

Critical appraisal of studies of diagnostic tests

Introduction

We have learned how to quantify the reliability (Chapter 2) and accuracy (Chapters 3 and 4) of diagnostic tests. In this chapter, we turn to critical appraisal of studies of diagnostic tests, with an emphasis on problems with study design that affect the interpretation or credibility of the results. After a general discussion of an approach to studies of diagnostic tests, we will review some common biases to which studies of test accuracy are uniquely or especially susceptible and conclude with an introduction to systematic reviews of studies of diagnostic tests.

General approach

A general approach to critical appraisal of studies of diagnostic tests is to break the study down into its component parts and consider strengths and weaknesses of each, as outlined in Table 5.1. Begin with the research question: is it a question to which you really want to know the answer? Is the test (or history or physical examination finding) being studied one that you have used or could use for your patients? Is the entity being diagnosed one that is important?

Next, consider the study design. All study designs (cross-sectional, case-control, cohort, randomized trial, etc.) have strengths and weaknesses. Watch out for studies of diagnostic tests with a case-control design, in which subjects with the disease are sampled separately from those without the disease. We previously mentioned that the separate sampling of those with and without disease can provide information on test characteristics (sensitivity and specificity) but generally cannot provide information about prior or posterior probability. Other problems with sampling those with and without the disease separately are that the process by which they were classified may have been influenced by the results of the test (“incorporation bias,” discussed in the next section) and studies with this design tend to select subjects in whom true disease

Table 5.1. Step-by-step critical appraisal of studies of diagnostic tests

Study Component	Examples	Issues for consideration
Research question: The question that the study is designed to answer.	<ul style="list-style-type: none"> • How accurate is a bedside test for influenza? • What is the inter-rater reliability of Pap smear readings? 	<ul style="list-style-type: none"> • Is the study question relevant? (Do you care what the answer is?) • Is the test one that you can use for your patients? • Are the outcome variables of interest to you or your patients?
Study design: How subjects were sampled, what variables were measured, and when.	<ul style="list-style-type: none"> • Cross-sectional study • Cohort study • Case-control study • Randomized trial 	<ul style="list-style-type: none"> • Are subjects sampled separately by disease status or test results? • Was the predictor variable measured before, at the same time, or after the outcome variable?
Subjects: How the subjects were identified and selected, and the inclusion and exclusion criteria.	<ul style="list-style-type: none"> • Consecutive patients 6 months to 6 years old admitted with fever and at least 1 of 4 additional symptoms • Women 35–75 years old presenting for routine Pap smear 	<ul style="list-style-type: none"> • Are the subjects (both with and without disease, if sampled separately) representative of those to whom you wish to generalize the results? • If not, in what direction will differences alter the results?
Predictor variable: For studies of test accuracy, the test result. How the test was done.	<ul style="list-style-type: none"> • QuickVue test result obtained by trained and certified nurses • Results of Pap smears read by 4 cytology technicians and 5 cytologists at 2 academic medical centers 	<ul style="list-style-type: none"> • How difficult is it to do the test? • If it requires skill or training, will the skill and training of those doing the test in your setting be similar to what was studied?
Outcome variable: For studies of test accuracy, the presence or absence of disease. For prognostic tests, the occurrence of a bad outcome, like relapse or death.	<ul style="list-style-type: none"> • Influenza diagnosed by viral culture or 2 consecutive positive Polymerase Chain Reaction (PCR) tests • Kappa statistic • 5-year all-cause mortality 	<ul style="list-style-type: none"> • Is the gold standard really gold? • Is it clinically relevant – i.e., how well does the gold standard correlate with what you really want to know? • Were those measuring it blinded to results of the test being evaluated?

(continued)

Table 5.1 (continued)

Study Component	Examples	Issues for Consideration
<p>Results and analysis: What the authors found at the end of the study. May include whether results vary in different subgroups of patients or by center or examiner.</p>	<ul style="list-style-type: none"> • For studies of reproducibility: kappa, Bland-Altman plots, etc. • For accuracy studies: sensitivity, specificity, predictive value, LRs, AUROC curve, all with confidence intervals 	<ul style="list-style-type: none"> • Were all the subjects analyzed, or were some (e.g., those with ambiguous results or some with negative results) excluded? • If sensitivity, specificity, or LRs were reported for ordinal or continuous tests, were standard cut-offs or intervals used? • If predictive value is reported, is the prevalence in the study representative of your patient population? • Were confidence intervals for relevant quantities included?
<p>Conclusions: The authors' conclusions regarding the research question, based on the results of the study.</p>	<ul style="list-style-type: none"> • Authors' conclusions often go beyond estimates of test accuracy or reliability and address whether or when the test is worth doing. 	<ul style="list-style-type: none"> • Do you believe the results are true in the population studied (internal validity)? • Do you believe they apply to patients you see (external validity)? • Did the test provide new information, beyond what was available without the test? • Given your estimates of prior probability and the costs of false-positive and false-negative results, do you agree with authors' conclusions on indications for the test?

status is more clear-cut than it is in clinical practice (“spectrum bias,” discussed later in this chapter).

As in any clinical research study, the extent to which findings can be generalized depends on how the subjects were sampled for the study. Are the prevalence and

severity of the disease (and of diseases that could be confused with it) similar to those in your clinical population? If not, in what direction would the differences change the results? If the study is of reproducibility, consider the sampling and representativeness of both the patients on whom the test was done and the people doing the test.

The predictor variable for studies of diagnostic tests is most often the test result. In appraising a study, it is important to look at exactly how the test was done. Are there factors, such as freshness or preparation of the sample, skill of those obtaining the sample, those doing or interpreting the test, or the quality of the equipment used, that might affect the results? If so, in what direction would results be affected?

The outcome variable is the outcome that the predictor variable is supposed to predict or affect. It may be the presence or absence of disease (determined by a gold standard), or the occurrence of a bad outcome, such as death or relapse of disease. Ideally, measurements of the outcome variable should be made by people blinded to the result of the predictor variable, although as will be discussed later, this is not always practical. Because the ultimate goal of testing is to improve outcomes by enhancing decision making, the best studies of diagnostic tests (admittedly few and far between) are those that compare outcomes in patients randomized to receive or not receive the test, as has been done for BNP in emergency department patients presenting with shortness of breath (Mueller et al. 2006) and pulmonary artery catheters in patients with acute lung injury (Wheeler et al. 2006).

The results of the study are what the authors found, generally presented using the parameters described in Chapters 2–4, including kappa, sensitivity, specificity, ROC curves, etc. Because there is a trade-off between sensitivity and specificity, watch for studies that only highlight one or the other; any test can be 100% sensitive if there is no lower limit of specificity and vice versa. These parameters should be accompanied by confidence intervals to quantify the precision of the estimates. We will discuss confidence intervals at length in Chapter 11; for now, we will just say that they show the range of values consistent with the study results.

