

Prognostic tests and studies

Introduction

In previous chapters, we discussed issues affecting evaluation of diagnostic tests: how the reason to make a diagnosis may determine which tests should be done, how test reliability and accuracy are assessed, how to combine the results of tests with prior information to estimate the probability that a patient has a disease, and how to assess studies of diagnostic tests. In this chapter, we consider those same kinds of questions, with respect to prognostic tests.

Prognostic versus diagnostic tests

Prognosis is “a forecasting of the probable course and termination of an illness” (Webster’s Unabridged Dictionary 2001). The main difference between prognostic tests and diagnostic tests is that, with prognostic tests, a time dimension is involved. With diagnostic tests, we are concerned with determining who does and does not have a disease. In this chapter, we begin with people who have a disease and try to predict their prognosis – that is, what will happen to them in the future. Thus, studies of prognostic tests generally have to be longitudinal in nature. That is, they need to follow a group of patients over time and allow measurement of incidence rather than just prevalence.

As was the case with studies of diagnostic tests, in which we compared test results among patients with and without the disease (a dichotomous variable), the outcome variable of a prognostic test study is generally also dichotomous – survival versus death, disease-free survival versus disease recurrence, and so on; each subject either does or does not develop that outcome. However, due to the time component, the result of a prognostic test is often not dichotomous. That is, studies of prognostic tests, even when the outcome variables are dichotomous, do not generally report results like “will die” and “will survive.” Instead, they generally report quantitative results – for example, when a given marker is present, 1-year mortality is 30%,

compared with 60% when it is absent. Thus, whereas a diagnostic test gives either the right or the wrong answer for an individual patient, the accuracy of prognostic tests, whose results are translated into probabilities of an outcome, can generally only be assessed in groups of patients followed over time. If a test suggests that a patient has a 30% chance of dying in the next year, it is not clear when the patient would need to die in order for the test to give the “right” answer.

An exception to this rule is that, if the time period of interest is very well defined (and usually short), accuracy can be assessed on individual patients. For example, if the time period of interest is the days following resuscitation from a cardiac arrest, one might measure predictors of survival to hospital discharge. A predictive index could be dichotomized or clinical prediction rule developed, which would, in fact, be either right or wrong. When the time course for a prognosis is short, prognostic tests are like diagnostic tests.

It is also possible to assess the accuracy of a prognostic test in individual patients when the outcome variable of interest is continuous. For example, you might predict that a woman with osteoporosis will lose 0.5 cm of height per year, or that a pregnant woman with diabetes will have a 4-kg baby. For patients with incurable disease, an estimated survival time (typically in months) is also a continuous outcome. In the case of a continuous outcome, the accuracy in individual patients can be assessed by the difference between what was predicted and what was observed, and the mean and distribution of these differences can be studied in groups of patients. A graph with the difference between observed and predicted outcomes on the y-axis versus observed outcome on the x-axis produces a calibration plot very similar to the Bland–Altman plots used for method comparison studies (discussed in Chapter 2).

Prognostic tests versus risk factors

In Chapter 6, we distinguished between screening for unrecognized symptomatic disease, presymptomatic disease, and risk factors. Evaluating the effectiveness of screening for risk factors generally requires larger and longer studies, comparing outcomes in the screened and unscreened groups. The study must allow time for the screening test to accurately identify those at higher risk, for those subjects to receive an effective intervention to decrease that risk, and for some of them to develop the outcome.

Because prognostic tests are predicting associations between test results and outcomes over time, they are more similar to risk factor-screening tests than to diagnostic tests. However, in contrast with risk factors, prognostic associations are of interest even if they are not causal. As we will discuss in the next chapter, the best way to estimate the probability of disease is often by combining information from several different tests. Similarly, the best way to estimate prognosis is often by combining information from several variables. Multivariate techniques, such as logistic regression, are often useful for studies of prognostic tests, just as they are for observational studies of risk factors. However, with prognostic tests, this is often to determine to what extent a new, more difficult or more expensive test adds information to what

was already available, rather than to distinguish between causal associations and those due to confounding, which is the priority for studies of risk factors.

Studies of risk factors for disease generally yield relative measures of association: risk ratios, odds ratios, and hazard ratios. These measures express how many times more likely a patient with the risk factor is to develop the outcome than a patient without the risk factor.¹ They do not tell us the patient's absolute risk of the outcome, such as how likely this particular patient is to develop the outcome over the next 5 years. But patients with diseases want to know their absolute risks. They do not just want to know if their chance of dying in the next 5 years is half (or twice) as high as someone else's; they want to know what their chance of dying actually *is* – that is, their prognosis.

Quantifying the accuracy of prognostic tests

Calibration versus discrimination

Prognostic test accuracy has two dimensions: calibration and discrimination. Calibration refers to how well the probability estimated from the test result matches the actual probability, whereas discrimination refers to how well the test differentiates between patients more and less likely to have the outcome.

Because each individual patient will either have the outcome or not, calibration is measured by comparing the predicted probability estimated from results of prognostic tests on a group of patients to the actual probability – that is, to the proportion of a group of patients that develops the outcome in a specified time period. For example, if we assemble a cohort of HIV+ patients with CD4 counts <500, the predicted 5-year mortality might be 45%. If the observed mortality in that group of patients after 5 years were 40% to 50%, calibration would be good: the observed probability of mortality in the group would match the expected probability.

