Correctly Using Sensitivity, Specificity, and Predictive Values in Clinical Practice: How to Avoid Three Common Pitfalls

David M. Naeger[1]
Maureen P. Kohi[1]
Emily M. Webb[1]
Andrew Phelps[1]
Karen G. Ordovas[1]
Thomas B. Newman[2]

OBJECTIVE. Radiology is the specialty of imaging-based diagnostic tests. Understanding the science behind evaluating diagnostic test performance is essential for radiologists because we provide care to patients and interact with our colleagues.

CONCLUSION. Here, we review the key terminology used and common pitfalls encountered in the literature and in day-to-day discussions of diagnostic test performance.

Radiology is the specialty of imaging-based diagnostic tests. The science behind evaluating diagnostic tests, so-called "evidence-based diagnosis" [1], is fundamental to our discipline [2–4]. Here, we review key terminology used to describe the performance of a diagnostic test (imaging and nonimaging), and we review three very common pitfalls encountered when these principles are discussed. Each pitfall is framed in the form of a case.

Case 1

You are covering the emergency department (ED) radiology service for the day and you encounter a minimally displaced scaphoid fracture in a patient who fell on her outstretched hand. You tell the ED physician your findings, and he asks you for the sensitivity of radiographs for such fractures. Having read a recent study, you reply "About 75%." The ED physician says, "Well, if you're that far away from 100% certain of a fracture, we might need an MRI before I call the surgeon. Can you arrange one?" What do you say?

What Exactly Do Sensitivity and Specificity Tell You?

Sensitivity and specificity describe how a test performs in people with known disease status. In clinical practice, we deal with patients of unknown disease status; this should immediately give us pause about how we use these terms.

Sensitivity expresses how a test performs in people known to have the disease. Highly sensitive tests tend to be positive in patients with disease. This parameter, therefore, depends on the biology of the disease (the chemical or anatomic abnormalities that result from the disease) and characteristics of the test (e.g., how well the machinery or chemical tests detect the abnormalities). For example, the sensitivity of ultrasound for gallstones depends on the underlying biology of gallstones (size and composition), the technology of the ultrasound machine, the technique of the sonographer, and the skill of the reader. Every step along the way affects the overall sensitivity of the test.

Specificity is how the test performs in people who are known to not have disease. Highly specific tests tend to be negative in patients without disease. This parameter also depends on the underlying biology of the disease and characteristics of the test. Even though sensitivity and specificity both involve the test and the underlying biology, they are inherently separate concepts. The specificity of ultrasound for gallstones, for example, involves the biology of healthy people (i.e., how often do patients without gallstones have high-density sludge, polyps, or other stone mimics) and the characteristics of the ultrasound machine, technician, or reader when nothing is truly there. For example, how often does the technologist capture images with artifacts over the lumen mimicking stones? Does the reader tend to overcall gallstones?

By definition, these are separate characteristics of tests. They are determined in different populations (people known to have disease vs not) and rely on different characteristics of the test (how good is it at finding the abnormality vs how likely is it to incorrectly suggest an abnormality).

Keywords: clinical epidemiology, evidence-based diagnosis, predictive values, sensitivity, specificity

DOI:10.2214/AJR.12.9888

Received August 29, 2012; accepted after revision October 23, 2012.

[1]Department of Radiology and Biomedical Imaging, University of California, San Francisco, 505 Parnassus Ave, M-391, Box 0628, San Francisco, CA 94143-0628. Address correspondence to D. M. Naeger (david.naeger@radiology.ucsf.edu).

[2]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA.

WEB
This is a Web exclusive article.

AJR 2013; 200:W566–W570

0361–803X/13/2006–W566

TABLE 1: Standard 2 × 2 Table for Diagnostic Test

	Disease Positive	Disease Negative
Test positive	*A*	*B*
Test negative	*C*	*D*
Total	*A* + *C*	*B* + *D*

Note—As per convention, true disease status as determined by reference standard is on top. Although many different study designs can be summarized in 2 × 2 table, if population is sampled irrespective of disease state (i.e., cross-sectional sampling), the proportion of study population with disease should mirror the proportion in the population. The proportion with disease (often called prevalence) equals those with disease (*A* + *C*) over everyone recruited (*A* + *B* + *C* + *D*). Thus, disease prevalence equals (*A* + *C*) / [(*A* + *C*) + (*B* + *D*)].