Finally, consider the conclusions of the study and decide whether they are justified by the results, given any limitations in the other components of the study. If a study concludes that a test is useful, pay particular attention to limitations in its methods that would tend to make the test look falsely good. On the other hand, studies that conclude a test is not useful should be scrutinized for biases that will make the test look worse in the study than it might be in practice.

Conclusions about usefulness of tests often require information and judgments that go far beyond the results of the study. For example, a study that estimates only sensitivity and specificity may conclude that a test is or is not worth doing when the answers to that question depend on the prior probability of the disease, the cost of the test, and the consequences of false-negative and false-positive results, all of which may vary in different populations and may depend on which decision the test is supposed to help with. History and physical examination findings, for example,

Box 5.1: Example of step-by-step appraisal of a diagnostic test study

The research question for a recent study (Nassar et al. 2006) was: what is the diagnostic accuracy of clinical examination for detection of noncephalic presentation in late pregnancy? "Noncephalic" means the fetal head is not pointed down. Diagnosing this prior to the onset of labor is important to help decide whether to try an "external version" (pushing on the uterus to turn the fetus) to avoid a breech delivery.

The study design was cross-sectional.

The subjects were 1,633 women with a singleton pregnancy at 35–37 weeks gestation attending antenatal clinics at a Women and Babies hospital in Australia. This represented 96% of the 1,707 eligible women who were approached.

The predictor variable was the clinical examination for presentation by one of more than sixty clinicians (residents or registrars 55%, midwives 28%, obstetricians 17%), with results classified as cephalic or noncephalic.

The outcome variable was presentation determined by portable handheld ultrasonography by operators blinded to the results of the clinical examination.

The results included noncephalic presentation in 130 women (prevalence = 8%), sensitivity $91/130 = 70\%$ (95% CI 62% to 78%), and specificity $1429/1503 = 95\%$ (95% CI 94% to 96%).

The conclusion in the abstract was: "Clinical examination is not sensitive enough for detection and timely management of noncephalic presentation." However, in the text, the authors make the point that costs, resource availability, and feasibility need to be considered before introducing routine ultrasonography to assess fetal presentation.

Critical appraisal: This is a nice, clearly reported study.¹ The research question is relevant and the study design is appropriate. The subjects are reasonable: we do not know of any reason why the fetal presentation of these women from Australia should be harder to determine than that of women elsewhere. Because the predictor variable was a clinical finding, and most examinations were done by residents and registrars, the level of skill of the examiners is relevant to generalizability; some readers of the article thought they could do better than the clinicians studied. In response to letters to the editor, the authors provided a breakdown of the sensitivity of different examiners, and there was no clear evidence of improvement with increasing level of experience. However, the authors acknowledge that accuracy could be improved by ongoing training and feedback. A relevant point made in a letter titled "*Doctors do not Dichotomize*" is that the design of the study did not allow the examiners to include a category for "can't tell." This relates to the decision that the physical examination is supposed to guide. Decisions about external version and operative delivery clearly require a more accurate test. But if the decision is whether to check the presentation with a sonogram, perhaps the examination could be made sufficiently sensitive by including a "not sure" category and then counting "not sure" as a positive test.

may not be sufficiently accurate to determine treatment, but may be sufficient to tip the balance toward or away from additional tests. An example of this is provided in Box 5.1.

¹ Except for the reported 95% confidence intervals, most of which are wrong.

Important biases for studies of diagnostic test accuracy

The general approach outlined above should help you appraise most clinical research studies of diagnostic tests. In this section, we turn to potential problems that are either unique or particularly important to studies of diagnostic test accuracy. In such studies, the test for which accuracy will be determined is called the “index test,” and, as described in Chapter 3, the patient’s true disease status is determined by the so-called gold standard. Four important biases in studies of diagnostic test accuracy are incorporation bias, verification bias, double gold standard bias, and spectrum bias. If a study estimates the accuracy of a dichotomous test by reporting sensitivity and specificity, these four biases affect the estimates of sensitivity and specificity in different directions, as summarized in Table 5.2.

Incorporation bias

In order for a study of a diagnostic test to be valid, the index test must be compared to an independent gold standard. If the gold standard is in any way subjective, it must be applied by observers blinded to the results of the index test. It is surprisingly common for the index test to be incorporated into the gold standard, leading to falsely high estimates of both sensitivity and specificity. For example, in their review of 109 studies of diagnostic technologies for acute cardiac ischemia, Lau et al. (2001a; 2001b) found that the studies rarely defined “unstable angina” and generally accepted the clinician’s diagnosis. This means that the diagnostic accuracy (sensitivity and specificity) of different technologies to diagnose unstable angina were probably overestimated due to incorporation bias, because the clinician’s diagnosis of unstable angina is likely to have been at least partly based on the results of the diagnostic technologies being evaluated.

Obviously, if you are assessing a test’s ability to detect disease, and you define disease partly by a positive test, the test is likely to look good. This does not mean that studies susceptible to incorporation bias are useless. Sometimes, in spite of the possibility of this bias, a test still does not look very good, in which case the results can be believed. And sometimes the key to interpreting such studies is simply to understand that they are answering a slightly different question: how well does the test predict the diagnosis of a disease? These studies thus may end up addressing questions about doctors’ understanding of the disease, rather than about the disease itself.

Verification bias

In a study of a diagnostic test, application of the gold standard should not depend on the result of the index test being evaluated. “Verification bias” (also known as referral or work-up bias) occurs when people who are positive on the index test are more likely to get the gold standard, and only those who receive the gold standard are included in the study. Consider a study evaluating the usefulness of ankle swelling to predict a fracture on x-ray in patients with ankle injuries. X-rays are less likely to be ordered in patients with no swelling, and the study includes only those with

Table 5.2. Biases in Studies of Diagnostic Test Accuracy

Bias Type	General description	Specific situations	Sensitivity is falsely. . .	Specificity is falsely. . .
Incorporation bias	Classification of disease status partly depends on the results of the index test. Gold standard incorporates the index test.		↑	↑
Verification bias	Patients with positive index tests are more likely to get the gold standard, and only patients who get the gold standard are included in the study.		↑	↓
Double gold standard bias	Patients with a positive index test are more likely to receive one (often invasive) gold standard, whereas patients with a negative index test are more likely to receive a different gold standard (often clinical follow-up). Bias occurs only if there is a subgroup where the two gold standards give different answers.	For disease that can resolve spontaneously.	↑	↑
		For disease that becomes detectable during the follow-up period.	↓	↓
Spectrum bias	Spectrum of disease and nondisease differs from clinical practice. Sensitivity depends on spectrum of disease. Specificity depends on spectrum of nondisease or of diseases that might mimic the disease of interest.	When disease is skewed toward higher severity than in clinical practice – “sickest of the sick.”	↑	NA
		When nondisease is skewed toward greater health – “wellest of the well.”	NA	↑

NA = Not Affected

x-rays. This design decreases the numbers of subjects with negative tests (no swelling), both with and without disease (fracture), as represented in cells (c) and (d) in Figure 5.1. Verification bias will tend to increase the sensitivity ($a/[a + c]$) and decrease the specificity ($d/[d + b]$) compared with what would have been obtained if the gold standard (x-rays) had been applied regardless of the index test result (presence of ankle swelling). Box 5.2 provides a numerical illustration of verification bias.

