Causal diagrams and diagnostic tests

How good is a test in diagnosing a condition or a disease (*D*)? What are the chances that I truly have the disease (D_{TRUTH} = Sick) if the test came back positive (D_{TEST} = Sick)? Should I stop worrying if the test was negative (D_{TEST} = Not sick)?

These questions and the like have been answered by estimating some of the following probabilities: sensitivity, specificity, false positive (or negative) "rate" (a common misnomer), and positive (or negative) predictive value. The computation is usually illustrated using a 2x2 table and simple notation:

Table 1.

		D_{TEST}	
		Sick (positive test)	Not sick (negative test)
D_{TRUTH}	Sick	a	b
	Not sick	c	d

Sensitivity	a/(a+b)
False negative "rate"	b/(a+b)
Specificity	d/(c+d)
False positive "rate"	c/(c+d)
Positive predictive value	a/(a+c)
Negative predictive value	d/(b+d)

The causal structure

A moment reflection, however, should make us wonder why the words "cause" and "effect" were not mentioned above, nor in countless lectures, articles, and textbooks on the topic. What is the causal structure behind the using of a test to diagnose a disease? Is the topic disconnected from the realm of causality?

No, it is not. And the causal diagram is simple, at least to begin with. If the binary variable D_{TRUTH} contains the true values and the binary variable D_{TEST} contains the test-based values, then $D_{TRUTH} \rightarrow D_{TEST}$. True disease status affects the test's classification, as in any classic measurement.[a]

[a] Note that if the test's result is not binary, we also have a an intermediary step: $D_{TRUTH} \rightarrow$ *TEST RESULT* $\rightarrow D_{TEST}$

Which takes us to the next question, highly relevant to judging the quality of a test: What is the magnitude of the effect of D_{TRUTH} on D_{TEST}? *The stronger the effect, the better the test* because a stronger effect of one binary variable on another implies that the value of the latter (here, D_{TEST}) is more likely to match the value of the former (D_{TRUTH}) — if a direct association is maintained (that is, being sick tends to generate a "positive" test.)[b] But neither sensitivity, nor specificity, nor any other proportion above is a measure of effect! They are all measures of frequency or probability, not measures of association.

How can we estimate the effect $D_{TRUTH} \rightarrow D_{TEST}$?

Assuming, for a start, no confounding or other biases, we may safely estimate the effect of D_{TRUTH} on D_{TEST} by the so-called marginal association between the two variables, by non-manipulated probability ratios or probability differences. (No computed association is truly "marginal", but I will leave this point aside.) Considering D_{TRUTH} = Sick as "exposed", and D_{TEST} = Sick as "diseased", we can learn about the quality of a test by computing the following measures of effect:

Probability ratio (PR) of a positive test:

$$\mathrm{PR}(D_{TEST} = \mathrm{Sick}) = \frac{a/(a+b)}{c/(c+d)}$$

Which is the sensitivity divided by the false positive proportion.

Probability ratio of a negative test:

$$\mathrm{PR}(D_{TEST} = \mathrm{Not\ sick}) = \frac{b/(a+b)}{d/(c+d)}$$

Which is the false negative proportion divided by the specificity.

Note that the effect on the probability ratio scale need not be identical for the two values of D_{TEST}. If you prefer to avoid reconciliation of two estimates, compute the odds ratio (OR) of a positive test:

$$\mathrm{OR}(D_{TEST} = \mathrm{Sick}) = \frac{a/b}{c/d}$$

[b] A strong effect may also imply a good test when the association is inversed, provided we interpret the test's result correctly.

The odds ratio of a negative test is the inverse:

$$\mathrm{OR}(D_{TEST} = \text{Not sick}) = \frac{c/d}{a/b}$$

The probability difference is also indifferent to the value of the effect variable:

$$\mathrm{PD}(D_{TEST} = \text{Sick}) = a/(a+b) - c/(c+d)$$

and

$$\mathrm{PD}(D_{TEST} = \text{Not sick}) = b/(a+b) - d/(c+d)$$

which equals $-\mathrm{PD}(D_{TEST} = \text{Sick})$

Here is a hypothetical example for coronary heart disease (CHD) and the result of a stress test:

Table 2.