TABLE 2: Standard 2 × 2 Table Broken Down Into Two 2 × 1 Columns

	Disease Positive			Disease Negative
Test positive	*A*		Test positive	*B*
Test negative	*C*		Test negative	*D*
Total	*A* + *C*		Total	*B* + *D*

Note—When you only know sensitivity and specificity of test, it may be helpful to think of the 2 x 2 table as two columns that can be evaluated separately. Sensitivity (*two left columns*) can only be determined in population confirmed to have disease. Total disease population tested is equal to *A* + *C*. *A* is number of diseased people in whom test is correct, and *B* is number in whom test is incorrect; thus, sensitivity equals *A* / (*A* + *C*). Specificity (*two right columns*) is the proportion of patients with correct test result (i.e., "negative") of all nondiseased people; thus, specificity equals *D* / (*B* + *D*).

TABLE 3: Patients With and Without Disease, by CSF Status (Case 2)

CSF Status	Disease Positive	Disease Negative	Total
High-density CSF	99	1	100
Normal-density CSF	1	99	100
Total	100	100	200

If you are very familiar with these concepts, you might argue that, although they are separate characteristics, there is a connection between sensitivity and specificity. It is true that extremely sensitive tests tend to have lower specificity and vice versa. This is not a rule, however. Also, for any given test in which a threshold is used to determine positivity, adjusting the threshold almost always improves sensitivity or specificity at the cost of the other.

Let us address an important question: how do we determine the sensitivity and specificity of a test if they can only be calculated in people with known disease status? After all, the point of developing a test is to diagnose the disease. These parameters are determined in research studies that use a reference standard to confirm the true disease status. What happens if the reference standard is imperfect or if the reference standard is a clinical determination that partly relies on the diagnostic test in question? These are great questions that may suggest that the reference standard is not entirely the "reference" and, therefore, a potential source of bias. Sensitivity and specificity calculated with poor reference standards may be inaccurate. As astute readers, we should closely evaluate the reference standard in any study reporting diagnostic test performance.

A common way to recruit patients for a study of a diagnostic test is to use case-control sampling. The word "sampling" can be thought of as a synonym for "recruitment" and specifically refers to the way in which patients are recruited, or sampled, from the underlying population. With case-control sampling, a group of diseased patients (case patients, in whom the disease is confirmed with the reference standard test) is compared with a group of nondiseased patients (control subjects, in whom the lack of disease is also confirmed with the reference standard). The number of control subjects studied is often chosen to achieve a one-to-one or other ratio to the number of cases. Both groups undergo the test. Although we can summarize the results of a study such as this in a standard 2 × 2 table (Table 1), it may be helpful to first think of the results summarized in two 2 × 1 tables (Table 2).

Case 1 Answer: Pitfall 1—Confusing Sensitivity or Specificity for Predictive Values

The conversation in case 1 contains a number of problems. With a positive radiograph, the real question is how likely is it that this positive radiograph reflects a true fracture versus something that looks like a fracture but is not (a false-positive)? If one would like to get an idea for how often this second possibility happens, a false-positive finding, one would ask about specificity (not sensitivity). Highly specific tests rarely result in false-positives (i.e., rarely is positive in those without disease), and low-specificity tests often result in falsely positive findings.

The second, larger, problem is that the ED physician was not trying to ascertain the specificity of the test to get an idea how often it may be falsely positive; he actually wanted a precise measure of "certainty" after receiving a positive test result. Neither specificity nor sensitivity tells us exactly how certain we are of a diagnosis after a diagnostic test result. That question depends on the precise balance of true-positives (actual fractures) versus false-positives (fracture "fake-outs"). With a positive or negative test in hand, the certainty of diagnosis is addressed with the concept of a "predictive value," as described in the next section.

Case 2

Suppose a new disease was recently discovered in people who live below high-voltage power lines. Apparently, only years and years of exposure cause the disease, but even then, it is very rare. Only 100 cases are known. There is a long presymptomatic phase, but once it is clinically evident, the disease is irreversible and debilitating. The government has encouraged people near power lines to move, but the costs are tremendous and residents are reluctant to leave because the disease is so rare. For years, an accurate diagnostic test for presymptomatic patients was lacking, until the field of radiology comes to the rescue. High-density CSF on head CT proves to be 99% sensitive and 99% specific for the disease. The landmark study describing this finding recruits all 100 clinically symptomatic cases known; 99 have high-density CSF. They also recruit 100 healthy subjects taken from nearby towns who do not live near the power lines; 99 of them do not have high-density CSF. Table 3 is an example of a 2 × 2 table created from these data.