Calibration is typically measured by dividing the population into groups (often deciles of risk) and comparing predicted and observed frequencies of the outcome. This is most appropriately done by visual inspection, either of the numbers themselves or of a graph of observed versus expected probabilities, as illustrated in Box 7.1. Statistical tests of calibration are also available, but they have the disadvantage that they provide P-values for the discrepancy between observed and predicted probabilities but no summary estimate of “effect size” – that is, how far apart these probabilities actually are. Thus, if the sample size is very large, P-values can be statistically significant even when the calibration is good, and conversely, if the sample size is too small, poor calibration will not be statistically significant. (We discuss this problem with P-values and the distinction between P-values and effect size in Chapter 11.)

Just good calibration is not sufficient, however. If you were to include an entire population in the test of calibration and simply use the population value for probability of death from the disease as your estimate for each person, your calibration would be perfect – each person in the population would have the same, correct

¹ We will review the difference between risk ratios and odds ratios, and when it is important, in Chapter 9.

Box 7.1: Calibration and discrimination for prognosis of low back pain

Dutch investigators studied predictors of prognosis in patients presenting to general practitioners with low back pain (Jellema et al. 2007). They developed a clinical prediction rule that provided an estimated probability of an “unfavorable course,” defined as back pain perceived by the patient as at most “slightly improved,” at subsequent follow-up visits. The prediction rule was based on answers to a baseline questionnaire covering things like radiation of the pain, previous history of back pain, and general health. (Clinical prediction rules are discussed in Chapter 8.) They also asked the general practitioners to estimate the probability of restricted functioning at 3 months to the nearest 10% (i.e., on an 11-point scale: 0%, 10%, 20%, 30% ... 100%). The calibration of the two methods is illustrated in Figures 7.1a and 7.1b. Calibration was good for both – most of the points are close to the line that represents perfect calibration. However, you can see that discrimination of the clinical prediction rule (Fig. 7.1a) was better than the practitioners’ estimates, because it yielded a wider range of predicted probabilities. Remember, discrimination is the ability to move probabilities away from the mean probability toward 0 and 1. This improved discrimination was reflected in a higher area under the ROC curve (AUROC): 0.75 (95% CI 0.69 to 0.81) for the clinical prediction rule, compared with 0.59 (95% CI 0.52 to 0.66) for the general practitioner’s estimate.

probability of dying. However, discrimination would be poor. You would be unable to tell which people were most likely to die and which were least likely to die.

Discrimination refers to how well the test can separate the subjects’ probability of the outcome from the average probability of the group to values closer to zero (no chance of developing the outcome) and 1 (certain to develop the outcome). In the example of HIV+ people with CD4 counts <500, discrimination could be improved

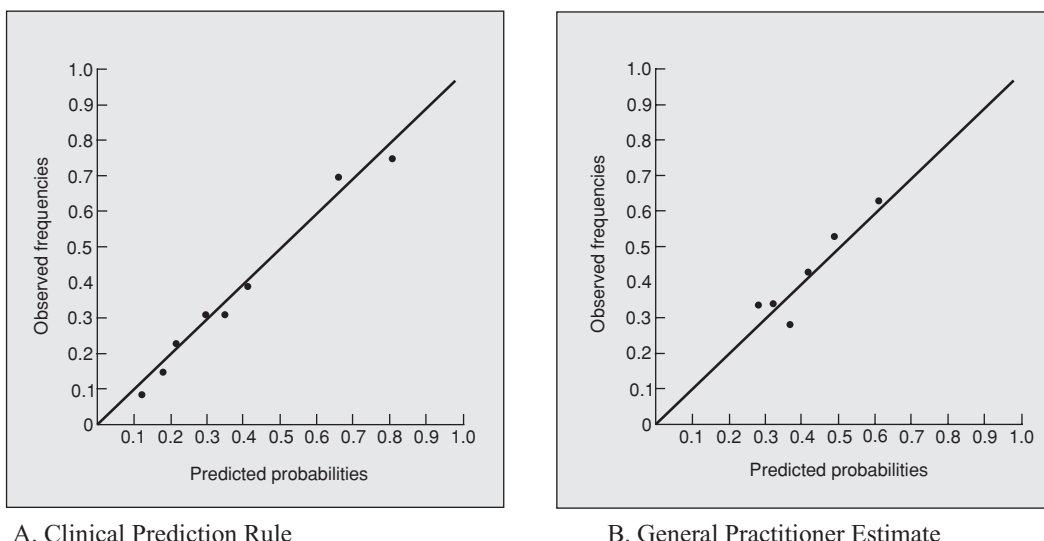


Figure 7.1 Predicted probability plotted against observed frequency of continued low back pain among patient seen by Dutch general practitioners. Both sets of predicted probabilities are well calibrated, but the clinical prediction rule discriminates better, as shown by a wider range of predicted probabilities (predicted probabilities closer to 0 and 1 for more subjects). From Jellema et al. (2007).

by further dividing the CD4 count into smaller categories, so that subjects with CD4 counts <50 , who have the worst prognosis, would not be lumped together with those with counts from 400 to 499, whose prognosis is better.

