

Multiple tests and multivariable decision rules

Introduction

In Chapter 3, when we introduced dichotomous tests, the LR of a positive result [LR(+)], and the LR of a negative result [LR(−)], we also made a point of distinguishing between prevalence and prior probability. Recall that prior probability is the more general term. The prior probability is equal to the prevalence of the disease in the population only when we do not know anything else about the patient. This is often the case for screening tests applied to large populations without obtaining information on individuals that allows differentiation between them. Although we tend to focus on laboratory or imaging tests, any new information about the patient can be used to update the prior probability of disease from what is known about the prevalence of disease in the population. As soon as we obtain individual-level information by taking a history and doing a physical examination, we develop a different estimate for the prior probability than the prevalence of the disease.

In this chapter, we discuss combining multiple types of information – elements of history, findings on physical examination, laboratory results, or radiographic images. We cover (at least theoretically) how we might get from prevalence to prior probability based on the history and physical examination, and then to posterior probability based on additional information from diagnostic tests. We begin by reviewing the concept of test independence, and then we discuss how to deal with departures from independence, which are probably the rule rather than the exception.

Test independence

Definition: Two tests are “independent” if the LR for any combination of results on the two tests is equal to the product of the LR for the result on the first test times the LR for the result on the second test.

Explanation: What independence means is that, *among people who have the disease*, knowing the result of Test 1 tells you nothing about the probability of a certain result

on Test 2, and that the same is true *among people who do not have the disease*. When we say the two tests are independent, we mean they are independent *once disease status is taken into account*. That is why we keep putting that part in italics. This is called “stratifying” on disease status. If we did not do this, then patients with an abnormal result on Test 1 would be more likely to be abnormal on Test 2 simply because they would be more likely to have the disease. Mathematically, the way to express this is to say the tests are “conditionally independent,” by which we mean they are independent once the condition of having the disease or not is accounted for.

Using probability notation, independence means that, for every possible result r_B of Test B, the probability of a patient with disease having that result, $P(r_B|D+)$, is the same regardless of the result that the patient has on Test A. If Tests A and B are dichotomous and the patient actually has the disease, independence requires that a false negative on Test B is no more likely because the patient had a false negative on Test A. It is easy to think of counterexamples – nonindependent tests – where a false negative on Test A makes a false negative on Test B more likely. For example, a patient with acute cardiac ischemia who does not have ST elevation on the electrocardiogram (ECG) is also less likely to have a positive troponin.¹

Similarly, in a patient without disease, independence means the probability of the result, $P(r_B|D-)$, is the same regardless of the result on Test A. For dichotomous tests on a patient without the disease, independence requires that a false positive on Test B is no more likely because the patient had a false positive on Test A. Again, counterexamples are numerous. An abdominal pain patient *without* appendicitis who nevertheless has a fever is also more likely to have an elevated WBC count.

If neither $P(r_B|D+)$ nor $P(r_B|D-)$ depends on the result of Test A, then the LR for result r_B on Test B, $P(r_B|D+)/P(r_B|D-)$, does not depend on the result of Test A. When this is the case, and tests are independent, we can start with any prior odds of disease and multiply by the LR for the result of Test A to get posterior odds of disease. Then, we use these odds as the prior odds for Test B, multiply by the LR for the result of Test B, and get the posterior odds after both Test A and Test B.

Perhaps it is easiest to understand independence by giving some examples of nonindependent tests. Suppose you are doing a study to identify predictors of pneumonia in nursing home residents with fever and cough. You determine that cyanosis has an LR of 5 and that an oxygen saturation of 85% to 90% has an LR of 6. If the patient is cyanotic *and* has an oxygen saturation of 87%, does that mean we can multiply the prior odds by $5 \times 6 = 30$ to get the posterior odds? No. Once we know that the patient is cyanotic, we do not learn that much more about the probability of pneumonia from the oxygen saturation, and vice versa.

There are at least three related reasons why tests can be nonindependent. The first is that they are measuring similar things. The cyanosis and low oxygen saturation example illustrates this. Some patients with pneumonia will be hypoxemic and some will not, and both the patient’s color and the oxygen saturation are giving

¹ Troponin is a serum marker for heart muscle damage. It is actually a continuous test, but for current purposes it can be viewed as dichotomous.

information on that one aspect of pneumonia: hypoxemia. Jaundice, dark urine, light stools, and direct hyperbilirubinemia provide a similar example of tests that are measuring the same basic pathophysiologic manifestation of hepatitis, and therefore will not be independent.

A second reason is that the disease is heterogeneous. Pneumonia is heterogeneous in that some cases are associated with hypoxemia and some are not. Similarly, some cases of hepatitis are icteric and some are not. But disease heterogeneity can lead to test nonindependence even when the tests do not measure the same pathophysiologic aspect of the disease. For example, another cause of heterogeneity is disease severity. The most severe acute coronary syndromes are ST elevation myocardial infarctions, and these are also the acute coronary syndromes most likely to result in elevated troponins. Varying disease severity is an obvious cause of nonindependence for diseases with an arbitrary definition. For example, if we define coronary heart disease based on at least 70% stenosis of a coronary vessel, patients with 71% stenosis are more likely to have false-negative results on most tests than those with 98% stenosis, regardless of what pathophysiologic alteration is actually being measured.

Third, the nondisease may be heterogeneous. Lack of coronary disease is going to be much more difficult to diagnose in a patient with 69% stenosis than it is in patients with 10% stenosis. Alternatively, the nondisease group could be heterogeneous because it includes patients with other diseases that make the tests falsely positive. For example, if we were looking at LRs for bacterial meningitis in patients with headache and fever, the comparison group might include both patients with no meningitis at all and patients with viral meningitis. If that were the case, we would expect findings that pointed to meningitis in general [e.g., headache, stiff neck, photophobia, cerebrospinal fluid (CSF) pleocytosis] also to be nonindependent, because all of these would be more likely to be falsely positive in the subset of non-bacterial meningitis patients who had viral meningitis. Above, we gave the example of fever and WBC count as tests for appendicitis in patients with abdominal pain. A patient without appendicitis who does have fever is also more likely to have a high WBC count, because the same nonappendicitis condition causing his fever is also likely to cause an increased WBC count.

Visualizing nonindependence using the LR slide rule

When tests are independent, their LRs can be multiplied, which is the same as adding on the LR slide rule's log(odds) scale (see Chapter 3). In other words, the arrows for their results can be laid end-to-end. If the tests are not independent – for example, if they measure the same pathophysiologic aspect of a disease – you cannot get the LR arrow for the combined results by laying the LR arrows for the individual results end-to-end.

For example, consider aspartate transaminase (AST) and lactate dehydrogenase (LD) enzyme levels² in the diagnosis of hepatocellular injury (liver inflammation). Conditions that cause a false-positive result on one (e.g., hemolysis, muscle injury)

² AST = Aspartate Transaminase, LD = Lactate Dehydrogenase. Both are important metabolic enzymes that are commonly elevated in hepatitis, but they can also be elevated when non-liver cells, such as blood cells

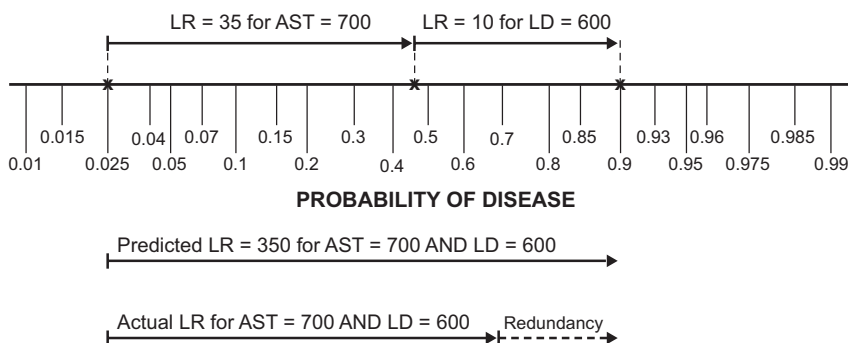


Figure 8.1 LR slide rule arrows demonstrate the concept of nonindependence. In diagnosing hepatocellular injury, the AST and LD are not independent because they can both be elevated in other conditions like hemolysis and muscle damage; therefore, the LR arrow for the combination of both AST = 700 and LD = 600 cannot be obtained by laying the LR arrows for each of these results end-to-end.

also make a false-positive result on the other more likely. Therefore, although the LR for AST = 700 might be 35, and for LDH = 600 might be 10, the LR for both AST = 700 *and* LD = 600 is likely to be a lot less than $(10 \times 35 =) 350$, because the tests are not independent. This is illustrated in Figure 8.1.

Combining the results of two dichotomous tests: an example

We start with an example of combining results from two prenatal sonographic tests for trisomy 21 (Down syndrome): nuchal translucency (NT) and examination for the nasal bone. Trisomy 21 can be established definitively by using chorionic villus sampling, but this is an invasive test that can accidentally terminate the pregnancy. Both NT and nasal bone examination are noninvasive tests done by ultrasound at approximately 13 weeks gestation. Examination of the nasal bone is a truly dichotomous test; absence of the nasal bone (NBA) is suggestive of trisomy 21 and therefore constitutes “positive” nasal bone exam for trisomy 21. NT is the measurement (in mm) of the subcutaneous edema between the skin at the back of the fetal neck and the soft tissue overlying the cervical spine. We pointed out in Chapter 4 that choosing a cut-off to make a continuous or multilevel test into a dichotomous test discards information. However, for purposes of exposition, we will use the cut-off of 3.5 mm to make NT a dichotomous test; we will consider an $NT \geq 3.5$ mm “positive” for trisomy 21.

Cicero et al. (2004) reported NTs and nasal bone examinations on 5556 fetuses. The tests were done prior to definitive determination of trisomy 21 versus normal karyotype via chorionic villus sampling. The results are shown in Table 8.1.

Assume that the fetuses screened have a trisomy 21 prevalence of 6%. If a fetus has a $NT \geq 3.5$ mm, the post-test probability of trisomy 21 is 31%. Ignoring the NT, if the fetus has NBA, the post-test probability is 64%. See if you can reproduce

and muscle cells are damaged. The units of both are International Units per Liter (IU/L). We left the units out in this text to improve readability.

Table 8.1. NT and NBA in fetuses with and without trisomy 21 as determined by chorionic villus sampling^a

		Trisomy 21		LR
		Yes	No	
NT ≥ 3.5 mm	Yes	212	478	7.0
	No	121	4745	0.4
Total		333	5223	

		Trisomy 21		LR
		Yes	No	
NBA	Yes	229	129	27.8
	No	104	5094	0.3
Total		333	5223	

^a From Cicero et al. (2004).

these calculations. They are displayed in Figure 8.2 on the LR Slide Rule’s Log(Odds) scale.

