not or exposed/not) or at best categorical (not diseased, possibly diseased, probably diseased, definitely diseased).

When a specimen is collected can influence study results. Many biomarkers reflect circadian rhythms; if the daily fluctuation within an individual is as large as that between individuals it is difficult to draw any inferences. One strategy is to standardize time of specimen collection. Urine varies in concentration depending upon time of day and amount of fluid the individual consumes. If the investigator wishes to measure concentration of a specific biomarker, participants might be directed to collect a first morning void, and to either record the amounts of liquid consumed or drink a specified amount of liquid the day before, which will enable the investigator to adjust for liquid proportional to the individual’s size. Alternatively, participants might be directed to collect all urine voided in a 24-hour period. The assay results can be further normalized to a metabolite that is excreted at a known rate.

How a specimen is collected can also influence study results. Bacteria grow in specific niches on the human host, so variation in specimen collection may change the probability of detection or of contamination from other sites. Bacteria grow around the urethral opening; in women, vaginal fluid containing bacteria may contaminate the urine specimen. If the investigator wishes to measure the concentration of bacteria in the bladder, avoiding contamination from the urethra or vaginal discharge, he or she may consider collecting urine directly from the bladder using a needle through the abdominal wall into the bladder. This may not be necessary. If the investigator can tolerate low levels of contamination, he or she might consider

| Procedure                                                                 | Potential for Error                                                                 |
|---------------------------------------------------------------------------|------------------------------------------------------------------------------------|
| Self-collection of rectal specimen using a swab, placed into transport media | Specimen does not contain fecal matter                                             |
|                                                                           | Swab placed in wrong vial                                                           |
|                                                                           | Swab/vial not labeled or improperly labeled                                        |
|                                                                           | Specimen not collected                                                              |
|                                                                           | Transport media improperly prepared or outdated                                      |
| Transport of specimen to laboratory                                        | Specimen lost during transport to laboratory                                        |
|                                                                           | Labeling lost during transport                                                      |
|                                                                           | Specimen heated or cooled during transport                                          |
|                                                                           | Delays or inconsistent transport time                                               |
| Culture specimen for GBS                                                  | Error in labeling                                                                   |
|                                                                           | Break in sterile technique                                                          |
|                                                                           | Culture too long/short; incorrect medium, temperature, or other conditions         |
| Identification of GBS                                                    | Error in identification                                                             |
|                                                                           | GBS isolate grown on plate, but not selected for further testing (identification)   |
|                                                                           | Incorrect recording of results                                                       |
|                                                                           | Break in sterile technique                                                          |
| Storage of isolates                                                      | Error in labeling                                                                   |
|                                                                           | Storage media not properly prepared                                                 |
|                                                                           | Incorrect storage media                                                             |
|                                                                           | Break in sterile technique                                                          |
|                                                                           | Storage at incorrect temperature or change in temperature that impacts isolate integrity |
|                                                                           | Storage location not recorded or recorded incorrectly                              |
collecting a clean-catch midstream urine specimen – urine is collected after cleaning the periurethral area, and urinating a small amount to minimize urethral bacteria – and/or asking women to insert a vaginal tampon before voiding to minimize vaginal discharge.

Once a specimen is collected, it will go through a variety of processing steps, depending on the study. Each step should be tracked to minimize specimen loss and to ensure that each specimen is processed appropriately. To continue with the group B Streptococcus example, the study should keep a record of (1) all individuals screened for participation and reason for refusal, (2) the completion of each participant of the enrollment forms (consent form, questionnaires), (3) the collection of specimens from each participant enrolled, (4) some assessment of the quality of the specimens and if no specimen was collected, (5) a listing of all materials sent to the laboratory for processing each day, (6) a listing of all materials received at the laboratory each day, (7) the processing of each specimen upon receipt, and (8) where specimens are stored.

Tracking is best done using an electronic database; but databases have their own pitfalls. The data can be lost if not backed up regularly. Data entry is not without error; numbers can be mistyped or entered into the wrong cells. I strongly recommend using a database rather than a spreadsheet to record data, as it is particularly easy to enter data in the wrong cell or to sort only a single column of a spreadsheet, resulting in errors that are very difficult to rectify. Using a barcoding system can be extremely useful, because it eliminates the need to type in a code and associated errors. However, barcodes are not foolproof, as the code can be linked to the wrong data, or the links between the barcode and the specimen identification number can be lost. Personnel must become accustomed to scanning each tube or rack before processing. Further, if the codes cannot be imported easily into the databases associated with high-throughput equipment used to test the specimen, the value is minimal.