		Stress test (effect)	
		Sick (positive test)	Not sick (negative test)
CHD (cause)	Sick	96	62
	Not sick	2,940	22,829

PR (Stress test = Sick): 5.3
PR (Stress test = Not sick): 0.4
OR (Stress test = Sick): 12
OR (Stress test = Not sick): 0.08 (=1/12)
PD (Stress test = Sick): 0.49
PD (Stress test = Not sick): – 0.49

So the effect is fairly strong by any measure, indicating a good test. On the probability ratio scale, the effect on a positive test (5.3) is somewhat stronger than the effect on a negative test when rescaled (2.5=1/0.4).

So far so good? Not really. Like most writers on this topic I assumed that we know the values of the variable D_{TRUTH}? Do we? Do we ever know – with logical certainty, I mean – the true value of any natural variable? No, we do not. Do we sometimes wholeheartedly believe that we have the right value, or that we almost always have the right value? Yes, we do. But science is not a poll of beliefs about true and false. Scientific methodology relies on precise, logical assertions.

Consider, again, the example of CHD and a stress test. A cardiologist may look at branches of the coronary arteries on angiography images and see plaques (CHD) or see no plaques (No CHD), but that's not an incontestable, correct classification of disease status. A plaque may be missed in a small branch and a small plaque may be as significant as a benign nevus. All those who tell you about "a gold standard" and "true disease status", against which to compare the quality of a test, are conveniently ignoring the fact that the values of D_{TRUTH} are beyond reach. We always compare the result of a test to the result of *another test*, which is thought to be a better measurement of disease status. We always compare one measurement to another. The causal structure behind our studies of the quality of a test is not $D_{\text{TRUTH}} \rightarrow D_{\text{TEST}}$. It is this:

Figure 1.

Where D^* is some "best" measurement of D_{TRUTH} such as the results of coronary angiography, tissue diagnosis of cancer, brain imaging, and the like.

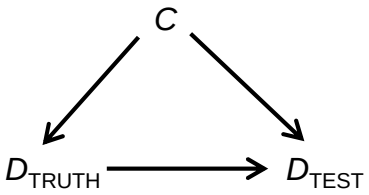
Again, oftentimes we strongly believe that the effect $D^* \leftarrow D_{\text{TRUTH}}$ is extremely strong, nearly deterministic, so the two variables are virtually identical. But we have no means to estimate that effect because the values of D_{TRUTH} remain unknown – forever. Strictly speaking, all that we may directly obtain is an estimate of the association between D^* and D_{TEST}. We are *always* at least one step removed from estimating the effect of interest. Information bias can never be excluded.

Other biases

If the quality of a diagnostic test is equated with the magnitude of the effect of D_{TRUTH} on D_{TEST}, we are facing the threat of other biases as well, when estimating that effect. I will focus on three: confounding, effect modification bias, and colliding bias. All three might interfere with our attempt to judge the quality of a test from a measure of association.

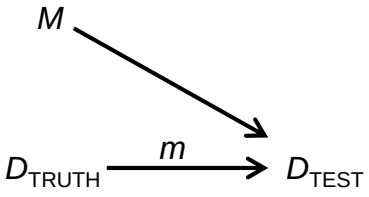
Confounding bias arises by shared causes of D_{TRUTH} and D_{TEST}, as shown below (Figure 2). If weight, for example, affects both CHD and the result of a stress test, the marginal association contains the unwanted contribution of a confounding path. To deconfound, we need to compute the association between D_{TRUTH} and D_{TEST} after conditioning on weight.

Figure 2.



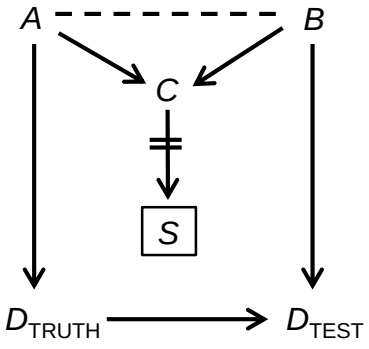
Effect modification bias arises when the effect of D_{TRUTH} on D_{TEST} is modified by some cause(s) of D_{TEST} (Figure 3), and we fail to estimate modified effects. For instance, the effect of CHD on the result of a stress test might differ between men and women (on some scale) – say, much stronger effect in men than in women. If we compute a marginal association on that scale, or a "gender-adjusted association", we have a biased estimator of the effect in one gender or both. Consequently, we do not have a good measure of the quality of the test in at least one gender.

Figure 3.