The calculations in Figure 8.2 apply if we consider either the NT ≥3.5 mm *or* NBA. What if we consider both? First, let us assume the two tests are independent. If the two tests are independent, we can multiply their LRs, so the LR for a combined positive result, NT ≥3.5 mm *and* NBA, would be $7.0 \times 27.8 = 194$. Using this LR and a pre-test probability of 6% results in a post-test probability of 92.5%. Figure 8.3 displays this calculation.

Now, rather than assuming independence, let us look at the actual data from the sample. If we consider both NT and the examination for the nasal bone together, there are four possible results. Table 8.2 shows the data and LRs associated with those four results.

Look at the top row of the table, where both tests are positive for trisomy 21. If both tests are positive, the LR is 68.8, not $7.0 \times 28.8 = 194$. Therefore, if the pre-test probability of trisomy 21 is 6% and both tests are positive, the post-test probability is 81%, not 92.5% (see Fig. 8.4).

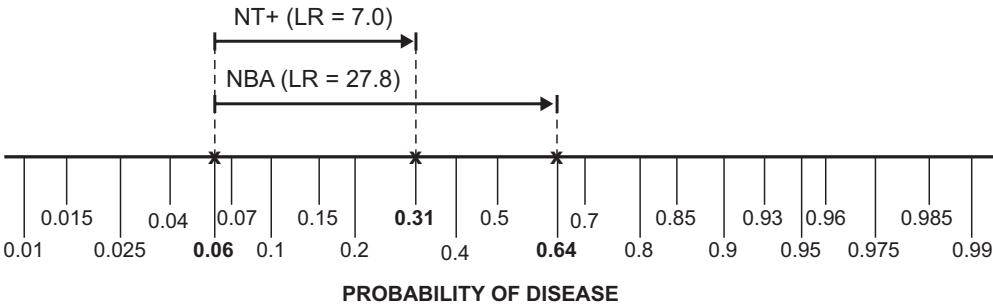


Figure 8.2 Starting with a 6% pre-test probability of trisomy 21, an NT ≥3.5 mm (NT+) increases the probability to 31%; ignoring the NT result, NBA increases the probability to 64%.

Table 8.2. The combination of NT and nasal bone examination results in fetuses with trisomy 21 and chromosomally normal fetuses

NT ≥ 3.5 mm	NBA	Trisomy 21				LR
		Yes	%	No	%	
Yes	Yes	158	47.4%	36	0.7%	68.8
Yes	No	54	16.2%	442	8.5%	1.9
No	Yes	71	21.3%	93	1.8%	12
No	No	50	15.0%	4652	89.1%	0.2
Total		333	100%	5223	100%	

^a Data from Cicero et al. (2004).

NBA does not tell you as much if you already know that the NT is ≥ 3.5 mm. Even in chromosomally normal fetuses, enlarged NT is associated with NBA. Of normal (D–) fetuses with a negative NT (<3.5 mm), only 2.0% had NBA. Of normal (D–) fetuses with a positive NT (≥ 3.5 mm), 7.5% had NBA. A false-positive NT makes a false positive on the nasal bone examination more likely. Ontologically, narrowing of the nuchal stripe and ossification of the nasal bone both occur as the fetus develops.

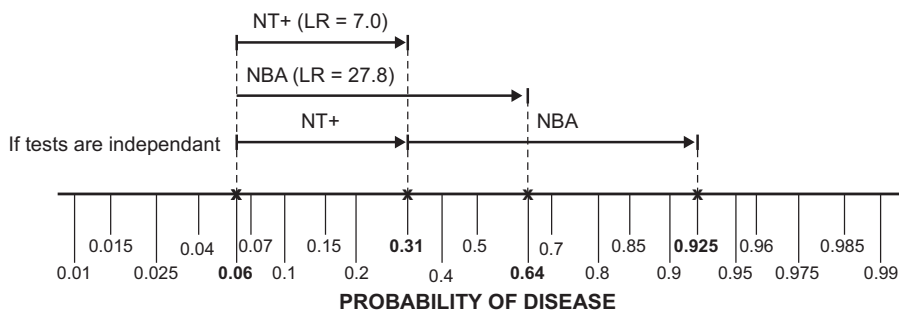


Figure 8.3 If the NT and nasal bone exams are independent, the LR of a combined positive result is the product of the LRs for a positive result on each test. On the log scale, multiplying LRs is the same as laying their arrows end-to-end.

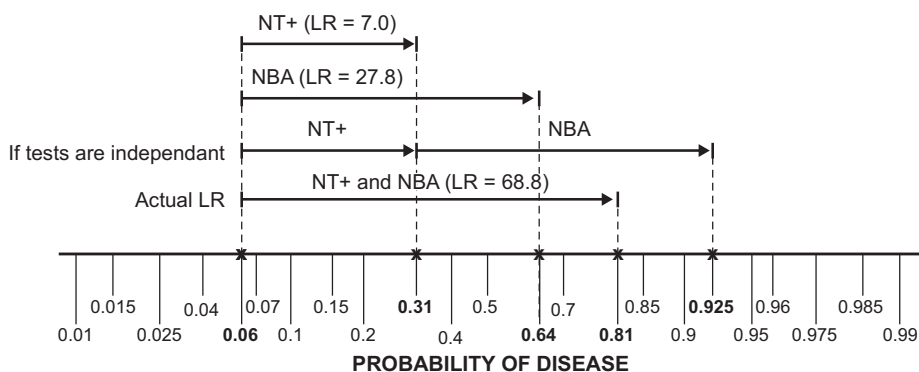


Figure 8.4 The LR associated with the combination of NT ≥ 3.5 mm (NT+) and NBA is less than the product of the LR for each result individually.

Box 8.1: Spectrum bias in estimating the sensitivity of the nasal bone examination for fetal chromosomal abnormalities

Examination of the nasal bone is a noninvasive test used to determine which fetuses are at high enough risk to warrant chorionic villus sampling, which is a test for all chromosomal abnormalities, not just trisomy 21. Accurately estimating the nasal bone exam's sensitivity requires a clinically realistic group of chromosomally abnormal (D+) fetuses. The previously presented data from Table 8.1 on the sensitivity of the nasal bone examination are:

		D+	D–
Nasal Bone	Yes	229	129
Absent	No	104	5,094
Total		333	5,223

Sensitivity = $229/333 = 69\%$.

The D+ group included only fetuses with trisomy 21 and excluded 295 fetuses with other chromosomal abnormalities, especially trisomy 18. If the purpose of the nasal bone exam is to determine when to get chorionic villus sampling, these 295 fetuses with chromosomal abnormalities other than trisomy 21 should be included in the D+ group. Of the 295, 95 (32%, not 69%) had NBA:

		D+	D–
Nasal Bone	Yes	$229 + 95 = 324$	129
Absent	No	$104 + 200 = 304$	5,094
Total		$333 + 295 = 628$	5,223

Sensitivity = $324/628 = 52\%$ (not 69%).

Including these 295 in the D+ group results in a sensitivity of 52%, not 69%, which constitutes a more clinically useful estimate of the sensitivity of NBA for chromosomal abnormalities.

Some chromosomally normal fetuses must develop more slowly than usual, resulting in both a positive NT and NBA.

Spectrum bias

The example of combining the NT and nasal bone examinations also illustrates the issue of spectrum bias that we discussed in Chapter 5. Our example assumes that the only fetal condition at issue is trisomy 21. In fact, the data we presented excluded from the D+ group 295 fetuses with chromosomal abnormalities other than trisomy 21. This biased upward the sensitivity of each test individually and also exaggerated the accuracy of the two tests when used in combination. Box 8.1 shows numerically how this spectrum bias exaggerates the sensitivity of the nasal bone examination for chromosomal abnormalities.

Combining the results of multiple dichotomous tests

We have demonstrated one way to handle the results of multiple tests: gather data on the LR for each possible combination of test results. For two dichotomous tests,

as in our example above, there are four possible results (+/+, +/−, −/+, and −/−). For three such tests, there are eight possible results; for four tests, sixteen results; and so on. Even with large samples, you might not have enough data to calculate LRs for the uncommon result combinations.

Another approach is to lump together all discordant results, calculating one LR for this category, while calculating separate LRs for the concordant results (all positive or all negative). In the case of two dichotomous tests, there would be an LR for “positive–positive (+/+),” “negative–negative (−/−),” and “discordant (+/− or −/+).” However, we saw in Chapter 3 that some tests are much more valuable when they are positive than when negative, or vice versa. A pathognomonic finding (specificity = 100%) should rule in disease when positive, regardless of other test results. Thus, if the pathognomonic finding is present and all the other tests are negative, it does not make sense to lump this together with other discordant results. Also, a single category for “discordant results” cannot accommodate multilevel or continuous tests.

A variant of the “lumping together” approach is to combine multiple tests into a decision rule that is considered positive if any one of the tests is positive. This approach has been used in the Ottawa Ankle Rule (Stiell et al. 1994) to determine which ankle-injury patients should get radiographs,³ the NEXUS (National Emergency X-Ray Utilization Study) Rule (Hoffman et al. 1998, 2000) to determine which neck-injury patients should get cervical spine films,⁴ and the San Francisco Syncope Rule (Quinn et al. 2004) to identify high-risk syncope patients requiring hospitalization.⁵ This strategy clearly maximizes sensitivity, though at the expense of specificity. The main issue is deciding which of many candidate tests to include in the rule – a topic to which we will return.

Recursive partitioning

Another approach is to use recursive partitioning to develop a fixed optimal sequence in which to do the multiple tests. “Recursive partitioning” (also called CART, for Classification and Regression Trees) is just what it sounds like – recursive meaning you do it over and over again, and partitioning meaning you divide up the data in different ways.

In our example of NT and examination for the nasal bone as tests for trisomy 21, which test should we do first? Figure 8.5 shows a tree of probabilities of trisomy 21 after each possible test result: (A) performing the nuchal translucency test first and (B) performing the nasal bone exam first. The nasal bone exam is better at discriminating between trisomy 21 and chromosomally normal fetuses. After a positive nasal bone

³ Radiographs are recommended if the patient has tenderness of either malleolus, navicular, or base of the fifth metatarsal; or the patient is unable to bear weight for four steps both at the time of injury and the time of evaluation.