A commonly used approach to quantifying the discrimination of a prognostic test is to pick a particular time period (e.g., 5 years) and then treat the outcome as a simple dichotomous variable (e.g., alive or dead at 5 years). We can then express the discrimination of a prognostic test with our old friend from Chapter 4, the AUROC, where instead of comparing test results in disease and nondiseased, the results are compared in those who did and did not develop the outcome (e.g., 5-year survival), and the varying definition of a positive test is what traces out the ROC curve (Box 7.2). A test with perfect discrimination would produce

Box 7.2

Mackillop and Quirt (1997) performed a cohort study of oncologists' ability to predict "cure" (5-year disease-free survival) in 96 cancer patients. At the beginning of the study, the doctors assigned a probability of cure to each patient. After 5 years of follow-up, 26 patients had survived without recurrence. Figure 7.2 (*top*) shows their distribution of doctor-assigned probability of cure. Figure 7.2 (*bottom*) shows this same distribution for the 70 patients who had a recurrence or died.

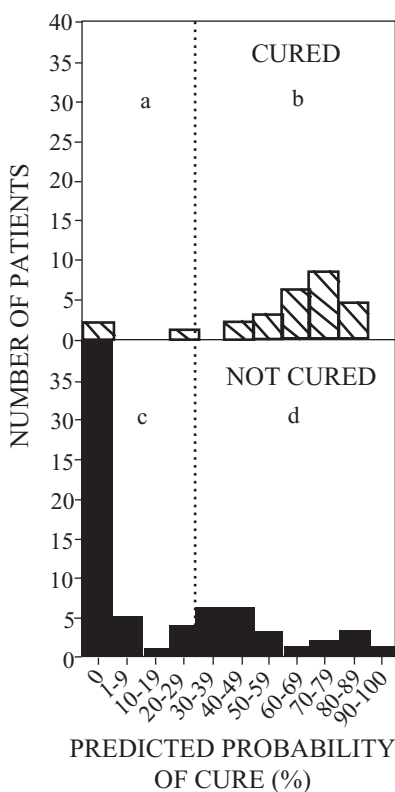


Figure 7.2 Frequency distribution of doctor-assigned, 5-year, disease-free survival ("cure") in 26 patients who did survive without recurrence (*top*) and 70 patients who suffered a recurrence or died (*bottom*). From Mackillop and Quirt (1997), Figure 3. Used by permission.

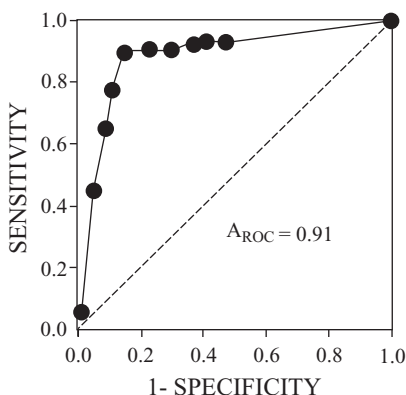


Figure 7.3 ROC curve summarizing doctors' ability to discriminate between patients who go on to survive for 5 years without recurrence and patients who die or have recurrences within 5 years. From Mackillop and Quirt (1997), Figure 4. Used by permission.

an ROC curve like a perfect diagnostic test: straight up and then straight across, with an AUROC of 1.0. We emphasize that the AUROC only measures how well the prognostic test discriminates between survivors and nonsurvivors; it says nothing about calibration. Recall from Chapter 4 that the ROC curve depends only on the ranking of individual measurements (in this case, risk estimates) and not their absolute values. Given any pair of patients – one who survived 5 years and one who did not survive – the AUROC is the probability that the survivor would be ranked more highly with regard to survival probability than the non-survivor.

For example, Figure 7.3 shows the ROC curve that results from sequentially raising the threshold for “will survive 5 years without recurrence” from 0% to 100% in the study described in Box 7.2. The AUROC is 0.91, meaning that, of all the ($26 \times 70 = 1820$) possible cured–uncured pairs, the cured patient was ranked by the oncologist as more likely to be cured than the uncured patient 91% of the time.

Although use of the AUROC to quantify discrimination is common and easy to understand, dichotomizing survival at one point leads to loss of information. It equates a death at 1 week with a death at 4.9 years, and a death at 5.1 years with >10-year survival. One approach to this problem is to make a whole family of ROC curves, for outcomes occurring at different time periods (e.g., 1-, 2-, 3-, 4-, 5-year survival, etc.).

Providing an estimate of discrimination is also difficult when not all outcomes are known. This can happen when subjects are lost to follow-up or die of other causes, as often happens in long-term studies of cause-specific mortality in older subjects. One approach to the latter problem is to focus on total mortality, but if many of the deaths are due to causes other than the disease of interest, the results will in part reflect the association (or lack thereof) between the prognostic factors and these other causes of death.

Risk ratios, rate ratios, and hazard ratios

Prognostic test studies often report results using risk ratios, rate ratios, or hazard ratios. These measures all express the likelihood of developing the outcome in people who have a risk factor compared with those who do not. As with the use of ROC curves, the use of risk ratios requires dichotomizing the outcome, which leads to loss of information about when the outcome occurred and about people who were lost to follow-up. Rate ratios and hazard ratios can take into account variable follow-up periods and times to the outcome, so these measures are preferable when follow-up time is variable and there are issues such as competing mortality. Rate ratios and hazard ratios will be biased if loss to follow-up is associated with both the prognostic factor and the outcome of interest. Also, as we noted above, risk, rate, or hazard ratios alone are less clinically useful, because most clinical decisions should be based on absolute rather than relative risks.

Assessing the value of prognostic information

Much prognostic information is available at little cost or risk; variables such as age, current symptom burden, extent of disease, and functional status are often highly predictive of prognosis. On the other hand, when we are considering risky or expensive prognostic tests, we should consider how to assess the value of such tests in order to make better decisions about their use.