	Fracture	No Fracture
Ankle Swelling	a	b
No Ankle Swelling	c↓	d↓

Figure 5.1 How verification bias leads to overestimation of sensitivity and underestimation of specificity by lowering numbers in cells (c) and (d).

Box 5.2: Numerical example of verification bias

We examine two hypothetical studies of ankle swelling as a predictor of fractures in patients with ankle injuries. The first study is a consecutive sample of 200 patients. In this study, all patients presenting to the emergency department with ankle injuries get x-rays, regardless of swelling. The sensitivity and specificity of ankle swelling are 80% and 75%, as shown in the following table:

	Fracture	No Fracture
Swelling	32	40
No Swelling	8	120
Total	40	160
	Sensitivity = $32/40 = 80\%$	Specificity = $120/160 = 75\%$

The second study is a **selected** sample, in which only half the patients without ankle swelling are x-rayed. Thus, the numbers in the “No Swelling” row will be reduced by half. This raises the apparent sensitivity from 32/40 (80%) to 32/36 (89%) and lowers the apparent specificity from 120/160 (75%) to 60/100 (60%), as shown in the next table:

	Fracture	No Fracture
Swelling	32	40
No Swelling	4	60
Total	36	100
	Sensitivity = $32/36 = 89\%$	Specificity = $60/100 = 60\%$

The verification bias that arises by excluding patients without swelling causes a false increase in sensitivity and decrease in specificity.²

Double gold standard bias

A bias related to verification bias occurs when two distinct gold standards exist and the results of the index test affect which is applied. People who are positive on the index test are more likely to get one gold standard (such as a surgical procedure), whereas people who are negative on the index test are more likely to get a different gold standard (such as clinical follow-up). We call this form of bias “double gold standard bias.”³ In some cases, a double gold standard is unavoidable for ethical or practical reasons. For example, a biopsy can be used as the gold standard in people

² This is a bit of an oversimplification, because ideally the subjects with no ankle swelling who do not receive x-rays are not a random sample, but rather a group (judged to be) at low risk based on other findings. Hence, the number of false negatives in this example might be closer to 6 or 7, rather than 4, resulting in a smaller effect on sensitivity than shown here.

³ Others call it “referral bias” or “verification bias” and do not distinguish this type of bias from what we called verification bias in the previous section.

with a positive result on a screening test and is hard to justify in those with negative results. But this application of different gold standards depending on the result of the test being studied can introduce problems.

Double gold standard bias is a common problem with cancer screening tests. We will see in Chapter 6 (on screening tests) that many cancers are clinically harmless; they can either resolve spontaneously or just sit there and never cause the patient any problem. If we look harder for cancer only in those whose screening test is positive, we will take credit for getting the right answer on these more benign cancers, no matter what answer the screening test gives. If these patients have a positive screening test and get a biopsy, they are counted as true positives; if they have a negative screening test and never get sick, they are counted as true negatives. Both sensitivity and specificity are falsely increased. In Chapter 6, we will show how this not only makes the test appear to be more accurate (our topic here), but also can make the test appear to reduce mortality among people with the disease. In that context, we will refer to this problem of detecting disease that will never cause clinical problems as “pseudodisease.”

The other possibility is that disease could be missed by the first (invasive) gold standard, but nonetheless detected on follow-up. This could occur if the disease was either not present or not detectable initially, as might occur with a fast-growing tumor that could become detectable and lead to symptoms in a short time. In that case, the screening test will always appear to give the wrong answer. If the test is initially positive, and the patient is referred for the invasive gold standard, the test will look like a false positive because the disease has not yet occurred or is not yet detectable by the gold standard. If the test is negative, the patient will be followed, the tumor will present with symptoms, and the test will be considered falsely negative.

With double gold standard bias, the degree of distortion of sensitivity and specificity depends on how closely correlated the test result is with the choice of which gold standard to use, and on the how often the two gold standards give different answers (which depends on the natural history of the disease). Box 5.3 gives a worked example of this type of bias for a disease that might resolve spontaneously.

Spectrum bias

Definition and explanation

The best studies of diagnostic tests are those that replicate the conditions of clinical practice, that is, those in which the disease status of the subjects is not known at the outset. Any test can be made to look good if it only needs to distinguish between the very sick and the very well. “Spectrum bias” is the name for the bias that occurs if the subjects for a study of a diagnostic test did not have a reasonable spectrum of the condition being tested for and of the “nondisease” that may mimic it.

We warned you in Chapter 1 that the assumption that disease was dichotomous is an oversimplification, and that in real life, diseased and nondiseased populations may be heterogeneous. In fact, we can be a bit more specific: sensitivity (or, for nondichotomous tests, the distribution of test results in the diseased group) will depend on the spectrum of disease, and specificity (or the distribution of results in

Box 5.3: Numerical example of double gold standard bias

In a study of ultrasonography to diagnose intussusception (a telescoping of the intestine upon itself) in young children (Eshed et al. 2004), all children with a positive ultrasound scan for intussusception received a contrast enema (Gold Standard #1), whereas the majority of children with a negative ultrasound were observed in the emergency department (Gold Standard #2). The results of the study are shown below:

	Intussusception	No Intussusception
Ultrasound+	37	7
Ultrasound–	3	104
Total	40	111
	Sensitivity = $37/40 = 93\%$	Specificity = $104/111 = 94\%$

The 104 subjects with a negative ultrasound listed as having “No Intussusception” actually included 86 who were followed clinically and did not receive a contrast enema. If about 10% of these subjects (i.e., 9 children) actually had an intussusception that resolved spontaneously but would still have been identified if they had a contrast enema, and all subjects had received a contrast enema gold standard, those 9 children would be considered false negatives rather than true negatives, with a resulting sensitivity of $37/49 = 76\%$ and specificity of $95/102 = 93\%$, as shown below:

	Intussusception	No Intussusception
Ultrasound+	37	7
Ultrasound–	$3 + 9 = 12$	$104 - 9 = 95$
Total	49	102
	Sensitivity = $37/49 = 76\%$	Specificity = $95/102 = 93\%$

Thus, compared with the single gold standard of the contrast enema, the double gold standard leads to higher estimates of both sensitivity and specificity because it counts as true negatives some subjects who would be considered false negatives by the contrast enema.

