DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
PUBLIC HEALTH SERVICE

NATIONAL INSTITUTES OF HEALTH
BETHESDA, MD. 20014
AREA CODE 301 TEL: 656-4000

Patty A. Keys
Tonis

July 17, 1970

Dr. Henry Blackburn
Laboratory of Physiological Hygiene
University of Minnesota
Stadium Gate 27
Minneapolis, Minnesota 55455

Dear Henry:

So far as could be told on the numbers involved the estimates of prevalence of an ECG abnormality would be essentially the same under both readings. There is some reassurance in this but not much.

Suppose you were to read ECGs from two populations. Suppose these ECGs were divided into only two categories, positive and negative. Suppose the proportion of positives differed in the two populations. In general

(1) The estimate of the proportion of positives in each population would be biased.

(2) The estimate of the difference between the two populations in the proportion of positives would be biased.

It may be instructive to pursue this simple model a bit further. Suppose

p = the probability of classifying a true positive as positive
r = the probability of classifying a true negative as positive
x = the proportion of true positives
y = the proportion of ECGs read as positive

then $x = \frac{y-r}{p-r}$, $y = px + r(1-x)$.

The interesting thing about this result is that in order to estimate the true prevalence of positive ECGs you would need exceedingly accurate estimates of the parameters p and r. Since these are not really constants and each has a probability distribution of its own

the problem (even in this simple model) is really a very sticky one.

Suppose, however, you were to consider the parameters as constants and suppose you wished to estimate the difference in prevalence between two groups. If you were to use the readers' estimates of prevalence then you would get as an estimate of the difference in prevalence between the two populations

$$y_1 - y_2 = (p-r)(x_1 - x_2)$$

That is, the estimate of the difference would be too small.

If you treat the problem a little more realistically it quickly gets out of hand. If ECGs are re-read a large number of times you will find that certain positives are almost always read as positives while others are read positive a much smaller proportion of the times. The same is true of negatives. In fact the problem is substantially more complicated with the negatives. For example, if you wanted to estimate the prevalence of characteristic 1.1, a 1.2 ECG is a negative and so is a 9.1 and for each negative class there is a different probability of being classified as 1.1. Hence a different distribution of types of positives and types of negatives is bound to alter the prevalence estimate you get.

The problem up to this point appears to me relatively simple in principle although it could involve some difficulties in practice. Our own situation leaves me more concerned. Essentially we are using the prevalence readings to designate, in a uniform fashion, a population free of CHD, in order to have a uniform basis for characterizing incidence. Suppose that you have two populations with substantially different prevalence levels of CHD. Suppose that the probability of classifying a true positive as negative is fairly high as some of the replicate readings suggest. Then, hidden in your "normal" populations are a fair number of people with CHD. Such people have a substantially elevated risk of a new episode of CHD and are therefore going to appear as incidence cases more frequently than persons who are truly negative. The result would be (I guess) that the estimate of the difference in incidence between these two populations would appear larger than it truly is.

All I am doing here is just poking at the problem. Sooner or later, I guess, we'll have to come to closer grips with it.

Sincerely yours,

Tavia Gordon
Supervisory Statistician
Biometrics Research Branch
National Heart and Lung Institute