Unfortunately, quantifying (or even estimating) the value of prognostic information is difficult, because the proper effect of prognostic information on decisions may be unclear. For example, if you inform a patient that her estimated 5-year survival from breast cancer is 80%, but a new test might change that to either 70% or 90%, how will that affect her subsequent decisions? The patient with only a 70% survival probability might opt for treatment with a more aggressive regimen, but this more aggressive treatment may not have been shown to be effective. If, in fact, the treatment is ineffective or even does harm, then the test that led to your giving that treatment was (at best) of no value. Similarly, if the test result led to treatment being withheld in the patient with 90% predicted survival, and that patient would have benefited from the treatment, again, the test was not valuable, and, in fact, caused the patient harm. Remember, the main value of a test is in allowing us to make better decisions. This implies enough knowledge of risks, costs, and benefits of treatment to create a treatment threshold. But if the treatment threshold is not known with any precision at all (e.g., it is thought that treating is worthwhile if the risk of death over the next 5 years is somewhere between 10% and 80%), then performing prognostic tests that allow better estimation of prognosis within this range does not help with the treatment decision.

The gap between prognostic information and the information needed to make decisions is illustrated by a study of breast cancer patients' estimates of their prognosis with and without adjuvant chemotherapy (Ravdin et al. 1998). The patients' estimates of recurrence risk and the benefits of treatment were much too high,

translating to an absolute risk reduction of about 30%, when the correct number was closer to 7.5%. However, the median absolute reduction of risk at which the women would choose adjuvant chemotherapy was only 0.5% to 1%, so in most cases, it does not appear that inaccurate estimates of prognosis affected their decisions. Similarly, a randomized trial that included provision of accurate prognostic information to patients and clinicians found no improvement in patient care or outcomes (SUPPORT 1995).

We should still make every effort to help patients obtain an accurate prognosis. Patients may value prognostic information beyond its ability to help with clinical decision making, and in some patients, accurate information could favorably affect decisions. But if the prognostic test is difficult or expensive, we should consider whether its result will be helpful and for what purpose.

You can think about Test–No test and Test–Treat thresholds for a prognostic test in the same way as for a diagnostic test if you dichotomize the outcome (e.g., 5-year mortality).

1. First, think about what treatment or management decision(s) the test is supposed to help resolve.
2. Then, consider the concept of the “treatment threshold.” This is the post-test probability of developing the outcome that determines whether or not to initiate treatment. For example, at what (posterior/post-test) probability of a bad outcome (e.g., death, recurrence, etc.) would you recommend more aggressive treatment?
3. Now use available clinical data to estimate a prior (pre-test) probability – in this case, not that the patient *has* a disease, but instead that the patient will *develop* the particular outcome you wish to avoid within a specified time period. This is based on your history, physical examination, and all other information except the prognostic test you are considering.
4. Now consider the probability that the additional prognostic information will change your decision, and estimate the value of that change. For example, if having a positive test for the latest pricey genetic marker would increase the estimated risk of recurrence from 30% to 60%, would that change your mind about what treatment you recommend? If so, what is the chance that the marker will be positive, and how confident are you that the change in treatment would improve the patient’s life expectancy?

All of this is difficult to do – much more difficult than deciding whether to perform a throat culture to help decide whether to treat with penicillin. But this is what you would need to do to be able to assign a value to a prognostic test. This difficulty is what makes us generally skeptical about expensive prognostic tests.

Critical appraisal of studies of prognosis and prognostic tests

As was the case with diagnostic tests, we summarize several issues that arise commonly in evaluating studies of prognosis. Most of these issues arise for other types of studies as well, so we will only highlight them briefly here.

Sample selection and generalizability

It is important to understand how the subjects in a study of prognosis were identified. Patients whose disease is identified by screening will generally do better than those who present to community physicians with symptoms, who will do better than those referred to tertiary centers. Prognostic studies that obtain their subjects from these different sources will yield different results. This occurs both as a result of the selection process leading to referral and as a result of differences in the time point in the course of the disease at which they are enrolled. Therefore, a key question to ask in appraising a study of a prognostic test is how the subjects came to be included in the study at the time they did, and whether the patients to whom you wish to generalize the results would have any reason to be under-represented in the study.

Effects of treatment

If the illness being studied is treatable, then its prognosis will be affected not only by the selection process for subjects in the study, but also by how they were treated. If you want to generalize to your patients, not only must the patients in the study be similar to yours at the onset of the study, they need to be treated similarly throughout the course of the study. Watch out for studies in which differences in treatment due to perceived differences in prognosis muddy the ability to determine what factors actually are most predictive of prognosis. A study of prognostic factors in elderly ICU patients would likely find associations with mortality either for factors that really do predict mortality or for factors that treating physicians strongly believe will predict mortality, because having many of the latter factors may lead to withdrawal of life support.