In this scenario, the authors report that the sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) are all 99% (each is 99/100). They state, "This is a

TABLE 4: Example 2 × 2 Table From Study Using Cross-Sectional Sampling [5]

Polyps Found on CT Colonography	Polyps Found on Colonoscopy	
	≥ 1 Polyp	0 Polyps
≥ 1 polyp	164	33
0 polyps	18	85
Total	182	118

Note—Three hundred patients were recruited who underwent CT colonography (test) and colonoscopy (reference standard). Patients were recruited sequentially before either test was performed (i.e., cross-sectional sample). Sensitivity (90% [164/182]) and specificity (72% [85/118]) were calculated. Disease prevalence in study (61% [182/300]) should mirror prevalence of standard referral population at institution (barring significant biases). Please note that this is simplified summary of one finding in a complex and detailed article.

TABLE 5: Determining Predictive Values From Standard 2 × 2 Table

	Disease Positive	Disease Negative	Total
Test positive	*A*	*B*	*A* + *B*
Test negative	*C*	*D*	*C* + *D*
Total	*A* + *C*	*B* + *D*	(*A* + *C*) + (*B* + *D*)

Note—Predictive values are determined by looking at rows of 2 × 2 table. Positive predictive value is proportion of true-positives (*A*) divided by all positive test results (*A* + *B*), or *A* / (*A* + *B*). Negative predictive value is proportion of true-negatives (*D*) divided by all negative test results (*C* + *D*), or *D* / (*C* + *D*). Disease prevalence equals (*A* + *C*) / [(*A* + *C*) + (*B* + *D*)].

ground-breaking finding. A positive test confers a tremendously high risk of disease (99%). Everyone near power lines should be screened and everyone with a positive test should be moved at any cost to avoid almost certain debilitation from this dreaded disease." Do you agree?

Putting the Two Columns Together: What Assumptions Are Made When Creating a 2 × 2 Table?

How do we move from a world where we only know sensitivity and specificity (which can be graphically depicted as two columns, as in Table 2) to a world where we put them together and can consider predictive values? The most intuitive use of a 2 × 2 table is when the numbers in each box come from a study design that uses cross-sectional sampling. Again, the word "sampling" refers to how patients are recruited, or sampled, from a predefined underlying population. In a study with cross-sectional sampling, a random or consecutive sample of the population is selected irrespective of disease or test status. With a screening test, the underlying population might be quite large (e.g., all people older than 50 years), but with targeted tests, the population may be quite small (e.g., only people with certain unusual symptoms). Regardless, patients are recruited from the underlying population without the investigators knowing their disease or test status. Once the patients are recruited, the authors then have to identify who has the disease using the reference standard and who has a positive test with the test in question. With rare diseases, a very large sample would be needed using this type of sampling, or else there would be very few people with the disease. Table 1 is a standard 2 × 2 table. An example study from the radiology literature is summarized in Table 4; the test in this case, CT colonography, is compared with a reference standard of optical colonoscopy [5]. Note in this example that the disease, having a polyp, is actually quite common, at 61% of the sampled subjects. The important point is, when a study recruits patients irrespective of disease status, the prevalence in the study population should reflect the prevalence in the underlying population.

Predictive Values

With a 2 × 2 table filled with cross-sectional sampling study data, we can now discuss PPVs and NPVs. These terms are test centric, meaning that they provide information about what the test means in patients with positive or negative test results. This is unlike sensitivity and specificity, which reflect the test's performance in people with known disease or nondisease (i.e., these are disease-centric terms). The PPV is the proportion of patients with a true-positive test among all patients with a positive test (which includes both true- and false-positive tests). The NPV is the proportion of patients with a true-negative test among all patients with a negative test (which includes both true- and false-negative tests). Note that the test result (positive or negative) defines the group of people in the denominator, from which we find the proportion of people who test correctly. As when we calculate sensitivity and specificity, the numerator is always the "correct answers."