Now consider the thirty-seven subjects with positive ultrasound scans, who had intussusception based on their contrast enema. Suppose about 10% (i.e., 4) of those intussusceptions would have resolved spontaneously if given the chance. Then, if the single gold standard were clinical observation, four children considered true positives by the contrast enema would become false positives, with a small decrease in specificity from 93% to 90%. The loss of these four true positives also decreases sensitivity a little, from 93% to 92%. Thus, compared with the single gold standard of clinical follow-up, the double gold standard again leads to higher estimates of both sensitivity and specificity because it counts as true positives some subjects who would be considered false positives by clinical follow-up.

	Intussusception	No Intussusception
Ultrasound+	$37 - 4 = 33$	$7 + 4 = 11$
Ultrasound–	3	104
Total	36	115
	Sensitivity = $33/36 = 92\%$	Specificity = $104/115 = 90\%$

Thus, for spontaneously resolving cases of intussusception, the ultrasound scan will appear to give the right answer whether it is positive or negative, increasing both its apparent sensitivity and specificity.

the nondiseased group) will depend on the spectrum of nondisease. A study that disproportionately includes patients with more severe disease will often have a falsely high sensitivity, whereas a study in which the patients without the disease are very healthy (or do not have anything resembling the target disease) will give a falsely high specificity. Conversely, sensitivity and specificity will be lower if patients with less severe disease are to be distinguished from patients with other, similar diseases. This comes as a bit of bad news for those who were hoping that they could begin tabulating a list of LRs for different results of various tests, because LRs may vary in different patient populations, depending on the typical severity of disease in that population, as well as the distribution and severity of other conditions that could mimic the disease in the nondiseased group.

As an example of spectrum bias, suppose you are interested in LRs for the erythrocyte sedimentation rate (ESR) for diagnosing appendicitis in patients with abdominal pain. The LR for a particular ESR result is $P(\text{result}|\text{appendicitis})/P(\text{result}|\text{no appendicitis})$. But $P(\text{result}|\text{no appendicitis})$ clearly depends on what the patients who do not have appendicitis actually do have. A study of the ESR in young women with abdominal pain who are at risk of acute salpingitis, a disease associated with high values of the ESR, will give different LRs from a study in children or in men. The distribution of ESRs in the no appendicitis groups will differ, even if the distribution of ESRs in subjects with appendicitis is the same.

Prevalence of disease, spectrum bias, and nonindependence of test characteristics

In this section, we show how changes in the spectrum of disease and nondisease may vary with disease prevalence. In previous chapters, we assumed that test characteristics (sensitivity, specificity, and LRs) do not vary with the prevalence of disease. When differences in disease prevalence are associated with differences in disease (and nondisease) severity, this assumption may be incorrect.

For example, in the United States, which is an area of relatively low prevalence of iron deficiency, possible tests for iron deficiency anemia, such as pallor on physical examination, a low hematocrit, or low mean corpuscular volume, are likely to have lower sensitivity than in Africa, where the prevalence of iron deficiency anemia is higher. This is because the severity of iron deficiency in Africa is likely to be greater, so that the African patients with iron deficiency will be more iron deficient, and the tests above are more likely to be abnormal in those with the disease (i.e., have higher sensitivity).

The same considerations apply to specificity, except that in this case, the “nondiseased” populations in the two regions are likely to differ. Specificity does not depend on the prevalence of the disease, but it does depend on the prevalence of diseases that can be confused with the disease in question. Specificity of the tests or findings for iron deficiency anemia could be lower in Africa because other diseases (like malaria or hookworm) that might make children anemic (and therefore pale) are more common there, and “tests” like pallor will be abnormal with these other diseases as well.

In the iron deficiency example, sensitivity increases with prevalence, because greater prevalence is associated with greater disease severity. But the opposite could

also be true. If the (apparent) prevalence of disease depends on the level of surveillance, then an area with high prevalence might also be an area where the average severity of disease is less, because the additional cases picked up by closer surveillance are likely to be milder than those that presented with symptoms. In that case, sensitivity of some tests could be lower in the high-prevalence area. For example, consider the sensitivity of digital rectal examination for detecting prostate cancer. In a place where prostate-specific antigen screening is widespread, the prevalence of prostate cancer would be higher, and the population of prostate cancer patients would presumably include many more in whom no tumor was palpable, leading to a lower apparent sensitivity of physical examination.

When you read a paper that tries to measure sensitivity and specificity, think about whether the spectrums of disease and nondisease in the study subjects are similar to those in patients you are likely to see. As a general rule, the more severe the disease in the patients who have it, the greater the sensitivity, whereas the healthier the “nondiseased” group, the greater the specificity.

Spectrum bias as a cause of test nonindependence

We just discussed examples where diseases had different prevalences and different test characteristics in different populations due to what we have called spectrum bias. In this section, we further generalize this point, showing that the populations in which test characteristics differ can be defined not only by time and place, but also by results of history, physical examination, or other tests. In this case, we see that spectrum bias is one reason for another problem, which we will encounter again in Chapter 8: nonindependence of tests.

In a classic article on spectrum bias (Lachs et al. 1992), the authors studied the leukocyte esterase and nitrite⁴ on a urine dipstick as predictors of a urinary tract infection (UTI), defined as a urine culture with greater than 10^5 bacteria/mL. They divided the 366 adults subjects in the study into those with high (>50%) and low ($\leq 50\%$) prior probability of UTI, based on the signs and symptoms recorded by clinicians before obtaining the urine dipstick result, which was classified as positive if either the leukocyte esterase or nitrite was positive. They found marked differences in both sensitivity and specificity in 2 groups defined by prior probability (Table 5.3).

How can we account for these results? One possibility is that the patients with higher prior probability of UTI had more severe UTIs. Thus, their UTIs were easier to diagnose, and sensitivity was higher. Similarly, maybe some of those with high

Table 5.3. Differences in test characteristics of the urine dipstick in women at high and low prior probability of UTI, based on signs and symptoms (from Lachs et al. 1992)

	Sensitivity	Specificity	LR+	LR–
High Prior Prob.	92%	42%	1.6	0.19
Low Prior Prob.	56%	78%	2.5	0.56

⁴ The leukocyte esterase is a test for white blood cells in the urine; the nitrite test is for bacteria.

prior probability of UTI had urine cultures with just under 10^5 bacteria/mL. In that case, their lack of UTI would be harder to diagnose, leading to a lower specificity.