Loss to follow-up

Patients lost to follow-up add uncertainty to estimates of prognosis. This is a particular problem if there is reason to believe their prognosis might be different from the prognosis of those whose outcome is known. One way to get some limits on the degree to which subjects lost to follow-up could affect the study results is to recalculate the proportion with the outcome (e.g., survival), first assuming that all those missing had the outcome, and then assuming none did. For example, consider a follow-up study of 200 patients, of whom 120 survived, 60 died, and 20 could not be accounted for at 5 years. If the subjects lost to follow-up are simply not counted, survival would be $120/180 = 67\%$. If all 20 patients lost to follow-up are assumed to have died, the observed survival would be $120/200 = 60\%$, and if none had died, it would be $140/200 = 70\%$. Thus, the largest effect of loss-to-follow-up in this example would be to increase apparent survival from 60% to 67%, or decrease it from 70% to 67%. If this very conservative approach still yields useful prognostic information, you are on firm ground. A less conservative approach would be to assign the missing subjects the lowest and highest plausible event rates (rather than the rates of 0% and 100% used above).

This same approach applies when estimating possible effects of loss-to-follow-up on an estimate of the relative risk associated with presence of a particular prognostic

factor. The largest possible effect that loss to follow-up could have can be estimated by assuming that all those lost to follow-up had a bad outcome in one group and a good outcome in the other group, or vice versa.

Blinding

We will discuss blinding in greater detail in Chapter 9. For now, we note that blinding is most important for subjective outcomes and when the prognostic factors of interest may affect treatment. For example, consider a study of predictors of the need for hospitalization in children presenting to the emergency department with acute asthma. If the goal is to study the prognostic value of the initial oxygen saturation, those making the decision to admit should be kept blinded to that value, so that it could not itself affect the decision to admit.

Overfitting

We mentioned this problem in Chapter 5, and drew an analogy with the Illinois 4th Congressional District (fig. 5.2). If you look at enough variables, you are bound to find some combination that is associated with adverse outcome in one particular sample. Similarly, if the investigators select the cut-off that defines abnormal results based on what works best in the sample, the value of the test will be overestimated. To be convincing, the prognostic factors identified in one study need to be restudied in another dataset, separate from the one from which they were derived, using the same cut-offs to define abnormal results (Hilsenbeck et al. 1992). The need for a *validation* dataset will be discussed again in Chapter 8.

Multiple and composite outcomes

A problem similar to overfitting occurs when multiple outcome variables are measured in a study, and the ones that are predicted by the test are selectively highlighted or reported. This multiple comparison problem also arises in randomized trials, and will be discussed in Chapters 9 and 11. Similarly, sometimes several outcomes are grouped together into a composite outcome – again, a common strategy in clinical trials. As in a randomized trial, it is important to know whether the investigators specified the composite outcome in advance, and whether there is evidence that the composite outcome variable might be dominated by a more common or more subjective but less important outcome, such as nonfatal myocardial infarction or the development of unstable angina rather than death. If more subjective outcomes are predicted best, you should double check to make sure that they were ascertained by blinded observers. (Such blinding can be difficult, as previously discussed in Chapter 6, when we reviewed “sticky diagnosis” bias.)

Sample size

Especially if bad outcomes are rare, there may be too few cases to be able to learn much about factors that affect prognosis. From the patient’s point of view, this does not necessarily make the study problematic, because patients are most interested in *absolute* risk. If a large study has very few bad outcomes, the confidence intervals

around relative risks may be very wide, suggesting that the study did not provide much information. But the confidence intervals around absolute risks may be narrow enough to be clinically meaningful. We discuss this issue at length in Chapter 11.

Quantifying new information

It is easy to identify findings and markers that statistically significantly predict prognosis. However, the key questions are how much new information a test provides, beyond what was already known, and how valuable that information is. Watch out for two ways the apparent predictive ability of a test can be inflated. First, if measurements of other variables that predict prognosis are coarse or imprecise, the apparent contribution of the new test will be larger, because information from the other variables will be incompletely taken into account in multivariate models. Second, the apparent predictive ability of a test can be inflated by comparing risk at extremes of the test, such as reporting the hazard ratio for a comparison between the highest and lowest quintiles of the measurement. Box 7.3 illustrates both of these problems.

Publication bias

Publication bias occurs when studies that have favorable results are published preferentially over those that do not. Although publication bias is a problem for all types of studies, it may be a particular problem for studies of prognostic markers. This is fairly understandable – it is hard to get very excited about submitting or reading a paper about factors that are worthless for predicting prognosis. On the other hand, if you look at enough possible prognostic factors in enough different ways, it is easy to find some that are good predictors of prognosis. These positive factors may be mentioned in the abstract of a paper, and other researchers will be able to easily find any previous studies in which they were predictive by doing a PubMed search. In contrast, all of the possible prognostic factors that were *not* associated with outcome in a study will be harder to find. They may or may not be listed in a table or mentioned in the “Methods” section of the current paper, but more significantly, evidence of their lack of association with outcome is unlikely to be found with a PubMed search. Publication bias is a significant problem for meta-analyses of studies of prognostic tests (Kyzas et al. 2005).

Keep in mind that clinically useful information about prognosis does not just come from studies that focus primarily on prognosis. Much valuable information can be obtained from the outcomes in either control or treated groups in randomized trials (depending on whether the patient of interest will be treated or not). Randomized trials (as discussed in Chapter 9) have the advantages that ascertainment of outcome is more complete and more objective than is typical of less rigorous designs.

Genetic tests

Because there seems to be so much interest and excitement (and hype!) about new genetic tests, we should clarify how they differ from other tests discussed in this book. A large part of the excitement about genetic tests relates to the possibility of greater understanding of underlying molecular mechanisms of disease. The hope is

Box 7.3: Example of a prognostic test study

Paik et al. (2004) reported on the ability of a multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. They used the assay to create a "recurrence index," which they then classified as low-, intermediate-, or high-risk. The 10-year Kaplan–Meier estimates of distant recurrence rates were 6.8%, 14.3%, and 30.5% in the three groups, respectively. When entered into a Cox proportional hazard model, the recurrence index was a strong, independent predictor of prognosis, with a hazard ratio of 3.21 per 50-point change in the index ($P < 0.001$).