**8.2 ASSESS THE POTENTIAL FOR ERROR AT EACH STEP, AND THE ERROR TOLERANCE**

Each step in specimen collection and processing is subject to error. Specimens can be collected by study personnel, medical personnel working with the study, or by the participant. Study personnel are under the investigators’ direct supervision; they can be trained and retrained periodically to ensure protocols are followed. Further, study personnel likely will be collecting specimens frequently. By contrast, medical personnel working with the study may be too busy to learn the study protocol well, and are likely to collect specimens from study participants relatively infrequently. Thus, making step-by-step instructions readily available is essential. Self-collected specimens depend on the extent that the study participant understands the directions; quality of self-collected specimens can be excellent if the protocol includes good training for participants and easy-to-follow instructions. These instructions should be given verbally along with written instructions. If specimens are collected on-site, the instructions might be taped to the wall of the specimen collection room. Regardless of who collects the specimens, having easy-to-follow instructions, color coding or numbering materials to match steps, and providing packets containing all materials and instructions together will reduce errors in specimen collection.

Molecular tests are increasingly sensitive in a laboratory sense, that is, able to detect exquisitely small amounts of material. The increase in sensitivity can translate to increased variability, as tests are able to detect previously undetectable variations among specimens. Whether the increased variation is attributable to technical rather than biological variation should be determined before publishing study results. There are distinct biological niches. For example, the bacteria found in the mouth vary by tooth surface, and bacteria on the skin vary by body site. Regardless of specimen collected, the goal of specimen collection should be to minimize variation due to differences in how, when, and where a specimen is collected, in order to maximize ability to detect true biological variation among study participants.
The investigator should determine both the minimum and optimal amount of a specimen required for a valid test. An optimal volume enables retesting (if required either because of test failure or for validation) and storage for future use. Specimen vials can be marked so that there is an easy visual check that sufficient volume was collected. For swabs, the investigator might institute visual checks that there is material on the swab, or include culture on nonselective media in addition to selective media to ensure adequate sample was collected. With an infectious disease process it is impossible to revisit a participant at the same disease state, so it is critical to get the specimen collected right each time. During the development of the study protocol, the investigator should look for key indicators that a specimen was appropriately collected and in sufficient amounts. These indicators can be recorded and analyzed as part of quality assurance procedures during study conduct.

Not only should the amount of specimen be sufficient, but the quality of the specimen must be appropriate. What constitutes good quality depends on how the specimen will be used. Requirements for a specimen that will be cultured are different than if the specimen will be processed and DNA extracted. For culture, there must be viable cells and minimal contamination; for DNA, there should be sufficient cells of interest present. Further, different assays are more or less tolerant of how the specimen is handled and processed. There are some DNA hybridization assays with high tolerance for presence of proteins in with the DNA, such as colony blots and in situ hybridizations. By contrast, for use in a microarray format, extensive preparation of the DNA may be necessary to remove all other materials. To avoid unnecessary costs and adverse effects on sample size associated with collecting specimens that cannot be used, the investigator should explore the tolerance of the intended assays to the intended specimen handling procedures and adjust either the assay or procedures accordingly.

All specimens must be properly labeled and stored appropriately until tested. Determining the required storage conditions and tolerance for storage is an important component of protocol development. Storage conditions also include the size of the vial and type of label. A small amount of specimen stored in a large vial may result in specimen loss due to evaporation. The label should be able to tolerate the storage conditions, because some labels fall off when a vial is frozen and thawed. A single specimen may be subject to several different assays, and the storage conditions can differ by assay, for example, one test requires storage on ice until testing and another test can be reliably conducted on a specimen stored in transport media for weeks; in this case the investigator might consider splitting the sample. Sometimes the most convenient collection method limits the ability to store the sample unprocessed. In this case it might be possible to do an initial stage of processing that makes storage less problematic.

Error tolerance is a measure of the acceptable level of error. If the error tolerance is set at 1%, it implies that a value of plus or minus 1% of the ideal is acceptable. Error always occurs; one way to minimize the impact of errors is to have inherent redundancies in the system. Invariably, some individuals will enroll in the study but be unable or refuse to give the required specimens or amount of specimen. Though there may be an optimal method of collecting a specimen, a less optimal alternative may be allowed. We conducted a study of urinary tract infection where *Escherichia coli* were isolated from vaginal specimens and urine. Vaginal specimens were self-collected using a tampon. Participants were instructed to insert a tampon before urinating. This had the added advantage of minimizing contamination of the urine specimen with vaginal secretions. Although almost all women were comfortable with this protocol some had never used a tampon and were unwilling to try. We offered self-collection of the vaginal specimen using a swab as an alternative, even though this increased the potential for contamination of the urine with vaginal secretions. Similarly, some women had emptied their bladder before meeting with our study recruiter. If unable to urinate, we offered women something to drink and a place to wait until they were able to urinate, even
though these urine samples would have had less time in the bladder and lower bacterial counts than if collected at a later point. However, increasing our tolerance for reasonable departures from the enrollment study protocol enabled us to maximize specimen collection.