⁴ Cervical spine films are recommended if the patient has any of the following: midline posterior cervical spine tenderness, alcohol or drug intoxication, abnormal alertness, focal neurologic deficit, or distracting painful injury.

⁵ Hospitalization is recommended if the patient has shortness of breath, history of congestive heart failure, triage systolic blood pressure less than 90 mm Hg, hematocrit less than 30%, or abnormal electrocardiogram.

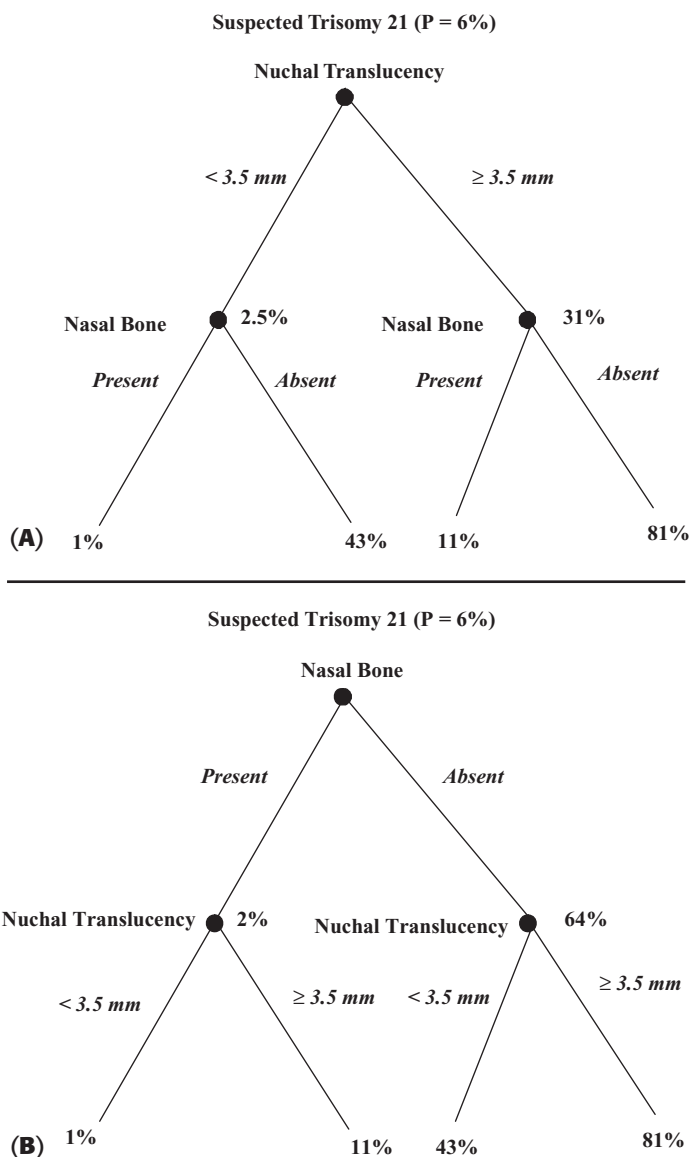


Figure 8.5 (A) Tree with branch-point probabilities of trisomy 21, assuming the NT test is performed first. (B) Tree with branch-point probabilities of trisomy 21 assuming the nasal bone exam is performed first.

exam (NBA), the probability of trisomy 21 is 64%; after a negative nasal bone exam (nasal bone present), the probability is 2%. This is as compared to 31% and 2.5% after positive and negative NTs, respectively. If your threshold probability (P_{TT}) for going on to chorionic villus sampling is 15%, you can stop after the nasal bone exam. After a positive nasal bone exam, a negative NT does not lower the probability of trisomy 21 below 15%, and after a negative nasal bone exam, a positive NT does not raise the probability of trisomy 21 above 15%. This suggests that you should do the nasal bone exam first.

For each sequence of tests, the LRs for the results of the second test are conditional on the result of the first test, the LR for the results of the third test are conditional on the results of the first and second tests, etc. At first, it would seem that you still have to work out the LRs for all possible combination of test results, but the algorithm is developed to “prune” branches of the tree and allow you to stop testing, or skip the next test(s) in the sequence. For example, after a negative result on a very sensitive test, you might stop testing – the probability has gotten so low that additional tests are not needed.

Whether to continue and do the NT test after the nasal bone exam depends on your actual threshold for chorionic villus sampling. We have seen that, if it is 15%, you can stop after the nasal bone exam. However, if it is 5% rather than 15% and the initial nasal bone exam is negative, you should continue with the NT test; a positive test will move the probability above the 5% threshold (Figs. 8.6A and 8.6B). If the initial nasal bone exam is positive, it does not make sense to do the NT test, because (at least as dichotomized here) the result cannot change your decision to proceed with chorionic villus sampling.

A classic example using recursive partitioning to develop a testing algorithm was developed by Goldman et al. to identify myocardial infarction in emergency department patients with chest pain (Goldman et al. 1988) (see Fig. 8.7). The percentages at each branch-point in Figure 8.7 represent the proportion of patients in that “partition” with acute myocardial infarction.

A much simpler example from the Pediatric Research in Office Settings Febrile Infant Study (Pantell et al. 2004) is shown in Figure 8.8. The percentages next to each branch-point in Figure 8.8 are the proportions of infants with bacteremia or meningitis.

Figures 8.5 through 8.8 display probabilities rather than LRs. We will return to this later, but if the LRs for a study like this are more generalizable than the actual probabilities, authors could publish just a subset of the LRs. Others could use these LRs with their own prior probabilities. For example, in the febrile infant example (Fig. 8.8), the only LRs related to temperature that would be needed would be those for $T \geq 38.6$ and $T < 38.6$ among infants who looked well and were at least 3.5 weeks old. This recursive partitioning algorithm ends up being a rule, like the Ottawa Ankle Rule and others mentioned above, that is considered positive if *any* of the individual tests is positive; that is, the infant is classified as “high risk” if he or she appears moderately ill, is < 3.5 weeks old, or has a temperature $\geq 38.6^\circ\text{C}$. This is a common, but by no means necessary, result of recursive partitioning algorithms; the chest pain algorithm in Figure 8.7 provides a counter-example.

Recursive partitioning handles continuous test results by selecting cut-offs to dichotomize the results. As we discussed in Chapter 4, selecting a fixed cut-off to dichotomize a test reduces the information to be gained from it, because a result just on the abnormal side of the cut-off is equated with a result that is maximally abnormal. However, with recursive partitioning, you are not necessarily done with a variable once you have dichotomized it. For example, an algorithm for predicting meningitis from CSF findings might first dichotomize the CSF WBC count

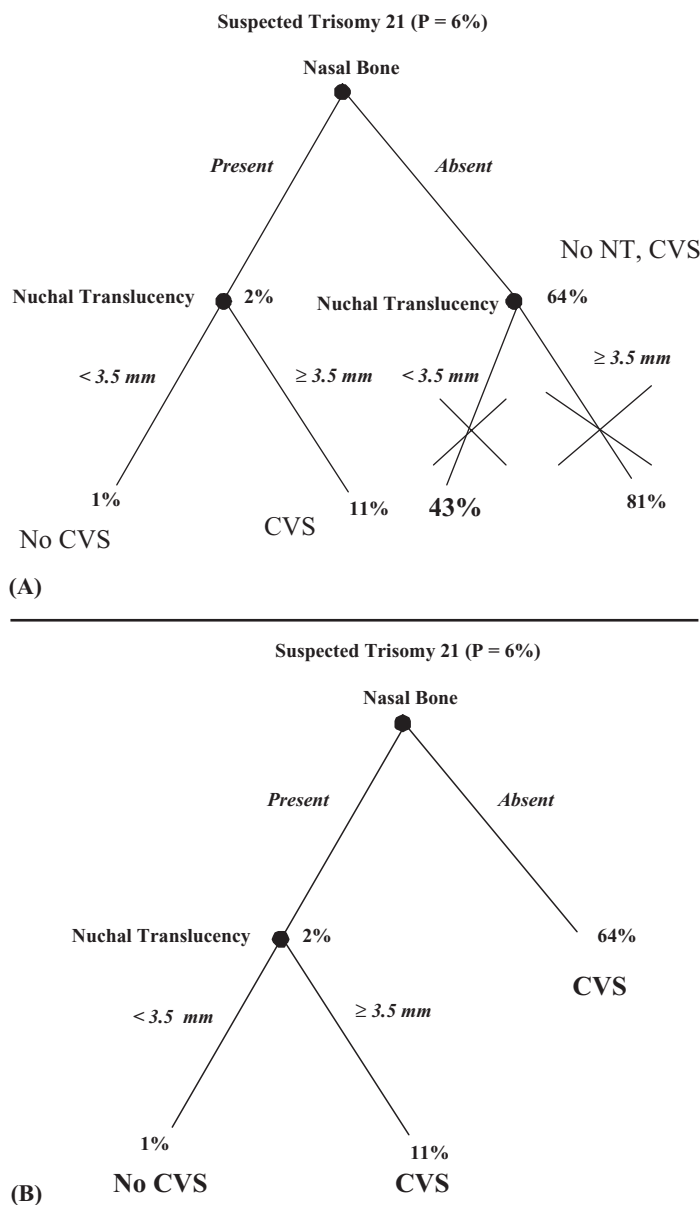


Figure 8.6 **(A)** If the probability threshold for chorionic villus sampling is 5% and the nasal bone is absent, a negative NT cannot change the decision to do chorionic villus sampling, so these branches in the tree may be pruned. **(B)** If the threshold probability for proceeding to chorionic villus sampling is 5%, the combination of nasal bone exam and NT becomes a two-test rule that is considered positive if either of the tests is positive.

(per mm^3) at 1,000; and then, if it was $<1,000$, dichotomize again at 100, where patients with CSF WBC count between 100 and 1,000 would be classified as high risk for meningitis if they had some other finding (e.g., low CSF glucose) as well.