A related possibility is that, distinct from disease severity, the index test (dipstick) is measuring something that has already been measured by another test: in this case, the clinical assessment based on signs and symptoms. Perhaps there is a subset of patients with UTI who have inflammation of the lower urinary tract. If this inflammation is what leads to both pain with urination and abnormal urine tests, then, in a way, painful voiding (obtained from the history) is measuring the same aspect of the disease (urinary tract inflammation) as the inflammation identified with dipstick for leukocyte esterase. In that case, we would expect the two tests – clinical assessment of dysuria and a dipstick positive for leukocyte esterase – to be nonindependent, as discussed in Chapter 8. Once you know that a woman has dysuria, you do not learn as much from finding out that she has leukocyte esterase on her urine dipstick, and vice versa. Nonindependence tends to make the sensitivity of the index test appear better, whereas the specificity will generally decrease. The results in Table 5.3 are quite consistent with this explanation.

Overfitting

“If you torture data sufficiently, it will confess to almost anything.”

– Fred Menger

“Overfitting” refers to use of models that are made overly complicated in order to fit the data that have been collected. It is analogous to gerrymandering of congressional districts, which provides perhaps the best way to visualize the problem (Fig. 5.2). Overfitting is mainly a problem when a combination of tests is chosen from many candidate tests to identify a disease or predict a prognosis, so we will discuss it in more detail in Chapter 8, which covers multiple tests and multivariable-decision rules. If you develop a prediction rule by choosing the best combination of tests and

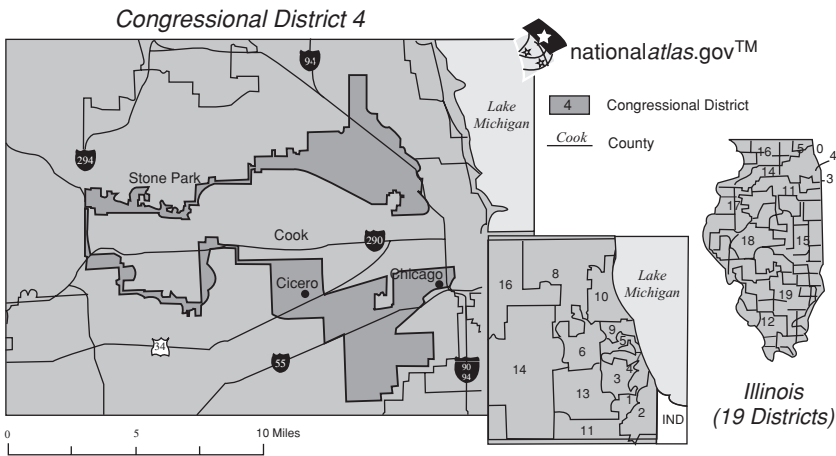


Figure 5.2 Gerrymandering, in which legislators choose their voters, rather than vice versa, provides a visual image of overfitting. This is the 4th Illinois Congressional District. (From: http://nationalatlas.gov/printable/images/pdf/congdist/IL04_110.pdf. Accessed 7/25/08.)

findings from a large number of such variables investigated, you can often develop what might look like a good prediction rule (especially if the sample size is small). But this takes advantage of chance variations in the data that will probably disappear if others try to validate the prediction rule on a new dataset.

A mild form of overfitting bias can occur in studies of a single continuous test when an “optimal” cut-off value is chosen to separate positive from negative results *after* looking at the data. In Chapter 4, we pointed out one problem with making a continuous or multilevel test dichotomous by choosing a fixed cut-off value: this practice discards information by equating slightly abnormal results with extremely abnormal results. Another problem is that the fixed cut-off is chosen to minimize some kind of cost function, often the total number of misclassifications.⁵ However, the cost-minimizing cut-off for one group of patients will not be the same as for another group of patients, both because the prevalence of disease will differ between the two datasets and because of chance variations in the distributions of test results.

Systematic reviews of diagnostic tests

Clinicians wishing to practice evidence-based diagnosis are often faced with a problem: when we look in the literature to find values for sensitivity, specificity, LRs, or other test characteristics, we find studies with varying results. Or, perhaps more commonly, we look in a textbook chapter or a typical review article and find statements like “the XYZ test has sensitivity from 63% to 100% and specificity from 34% to 88%,” followed by a string of references. For many tests, the range of reported estimates for sensitivity and specificity is so large that the resulting LRs could be consistent with either an informative or useless test. What do we do?

One approach is to pull all of the articles and critically appraise them, using the general approach you have learned in this book or by using a checklist for diagnostic test study quality (Whiting et al. 2006; Bossuyt et al. 2003; Straus et al. 2005). However, most of us do not have time to do this, and even if we did, it would be hard to synthesize the results. To address this problem, systematic reviews of diagnostic tests are starting to appear, although methods and standards for them are still developing (Deeks 2001; Pai et al. 2004; Mallett et al. 2006). As with other systematic reviews, systematic reviews of diagnostic tests should have four key features: 1) a systematic and reproducible approach to finding and selecting the relevant studies; 2) a summary of the results of each of the studies; 3) an investigation seeking to understand any heterogeneity between the studies; and 4) a summary estimate of results, if appropriate.

One difference between systematic reviews of studies of diagnostic tests and other systematic reviews is that reviews of diagnostic tests commonly attempt to estimate two parameters (sensitivity and specificity), rather than one (e.g., a risk ratio). These two parameters are related: as one goes up, the other often goes down, especially

⁵ Choosing a cut-off that minimizes total number of misclassifications assumes that misclassifying a normal individual as diseased is just as bad as misclassifying a diseased individual as normal. It is often the case that it is much worse to misclassify a diseased individual as normal than vice versa.