8.3 DETERMINE THE RELIABILITY OF THE SELECTED MEASURE ACROSS A RANGE OF VALUES

Reliability has two components: repeatability, when repeated testing of the same specimen under the same conditions yields the same result; and reproducibility, when repeated testing of the same specimen in different laboratories yields the same result. A highly reliable test will give very similar results on repeated tests; in a statistical sense the variance is small. There is generally a range for which results of a specific test are most reliable: the reliability may vary with the assessed value, particularly at the ranges of detection (very low or very high). Thus the investigator should conduct reliability assessments across a range of values. In a reliability assessment of dot blot hybridization, the variability, although not the interpretation, was greater for \textit{E. coli} that had greater signal intensity upon hybridization. The increased variability was partly attributable to different copy numbers of the gene under detection.\(^2\)

A general strategy for conducting reliability assessments is shown in Table 8.3. Note that the process is iterative: there is ongoing assessment to identify sources of technical variation followed by modification of the protocol until the desired level of reliability is achieved. Reproducibility uses a similar protocol except that a set of standard unknowns is evaluated in different laboratories using the same assay (discussed in Section 8.5).

Reliability is assessed in several ways depending on whether the measure is continuous or categorical. There is no standard for comparison; we are looking for agreement between measures. For continuous variables the metric may be the standard error, or the coefficient of variation (standard deviation/mean), or the intraclass coefficient of reliability. For categorical measures we use measures of agreement, the most common being the kappa statistic, a chance-corrected measure of agreement. Most spreadsheets and software packages enable calculation of these statistics; the formulas can be examined in the software documentation or in standard textbooks of statistics.

For proper interpretation of study results, further reliability assessment is required to determine the variability from repeated samples from the same individual, and variation among individuals. Repeated samples from the same individual will indicate if the measure varies with time of day, menstrual cycle, or consumption of food or liquids; the extent that this impacts the reliability will dictate if the protocol should stipulate timing of specimen collection. The dynamics of colonization of human body sites by microbes are essentially unknown. Currently there are few estimates in the literature of how frequently there is a change in the bacterial strains (or other colonizing microbes) that commonly colonize the human gut, mouth, vaginal cavity, and skin. Also unknown is the average duration of carriage. There are some estimates for group B \textit{Streptococcus}; colonization is very dynamic, with an average duration of carriage of \(\approx14\) weeks among women.\(^3\) This suggests that

### TABLE 8.3 Conducting Reliability Assessments

| Step | Description |
|------|-------------|
| 1.   | Select representative samples from the range of values (high, medium, low, and negative) |
| 2.   | Conduct experiments on \(~10\) replicates of each sample |
| 3.   | Compare results from replicates tested in the same and different experiments |
| 4.   | Compare results from replicates tested on the same and different equipment |
| 5.   | Identify sources of technical variation and modify protocol |
| 6.   | Repeat 2 through 5 until desired level of reliability is achieved (\(-\)coefficient of variation \(<5\%)\) |
assessments of reliability must be done over a short time, and that loss over a 2-week period could as easily reflect true loss as sampling error. Without similar estimates to set the sampling intervals, it is difficult to interpret whether the tests themselves are less reliable than desirable or whether there is biological variation over the testing interval. Further, as tests become increasingly sensitive it is possible that the tests will detect differences due to the test itself: collection with a swab or lavage may inadvertently modify the biota of interest.

**8.4 DETERMINE THE VALIDITY OF THE SELECTED MEASURE**

The term validity is used in multiple ways in epidemiology. The term may apply to the study itself (internal validity), generalizations from the study to other populations (external validity), or to the characteristics of a specific measure. In this chapter, the focus is on determining the validity of a specific measure.

Ideally, the validity or accuracy of a specific measure is determined relative to a gold standard; that is, a measure that reflects the true value of what we wish to measure. Because the true value is often not known, different criteria have been developed to assess whether a measure is valid (Table 8.4). The weakest criterion is content validity; this criterion is generally invoked when developing a new measurement and appears more often in the social science literature than in assessments using laboratory techniques. The assessment of content validity is by professional judgment or consensus of the field. Assessing construct and criterion validity requires conduct of studies relative to some standard. Generally the investigator uses a case–control design, and selects specimens from individuals known to have the outcome of interest and those known not to have the outcome of interest. For construct validity the specimens will be positive or negative for some relevant characteristics of the phenomena, such as another test, and the results of the study will be the correlation or agreement between the two measures. For criterion validity, the specimens will be positive or negative for the phenomena as defined by the current standard (“truth”), and the results will predict the sensitivity, specificity, and predictive value of the new test relative to truth.

Antibiotic resistance is, at this writing, usually assessed using a phenotypic test, that is, a test assessing whether a bacterial species can grow in the presence of an antibiotic. As we move increasingly toward rapid testing using nonculture techniques like PCR, phenotypic tests become less practical because they require more time because the microbe must be grown. Alternatives, such as identifying the presence of a gene that causes resistance, can be used if the gene is known. *Streptococcus pneumoniae* resistance to penicillin is known to be caused by variations in penicillin binding proteins. A PCR-based test that assessed these variations would have content validity, because penicillin is known to bind to bacterial cell walls at the site of penicillin binding proteins. If presence of mutations in binding proteins correlates with resistance phenotype, then the test would have construct validity. Finally, if variants in penicillin binding proteins predict resistance based on a phenotypic test, the measure would have criterion validity.