Recursive partitioning software works by trying all different ways of splitting the data, and then selecting the way that leads to the least misclassification. The user is

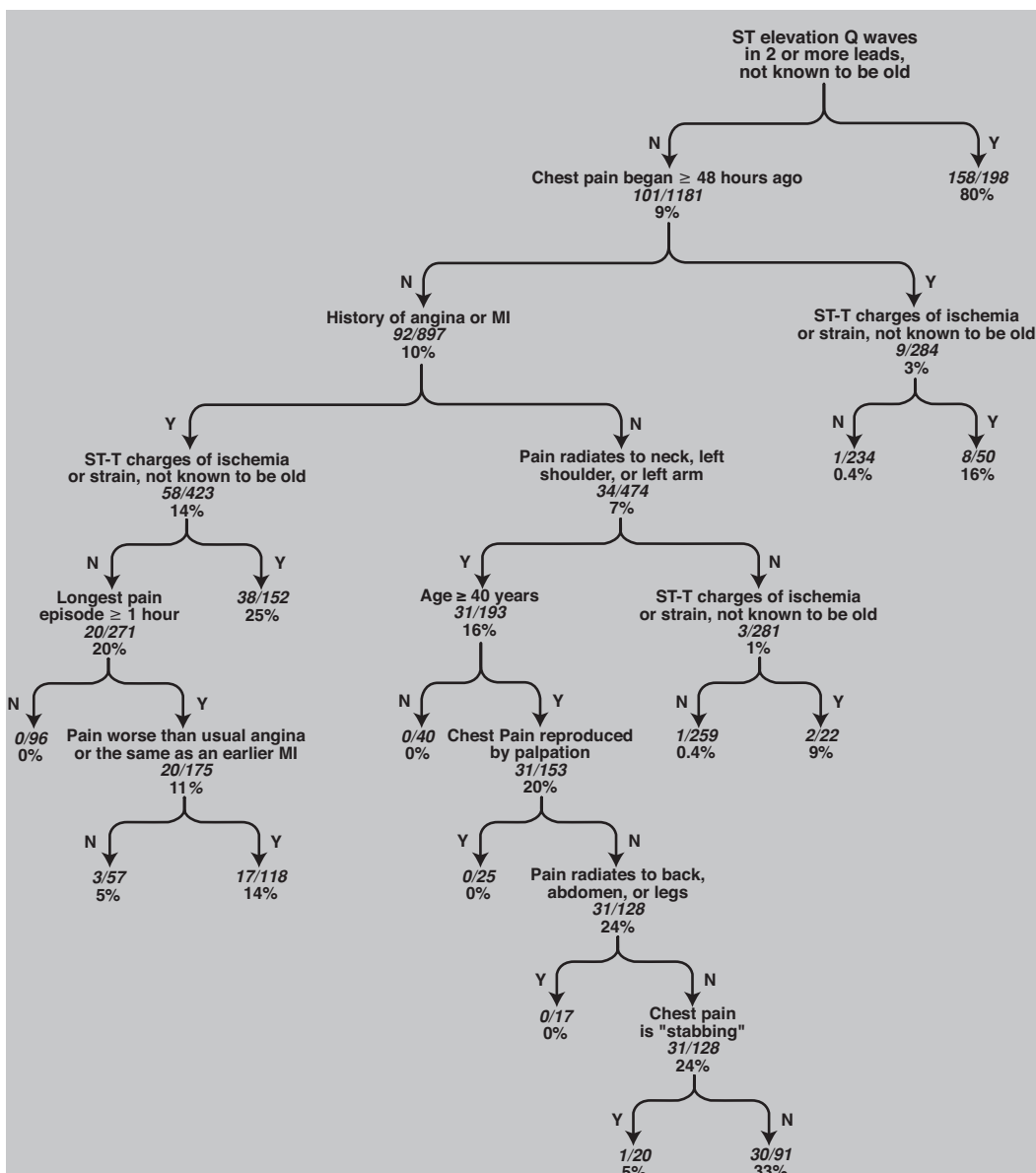


Figure 8.7 Recursive partitioning to predict the likelihood that a chest pain patient has myocardial infarction (Goldman et al. 1988; Lee et al. 1991).

allowed to specify the ratio of misclassification costs – that is, how much worse it is to have a false-negative than a false-positive result. For the febrile infant study example above (Fig. 8.8), the tree resulted from an analysis with the ratio of false-negative to false-positive misclassification costs set at 50:1.

Recursive partitioning does not assume that the risk of disease changes monotonically with a continuous test result. For example, in Chapter 4 we assumed that the probability of bacteremia increases as the peripheral WBC count increases, but this is not necessarily true in very young infants. A WBC count <5,000 has as high an

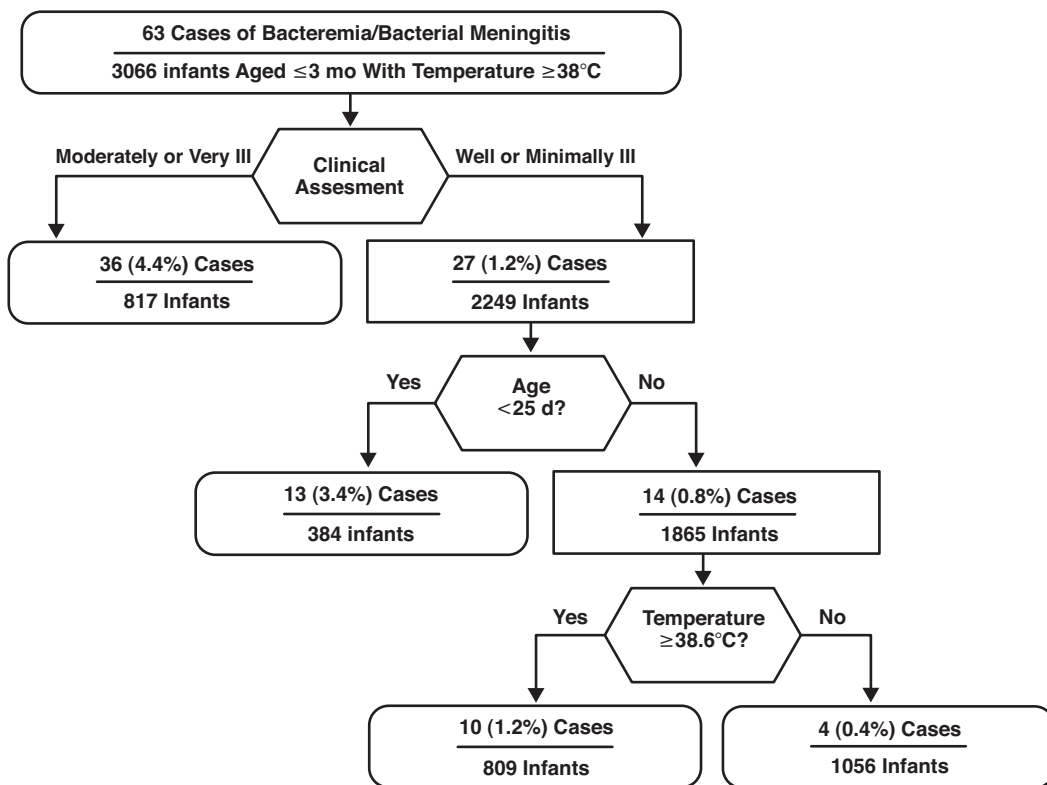


Figure 8.8 Recursive partitioning combining general appearance, age in weeks, and temperature to determine likelihood of bacteremia or bacterial meningitis in febrile infants <3 months old (Pantell et al. 2004). Used with permission.

LR for bacteremia as a WBC count between 25,000 and 30,000 (see Problem 4–5 for an illustration). However, if risk does change monotonically with a continuous test result, logistic regression (which we will discuss in the next section) provides a more efficient use of the data in predicting the risk of disease.

Logistic regression

Partially because recursive partitioning deals less efficiently with continuous variables than with discrete variables, the most popular way to accommodate the results of multiple diagnostic tests where at least some results are continuous is multiple logistic regression modeling (Wasson et al. 1985; Laupacis et al. 1997; Katz 1999). In Chapter 3, we used odds instead of probabilities in Bayes's Theorem. Unlike probabilities, odds do not have an upper bound of 1, and pre-test odds can be multiplied by the LR of a test result to get post-test odds. Also in Chapter 3, we converted this multiplication into addition by replacing odds with their logarithms on the LR slide rule. Logistic regression takes advantage of these desirable properties of odds (compared with probabilities) and models the natural logarithm of the odds of disease $[\text{Ln}(\text{odds})]$ as a linear function of the test results.

Box 8.2: How to calculate the OR for NBA in the diagnosis of trisomy 21

Here are the data on the nasal bone exam in fetuses with and without trisomy 21:

		Trisomy 21		
		Yes	No	Odds
NBA	Yes	229	129	$229/129 = 1.775$
	No	104	5,094	$104/5,094 = 0.020$
		Odds	$229/104$	$129/5,094$
		2.202	0.025	

The OR is

$$\frac{\text{Odds of Disease in those with a Positive Test}}{\text{Odds of Disease in those with a Negative Test}} = \frac{\text{Odds}(D+|+)}{\text{Odds}(D+|-)} = \frac{1.775}{0.020} = 87$$

Because of the symmetry of the odds ratio, this is the same as

$$\frac{\text{Odds of a Positive Test in those with Disease}}{\text{Odds of a Positive Test in those without Disease}} = \frac{\text{Odds}(+|D+)}{\text{Odds}(+|D-)} = \frac{2.202}{0.025} = 87$$

Odds ratios

The logistic regression coefficient for each test result is the natural logarithm of its multivariate odds ratio (OR). In Chapter 9, we will return to the OR in the context of quantifying the benefits of a treatment. (The OR is often used inappropriately to quantify treatment effects in randomized trials.) Here, we discuss how ORs are used to quantify the information provided by a positive test result or presence of a risk factor. ORs are easiest to understand when the test is dichotomous; in this case, the OR is the quotient of the odds of disease in those with a positive test divided by the odds of disease in those with a negative test:

$$\text{Odds Ratio} = \frac{\text{Odds of Disease in those with a Positive Test}}{\text{Odds of Disease in those with a Negative Test}}$$

In contrast with probabilities, odds work symmetrically so that this is also the quotient of the odds of a positive test in those with disease divided by the odds of a positive test in those without disease:

$$\text{Odds Ratio} = \frac{\text{Odds of a Positive Test in those with Disease}}{\text{Odds of a Positive Test in those without Disease}}$$

Box 8.2 shows the calculation of the OR for NBA in the diagnosis of trisomy 21.

The OR for a dichotomous test is also the LR of a positive result divided by the LR of a negative result. ORs and LRs are frequently confused. For test results, LRs are generally more appropriate to use than ORs, but when assessing risk factors with widely varying prevalences from population to population, the OR may be more useful, as shown in Box 8.3.