In a 2 × 2 table, predictive values are terms that involve looking at the numbers along the rows of the 2 × 2 table, not the columns. This concept is illustrated in Table 5.

Using (and Misusing) Predictive Values

PPVs and NPVs are the holy grail of diagnostic test research. Why? These parameters can be directly applied to patients in clinical settings [6]. For example, the referring physician in case 1 was, in truth, trying to ascertain the PPV to help him determine the next step in his management. When we do not have predictive values readily available, physicians consciously or unconsciously tend to estimate them and apply them clinically. The importance of predictive values (known or estimated) cannot be overstated.

What makes these values so important? When we are using diagnostic tests on clinical patients, we do not know their disease status. Presumably, patients undergoing diagnostic tests have not had the reference standard. If the reference standard were cheap and easy, there would be no reason to use some other diagnostic test at all. With clinical patients, we want to know the likelihood that the patient has the disease after a positive or negative diagnostic test result. This is exactly what the predictive values tell us.

Predictive values can be misused, as occurred in case 2. The misapplication is often not in the math but rather in calculating it with the wrong type of data, most commonly data from studies with case-control sampling (again, "sampling" refers to how the patients were recruited). These types of studies are often easier to perform than studies with cross-sectional sampling and are quite common in the radiology literature. Their ease comes from usually small sample sizes and the guarantee of having a reasonable number of both sick and healthy patients. Cross-sectional sampling may require large sample sizes, particularly if the disease is rare, and can therefore be very costly. The benefit to cross-sectional sampling, however, is it can yield sensitivity, specificity, and predictive values, as we showed already when we reviewed predic-

TABLE 6: People Living Under Power Lines, With and Without Disease, by CSF Status (Case 2)

CSF Status	Disease Positive	Disease Negative	Total
High-density CSF	99	9,990	10,089
Normal-density CSF	1	89,910	89,911
Total	100	99,900	100,000

TABLE 7: Determining Predictive Values and Chance of Becoming Infected, by Presence or Absence of Pericardial Calcification (Case 3)

Pericardial Calcification	Disease Positive	Disease Negative	Total
Present	99	99	198
Absent	1	9801	9802
Total	100	9900	10,000

tive values. Case-control sampling can only directly yield sensitivity and specificity.

What happens when predictive values are inappropriately calculated from a study with case-control sampling? Case-control–sampled studies usually have a disease prevalence in the study that is higher than that in the underlying population (i.e., reality). For example, one-to-one case-control recruitment always results in a study disease prevalence of 50%, which is a higher prevalence than observed with most diseases. A PPV inappropriately calculated from a study with case-control sampling (with artificially high disease prevalence) will be artificially high; this is because the higher disease prevalence in the study results in a greater abundance of true-positives. An NPV inappropriately calculated from a study with case-control sampling (with artificially high disease prevalence) is artificially low; this is because the artificially high study disease prevalence results in fewer healthy patients and fewer true-negatives.

Case 2 Answer: Pitfall 2—Inappropriately Calculating Predictive Values From Studies Using Case-Control Sampling

The case 2 study is one with case-control sampling. This is a very rare disease, and all 100 known cases are recruited. With 100 control subjects, we have an apparent disease prevalence in the study of 50%. That is obviously not representative of the population as a whole. As a consequence, the inappropriately calculated predictive values do not represent the predictive values of the test if it were applied to the underlying population in question. However, the calculated sensitivity and specificity are entirely correct and are the main contribution of this study. To illustrate this point, assume that the disease prevalence is actually 0.1% in those who would be tested (i.e., people living under power lines). To get 100 cases, one would have to recruit 100,000 subjects and perform 100,000 head CT examinations. The 2 × 2 table would look like Table 6.

The PPV in this case would be 0.009 (99/10,089), so the researchers in case 2 incorrectly overestimated the PPV. If this test were applied to the true population, a person with a positive test result would only have a 0.9% chance of actually having the disease. Although the test is very sensitive (i.e., nearly all diseased individuals will be detected), there are so many nondiseased individuals in the population being tested that the number of false-positive results becomes a real problem. If the government were to adopt screening with this test and move all individuals with a positive test result, 100 people would have to move to prevent 1 person from developing the disease.