| Type of Validity | Definition | Assessment |
|-----------------|------------|-----------|
| Content         | Pertains to underlying biologic phenomena | Professional judgment, consensus |
| Construct       | Correlates with relevant characteristics of the phenomenon | Correlation or agreement |
| Criterion       | Predicts an aspect of the phenomenon | Sensitivity, specificity, and predictive value |
Although nonculture techniques show great promise for rapid detection of antibiotic resistance for many organisms – meeting all three types of validity – they have some disconcerting limitations. Phenotypic tests for antibiotic resistance detect resistance regardless of mechanism. A PCR-based test is limited to a specific genetic mechanism of resistance. Thus, if not all mechanisms of resistance are included in a test or a new mechanism emerges, the test will be incapable of detecting that the bacterium is resistant. Further, a mechanism may be present and not active because the gene is degenerate, there is a missing regulator, or there is some other reason. In this case the test would be falsely positive.

The extent that a measure is valid is generally estimated using sensitivity and specificity (Table 8.5). Should there be no reference standard (gold standard) the extent that the new measure agrees with the old or the correlations between the measures might be reported. Ideally, the sensitivity and specificity should approach 100%. There is a trade-off between the two, however, and the cutpoint chosen depends on the type of test, the population under study, and whether the test is for screening or diagnostic purposes.

Even a test that approaches the ideal may result in error, and the extent of that error depends on the prevalence of the item of interest in the study population. This is measured by the predictive value positive, which is interpreted as the probability that the result is truly positive given a positive test. Consider a test that is 99.9% sensitive and 99.9% specific. If the prevalence of the item in the study population is 5% the predictive value positive is 99.1%. This translates to 9.5 truly negative individuals out of every 10,000 screened being misdiagnosed as positive. The predictive value negative is even higher, 99.9%, but 0.5 truly positive individuals will be misdiagnosed as negative. For a prevalence of 1%, the predictive value positive is 91.0% and predictive value negative remains 99.9%. Should the sensitivity and specificity be only 95% each, the predictive value positive falls to 50% if the prevalence is 5%, although the predictive value negative is 99.7%. Should the prevalence be 1%, the predictive value positive drops to 16%.

For many tests there is no gold standard available. This is often the case with new tests that assay a characteristic that was previously unmeasurable, such as gene expression profiles. In this situation, after the reliability of the test is assessed, it is standard to assess the predictive validity of the test. Predictive validity is the extent that the test predicts an outcome of interest. This entails obtaining prospective specimens for testing and observing the extent the test can discriminate between those with and without the outcome of interest. Leishmaniasis is a vector-borne disease of humans and animals caused by a parasitic protozoan of the genus *Leishmania*. No gold standard is available, and there are questions regarding the validity of each of the available methods. To assess the validity of various methods, Rodriguez-Cortes and colleagues used an experimental model and compared the predictive validity of each method. This assessment also pointed out the strengths and weaknesses of each test for different purposes; the best test for diagnosis was an enzyme-linked immunosorbent assay (ELISA). Quantitative PCR was useful for tracking parasite load.

### Table 8.5: Sensitivity, Specificity, and Predictive Value

| Test positive | A | B | A + B |
|---------------|---|---|-------|
| Test negative | C | D | C + D |
|               | B + D | A + B + C + D |

*Sensitivity: Probability test positive, given truly positive: A/(A + C).*

†*Specificity: Probability test negative, given truly negative: D/(B + D).*

‡*Predictive value positive: Probability truly positive, given test positive: A/(A + B).*

§*Predictive value negative: Probability truly negative, given test negative: D/(C + D).*
but overall was less predictive than the ELISA test. The different tests were compared using receiver operating curves (ROCs).

### 8.4.1 Receiver Operating Curves

Receiver operating curves graphically display the trade-off between sensitivity and specificity for various cutpoints of diseased, nondiseased, exposed, nonexposed, or any test that dichotomizes a population. ROCs are used frequently in the diagnostic literature, because they enable comparison of tests and a quick evaluation of whether a test classification is better than might be achieved by chance alone. Researchers are also beginning to use them with classification rules applied to microarray data. ROCs enable comparisons between tests compared to the same reference values; such comparisons might be used to either choose between tests or determine optimal test order if tests are used in series.

Any test classification has some error, because the distributions of diseased/nondiseased or exposed/nonexposed groups overlap; dichotomizing has tremendous utility but is somewhat arbitrary. The distribution of diseased and nondiseased individuals might be as shown in Figure 8.2. Setting the cutpoint at a test value of 1 erroneously places some diseased persons into the nondiseased group and vice versa.

If the threshold value was moved to a test value of −3, no cases would be missed (100% sensitive), but given the distribution of nondiseased, virtually all nondiseased would be classified as diseased (little specificity). Alternatively if the threshold was set at 3, the reverse would be the case. All positives would have disease (100% specificity) but many cases would be missed (poor sensitivity). The desired cutpoint depends on the application and what other tests are available. Using multiple tests in parallel increases the validity; a typical strategy used in diagnostic testing.