When the test is dichotomous, the farther the OR is from 1, the stronger the association between the test result and the disease. For continuous tests, the OR from logistic regression is the amount the odds of disease change per unit increase

Box 8.3: Understanding the difference between ORs and LRs

If we start with the prior probability of disease, $P(D+)$, we can convert to prior odds, $\text{Odds}(D+)$, and then multiply by the $\text{LR}(+)$ or $\text{LR}(-)$ to get the posterior odds:

Odds of disease given a positive test or exposure = $\text{Odds}(D+|+) = \text{Odds}(D+) \times \text{LR}(+)$

Odds of disease given a negative test or no exposure = $\text{Odds}(D+|-)$
 $= \text{Odds}(D+) \times \text{LR}(-)$

The OR is the ratio of the posterior odds in those who test positive (or are exposed to a risk factor) to those who test negative (or are unexposed). Because the prior odds cancel out of that ratio, the OR is just $\text{LR}(+)/\text{LR}(-)$.

$$\begin{aligned}\text{OR} &= \text{Odds}(D+|+)/\text{Odds}(D+|-) = [\text{Odds}(D+) \times \text{LR}(+)]/[\text{Odds}(D+) \times \text{LR}(-)] \\ &= \text{LR}(+)/\text{LR}(-)\end{aligned}$$

If you want the odds of disease in a patient with a positive test result or exposure to a risk factor, you can either multiply the odds of disease *in the overall population* by the $\text{LR}(+)$, or multiply the odds of disease *in the test-negative or unexposed population* by the OR. In other words, if you start with the overall odds of disease, you use the $\text{LR}(+)$; if you start with the odds of disease in the test-negative or unexposed group, you use the OR. The only time the OR approximates the $\text{LR}(+)$ is when the prevalence of a positive test or an exposure is so low that it doesn't significantly affect the population prevalence of disease. [In that case, the $\text{LR}(-)$ will be very close to 1.]

The difference between ORs and LRs is illustrated in Figure 8.9. Consider a disease that has a strong risk factor, the prevalence of which varies widely in different populations. An example one of us has studied is UTIs in young febrile infant boys (Newman et al. 2002). The OR for UTI in uncircumcised boys, compared with circumcised boys, is about 10. What would be the LRs? The answer is that the LRs will depend on the proportion of the population that is circumcised. In Figure 8.9A, most of the boys in the population are circumcised. Therefore, the prior odds of UTI in a febrile boy will be low, and if he is circumcised (which we are calling being unexposed to the risk factor), the odds will not decline very much because they already start out low. On the other hand, if he is one of the few who is uncircumcised, the $\text{LR}+$ will be high and significantly increase his posterior odds.

Now consider the situation in a population where hardly any boys are circumcised (Fig. 8.9B). The prior odds start out much higher, reflecting this high prevalence of a strong risk factor for UTI. However, in this case, the odds change much more if the boy is circumcised than if he is not. Thus, for this type of clinical situation, LRs have the disadvantage that they are unlikely to be generalizable from one population to another. This is not the case for the sort of predictive factors that LRs were designed for: clinical tests, such as laboratory and imaging tests. For example, the prevalence of abnormality on laboratory tests (like the urinalysis) does not vary widely in populations as the prevalence of circumcision does. Hence, LRs for laboratory tests are more likely to be generalizable than LRs for risk factors or behaviors, the prevalence of which may vary widely across populations.

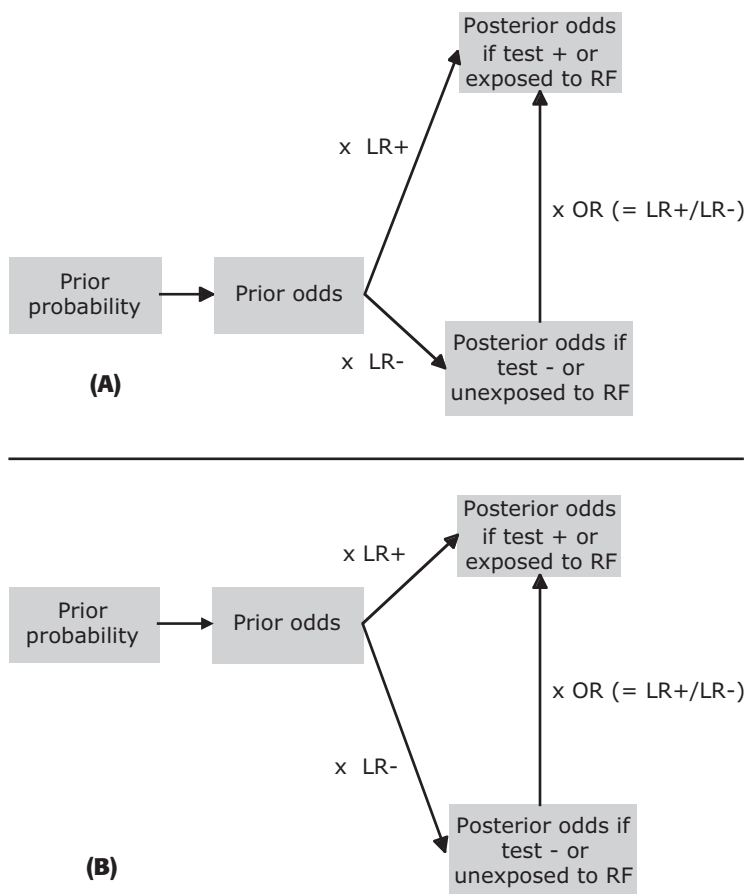


Figure 8.9 Relationship between prior odds, LRs (LR+ and LR-), posterior odds, and the OR. **(A)** Low prevalence of strong risk factor. **(B)** High prevalence of strong risk factor. The length of an LR arrow corresponds to the magnitude the LR (actually the logarithm of the LR). The LR magnitudes change depending on the prevalence of the risk factor, whereas their ratio, the OR, remains the same.

in the test result. Thus, the OR for fever per degree will differ depending on whether the temperature is measured in Centigrade or Fahrenheit. (It will be farther from 1 for temperature measured in Centigrade.) If the units of measurement vary, it is not true that ORs farther from 1 always mean a stronger association with disease.

Logistic regression modeling

We applied a logistic regression approach to the NT and nasal bone exam data, using $NT \geq 3.5$ mm and NBA as dichotomous predictors of trisomy 21. The dataset included 5,556 records, one for each fetus evaluated. The variable for NT was valued “1” for $NT \geq 3.5$ mm and “0” for $NT < 3.5$ mm; the variable for NBA was similarly valued “1” if nasal bone absent or “0” if nasal bone present. The binary outcome variable for trisomy 21 was also coded in standard fashion. We entered the frequencies from Table 8.2 into our favorite statistical program (Stata). The results, as they might appear in a journal article, are shown in Table 8.3.

Table 8.3. Multivariate ORs resulting from a logistic regression model using fetal NT and NBA as dichotomous predictors of trisomy 21

	Multivariate OR for trisomy 21	95% CI for the OR
NT ≥ 3.5 mm	8.7	6.3–11.8
NBA	53.0	38.7–72.7

The multivariate OR for NBA is much greater than the multivariate OR for NT. This allows us to say that, when both are available, NBA is a more important predictor of trisomy 21 than NT.

A multiple logistic regression model adjusts the OR associated with one dichotomous test for the fact that one or more additional tests are performed. Based on the data in Table 8.1, the univariate OR for NBA is 87.0 (calculated in Box 8.2) and the univariate OR for NT is 17.4. Because the two tests are not independent, the multivariate ORs are lower when both variables are included together than they are for each variable separately.

Box 8.4: Advanced material on logistic regression: interaction terms and goodness of fit

The logistic model presented in Table 8.3 does not include an interaction term. An “interaction term” is an additional term that distinguishes when both tests are positive from when only one or the other is positive. In the above model, the interaction term would be NT $\times$ NBA, which would equal “1” only if both tests were positive. Unless a logistic regression model includes interaction terms, the result of any given test changes the Ln(odds) of disease by the same amount, regardless of how the other tests came out. For this reason, it is important to assess how well the logistic model fits the data – the so-called goodness of fit. The lack of an interaction term means our model for predicting trisomy 21 assumes that NBA has the same effect on the Ln(odds) regardless of whether NT is ≥ 3.5 mm or < 3.5 mm. As shown in the table below, this model fits the data reasonably well, and an interaction term is probably unnecessary.

Comparison of Actual Probability of Trisomy 21 with Probability Predicted by the Logistic Regression Model Without an Interaction Term

Nuchal Tranlucency	Nasal Bone	Probability of Trisomy 21	
		Actual (%)	Predicted (%) ^a
≥ 3.5	Absent	81	85
	Present	11	10
	Absent	43	39
< 3.5	Present	1	1

^a Logistic regression model with no interaction term.

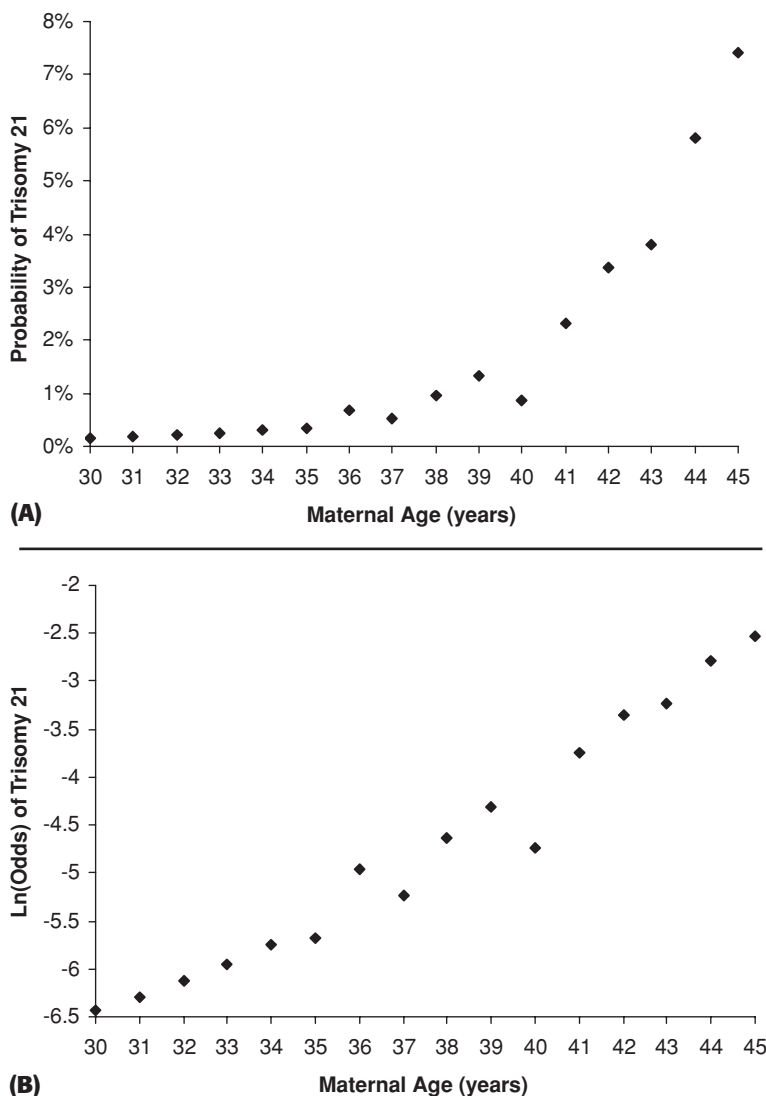


Figure 8.10 Probability of trisomy 21 as a function of maternal age (data from Snijders et al. 1999, Table 1). **(A)** Plot of probability versus maternal age. **(B)** Plot of Ln(odds) versus maternal age.

