POPULATION MEDICINE NEWS

Epidemiology, Preventive Medicine, Public Health
Production Medicine, Computer Applications in Vet Med

September 16, 1991 4(22)

CRITICAL EVALUATION OF THE SCIENTIFIC LITERATURE

Evaluating Precision and Power

Biomedical sciences are fundamentally empirical; all our knowledge comes from data. The trouble with data is that it never lets us see truth, only fuzzy reproductions of it. Special tools are required to avoid being misled by authors who are not aware of the concepts of precision and power.

The heart of the Scientific Method is the model and the hypotheses that we derive from it. Consider the simplest model:

$$Y_i = U + BX_i + E$$

where U stands for an overall population average; Y stands for the dependent variable--the thing that we would like to have an effect on (e.g., morbidity or some measure of performance. The subscript $_i$ indicates the value of Y for the ith animal). X stands for the independent variable--the thing we are manipulating to determine its effect on Y (e.g., a particular vaccination or treatment). B is the effect estimate associated with X (e.g., the amount this vaccine reduces morbidity). E stands for error--the unexplained noise inherent in any complex system.

Error comes mainly from imperfect measurement and from the action of extraneous variables, known and unknown, that influence Y (Fig 1). E interacts with B in several deleterious ways one of which is imprecision which blurs our vision making it difficult to resolve the magnitude of B. The power of a study is the ability to avoid imprecision and thus to resolve treatment effects that truly exist.

To illustrate the concept of precision, assume we want to determine the effect of a particular vaccination program on average daily gain (ADG) in a cattle operation. Just for the sake of argument, assume the difference in ADG between vaccinates and controls is exactly 0--that is, vaccinates and non-vaccinates have exactly the same ADG. Do we expect to observe an effect of exactly 0 in our trial? No, because animal to animal variation exists in ADG and, even under the assumption that treatment has no effect, we are unlikely to get animals with precisely identical ADG's in 2 groups. How large does the observed effect have to be to convince us that it's real and not just due to chance? That depends on sample size.

The 3 curves in Fig. 2 represent the sampling distributions expected from trials comparing ADG (assuming a standard deviation of 0.3 lb/day) under the assumption of no treatment effect (B=0). A sampling distribution gives the probability of observing a particular difference in ADG by "chance" even though there is no true difference. Note that the distribution becomes narrower with increasing sample size: we are unlikely to observe a large difference, say .15 lb/day, just by chance in the trial with 500 animals per group. A difference of even .20 lb/day could occur by chance in the trial with 10 animals per group.

<u>Statistical significance</u>. Consider now the more practical situation where we don't know whether or not an effect of a vaccine on ADG exists. If we perform a trial with 10 animals per group and observe a difference of, say, .15 lb/day in favor of vaccinates, what is our conclusion? Since the observed difference is well within the range of differences that we might expect even if no true difference existed (Fig. 2), we would have to conclude that we have insufficient evidence to conclude that vaccine has a beneficial effect. In statistical parlance, we would say that the observed difference was not significant since one as large or larger than it <u>could</u> have occurred "by chance" even if no true difference existed. This might be denotes as so: ($P > .10$) meaning that there was a greater than 10% probability that differences as large as that observed could have occurred by chance even if no true effect exists. If, on the other

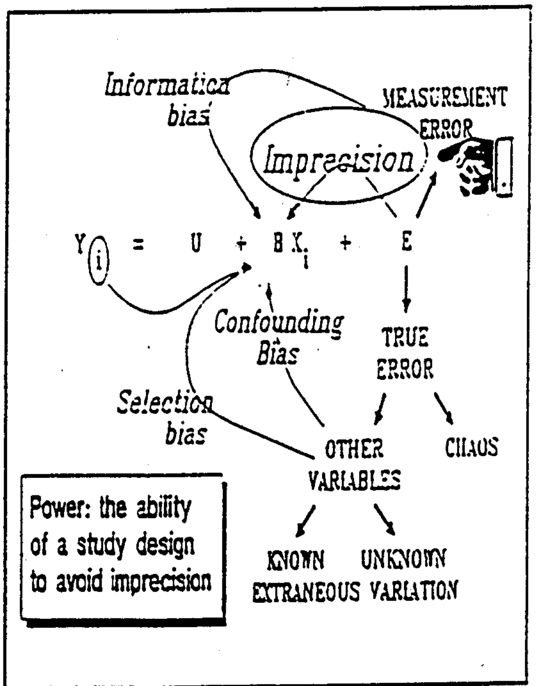

Figure 1. Error affects the precision and validity of effect estimates. Precision is a direct effect of error. Validity is determined by freedom from the 3 forms of bias shown.

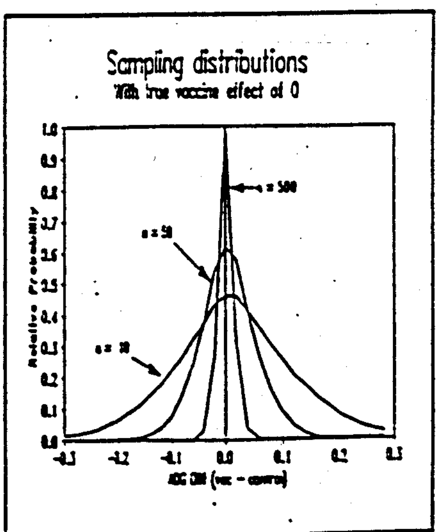

Figure 2. Large "chance" differences can occur in small sample size studies; such studies cannot resolve important differences and are thus said to have low power.