Disease stage or extent of exposure can modify the validity. For example, in persons with HIV, antibody titers are only detectable after ~3 months postinfection. Study population

![Figure 8.2](image-url)

**FIGURE 8.2**

Distributions of diseased (D) and nondiseased individuals, shown as the probability density function, for a diagnostic test. If a test value of 1 is used to classify diseased from nondiseased, the sensitivity (the shaded area under the diseased distribution A) is 72%, and specificity (the shaded area under the nondiseased distribution B) is 84%. Source: Reproduced, with permission, from Zou et al.⁵ (2007) and Mauri et al.⁷ (2005).
variables may also modify validity so the investigator should ensure that the selected cutpoint is appropriate for the desired purpose. Some ethnic groups have different ranges of normal values than others. Tests can also be done in series, also increasing the validity as is generally done for screening. HIV screening is done in series. All individuals positive by the first test are retested with a more definitive test. Because the costs of missing a positive case are high, the sensitivity of the HIV screening test is set very high. Further tests are more specific, but less sensitive. Because the population screened using additional tests is enriched for potential cases (higher prevalence) the predictive value positive and negative will be improved. This strategy misses the false negatives, as only those screening positive are retested.

These trade-offs are visualized by plotting the sensitivity (true positive rate) versus 1 – specificity (the false positive rate). An ideal test hugs the y-axis and then moves parallel to the x-axis at the highest value of the y-axis, line A on Figure 8.3. Line A represents the gold standard. A typical test (after smoothing) looks like line B; line C is the chance line, because a test that fits that line classifies no better than chance alone. The area under the curve (AUC) gives a quantitative assessment of the accuracy of the test. If the test is perfect the AUC is 1.0; an AUC of 0.5 is considered poor. Guinto-Ocampo and colleagues compared three laboratory indicators, white blood cell count, percent of lymphocytes, and absolute lymphocyte count (ALC) to a PCR test for pertussis among 141 infants who were tested for pertussis; 18 infants (13%) tested positive. The ROC curves were not smoothed (Figure 8.4). Notice that ALC is closest to the upper left-hand corner, and thus predicted best of the three. The area under the curve was 81% (95% CI: 72%, 90%). The point maximizing sensitivity and specificity was an ALC cutoff of 9400; using this cutpoint the sensitivity was 89%, specificity was 75%, the positive predictive value was 44%, and the negative predictive value was 97%. Although ALC was not a strong predictor of pertussis, in the absence of a licensed PCR test for pertussis, using the ALC cutpoint provides a reasonable guide to patient management, at least while awaiting results of culture, which can take up to 10 days.
8.5 DETERMINE THE INTRALABORATORY AND INTERLABORATORY RELIABILITY

Intra- means within and inter- means among or between. Random variation occurs within and between laboratories; the smallest variation is observed when replicate samples are tested in the same experiment, under identical conditions, and the largest variation is observed when samples are tested in different laboratories using different techniques (Table 8.6). There are some situations in which all specimens may be tested simultaneously in the same experiment, for example, molecular fingerprinting of bacterial isolates from a small disease outbreak. The accuracy and reliability of many rapid typing techniques, such as PCR-based techniques using random primers or repetitive elements, often depends on the ability to test isolates in a single experiment, as there can be considerable variation from experimental run to experimental run, although the findings within a run will be informative.

Intralaboratory reliability generally means assessing reliability among the technicians who will be conducting the experiments, and if there are multiple pieces of equipment, among equipment. The strategy is the same as shown in Table 8.3.

Determining interlaboratory variation is important for comparing results across geographic areas. In our global economy the potential for rapid spread of infectious agents is real, as demonstrated by severe acute respiratory syndrome (SARS). Antibiotic resistance also may emerge locally but be spread by travelers or via food transported within and between countries. Hageman and associates9 (2003) conducted an assessment of laboratories participating in the U.S. National Nosocomial Surveillance System to validate antimicrobial testing results; 193 laboratories participated from 39 states. Laboratories were sent test organisms including an oxacillin-resistant *Staphylococcus aureus* and vancomycin-resistant *Enterococcus faecalis*; *S. aureus* and *E. faecalis* are important causes of hospital-acquired infections (nosocomial). All laboratories were able to correctly identify the *S. aureus* as oxacillin resistant; and although 88% of the laboratories identified the *E. faecalis* as resistant to vancomycin, the accuracy depended heavily on the testing method. Only 28% of laboratories using the disk diffusion methods correctly reported the result compared to 94% using minimum inhibitory concentration methods. These types of assessment studies highlight which methods work best in practice; providing feedback to participating laboratories improves diagnostic capability everywhere and the accuracy of surveillance data. Regular assessment of the validity and reliability of laboratory testing within and among participating laboratories ensures overall data quality. This topic is addressed in more detail in Chapter 10 on study conduct.