Logistic regression using the results of a single continuous test

So far in this chapter, we have violated the principle of Chapter 4 and discarded information by dichotomizing NT at 3.5 mm, calling an NT <3.5 mm “negative” and ≥ 3.5 mm “positive” for trisomy 21. In fact, an NT of 6 mm is much more suggestive of trisomy 21 than an NT of 3.5 mm. One of the main reasons to use logistic regression is to accommodate one or more continuous test results. To see why logistic regression models Ln(odds) instead of probability, consider another predictor of trisomy 21: maternal age. Figure 8.10A shows the probability of trisomy 21 (at 16 weeks gestation) by maternal age. (Snijders 1999) Because probabilities are

bounded by 0 and 1, the relationship between probability and a continuous variable, such as maternal age, is a distinctly nonlinear curve. If, instead of probability, we graph the $\text{Ln}(\text{odds})$ as a function of maternal age, as in Figure 8.10B, we tend to get a simple linear relationship. This is why logistic regression models $\text{Ln}(\text{odds})$ instead of probability as a linear function of test results.

As discussed in Chapter 4, we sometimes choose a cut-off value for a continuous test to trigger some action. In maternal–fetal medicine, the cut-off for obtaining a fetal karyotype by chorionic villus sampling or amniocentesis has been arbitrarily set at a 1 in 300 (0.33%) risk of trisomy 21. Based on logistic regression models used to fit data like those displayed in Figure 8.10, the maternal age cut-off should therefore be 35 years old.

Logistic regression using the results of two continuous tests

The situation becomes more complex when logistic regression models use more than one continuous test to determine the patient's probability of disease. For example, a decision rule about proceeding to chorionic villus sampling might consider NT as well as maternal age. Now, we move from a single-variable logistic regression model to a multivariable model. The single cut-off value (35 years old) is replaced by a cut-off line or curve (Fig. 8.11). The line represents the NT cut-off at each maternal age. We expect this line to have a negative slope; the NT threshold should decrease as the maternal age increases.

For Figure 8.11, we defined high risk of trisomy 21 as probability greater than 1%. In a 21-year-old woman, a fetus with NT of 3 mm is considered low risk (<1%

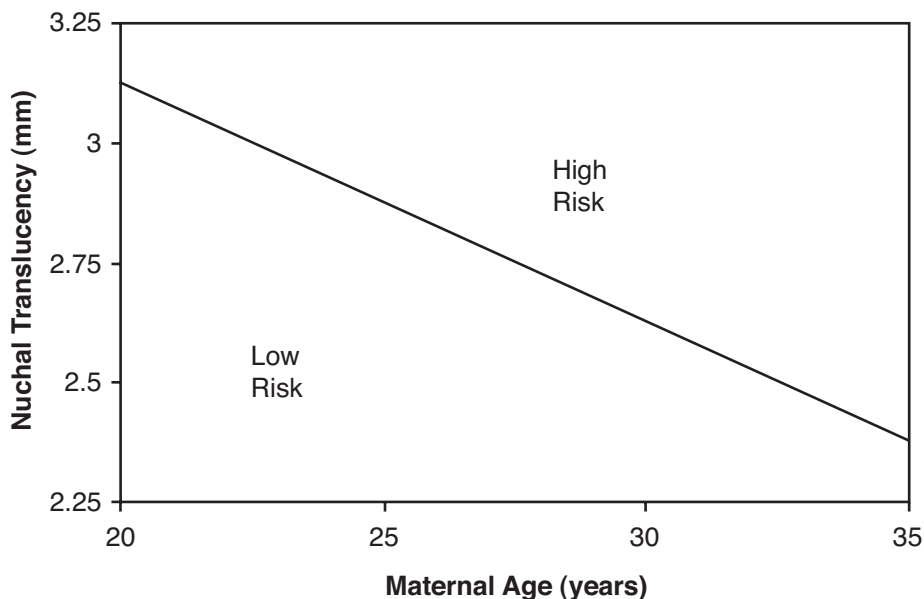


Figure 8.11 A hypothetical nomogram showing the combinations of maternal age and NT that identify fetuses at high risk for trisomy 21. In this nomogram, “high risk” is greater than 1% probability of trisomy 21 (Data abstracted from Nicolaides 2004, Figure 6, page 20).

probability of trisomy 21), but in a 34-year-old woman, a lower NT of 2.5 mm is considered high risk ($>1\%$ probability).

Clinical decision rules developed using logistic regression

Like the rule of Goldman et al. for predicting myocardial infarction, developed using recursive partitioning, a famous example of a clinical decision rule developed using logistic regression is also for predicting myocardial infarction, as well as unstable angina. This is the Acute Coronary Ischemia–Time Insensitive Predictive Instrument (ACI-TIPI) (Selker et al. 1991, 1998). The predictors in this logistic model include sex, age, existence/importance of chest pain as a presenting symptom and multiple ECG findings (Table 8.4).

As an example, a 55-year-old man with chest pain as his major symptom and new Q waves on his ECG but no ST or T wave changes would have $\text{Ln}(\text{odds})$ of acute coronary ischemia of $-3.93 + 1.23 + 0.88 + 0.71 + 0.67 - 0.43 + 0.62 = -0.25$, so the odds would be $e^{-0.25} = 0.78$ and the probability would be $0.78/1.78 = 44\%$. Although this is not practical for a clinician to calculate, the rule can be programmed into an ECG machine so that, if the technician enters a few items from the history, the estimated probability of acute coronary ischemia can be printed with the automated ECG analysis.

Another famous use of multiple logistic regression was the development of the PORT Pneumonia Score (Fine et al. 1997) to predict death in patients with

Table 8.4. Logistic regression coefficients from the ACI-TIPI model^{a,b}

Intercept	Coefficient	Multivariate OR
	−3.93	
Presence of chest pain	1.23	3.42
Pain major symptom	0.88	2.41
Male sex	0.71	2.03
Age ≤ 40	−1.44	0.24
Age > 50	0.67	1.95
Male > 50 years ^c	−0.43	0.65
ST elevation	1.314	3.72
New Q waves	0.62	1.86
ST depression	0.99	2.69
T waves elevated	1.095	2.99
T waves inverted	1.13	3.10
T wave + ST changes ^c	−0.314	0.73

^a From Selker et al. (1991).

^b The multivariate OR is obtained by exponentiating the coefficients.

^c This score includes two interaction terms. Male sex has an OR of 2.03 and age > 50 years has an OR of 1.95. Without an interaction term, the OR for being both male and over 50 would be $2.03 \times 1.95 = 3.96$. The OR of 0.65 for being both male and over 50 indicates that 3.96 is too high and that $0.65 \times 3.96 = 2.57$ is a better estimate.

Table 8.5. Calculation of the PORT score to predict likelihood of death among patients with pneumonia

Characteristic	Points assigned ^a
Demographic factor	+Age (years)
Age	
Sex	
Women	−10
Nursing home resident	+10
Coexisting illness	
Neoplastic disease	+30
Liver disease	+20
Congestive heart failure	+10
Cerebrovascular disease	+10
Renal disease	+10
Physical-examination findings	
Altered mental status	+20
Respiratory rate $\geq 30/\text{min}$	+20
Systolic blood pressure < 90 mm Hg	+20
Temperature $< 35^\circ\text{C}$ or $\geq 40^\circ\text{C}$	+15
Pulse $\geq 125/\text{min}$	+10
Laboratory and radiographic findings	
Arterial pH < 7.35	+30
Blood urea nitrogen ≥ 30 mg/dL	+20
Sodium < 130 mEq/L	+20
Glucose ≥ 250 mg/dL	+10
Hematocrit $< 30\%$	+10
Partial pressure of arterial oxygen < 60 mm Hg	+10
Pleural effusion	+10

^a A total point score for a given patient is obtained by summing the patient's age in years (age − 10 for women) and the points for each applicable characteristic. The points assigned to each predictor variable were based on coefficients obtained from the logistic-regression model.

pneumonia. The authors used the coefficients from their logistic regression model to create the point scoring system shown in Tables 8.5 and 8.6.

What happened to pre-test probability and misclassification costs? The clinician versus the decision rule

The output of the ACI-TIPI model as printed on the ECG header is a probability of acute coronary ischemia. Similarly, the recursive partitioning algorithms depicted in Figures 8.6 through 8.8 show disease probabilities next to the nodes. In Chapters 3 and 4, we always combined the test result with the pre-test likelihood of disease. In this chapter, we have estimated the probabilities from multivariable models assuming

Table 8.6. Mortality according to the PORT score, excluding lowest risk "Class 1" patients^a

Score	30-Day mortality
<71	0.6%
71–90	2.8%
91–130	8.2%
>130	29.2%

^a From Fine et al. (1997).

that our patients have a pre-test probability of disease that is the same as that in the sample used to develop the models. This is reasonable if the sample population is similar to our own or the model takes account of all the important variables that we would use to adjust our subjective pre-test probability estimate.

The clinician's advantage over a multivariable decision rule is the ability to adjust interpretation of test results based on the patient's pre-test probability of disease, if she knows that important variables have been left out of the model. The clinician's disadvantage is that this adjustment is done intuitively rather than mathematically, and without the benefit of the large dataset used to develop the decision rule.

The Ottawa Ankle Rules (Stiell et al. 1994) suggesting when radiographs can be deferred in patients with ankle injuries have been shown in a variety of settings to have high sensitivity for fracture while substantially reducing radiographs, relative to clinicians working without the benefit of the rules. This may be explained by the relative homogeneity (in terms of pre-test probability) of the population of patients with ankle injuries, or equivalently, by the rules' accounting for all the important predictors of finding a fracture on x-ray. It may be more difficult to account for all the important predictors of disease when the decision is whether to hospitalize a febrile child, or an adult with chest pain, syncope, or community-acquired pneumonia.