You might think the world of research is doomed: studies with case-control sampling give the wrong PPV and NPV, and studies with cross-section sampling of rare diseases must be prohibitively huge. The truth is that there are ways around these problems: sensitivity and specificity can be combined mathematically with the underlying risk of disease (which must be known) to derive predictive values. The mechanics of this are discussed in detail elsewhere [1, 7].

If you are thinking to yourself that this is an overstated problem and that no one would make this mistake, here are two examples from the literature [8, 9]. In the first example, an inappropriately calculated predictive value is reported in the abstract [8], and in the second, inappropriately calculated predictive values are reported in a table [9].

Case 3

Suppose a frightening new infection is striking the nation. Approximately 1% of all Americans become infected before a cure could be found. The cure is extraordinarily expensive, as is the only known test for the disease (i.e., the reference standard), which also takes weeks of processing. The medical community begins searching for a less expensive and quicker way to identify those with early treatable disease. The field of radiology comes to the rescue. It is discovered that pericardial calcification detected on chest radiograph is an excellent sign, with 99% sensitivity and 99% specificity. The landmark study reports these values and also highlights the impressively low false-negative and false-positive rates: both 1%. Your institution immediately begins screening. You get a call from an upset patient who states, "Before I was tested, I had a 1% chance of having this dreaded infection [the prevalence of the disease]. I have a negative x-ray, which has a 1% false-negative rate, so that means I still have a 1% chance of having the disease. What good was this test?" What do you say?

False-Positive and False-Negative Rate Confusion

In a standard 2 × 2 table (Table 1), there is no confusion about who are the false-positives (group C) and false-negatives (group B). However, people like to think in terms of proportions and will often refer to a false-positive or false-negative rate.

The word rate itself is often confusing in medical writing and can have many meanings, usually referring to the frequency with which events happen over a period time. In this context, the word rate simply implies a proportion or percentage. Though we very much dislike it, we will use the word rate here because it is quite commonly used.

Even if we correctly understand the meaning of rate, we are still left with confusion. What is the denominator used to calculate

a false-positive or false-negative rate? Take the false-negative rate as an example: the number of false-negatives can be divided by the number of all negative test results (i.e., this would be a test-centric way of thinking about it, and the rate would be equal to 1 – NPV) or the denominator could be all truly disease-positive patients (i.e., this would be a disease-centric way of thinking about it, and the rate would be equal to 1 – sensitivity). In clinical practice, the test-centric value (1 – NPV) is the most useful, because it addresses the question of all the negative test results—that is, how many of those individuals actually have the disease despite having negative test results. The more conventional usage is actually the other, the disease-centric value (1 – sensitivity), probably because of the ease of calculating it from relatively ubiquitous sensitivity values.

Unfortunately, there are not unique terms to distinguish to which denominator a false-positive or false-negative rate refers. Although convention tends toward the disease-centric definition (with false-positive and false-negative rates implying 1 – specificity and 1 – sensitivity, respectively), many people intentionally or unintentionally use the other definition. One should not make assumptions when a speaker or author uses the phrase "false-negative rate"; rather, we should attempt to determine how the value was calculated. Also, as a writer or speaker yourself, it is important to be clear about which one you mean.

Case 3 Answer: Pitfall 3—False-Negative (and -Positive) Confusion

The article about the outbreak reports a 1% false-negative rate. Given that the sensitivity is reported as 99%, we assume that the calculation was made using the disease-centric understanding (1 – sensitivity). The patient, however, interpreted the rate to mean the more clinically relevant and test-centric understanding (1 – NPV), which would be the proportion of patients with a negative test (like this patient) who actually have the disease despite the negative test. To determine that number, we need to first know the NPV, which is not reported in the scenario, but which we can derive from the information in the problem. We could fill in a theoretic 2 × 2 table assuming we had a large population (say, 10,000 people). We know the prevalence is 1% (so we fill in the lower row with 100 and 9900). We are also told the sensitivity and specificity, so we can fill in the remaining cells, as in Table 7.

The NPV is 0.9999 (9801/9802), meaning a negative test confers a 99.99% chance of not having the disease. The risk of actually still having the disease given a negative test is 0.01% (100% – 99.99%). Even without doing the math, when presented with a situation such as this, we can realize two different meanings of the "false-negative rate" are being used, leading to confusion.

