Needles, Haystacks and p-values

By Chuck Dinerstein — May 8, 2018

Science and medicine often deal with issues that are about finding “the needle in the haystack.” We might get a better handle on p-values and science’s “reproducibility” problem by considering, metaphorically, those haystacks and needles.

In searching for that needle there are three inter-related factors.

- First, how many needles are in the haystack? Intuitively we know that the more needles there are, the easier they will be to find.

- Second, how much do the needles resemble hay? When there are significant differences, needles should be easier to find. But when the needle is disguised, and the difference between needle and hay is that one is a little sharper than the other, those needles may be more difficult to identify.

- Third, how good am I at sorting needles and hay? You know the first hour, or so, I should be spot on; but after a few hours of sorting, I am going to be less attentive and make mistakes. Sometimes I will put a needle in the hay pile and sometimes hay will wind up with the needles.

Epidemiology and other studies of healthcare interventions often are searching for those needles and are plagued by the same problems. And perhaps if we think about it as needles and haystacks, it might be easier to see their interactions. For example:

- For any given size haystack, the accuracy of my sorting will improve when needles are very different from hay. My stacks will also be more accurate if I am good at differentiating a
needle from hay.

- For any given degree of difference between needles and hay, I will have more accurate stacks if I am good at telling needles apart from hay and if there are lots of needles to be found in the haystack.

- Finally, for any given ability at sorting, I will get more accurate stacks when the difference between needles and hay is large and when there are lots of needles in the haystack to find.

There are epidemiologic and statistical terms for this hay and needles. The number of needles in the haystack is the prevalence of a disease. The differences in needle and hay are equivalent to the power of the test. My ability to sort needles from hay is described by sensitivity and specificity. Sensitivity is how often I can correctly identify only the needles; specificity is only correctly identifying the hay. And you can see how these qualities interact, the more different needles and hay appear (a test’s power), the more likely I can separate them and the higher my specificity and sensitivity will be. And the more needles are to be found (prevalence) the more likely that my sorting skills improve.

When screening for a disease or trying to identify a treatment’s effect, we are still looking for the needle. Consider mammography, the search in the haystack of women at risk for the needle of breast cancer. It is a very good screening test, correctly identifying the needle 87% of the time and correctly identifying hay 95% of the time.

So when I call telling you that I, unfortunately, found that needle indicating breast cancer, how often am I correct? That is the real question, you as my patient want to know, am I right, do you have cancer. Looking at the few numbers I have offered you might guess I am right at least 87% of the time although most people, and I include a majority of physicians would say 95%. But both of those answers are incorrect, in reality, I am right about 70% of the time, and I unnecessarily upset 1 out of the three patients I call. For the mathematically inclined the calculations can be found in the footnote.[1]

Part of science’s reproducibility problem is the confusion between what I say and what I mean, in this case confusing statistical and clinical significance. The ability to correctly identify hay, at a p-value of 0.05 is statistically significant, but what I mean, what is clinically significant, is that I am right about 70% of the time. Which of those choices, what I say or what I mean instills more confidence?

We think that significance in matters of epidemiology and healthcare reside among sensitivity and p-values, correctly identifying the hay – in what the test says, when instead we should recognize that significance, clinically relevant significance is associated with another statistical term, the positive predictive value – what the test means.
[1] Using a population of 10,000, incidence of 12.4%, sensitivity of 87% and specificity of 95% we can do the following calculations:

Patients in population with cancer = (Population) (Incidence) = 1240

Patients in population without cancer = (Population) (1-Incidence) = 8760

Patients correctly identified as without cancer are True Negatives = (Population without cancer) (Specificity) = 8322

Patients correctly identified with cancer are True Positives = (Population with cancer) (Sensitivity) = 1079

Patients incorrectly identified as having cancer are False Positives or False Discovery = Population without cancer – True Negatives = 438

Patients incorrectly identified as not having cancer are False negatives = Population with cancer – True Positives = 161

Positive Predictive Value = True Positives/(False Positives + True Positives) = .71 or 72% of the time.

False Discovery = 1-Positive Predictive Value = .29 or 29% of the time