Lastly, colliding bias is also possible, albeit under unusual circumstances. The bias might arise because every study, including a study of the quality of a test, is conducted in a sample (*S*=selected), so conditioning on *S* is inevitable (Figure 4). If some selection criterion, *C*, shares one cause with D_{TRUTH} (*A*) and another cause with D_{TEST} (*B*), a path of colliding bias will be created (D_{TRUTH}←*A*---*B*→D_{TEST}).

Figure 4.



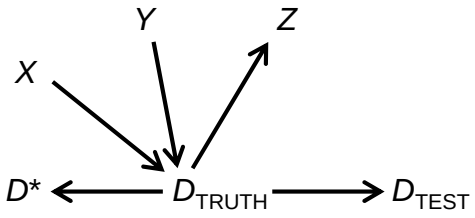
Predictions

The quality of a test, quantified by the effect D_{TRUTH}→D_{TEST}, is a property of a test. It should not be confused with the test's ability to predict (that is, help us guess) whether someone is diseased or not. As we'll see next, the domain of prediction – guessing the value of D_{TRUTH} from the value of D_{TEST} – is definitely connected with a causal structure, but it is not a matter of unbiased estimation of any effect.

Refer, again, to the basic diagram in Figure 1 (D^*←D_{TRUTH}→D_{TEST}). Since D^* is associated with D_{TEST}, it is possible to guess – not always correctly, of course – the value of D^* from the value of D_{TEST}. And that's exactly what positive predictive value, Pr(D^*=Sick|D_{TEST}=Sick) and negative predictive value, Pr(D^*=Not sick|D_{TEST}=Not sick) help us do. Or generically, we seek help from a conditional probability: Pr(D^*|D_{TEST}).

Good prediction, however, is rarely based on a single predictor. Notice, for instance, how poorly we guess CHD from a positive stress test in my hypothetical example of a good quality test: the probability of having CHD, given a positive test is only 0.03 (=96/3,036). In fact, it is unclear why anyone would always rely on D_{TEST} alone to guess the value of D_{TRUTH} (or actually D^*). Can't we sometimes meaningfully improve our guess by adding values of some easily measured causes (*X*, *Y*) or effects (*Z*) of D_{TRUTH} (Figure 5)?

Figure 5.



We surely can, at least sometimes. For instance, we may fit the regression model below to data from some sample and estimate the coefficients.

$$\ln \Pr(D^*=\text{Sick}) = \beta_0 + \beta_1 D_{TEST} + \beta_2 X + \beta_3 Y + \beta_4 Z$$

$$\Pr(D^*=\text{Sick}) = \exp(\beta_0 + \beta_1 D_{TEST} + \beta_2 X + \beta_3 Y + \beta_4 Z)$$

Then, we can use the coefficients, along with the values of the regressors, to compute the probability that a person is diseased, or will be diseased. This simple idea, widely known as risk prediction, has been regularly forgotten in writing about positive and negative predictive values of a test.

Actually, it was not entirely forgotten.

It is well known that the predictive values of a screening test depend on the prevalence of the disease in the screened population. For instance, the higher the prevalence, the larger the positive predictive value, which means a better guess of D_{TRUTH}=Sick from D_{TEST}=Sick. It is therefore prudent to screen a high-risk population, a population in which the prevalence of the disease is expected to be high.

But what is a "high-risk population" if not a population in which many people have unfavorable values of some causes of D_{TRUTH}=Sick (*X* and *Y* in Figure 5)? And what do these unfavorable values imply if not a higher probability of D_{TRUTH}=Sick (for many members of the population)? The screening of a high-risk population implicitly relies, albeit only crudely, on other predictors of D_{TRUTH}.

There is room for improvement, however, even when a high-risk population is screened. Instead of inferring D_{TRUTH}=Sick from D_{TEST}=Sick alone, we can also use the regression coefficients above to estimate the probability of D_{TRUTH} = Sick for people in a high-risk population, given D_{TEST} = Sick *and* their actual values of *X*, *Y*, and *Z* (Figure 5). That's the kind of "double dipping" about which no one should complain.

Epilogue

A fresh look at diagnostic tests reveals a tight connection to causal diagrams, estimation of effects, and prediction models. Hardly surprising – for anyone who understands that causality is the building block of reality; that thoughtful prediction is anchored in a causal structure; and that a causal diagram is an indispensable methodological tool.