Clinical decision rules assume that the cost of failing to treat in the presence of disease and the cost of treatment in the absence of disease (B and C from Chapter 3) are the same from patient to patient. We have mentioned the clinician's ability to modify the decision threshold based on the pre-test probability of disease. But, the clinician also has the ability to adjust the decision threshold based on differing consequences of error. For example, failing to treat bacteremia in a 1-month-old has more serious consequences than failing to treat bacteremia in a 3-year-old; failing initially to treat bacteremia may also be worse if the family lives far from the hospital or has no home telephone. The above-mentioned risk threshold for fetal diagnostic procedures of 1 in 300 does not allow that failing to diagnose trisomy 21 may have different consequences for different women/couples/families. The ability to adjust for these differences is another potential advantage of the clinician over the decision rule.

Selecting tests to include in a decision rule

Thus far, we have focused on how to combine the results of several tests, not on which tests to include in a clinical decision rule. We want to include those tests with the greatest ability to discriminate between D+ and D− individuals (at reasonable cost and risk). These are also the tests that we want to do first in a recursive partitioning algorithm. As an oversimplified example, if your “rule” for predicting fetal trisomy 21 can only consist of one sonographic screening test, it should clearly be examination of the nasal bone rather than NT (at least as dichotomized at 3.5 mm). There are always a number of candidate variables that may be important in determining the probability of disease. In developing a clinical decision rule, we have to choose just a few of these variables. This variable selection is best done based on biological understanding and the results of past studies. Often, however, research studies measure many predictor variables, and there is no strong basis for narrowing down the large number of candidate variables to the handful that provide the most predictive power. Stepwise multivariate models can help with this: they either start with a large number of variables in the model and remove the least statistically significant variables one at a time (backwards), or start with no predictor variables and add variables one at a time, each time adding the one that is most significant (forwards). The resulting models may be those that best predict outcome in the particular dataset from which they were derived, but they generally will do less well in other datasets, as discussed below.

Importance of a derivation and a validation set

If you are allowed to choose combinations of tests and findings from a large number investigated, and then choose the cut-off that optimizes discrimination, you can often develop a prediction rule that works well in a specific sample (especially if the sample size is small). But this variable selection takes advantage of chance variations in the data that will no longer work to your advantage if you try to validate the prediction rule on a new dataset. As mentioned in Chapter 5, this is called “overfitting.” For example, Oostenbrink et al. (2000) used four history variables, four laboratory variables, and ultrasound result to predict vesicoureteral reflux among 140 children (5 years and younger) who had their first UTI. Their final prediction rule had an AUROC of 0.78; at the cut-off they chose, it had 100% sensitivity and 38% specificity for Grade III or higher reflux, which was found in 28 subjects in their sample. When another group attempted to validate the rule on a similar group of 143 children, sensitivity and specificity at the same cut-off were only 93% and 13% respectively, neither clinically nor statistically significant (Leroy et al. 2006).⁶

The way to avoid (or at least quantify) overfitting is to develop a clinical prediction rule on one (generally randomly selected) group of patients, called the “derivation set” and then test it on a second group, called the “validation set.” If overfitting occurred, the performance on the validation set will be substantially worse. If derivation and validation sets came from the same study, the investigator might be tempted to try again, tweaking the prediction rule so it performs better in the validation set. But, of course, this defeats the purpose of the validation set, and, in effect, makes the

⁶ A quick shortcut: any time sensitivity and specificity sum to 1, the test is useless. In this case the sum is 1.06.

whole study a derivation set. (There is a subtle example in this chapter's Problem 2.) Finally, even if a prediction rule performs well in a validation set randomly selected from the study population, additional validation is helpful to determine how well it performs in different populations and different clinical settings.

Summary of key points

1. When combining the results of multiple tests for a disease, it is only valid to multiply the LRs for the individual test results if the tests are independent.
2. Tests for the same disease are often nonindependent for three inter-related reasons:
 - a) they measure the same pathophysiologic aspect of the disease;
 - b) the diseased group is heterogeneous; and
 - c) the nondiseased group is heterogeneous.
3. The ideal way to use results from multiple different tests would be to empirically define an LR for each possible combination of results. However, the number of possible combinations often makes this impossible.
4. The main two methods used to combine results of multiple tests are recursive partitioning and multivariable logistic regression.
5. Developing a decision rule for combining multiple tests often involves variable selection – that is, choosing which tests to include in the rule.
6. The choice of variables when deriving a decision rule is particularly subject to chance variations in the sample (derivation) dataset, and therefore, validation of the rule in a separate, independent population is important.

References

- Cicero, S., G. Rembouskos, et al. (2004). "Likelihood ratio for trisomy 21 in fetuses with absent nasal bone at the 11–14-week scan." *Ultrasound Obstet Gynecol* **23**(3): 218–23.
- Fine, M. J., T. E. Auble, et al. (1997). "A prediction rule to identify low-risk patients with community-acquired pneumonia [see comments]." *N Engl J Med* **336**(4): 243–50.
- Goldman, L., E. F. Cook, et al. (1988). "A computer protocol to predict myocardial infarction in emergency department patients with chest pain." *N Engl J Med* **318**(13): 797–803.
- Hoffman, J. R., A. B. Wolfson, et al. (1998). "Selective cervical spine radiography in blunt trauma: methodology of the National Emergency X-Radiography Utilization Study (NEXUS) [see comments]." *Ann Emerg Med* **32**(4): 461–9.
- Hoffman, J. R., W. R. Mower, et al. (2000). "Validity of a set of clinical criteria to rule out injury to the cervical spine in patients with blunt trauma. National Emergency X-Radiography Utilization Study Group." *N Engl J Med* **343**(2): 94–9.
- Katz, M. H. (1999). *Multivariable Analysis: A Practical Guide for Clinicians*. Cambridge, Cambridge University Press.
- Laupacis, A., N. Sekar, et al. (1997). "Clinical prediction rules. A review and suggested modifications of methodological standards." *JAMA* **277**(6): 488–94.
- Lee, T. H., G. Juarez, et al. (1991). "Ruling out acute myocardial infarction. A prospective multicenter validation of a 12-hour strategy for patients at low risk." *N Engl J Med* **324**(18): 1239–46.
- Leroy, S., E. Marc, et al. (2006). "Prediction of vesicoureteral reflux after a first febrile urinary tract infection in children: validation of a clinical decision rule." *Arch Dis Child* **91**(3): 241–4.

- Newman, T. B., J. A. Bernzweig, et al. (2002). "Urine testing and urinary tract infections in febrile infants seen in office settings: the Pediatric Research in Office Settings' Febrile Infant Study." *Arch Pediatr Adolesc Med* **156**(1): 44–54.
- Nicolaides, K. H. (2004). *The 11–13+6 Weeks Scan*. London, Fetal Medicine Foundation.
- Oostenbrink, R., A. J. Van Der Heijden, et al. (2000). "Prediction of vesico-ureteric reflux in childhood urinary tract infection: a multivariate approach." *Acta Paediatr* **89**(7): 806–10.
- Pantell, R. H., T. B. Newman, et al. (2004). "Management and outcomes of care of fever in early infancy." *JAMA* **291**(10): 1203–12.
- Quinn, J. V., I. G. Stiell, et al. (2004). "Derivation of the San Francisco Syncope Rule to predict patients with short-term serious outcomes." *Ann Emerg Med* **43**(2): 224–32.
- Selker, H. P., J. L. Griffith, et al. (1991). "A tool for judging coronary care unit admission appropriateness, valid for both real-time and retrospective use. A time-insensitive predictive instrument (TIPI) for acute cardiac ischemia: a multicenter study." *Med Care* **29**(7): 610–27. [For corrected coefficients, see <http://medg.lcs.mit.edu/cardiac/tipicoef.htm>.]
- Selker, H. P., J. R. Beshansky, et al. (1998). "Use of the acute cardiac ischemia time-insensitive predictive instrument (ACI-TIPI) to assist with triage of patients with chest pain or other symptoms suggestive of acute cardiac ischemia. A multicenter, controlled clinical trial." *Ann Intern Med* **129**(11): 845–55.
- Snijders, R. J., K. Sundberg, et al. (1999). "Maternal age- and gestation-specific risk for trisomy 21." *Ultrasound Obstet Gynecol* **13**(3): 167–70.
- Stiell, I. G., R. D. McKnight, et al. (1994). "Implementation of the Ottawa ankle rules." *JAMA* **271**(11): 827–32.
- Wasson, J. H., H. C. Sox, et al. (1985). "Clinical prediction rules. Applications and methodological standards." *N Engl J Med* **313**(13): 793–9.

Chapter 8 Problems: multiple tests

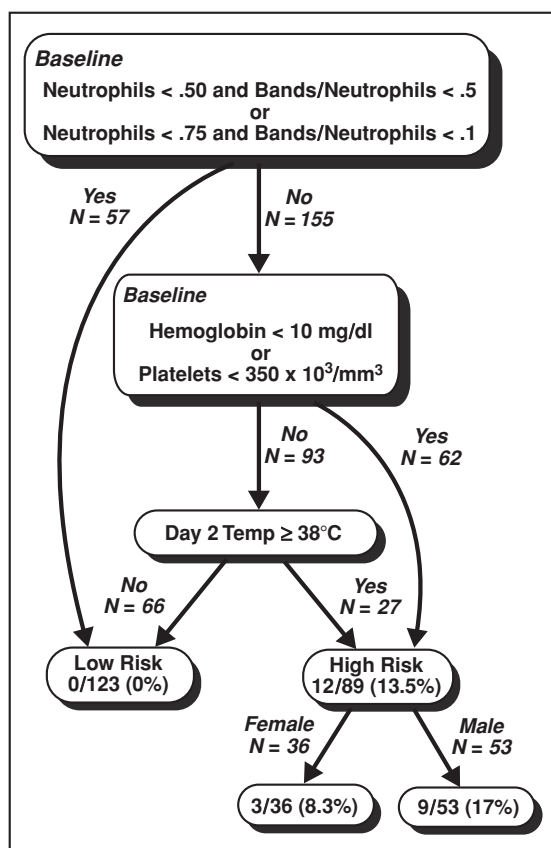
- Kaiser et al. (2002) performed a randomized study to evaluate CT and US as diagnostic tools for acute appendicitis in children. A total of 600 children with high clinical suspicion for acute appendicitis were enrolled; 283 were randomized to undergo US imaging only, and 317 children had both US and CT imaging (US was always done prior to CT scan).
 - Use the results summarized in the Table below to create two 2×2 tables to calculate the positive LR for CT and US results alone.