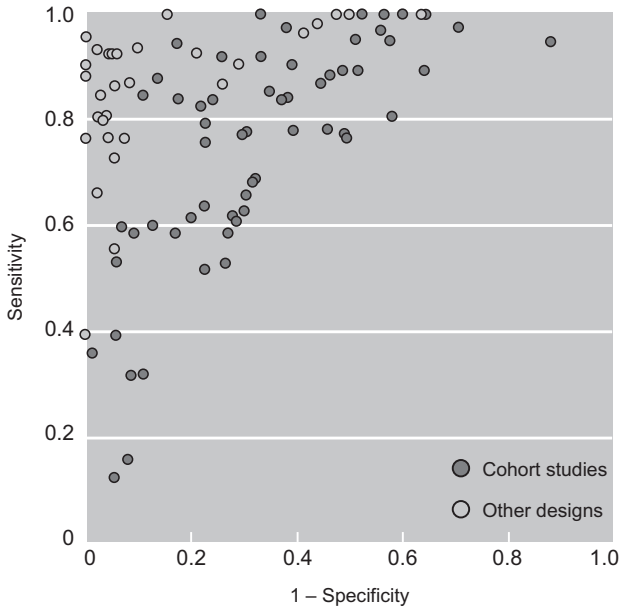


Figure 5.3 Studies of MRI for the diagnosis of multiple sclerosis. Cohort studies (solid circles) produced lower estimates of accuracy than studies using other designs. (From Whiting et al. 2006, used by permission)

if one of the reasons for differing estimates is a difference in the cut-off (or some underlying hidden threshold⁶) used to define a positive result. One approach to this is to plot the sensitivity and specificity obtained from different studies on the same axes used to draw an ROC curve (Sensitivity vs. 1 – Specificity; see Chapter 4). This gives a visual representation of the extent to which differences in reported sensitivity and specificity could be the result of differences in the threshold for a positive test.

It is particularly helpful if characteristics of the studies help explain the location of their points on the ROC plane. For example, Figure 5.3 is taken from a systematic review of magnetic resonance imaging for the diagnosis of multiple sclerosis (Whiting et al. 2006). It shows that studies with a cohort design tend to have lower accuracy estimates: almost all of the points in the upper left corner of the ROC plane (corresponding to the highest estimates of accuracy) came from studies with weaker designs.

There are methods of drawing Summary Receiver Operating Characteristic (SROC) curves through groups of studies, plotted as in Figure 5.3 (Littenberg and Moses 1993; Macaskill 2004). Generally, this is most appropriate if studies with similar designs, in similar populations, and/or with similar gold standard definitions are grouped together, making the results more homogeneous. Possible reasons for heterogeneity in accuracy estimates can also be investigated statistically, using analyses in which each study constitutes an observation, characteristics of the study (like design, blinding, spectrum of disease, etc.) are the predictor variables, and the results of the study are the outcomes. Whether the review uses these sophisticated methods,

⁶ Even apparently dichotomous tests can have different thresholds. For example, some observers may call a urine dipstick positive for only a minimal color change, while others may require the color change to be more definite.

or simply identifies and summarizes studies, your goal as the reader of a systematic review of a diagnostic test is to obtain estimates of test characteristics based on the most valid studies, in populations and under testing conditions that best duplicate the conditions under which you would be using the test.

Making use of biased studies

A key step in reading journal articles is not just to identify potential biases, but to determine how the biases could affect the conclusions. A problem with some checklists for critically appraising studies of diagnostic tests is that they can lead to rejection of some studies as “flawed,” even though the studies may provide useful information. For example, if a study concludes that a diagnostic test is not useful in a particular situation, and biases in the design of the study would have led to the test looking better than it really is, the study’s conclusion is not undermined. On the other hand, if biases in the study design would tend to make the test look bad, the conclusion that the test is not useful may simply be due to these biases.

For example, if a test distinguishes poorly between people with severe disease and healthy medical students, it is likely to do even worse in patients with a more clinically relevant spectrum of disease and nondisease. Similarly, if a study subject to verification bias still reports that sensitivity is poor, that conclusion is probably valid. In these examples, the key is to notice that the potential bias would make the test look falsely good. On the other hand, consider the study of ultrasonography to diagnose intussusception (Eshed et al. 2004). The ultrasonographers were not the world experts; in fact, many of them were junior radiology residents new to the procedure. If the authors had reported poor accuracy, the generalizability of the results to setting with more experienced ultrasonographers would have been questionable. However, since the reported accuracy was good, this lack of ultrasonographer experience is of less concern.

Summary of key points

1. As with any clinical research study, critical appraisal of a study of a diagnostic test starts with consideration of the research question, study design, study subjects, predictor variable, and outcome variable.
2. In studies of diagnostic test accuracy, the predictor variable is typically the result of the index test (the test being studied) and the outcome variable is the patients true disease state as determined by the gold standard.
3. Studies evaluating diagnostic tests are susceptible to particular biases.
4. Incorporation bias occurs when classification of the patient as diseased depends partly on the result of the index test. It biases both sensitivity and specificity up.
5. Verification bias occurs when patients who are positive on the index test are more likely to be referred for the gold standard and hence to be included in the study. It biases sensitivity up and specificity down.

6. Double gold standard bias occurs when there are two different gold standards applied selectively based on index test results – for example, an invasive test that is applied when the index test is positive and clinical follow-up that is applied when the index test is negative. If there is a subgroup of patients for whom the invasive test would be positive but the clinical follow-up would be negative (e.g., because of spontaneous resolution of disease), the use of these two gold standards, instead of one or the other, will bias both sensitivity and specificity up. On the other hand, if there are subjects for whom the invasive gold standard is negative, but clinical follow-up is positive (e.g., because of rapidly progressive disease not initially present), both sensitivity and specificity will be biased down.
7. Spectrum bias occurs when the spectrum of disease and nondisease in the study population differs from that in the clinical population in which the test will be used. If the group of patients with the disease has severe disease (“the sickest of the sick”), sensitivity will be biased up. If the group of patients without the disease is very healthy (“the wellest of the well”), specificity will be biased up.
8. When there are multiple studies of the same test, it may be possible to do a systematic review and develop summary estimates of test sensitivity and specificity and to summarize the results using an SROC curve.
9. Even flawed studies of diagnostic tests can be useful as long as the flaws affect sensitivity and specificity in predictable ways.