95% CONFIDENCE INTERVAL
Difference between 2 means

$$SE = \text{standard deviation} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

95% confidence interval of difference = Observed difference +/- 2 X SE

Example:
Observed difference = .15 lb/day ADG
SD = .3 lb/day
$n_1 = n_2 = 50$

$$SE = .3 \times \sqrt{\frac{1}{50} + \frac{1}{50}} = .06$$

.15 +/- 2 X .06

.03 to .27 95% confidence interval

Figure 3. Method for calculating 95% confidence interval for the difference between 2 means.

hand, we observed a difference of .15 lb/day in a trial with 500 animals per group, we could confidently conclude that a true vaccine advantage does exist ($P < .001$). Even for the trial with 50 animals per group, a difference of .15 lb/day is unlikely ($P < .05$) and thus would be statistically significant.

Effect "not significant". Suppose we are reading a report of a vaccine trial with 10 animals per group that reports a difference in ADG between 2 groups of 0.15 lb/ day ($P > .10$, not significant). Choose 1 of the following: (a) The difference was produced by chance; there is greater than 95% probability that no true effect exists. (b) The difference could have been produced by chance; but .15 lb/day is below the resolving power of a study with only 10 animals per group, thus there could be a true difference that escaped detection. (c) If a larger sample size is used, a significant effect will be found for this vaccine.

The answer is b. In small sample size trials (those with low power), a non-significant difference provides evidence for neither the existence or non-existence of a true treatment effect. Understanding this seeming paradox--that a study can fail to provide evidence in either direction--is fundamental to the correct interpretation of the scientific literature. A metaphor will help us understand and remember it. Suppose you are looking for BVD virus in nasal secretions. You hold up a petri dish of secretions to the light and examine it with your naked eye. You say: "I cannot see any BVD virus." Does this constitute evidence for or against the presence of BVD virus in the sample? No, because the power of the observing instrument is insufficient to the task. Appropriate statistical power must be selected for a study in the same manner as the appropriate power of the magnifying instruments that will be used. Ideally, scientific researchers would plan the power of their study designs so as to protect readers from low power studies. But most don't, so we'll have to do it ourselves in using confidence intervals to be demonstrated shortly.

Estimating magnitude of effect (3). Let's go back to the feedlot example for a moment and assume that in a trial with 50 animals per group we observe a difference in ADG of .15 lb/day in favor of vaccinates. We are thinking about buying a few hundred thousand doses but would like to do some cost accounting to see if it will pay. Can we count on the true effect being .15 lb/day? No, as any good empiricist would say, "The exact effect unknowable." The best we can do is put a bound of uncertainty around the estimate, the width of which depends on the resolving power of the study design. In Fig. 3, the formula is shown for computing the 95% confidence interval for the difference between 2 means, and the confidence interval is computed for the present example. Note that, with a sample size of 50 per group, the interval is wide. It is quite possible that the vaccine effect on ADG could be as low as .03 lb/day or as high as .27 lb/day. Although we can never know the exact effect, we are 95% certain that the true effect is somewhere within this bound. How do we get a more precise estimate? Provide for a larger sample size. If a difference of .15 lb /day ADG was observed in a 500 per group vaccination trial, the 95% confidence interval would be .11 to .19 lb/day ADG.

These computations depend on an estimate of the standard deviation. This will commonly be reported separately for each group. To estimate the pooled standard deviation, take a weighted average of the standard deviations for the groups. If reported, the square root of the mean square error also provides an estimate of standard deviation.

Computing confidence intervals is a pastime you may choose to forgo where significant differences are reported, but it is essential in studies where differences are found "not significant." Consider, for example, the confidence interval for an observed difference in ADG between vaccinates and controls of .15 lb/day in a 10 animal per group trial:

-.11 to .41 lb/day.

95% CONFIDENCE INTERVAL
Difference between 2 proportions

$$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

95% Confidence interval of difference = Observed difference +/- 2 X SE

Example:

	Morbidity	n
Vaccinates	.20	50
Controls	.30	50

$$SE = \sqrt{\frac{.20(1-.20)}{50} + \frac{.30(1-.30)}{50}}$$

.10 +/- 2 X .086

-.07 to .27 95% confidence interval

Figure 4. Method for calculating 95% confidence interval for the difference between 2 proportions.

Note that this interval overlaps 0 which reflects the lack of statistical significance--that a difference as large as .15 lb/day could be produced "by chance" in a 10 animal per group trial even if no true difference existed. However, what is usually missed by authors of low power studies and sometimes by their readers, is that the observed data are also compatible with a large and important effect. The observed data are as compatible with a .30 lb/day difference as with a 0 difference; and a true difference as large as .41 lb/day is plausib[le]. Most people would count a .30 lb/ increase in ADG as very important. Thus, equating "no significant" difference as evidence of lack of an important difference would be erroneous in this example or in any low power study. The only legitimate thing we can say after reading a low power study is the same as we'd say to the fellow looking for virus with his naked eyes: "Get real, Bozo."

Fig 4 shows the formula for computing the confidence interval for the difference between 2 proportions. Though it may seem a lot of trouble to use such formulae (really just <2 minutes), they are the only protection we have against the low power studies which are, unfortunately, very common in the scientific literature.

POPULATION MEDICINE NEWS is produced by the Washington State University Field Disease Investigation Unit, Pullman, WA 99164-6610 (509-335-0711) and is supported in part by the WSU Cooperative Extension Service. Viewpoints expressed are not necessarily those of the Univers[ity] or its Administration. newsletter may be copied by anyone. Editors: Dale Hancock and Susan Waller.