8.6 DETERMINE THE APPROPRIATE INTERPRETATION OF THE MEASUREMENT

The reliability and validity of a test are determined by the test developers, but these determinations are only a guide. What might be achieved in a given laboratory depends

| Variation | Technique | Experiment | Day | Laboratory |
|-----------|-----------|------------|-----|------------|
| Least     | Same      | Same       | Same| Same       |
|           | Same      | Different  | Same| Same       |
|           | Same      | Different  | Different| Same |
|           | Different | Different  | Same| Different  |
|           | Same      | Different  | Different| Same |
|           | Different | Different  | Different| Same |
| Most      | Different | Different  | Different| Different|
on the technical expertise, equipment, and population under study. It is essential that the investigator determine the reliability and validity of the test in his or her laboratory, and that these levels are monitored (via use of duplicate samples, and positive and negative controls) throughout the conduct of the study. Ideally, laboratory personnel will not know which samples are duplicates nor the exposure or disease status of the person from whom the samples were collected (a procedure known as masking or blinding). Once the investigator is confident that the test is being conducted properly, the next step is to determine the appropriate interpretation of test results. Determining the correct interpretation depends on whether the study purpose is diagnostic or exposure assessment.

Often, the intended result of a test is diagnostic – is the individual healthy or ill? As mentioned in the discussion of validity, the investigator should consider a variety of factors when interpreting the results of a particular diagnostic test, especially how the results might be modified by the characteristics of the study population. Clinical decisions are generally based on reference ranges specific to the local population, values that reflect the observed mean and 95% confidence interval. Determining reference values is standard procedure for clinical laboratories; these values are generally reported to clinicians along with test results, reflecting the local population as well as local laboratory conditions that influence test values. Multilaboratory studies thus must consider local variations in values; all participating laboratories will use the same standards (positive and negative controls), which can be used to normalize values across sites.

All tests involve some error. *Ascaris* is a parasitic round worm that lives in the intestine, consuming partially digested food. If eggs are detected in stool samples it is certain that the individual is infected. But for a variety of reasons – including the number of worms present and the distribution of eggs in the stool – parasite eggs are not always found in stool even though the parasite is present. Thus clinically, it is essential to consider the entire clinical picture, rather than the result of a single test. In some populations, presence of a single parasite ovum in the stool will not necessarily correspond to clinical disease. This does not mean that there is not some effect of parasite load on health, only that in endemic populations low levels might be considered in the normal range.

That population-specific norms are important for the clinical interpretation of a measure should be kept firmly in mind as new molecular measures are developed to characterize health and disease. Limited predictive value for a measure in one population does not preclude its ability to predict in another or across populations, perhaps indicating a causal relationship. Where malaria is endemic, a blood smear positive for plasmodia is not predictive of disease symptoms except among the very young, pregnant women, and those visiting from malaria-free areas. A cross-sectional study in an endemic population will find most individuals to have malarial parasites present. In this case, the presence of parasites in the blood might be dismissed as an incidental finding; those without symptoms also have them. Therefore it would be difficult using an endemic population to demonstrate an association between parasite presence and symptoms or even to use parasite load as a meaningful indication for therapy. When developing a diagnostic test or using a test as a potential marker of exposure or outcome in explorations of causality, the epidemiologist must consider the entire population picture.

Similar to clinical measures, many molecular measures are only appropriately interpreted within a population context. Molecular fingerprints (described in detail in Chapter 6), are an important adjunct to epidemiologic outbreak investigation. When the epidemiologic evidence is strong, the fingerprinting method is considered confirmatory, verifying that all affected individuals were infected by a microbe with the same fingerprint and it was found in the putative source (in a common source outbreak). However, in the absence of epidemiologic evidence, the presence of a common fingerprint is a much poorer predictor of true linkage, as the probability that the fingerprint might occur in the population by
chance alone is much higher. This might be remedied by studies describing the distribution of fingerprints for that organism in the population for comparison. Because the generation time for microbes is much shorter than for humans, and the genomes of many pathogenic bacteria are quite heterogeneous, this would have to be an ongoing project.

Interpretation is also influenced by how close the measurement is to the construct of interest. Finding the presence of genes that code for a specific factor in an infectious agent does not guarantee that the gene is expressed in vivo. It may only be expressed under certain circumstances, the gene may not be functional because the code has been modified in ways not detectable by the test, or there must be other genes present and active for expression to occur. This is a limitation of rapid assessments for antibiotic resistance based on gene presence rather than gene expression. Tests based on gene expression also have limitations. Gene expression does not have a one-to-one correlation with protein production. Most processes require multiple steps; a test will generally focus on expression indicating one step. If other steps are not functioning, the protein of interest will not be made, or may be made but not in proper form. Even detection of the phenotype under laboratory conditions may not have a direct correlation to what is observed in vivo. Treatment of a bacterium with an antibiotic to which it is resistant in vitro may still result in clinical cure. The antibiotic may concentrate where the bacteria clusters, or the resistance mechanism is overexpressed under laboratory conditions invalidating test results, or there may be some other reason.