References

- Bossuyt, P. M., J. B. Reitsma, et al. (2003). “Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative.” *Br Med J* **326**(7379): 41–4.
- Deeks, J. J. (2001). “Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests.” *Br Med J* **323**(7305): 157–62.
- Eshed, I., A. Gorenstein, et al. (2004). “Intussusception in children: can we rely on screening sonography performed by junior residents?” *Pediatr Radiol* **34**(2): 134–7.
- Lachs, M. S., I. Nachamkin, et al. (1992). “Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection.” *Ann Intern Med* **117**(2): 135–40.
- Lau, J., J. P. Ioannidis, et al. (2001a). “Diagnosing acute cardiac ischemia in the emergency department: a systematic review of the accuracy and clinical effect of current technologies.” *Ann Emerg Med* **37**(5): 453–60.
- Lau, J., J. P. Ioannidis, et al. (2001b). *Evaluation of Technologies for Identifying Acute Cardiac Ischemia in Emergency Departments*. Rockville, MD, The Agency for Healthcare Research and Quality. Available at <http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=hstat1.chapter.37233>.
- Littenberg, B., and L. E. Moses (1993). “Estimating diagnostic accuracy from multiple conflicting reports: a new meta-analytic method.” *Med Decis Making* **13**(4): 313–21.
- Macaskill, P. (2004). “Empirical Bayes estimates generated in a hierarchical summary ROC analysis agreed closely with those of a full Bayesian analysis.” *J Clin Epidemiol* **57**(9): 925–32.
- Mallett, S., J. J. Deeks, et al. (2006). “Systematic reviews of diagnostic tests in cancer: review of methods and reporting.” *Br Med J* **333**(7565): 413.
- Mueller, C., K. Laule-Kilian, et al. (2006). “Cost-effectiveness of B-type natriuretic peptide testing in patients with acute dyspnea.” *Arch Intern Med* **166**(10): 1081–7.

- Nassar, N., C. L. Roberts, et al. (2006). "Diagnostic accuracy of clinical examination for detection of non-cephalic presentation in late pregnancy: cross sectional analytic study." *Br Med J* **333**(7568): 578–80.
- Pai, M., M. McCulloch, et al. (2004). "Systematic reviews of diagnostic test evaluations: What's behind the scenes?" *ACP J Club* **141**(1): A11–3.
- Straus, S., W. Rihardson, et al. (2005). *Evidence-Based Medicine: How to Practice and Teach EBM*. New York, Elsevier/Churchill Livingstone.
- Wheeler, A. P., G. R. Bernard, et al. (2006). "Pulmonary-artery versus central venous catheter to guide treatment of acute lung injury." *N Engl J Med* **354**(21): 2213–24.
- Whiting, P., R. Harbord, et al. (2006). "Accuracy of magnetic resonance imaging for the diagnosis of multiple sclerosis: systematic review." *Br Med J* **332**(7546): 875–84.

Chapter 5 Problems: studies of diagnostic tests

1. Consider the following excerpt from the abstract of an article about diagnosing Von Willebrand's disease (Werner et al. 1992):

ABSTRACT

... To determine the best tests to identify patients with von Willebrand disease (vWD), we reviewed the laboratory studies of 24 children with vWD ... The diagnosis of vWD required the presence of a personal and family history of bleeding symptoms and a documented abnormality of vWF [von Willebrand Factor] activity or vWF antigen. vWF activity, vWF antigen, factor VIII procoagulant (factor VIII:c) and blood type were [also] determined in 104 symptom-free children.

... The vWF activity, vWF antigen, and factor VIII:c were abnormal in 79%, 58%, and 33% [of vWD patients], respectively. Receiver-operating-characteristic analysis showed the vWF activity to be superior to either the vWF antigen or factor VIII:c in establishing the diagnosis of vWD. The combination of the activity, bleeding time, and partial thromboplastin time successfully identified 92% of the patients as abnormal. Determination of vWF activity should be included routinely in the evaluation of hemostasis in children with symptomatic disease.

- a) What were the index tests in this study?
 - b) How was disease status determined – that is, what was the gold standard?
 - c) What parameter (sensitivity, specificity, predictive value, etc.) are the percentages (79%, 58%, and 33%) reported in the second paragraph?
 - d) What is the problem with the study design and how would it affect the estimates in part c?
2. Below are excerpts from the abstracts from two studies that gave discrepant answers to the question: Is urine microscopy needed in patients whose urine dipstick is negative?

#1. Morrison, M.C., and G. Lum. (1986). “Dipstick testing of urine: can it replace urine microscopy?” *Am J Clin Pathol* **85**(5): 590–4.

Abstract. One thousand consecutive urine specimens were studied to assess the sensitivity of a commercially available dipstick to predict the presence or absence of microscopic abnormalities . . . The Chemstrip-9™ had a sensitivity of 82%, specificity of 42%, and a false negative rate of 36%. Clinical review of patients with false negative results showed that approximately one-third to one-half of these patients had either spinal cord injury or genitourinary problems . . . Our data suggest that in our patient population, we should not eliminate microscopic urine examination based on abnormal dipstick findings.

#2. Hamoudi, A.C., S.C. Bubis, C. Thompson. (1986). “Can the cost savings of eliminating urine microscopy in biochemically negative urines be extended to the pediatric population?” *Am J Clin Pathol* **86**(5): 658–60.

Abstract. The authors determined the value of performing urine microscopy on biochemically negative urine specimens in a pediatric population. Four reactions of the Chemstrip-9™ were used as biochemical indicators, namely, protein, occult blood, leukocyte esterase, and nitrite. Out of 1,016 urine specimens thus studied, 310 were true positive. Eleven specimens reacted biochemically in the absence of significant microscopic findings (false positive), 668 specimens were negative by the Chemstrip-9™ and were either negative microscopically or had less than five white blood cells (WBCs) per high power field (HPF) and were considered true negatives. Twenty-seven specimens had negative biochemical indicators, in spite of positive microscopy . . . The sensitivity of the four parameters for predicting significant microscopy of urinary sediment is 92% and the specificity is 98%. The predictive value of a negative result is 96.1%, and that of a positive result is 96.5%. The authors therefore conclude that urine microscopy is unnecessary in biochemically negative urine specimens from pediatric patients who are asymptomatic for urinary tract disease.

- a) Make 2×2 tables for the Chemstrip-9 compared with urine microscopy for both studies. Put the “disease” (urine microscopy result: abnormal or normal) on the top and the test (urine dipstick) on the left side. (Note: In the Morrison study, 573 of the 1000 specimens had abnormal urine microscopy. Because of rounding, your numbers will only be approximate.)
 - b) What is meant by the “false negative rate of 36%” in the Morrison study? (Is this the standard definition?)
 - c) What are at least three possible explanations for the apparently discrepant results of these two studies? Which do you think are most likely to explain the results? (Hint: Use the structure suggested at the beginning of the chapter.)
3. Women presenting to the emergency department with abdominal pain and a positive pregnancy test may have an ectopic pregnancy (about 10%), an

abnormal intrauterine pregnancy (about 30%), or a normal intrauterine pregnancy (about 60%). An excerpt from the abstract of a study on this topic is reprinted below (Marill et al. 1999):

The objectives of this study were to determine the optimal cutoff value and utility of a single serum beta human chorionic gonadotropin hormone (HCG) level in assessing the likelihood of ectopic pregnancy. A retrospective chart review was performed at an urban county hospital. The optimal cutoff value was determined by comparing all available patients diagnosed with ectopic pregnancy and ***patients diagnosed with threatened abortion in the Emergency Department who subsequently delivered a baby at the same hospital . . . [emphasis added]***

The “patients diagnosed with threatened abortion . . . who subsequently delivered a baby at the same hospital” were patients who presented with abdominal pain but ultimately turned out to have a normal intrauterine pregnancy. These patients tend to have higher HCGs than patients with abnormal intrauterine (non-ectopic) pregnancies. No women with abnormal intrauterine pregnancies were included in the non-ectopic study sample. The authors found the sensitivity of an HCG <40,000 mIU/mL for ectopic pregnancy was 99%, and the specificity was 85%. (That is, 85% of the women who subsequently delivered a baby at the same hospital had an HCG $\geq 40,000$ mIU/mL.)