Do you think that this does not happen in practice? Here are two articles reporting the accuracy of a rapid influenza test against a reference standard: the Centers for Disease Control and Prevention website [10] reports a false-positive rate equal to 1 minus PPV, whereas the journal article [11] reports a false-positive rate equal to 1 minus specificity.

Conclusion

We have reviewed the key terms used to describe diagnostic test performance, which is the cornerstone of radiology. We have reviewed three very common pitfalls.

The first pitfall is confusing sensitivity and specificity with predictive values. Sensitivity and specificity describe the performance of a diagnostic test and are estimated in research studies. In daily radiology practice, predictive values are of the greatest utility because they can be applied to individual patients after they are tested (e.g., "What is the significance of this test result?").

The second pitfall is inappropriately calculating predictive values from studies with case-control sampling. When studies select an arbitrary number, or ratio, of subjects with and without the disease, predictive values cannot be directly calculated.

The third pitfall is confusion regarding false-negative and false-positive "rates." The proportion of patients with an incorrect test result (often confusingly referred to as a "rate") can be calculated with two different denominators, either all the people with the same test result (positive or negative) or all the people of the same disease status (with or without the disease). You must dig deeper if you encounter these terms to figure out which is being used; usually "rates" that refer to 1 minus predictive value are more clinically relevant.

Although the basics of epidemiology are taught in medical school, and are even tested on national licensure examinations, we find that these pitfalls are fairly common both in daily discussions and in the literature. Being familiar with evidence-based diagnosis will help us serve our patients and referring clinicians. For interested readers, additional articles are available describing these terms in intuitive ways [6, 12–15], and many textbooks that provide more in-depth discussions are available [1, 7, 16].

References

1. Newman TB, Kohn MA. *Evidence-based diagnosis*. New York, NY: Cambridge University Press, 2009
2. Evidence-Based Radiology Working Group. Evidence-based radiology: a new approach to the practice of radiology. *Radiology* 2001; 220:566–575
3. Dixon AK. Evidence-based diagnostic radiology. *Lancet* 1997; 350:509–512
4. Erturk SM, Ondategui-Parra S, Otero H, Ros PR. Evidence-based radiology. *J Am Coll Radiol* 2006; 3:513–519
5. Yee J, Akerkar GA, Hung RK, Steinauer-Gebauer AM, Wall SD, McQuaid KR. Colorectal neoplasia: performance characteristics of CT colonography for detection in 300 patients. *Radiology* 2001; 219:685–692
6. Altman DG, Bland JM. Diagnostic tests 2: predictive values. *BMJ* 1994; 309:102
7. Szklo M, Nieto FJ. *Epidemiology: beyond the basics*, 2nd ed. Sudbury, MA: Jones and Bartlett, 2007
8. Davis DP, Wold RM, Patel RJ, et al. The clinical presentation and impact of diagnostic delays on emergency department patients with spinal epidural abscess. *J Emerg Med* 2004; 26:285–291
9. Zangwill KM, Hamilton DH, Perkins BA, et al. Cat scratch disease in Connecticut: epidemiology, risk factors, and evaluation of a new diagnostic test. *N Engl J Med* 1993; 329:8–13
10. Centers for Disease Control and Prevention. Rapid diagnostic testing for influenza: information for clinical laboratory directors. Centers for Disease Control and Prevention website. www.cdc.gov/flu/professionals/diagnosis/rapidlab.htm. Published July 6, 2011. Updated January 3, 2012. Accessed August 27, 2012
11. Waner JL, Todd SJ, Shalaby H, Murphy P, Wall LV. Comparison of Directigen FLU-A with viral isolation and direct immunofluorescence for the rapid detection and identification of influenza A virus. *J Clin Microbiol* 1991; 29:479–482
12. Loong TW. Understanding sensitivity and specificity with the right side of the brain. *BMJ* 2003; 327:716–719
13. Deeks JJ, Altman DG. Diagnostic tests 4: likelihood ratios. *BMJ* 2004; 329:168–169
14. Altman DG, Bland JM. Diagnostic tests 3: receiver operating characteristic plots. *BMJ* 1994; 309:188
15. Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. *BMJ* 1994; 308:1552
16. Hulley SB. *Designing clinical research*, 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins, 2007