Comparison of imaging findings with outcomes of patients who underwent both US and CT

Outcome	Concordant results (n = 267)		Discordant results (n = 50)	
	US and CT positive	US and CT negative	US positive, CT negative	US negative, CT positive
Appendicitis	100	2	2	31
No appendicitis	3	162	8	9

Note: Data are numbers of patients, out of a total of 317 patients. Negative = findings were negative for appendicitis, positive = findings were positive for appendicitis. From Kaiser et al. (2002).

- b) If your patient's prior probability of acute appendicitis was 10%, what would be the posterior probability, given positive (both) US and CT scan results? Assume that the two tests are independent.
 - c) Now create a 4×2 table like Table 8.2 and calculate the LR_s for each possible US and CT result combination. Recalculate the posterior probability for a patient with a prior probability of 10% and positive results on both tests.
 - d) Are the two tests independent? Please explain how you know and suggest a possible biologic reason for your answer.
2. Kawasaki disease is an acute febrile illness in children of unknown cause that includes a rash, conjunctivitis, inflammation of mucous membranes, adenopathy, and swelling of hands and feet. Affected children are treated with intravenous immune globulin (IVIG) to prevent coronary artery aneurysms, the most serious complication of the disease. Using data from the intervention groups of two randomized controlled trials of IVIG, Beiser et al. (1998) developed an instrument to predict which children with Kawasaki disease would develop coronary artery aneurysms. The predictive instrument they developed is shown below (from Beiser et al. 1998. Used by permission.):



Neutrophils (also known as polymorphonuclear leukocytes) are one kind of WBC. Bands are immature neutrophils. "Neutrophils <0.5" means that, based on the WBC count differential, <50% of the white cells are neutrophils. "Bands/neutrophils <.5" means that, of all the neutrophils, fewer than 50% are bands.

- a) What type of analysis do you think the investigators did to come up with this figure?
- b) Assume you are treating a child like those included in the study. His initial CBC shows a hemoglobin of 11.2 g/dL, 600,000 platelets, and 13,000 WBCs/mm³, with 8,000 (61.5%) neutrophils of which 1,000 (1,000/8,000 = 12.5%) are bands. On day 2 of the illness, his temperature is 38.1°C. Would you classify him as high or low risk?
- c) Now imagine the same patient, only this time he only has 5,000 neutrophils/mm³. He therefore is at low risk, regardless of his temperature on day 2. How does this help you manage your patient? For example, does this mean you don't need to treat him with IVIG?
- d) In a study such as this, it is important that the clinical prediction rule be validated on a group of patients separate from the group used to derive it. The abstract of the study states:

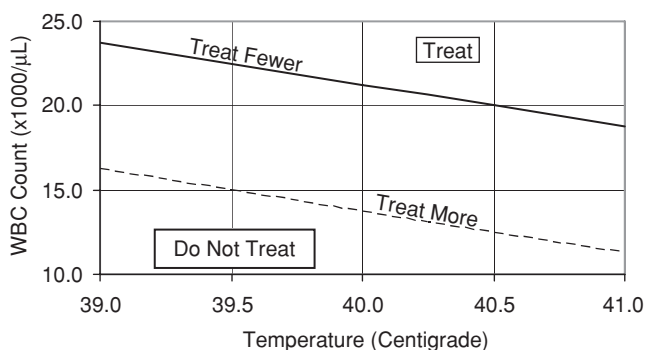
"The instrument was validated in 3 test data sets . . . [it] performed similarly in the 3 test data sets; no patient in any data set classified as low risk developed coronary artery abnormalities."

However, the "Methods" section states:

"We developed many such [sequential classification] processes, each using a different combination of risk factors . . . Instruments that performed well on the development data set were validated using each of the 3 test data sets."

Is there a problem here? If so, what is it and how would it affect the results?

3. Based on results reported by Lee and Harper (1998), we simulated a dataset of 8,756 well-appearing, febrile children, 3 to 36 months of age, with no obvious source of infection (Kohn and Newman 2001). Using logistic regression, we developed two prototype decision rules combining temperature and WBC count to determine the need for antibiotic treatment. The figure below (from Kohn and Newman 2001) presents these two decision rules.

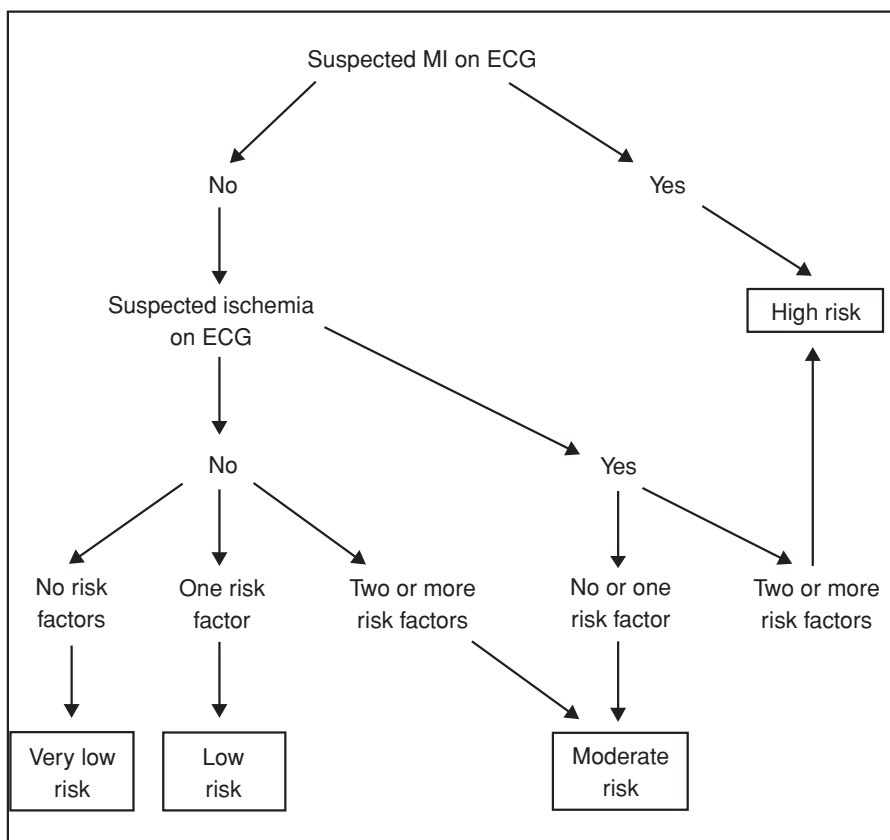


One rule ("treat more") prompts treatment when the odds of bacteremia are only 1:100; the other rule ("treat fewer") requires odds that are four times as high (1:25) before recommending treatment. In other words, the "treat more"

rule assumes that failing to treat a bacteremic child is 100 times worse than unnecessarily treating a child who is not bacteremic. The “treat fewer” rule assumes that such a false negative is only 25 times as bad as a false positive.

- Assume that you agree with the “treat fewer” rule, that is, that your threshold odds for treatment are 1:25. Would you treat a 5-month-old boy with a fever of 39.5°C and a WBC count of $20 \times 10^3/\mu\text{L}$ if his prior probability was similar to that in the study from which the figure above was derived?
 - What if the fever was 41°C and the WBC count still $20,000/\mu\text{L}$?
 - Explain why the slope of the line is negative (downward).
 - Explain why the “treat more” line is lower than the “treat fewer” line.
4. Goldman Prediction of Need for ICU.

Goldman et al. (1996) used recursive partitioning with a derivation dataset of 10,682 patients to divide emergency department chest pain patients into four risk groups: high, moderate, low, and very low. “Risk” refers to the risk of a major event (such as cardiac arrest, complete heart block, or cardiogenic shock) requiring intensive care.



Myocardial infarction (MI) was suspected on the ECG if it showed ST-segment elevation of 1 mm or more or pathologic Q waves in two or more leads, and these findings were not known to be old. Ischemia was suspected if the ECG showed

ST-segment depression of 1 mm or more or T-wave inversion in two or more leads, and these findings were not known to be old. Risk factors included systolic blood pressure below 110 mm Hg, rales heard above the bases bilaterally on physical examination, and known unstable ischemic heart disease, defined as a worsening of previously stable angina, the new onset of postinfarction angina or angina after a coronary-revascularization procedure, or pain that was the same as that associated with a prior MI. The difference between each adjacent pair of risk groups was significant ($P < 0.001$).

- a) A 67-year-old man presents to the Emergency Department with chest pain that he reports is “just like when I had my heart attack two years ago.” His BP is 140/80 mm Hg, lungs are clear, and ECG is unchanged from one done in the cardiologist’s office 6 weeks ago. What is his risk group?
- b) The data (from the derivation dataset) on risks in each of the four risk groups are given in this table:

Risk	Major Event	No Major Event	Total
High	222	812	1,034
Moderate	158	1,791	1,949
Low	55	1,456	1,511
Very Low	48	6,140	6,188
Total	483	10,199	10,682

What is the overall risk of a major event in this population?

- c) Assume that this system of risk stratification is reliable and that your population of chest pain patients has the same baseline risk of “major events” and is otherwise similar to Goldman’s derivation set. What is your patient’s risk?
- d) As it turns out, the validation set for this risk stratification system had a different overall prevalence of major events. It was 168/4,676 or 3.6%. The LRs, however, were consistent between derivation and validation. If your population risk (prior probability) is more like the validation set, what would be the risk for a high-risk patient?

References for problem set

- Beiser, A. S., M. Takahashi, et al. (1998). “A predictive instrument for coronary artery aneurysms in Kawasaki disease. US Multicenter Kawasaki Disease Study Group.” *Am J Cardiol* **81**(9): 1116–20.
- Goldman, L., E. F. Cook, et al. (1996). “Prediction of the need for intensive care in patients who come to the emergency departments with acute chest pain.” *N Engl J Med* **334**(23): 1498–504.
- Kaiser, S., B. Frenckner, et al. (2002). “Suspected appendicitis in children: US and CT—a prospective randomized study.” *Radiology* **223**(3): 633–8.

- Kohn, M. A., and M. P. Newman (2001). "What white blood cell count should prompt antibiotic treatment in a febrile child? Tutorial on the importance of disease likelihood to the interpretation of diagnostic tests." *Med Decis Making* **21**(6): 479–89.
- Lee, G. M., and M. B. Harper (1998). "Risk of bacteremia for febrile young children in the post-Haemophilus influenzae type b era." *Arch Pediatr Adolesc Med* **152**(7): 624–8.