A strength of this study is that all of the decisions about how to create the index from the results of individual gene tests, including the cut-offs, were made in advance. This should reduce overfitting. However, the reported hazard ratio of 3.2 is impossible to interpret without knowledge of the meaning of a 50-point change in the index. (The hazard ratio for a 25-point change would be $\sqrt{3.2}$, or about 1.8.) In this study, a 50-point difference in the index was a large difference: 51% of the subjects had scores less than 18 and only 12% had scores over 50. On the other hand, the authors simply dichotomized age (at 50 years) and tumor size (at 2 cm).² By failing to capture all information in these covariates, they may have inflated the apparent predictive power of their new index. A Letter to the Editor by Goodson (2005) brings up a similar point with respect to the pathological grading of the tumors. Again, if the pathologists grading the tumors are not very good at that task, the recurrence index will look better in comparison. Supplementary appendices to the paper indicate that the agreement on tumor differentiation (in three categories) was only fair ($\kappa = 0.34\text{--}0.37$), supporting Goodson's concern. Finally, the authors do not indicate the degree to which adding their test to what was already available improved discrimination, how this would improve decisions, or how these better decisions might improve outcomes. These are relevant considerations because, at the time of the study, the test (patented and/or owned by many authors of the study; Paik et al. 2004) was being sold for \$3,500 (Tanvetyanon 2005).

that, by identifying alleles of specific genes that cause or predispose to disease, we may be able to learn what these genes do and understand how variations in their expression can lead to ill health. Although so far the track record of successes in this area is underwhelming, there is no doubt that some genetic tests have value for this purpose. Because the goal in that situation is improved understanding of disease rather than assisting with clinical decisions, assessment of these tests and the studies that describe them requires specific content knowledge about the underlying biology and is not covered in this book.

In contrast, other genetic tests may have the potential to improve health by allowing better estimates of the probability of various diseases either being present already or developing in the future or of the prognosis of existing disease. The evaluation and interpretation of these genetic tests is the same as for any other test – it involves asking the same questions about the information different test results provide: how likely a particular patient is to have a result that is informative, how the test will improve clinical decisions, and the estimated impact of these improved decisions on clinically relevant outcomes.

² The investigators could have treated tumor size the way they treated their recurrence index, as a continuous variable, and reported the hazard ratio per 10-cm increase in tumor size!

In interpreting studies of genetic tests, and gauging which of the two purposes above may be most relevant, it is helpful to ignore low P-values and look for clinically meaningful measures of effect size. For example, consider a recent report of risk alleles for multiple sclerosis (MS) identified by a genome-wide study (Hafler et al. 2007). No disease-causing mutations for MS have been identified; it is thought that multiple common polymorphisms work in concert to increase susceptibility to the disease. The investigators reported associations between MS and multiple single-nucleotide polymorphisms. Most P values for the single-nucleotide polymorphisms they found were in the 10^{-4} to 10^{-8} range, although the authors reported a P-value of 8.94×10^{-81} for the HLA-DRA locus.³ However, the corresponding odds ratios for most of the risk alleles were only 1.08 to 1.25, and the odds ratio for the HLA-DRA locus was only 1.99. It is hard to make a case that odds ratios of this magnitude could be helpful clinically, and the authors do not do so. Rather, the hope is that these results may contribute to better understanding of the pathogenesis of MS.

Summary of key points

1. Prognostic tests differ from diagnostic or screening tests because their goal is to predict events that may happen in the future, rather than to identify conditions already present.
2. Studies of the value and accuracy of prognostic tests generally require longitudinal follow-up of groups of patients.
3. How the groups are selected and the completeness of follow-up are important aspects of the critical appraisal of such studies.
4. The potential value of prognostic tests is related to both their *calibration* to the actual prognosis of the patient and their *discrimination* between those more and less likely to develop the outcome.
5. Prognostic information can be summarized with baseline (absolute) risk, risk ratios, rate ratios, hazard ratios, and/or ROC curves for outcomes at various points in time.
6. The value of invasive or expensive tests used to assess prognosis depends on what decisions the additional prognostic information will help with, the importance of these decisions, and the likelihood that they will be changed by a more accurate estimate of prognosis.
7. Genetic tests whose purpose is to inform clinical decision making are critically appraised and used in the same way as other prognostic tests.

References

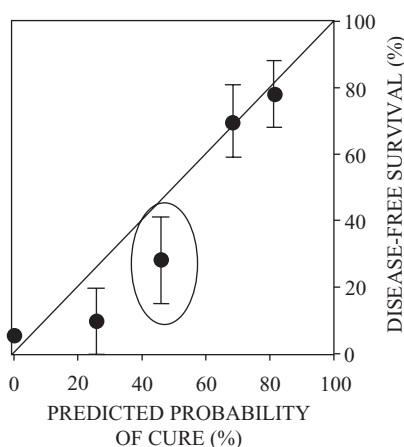
- Goodson, W. H., 3rd (2005). "Molecular prediction of recurrence of breast cancer." *N Engl J Med* 352(15): 1605–7; author reply 1605–7.
- Hafler, D. A., A. Compston, et al. (2007). "Risk alleles for multiple sclerosis identified by a genomewide study." *N Engl J Med* 357(9): 851–62.