It is also critical to have a clear fix on what constitutes a negative test versus a failed test. A negative test implies that the characteristic of interest is not present. A failed test can look like a negative test but it fails to assess anything. The difference is enormous when interpreting study results. If a positive test is based solely on the presence of a PCR product, it is impossible to determine what the absence of the product means. Some proportion, hopefully all, are true negatives. Without proper controls, such as including primers for a genetic sequence that should always be present that gives a PCR product of a different size, it is impossible to determine if the lack of product was due to experimental error.

Proper interpretation of exposure measurement is similar to that of outcome measures. It entails an understanding of the biological variability of the measure, the inherent variability due to testing, and the construct being measured. Biological measures may vary markedly within an individual. After a night’s sleep, we are all taller than after a day spent upright. If there is almost as much variation within an individual over the course of a week as observed between two individuals at a single point in time it is difficult to draw any valid conclusions.

A highly discriminatory measure makes it possible to distinguish between groups, often with great precision and statistical significance. This level of precision may be misleading, as the scale of measurement can imply a greater level of precision than is inherent in the data. The investigator is cautioned to have clear criteria for determining validity, lest an excellent test is consigned to the trash can because the measuring scale suggests it is too variable. Envision an ELISA that quantifies the amount of human chorionic gonadotropin (hCG) present in the urine. hCG is only found in the urine if a woman is pregnant. The added level of precision that quantifying the amount of hCG has no real meaning if the construct being measured is pregnancy. However, a high degree of variation may suggest a poor test, when actually it reflects stage of fetal development.

**8.7 USING STORED MATERIALS**

Assembling and following a large human cohort requires significant effort and expense, especially if biological specimens are collected. Using existing data collections and piggybacking on ongoing efforts is thus extremely attractive. However, using existing collections is not without limitations (Table 8.7). Similarly, using specimens collected for clinical testing for developing or validating new diagnostics has the advantage of
convenience, but there are inherent limitations that should be taken into account in data interpretation. Further, the type of specimen, length of time in storage, and method of specimen storage all influence the use of the specimens in future studies. In general, none of these are fatal flaws, but any one may substantially limit the study generalizability.

The original study design constrains what design can be used for a study of existing data. Depending on the sampling scheme, specimens from a cohort study might be analyzed as a cohort study, a cross-sectional study, or a case–control study for a variety of outcomes and exposures independent of the original study purpose. Other study designs impose stronger constraints: specimens from a case–control study are generally limited to further examining the same outcome, although the case definition might be refined after specimens are tested. If the controls were population-based or sampled from a cohort, the controls might be analyzed to give insight into the population prevalence of a variety of variables, similar to a cross-sectional study.

Different study designs impose different sampling schemes that limit the parameters that can be estimated, and the generalizability of results (see Chapter 9). Inferences made using existing specimens will have the same limitations. A study population may have been limited by age, gender, or racial/ethnic group for reasons associated with the primary study purpose. Controls may have been selected to match case characteristics; while optimizing ability to test the primary study hypotheses, specimens from controls will be highly selected and not reflect the general population. Disease risk is often modified by individual behavior; the effects of exposure may be modified or mediated by other exposures not easily ascertained from biological assays. These types of data may not have been collected and may not be easily ascertained from records.

Even a well-conducted study has limitations. Not all data will be collected for all participants for all variables and all time points. Missing data and loss to follow-up result in less than the entire sample included in analyses, often substantially reducing the effective sample size. Not only does this reduce the power to detect significant effects and the precision of estimates, patterns of loss usually are not random. The effects of such losses can bias inference and generalizability. The potential for bias can be evaluated by comparing results from analyses of the subset with complete data to a set where missing values are inputted. Selection biases occur even within repositories and data banks; reasons for participation or refusal are associated with health, access to medical care, clinical manifestations of illness, socioeconomic status, and age, which in turn are associated with disease risk.

Not all variables can be accurately assessed from existing specimens. For many variables that may modify the study interpretation and generalizability of study results, the investigator will be dependent upon the quality of the data already collected. Because the purpose of the secondary analysis may be somewhat or even quite different from that of the original study, important modifying variables may not be measured with the optimum level of precision. This in turn can lead to uncontrolled confounding adversely affecting the validity of the secondary study findings.

Specimens are generally collected over time, even for a cross-sectional study, and therefore are stored for varying lengths of time. In a case–control study, controls may be identified at

### TABLE 8.7 Considerations When Using Stored Specimens

- Constraints imposed by study design
- Limitations imposed by study conduct (study population, selection biases)
- Reliability and validity of existing diagnosis and exposure variables
- Sample quality
- Sample quantity
- Effects of storage on substrate of interest
the same time as cases or only after the case groups is assembled. If the latter, and the assay is sensitive to this time difference, the results could suggest a difference between groups solely attributable to storage. Similarly, if specimens are sensitive to storage and are collected from individuals each year, degradation over time might be erroneously interpreted as an increase in the item of interest with time: with less time to degrade, the most recent specimens will have higher levels. These effects are difficult to assess, but some insight can be gained, if there is sufficient sample available, by retesting for measures tested before storage.