- a) Is spectrum bias a potential problem in this study? Why or why not?
 - b) Is the sensitivity estimate too high, too low, or about right?
 - c) Is the specificity estimate too high, too low, or about right?
4. The abstract of the study of ultrasound to diagnose intussusception summarized in Box 5.3 is excerpted below (Eshed et al. 2004).

BACKGROUND. Ultrasonography (US) is an important tool in the screening and diagnosis of patients with suspected intussusception.

MATERIALS AND METHODS. Between January 1999 and February 2003, 151 patients with suspected intussusception underwent screening US. The mean age of the patients was 13.8 months . . .

RESULTS. Sixty-five patients had both US and air enema. Forty-four patients had a positive US result; 37 (84%) were true positive and 7 (16%) were false positive. Twenty-one patients had a negative US result; 18 (86%) were true negative and 3 (14%) were false negative. Eighty-six patients [with negative ultrasound scans] underwent screening US only and were then kept under observation in the emergency room. They were all diagnosed as having a non-surgical condition [i.e., as true negatives]. The total accuracy rate was 93%, sensitivity was 84%, specificity was 97%, positive predictive value was 93% and negative predictive value was 94% . . .

(For the questions below, assume that the statement, “They were all diagnosed as having a non-surgical condition,” means that none of the 86 patients who only had a negative screening ultrasound and clinical follow-up (but no air enema) were felt to have an intussusception. Also assume that, if air enema was performed at all, it was performed immediately after the ultrasound. Parts (a–c) review material covered in Chapter 3.)

- a) Create a 2×2 table that summarizes the results of the study. (You can check your answer in Box 5.1.)
 - b) Check the authors’ calculations of sensitivity, specificity, and accuracy of ultrasound for diagnosis of intussusception in this study.
 - c) Can positive and/or negative predictive value be estimated from a study with this sampling scheme? If so, what are they? If not, why not?
 - d) The authors did not perform air enemas on all of the children; in some, they just watched them. Name the bias this could cause (using the terminology from this text).
 - e) Assume that intussusception never resolves spontaneously – that is, that nobody who would have had a positive enema (if one were done) would ever have negative clinical follow-up. Also assume no new cases of intussusception develop after the air enema. What would be the effect of the bias you named above on estimates of sensitivity and specificity of ultrasound from this study?
 - f) Now repeat part (e) assuming that intussusception does sometimes spontaneously resolve – that is, that some of those with negative clinical follow-up would have had a positive enema, if one were done. (Maintain the assumption that no new cases develop.)
 - g) Now repeat part (e), only this time assume that intussusception can develop during a short follow-up period after the enema has been done.
5. In Chapter 4, Problem 3, we mentioned the B-natriuretic peptide (BNP) test for congestive heart failure (CHF) and the study (Maisel et al. 2002) that reported that, using a cutoff of 100 pg/mL, the test had a sensitivity of 90%, a specificity of 76%, and an LR(+) of about 3.8. In Chapter 4, we learned that using a single cut-off to dichotomize a continuous test like BNP is unwise; the range of BNP values should be divided into several intervals, each with an associated LR.
- a) Another issue (Schwam 2004) with the study is that some of the dyspnea patients without acute CHF didn’t have any realistic clinical chance of having CHF. They had an obvious diagnosis of asthma, upper respiratory infection, or pneumonia. Also, the authors seem to have excluded from the analysis 72 patients with a history of left ventricular dysfunction but no acute CHF. These excluded patients had higher BNP levels than other non-CHF (D–) patients (average 346 pg/mL vs. 110 pg/mL). How would these problems affect the reported sensitivity, specificity, and LR(+)?
 - b) The gold standard in this study was the consensus diagnosis of two cardiologists who reviewed the patient’s chart, including medical history, ECG, chest x-ray, and follow-up studies. The two cardiologists were blinded to BNP and to the

emergency department diagnosis. Ignoring BNP for the moment, it turned out that increased heart size on chest x-ray had high sensitivity and specificity for CHF. Do you think these estimates of sensitivity and specificity are biased? If so, name the bias and how it affects the estimates.

6. Kharbanda et al (2005) reported sensitivity and specificity of several history and physical findings to diagnose appendicitis among children presenting to an urban emergency room. Children were included if they underwent surgical consultation for possible appendicitis. The “gold standard” was pathology for patients who had an appendectomy and phone follow-up for those who did not have surgery. The finding “pain with percussion, hopping or cough” had sensitivity of (only) 78%. Indicate whether the following statements are true or false and explain your answer.
- The sensitivity may be falsely low because children who could hop around without pain would be less likely to receive surgical consultation, and hence would be under-represented in the study (verification bias).
 - The sensitivity may be falsely low because of double gold standard bias: the gold standard was different for those who did and did not receive surgery, and a child who could hop without pain would be unlikely to receive surgery, giving what was probably a mild case of appendicitis time to resolve on its own.

References for problem set

- Eshed, I., A. Gorenstein, et al. (2004). “Intussusception in children: can we rely on screening sonography performed by junior residents?” *Pediatr Radiol* **34**(2): 134–7.
- Maisel, A. S., P. Krishnaswamy, et al. (2002). “Rapid measurement of B-type natriuretic peptide in the emergency diagnosis of heart failure.” *N Engl J Med* **347**(3): 161–7.
- Marill, K. A., T. E. Ingmire, et al. (1999). “Utility of a single beta HCG measurement to evaluate for absence of ectopic pregnancy.” *J Emerg Med* **17**(3): 419–26.
- Schwam, E. (2004). “B-type natriuretic peptide for diagnosis of heart failure in emergency department patients: a critical appraisal.” *Acad Emerg Med* **11**(6): 686–91.
- Werner, E. J., T. C. Abshire, et al. (1992). “Relative value of diagnostic studies for von Willebrand disease.” *J Pediatr* **121**(1): 34–8.