³ We find it amusing that the significance was reported with 3 digits when the exponent was – 81!

- Hilsenbeck, S. G., G. M. Clark, et al. (1992). "Why do so many prognostic factors fail to pan out?" *Breast Cancer Res Treat* **22**(3): 197–206.
- Jellema, P., D. A. Van Der Windt, et al. (2007). "Prediction of an unfavourable course of low back pain in general practice: comparison of four instruments." *Br J Gen Pract* **57**(534): 15–22.
- Kyzas, P. A., K. T. Loizou, et al. (2005). "Selective reporting biases in cancer prognostic factor studies." *J Natl Cancer Inst* **97**(14): 1043–55.
- Mackillop, W. J., and C. F. Quirt (1997). "Measuring the accuracy of prognostic judgments in oncology." *J Clin Epidemiol* **50**(1): 21–9.
- Paik, S., S. Shak, et al. (2004). "A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer." *N Engl J Med* **351**(27): 2817–26.
- Ravdin, P. M., I. A. Siminoff, et al. (1998). "Survey of breast cancer patients concerning their knowledge and expectations of adjuvant therapy." *J Clin Oncol* **16**(2): 515–21.
- SUPPORT (1995). "A controlled trial to improve care for seriously ill hospitalized patients. The study to understand prognoses and preferences for outcomes and risks of treatments (SUPPORT). The SUPPORT Principal Investigators." *JAMA* **274**(20): 1591–8.
- Tanvetyanon, T. (2005). "Molecular prediction of recurrence of breast cancer." *N Engl J Med* **352**(15): 1605–7; author reply 1605–7.
- Webster's unabridged dictionary (2001). New York, Random House Reference.

Chapter 7 Problems: prognostic tests

- Box 7.2 presents the findings from a study of oncologists' ability to discriminate between those cancer patients who are likely to survive, disease-free, for 5 years and those who are likely to die or have a recurrence (Mackillop and Quirt 1997). The authors assessed calibration as well as discrimination by comparing the predicted probability of cure to the actual disease-free survival rate in 5 subgroups of the 96 patients.



From Mackillop and Quirt (1997). Used by permission.

- The circled point represents doctor-predicted versus actual cure for 12 patients to whom the doctors assigned a 40% to 59% chance of cure. Were the doctors overly optimistic or pessimistic?

- b) What about the subgroup of patients ($N = 42$) to whom the doctors assigned a 0% chance of cure (i.e., classified as incurable)? Were the doctors right?
- 2. Cytomegalovirus (CMV), like the chickenpox virus, varicella, is a virus to which most women have already been exposed before pregnancy. However, as with varicella, a pregnant woman with a first-time infection can pass the virus to the fetus. This congenital infection can lead to complications, including hearing loss. One way to test for congenital infection is to look for the virus in the baby's urine (viuria). Fowler et al. (1992) studied the risk of hearing loss in infants and children with congenital CMV.⁴

Here is an excerpt from the paper's Methods section.

Of the children in the study, 172 were identified through the obstetrical services at hospitals where we screened newborns for viuria to detect congenital CMV infection. Twenty-five additional congenitally infected newborns were referred to us by other hospitals or because the mother had evidence of infection on serologic screening or had illness during pregnancy or because elevated cord-blood levels of IgM or rheumatoid factor or symptoms of congenital infection in the newborn led to virologic testing.

- a) Comment on the inclusion criteria for the $172 + 25 = 197$ subjects in the study. What effect might the sample selection process have had on the results?
- b) Overall, hearing loss occurred in 24 of 197 (12%) infants in this study. If you are seeing an infant with congenital CMV picked up on routine screening, what would you calculate as the lowest and highest risks of sensorineural hearing loss consistent with this study? (Ignore random error, i.e., do not do confidence intervals. We are just looking for the effect of the sampling scheme.)
- c) Repeat part (b), but this time assume the infant was referred because of symptoms of congenital CMV.
- 3. In Box 7.1, we summarized a Dutch study of prognosis of low back pain (Jellema et al. 2007). In that study, a clinical prediction rule derived by the authors on the 314 subjects in the study had better discrimination than the general practitioner's (GP's) estimate of prognosis. Nonetheless, the authors concluded that "risk estimation by GPs . . . at present, seems to be the best available option."
- a) Why do you suppose this conclusion was so cautious? Do you agree?
- b) A close look at Figure 7.1b shows that there are only 6 points on the graph of predicted versus observed probabilities of a bad outcome. Normally, predicted probabilities are divided into deciles for these plots. If you assume that is the case, can you explain why there are only 6 points on this graph? (Hint: Recall the GPs rated their estimated probability rounding to the nearest 10%.)
- 4. Greenland et al. (2004) recently compared the Framingham Risk Score (FRS), obtained from history and physical examination and lipid levels, with a Coronary

⁴ Thanks to Ruth Gilbert and Stuart Logan then of the Evidence-based Child Health Centre in London for this example.

Artery Calcium Score (CACS) obtained from CT scanning in 1,461 asymptomatic adults at least 45 years old.

The FRS is an estimate of the 10-year risk of nonfatal myocardial infarction or death. The authors found that the CACS was predictive of this combined outcome among those with a FRS of more than 10%, but not in those with an FRS less than 10%, and they recommended against doing CACS when the FRS was less than 10%.