Specimens are stored in a variety of ways. They may be placed in various media or solutions and frozen, or preserved or dried and stored at room temperature. The type and extent of processing and storage all can influence the experimental results; the severity of the effect depends on whether the item of interest is DNA, RNA, protein, or a microbe that the investigator plans to regrow. DNA, RNA, or protein may alter or degrade with time. Microbes may die, or when regrown following storage display different phenotypes. Any changes in results attributable to storage are problematic because they add an additional source of variation beyond that due to technical and biological variation. Sample quality depends on how the original sample was collected, the processing it was subjected to, and the storage conditions – including times thawed and refrozen. Different laboratory tests are more or less sensitive to these factors; when using existing specimens the investigator must judge the potential impact on any inferences made. Thus it is critical to assess the validity and reliability of the planned assays on the specimens. As stored specimens are precious and often irreplaceable, these studies must be planned carefully to maximize the assessment while minimizing any losses.

Laboratory procedures might be optimized using freshly collected specimens subjected to the same handling and processing as those from the repository. Although the same length of storage probably cannot be duplicated, specimens might also be frozen and thawed, for example, to assess any effects on results relative to fresh specimens. Initial assessments of repository specimens might be made using material left over from another use, or on samples from which multiple replicates are available. These assessments can detect if the specimen contains material that inhibits or modifies the reactions in the planned experimental procedures, and, if present, to identify additional processing steps that might minimize these results. It also can determine if the specimens are of sufficient quality to be reliably and accurately assessed using the planned testing procedures.

Most tests require a minimal amount of sample for testing. If there is limited amount of sample available that is not renewable, or if there are concerns about affecting specimen integrity if thawed and refrozen, it may be difficult to obtain access to stored samples. Bacterial and viral collections are generally renewable resources, and human DNA can be immortalized, making it effectively renewable. But other microbes and biological specimens (blood, cells, tissue) generally are not renewable. The investigator will want to optimize use of nonrenewable resources, grouping secondary tests together, which may limit the amount of sample available. Even if collections are renewable, investigators are appropriately concerned that other users be aware of the collections strengths and weaknesses, and appropriately take into account any design limitations. Ensuring proper use and data interpretation takes time and effort even if data are well documented. Moreover, the group collecting the data quite rightly would like to have first crack at using it. Finally, identifying, packing, and shipping specimens to another site for testing is not without cost. Thus, it is not surprising that repositories tend to closely guard access to their collections, often requiring potential users to complete some sort of application process, and tending to favor those with whom they have some sort of social connection. Unfortunately, this can mean that specimens are not used to optimum effect, because the zeal to protect the collection can lead to little or no use.
References

1. N.T. Atkins, More on Bias (systematic) and Random Errors [website]. Lyndonville: Lyndon State College. Available at: http://apollo.lsc.vsc.edu/classes/remote/lecture_notes/measurements/bias_random_errors.html (Updated 7.1.2000; accessed 7.1.10).

2. L. Zhang, B.W. Gillespie, C.F. Marrs, et al., Optimization of a fluorescent-based phosphor imaging dot blot DNA hybridization assay to assess E. coli virulence gene profiles, J. Microbiol. Methods 44 (3) (2001) 225–233.

3. B. Foxman, B. Gillespie, S.D. Manning, et al., Incidence and duration of group B Streptococcus by serotype among male and female college students living in a single dormitory, Am. J. Epidemiol. 163 (6) (2006) 544–551.

4. A. Rodriguez-Cortes, A. Ojeda, O. Francino, et al., Leishmania infection: laboratory diagnosing in the absence of a gold standard, Am. J. Trop. Med. Hyg. 82 (2) (2010) 251–256.

5. S.G. Baker, B.S. Kramer, Identifying genes that contribute most to good classification in microarrays, BMC Bioinformatics 7 (2006) 407.

6. K.H. Zou, A.J. O’Malley, L. Mauri, Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models, Circulation 115 (5) (2007) 654–657.

7. L. Mauri, E.J. Orav, A.J. O’Malley, et al., Relationship of late loss in lumen diameter to coronary restenosis in sirolimus-eluting stents, Circulation 111 (3) (2005) 321–327.

8. H. Guinto-Ocampo, J.E. Bennett, M.W. Attia, Predicting pertussis in infants, Pediatr. Emerg. Care 24 (1) (2008) 16–20.

9. J.C. Hageman, S.K. Fridkin, J.M. Mohammed, et al., Antimicrobial proficiency testing of National Nosocomial Infections Surveillance System hospital laboratories, Infect. Control and Hosp. Epidemiol. 24 (5) (2003) 356–361.

10. M. Porta, I. Hernandez-Aguado, B. Lumbreras, et al., Omics research, monetization of intellectual property and fragmentation of knowledge: can clinical epidemiology strengthen integrative research? J. Clin. Epidemiol. 60 (12) (2007) 1220–1225 discussion 1226–8.