In a Letter to the Editor, Pletcher et al. (2004) wrote:

“In fact, such an interaction would be difficult to detect, and this study adds little evidence, given the low number of persons in the study with an FRS < 10% ($n = 98$) and the low number of events in this subgroup ($n = 1$).”

- a) Do you think this lack of power affects the conclusion that CACS is not indicated in this low-risk group?
 - b) Pletcher et al. also point out that the FRS was less predictive of events in the Greenland et al. study, compared with previous studies, and postulated that this could occur if treatment decisions based on the FRS blunted its predictive ability. What could the authors do to address this possibility?
5. People with cancer that spreads to bone can have fractures and severe bone pain. Brown et al. (2003) investigated a measure of bone resorption, urinary N-telopeptide excretion (Ntx), as a predictor of these complications. Bisphosphonates are drugs used to strengthen bones and reduce fractures in people with osteoporosis.

From the abstract:

A total of 121 patients had monthly measurements of Ntx during treatment with bisphosphonates. All skeletal-related events, plus hospital admissions for bone pain and death during the period of observation, were recorded . . . **Patients with baseline Ntx values ≥ 100 nmol/mmol creatinine (representing clearly accelerated bone resorption) were 19.48 times (95% CI 7.55, 50.22) more likely to experience a skeletal-related event/death during the first 3 months than those with Ntx < 100 ($P < 0.001$).** In a multivariate logistic regression model, Ntx was highly predictive for events/death. N-telopeptide appears useful in the prediction of patients most likely to experience skeletal complications and thus benefit from bisphosphonate treatment.

- a) The “events” that the Ntx predicted included death. How might this have affected the results?
- b) Here are the study results at 3 months:

NTX	Skeletal Complication (0-3 months)		Total
	Yes	No	
≥ 100	41	15	56
<100	8	57	65
Total	49	72	121

Do you agree with the authors’ statement? (“Patients with baseline Ntx values ≥ 100 nmol/mmol creatinine . . . were 19.48 times . . . more likely to experience

a skeletal-related event/death during the first 3 months than those with Ntx <100...”)

- c) The last sentence of the abstract states: “N-telopeptide appears useful in the prediction of patients most likely to experience skeletal complications and thus benefit from bisphosphonate treatment.”

Do you agree that the study provides information on who might most benefit from bisphosphonates? Why or why not?

6. *TP53* is the gene for tumor-suppressor protein p53. In a multicenter, 7-year prospective cohort study, disruptive *TP53* mutations in tumor DNA (i.e., mutations leading to loss of function of p53) were associated with reduced survival after surgical resection in patients with squamous-cell cancer of the head and neck (Poeta et al. 2007).

- a) Of the 420 subjects, 232 had died by the end of the follow-up period. Of these, 121 died from head and neck cancer, 62 from other causes, and 49 from unknown causes. The authors used overall survival as the outcome for all analyses. How would the use of overall (vs. cause-specific) survival affect the results?
- b) How else could the authors have handled the subjects who died of other and unknown causes?
- c) One question that arises for genetic tests is how much *new* information they provide. For example, if disruptive *TP53* mutations worsened prognosis by leading to more advanced stage at presentation, much of the prognostic information from *TP53* might be captured from stage at presentation. In fact, in this study, the nodal stage at presentation was highly predictive of survival. Bivariate (just one variable plus the outcome) and multivariate hazard ratios for nodal stage (N1–N3 vs. N0 or NX) and *TP53* (disruptive mutation vs. no mutation) are shown in the table below.

Prognostic Factor	HR (95% CI)	
	Bivariate	Multivariate
Nodal Stage N1–N3	2.0 (1.4–2.4)	2.4 (1.8–3.3)
Disruptive <i>TP53</i> mutation	1.7 (1.3–2.4)	1.7 (1.3–2.4)

What can you conclude about whether the *TP53* gene provides *new* information about prognosis in head and neck cancer patients?

- d) Assuming the hazard ratios reported in this study are valid and generalizable, what else would you need to know in order to decide whether to order this test on your patients?

References for problem set

- Brown, J. E., C. S. Thomson, et al. (2003). “Bone resorption predicts [0.3k] for skeletal complications in metastatic bone disease.” *Br J Cancer* **89**(11): 2031–7.

- Fowler, K. B., S. Stagno, et al. (1992). "The outcome of congenital cytomegalovirus infection in relation to maternal antibody status." *N Engl J Med* **326**(10): 663–7.
- Greenland, P., L. LaBree, et al. (2004). "Coronary artery calcium score combined with Framingham score for risk prediction in asymptomatic individuals." *JAMA* **291**(2): 210–5.
- Jellema, P., D. A. Van Der Windt, et al. (2007). "Prediction of an unfavourable course of low back pain in general practice: comparison of four instruments." *Br J Gen Pract* **57**(534): 15–22.
- Mackillop, W. J., and C. F. Quirt (1997). "Measuring the accuracy of prognostic judgments in oncology." *J Clin Epidemiol* **50**(1): 21–9.
- Pletcher, M. J., J. A. Tice, et al. (2004). "Use of coronary calcification scores to predict coronary heart disease." *JAMA* **291**(15): 1831–2; author reply 1832–3.
- Poeta, M. L., J. Manola, et al. (2007). "TP53 mutations and survival in squamous-cell carcinoma of the head and neck." *N Engl J Med* **357**(25): 2552–61.